M

MAC in Cognitive Radio Networks

Juncheng Jia
Soochow University, Suzhou, China

Synonyms

Media access control protocol for cognitive radio networks

Definitions

In cognitive radio networks, the media access control (MAC) protocols need to determine the availability of spectrum resource through spectrum sensing and coordinate with the other secondary users (SUs) for spectrum access, so that the spectrum resource is efficiently utilized without harmful interference to the primary users (PUs).

Historical Background

Cognitive radio has been first proposed in Mitola and Maguire (1999) and Mitola (2000) as a way to promote the efficient use of the spectrum by exploiting spectrum opportunities. SUs with cognitive radios are able to exploit information about the wireless environment and adapt their transmission parameters to access available channels while avoiding harmful interference to PUs (Haykin 2005).

In general, there are three transmission schemes for cognitive radios: underlay, overlay, and interweave (Goldsmith et al. 2009). In underlay scheme, SUs are allowed to transmit signal as long as the generated interference stays below a certain threshold. In 2003 the FCC defined the interference temperature as a way to measure and limit the interference at PUs. However, this model has been abandoned by the FCC in 2007 due to the implementation issues. In overlay scheme, SUs exploit the knowledge of PUs' messages to either limit the interference. In interweave scheme, SUs can only transmit in spectrum holes; whenever a SU detects PUs, the SU vacates its channel to avoid harmful interference to PUs.

In order to realize the functions of cognitive radios, MAC should possess the ability of channel sensing, resource allocation, spectrum sharing, and spectrum mobility (Akyildiz et al. 2006). Channel sensing is the ability of a cognitive radio to collect information about spectrum usage. Resource allocation is employed to opportunistically assign available channels to SUs according to QoS requirements. Spectrum sharing deals with contentions between heterogeneous PUs and SUs in order to avoid harmful interference. Spectrum mobility allows a SU to vacate its channel, when a PU is detected, and to access an idle band where it can reestablish the communication link. The research works

X.(S.) Shen et al. (eds.), *Encyclopedia of Wireless Networks*,
https://doi.org/10.1007/978-3-319-78262-1

of cognitive radio MAC mainly focus on these aspects, which is reviewed in several surveys (Krishna and Amitabha 2009; De Domenico et al. 2012; Cormio and Chowdhury 2012). Apart from the close coupling of physical layer and MAC layer, there also exist interactions between the network and transport layers with MAC layer, such as joint spectrum and route decisions, congestion-free end-to-end reliability, spectrum and node mobility, etc. (Akyildiz et al. 2009).

On the standardization side of cognitive radio, IEEE 802.22 represents the major effort, which was started in 2004 and finalized in 2011 (Cordeiro et al. 2005). 802.22 is designed for wireless regional area networks (WRANs) which takes advantage of the favorable transmission characteristics of the VHF and UHF TV bands to provide broadband wireless access over a large area up to 100 km from the transmitter. 802.22 MAC uses time division multiplexing, with structures of superframe and frame, where a superframe is composed of multiple MAC frames preceded by the frame preamble. In addition to 802.22, the IEEE has standardized another white space cognitive radio standard, 802.11af, which was approved in 2014 (Flores et al. 2013). 802.11af is a wireless LAN standard designed for ranges up to 1 km. 802.11af uses carrier sense multiple access with collision Avoidance (CSMA/CA)-based MAC protocol.

Foundations

Due to the unique requirement of protection for PUs, the design of cognitive radio MAC protocols is different from MAC protocols of most of the other wireless systems. The cognitive radio MACs are based on the closed coupling with the physical layer and the hardware support on the radio device.

Spectrum sensing is one of the key enabling functions in cognitive radio networks that is used to explore vacant spectrum opportunities and to avoid interference with the PUs. While the physical layer of spectrum sensing is responsible for the signal processing work, the MAC layer is more involved in the higher-level functions of spectrum sensing coordination. For example, MAC layer needs to optimize sensing and transmission durations, the order in which channels are sensed, and the number of channels to be sensed before transmission (Kim and Shin 2008; Lee and Akyildiz 2008; Jia et al. 2008). Furthermore, the reliability of spectrum sensing can be improved if SUs share their sensing results with neighbors due to multiuser spatial diversity, which is called cooperative spectrum sensing. However, the improvement of spectrum sensing comes at the cost of increased latency and communication overhead, where the MAC layer tries to coordinate multiple SUs and reduce the cost.

Since cognitive radio networks usually use multiple channels, control channel establishment needs to be carefully considered, especially for distributed networks. Some MAC designs assume there exists a global control channel available for all SUs, which simplifies the implementation (Ma et al. 2005). When a global control channel is not available or reliable, local control channels can be used, since it is possible that neighboring SUs may share some commonly available channels (Zhao et al. 2005; Chen et al. 2005). Besides, some designs do not use a dedicated control channel at all, which is based on channel hopping. The coordination of channel hopping in such context is usually referred to as rendezvous (Lin et al. 2011; Bian et al. 2011).

Spectrum allocation is an important function of the MAC protocol in cognitive radio networks. Since the spectrum availability and transmission requirement at SUs could be different, SUs have to share their spectrum availability information with neighbors and conduct efficient spectrum allocation to maximize certain utility. MAC protocols exploit advanced optimization algorithms to realize intelligent, fair, and efficient allocation of the available spectrum. Each SU adapts its transmission parameters to changes of the wireless environment, in order to efficiently exploit the available resource. Considering the high complexity of centralized optimization, decentralized approaches in which each SUs acts based on partial knowledge of network status have been

proposed based on graph coloring theory (Zheng and Peng 2005), game theory (Wang et al. 2010), stochastic theory (Zhao et al. 2007), etc.

Spectrum access enables multiple SUs to share the spectrum resource by determining who will access the channel or when a user accesses the channel. Contentions between SUs can be avoided through coordinated access both in centralized and distributed architectures. When coordination is absent, a random approach could be exploited to contend for access to available channels (Zhao et al. 2007). Otherwise, SUs may exchange signalling messages in order to reserve the access to a data channel (Su and Zhang 2008). The control handshaking mechanism, however, does not completely solve the hidden terminal problem; hence the busy tone scheme is often exploited to prevent hidden nodes (Ma et al. 2005).

When a PU is detected, to realize seamless transmission, a cognitive radio vacates its channel and reconstructs a transmission link on a different channel. The procedure that permits this transition from a channel to another with minimum performance degradation is called handoff. Spectrum mobility in MAC layer is to reduce delay and loss during spectrum handoff. In general, there are two different strategies for spectrum mobility: proactive spectrum handoff and reactive spectrum handoff (Akyildiz et al. 2009). In proactive spectrum handoff, SUs predict future activity in the current link and determine a new spectrum while maintaining the current transmission and then perform spectrum switching before the link failure happens. For proactive spectrum handoff, the spectrum switching is faster but requires complex algorithms. On the other hand, in reactive spectrum handoff, SUs perform spectrum switching after detecting link failure due to spectrum mobility. This method requires immediate spectrum switching without any preparation time, resulting in significant quality degradation in ongoing transmissions. It is also possible to design a hybrid spectrum handoff strategy, which combines proactive strategy and reactive strategy by applying proactive spectrum sensing and reactive handoff action (Christian et al. 2012).

Key Applications

MAC protocol is a fundamental component for cognitive radio networks.

Cross-References

- Cognitive Heterogeneous Networks
- Millimeter Wave MAC Layer
- QoS-Aware MAC

References

Akyildiz IF, Lee WY, Vuran MC, Mohanty S (2006) NeXt generation/dynamic spectrum access/cognitive radio wireless networks: a survey. Comput Netw J 50:2127–2159

Akyildiz I, Lee W, Chowdhury K (2009) CRAHNs: cognitive radio ad hoc networks. Ad Hoc Netw 7:810–836

Bian K, Park J, Chen R (2011) Control channel establishment in cognitive radio networks using channel hopping. IEEE J Sel Areas Commun 29:689–703

Chen T, Zhang H, Maggio G, Chlamtac M (2005) CogMesh: a cluster-based cognitive radio network. In: IEEE DySPAN, pp 168–178

Christian I, Moh S, Chung I, Lee J (2012) Spectrum mobility in cognitive radio networks. IEEE Commun Mag 50:114–121

Cordeiro C, Challapali K, Birru D, Shankar S (2005) IEEE 802.22: the first worldwide wireless standard based on cognitive radios. In: IEEE DySPAN, pp 328–337

Cormio C, Chowdhury K (2012) A survey on MAC protocols for cognitive radio networks. Ad Hoc Netw 7:1315–1329

De Domenico A, Strinati EC, Di Benedetto MG (2012) A survey on MAC strategies for cognitive radio networks. IEEE Commun Surv Tutor 14:21–44

Flores A, Guerra R, Knightly E, Ecclesine P, Pandey S (2013) IEEE 802.11 af: a standard for TV white space spectrum sharing. IEEE Commun Mag 51:92–100

Goldsmith A, Jafar SA, Maric I, Srinivasa S (2009) Breaking spectrum gridlock with cognitive radios: an information theoretic perspective. Proc IEEE 97:894–914

Haykin S (2005) Cognitive radio: brain-empowered wireless communications. IEEE J Sel Areas Commun 23:201–220

Jia J, Zhang Q, Shen X (2008) HC-MAC: a hardware constrained cognitive MAC for efficient spectrum management. IEEE J Sel Areas Commun 26:106–117

Kim H, Shin K (2008) Efficient discovery of spectrum opportunities with MAC-layer sensing in cognitive radio networks. IEEE Trans Mob Comput 7:533–545

Krishna TV, Amitabha D (2009) A survey on MAC protocols in OSA networks. Comput Netw 9:1377–1394
Lee W, Akyildiz I (2008) Optimal spectrum sensing framework for cognitive radio networks. IEEE Trans Wirel Commun 7:845–857
Lin Z, Liu H, Chu X, Leung Y (2011) Jump-stay based channel-hopping algorithm with guaranteed rendezvous for cognitive radio networks. In: IEEE INFOCOM, pp 2444–2452
Ma L, Han X, Shen CC (2005) Dynamic open spectrum sharing MAC protocol for wireless ad hoc networks. In: IEEE DySPAN, pp 203–213
Mitola J (2000) Cognitive radio: an integrated agent architecture for software defined radio. PhD thesis, Royal Institute of Technology (KTH), Stockholm
Mitola J, Maguire G (1999) Cognitive radio: making software radios more personal. IEEE Pers Commun 6:13–18
Su H, Zhang X (2008) Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks. IEEE J Sel Areas Commun 26:118–129
Wang B, Wu Y, Liu K (2010) Game theory for cognitive radio networks: an overview. Comput Netw 54:2537–2561
Zhao J, Zheng H, Yang GH (2005) Distributed coordination in dynamic spectrum allocation networks. In: IEEE DySPAN, pp 259–268
Zhao Q, Tong L, Swami A, Chen Y (2007) Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: a POMDP framework. IEEE J Sel Areas Commun 25:589–600
Zheng H, Peng C (2005) Collaboration and fairness in opportunistic spectrum access. In: IEEE ICC, pp 3132–3136

Machine Learning

▸ Application of Machine Learning in Wireless Sensor Network
▸ Big Data in 5G
▸ Extending WSN Lifetime Based on Evolutionary Clustering Algorithm
▸ Machine Learning in Wireless Sensor Networks for the Internet of Things

Machine Learning Algorithms

▸ Machine Learning Paradigms in Wireless Network Association

Machine Learning in Wireless Sensor Networks for the Internet of Things

Abdallah Jarwan[1], Ayman Sabbah[2], and Mohamed Ibnkahla[3]
[1]Carleton University, Ottawa, ON, Canada
[2]Spectrum and Telecommunication Sector (STS) - Innovation, Science, and Economic Development (ISED)/Government of Canada, Ottawa, ON, Canada
[3]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada

Synonyms

Artificial intelligence; Machine learning

Definition

In usual ways of programming, a program is built by building instructions in order to reach desired outputs from inputs. However, in machine learning (ML), that process is flipped. The inputs and desired set of outputs are given, and the program should learn what instructions or policy should be followed. ML is concerned with algorithms that observe data, learn from it, grow up, and make more intelligent decisions. Based on the way that machines accumulate knowledge and become able to function as needed, ML can be classified into supervised, semi-supervised, unsupervised, and reinforcement learning.

Historical Background

Internet of Things (IoT) systems are built on thousands of wireless sensor networks (WSNs). WSNs suffer from various limitations such as finite energy, hardware inaccessibility, uncontrolled environment, uncontrolled topology, limited computational power and storage capability,

and heterogeneity in systems. Recent developments of WSN technologies have targeted energy efficiency, lifetime, coverage, connectivity, traffic balancing, spectrum management, fault tolerance, latency, reliability, and security, among others. These targets can be improved by considering controlled deployment and mobility, routing, scheduling, power and rate control, clustering, data aggregation, in-network processing, and traffic control. One of the most important aspects of WSNs and IoT systems is traffic control through traffic reduction. Traffic reduction is done by reducing the amount of data or packets that need to be transferred across WSNs and various IoT layers. This improves IoT systems in terms of energy and spectrum, which are the main resources, and increases lifetime. Traffic control can be implemented in IoT systems using machine learning (ML) algorithms.

ML algorithms observe data, learn from it, grow up, and make intelligent decisions. Through training, machines build policies to perform complex tasks that mimic intelligent human behavior. ML tools have the potential to provide lower-complexity alternatives for sophisticated algorithms that consume time and resources. Also, ML makes it easy to deduce hidden correlations, understand context, and make predictions from the collected data. It also provides autonomous and intelligent decision making schemes with minimal human intervention. The importance of applying ML approaches in IoT systems and WSNs is mainly due to the dynamicity and uncontrollability of the surrounding environment (Alsheikh et al. 2014). Deep neural networks (DNNs) and long short-term memory networks (LSTMs) are supervised ML techniques. They are used to improve WSNs in various applications. In this chapter, both will be studied and proposed as a key application for traffic reduction in IoT systems.

The remainder of this chapter is organized as follows. First, mathematical computational graphs of DNNs and LSTMs are used to illustrate how they work. After that, their applications to WSNs and IoT systems are studied. Finally, both techniques are used to improve traffic reduction through IoT systems.

Foundations

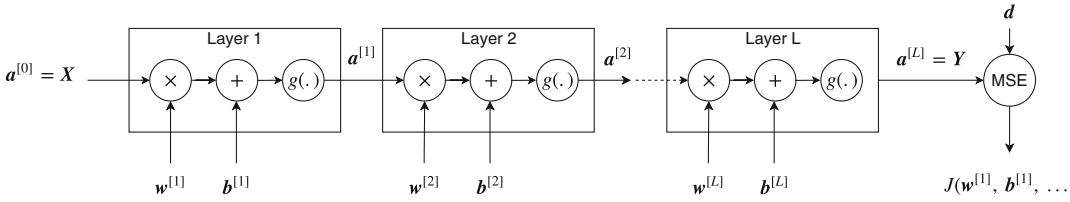
Deep Neural Networks (DNNs)

Deep learning is a machine learning technique where learning is done through multiple levels of representations (Chen and Lin 2014). Deep neural networks (DNNs) mimic how humans' brains work in the sense that they are composed of many abstraction layers. Each layer draws conclusions based on the previous layer outputs. They are widely used in feature extraction, modeling complex relationships between inputs and outputs, and context understanding.

DNNs are represented in computational graphs as shown in Fig. 1. The illustrated DNN consists of L layers. The lth layer takes an $n^{[l-1]} \times 1$ vector $\boldsymbol{a}^{[l-1]}$ and outputs another $n^{[l]} \times 1$ vector $\boldsymbol{a}^{[l]}$ that goes as an input to the next layer, where $n^{[l]}$ is the number of neurons at the lth layer. The output of each layer is computed from its input. In the figure, $\boldsymbol{w}^{[l]}$ is the lth layer $n^{[l]} \times n^{[l-1]}$ weight matrix, $\boldsymbol{b}^{[l]}$ is its $n^{[l]} \times 1$ bias, and $g(.)$ is an element-wise activation function. The activation function $g(.)$ can be Sigmoid, tanh, rectified linear unit (ReLU) functions, among others.

As a supervised learning approach, the DNN is trained using a labeled data set. The goal of building a DNN is to map a new input $\boldsymbol{X}$ to an output $\boldsymbol{Y}$ based on the labeled inputs. During this phase, a cost function $J(\boldsymbol{w}^{[1]}, \boldsymbol{b}^{[1]}, \boldsymbol{w}^{[2]}, \boldsymbol{b}^{[2]}, \ldots)$ is defined in terms of outputs $\boldsymbol{Y}$ and labels $\boldsymbol{d}$ as shown in Fig. 1. For example, the cost can be defined as the mean square error (MSE).

The DNN output $\boldsymbol{Y}$ is computed using the computational graph in Fig. 1. This process is called forward propagation. During training, the gradients $\mathrm{d}\boldsymbol{w}^{[l]} = \partial J/\partial \boldsymbol{w}^{[l]}$ and $\mathrm{d}\boldsymbol{b}^{[l]} = \partial J/\partial \boldsymbol{b}^{[l]}$ at every layer are computed. After that, they are used to update (tune) the weights and biases in a process that is called backward propagation. Weights and biases updates are done using various iterative optimization algorithms such as gradient descent (GD) or Adam optimizer. In GD algorithm, the updates are done using $\boldsymbol{w}^{[l]} := \boldsymbol{w}^{[l]} - \alpha \mathrm{d}\boldsymbol{w}^{[l]}$ and $\boldsymbol{b}^{[l]} := \boldsymbol{b}^{[l]} - \alpha \mathrm{d}\boldsymbol{b}^{[l]}$, where := denotes the update operator and α represents the learning rate. Gradients and updates are done



Machine Learning in Wireless Sensor Networks for the Internet of Things, Fig. 1 DNNs computational graph architecture

numerically using frameworks such as TensorFlow (Rampasek and Goldenberg 2016). It is emphasized that TensorFlow is optimized to calculate gradients because it is based on computational graphs where the chain rule is used to compute derivatives.

Long Short-Term Memory Networks (LSTMs)

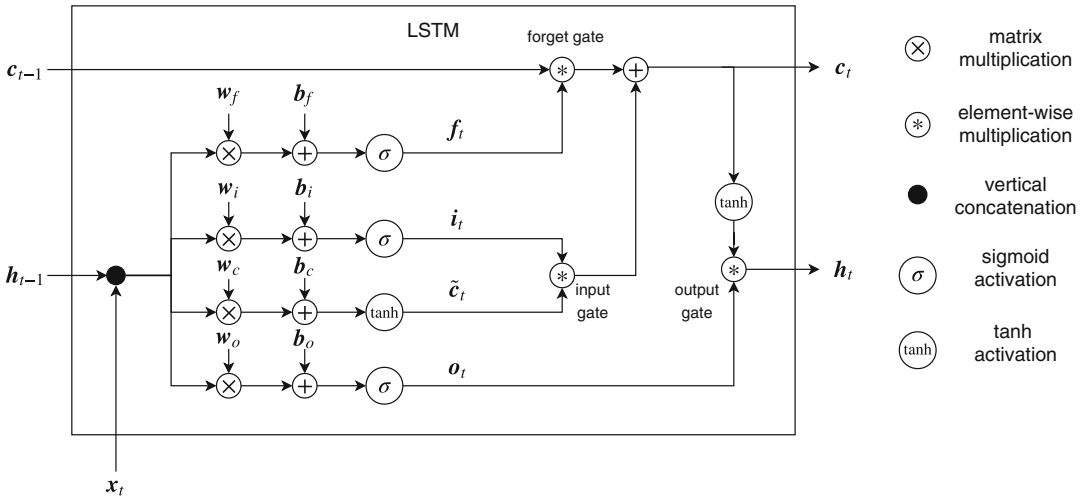
Long Short-Term Memory networks (LSTMs) represent a special kind of recurrent neural networks (RNNs). LSTMs are good in handling sequence dependencies where the inputs are represented in terms of time, such as in prediction and language processing problems. An LSTM network is composed of a single LSTM cell that has a feedback connection from its output to its input. The structure of an LSTM cell is shown in Fig. 2. It is shown that at any time instance t, the LSTM cell input is composed of the current input ($\boldsymbol{x}_t$) and the outputs from the previous time instance ($\boldsymbol{h}_{t-1}$ and $\boldsymbol{c}_{t-1}$). LSTMs can be mathematically considered as multiple cells that are connected in cascade as shown in Fig. 3. However, it differs from DNNs in the sense that all LSTM cells in the cascade connection use the same weights and biases.

The LSTM cell contains four NNs that are connected to form three gates: forget (f), input (i), and output (o) gates as shown in Fig. 2. Each NN comprises one neurons layer with H number of neurons. The weights and biases of these NNs are denoted by $\boldsymbol{w}_f$, $\boldsymbol{b}_f$, $\boldsymbol{w}_i$, $\boldsymbol{b}_i$, $\boldsymbol{w}_c$, $\boldsymbol{b}_c$, $\boldsymbol{w}_o$, and $\boldsymbol{b}_o$. In general, during the training phase, these variables are updated so that the LSTM performs as required. Each LSTM cell has a state $\boldsymbol{c}_t$ and an output $\boldsymbol{h}_t$ at time instance t. Both cell state and output are fed to the LSTM in the next time instance $t + 1$. The new cell state and output are calculated from previous state, previous output, and current input using the computational graph in Fig. 2. The cost function and variable tuning (backward propagation) can be done similar to what has been illustrated in DNNs.

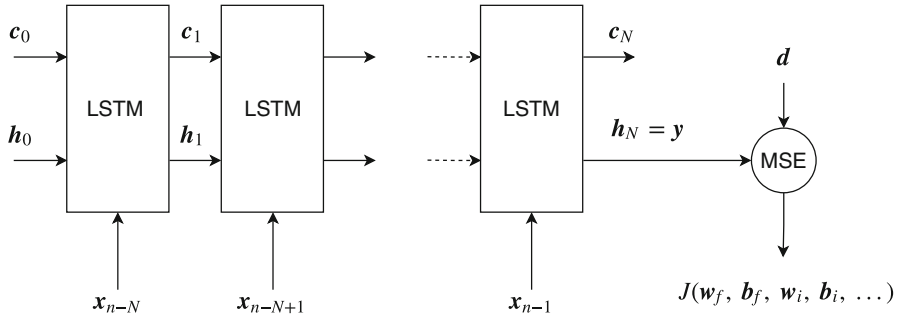
Key Applications

Applications of DNNs and LSTMs in IoT Systems

DNNs are used to tackle many applications in IoT (Chen et al. 2017). For instance, DNNs are powerful for big data analytics (Mohammadi et al. 2018). Therefore, they are used to extract patterns and conclusions from data gathered by IoT devices. For example, DNNs can be used in context awareness applications such as human activity recognition in wearable WSNs. Moreover, DNNs are used to make predictions using simple information as input. Also, they can be used for classification problems which can be applied to many aspects in WSNs and IoT. Furthermore, DNNs have many applications in automation and building self-organizing networks which are the essence of IoT systems. The recent literature has proposed many applications of DNNs in applications related to IoT systems. In Bande and Shete (2017), DNNs are used for flood management in IoT systems to predict floods. The proposed framework has been applied to data collected from Chennai region. In Kumar and Jain (2018), localization in IoT scenarios has been implemented using DNNs. Localization is critical to many application where nodes location should be monitored, while GPS devices are not feasible to be used. DNNs are



Machine Learning in Wireless Sensor Networks for the Internet of Things, Fig. 2 Computational graph of a single LSTM cell



Machine Learning in Wireless Sensor Networks for the Internet of Things, Fig. 3 Prediction architecture using LSTMs

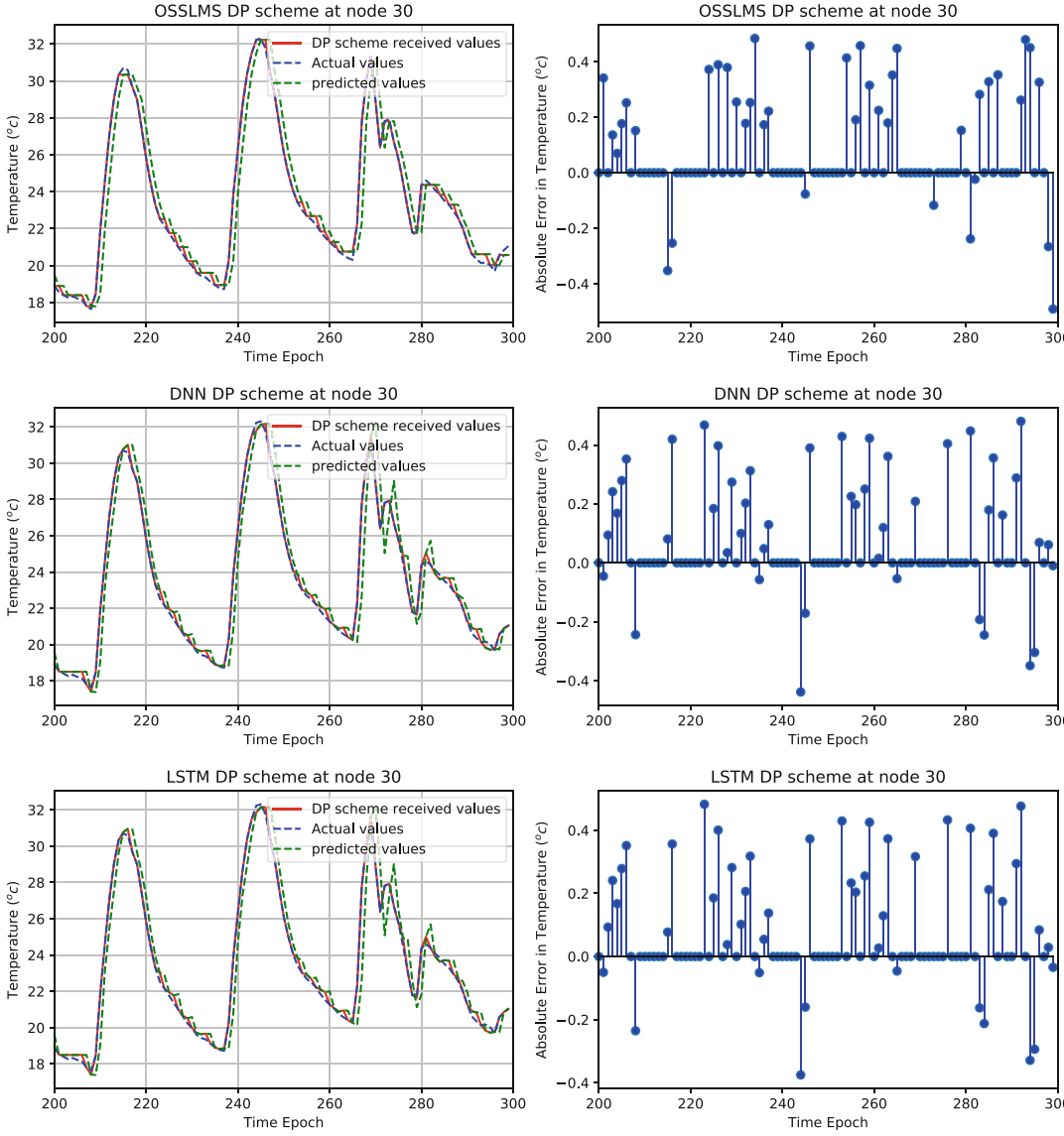
used to improve tracking in WSNs as well, as discussed in Luo et al. (2016).

Due to their gated and memory architecture, LSTMs are suitable for extracting patterns where the input data spans over long sequences. Therefore, LSTMs are suitable for prediction and time-series analytics (Jansen et al. 2018). They can be also used to extract features from long sequences (Liu et al. 2016). Moreover, LSTMs have some implementations for filling missing data points in data sets (Lipton et al. 2016). In literature, many LSTM applications can be found in WSNs and IoT research. For instance, LSTMs are commonly used architectures for short-term traffic flow prediction as illustrated in Ali and Mahmood (2018). In Aydin and Guldamlasioglu (2017), LSTMs are used for maintenance systems in predicting the condition of components in factories and autonomous IoT systems.

In order to illustrate the applicability of using DNNs and LSTMs in IoT systems, they are applied to traffic reduction problem. Next, the dual prediction scheme and its performance when applying DNNs and LSTMs will be studied.

Traffic Reduction in IoT Systems Using Dual Prediction

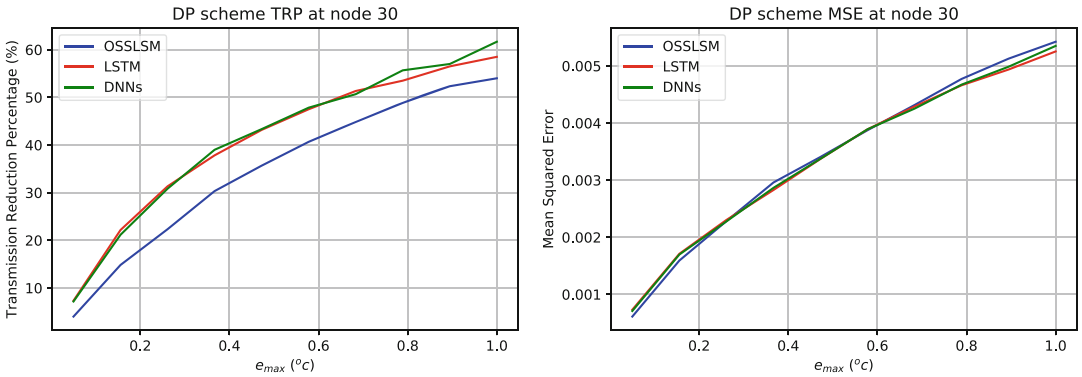
Most of the collected data is usually redundant and can be mined from other observations due to the existence of a temporal correlation. Preventing the unnecessary transmissions in a WSN has a significant impact on reducing



Machine Learning in Wireless Sensor Networks for the Internet of Things, Fig. 4 DP scheme at Node 30 using OSSLMS, DNNs, and LSTMs

energy consumption and bandwidth usage. One approach to reduce data transmission between any two endpoints is to use dual prediction (DP) (Wu et al. 2016). The endpoints can be a node-to-node, node-to-gateway, or node-to-cloud. The DP scheme is based on data prediction algorithms that can be implemented using DNNs or LSTMs.

The two ends of the DP scheme are a source of data (S) and a destination (D). Let us assume that a data buffer of size N is used to hold the last N observations at S and D. The DP scheme at S can be explained as follows. At the nth time slot, the data buffer is represented as $\boldsymbol{X}_n = [x_{n-1}, x_{n-2}, \ldots, x_{n-N}]$ where x_{n-i} represents the data point in the buffer at the $(n-i)$th time slot. Initially, for the first N time slots, collected observations are transmitted from S to D and placed in the data buffer $\boldsymbol{X}_n$. After that, when observation O_n is made at time slot n, the



Machine Learning in Wireless Sensor Networks for the Internet of Things, Fig. 5 DP scheme performance while using different accuracy levels

information in $\boldsymbol{X}_n$ is used to predict the observed value. The implemented prediction algorithms take $\boldsymbol{X}_n$ as input and generates a prediction P_n at time slot n. If the predicted value P_n was not close enough to the observed value O_n ($|P_n - O_n| > e_{\max}$ where $e_{\max}$ is the maximum acceptable prediction error), then the data buffer will be updated as $\boldsymbol{X}_{n+1} = [O_n, x_{n-1}, \ldots, x_{n-N+1}]$. Also, O_n is transmitted to D, and the value O_n is used to update the prediction model variables. However, if the predicted value P_n was close enough to the observed value O_n ($|P_n - On| \leq e_{\max}$), then $\boldsymbol{X}_{n+1} = [P_n, x_{n-1}, \ldots, x_{n-N+1}]$ and no transmission occurs because this observation can be predicted accurately. Also, the value P_n is used to update the prediction model. Following this scheme at S eliminates the need for transmitting the observations that can be accurately predicted at D.

The DP scheme at D is implemented as follows. During the first N time slots, the received observations are just saved at buffer $\boldsymbol{X}_n$. After that, whenever an observation O_n is received, it is saved in the buffer as the newest data point and used to update the prediction model. However, if no observation is received at time slot n, then it means that this point can be predicted accurately; therefore, the prediction model is used to generate the corresponding observation. After that, the predicted point is used to update the prediction model and is added to the data buffer. It should be noted that the transmitting and receiving ends use the same prediction model. Also, they perform the same model updates; therefore, both models are synchronized.

To evaluate our proposed models, the results are compared with optimal step-size least mean square (OSSLMS) algorithm that is discussed in Wu et al. (2016). The simulation has been performed on sample data collected by the Intel Berkeley research lab during a 1-month period that includes temperature, humidity, and light intensity. It is assumed that one node in Intel's WSN tries to send temperature readings periodically to another end which can be a cluster head, edge node, or cloud. It is assumed to be the cloud here. In any case, the DP scheme has the same performance.

The following parameters have been set for evaluation: the acceptable error threshold is $e_{\max} = 0.5\,^{\circ}\mathrm{C}$ and the data buffer size $N = 10$. Figure 4 shows the performance of the DP scheme when using OSSLMS, DNNs, and LSTMs for predictions. It shows how the observed (at Node 30), predicted (at Node 30 and the cloud), and perceived (at the cloud) data changes over 100 time epochs. It can be seen that DP scheme's perceived data points (red) are equal to the predicted values (green) when the prediction is accurate and equal to the observed values (blue) when the prediction is not. Also, the error between the observed values (at Node 30) and the perceived ones (at the cloud) does not exceed $e_{\max}$ at any point. Figure 5 shows the percentage of traffic reduction and the MSE while varying the acceptable

M

error value e_{max}. It shows how the proposed DNN and LSTM algorithms outperform recent literature algorithms such as OSSLMS. While all have the same MSE performance, DNN and LSTM outperform OSSLMS in terms of traffic reduction.

Cross-References

- Application of Machine Learning in Wireless Sensor Network
- Machine Learning in Wireless Sensor Networks for the Internet of Things

References

Ali U, Mahmood T (2018) Using deep learning to predict short term traffic flow: a systematic literature review. In: Intelligent transport systems – from research and development to the market uptake. Springer International Publishing, Cham, pp 90–101

Alsheikh MA, Lin S, Niyato D, Tan HP (2014) Machine learning in wireless sensor networks: algorithms, strategies, and applications. IEEE Commun Surv Tutor 16(4):1996–2018

Aydin O, Guldamlasioglu S (2017) Using LSTM networks to predict engine condition on large scale data processing framework. In: 2017 4th international conference on electrical and electronic engineering (ICEEE), pp 281–285

Bande S, Shete V (2017) Smart flood disaster prediction system using IoT neural networks. In: International conference on smart technologies for smart nation (SmartTechCon), pp 189–194

Chen XW, Lin X (2014) Big data deep learning: challenges and perspectives. IEEE Acc 2:514–525

Chen M, Challita U, Saad W, Yin C, Debbah M (2017) Machine learning for wireless networks with artificial intelligence: a tutorial on neural networks. arXiv preprint:171002913

Jansen F, Holenderski M, Ozcelebi T, Dam P, Tijsma B (2018) Predicting machine failures from industrial time series data. In: 5th international conference on control, decision and information technologies (CoDIT), pp 1091–1096

Kumar A, Jain V (2018) Feed forward neural network-based sensor node localization in internet of things. In: Pattnaik PK et al (eds) Progress in computing, analytics and networking. Springer, Singapore, pp 795–804

Lipton ZC, Kale D, Wetzel R (2016) Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In: Proceedings of Machine learning for healthcare conference, PMLR, vol 56, pp 253–270

Liu J, Shahroudy A, Xu D, Wang G (2016) Spatio-temporal LSTM with trust gates for 3D human action recognition. In: European conference on computer vision. Springer, Amsterdam, pp 816–833

Luo X, Lv Y, Zhou M, Wang W, Zhao W (2016) A laguerre neural network-based ADP learning scheme with its application to tracking control in the internet of things. Pers Ubiquit Comput 20(3):361–372

Mohammadi M, Al-Fuqaha A, Sorour S, Guizani M (2018) Deep learning for IoT big data and streaming analytics: A survey. IEEE Communications Surveys Tutorials Early Access

Rampasek L, Goldenberg A (2016) Tensorflow: biology's gateway to deep learning? Cell Syst 2(1):12–14

Wu M, Tan L, Xiong N (2016) Data prediction, compression, and recovery in clustered wireless sensor networks for environmental monitoring applications. Info Sci 329:800–818, special issue on Discovery Science

Machine Learning Paradigms in Wireless Network Association

Jingjing Wang[1] and Chunxiao Jiang[2]
[1]Department of Electronic Engineering, Tsinghua University, Beijing, People's Republic of China
[2]Department of Electronic Engineering, Tsinghua Space Center, Tsinghua University, Beijing, People's Republic of China

Synonyms

Cooperative wireless networks; Machine learning algorithms; Resource allocation

Definitions

Investigating machine learning aided cooperative resource allocation and network association mechanisms for wireless networks.

Introduction

Wireless network association has received substantial attention both in the academic

and industrial communities. One of their driving forces is that of efficiently providing unprecedented data rates for supporting radical new applications (Jiang et al. 2017). Specifically, a range of network association schemes are expected to learn the diverse and colorful characteristics of both the users' ambience and the human behaviors, in order to autonomously determine the optimal system configurations for the sake of conserving energy as well as maximizing network capacity. Moreover, smart mobile terminals have to rely on sophisticated learning and decision-making. Machine learning, as one of the most powerful artificial intelligence tools, constitutes a promising solution for wireless network association (Alsheikh et al. 2014).

Machine learning has found wide-ranging applications in image/audio processing, finance and economics, social behavior analysis, etc. Explicitly, a machine learns the execution of a particular task **T**, with the goal of maintaining a specific performance metric **P**, based on a particular experience **E**, where the system aims for reliably improving its performance **P** while executing task **T**, again by exploiting its experience **E**. Machine learning algorithms can be simply categorized as *supervised learning* and *unsupervised learning*, where the adjectives "supervised/unsupervised" indicate whether there are labeled samples in the database. Later, *reinforcement learning* emerged as a new category, which was inspired by behavioral psychology. It is concerned with an agent's certain form of reward/utility, who is connected to its environment via perception and action. Relying on a cascade of multiple layers of nonlinear processing units for feature extraction and transformation, *deep learning* was applied to network association for the sake of its expressive capacity and convenient optimization capability compared with conventional machine learning algorithms.

In wireless networks, machine learning can be widely used in modeling various technical problems of large-scale MIMOs, device-to-device (D2D) networks, heterogeneous networks (HetNets), cognitive radio, etc. (Clancy et al. 2007). In this entry, we will introduce the basic concept of machine learning algorithms and their corresponding applications in network association according to the category of supervised, unsupervised, reinforcement learning, and deep learning.

Supervised Learning in Network Association

Models

The *k-nearest neighbor* (KNN) and *support vector machines* (SVM) are conceived for classification of points/objects relying on a D-dimensional vector **x** of input variables. In KNN, an object is classified into a specific category by a majority vote of the object's neighbors, with the object being assigned to the class that is most common among its k-nearest neighbors. By contrast, the SVM relies on a nonlinear mapping, which transforms the original training data into a higher dimension where it becomes separable and then it searches for the optimal linear separating hyperplane that is capable of separating one class from another in this higher dimension.

The philosophy of *Bayesian learning* is to compute the a posteriori probability distribution of the target variables conditioned on its input signals and on all of the training instances. Some simple examples of generative models that may be learned with the aid of Bayesian techniques include, but are not limited to, the *Gaussian mixture model* (GM), *expectation maximization* (EM), and *hidden Markov models* (HMM). Specifically, GM is a model where each data point belongs to one of several clusters or groups, and the data points within each cluster are Gaussian distributed. EM is a generalization of maximum likelihood estimation, which iteratively finds the most likely solutions or parameters. It is characterized by two steps, i.e., the "E" step and the "M" step. HMM is a tool designed for representing probability distributions of sequences of observations. It can be considered a generalization of a mixture-based model, where the hidden variables, which control the specific mixture of the component to be selected for each observation, are related to

each other through a Markov process, rather than being independent of each other.

Applications

In the following, some applications of supervised learning aided network association algorithms are elaborated, which are also summarized in Table 1. KNN algorithms are beneficial for traffic prediction, for anomaly detection, as well as for modulation classification. Specifically, in order to capture the dynamic characteristics of radio resource demands, a weighted KNN model was proposed based on a large-scale historical data set from cellular operator networks, which was used to explore both temporal and spatial characteristics of radio resource margins to enhance the load balance (Feng et al. 2017). Moreover, the authors of Onireti et al. (2016) proposed a KNN-based anomaly detecting algorithms for improving cell outage detection accuracy. In Aslam et al. (2012), genetic programming and KNN were combined in order to improve the modulation classification accuracy, which provided a reliable modulation classification scheme for the secondary users (SU) in cognitive radio networks.

Furthermore, the SVM-aided learning models can be used for estimating radio parameters, for predicting mobile terminal's usage pattern, and for guiding channel selection. For example, in order to generalize the SVM function for employment in data classification problems, its hierarchical version referred to as H-SVM was proposed in Feng and Chang (2012), where each hierarchical level consisted of a finite number of SVM classifiers. This regime was used for the estimation of the Gaussian channel's noise level in a MIMO-aided wireless network. By exploiting the training data, the H-SVM model was trained for the estimation of the channel noise statistics. SVM can also be used for learning the mobile terminal's specific usage pattern in diverse spatiotemporal and device contexts, as discussed in Donohoo et al. (2014). This may then be exploited for prediction of the configuration to be used in the location-specific interface for HetNets constituted by diverse cells. In Thilina et al. (2016), the common control channel for SUs during a given frame was selected by employing an SVM-based learning technique for a cognitive radio network, which was capable of implicitly learning the surrounding environment cooperatively in an online fashion.

The Bayesian learning model may be readily invoked for spectral characteristic learning and estimation. The authors of Wen et al. (2015) estimated both the channel parameters of the desired links in a target cell and those of the interfering links of the adjacent cells, relying on the sparse Bayesian learning techniques, where the channel component was first modeled by a GM, and then estimated with the aid of the EM algorithm. In Choi and Hossain (2013), a cooperative wideband spectrum sensing scheme based on the EM algorithm was proposed for the detection of a primary user (PU) supported by a multi-antenna assisted cognitive radio net-

Machine Learning Paradigms in Wireless Network Association, Table 1 Applications of supervised learning aided network association algorithms

Method	Scenario	Applications	Reference
KNN	Cognitive radio	Radio resource capture	(Feng et al. 2017)
		Modulation classification	(Aslam et al. 2012)
	HetNets	Anomaly detection	(Onireti et al. 2016)
SVM	MIMO	Channel noise estimation	(Feng and Chang 2012)
	Cognitive radio	Channel selection	(Thilina et al. 2016)
	HetNets	Usage pattern prediction	(Donohoo et al. 2014)
Bayesian learning	MIMO	Channel parameter estimation	(Wen et al. 2015)
	Cognitive radio	Spectrum sensing	(Choi and Hossain 2013)
		Signal estimation and detection	(Assra et al. 2016)

work. Furthermore, in Assra et al. (2016), the authors constructed a HMM relying on a two-state hidden Markov process, where the PUs are present or absent, and a two-state observation space, indicating whether the PUs are present or absent. Furthermore, the EM algorithm was invoked for finding the true channel parameters, such as the sojourn time of the available channels, the inactive states of the PUs, as well as the PUs' signal strength.

Unsupervised Learning in Network Association

Models

K-means clustering aims for partitioning n observations into k clusters, where each observation belongs to the closest cluster. It defines the centroid of a cluster as the center of gravity, i.e., the mean value of the points within the cluster. The clustering algorithm proceeds in an iterative manner, where an object is assigned to the specific cluster whose centroid is nearest to the object based on the Euclidean distance, and then the in-cluster differences are minimized by iteratively updating the cluster centroid, until convergence is achieved.

Principal component analysis (PCA) transforms a set of potentially correlated variables into a set of uncorrelated variables referred to as the principal components, where the number of principal components is less than or equal to the number of original variables. Basically, the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The principal components are orthogonal, because they are the eigenvectors of the covariance matrix, which is symmetric. By contrast, *independent component analysis* (ICA) is a statistical technique conceived for revealing hidden factors that underlie sets of random variables, measurements, or signals. In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed to be non-Gaussian and mutually independent, and they are referred to as the independent components of the observed data, which can be found by ICA.

Applications

Table 2 summarizes some applications relying on unsupervised learning assisted network association schemes. To elaborate a little further, clustering is a common problem in wireless networks, especially in heterogeneous scenarios associated with diverse cell sizes as well as Wi-Fi and D2D networks. For example, the small cells have to be carefully clustered for avoiding interference using coordinated multipoint transmission, while the mobile users are clustered for obeying an optimal offloading policy; the devices are clustered in D2D networks for the sake of achieving high energy efficiency; the Wi-Fi users are clustered for maintaining an optimal access point association, etc. A mixed integer programming problem was formulated for jointly optimizing both the gateway partitioning and the virtual channel allocation based on classic k-means clustering, which was employed for partitioning the mesh access points into several groups in a hybrid optical/wireless network scenario for the sake of reducing the overall wireless tele-traffic by encouraging the utilization of the high-capacity optical infrastructure (Xia et al. 2012). In Hajjar et al. (2017), a K-means-based relay selection algorithm was proposed for creating low-power small cells in an LTE macro cell within a multi-cell scenario, where the network's total capacity was increased without the need of additional infrastructure.

Both the PCA and ICA constitute powerful statistical signal processing techniques devised for recovering statistically independent source signals from their linear mixtures. One of their major applications may be found in the area of anomaly-, fault-, and intrusion-detection problems of wireless networks, which rely on traffic monitoring. Furthermore, similar problems may also be solved in sensor networks, mesh networks, etc. They can also be invoked for the physical layer signal dimension reduction of massive MIMO systems or for classifying

Machine Learning Paradigms in Wireless Network Association, Table 2 Applications of unsupervised learning aided network association algorithms

Method	Scenario	Applications	Reference
K-means clustering	Hybrid network	Virtual channel allocation	(Xia et al. 2012)
	LTE network	Relay selection	(Hajjar et al. 2017)
	MIMO	Spectral efficiency	(Liang et al. 2016)
	Optical networks	Signal detection	(Zhao et al. 2006)
	WSN	Optimal sensor deployment	(Ateş et al. 2017)
PCA	Wi-Fi	User detection	(Zhu et al. 2017)
	WSN	Efficient localization	(Li et al. 2017)
ICA	Cognitive radio	Signal sources detection	(Nguyen et al. 2013)
	Smart grid	Energy efficiency	(Qiu et al. 2011)

the primary users' behaviors in cognitive radio networks. In Qiu et al. (2011), PCA and ICA were applied in a smart grid scenario for recovering the simultaneous wireless transmissions of smart utility meters installed in each home, which were capable of enhancing both the transmission efficiency by avoiding the channel estimation in each frame and the data security by eliminating any wideband interference or jamming signals. Another pertinent example is found in cognitive radio scenarios, where the so-called Boolean ICA relied on the Boolean mixing of OR, XOR, and other functions of binary signals (Nguyen et al. 2013), where an iterative algorithm, termed as the binary ICA, was developed for determining the activities of the underlying latent signal sources, such as the PUs. It was demonstrated that given m monitors or SUs, the activities of up to $(2m - 1)$ distinct PUs can be inferred.

Reinforcement Learning in Network Association

Models

Multiarmed bandit relies on a player at a row of slot machines who has to decide which machines to play on and how many times to play on each. He/she will receive a random reward provided by the selected machine. The objective of the game is to maximize the sum of rewards earned through a sequence of lever pulls. It has become one of the popular examples of sequential decision-making problems striking an exploration-exploitation trade-off relying on limited information.

Markov decision process (MDP) provides a mathematical framework for modeling decision-making in specific situations, where the outcomes are partly random and partly under the control of a decision-maker. At each time step, the process is in some state S, and the decision-maker may opt for any of the legitimate actions A that is available in state S. The process responds at the next time step by randomly moving into a new state S' and giving the decision-maker a corresponding reward $r(S, A)$. The probability that the process moves into its new state S' is influenced both by the specific action chosen and by the system's inherent transitions. As an important member of MDP family, *partially observable Markov decision process* (POMDP) may be suitable for a general scenario, where the agent is unable to directly observe the underlying state transitions and hence only has partial knowledge. The agent has to keep track of both the probability distribution of the legitimate states, based on a set of observations, and the observation probabilities and of the underlying MDP.

Q-learning may be invoked for finding an optimal action policy for any given finite Markov decision process, especially when the system model is unknown. It is a model-free reinforcement learning technique, and as such it can be used in conjunction with MDP models. In such a case, the Q-learning model is also comprised of an agent, of the states, and of a set of actions per state. By executing an action in a

Machine Learning Paradigms in Wireless Network Association, Table 3 Applications of reinforcement learning aided network association algorithms

Method	Scenario	Applications	Reference
Multiarmed bandit	D2D network	Channel selection	(Maghsudi and Stańczak 2015)
	Li-Fi network	AP selection	(Wang et al. 2018)
MDP/POMDP	WSN	Power control	(Aprem et al. 2013)
	Super Wi-Fi	Energy harvesting	(Wang et al. 2016)
	D2D	Network coding	(Karim et al. 2017)
Q-learning	Femtocell network	Resource allocation	(Alnwaimi et al. 2015)
	HetNets	Admission control	(Chen et al. 2009)
	Dense cellular network	Traffic offloading	(Fakhfakh and Hamouda 2017)

specific state, the agent gleans a reward and the goal is to maximize its accumulated reward. Such a reward is illustrated by a Q-function, where "Q" is initialized to be a fixed value. Then, "Q" is updated in an iterative manner after the agent carries out an action and observes the resultant reward as well as the associated new state at each time instant.

Applications

Table 3 lists a range of applications of reinforcement learning assisted network association schemes. Generally, multiarmed bandit may be beneficially used in multiplayer adaptive decision-making problems, where selfish players infer an optimal joint action profile from their successive interactions with a dynamic environment and finally settle at some equilibrium point. This problem has indeed been encountered in many wireless networking scenarios, with a compelling one being the channel selection problem (Maghsudi and Stańczak 2015) and another one in the context of access point (AP) selection problem (Wang et al. 2018). Specifically, every mobile user was modeled as a player of the multiarmed bandit game, while the channels/APs were regarded as arms and choosing a channel or an AP corresponds to pulling an arm. The authors of Maghsudi and Stańczak (2015) proposed a channel selection strategy consisting of two main blocks for D2D communication system integrated into a cellular network, namely, the calibrated forecasting and the no-regret bandit learning strategies. Moreover, the authors of Maghsudi and Stańczak (2015) proposed a multiarmed bandit-aided AP selection algorithm for a visible light communication (VLC) and Wi-Fi hybrid system (Li-Fi).

The family of MDP/POMDP models constitutes ideal tools for supporting decision-making in wireless networks, where the users may be regarded as agents and the network constitutes the environment. For instance, in Aprem et al. (2013), the transmission power control problems in energy harvesting wireless sensor networks (WSN) were investigated using the POMDP model, where the state space was defined by including the battery state, the channel state, the packet transmission/reception states, and an action by the node, which corresponded to sending a packet at a certain power level. In Wang et al. (2016), a POMDP-aided AP section algorithm was proposed for improving the energy efficiency of the super Wi-Fi system, including a low-complexity energy function-based solution.

Q-learning has also been extensively applied in HetNets, usually in conjunction with the aforementioned MDP models. In Alnwaimi et al. (2015), the authors presented a heterogeneous fully distributed multi-objective strategy based on a reinforcement learning model constructed for the self-configuration/optimization of femtocells. The model was supposed to solve both the resource allocation and interference coordination problem in the downlink of femtocell networks under carefully constructed restrictions for avoiding interference and for meeting the quality of service (QoS) requirements. Furthermore,

in Chen et al. (2009), a fuzzy Q-learning admission control for wideband code division multiple access (WCDMA)/wireless local area network (WLAN) HetNets was proposed, which considered not only QoS requirements but also multiple system measures, such as interferences, the numbers of real-time and non-real-time users, user's mobility, etc.

Deep Learning in Network Association

Models

Artificial neural networks are a set of algorithms inspired by the biological neural networks that constitute animal brains, which help us cluster and classify. The *deep neural network* (DNN) is a kind of artificial neural network, with multiple layers between the input and output layers, which can model complex nonlinear relationships. The layers are made of nodes, where a node is just a place where computation happens, loosely patterned on a neuron in the human brain. A node combines input from the data with a set of weights that either amplify or dampen that input, thereby assigning significance to inputs for the task considered. The nodes residing in "deeper" layer of the neural net are capable of recognizing more complex features, since they aggregate and recombine features from the previous layer. In other words, the extra layers enable composition of features from lower layers, potentially modeling complex data with fewer units than a similarly performing shallow network.

The *recurrent neural network* (RNN) is another class of artificial neural network where connections between units form a directed cycle, which allows it to exhibit dynamic temporal behavior. Unlike traditional neural networks where all inputs are independent of each other, RNN is called "recurrent" because they perform the same task for every element of a sequence, with the output being depended on the previous computations. Hence, RNN can be viewed to have a "memory" which captures information about what has been calculated so far. RNNs can use their internal memory to process arbitrary sequences of inputs.

Applications

In Table 4, we summarize some typical applications of deep learning aided network association algorithms. DNN algorithms are readily used for traffic control, wireless localization, activity recognition, resource allocation, and anomaly detection. Specifically, in order to improve traffic control in HetNets, the authors in Kato et al. (2017) first characterized the HetNet traffic and then proposed a supervised DNN system, which exploited the extracted characterizations to achieve excellent signaling overhead, throughput, and delay. Moreover, in a cognitive radio network, the DNN-based algorithm was proposed to learn features that indicated the influence of the target on the wireless signals, which substantially improved wireless localization and activity recognition (Wang et al. 2017a).

Furthermore, RNN can be applied to traffic classification, interference cancellation, and channel equalization. For instance, in Lopez-Martin et al. (2017), a combination model of a RNN and a convolutional neural network (CNN) was proposed to classify both the traffic and the

Machine Learning Paradigms in Wireless Network Association, Table 4 Applications of deep learning aided network association algorithms

Method	Scenario	Applications	Reference
DNN	HetNets	Traffic control	(Kato et al. 2017)
	Cognitive radio	Wireless localization	(Wang et al. 2017a)
	Smart grid	Attack detection	(He et al. 2017)
	WSN	Sensor calibration	(Wang et al. 2017b)
RNN	IoT network	Traffic classification	(Lopez-Martin et al. 2017)
	MIMO	Interference cancellation	(Mostafa 2017)
		Channel equalization	(Zhao et al. 2011)
	Cellular network	Power control	(Chen et al. 2006)

behavior of heterogeneous devices and services in IoT networks, which provided a superior detection result without requiring any feature engineering. Besides, as discussed in Mostafa (2017), RNN was beneficially invoked for interference cancellation in MIMO systems. The proposed RNN algorithm showed good performance in terms of suppressing impulsive noise while assuring the Lyapunov stability simultaneously.

Conclusions

This entry reviewed the benefits of artificial intelligence-aided wireless systems equipped with machine learning. We introduced the major families of machine learning algorithms and discussed their applications in the context of massive MIMOs, the smart grid, cognitive radios, HetNets, small cells, D2D networks, etc. The classes of supervised, unsupervised, reinforcement, and deep learning tools were investigated, along with the corresponding modeling methodology and possible future applications in wireless network association. In a nutshell, machine learning is an exciting area for artificial intelligence-aided networking research!

References

Alnwaimi G, Vahid S, Moessner K (2015) Dynamic heterogeneous learning games for opportunistic access in LTE-based macro/femtocell deployments. IEEE Trans Wirel Commun 14(4):2294–2308

Alsheikh MA, Lin S, Niyato D, Tan HP (2014) Machine learning in wireless sensor networks: algorithms, strategies, and applications. IEEE Commun Surv Tutorials 16(4):1996–2018

Aprem A, Murthy CR, Mehta NB (2013) Transmit power control policies for energy harvesting sensors with retransmissions. IEEE J Sel Top Sign Proces 7(5):895–906

Aslam MW, Zhu Z, Nandi AK (2012) Automatic modulation classification using combination of genetic programming and KNN. IEEE Trans Wirel Commun 11(8):2742–2750

Assra A, Yang J, Champagne B (2016) An EM approach for cooperative spectrum sensing in multiantenna CR networks. IEEE Trans Veh Technol 65(3):1229–1243

Ateş E, Kalayci TE, Uğur A (2017) Area-priority-based sensor deployment optimisation with priority estimation using K-means. IET Commun 11(7):1082–1090

Chen YS, Chang CJ, Hsieh YL (2006) A channel effect prediction-based power control scheme using PRNN/ERLS for uplinks in DS-CDMA cellular mobile systems. IEEE Trans Wirel Commun 5(1):23–27

Chen YH, Chang CJ, Huang CY (2009) Fuzzy Q-learning admission control for WCDMA/WLAN heterogeneous networks with multimedia traffic. IEEE Trans Mob Comput 8(11):1469–1479

Choi KW, Hossain E (2013) Estimation of primary user parameters in cognitive radio systems via hidden Markov model. IEEE Trans Signal Process 61(3):782–795

Clancy C, Hecker J, Stuntebeck E, O'Shea T (2007) Applications of machine learning to cognitive radio networks. IEEE Wirel Commun 14(4):47–52

Donohoo BK, Ohlsen C, Pasricha S, Xiang Y, Anderson C (2014) Context-aware energy enhancements for smart mobile devices. IEEE Trans Mob Comput 13(8):1720–1732

Fakhfakh E, Hamouda S (2017) Optimised Q-learning for WiFi offloading in dense cellular networks. IET Commun 11(15):2380–2385

Feng VS, Chang SY (2012) Determination of wireless networks parameters through parallel hierarchical support vector machines. IEEE Trans Parallel Distrib Syst 23(3):505–512

Feng Z, Li X, Zhang Q, Li W (2017) Proactive radio resource optimization with margin prediction: a data mining approach. IEEE Trans Veh Technol 66(10):9050–9060

Hajjar M, Aldabbagh G, Dimitriou N, Win MZ (2017) Hybrid clustering scheme for relaying in multi-cell LTE high user density networks. IEEE Access 5:4431–4438

He Y, Mendis GJ, Wei J (2017) Real-time detection of false data injection attacks in smart grid: a deep learning-based intelligent mechanism. IEEE Trans Smart Grid 8(5):2505–2516

Jiang C, Zhang H, Ren Y, Han Z, Chen KC, Hanzo L (2017) Machine learning paradigms for next-generation wireless networks. IEEE Wirel Commun 24(2):98–105

Karim MS, Sorour S, Sadeghi P (2017) Network coding for video distortion reduction in device-to-device communications. IEEE Trans Veh Technol 66(6):4898–4913

Kato N, Fadlullah ZM, Mao B, Tang F, Akashi O, Inoue T, Mizutani K (2017) The deep learning vision for heterogeneous network traffic control: proposal, challenges, and future perspective. IEEE Wirel Commun 24(3):146–153

Li X, Ding S, Li Y (2017) Outlier suppression via non-convex robust PCA for efficient localization in wireless sensor networks. IEEE Sensors J 17(21):7053–7063

Liang HW, Chung WH, Kuo SY (2016) Coding-aided K-means clustering blind transceiver for space shift keying MIMO systems. IEEE Trans Wirel Commun 15(1):103–115

Lopez-Martin M, Carro B, Sanchez-Esguevillas A, Lloret J (2017) Network traffic classifier with convolutional and recurrent neural networks for internet of things. IEEE Access 5:18042–18050
Maghsudi S, Stańczak S (2015) Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting. IEEE Trans Wirel Commun 14(3):1309–1322
Mostafa M (2017) Stability proof of iterative interference cancellation for OFDM signals with blanking nonlinearity in impulsive noise channels. IEEE Signal Process Lett 24(2):201–205
Nguyen H, Zheng G, Zheng R, Han Z (2013) Binary inference for primary user separation in cognitive radio networks. IEEE Trans Wirel Commun 12(4):1532–1542
Onireti O, Zoha A, Moysen J, Imran A, Giupponi L, Imran MA, Abu-Dayya A (2016) A cell outage management framework for dense heterogeneous networks. IEEE Trans Veh Technol 65(4):2097–2113
Qiu RC, Hu Z, Chen Z, Guo N, Ranganathan R, Hou S, Zheng G (2011) Cognitive radio network for the smart grid: experimental system architecture, control algorithms, security, and microgrid testbed. IEEE Trans Smart Grid 2(4):724–740
Thilina KGM, Hossain E, Kim DI (2016) DCCC-MAC: a dynamic common-control-channel-based MAC protocol for cellular cognitive radio networks. IEEE Trans Veh Technol 65(5):3597–3613
Wang J, Jiang C, Han Z, Ren Y, Hanzo L (2016) Network association strategies for an energy harvesting aided super-wifi network relying on measured solar activity. IEEE J Sel Areas Commun 34(12):3785–3797
Wang J, Zhang X, Gao Q, Yue H, Wang H (2017a) Device-free wireless localization and activity recognition: a deep learning approach. IEEE Trans Veh Technol 66(7):6258–6267
Wang Y, Yang A, Chen X, Wang P, Wang Y, Yang H (2017b) A deep learning approach for blind drift calibration of sensor networks. IEEE Sensors J 17(13):4158–4171
Wang J, Jiang C, Zhang H, Zhang X, Leung VC, Hanzo L (2018) Learning-aided network association for hybrid indoor LiFi-WiFi systems. IEEE Trans Veh Technol 67(4):3561–3574
Wen CK, Jin S, Wong KK, Chen JC, Ting P (2015) Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning. IEEE Trans Wirel Commun 14(3):1356–1368
Xia M, Owada Y, Inoue M, Harai H (2012) Optical and wireless hybrid access networks: design and optimization. J Opt Commun Netw 4(10):749–759
Zhao T, Nehorai A, Porat B (2006) K-means clustering-based data detection and symbol-timing recovery for burst-mode optical receiver. IEEE Trans Commun 54(8):1492–1501
Zhao H, Zeng X, Zhang J, Li T (2011) Equalisation of non-linear time-varying channels using a pipelined decision feedback recurrent neural network filter in wireless communication systems. IET Commun 5(3):381–395
Zhu H, Xiao F, Sun L, Wang R, Yang P (2017) R-TTWD: robust device-free through-the-wall detection of moving human with WiFi. IEEE J Sel Areas Commun 35(5):1090–1103

Machine to Machine (M2M)

▸ Machine-Type Communication

Machine-Learning (ML) in 4G Networks

▸ Survey on Machine-Learning Techniques in LTE Networks

Machine-Type Communication

Chia-Peng Lee[1] and Phone Lin[2]
[1]Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan
[2]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, Republic of China

Synonyms

Internet of things; Machine to machine (M2M)

Definition

The machine-type communication (MTC) is a new type of communication technologies for the machine-to-machine (M2M) or Internet of Things (IoTs) with few or without human intervention (3GPP 2014), and it is proposed by the Third Generation Partnership Project (3GPP) working group.

Historical Background

The standardization work for the MTC was started by 3GPP TSG SA1 working group in early 2005. In 2007, a technical report on facilitating M2M communication in the 3GPP systems was exported (3GPP 2007). The 3GPP also identified the service requirements and system improvements for the MTC in release 10 (3GPP 2014), and a specific protocol implementation is described in release 11 (3GPP 2012).

The features of MTC specified by the 3GPP are listed as follows (3GPP 2014):

- Low mobility: The MTC devices have low mobility or move only within a specific area.
- Time controlled: The MTC devices transmit the data during specific time intervals.
- Small data transmissions: The applications have a small amount of data transmission or reception.
- Mobile originated only (or infrequent mobile terminated): Most of application calls are initiated by the MTC device.
- MTC monitoring: Most of applications monitor the behaviors and the events of the MTC devices.
- Priority alarm: The alerting function in applications reports anomaly events, e.g., the MTC devices were stolen or destroyed.
- Group-based MTC features: The system is optimized with grouped MTC devices.

Figure 1 shows the architecture of the MTC in release 11 (3GPP 2012, 2018), where an MTC device connects to the LTE core network through the LTE-Uu air interface. The network entities and interfaces are described as follows:

- The MTC Interworking Function (MTC-IWF) is a network element that receives external application service signaling through the Tsp interface.
- The Service Capability Server (SCS) communicates with MTC devices through the MTC-IWF. The MTC devices can access one or more application services controlled and managed by the Application Servers (ASs) through the SCS.
- The S6m interface exercises between the MTC-IWF and the Home Subscriber System (HSS). The interface is used by the MTC-IWF to obtain the IDs of an MTC device from the HSS, which include the International Mobile Subscriber Identity (IMSI) and Mobile Subscriber Integrated Services Digital Network Number (MSISDN).
- The MTC-IWF sends trigger messages to the Short Message Service Center (SMS-SC) through the T_4 interface.

Based on the above network architecture, the 3GPP defines three modes for the core network to connect to the MTC ASs, including the indirect mode, the direct mode, and the hybrid mode. The executions of the three connected modes are described below:

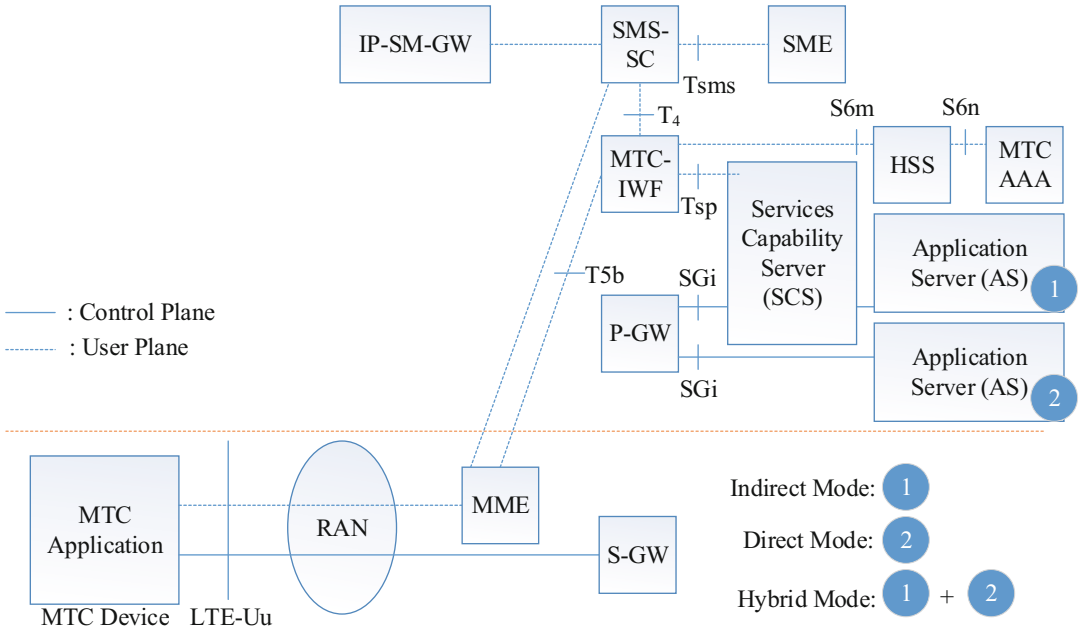
- Indirect mode: The AS connects to the core network through the SCS.
- Direct mode: The AS connects the core network through the PDN Gateway (P-GW) that is a router in the core network and provides IP connectivity between the core network and external data network.
- Hybrid mode: The LTE network operator can provide both of indirect and direct modes.

Key Applications

The 3GPP release 14 (3GPP 2017) lists the key applications that fall into the following service areas.

- The following MTC applications belong to the security service area:

  1. An alarm system includes a set of MTC devices that can trigger an alarm. Examples of the MTC devices can be gas sensors, temperature sensors, etc.
  2. A surveillance system uses the MTC devices (e.g., camera, closed-circuit



Machine-Type Communication, Fig. 1 The architecture of LTE MTC network

television; CCTV) to observe or monitor users' behaviors and activities.

3. The backup-for-landline system is usually used for home security, which is equipped with the MTC devices to monitor the landline. If the landline is broken, the MTC devices will send a warning message to the monitor system and notify that the landline has something wrong and needs to fix.
4. The physical access control system accommodates a set of physical equipment (e.g., RFID door and gate controller). The physical access control system is implemented in a building and authorizes legal access.
5. With the car/driver security system, in a connected vehicle, the MTC devices can retrieve the information (e.g., revolution per minute; RPM), based on which, the system can diagnose whether the car is safe or not. It also ensures the safety of the driver.

- The following applications fall into the tracking and tracing service area:

1. With the fleet management system, a car installed with the MTC devices which combine GPS and short message service to keep track immediate location and history trajectory of a car.
2. In the pay-as-you-drive service, the car insurance fees are easy to collect based on the information in car such as speed, history trajectory, and driving time, and then the collected data transmit to the insurance company.
3. With the asset tracking system, every asset installs an MTC device with GPS service, and the MTC device sends the location to the asset tracking system for easy tracking the location of the asset.
4. In the navigation service, the drivers can change the driving path in advance through the geographical information and other information reported by the MTC devices to avoid the traffic jams.
5. With traffic information system, the car equips with MTC devices combined GPS function that reports location information

to the MTC server. Then, based on the reported location, the MTC server provides the traffic information of that of reported location to inform driver to avoid the traffic jam.

6. With the road tolling system, the roadside units (RSUs) are deployed along the freeway. When a car equipped with the MTC device passes by that of the freeway, the RSU senses the equipped MTC device and then makes a charge.
7. With the traffic optimization/steering system, based on the location collected by MTC device in a car, it is easy to provide the guideline to the driver to achieve the traffic optimization and steering.

- The following MTC applications belong to the payment service area:

1. With the point-of-sales system, the MTC device reads commodity information (e.g., good name, price, and qualify) and then sends the information to relevant departments through communication networks for analysis to improve operating efficiency of the system.
2. The vending machine system is equipped with MTC devices to inform the sales that the items have been sold out by email or short messages, and they should replenish the items immediately. Another application is if the vending machine is broken down, the MTC device is triggered to send an alert (e.g., short message) to the sales.

- The following applications fall into the health service area:

1. With the monitoring vital signs system, the person wears MTC devices that monitor the changing of vital signs (e.g., the pulse, temperature, breathing, and blood pressure). If the monitoring vital signs change, the MTC devices send an alert message to the monitoring vital signs system.
2. In the aged and handicapped system, the MTC devices are deployed in the home environment to keep track of the behavior of elderly people.
3. With the remote health diagnostics system, the doctor can remotely monitor the health status and chronic conditions of patients anywhere anytime through the MTC devices and make reaction immediately when an emergency event occurs.

- The following applications fall into the remote maintenance and control service area:

1. In the programmable logic controllers (PLCs) system, MTC devices can control PLC by predefined instructions (e.g., turning on the light at home) which are stored in PLCs memory.
2. In the sensor system, the MTC devices are deployed as sensors (e.g., power, fire, smoke, gas, water, wind, etc.) to monitor the environment.
3. In the lighting system, the MTC devices can detect environmental change (e.g., from day to night) and send a notification to the lighting system to turn on the light.
4. In the vending machine control system, the vending machine is installed with MTC devices to determine whether the vending machine operates normally or abnormally. If the detected result is abnormal, the MTC devices send a warning message to the vending machine control system.
5. In the vehicle diagnostics system, the MTC devices are installed in a connected car to collect the information from the car, and then send the information to the vehicle diagnostics system, based on which the diagnostics system analyzes whether the car is in the safe status or the dangerous status.

- The following MTC applications belong to the metering service area:

1. In the gas system or the water system, the MTC devices are installed as smart meters,

which collect the information about the usage of gas or water. The information is referenced to charge the users for the usage of gas or water.
2. In the heating system, the MTC devices are deployed as sensor to sense the ambient temperature to control the heating automatically.
3. In the grid control system, the MTC devices collect information of power supply status in the home environment, and send the information to the grid control system to adjust the production and transmission of electricity to reduce the power consumption for power saving.

Cross-References

- ▸ Internet of Things
- ▸ Machine to Machine (M2M)

References

3 GPP (2007) Study on facilitating machine to machine communication in 3 GPP systems: TR 22.868
3GPP (2012) System improvements for Machine-Type Communications (MTC):TR 23.888
3GPP (2014) Service requirements for Machine-Type Communications (MTC): TS 22.368
3GPP (2017) Service requirements for Machine-Type Communications (MTC): TS 22.368
3GPP (2018) Architecture enhancements to facilitate communications with packet data networks and applications: TS 23.682

Recommended Reading

3GPP (2013) Study on Machine-Type Communications (MTC) and other mobile data applications communications enhancements: TR 23.887
Jain P et al (2012) Machine type communications in 3GPP systems. IEEE Commun Mag 50(11):28–35. https://doi.org/10.1109/mcom.2012.6353679
Lien S-Y et al (2011) Toward ubiquitous massive accesses in 3 GPP machine-to-machine communications. IEEE Commun Mag 49(4):66–74. https://doi.org/10.1109/mcom.2011.5741148
Taleb T, Kunz A (2012) Machine type communications in 3GPP networks: potential, challenges, and solutions. IEEE Commun Mag 50(3):178–184. https://doi.org/10.1109/mcom.2012.6163599

Macrocell–Small Cell Systems

- ▸ Resource Management in Macrocell–Small Cell Systems and D2D-Assisted Cellular Systems

Malware

- ▸ New Variants of Mirai and Analysis

Maritime Internet of Vessels

Wang Zhen[1] and Bin Lin[2]
[1]Dalian Neusoft University of Information, Dalian Maritime University, Dalian, China
[2]School of Information Technology, Dalian Maritime University, Dalian, China

Synonyms

Internet of Things for vessels; Networking of marine vessels

Definitions

Maritime Internet of Vessels is a kind of information service network for intelligent shipping which integrates Internet of Things technology. It takes vessels, waterways, and land-shore facilities as basic nodes and information sources and connects vessels through wireless communication technology, aiming at realizing the refinement of shipping management, the comprehensiveness of industry services, and the humanization of user experience.

Key Points

Historical Background

With the development of shipping industry, the amount of data and information acquired from

ships, cargo, operators, and navigation environment is increasing at the speed of TB and PB, and the traditional maritime communication systems can hardly afford for the development requirements at present. How to transmit, store, and manage the massive data effectively has become a research hotspot. At present, more than 80% of the world's trade is completed by shipping, and more than 50,000 merchant ships are engaged in international trade, maintaining the stability of global material transport. However, accidents related to navigation continue to occur despite the development and availability of a number of ship- and shore-based technologies that promise to improve situational awareness and decision-making. In 2018, there were 255 pirate attacks and harassment incidents worldwide, and 126 crew members were kidnapped by pirates worldwide. Safety, efficiency, and environmental protection of maritime communications have always been the core issues in the development of intelligent shipping, as well as the key points in ship collision avoidance, navigation assistance, and maritime positioning.

In recent years, the "intelligent" manufacturing model and products have extended to various fields, and many high tech and equipment have been greatly developed and widely used in the field of navigation, especially on board ships. "Intelligent shipping" and "e-Navigation" are respectively promoted on two independent main lines of development to jointly realize the safety, efficiency, and environmental protection of maritime shipping. In addition, the Maritime Internet of Vessels will bring shipping into a new era in which vessels and vessels, vessels and waterways, and vessels and people are interconnected and the original risk-ridden and uncertain voyages begin to become intelligent and controllable. To better comprehend the development of the Maritime Internet of Vessels, this entry makes a detailed investigation about the existing networking modes for Maritime Internet of Vessels and puts forward the further development. With different kinds of possible network access and interconnection, the ubiquitous links between vessels and vessels and people can be realized; the intelligent perception, recognition, and management of vessels and processes can also come true; and the maritime navigation will become more intelligent, controllable, safe, and environmentally friendly.

The Existing Networking Modes for Maritime Internet of Vessels

At the 81st meeting of the Maritime Safety Committee (MSC) of the International Maritime Organization (IMO), an e-Navigation proposal was put forward jointly in order to integrate and harmonize the maritime communication systems and better exchange and unify information between shore-based and shipboard users through electronic means. The IMO has defined e-Navigation as "the harmonized collection, integration, exchange, presentation and analysis of marine information on board and ashore by electronic means to enhance berth to berth navigation and related services for safety and security at sea and protection of the marine environment" (Weintrit and Weintrit 2011). The establishment of e-Navigation makes the information exchange between ship and ship and ship and shore safer, more convenient, and more efficient. On this basis, the concept of Maritime Service Portfolios (MSP) came into being which mainly refers to a set of standardized, operational, and technical maritime service information provided by the coastal side to navigators in a given sea area, waterway, port, and other similar areas. According to different applications and services, the networking modes for Maritime Internet of Vessels will be described from the perspective of conventional communications and emergency communications:

- The conventional communications

  The conventional communications are mainly used for maritime applications such as electronic chart display and information system (ECDIS), normal navigation, voyage planning, crew and passengers' communications, etc. These systems mainly operate at medium- and high-frequency systems and very-high-frequency (MF, HF, VHF) band to achieve a certain degree of voice and data communication.

(1) Digital selective calling (DSC) (MF/HF & VHF bands)

Digital selective calling or DSC is a standard for sending pre-defined digital messages to another station or group of stations for distress or general communications via the medium-frequency (MF), high-frequency (HF), and very-high-frequency (VHF) maritime radio systems (de Sousa and Pelas 2011). The International Radio Advisory Committee (CCIR) has allocated a DSC channel in the 2 MHz band of MF; the 4, 6, 8, 12, and 16 MHz band of HF; and the VHF marine channel 70 (156.525 MHz) for the DSC distress call (alarm) frequency and the radio telephone, radio distress, and secure communication frequency.

(2) Narrowband direct printing (NBDP or radio telex, MF)

NBDP is a technology that automates radio signals to telegraphy which has been used for ship location reporting and the issuance of weather warnings and forecasts at coastal stations, while the use of NBDP for general communications is decreasing. As an integral part of GMDSS, NBDP is a FSK modulated to 0.5 kHz HF channel, which supports low-speed data transmission (100 bps) over 1.6–26.5 MHz for maritime mobile services. It can be used as text-based distress tracking communication and general communication between ship-to-ship, ship-to-shore, and shore-to-ship, especially to overcome language difficulties.

(3) Navigational Telex (NAVTEX, MF)

NAVTEX (Navigational Telex), as a major element of the global maritime distress and safety system, is diffusely used for delivery of navigational and meteorological warnings and forecasts, as well as urgent maritime safety information to ships. It mainly provides a low-cost, simple, and automated means of receiving this information aboard ships at sea within approximately 370 km off shore. It adopts FSK modulation scheme and broadcasts messages in English on MF band of 518 kHz, while 490 and 4209.5 kHz are used to broadcast in English and/or local language.

(4) Automatic identification system (AIS, VHF)

The automatic identification system (AIS) is an automatic tracking system aimed at improving the safety of navigation by assisting ships in effective navigation, environmental protection, and operation of vessel traffic services (VTS). It provides a means for ships to electronically exchange ship data including identification, position, course, speed and navigational status with other nearby ships and AIS base stations, and satellites. AIS uses VHF channels and adopts a technology called self-organized time division multiple access (SOTDMA) to ensure that the VHF transmissions of different transceivers do not occur at the same time. Besides, AIS is also integrated into other maritime devices such as aids to navigation (A to N), search and rescue transmitter, and man overboard units (Lázaro et al. 2017).

(5) VHF data exchange system (VDES, VHF)

VDES is an enhanced and upgraded system of AIS for alleviating the pressure in the existing frequency band of the AIS system and ensuring the implementation of the normal function in the field of maritime mobile services. On the basis of integrating existing AIS functions, VDES adds special application message (ASM) and broadband VHF data exchange (VDE) functions, which can effectively alleviate the pressure of existing AIS data communication and provide effective auxiliary means for the protection of ship navigation safety. At the same time, it will comprehensively improve the ability and frequency efficiency of marine data communication, which is of great significance to promote the development

of marine radio digital communication industry. Furthermore, the standardization process of VDES is quite mature, and it is expected to become an ITU standard soon (Lázaro et al. 2017).

- The emergency communications

  Besides general communications, the emergency communications also focus on security issues of maritime applications which are related to the safety of life and property at sea, distress and urgency alerting/calling, collision avoidance, long-range identification and tracking (LRIT), etc. The typical systems are the global maritime distress and safety system (GMDSS), the global navigation satellite systems, and the marine satellite communication systems.

  (1) GMDSS

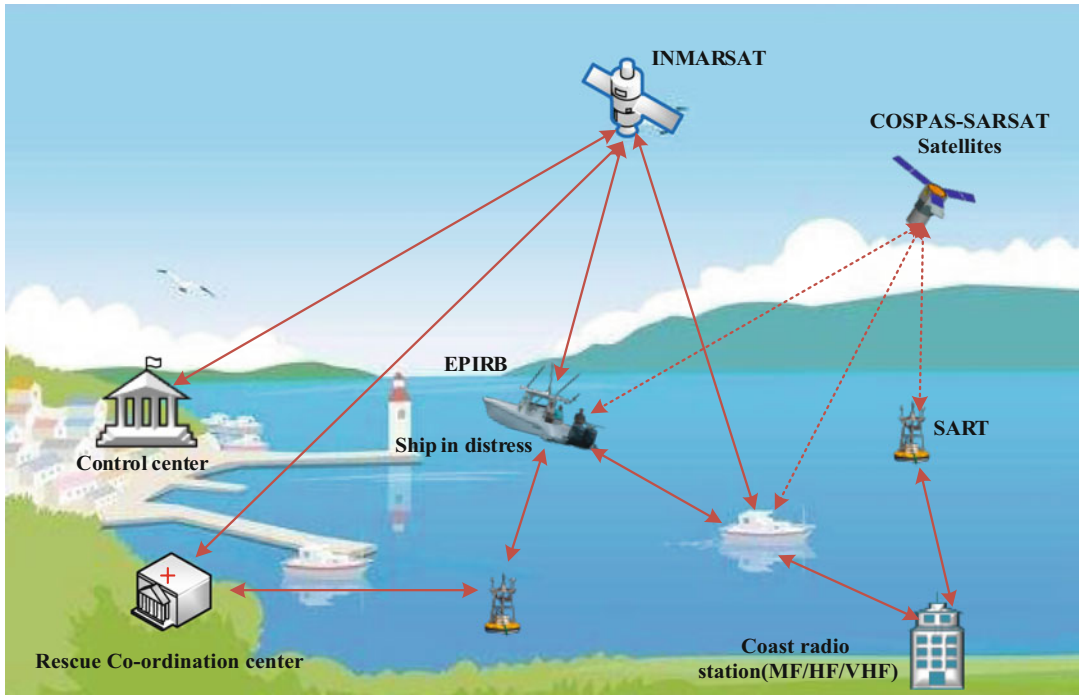
  The global maritime distress and safety system (GMDSS), described in the Safety of Life at Sea (SOLAS) Chapter IV Convention, is an immense integrated global communication network mainly composed of three parts: satellite communication systems (INMARSAT (International Maritime Satellite Organization) and COSPAS-SARSAT (Global Satellite Search and Rescue System)), terrestrial radio communication systems (i.e., coastal radio station), and maritime safety information broadcasting systems. Figure 1 shows its concept map. The system is intended to perform the following functions: alerting (including position determination of the unit in distress), search and rescue coordination, locating, maritime safety information broadcasts, general communications, and bridge-to-bridge communications. Once distress alerts appear, onshore search and rescue agencies and nearby shipping departments will promptly issue alerts through satellite communications and ground communications technology so that they can assist in coordinating search and rescue operations. According to the GMDSS, all cargo or passenger ships with a total tonnage of 300 or more must be equipped with radio equipment that conforms to international standards which guarantees maritime safety to a certain extent.

  (2) The global navigation satellite systems

  The global navigation satellite systems (GNSS) is a space-based radio navigation and positioning system that uses satellites to provide autonomous geo-spatial positioning on the earth's surface or near-Earth space. It can be used for providing position and navigation or for tracking the position of something fitted with a receiver (satellite tracking) which is widely applied for the safety of maritime navigation. Common systems include the United States' Global Positioning System (GPS), Russia's GLONASS, China's BeiDou Navigation Satellite System (BDS), and the European Union's GALILEO. Table 1 shows the details.

  (3) The marine satellite communications

  The marine satellite communications in the UHF band have global or regional coverage and are commonly deployed on ships and vessels to fulfill different communication requirements despite their high cost. As the commercially provided services, they can implement multiple communication services (voice, data, e-mail, SMS, crew calling, telex, facsimile, remote monitoring, tracking (position reporting), etc.) to guarantee the normal operation of daily cruise and satisfy the needs of crew members to contact with their families. According to the status of satellites, they can be geostationary or non-geostationary. Worldwide, typical marine satellite communication systems mainly include VSAT, Inmarsat systems,



Maritime Internet of Vessels, Fig. 1 Concept map of GMDSS

Maritime Internet of Vessels, Table 1 The global navigation satellite systems

System	BDS	GPS	GLONASS	GALILEO
Coding	CDMA	CDMA	FDMA	CDMA
Frequency (GHz)	1.561098 (B1) 1.589742 (B1–2) 1.20714 (B2) 1.26852 (B3)	1.563–1.587 (L1) 1.215–1.2396 (L2) 1.164–1.189 (L5)	1.593–1.610 (G1) 1.237–1.254 (G2) 1.189–1.214 (G3)	1.559–1.592 (E1) 1.164–1.215 (E5a/b) 1.260–1.300 (E6)
Precision	10 m (public) 0.1 m (encrypted)	15 m (no DGPS or WAAS)	4.5–7.4 m	1 m (public) 0.01 m (encrypted)

iridium satellite system, etc. Table 2 mainly summarizes the correlative systems widely employed in maritime communication.

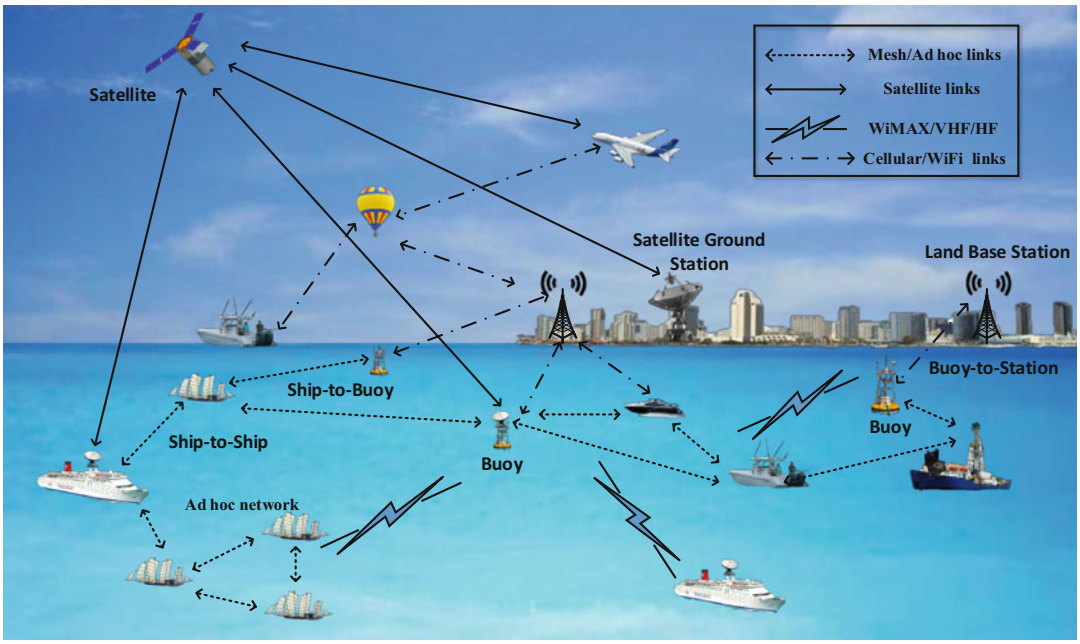
In addition, it should be noted that as a supplement to conventional communications and emergency communications, coastal mobile communications mixing with cellular or other wireless communication technologies have unique communication advantages in offshore areas. The offshore coverage of these shore-based communication systems could provide reliable communication guarantee for ports, wharfs, channel management, mariculture, and salvage at sea (Minghua et al. 2017), such as GSM/GPRS, 3G, 4G, Wi-Fi, WiMAX, mesh networks, ad hoc networks, and short-range devices like ZigBee and Bluetooth links. Although they could offer higher transmission rate, the communication distance is limited in range. A heterogeneous network integrated with above communication technologies is shown in Fig. 2.

Further Development

Currently, with the high development of maritime industry and activities, the demand for maritime communications services is increasing, which

Maritime Internet of Vessels, Table 2 The marine satellite communication systems

Systems	Services	Data rate	Communication distance
EPIRB	Alarm, identification, positioning, and locating	400 bps	Long distance
Iridium	Positioning, voice, data, paging, short messages, Internet services	Data transmission 2.4 kbps; Internet services 9.6 kbps	
VSAT	Data, voice, image, video, Internet services	Downlink 64 MbpsUplink 2 Mbps	
Inmarsat	Voice, fax, data, video, etc.	9.6~128 kbps	
Global Xpress(Inmarsat-5)	Voice, data, video, VOIP, etc.	Downlink 50 MbpsUplink 5 Mbps	
Globalstar	Voice, fax, data, short information, location, etc.	Voice 2.4/4.8/9.6 kbpsData 7.2 kbps	



Maritime Internet of Vessels, Fig. 2 A heterogeneous network for Maritime Internet of Vessels

puts forward higher requirements for the safety, reliability, and effectiveness of maritime communications. The traditional communication systems can hardly afford for the extreme volume of data generated by navigation services, video monitoring, and multimedia downloading, and new technologies and schemes are in the pipeline. Following the Internet of Things and Internet of Vehicles, the Maritime Internet of Vessels has come into people's vision. Maritime autonomous surface ships (MASS) are playing a very important role in the development of Maritime Internet of Vessels, and many new technologies are being explored to enhance the communication efficiency for the Maritime Internet of Vessels.

- MASS

  In recent years, ship intellectualization has become the general trend of global intelligent shipping. In order to reduce the difficulty of ship control and management, reduce

man-made misoperation, improve the safety of equipment and ship operation, optimize ship navigation, reduce costs, and increase profits, the research on MASS has been carried out worldwide (http://www.sumia.org/newsdetails.aspx?id=2760).

According to the definition of "intelligent ship specification" promulgated by China Classification Society (CCS) in 2016, "intelligent ships" refers to the use of sensors, communications, Internet, and other technical means to automatically perceive and obtain information and data of the ship itself, marine environment, logistics, ports, and other aspects. Based on computer technology, automatic control technology, and large data processing and analysis technology, the intelligent operation of ships in the aspects of navigation, management, maintenance, and cargo transportation should be realized, so as to make the ships safer, more environmentally friendly, more economical, and more reliable. The functions of intelligent ships are composed of intelligent navigation, intelligent hull, intelligent engine room, intelligent energy efficiency management, intelligent cargo management, and intelligent integrated platform (https://blog.csdn.net/weixin_37978606/article/details/80957137). Figure 3 summarized four development phases for MASS. Now it is in the transitional stage from the first phase to the second phase, and the fourth phase is the era of full automation.

In order to realize the full intellectualization of ships, there are still many problems that had to be overcome. First of all, the communication volume is large because of the need to transmit a large number of sensor information and equipment status information, as well as radar images, sea video, and so on. The maritime communication systems should be able to fulfill the requirements of high bandwidth, low delay, low cost, etc. Besides, adequate attention should be paid to the maritime network security. With the increasing of ship-shore communication, more and more ships are exposed which makes the security of maritime communication systems becomes increasingly important. How to avoid hacker attacks, key information leaks, and emergency plans of MASS in network attacks are all issues that need further study.

- Application of new technologies

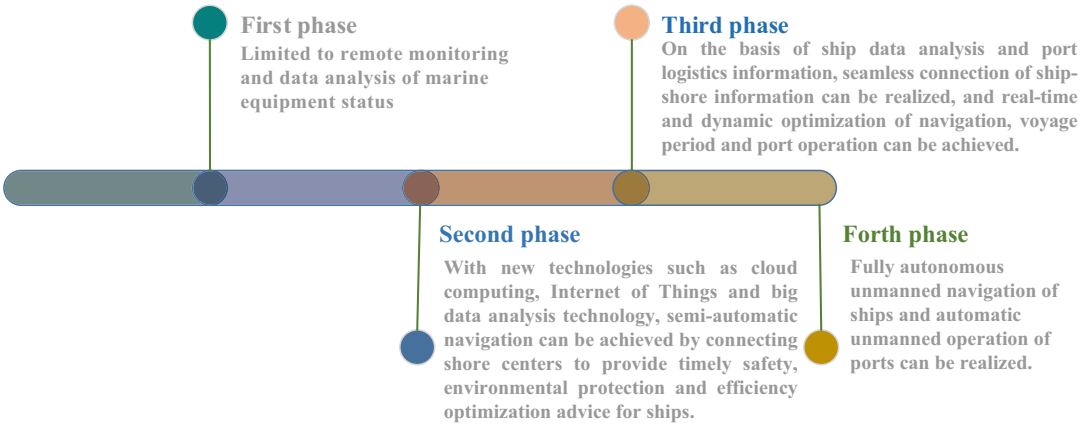
With the refinement, comprehensiveness, and humanization of ship management, the amount of data is increasing rapidly. How to store and manage massive data effectively has become more and more important. The establishment of Maritime Internet of Vessels needs to be data-centric, combined with Internet technologies and Internet of Things technologies, to realize the interconnection between vessels and vessels, people and vessels, vessels and cargo, and vessels and shore-based facilities. For example, the fog/edge computing, information perception and interaction technologies, and big data fusion and analysis technologies can be applied in the field of Maritime Internet of Vessels.

(1) Fog/edge computing

Currently, continuous development of maritime business has fuelled faster growth in the amount of data and information acquired from ships, cargo, operators, and navigation environment. It has become increasingly important to achieve effective transmission, storage, and management of these massive data although they require massive computing resources, storage space, and communication bandwidth. As an extension of cloud computing, fog/edge computing can be employed to extend computing, storage, and networking resources to the network edge and alleviate the heavy burden of large amounts of data on the network (Ni et al. 2018).

(2) Information perception and interaction technology

The application of Maritime Internet of Vessels cannot be separated from the support of information sensing technology. With the increasing concentration of maritime traffic routes, the increasing total



Maritime Internet of Vessels, Fig. 3 Development of MASS (http://www.sumia.org/newsdetails.aspx?id=2760)

number of vessels, and the sharp rise of communication equipment, information transmission and interaction between different equipment become more and more important, which puts forward higher requirements for information collection, identification, management, and integration. In order to reduce information redundancy and error, improve the information system of ship networking, and improve the intelligent level of Maritime Internet of Vessels, it is necessary to strengthen the research of information perception and interaction technology.

(3) Software-defined network

SDN (software-defined network) is a new network architecture, which can separate network control from traditional hardware devices. Since most of the maritime communication nodes are in motion, the business environment of the network is constantly changing. For the traditional network, if the business needs change, it is quite tedious to modify the configuration of network equipment. Therefore, in order to better meet the special requirements of Maritime Internet of Vessels, the flexibility and agility of the network become more and more important, and SDN can be taken into account to address such issues.

(4) Big data analysis technology

With the continuous development of maritime communications, marine information sensing and communicating equipment, which covers a variety of information transmitting means such as acoustic, optical, and electromagnetic, is continuously generating massive data which have shown the characteristics of huge volume, various data types, fast flow speed, and high value (You and Wei 2018). How to analyze and process these data and improve the breadth and depth of data utilization is a critical issue. It is urgent to strengthen the research of big data analysis technologies, so as to promote the rapid development of Maritime Internet of Vessels.

In addition, the layout of customized systems for Maritime Internet of Vessels should be accelerated, the integration of ship-shore intelligent information should be realized, and a visual sharing platform of information for Maritime Internet of Vessels to ensure the safety of ship navigation and intra-ship communication can also be taken into account. At the same time, the goals of saving energy, reducing operation and maintenance costs, and improving operational efficiency can be achieved. With all kinds of possible network access and interconnection, the ubiquitous links

between vessels and vessels and people can be realized; the intelligent perception, recognition, and management of vessels and processes can also come true, which makes vessels and vessels, vessels and waterways, and vessels and people more interoperable; and the maritime navigation will become more intelligent, controllable, safe, and environmentally friendly.

Conclusion

With the development of shipping industry, the amount of data and information acquired from vessels, cargo, operators, and navigation environment is increasing, and the traditional maritime communication systems can hardly fulfill the development requirements at present. This entry makes a detailed investigation about the existing networking modes for Maritime Internet of Vessels and points out the further development. In 5G era, great efforts should be made to connect more information units for Maritime Internet of Vessels and provide more information sharing services through real-time interconnection, and then the development of intelligent shipping can be further promoted, so that the safety and efficiency of shipping become predictable.

Cross-References

- ▸ Edge Data Centers in Fog Computings
- ▸ Maritime Internet of Vessels
- ▸ MIMO Satellite
- ▸ Software-Defined Wireless Networking

References

de Sousa MM, Pelas EG 2011 Digital selective calling system for controlling of communications by terrestrial waves on the Maritime mobile service. In: 2011 SBMO/IEEE MTT-S international microwave and optoelectronics conference (IMOC 2011), Natal, pp 612–617

Lázaro F, Raulefs R, Wei W et al (2017) VHF data exchange system (VDES): an enabling technology for maritime communications. Ceas Space J 2017(4): 1–9

Minghua X, Youmin Z, Erhu C et al (2017) Current status and challenges of marine communications. Chin Sci Inf Sci 2017(06):5–23

Ni J, Zhang K, Lin X, Shen XS (2018) Securing fog computing for internet of things applications: challenges and solutions. IEEE Commun Surv Tutorials 20(1):601–628

Weintrit A, Weintrit A (2011) Development of the IMO e-Navigation concept – common Maritime data structure[J]. Commun Comput Inf Sci 239:151–163

You H, Wei Z (2018) Big data technology for maritime information sensing. Command Inf Syst Technol 9(2):1–7

Market Entry

Xuehe Wang and Lingjie Duan
Engineering Systems and Design Pillar, Singapore University of Technology and Design, Singapore, Singapore

Synonyms

Market entry strategy

Definitions

Market entry is the activities associated with bringing a product or service to a targeted market. A market can denote the *primary market* and *secondary market*. The *primary market* refers to the market where the companies sell new products to the customers for the first time, while the *secondary market* is one in which the customers trade the products among themselves.

Historical Background

As the rapid development of the wireless technology, numerous wireless-related services (e.g., 4G wireless service, femtocell service, mobile data trading service) are emerging. When facing the new services, asymmetric behaviors of the wireless service providers (WSPs) and users are observed. For example, the providers may have different timing of introducing the new wireless

services to the market, and the heterogenous users have different response to these services. Game theory provides solid mathematic tools for analyzing the behaviors of providers and users in wireless networks (Han 2012). One of the most popular examples of game theory in wireless networks is the power control problem, where the transmit power level of a wireless user can pose a positive or negative impact on the transmission rate and quality of service (QoS) of the other users due to interference. By noting the competition between the users, Shah et al. (1998) first formulate the interactions of wireless users as a non-cooperative game and obtain the equilibrium transmit power of all users. Following the pioneer work on non-cooperative games in power control, game theory has been implemented to plenty of application areas in wireless network market entry, e.g., the adoption of femtocell service (Shetty et al. 2009; Yun et al. 2011), new wireless technology introduction (Sen et al. 2010), and secondary mobile data trading market (Zheng et al. 2015; Wang et al. 2016). Musacchio et al. (2006) develop a game theoretic model for studying the upgrade timing of two interconnected Internet service providers (ISPs), where the network upgrade of one ISP has positive network effect on the other. This "free-rider" effect makes it reluctant for the providers to upgrade and instead enjoy the benefits of other providers' upgrading. Duan et al. (2015) further discuss the upgrade timing of 4G network by taking the user churn into consideration, which weakens the free-riding effect. When introducing new technologies to the network, the competition between entrant and incumbent network technologies may exist. Dynamic games can be used to analyze the users' adoption dynamics of a new network technology in the presence of an incumbent technology (Sen et al. 2010). To understand the interactions between the providers and their users, Stackelberg games provide an ideal tool to study the hierarchical decision-making among the players, in which the providers (leaders) declare and announce their strategies first and then the users (followers) react selfishly based on the strategies of the providers. Using non-cooperative Stackelberg games, Duan et al. (2013, 2016) study the economic incentive for an operator to introduce the femtocell service on top of its existing macrocell service centrally and distributedly, in which users submit their bandwidth demands to the operator based on the service and corresponding price and the macrocell operator and femtocell operator (same in central case) can adjust the bandwidth price to maximize their profits. Wang et al. (2016) also use Stackelberg games to study a secondary market in which mobile data users are allowed to trade unused data.

Foundations

Game theory is a basic theory in economics used for the analysis of strategic decision-making of intelligent rational players. The normal-form representation of a game consists of three parts (Gibbons 1992):

- The game players $\mathcal{N} = \{1, 2, \ldots, N\}$.
- The available strategy set S_i of each player $i \in \mathcal{N}$. Denote the strategy chosen by player i as $s_i \in S_i$ and the strategies of all players except player i as s_{-i}. Let $S = \prod_i S_i$ denote the set of all possible strategy profiles.
- The utility U_i received by player i for each possible combination of strategies $s \in S$.

Non-cooperative Games

There are many types of games possessing different properties suited for analyzing variety of situations. *Non-cooperative games* are used ubiquitously for the analysis of market entry in wireless networks, in which many players compete for finite resources and each player behaves selfishly and independently by maximizing his own utility, given the possible strategies of other players and their impact on the player's utility (Osborne 2004). For example, the wireless providers compete over new wireless technologies and pricing strategies to attract users, and the users are involved in many non-cooperative situations such as allocation of bandwidth, transmit power, etc.

To describe the best strategy for a typical player under any given strategy profile of all

other players, the concept of *best response* is introduced.

Definition 1 For each player i, the best response is the strategy such that

$$BR_i(s_{-i}) := \{s_i \in S_i : U_i(s_i, s_{-i}) \geq U_i(s_i', s_{-i}), \forall s_i' \in S_i\} \quad (1)$$

Nash equilibrium is the stable outcome of a game, which is defined as the strategy profile of all players where none of the players can improve its utility by a unilateral move.

Definition 2 A profile $s^* \in S$ is called a pure strategy Nash equilibrium, if for each player $i \in \mathcal{N}$:

$$U_i(s_i^*, s_{-i}^*) \geq U_i(s_i, s_{-i}^*), \quad \forall s_i \in S_i. \quad (2)$$

Stackelberg Games

In many non-cooperative games, a hierarchy among the players might exist where a fraction of the players (leaders) make decisions first and the rest of the players (followers) react selfishly based on the strategies of the leaders. This kind of games forms special classes of non-cooperative games called *Stackelberg games*, which is widely used in economic literature (He et al. 2007).

For a two-stage Stackelberg game, let $\mathcal{N}_L$ denote the set of leaders who announce their strategies first at Stage I and $\mathcal{N}_F$ denote the set of followers who make decisions accordingly at Stage II. Let $s_{i,L}$ denote the strategy of leader $i \in \mathcal{N}_L$ and s_L denote the strategy profile of all leaders. Let $s_{i,F}$ denote the strategy of follower $i \in \mathcal{N}_F$ and s_F denote the strategy profile of all followers. Then, the concept of Stackelberg equilibrium can be characterized as follows.

Definition 3 A profile $\{s_L^*, s_F^*\} \in S$ is a Stackelberg equilibrium strategy, if for each follower $i \in \mathcal{N}_F$:

$$U_i(s_{i,F}^*, s_{-i,F}^*, s_L^*) \geq U_i(s_{i,F}', s_{-i,F}^*, s_L^*), \quad \forall s_{i,F}' \in S_i, \quad (3)$$

and for each leader $i \in \mathcal{N}_L$:

$$U_i(s_{i,L}^*, s_{-i,L}^*, s_F^*(s_{i,L}^*, s_{-i,L}^*)) \geq U_i(s_{i,L}', s_{-i,L}^*, s_F^*(s_{i,L}', s_{-i,L}^*)), \quad \forall s_{i,L}' \in S_i. \quad (4)$$

Generally, the non-cooperative Stackelberg game is used to analyze the strategic interaction between the wireless providers (who decide the market entry timing and pricing of the services) and wireless users (who decide whether to subscribe to the new services and how much to buy based on their utilities) in the wireless market.

Key Applications

Network economics and pricing can be used to model and analyze many market entry problems in wireless networks. Some key applications are summarized in the following.

Market Entry Timing of New Technologies

As with the emergence of new technologies in wireless networks, the wireless providers should decide whether to deploy the new technology to improve the QoS and the upgrade timing. Generally, the cost of new technology upgrade decreases over time as the technology matures. An existing user can switch to the new technology service of the same provider or a different provider according to the switching cost. Being the first to upgrade to new technology, a provider increases its market share but spends more money on the upgrade. Therefore, the provider decides its upgrade time by trading off increased market share and upgrade cost.

Duan et al. (2015) study the upgrade timing of the fourth generation (4G) technology on the basis of third generation (3G) of cellular wireless networks by developing a game theoretic model for analyzing the competitive providers' interactions. The main results show that:

- When the upgrade cost is low, the 4G monopolist upgrades at the earliest available time; otherwise it postpones its upgrade.
- Under no inter-network switching case, i.e., no user switches operators duo to high switching cost, competitive operators upgrade simultaneously, no matter how much the upgrade cost is.
- Under practical inter-network switching case, competitive operators select different upgrade times to avoid severe competition. By upgrading early, an operator captures a large market share and the 4G's QoS improvement which can compensate for the large upgrade cost. The other operator, however, postpones its upgrade to avoid severe competition and benefits from cost reduction. The availability of 4G upgrade may decrease both operators' profits because of the increased competition, and paradoxically, their profits may increase with the upgrade cost.

Secondary Service Entry

To improve the QoS and attract more users, the WSPs are eager to bring new technologies and services to the wireless market on top of the existing services, e.g., the introduction of femtocell service on top of the existing macrocell service. Femtocell technology is introduced to solve the poor signal reception problem for indoor users, as the high-frequency and low-power wireless signals from macrocell base stations often have difficulty in effectively traveling through walls. By bringing the femtocell service to the market, the macrocell operator should consider two main questions:

- Can the macrocell operator benefit from the femtocell service?
- How should the macrocell operator allocate and price the spectrum bands?

Users experience different channel conditions and spectrum efficiencies with the macrocell service, but all of them achieve a high spectrum efficiency with the femtocell service by deploying a femtocell at home. Thus, users have different preferences between macrocells and femtocells and should decide which service to choose and how much bandwidth to request based on the price of each service.

The macrocell operator can provide the femtocell service itself or lease spectrum to a femtocell operator. For the situation when femtocells are managed by the same operator that controls the macrocells, Duan et al. (2013) model the interactions between the operator and users as a two-stage Stackelberg game. In Stage I, the operator determines bandwidth allocated to femtocell and macrocell service and the corresponding prices. In Stage II, each user decides which service to choose and how much bandwidth to purchase. It reveals that:

- If femtocell service has the same maximum coverage as macrocell service, the operator will choose to provide femtocell service only, as this leads to a better user quality of service and a higher operator profit.
- If all users have reservation payoffs which are what they can achieve with the original macrocell service, then the operator will always continue providing the macrocell service (with or without the femtocell service) to ensure that all users achieve payoffs no worse than their reservation payoffs.
- When multiple femtocells can reuse the same spectrum resource, the operator has more incentives to allocate spectrum to femtocell, and the femtocell price will be reduced. Moreover, when femtocell service incurs operational cost to the operator or has a smaller coverage than the macrocell service, the operator will always serve users by dual services. Furthermore, as cost decreases or coverage increases, more users are served by the femtocell service.

For the situation when the macrocell operator leases spectrum to independent femtocell operators, Duan et al. (2016) model the interactions between a macrocell operator, a femtocell operator, and end users as a three-stage dynamic game.

In Stage I, the macrocell operator decides bandwidth allocations to femtocell service and macrocell service and the macrocell price. In Stage II, the femtocell operator decides how much bandwidth to lease from the macrocell operator and the femtocell retail price. In Stage III, each user decides which service to choose and how much bandwidth to request. It reveals that:

- The macrocell operator has more incentive to lease spectrum to the femtocell operator when its spectrum capacity is small, as femtocell services can help cover more users and improve the utilization efficiency of the limited spectrum resource. Otherwise, the macrocell operator has less incentive to enable femtocell service due to the significant competition.
- In general, femtocell service can increase both the total consumer surplus and social welfare. However, some users who originally receive good service in macrocell might experience a payoff drop after the femtocell service is introduced, due to fewer resources being allocated to the macrocell service and a higher macrocell price.

Secondary Data Trading Market Entry

Another key application is the secondary data market entry for mobile data trading. In reality, a lot of data users subscribe to the data plans with certain amount of data quota which will expire in a billing cycle (e.g., 1 month). Some users can easily use up their data quota and may pay for costly data over-usage, while some users cannot use up their data quota and have leftover data. Motivated by users' diverse usage behavior (more or less than the subscribed data quotas), the secondary data trading market appears in which cellular users can trade data plans with each other, e.g., 2CM (2nd exchange market) data trading platform introduced by China Mobile Hong Kong. Yu et al. (2015), Zheng et al. (2015), and Yu et al. (2017a) discuss such a market, where data users submit bids to buy and sell data. The provider serves as the middleman to match buyers and sellers, as well as charges administration fees and/or pockets the differences between the buyer and seller prices. Moreover, Yu et al. (2017b) study the mobile data trading problem under future data demand uncertainty. Taking the latest market and usage information into consideration, an algorithm is designed to help estimate the users' risk preference and provide trading recommendations dynamically. Wang et al. (2016) further propose a new type of user-initiated network for cellular users to freely trade data plans by leveraging personal hotspots (PHs) without the wireless provider's involvement. A user with data surplus can set up a PH and share the cellular data connection to another user with data deficit in the vicinity. Moreover, the WSP's countermeasures to the PH market and the competition between WSPs are investigated. Following this work, Wang et al. (2017) study the PH-enabled data-plan sharing between local users and travelers by taking into account the information uncertainty at the traveler side. Lacking the selfish PHs' information, the expected cost of the traveler is higher than that under the complete information, but the gap diminishes as the PHs' spatial density increases.

Future Directions

The economics of market entry can be applied to analyze the upgrading time and interactions between the WSPs and users for many new technologies and services, e.g., the introduction of 5G, smart home.

Cross-References

- ▸ Behavioral Economics
- ▸ Dynamic Pricing
- ▸ Game Theory

References

Duan L, Huang J, Shou B (2013) Economics of femtocell service provision. IEEE Trans Mobile Comput 12(11):2261–2273

Duan L, Huang J, Walrand J (2015) Economic analysis of 4G upgrade timing. IEEE Trans Mobile Comput 14(5):975–989
Duan L, Shou B, Huang J (2016) Capacity allocation and pricing strategies for new wireless services. Prod Oper Manag 25(5):866–882
Gibbons R (1992) A primer in game theory. Harvester Wheatsheaf, New York
Han Z (2012) Game theory in wireless and communication networks: theory, models, and applications. Cambridge University Press, Cambridge
He X, Prasad A, Sethi SP, Gutierrez GJ (2007) A survey of stackelberg differential game models in supply and marketing channels. J Syst Sci Syst Eng 16(4):385–413
Musacchio J, Walrand J, Wu S (2006) A game theoretic model for network upgrade decisions. In: Proceedings of the 44th annual allerton conference on communication, control, and computing, Monticello, pp 191–200
Osborne MJ (2004) An introduction to game theory, vol 3. Oxford university Press, New York
Sen S, Jin Y, Guérin R, Hosanagar K (2010) Modeling the dynamics of network technology adoption and the role of converters. IEEE/ACM Trans Netw (TON) 18(6):1793–1805
Shah V, Mandayam NB, Goodman DJ (1998) Power control for wireless data based on utility and pricing. In: Proceedings of the ninth IEEE international symposium on personal, indoor and mobile radio communications, vol 3. IEEE, pp 1427–1432
Shetty N, Parekh S, Walrand J (2009) Economics of femtocells. In: Proceedings of IEEE global telecommunications conference (GLOBECOM). IEEE, pp 1–6
Wang X, Duan L, Zhang R (2016) User-initiated data plan trading via a personal hotspot market. IEEE Trans Wirel Commun 15(11):7885–7898
Wang F, Duan L, Niu J (2017) Optimal pricing of user-initiated data-plan sharing in a roaming market. In: Proceedings of IEEE global telecommunications conference (GLOBECOM). IEEE, pp 1–6
Yu J, Cheung MH, Huang J, Poor HV (2015) Mobile data trading: a behavioral economics perspective. In: 13th international symposium on modeling and optimization in mobile, ad hoc, and wireless networks (WiOpt). IEEE, pp 363–370
Yu J, Cheung MH, Huang J (2017a) Economics of mobile data trading market. In: 15th international symposium on modeling and optimization in mobile, ad hoc, and wireless networks (WiOpt). IEEE, pp 1–8
Yu J, Cheung MH, Huang J, Poor HV (2017b) Mobile data trading: behavioral economics analysis and algorithm design. IEEE J Sel Areas Commun 35(4):994–1005
Yun S, Yi Y, Cho DH, Mo J (2011) Open or close: on the sharing of femtocells. In: Proceedings of IEEE conference on computer communications (INFOCOM). IEEE, pp 116–120
Zheng L, Joe-Wong C, Tan CW, Ha S, Chiang M (2015) Secondary markets for mobile data: feasibility and benefits of traded data plans. In: Proceedings of IEEE conference on computer communications (INFOCOM). IEEE, pp 1580–1588

Market Entry Strategy

▸ Market Entry

Massive MIMO

Erik G. Larsson and Emil Björnson
Department of Electrical Engineering (ISY), Linköping University, Linköping, Sweden

Synonyms

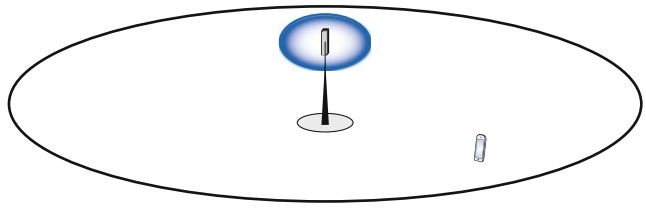
Full-dimension MIMO; Large-scale antenna systems (LSAS); Large-scale MIMO; Massive multiuser MIMO; Very large MIMO

Definition

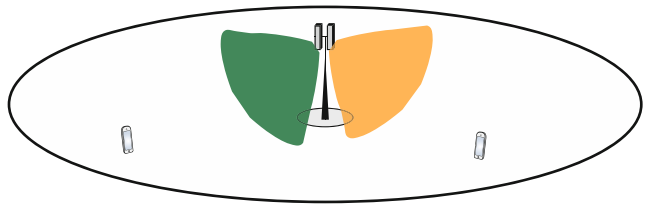
Technology that uses large numbers of phase-coherently operating antennas at wireless base stations to serve a multitude of users by spatial multiplexing.

Historical Background

The wireless transmission from a single antenna has a predefined directivity, independent of where the users is located; see Fig. 1. The idea of using multiple antennas at wireless transceivers goes back a long time, at least to Alexanderson (1919), where analog transmit beamforming was used to direct the signal toward the receiving user and Peterson et al. (1931) where spatial diversity from receive beamforming was discussed. Modern incarnations of the idea have particularly included the concept of spatial-division multiple access (SDMA), proposed in the early 1990s by Swales et al. (1990), Anderson et al. (1991), and Roy and Ottersten (1991), and techniques for the suppression of interference using antenna arrays by Winters (1984). The novelty of these methods was to enable users to be multiplexed spatially,

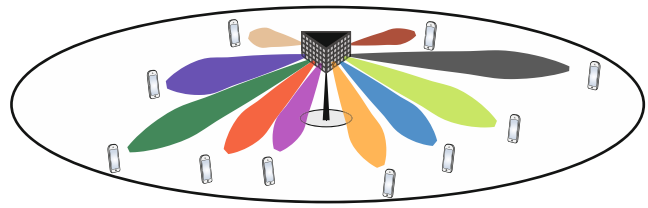


Massive MIMO, Fig. 1 In single-antenna transmission, the transmission has a predefined directivity, independent of the location of the receiving user. One example is omnidirectional transmission, as shown in the figure



Massive MIMO, Fig. 2 In multiuser MIMO transmission with few antennas, the spatially multiplexed signals are broadly directed with beamforming toward the respective receiving users

Massive MIMO, Fig. 3 In Massive MIMO transmission, many users are spatially multiplexed with narrow beamforming, using an array of very many antennas



instead of in time or frequency, to increase the spectral efficiency of a wireless network; see Fig. 2.

Early attempts to efficiently utilize multiple transceiver antennas did not succeed well. There are many potential reasons for this: first, most methods made simplifying assumptions on the propagation environment that do not hold in practice; second, the hardware was not sufficiently advanced at the time; and, third, the corresponding information theory was not yet developed and hence important insights into the significance and difficulties in acquiring accurate channel state information were largely missing. An information-theoretic framework first took form in Caire and Shamai (2003), Goldsmith et al. (2003), and Viswanath and Tse (2003). Subsequently, SDMA became more commonly known as multiuser multiple-input multiple-output (MIMO), which reflects the fact that the base stations have multiple antennas and there are multiple users.

Despite the mentioned difficulties, elementary forms of SDMA technology were successfully showcased in the mid-1990s by Anderson et al. (1996). ArrayComm in cooperation Kyocera made the first commercial deployment of SDMA as an overlay to the Personal Handyphone System (PHS) in a limited part of Southeast Asia.

Foundations

Massive MIMO is an evolved and practically useful form of multiuser MIMO. The key concept is to use a very large number of phase-coherently operating antennas at the base station, as illustrated in Fig. 3. Since its inception in Marzetta (2010), it has quickly evolved from

a wild theoretical concept with an "unlimited" number of antennas to a practical technology that has been demonstrated in field trials, as described in Harris et al. (2017) and Malkowsky et al. (2017) among others and included as one of the main features of the 5G New Radio; see Parkvall et al. (2017).

Exact definitions of "Massive MIMO" differ slightly among different researchers, but the underlying idea is always to equip base stations with arrays of many electronically steerable antennas. These antennas serve many users simultaneously, in the same time-frequency resource. A minimum of 64 base station antennas is commonly considered, and the goal is to support spatial multiplexing of tens of users, each equipped with one or multiple antennas. The very large number of antennas is a design choice to limit inter-user interference and provide a beamforming gain wherever the user is. The textbooks by Marzetta et al. (2016) and Björnson et al. (2017) advocate the use of time-division duplexing (TDD) operation, which permits exploitation of the uplink-downlink reciprocity of the radio propagation. Specifically, the base station uses channel estimates obtained from uplink pilots transmitted by the users to learn the channels in both directions. These estimates are then used for multiuser precoding in the downlink and multiuser detection in the uplink. With TDD operation, Massive MIMO is entirely scalable with respect to the number of base station antennas, in the sense that the amount of resources spent on channel estimation is independent of the number of antennas.

Less restrictive definitions of Massive MIMO are also common. While the preferred operation of Massive MIMO is in TDD mode, where reciprocity between the uplink and downlink holds, operation in frequency-division duplex (FDD) mode is possible in some cases, e.g., using the concepts provided by Adhikary et al. (2013) and Choi et al. (2014), but Flordelis et al. (2018) show that there is a significant performance penalty. While the propagation models in much Massive MIMO literature suggest sub-6 GHz operation, see e.g., Marzetta et al. (2016), Björnson et al. (2017), and references therein variants of Massive MIMO are also considered for millimeter wave frequency bands.

Three key features of Massive MIMO are as follows:

- High spectral efficiency (measured in bits/s/Hz), primarily attributed to the spatial multiplexing gain
- Provision of uniformly good quality of service in the entire coverage area, by means of beamforming and power control
- Support for high user mobility by efficient channel acquisition (primarily in TDD operation)

The word "massive" refers to the number of antennas and not to the physical size of a base station. Since low-gain antennas are typically used, the antenna arrays can have attractive form factors, e.g., in the 2 GHz band, a half-wavelength-spaced rectangular array with 200 dual-polarized elements is on the order of 1.5×0.75 meters large. At higher carrier frequencies, the arrays will be even smaller, since the dimensions scale with the carrier wavelength. No specific array geometry is required or assumed in Massive MIMO, and distributed antenna deployments are feasible.

The massive number of antennas can give rise to two desirable propagation phenomena. The first is channel hardening, which is a form of massive spatial diversity implying that the link performance fluctuates very little over time and frequency, even when communicating over a fading channel. The second is favorable propagation, which is the ability for the base station to separate the users in the spatial domain; technically, the channel responses associated with different users become nearly orthogonal with an increasing number of antennas.

Information theory, signal processing, and power control for Massive MIMO are well understood, as demonstrated by Marzetta et al. (2016) and Björnson et al. (2017). An important insight is that the link performance over a fading channel can be rigorously quantified in terms of an effective signal-to-interference-and-noise

ratio (SINR), which depends only on the average channel properties between the users and the base stations and on certain system parameters. These expressions take into account the effects of all significant physical phenomena: small-scale and large-scale fading, fading correlation, intra- and intercell interference, channel estimation errors, and pilot interference from pilot reuse in the network (also known as pilot contamination). In the moderately large number-of-antennas regime, the closed-form capacity bounds that result from these effective SINR expressions become convenient proxies for the link performance achievable with practical coding and modulation, power control, and resource allocation schemes.

Massive MIMO is not only a groundbreaking technology for wireless communications, which can vastly increase the spectral efficiency and uniformity of the quality-of-service over previous technologies. It is also an elegant and mathematically rigorous approach to teaching of wireless communications. A comprehensive analytical understanding of the key properties that determine wireless communication link and system performance is facilitated by the capacity bound analysis of Marzetta et al. (2016). Such analyses have simply not been possible before as the corresponding information theory was too complicated for any practical use.

Key Applications

Fifth-generation (and beyond) mobile access networks; Wireless communications with autonomous vehicles; Backhauling in small-cell networks

Cross-References

- ▸ Massive MIMO BDMA Transmission
- ▸ Millimeter Wave Massive MIMO
- ▸ Omnidirectional Transmission for Massive MIMO
- ▸ Per-Beam Synchronization for Millimeter-Wave Massive MIMO
- ▸ Pilot Reuse for Massive MIMO

References

Adhikary A, Nam J, Ahn JY, Caire G (2013) Joint spatial division and multiplexing–the large-scale array regime. IEEE Trans Inf Theory 59(10):6441–6463

Alexanderson EFW (1919) Transoceanic radio communication. Presented at a joint meeting of the American Institute of Electrical Engineers and the Institute of Radio Engineers, New York

Anderson S, Millnert M, Viberg M, Wahlberg B (1991) An adaptive array for mobile communication systems. IEEE Trans Veh Technol 40(1):230–236

Anderson S, Forssén U, Karlsson J, Witzschel T, Fischer P, Krug A (1996) Ericsson/Mannesmann GSM field-trials with adaptive antennas. In: Proceedings of IEE colloquium on advanced TDMA techniques and applications, London

Björnson E, Hoydis J, Sanguinetti L (2017) Massive MIMO networks: spectral, energy, and hardware efficiency. Found Trends® Signal Process 11(3–4):154–655. https://doi.org/10.1561/2000000093

Caire G, Shamai S (2003) On the achievable throughput of a multiantenna Gaussian broadcast channel. IEEE Trans Inf Theory 49(7):1691–1706

Choi J, Love D, Bidigare P (2014) Downlink training techniques for FDD massive MIMO systems: open-loop and closed-loop training with memory. IEEE J Sel Topics Signal Process 8(5):802–814

Flordelis J, Rusek F, Tufvesson F, Larsson EG, Edfors O (2018) Massive MIMO performance-TDD versus FDD: what do measurements say? IEEE Trans Wirel Commun 17(4):2247–2261

Goldsmith A, Jafar SA, Jindal N, Vishwanath S (2003) Capacity limits of MIMO channels. IEEE J Sel Areas Commun 21(5):684–702

Harris P, Malkowsky S, Vieira J, Bengtsson E, Tufvesson F, Hasan WB, Liu L, Beach M, Armour S, Edfors O (2017) Performance characterization of a real-time Massive MIMO system with LOS mobile channels. IEEE J Sel Areas Commun 35(6):1244–1253

Malkowsky S, Vieira J, Liu L, Harris P, Nieman K, Kundargi N, Wong IC, Tufvesson F, Öwall V, Edfors O (2017) The world's first real-time testbed for Massive MIMO: design, implementation, and validation. IEEE Access 5:9073–9088

Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans Wirel Commun 9(11):3590–3600

Marzetta TL, Larsson EG, Yang H, Ngo HQ (2016) Fundamentals of massive MIMO. Cambridge University Press, Cambridge

Parkvall S, Dahlman E, Furuskär A, Frenne M (2017) NR: the new 5G radio access technology. IEEE Commun Stand Mag 1(4):24–30

Peterson HO, Beverage HH, Moore JB (1931) Diversity telephone receiving system of R.C.A. communications, Inc. Proc IRE 19(4):562–584
Roy R, Ottersten B (1991) Spatial division multiple access wireless communication systems. US Patent No. 5,515,378
Swales SC, Beach MA, Edwards DJ, McGeehan JP (1990) The performance enhancement of multibeam adaptive base-station antennas for cellular land mobile radio systems. IEEE Trans Veh Technol 39(1):56–67
Viswanath P, Tse DNC (2003) Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality. IEEE Trans Inf Theory 49(8):1912–1921
Winters JH (1984) Optimum combining in digital mobile radio with cochannel interference. IEEE J Sel Areas Commun 2(4):528–539

Massive MIMO BDMA Transmission

Chen Sun, Xiqi Gao, and Li You
National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

Synonyms

Beam division multiple access (BDMA) transmission; Massive MIMO; Statistical channel state information

Definitions

Beam division multiple access (BDMA) transmission refers to a wireless multiple access scheme, where a base station employs massive antennas to generate beams and communicates with several users simultaneously. Multiple access is achieved in BDMA by assigning beams to individual users, and beams for different users are orthogonal (nonoverlapping).

Historical Background

In developing next-generation wireless networks, massive multiple-input multiple-output (MIMO) is a promising technology to deliver high spectral and power efficiencies (Rusek et al. 2013; Wang et al. 2014). In massive MIMO system, a base station (BS) employs a large number of antennas to simultaneously and jointly serve multiple mobile users. Assuming an environment of independent and identically distributed propagation MIMO channels, such a massive MIMO configuration has several attractive features (Marzetta 2010) including: (1) downlink channel vectors for different users become asymptotically orthogonal; (2) independent user beamforming can mitigate intra-cell interferences and uncorrelated noises; and (3) transmit power can be asymptotically low.

The aforementioned benefits of massive MIMO depend heavily on the availability of channel state information at the transmitter (CSIT). In practice, the CSIT availability hinges on time-division duplex (TDD) operation through uplink training and the channel reciprocity assumption. Yet, training overhead scales linearly with the total number of user antennas. When the number of users is large or when each user has multiple antennas, the overhead becomes prohibitively high. In addition, as user mobility increases, channel coherence time becomes relatively short. As a result, it becomes much more difficult to obtain accurate instantaneous CSIT for medium- or high-mobility user applications. Although the propagation channel itself is likely reciprocal in TDD, the uplink/downlink RF hardware chains at both BS and mobile transceivers are often not reciprocal (Choi et al. 2014). These practical limitations pose serious challenges to accurate CSIT acquisition in TDD systems, especially when dealing with high user mobility. For frequency-division duplex (FDD) systems without instantaneous channel reciprocity, the required number of independent pilot symbols for CSIT acquisition scales with the number of BS antennas, as does the CSIT feedback overhead (Jindal 2006). Therefore, it is likely to be more difficult to acquire global real-time CSIT for precoding in FDD.

Instantaneous CSIT acquisition represents a significant bottleneck; it is more desirable to exploit statistical CSIT for massive MIMO

systems. Statistical CSI varies over much larger time scales than instantaneous channel parameters. Moreover, the uplink and downlink statistics are usually reciprocal in both FDD and TDD systems (Barriac and Madhow 2004). Therefore, the statistical channel information can be more easily obtained by exploiting reciprocity, even for terminals equipped with multiple antennas.

Beam division multiple access (BDMA) transmission scheme is an effective technology (Sun et al. 2015), which can approach optimal transmission, when only statistical CSIT is available at the BS. In the beam domain, different beams transmit independent signals, and one beam transmits signal to at most one user. The BDMA transmission consists of the following steps: (1) acquiring channel statistics, (2) scheduling users and beams, (3) transmitting pilot and data in the uplink and downlink. By user and beam scheduling, beams assigned to different users are nonoverlapping (orthogonal). Thus, BDMA transmission decomposes massive MU-MIMO channel links to multiple small-scale SU-MIMO interference channel links. For each link, the number of transmit beams is much smaller than the number of antennas, and the channel state information between beams and users can be obtained via channel training. BDMA transmission can cope with the difficulties of instantaneous CSIT acquisition, reduce the inter-user interference, and increase the spectrum efficiency. Moreover, BDMA transmission can reduce the complexity of transceiver design and can be applied in TDD or FDD systems.

Beam Domain Channel

A single-cell system consists of one BS equipped with M antennas, and K users, each with N antennas, randomly distributed in the cell. For a physical channel model, suppose that there are P physical paths between the BS and users, and the pth path of the kth user has an attenuation of $a_{p,k}$, an angle $\varphi_{p,k}$ with the transmit antenna array, and an angle $\theta_{p,k}$ with the receive antenna array. For a wideband channel, after OFDM operation, the channel frequency response matrix in the ℓth subcarrier of the kth user is given by (Tse and Viswanath 2005):

$$\mathbf{H}_{k,\ell}^{d} = \sum_{p=1}^{P} a_{p,k} e^{-j2\pi d_{p,k}/\lambda_c} \mathbf{e}_r(\theta_{p,k}) \mathbf{e}_t^H(\varphi_{p,k}) e^{-j2\pi \ell \tau_{p,k}}, \quad (1)$$

where the superscript d means downlink, $d_{p,k}$ is the physical distance between transmit antenna 1 and receive antenna 1 along path p, λ_c is the carrier wavelength, and $\tau_{p,k}$ is the propagation delay associated with the pth path. Moreover, $\mathbf{e}_r(\theta) \in \mathbb{C}^{N\times 1}$ satisfying $\|\mathbf{e}_r(\theta)\|_2 = 1$ is the user antenna array response vector corresponding to the angle of arrival (AoA) θ, and $\mathbf{e}_t(\varphi) \in \mathbb{C}^{M\times 1}$ satisfying $\|\mathbf{e}_t(\varphi)\|_2 = 1$ is the BS antenna array response vector corresponding to the angle of departure (AoD) φ.

Assume that the array response vectors corresponding to distinct AoDs are asymptotically orthogonal with infinite number of antennas at the BS (You et al. 2015), i.e.:

$$\lim_{M\to\infty} \mathbf{e}_t^H(\varphi)\mathbf{e}_t(\eta) = \delta(\varphi - \eta). \quad (2)$$

At the user side, the angle $\theta_{n,k}$ is the sampling of receive signals. When the response vectors are orthogonal, i.e.:

$$\mathbf{e}_r^H(\theta_{n,k})\mathbf{e}_r(\theta_{n',k}) = \delta(n - n'), \quad (3)$$

users can separate these orthogonal direction signals perfectly. For uniform linear antenna arrays, uniform sampling of $\sin(\theta_{n,k})$, i.e., $\sin(\theta_{n,k}) = n/N$, is a classical choice for spatial angles (Sayeed 2002).

The channel matrix in (1) can be rewritten as

$$\mathbf{H}_{k,\ell}^{d} = \sum_{n=1}^{N}\sum_{m=1}^{M} \left[\tilde{\mathbf{H}}_{k,\ell}^{d}\right]_{nm} \mathbf{e}_r(\theta_{n,k})\mathbf{e}_t^H(\varphi_m) = \mathbf{U}_k \tilde{\mathbf{H}}_{k,\ell}^{d} \mathbf{V}^H \quad (4)$$

where $\mathbf{U}_k = \left[\mathbf{e}_r(\theta_{1,k}), \mathbf{e}_r(\theta_{2,k}), \ldots, \mathbf{e}_r(\theta_{N,k})\right] \in \mathbb{C}^{N\times N}$ is a unitary matrix and $\mathbf{V} = [\mathbf{e}_t(\varphi_1), \mathbf{e}_t(\varphi_2), \ldots, \mathbf{e}_t(\varphi_M)] \in \mathbb{C}^{M\times M}$. As the

number of BS antennas M tends to infinity, $\mathbf{V}$ becomes an asymptotically unitary matrix. We call $\tilde{\mathbf{H}}_{k,\ell}^{d}$ the *beam domain channel matrix* with one direction of AoD called a beam. Following Sayeed (2002), we have the following sampling approximation:

$$\left[\tilde{\mathbf{H}}_{k,\ell}^{d}\right]_{nm} \approx \sum_{p\in S_{r,n}\bigcap S_{t,m}} a_{p,k}e^{-j2\pi d_{p,k}/\lambda_c}e^{-j2\pi\ell\tau_{p,k}} \quad (5)$$

where $S_{r,n}$ is the set of all paths whose receive angles are nearest to the angle $\theta_{n,k}$ and $S_{t,m}$ is the set of paths whose transmit angle is exactly equal to the angle φ_m. In the beam domain, different elements of the channel matrix represent the signals of different transmit and receive angles. Different from the summation of all path signals in the physical domain, the beam domain channel can separate the paths of different angles by different beams.

Define the eigenmode channel coupling matrix as (Gao et al. 2009)

$$\boldsymbol{\Omega}_{k,\ell}^{d} = \mathbb{E}\left\{\tilde{\mathbf{H}}_{k,\ell}^{d} \odot (\tilde{\mathbf{H}}_{k,\ell}^{d})^{*}\right\} \quad (6)$$

whose elements specify the mean amount of energy that is coupled from the mth eigenvector of the BS to the nth eigenvector of users. The eigenmode channel coupling matrix $\boldsymbol{\Omega}_{k,\ell}^{d}$ is likewise independent of subcarriers. Then, when the BS has access only to the channel coupling matrix, the optimal transmit strategies are the same for all subcarriers, and the subscript ℓ can be omitted. This observation will significantly decrease the overhead of CSIT acquisition.

Beam Division Multiple Access Transmission

As the channel gains of different paths are independent, the beam domain channel matrix $\tilde{\mathbf{H}}_{k}^{d}$ (whose subscript ℓ is omitted) has uncorrelated elements. The optimal transmission in the beam domain is that BS transmits independent signals through different beams and that different set of beams transmit signals to different users, which is called beam division multiple access (BDMA) transmission (Sun et al. 2015, 2017; You et al. 2017). In the BDMA transmission, the beams for different users are nonoverlapping (orthogonal), and one beam transmits at most one user's signal.

Figure 1 illustrates the BDMA downlink and uplink transmission. In BDMA downlink transmission, as shown in Fig. 1a, BS employs different beam sets transmitting different user's signal, and the received signal at user k is

$$\begin{aligned}\mathbf{y}_{k}^{d} &= [\tilde{\mathbf{H}}_{k}^{d}]^{\mathcal{B}_k}\tilde{\mathbf{x}}_k + \sum_{k'\neq k}[\tilde{\mathbf{H}}_{k}^{d}]^{\mathcal{B}_{k'}}\tilde{\mathbf{x}}_{k'} + \mathbf{n}_k \\ &= [\tilde{\mathbf{H}}_{k}^{d}]^{\mathcal{B}_k}\tilde{\mathbf{x}}_k + \mathbf{n}_{k}^{d}, \end{aligned} \quad (7)$$

where $\mathcal{B}_k$ is the beam set for user k, $\tilde{\mathbf{x}}_k = \mathbf{V}^H\mathbf{x}_k$ is the beam domain transmitted signal, and $\mathbf{n}_k^d$ is the aggregate interference plus noise.

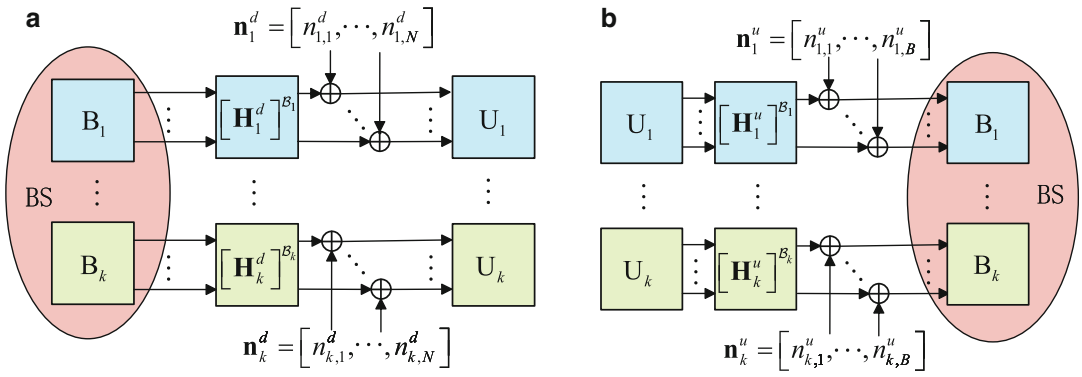
In BDMA uplink transmission, as shown in Fig. 1b, BS employs beam set $\mathcal{B}_k$ to receive signal of user k, and the received signal of user k at the BS is

$$\begin{aligned}\mathbf{y}_{k}^{u} &= [\tilde{\mathbf{H}}_{k}^{u}]^{\mathcal{B}_k}\tilde{\mathbf{x}}_k + \sum_{k'\neq k}[\tilde{\mathbf{H}}_{k'}^{u}]^{\mathcal{B}_k}\tilde{\mathbf{x}}_{k'} + \mathbf{n} \\ &= [\tilde{\mathbf{H}}_{k}^{u}]^{\mathcal{B}_k}\tilde{\mathbf{x}}_k + \mathbf{n}_{k}^{u}, \end{aligned} \quad (8)$$

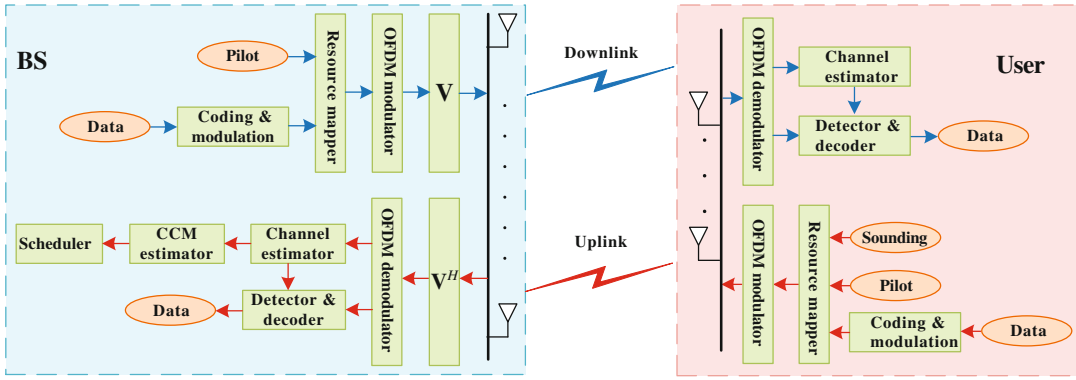
where the superscript u means uplink and $\mathbf{n}_k^u$ is the aggregate interference plus noise of user k in the uplink.

BDMA transmission scheme for MU-MIMO in the beam domain is illustrate in Fig. 2, which consists of the following key components:

1. The BS acquires channel coupling matrices from all users in its own cell. The channel coupling matrices acquisition consists of two steps: first in the uplink, different users transmit sounding signals. Secondly, the BS estimates the instantaneous CSI to calculate channel coupling matrices. Since the statistical CSI varies much slower than the instantaneous CSI, we do not need to estimate the whole channel realizations, and therefore, the



Massive MIMO BDMA Transmission, Fig. 1 (**a**) BDMA downlink transmission and (**b**) BDMA uplink transmission



Massive MIMO BDMA Transmission, Fig. 2 Block diagram of a BDMA transmission scheme

overhead of statistical CSI is much less than that of instantaneous CSI.

2. The BS schedules users to be sufficiently separated in beam domain. Based on the channel coupling matrices, the user scheduler selects users for maximizing the sum-rate with the constraints that different user beams are nonoverlapping. After user scheduling, beams at the BS are divided into different sets, leading to decomposing the massive MU-MIMO link into small dimensional SU-MIMO links.
3. Uplink and downlink transmissions consist of pilot training and data transmission. In the training step, since users are separated by different beam sets at the BS, uplink and downlink pilot sequences do not need to be orthogonal. According to the number of selected users, optimal pilot sequences are designed to minimize the channel estimation error. The receivers in the uplink and downlink estimate the reduced dimensional channel matrices, as well as the aggregate interference covariance matrix for data detection. With channel information, the receivers utilize the iterative soft-input and soft-output (SISO) detection and SISO decoding.

The BDMA scheme is suitable for FDD systems with only statistical CSI at the BS. Note, however, that the approach also applies to TDD systems, when instantaneous CSI at the BS is not available, e.g., for cases involving high user mobility.

Conclusion

BDMA transmission is a multiple access scheme in massive MIMO communication systems. In

the beam domain, the elements of the channel matrix represent the channels between different transmit and receive angles. According to the channel statistics, BS assigns nonoverlapping beams to different users, and users are separated by different beams. Massive MU-MIMO transmission link is decomposed into multiple small dimensional SU-MIMO links, to cope with the difficulties of instantaneous CSIT acquisition and reduce the complexity of transceiver design.

Key Applications

BDMA transmission is an optimal and effective multi-user transmission scheme, when only statistical CSIT is available in both TDD and FDD systems. Employing massive antennas at the BS, BDMA transmission explores the spatial resource and provides high spatial resolution by different beams. BDMA transmission helps to provide high spectral efficiency and reduce the complexity of transceiver design.

Cross-References

- Massive MIMO
- Per-Beam Synchronization for Millimeter-wave Massive MIMO
- Pilot Reuse for Massive MIMO

References

Barriac G, Madhow U (2004) Space-time communication for OFDM with implicit channel feedback 50(12):3111–3129

Choi J, Love DJ, Bidigare P (2014) Downlink training techniques for FDD massive MIMO systems: open-loop and closed-loop training with memory. IEEE J Sel Areas Signal Process 8(5):802–814

Gao XQ, Jiang B, Li X, Gershman AB, McKay MR (2009) Statistical eigenmode transmission over jointly correlated MIMO channels. IEEE Trans Inf Theory 55(8):3735–3750

Jindal N (2006) MIMO broadcast channels with finite-rate feedback. IEEE Trans Inf Theory 52(11):5045–5060

Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans Wirel Commun 9(5):3590–3600

Rusek F, Persson D, Lau BK, Larsson EG, Marzetta TL, Edfors O, Tufvesson F (2013) Scaling up MIMO: opportunities and challenges with very large arrays. IEEE Signal Process Mag 30(1):40–60

Sayeed AM (2002) Deconstructing multi-antenna fading channels. IEEE Trans Signal Process 50(10):2563–2579

Sun C, Gao X, Jin S, Matthaiou M, Ding Z, Xiao C (2015) Beam division multiple access transmission for massive MIMO communications. IEEE Trans Commun 63(6):2170–2184

Sun C, Gao X, Ding Z (2017) BDMA in multicell massive MIMO communications: power allocation algorithms. IEEE Trans Signal Process 65(11):2962–2974

Tse D, Viswanath P (2005) Fundamentals of wireless communication. Cambridge University Press, New York

Wang CX, Haider F, Gao XQ, You XH, Yang Y, Yuan D, Aggoune H, Haas H, Fletcher S, Hepsaydir E (2014) Cellular architecture and key technologies for 5G wireless communication networks. IEEE Commun Mag 52(2):122–130

You L, Gao XQ, Xia X, Ma N, Peng Y (2015) Pilot reuse for massive MIMO transmission over spatially correlated Rayleigh fading channels. IEEE Trans Wirel Commun 14(6):3352–3366

You L, Gao XQ, Li GY, Xia XG, Ma N (2017) BDMA for millimeter-wave/Terahertz massive MIMO transmission with per-beam synchronization. IEEE J Sel Areas Commun 35(7):1550–1563

Massive MIMO Channel Estimation

Jie Yang[1], Shi Jin[1], Chao-Kai Wen[2], and Tao Jiang[3]
[1]National Mobile Communications Research Laboratory, Southeast University, Nanjing, China
[2]National Sun Yat-Sen University, Kaohsiung, Taiwan
[3]Huazhong University of Science and Technology, Wuhan, China

Synonyms

Hyper MIMO channel estimation; Large-scale multiple-antennas channel estimation

Definitions

Massive MIMO channel estimation is a procedure to obtain channel state information (CSI) under systems consisting of large-scale multiple antennas at both transmitter and receiver before detecting the transmitted data. CSI refers to the transmission matrix between the transmitter and receiver, which includes scattering, fading, and power decay effects of a communication link. Massive MIMO channel estimation performance determines the suitability of massive MIMO techniques, offers the possibility of improving link reliability and network capacity, and reduces latency and interference.

Historical Background

Massive MIMO technology has been widely studied and incorporated into many standards, such as long-term evolution (LTE) and IEEE802.11 (Wi-Fi), because it can significantly improve the capacity and reliability of wireless systems Larsson et al. (2014).

Over the past two decades, the speed of wireless networks has improved, starting from SISO in 2000, to SU-MIMO in 2009, and to MU-MIMO in 2012. In the near future, around 2020, massive MIMO aims to achieve a data rate of more than 10 Gbps Andrews et al. (2014). Massive MIMO is a bold vision of development from classical MIMO by using very large-scale antennas at both ends, which are fully coherent and adaptive. Wireless channel diversity is applied to provide links with higher speed and reliability.

The larger the number of antennas placed at both transmitter and receiver, the more terminals can be served by the increasing data streams. This condition is expected to bring significant improvements in throughput. Furthermore, transmitters can easily avoid transmission into undesired directions, thereby alleviating harmful interference. This condition can enhance the interference cancelation and system reliability through spatial diversity Björnson et al. (2016).

The aforementioned advantages and improvements in the massive MIMO technology are based on perfect knowledge of the channel matrix, which has to be estimated. However, the massive MIMO channel estimation faces serious challenges. First, the wireless channel for massive MIMO is highly complicated, being characterized by selectivities in frequency, time, and space Lu et al. (2014). Second, the channel information is acquired by probing finite-length pilot sequences into the wireless channel because with the existence of intercell interference, the reused pilot sequences from neighboring cells would contaminate one another. According to Marzetta T.L. (2010), pilot contamination does not vanish with an unlimited number of antennas. Moreover, the training and feedback overhead become overwhelming with the increase in the number of antennas. These problems emphasize the need for alternative channel estimation schemes for increasingly large transceiver arrays.

The development pathway of the massive MIMO channel estimation should highlight the following popular techniques: To deal with the pilot contamination, Yin et al. (2013) present a novel approach that enables low-rate coordination between cells during the channel estimation phase. This approach can offer a powerful method of discriminating across mutual interfering users with strongly correlated pilot sequences. Furthermore, Yang et al. (2014) propose a new two-way training scheme for discriminatory channel estimation (DCE) by exploiting a whitening-rotation-based semi-blind method, which achieves better DCE performance than the existing DCE schemes based on linear minimum mean square error (LMMSE) channel estimator; this proposed method is robust against pilot contamination attacks. In particular, Wen et al. (2015a) introduce the estimation not only of the channel parameters of the desired links in a target cell but also those of the interference links from adjacent cells. This channel estimation method is based on sparse Bayesian learning methods and utilizes the approximately sparse property of the beamspace channel. In addition, a subspace method for channel estimation, which considers a highly sensible finite-dimensional physical

channel model, has been proposed to address the pilot contamination effect Teeti et al. (2015).

Extensive research has been conducted to exploit compressive sensing (CS) techniques to reduce the training overhead and feedback cost, especially for FDD massive MIMO systems. In Rao et al. (2014), a distributed compressive channel state information at transmitter (CSIT) estimation scheme is presented by exploiting the hidden joint sparsity structure in the user channel matrices. Additionally, a discrete Fourier transform (DFT)-aided spatial basis expansion model is introduced by Xie et al. (2016) to represent the channel, which can be applied to TDD and FDD systems under certain conditions. However, the DFT-based methods are only applicable to uniform linear arrays and have power leakage problems. Subsequently, Dai et al. (2018) design an off-grid model for downlink channel sparse representation with arbitrary 2D-array antenna geometry and apply an efficient sparse Bayesian learning (SBL) approach for sparse channel recovery and off-grid refinement. Also, CS-based massive MIMO channel estimation, such as structured compressive sensing-based spatiotemporal joint channel estimation scheme Gao et al. (2016) and beam-blocked compressive channel estimation scheme Huang et al. (2017), has been effective in reducing the required pilot overhead.

Computational complexity is one of the main challenges for massive MIMO systems. To reduce the computational complexity of the channel estimation, a set of low-complexity Bayesian channel estimators, called polynomial expansion channel estimators, are demonstrated by Shariati et al. (2014). To further reduce the computational complexity and improve the accuracy of channel estimation, Wen et al. (2015b) adopt a joint channel-and-data estimation method based on Bayes optimal inference.

Furthermore, other concepts, such as beam-domain channel estimation Duly et al. (2014); Xiong et al. (2017), eigenvalue-decomposition-based channel estimation Xu et al. (2017), and blind channel estimation Ngo et al. (2015), have also received considerable attention.

Foundations

Massive MIMO channel estimation methods are usually divided into two categories Hassan et al. (2017):

- *Training-based channel estimation,* which transmits known training symbols (also called pilots) at predetermined times and frequencies known by the receiver. Then, the receiver uses the known information to obtain the CSI (including the gain and phase rotation imparted by the channel at each point in time and frequency) based on the characteristics of the received training symbols.
- *Blind-based channel estimation,* which estimates the channel without the assistance of known training symbols. This method can save timefrequency resources and offer higher bandwidth efficiency. As the blind-based channel estimation method has lower speed and poorer performance than the training-based method, researchers pay closer attention to the latter. Thus, the later part of this section considers only the training-based channel estimation methods.

Two common modes for massive MIMO systems exist, namely, time division duplexing (TDD) and frequency division duplexing (FDD) Lu et al. (2014). In FDD-mode massive MIMO systems, the uplink and downlink use different frequency bands. Therefore, the uplink and downlink CSI is different. The uplink channel estimation is conducted at the base station (BS) by enabling all users to send different pilot sequences. The downlink channel estimation is followed by two steps: the BS sends pilot sequences to all users, and then the users feed the estimated downlink CSI back to the BS. As the number of antennas increases, the consumption of time resource required to transmit pilot sequences also increases in the FDD mode, whereas the TDD mode is more efficient in this aspect. In TDD systems, based on the assumption of channel reciprocity, only the uplink channel estimation is needed. According to this protocol,

the BS estimates CSI by using the pilot sequences sent by all users to estimate CSI, and then the BS uses the estimated CSI to detect the uplink data and generate beamforming vectors for downlink data transmission. The time required to transmit pilot sequences is less in the TDD mode than in the FDD mode; thus, the TDD mode is usually assumed in the study.

Pilot design is one of the key steps in massive MIMO channel estimation. Pilot placement has to comply with the following principles: the time interval between pilot symbols should be less than the coherence time of the channel and the frequency spacing should be less than the coherence bandwidth Hampton (2013). These conditions mean that in a fast-fading environment, pilots have to be placed relatively often in time, and in a highly frequency-selective channel, pilots have to be placed close together in the frequency dimension. Furthermore, based on meeting the aforementioned principles, pilots should be spaced as far apart as possible to reduce the training overhead. In massive MIMO systems, the received signal at each antenna is the superposition of the signals from all of the transmission antennas. Thus, the pilot sequences have to maintain temporal, frequency, or signal orthogonality to avoid interfering with one another. A common pilot placement strategy is packet-based in which synchronization and pilot sequences are placed in the header of the packet and then the body of the packet contains the data. Based on the hypothesis that the channel only changes slightly in the duration of a packet, the receiver can estimate the channel at the beginning of a packet and then use that estimation during the body of the packet to detect data.

Massive MIMO channel estimation techniques can be elaborated in terms of sub-6 GHz and millimeter wave (mmWave). The foundational channel estimation techniques for sub-6 GHz are introduced as follows:

- *Narrowband MIMO channel estimation:* For the quasi-static fading channel, where the packet length is shorter than the coherence time, the channel is assumed to be changeless over one packet. Therefore, the transmitter can transmit an N_t-symbol packet, including N_p pilots, every T_coh seconds. Then, the receiver can estimate the channel through pilots and use the estimated channel to detect the data in the same packet. The received data are given by

$$\mathbf{Y} = \mathbf{HX} + \mathbf{N}, \tag{1}$$

where $\mathbf{H}$ is the $N_r \times N_t$ channel matrix, $\mathbf{X}$ is the $N_t \times p$ pilot matrix, and $\mathbf{N}$ is the $N_r \times p$ noise matrix. The commonly used estimation techniques are the following:

– **Maximum likelihood (ML)** channel estimation aims to maximize the likelihood function of $\mathbf{H}$, namely, $\hat{\mathbf{H}}_{ML} \triangleq \arg\max_{\{\mathbf{H}\}} p(\mathbf{Y}|\mathbf{H})$. Then, after simplification, the solution of $\hat{\mathbf{H}}_{ML}$ is given by $\hat{\mathbf{H}}_{ML} = \left(\mathbf{YX}^H\right)\left(\mathbf{XX}^H\right)^{-1}$.

– **Least-square (LS)** channel estimation minimizes the squared error between the actual received signal $\mathbf{Y}$ and the estimated received signal $\hat{\mathbf{Y}} = \hat{\mathbf{H}}\mathbf{X}$, that is, $\hat{\mathbf{H}}_{LS} \triangleq \arg\min_{\{\hat{\mathbf{H}}\}} \left\|\mathbf{Y} - \hat{\mathbf{Y}}\right\|_F^2$. Then, after simplification, the solution of $\hat{\mathbf{H}}_{LS}$ is expressed by $\hat{\mathbf{H}}_{LS} = \left(\mathbf{YX}^H\right)\left(\mathbf{XX}^H\right)^{-1}$, which is equivalent to the ML channel estimation.

– **Linear minimum mean square error (LMMSE)** channel estimation minimizes the mean square error between the actual and estimated channels; this error is expressed as $\hat{\mathbf{H}}_{LMMSE} \triangleq \arg\min_{\{\hat{\mathbf{H}}\}} \mathsf{E}\left\{\left\|\mathbf{H} - \hat{\mathbf{H}}\right\|_F^2\right\}$. After simplification, the solution of $\hat{\mathbf{H}}_{LMMSE}$ is given by $\hat{\mathbf{H}}_{LMMSE} = \mathbf{Y}\left(\mathbf{I} + \mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{X}^H$.

– **Maximum a posteriori (MAP)** channel estimation is designed to maximize the posterior probability distribution function, that is, $\hat{\mathbf{H}}_{MAP} \triangleq \arg\max_{\{\mathbf{H}\}} p(\mathbf{H}|\mathbf{Y})$.

The ML and LS channel estimations are unbiased estimation techniques as long as the

noise has a zero-mean symmetrical distribution. The LMMSE channel estimation is not unbiased, whereas with the increase in signal-to-noise ratio, it becomes unbiased and is equivalent to the ML/LS estimation.

- *Broadband MIMO channel estimation:* When the bandwidth of the signal is greater than the channel coherence bandwidth, the frequency-selective fading channel occurs. The orthogonal frequency division multiplexing (OFDM) technique is commonly used in broadband massive MIMO applications. In OFDM systems, the pilot is placed in both time and frequency domains, and the frequency domain channel is usually estimated by ML/LS/LMMSE channel estimation techniques.
- *CS-based mmWave massive MIMO channel estimation:* CS techniques bring numerous benefits to mmWave massive MIMO beamspace channel estimation, by using CS channel estimation algorithms. The system operation requires a relatively small training overhead. To adapt to various channel estimation scenarios, several CS algorithms have been introduced, such as LASSO Tibshirani (1996), orthogonal matching pursuit Cai et al. (2011), support detection Gao et al. (2017), and sparse non-informative parameter estimator-based cosparse analysis approximate message passing for imaging Yang et al. (2018) channel estimation algorithms.

Finally, the bit error ratio and normalized mean square error are commonly used to study the performance of massive MIMO channel estimation techniques.

Key Applications

Massive MIMO channel estimation is involved in most wireless communication systems, especially in 5G mobile communications.

Cross-References

▸ Massive MIMO
▸ Millimeter wave massive MIMO

References

Andrews JG, Buzzi S, Choi W, Hanly SV, Lozano A, Soong AC, Zhang JC (2014) What will 5G be? IEEE J Sel Areas Commun 32(6):1065–1082

Björnson E, Larsson EG, Marzetta TL (2016) Massive MIMO: ten myths and one critical question. IEEE Commun Mag 54(2):114–123

Cai TT, Wang L (2011) Orthogonal matching pursuit for sparse signal recovery with noise. IEEE Trans Inf Theory 57(7):4680–4688

Dai J, Liu A, Lau VKN (2018) FDD massive MIMO channel estimation with arbitrary 2D-array geometry. IEEE Trans Signal Process 99(99):1–1

Duly AJ, Kim T, Love DJ, Krogmeier JV (2014) Closed-Loop beam alignment for massive MIMO channel estimation. IEEE Commun Lett 18(8):1439–1442

Fang J, Li X, Li H, Gao F (2017) Low-rank covariance-assisted downlink training and channel estimation for FDD massive MIMO systems. IEEE Trans Wirel Commun 16(3):1935–1947

Gao Z, Dai L, Dai W, Shim B, Wang Z (2016) Structured compressive sensing-based spatio-temporal joint channel estimation for FDD massive MIMO. IEEE Trans Commun 64(2):601–617

Gao X, Dai L, Han S, I C-L, Wang X (2017) Reliable beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array. IEEE Trans Wirel Commun 16(9):6010–6021

Hampton JR (2013) Introduction to MIMO communications. Cambridge University Press, Cambridge

Hassan N, Fernando X (2017) Massive MIMO wireless networks: an overview. Electronics 6:63

Huang W, Huang Y, Xu W, Yang L (2017) Beam-blocked channel estimation for FDD massive MIMO with compressed feedback. IEEE Access 5:11791–11804

Larsson E, Edfors O, Tufvesson F, Marzetta T (2014) Massive MIMO for next generation wireless Systems. IEEE Commun Mag 52(2):186–195

Lu L, Li GY, Swindlehurst AL, Ashikhmin A, Zhang R (2014) An overview of massive MIMO: benefits and challenges. IEEE J Sel Top Signal Process 8(5):742–758

Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbersof base station antennas. IEEE Trans Wirel Commun 9(11):3590–3600

Ngo HQ, Larsson EG (2015) Blind estimation of effective downlink channel gains in massive MIMO. In: Proc ICASSP 2015, Brisbane, pp 2919–2923
Rao X, Lau VK (2014) Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO Systems. IEEE Trans Signal Process 62(12):3261–3271
Shariati N, Bjornson E, Bengtsson M, Debbah M (2014) Low-complexity polynomial channel estimation in large-scale MIMO with arbitrary statistics. IEEE J Sel Top Signal Process 8(5):815–830
Teeti M, Sun J, Gesbert D, Liu Y (2015) The impact of physical channel on performance of subspace-based channel estimation in massive MIMO systems. IEEE Trans Wirel Commun 14(9):4743–4756
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol) 58:267–288
Wen CK, Jin S, Wong KK, Chen JC, Ting P (2015a) Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning. IEEE Trans Wirel Commun 14(3):1356–1368
Wen CK, Wang CJ, Jin S, Wong KK, Ting P (2015b) Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs. IEEE Trans Signal Process 64(10):2541–2556
Xie H, Gao F, Zhang S, Jin S (2016) A simple DFT-aided spatial basis expansion model and channel estimation strategy for massive MIMO systems. In: Proceedings of GLOBECOM, 2016, Washington, DC, pp 1–6
Xiong X, Wang X, Gao, X, You X (2017) Beam-domain channel estimation for FDD massive MIMO systems with optimal thresholds. IEEE Trans Wirel Commun 16(7):4669–4682
Xu W, Xiang W, Jia Y, Li Y, Yang Y (2017) Downlink performance of massive-MIMO systems using EVD-based channel estimation. IEEE Trans Veh Technol 66(4):3045–3058
Yang J, Xie S, Zhou X, Yu R, Zhang Y (2014) A semiblind two-way training method for discriminatory channel estimation in MIMO systems. IEEE Trans Commun 62(7):2400–2410
Yang J, Wen CK, Jin S, and Gao F (2018) Beamspace channel estimation in mmwave systems via corsparse image reconstruction technique. IEEE Trans Commun, to be published, DOI 10.1109/TCOMM.2018.2805359
Yin H, Gesbert D, Filippou M, Liu YA (2013) Coordinated approach to channel estimation in large-scale multiple-antenna systems. IEEE J Sel Areas Com 31(2): 264–273

Massive Multiuser MIMO

▸ Massive MIMO

Matching for Cooperative Spectrum Sharing

Lin Gao[1], Lingjie Duan[2], Shimin Gong[3], and Qinyu Zhang[1]
[1]School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, China
[2]Engineering Systems and Design Pillar, Singapore University of Technology and Design, Singapore, Singapore
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Synonyms

Cooperative spectrum sharing; Dynamic spectrum access; Two-sided matching

Definition

Dynamic spectrum access, also known as opportunistic spectrum access, is a new spectrum access paradigm, which allows unlicensed secondary users (SUs) to access the licensed spectrum of primary users (PUs) in an opportunistic way, and hence can effectively increase the spectrum efficiency and alleviate the spectrum scarcity. *Cooperative spectrum sharing* is an effective form of dynamic spectrum access, where SUs relay the transmissions of PUs in exchange for the opportunity of accessing the licensed spectrum of PUs, and thus, it can offer the necessary incentives for both PUs and SUs in dynamic spectrum access. *Two-sided matching*, or matching, is an efficient framework for analyzing the inter-dependent interactions between two disjoint player sets such as PUs and SUs in cooperative spectrum sharing.

Historical Background

Nowadays, wireless spectrum is becoming increasingly congested and scarce with the

explosive development of wireless devices and services. Dynamic spectrum access (DSA) is a promising approach to increase the spectrum efficiency and alleviate the spectrum scarcity, by allowing unlicensed secondary users (SUs) to opportunistically access the spectrum licensed to primary users (PUs) (Zhao and Sadler 2007; Akyildiz et al. 2008). Hence, it has become one of the important information and communication technologies (ICT) for future wireless networks. To successfully implement DSA in practice, it is important to offer necessary incentives for PUs to open their spectrum for SUs' utilizations and for SUs to access the PUs' spectrum despite the potential costs (Niyato and Hossain 2008; Gao et al. 2011a, b, 2013; Huang et al. 2006; Huang 2013; Luo et al. 2015).

Cooperative spectrum sharing (CSS) has been recently proposed as an effective approach to offer necessary incentives for both PUs and SUs in DSA (Simeone et al. 2008; Zhang and Zhang 2009; Wang et al. 2010; Duan et al. 2011, 2014). The key idea is to enable SUs to act as relay and help the transmission of PUs in exchange for the opportunities of accessing the PUs' licensed spectrum. In such a way, PUs can increase the transmission efficiency with the help of SUs (and hence can save some spectrum resource for SUs), and SUs can obtain the desired transmission opportunity on the PUs' licensed spectrum (at the cost of consuming energy resource to help PUs). Thus, it will lead to a win-win situation for PUs and SUs.

Existing work on CSS mainly focused on the interactions between *one* PU and one or multiple SUs (Simeone et al. 2008; Zhang and Zhang 2009; Wang et al. 2010; Duan et al. 2011, 2014). The PU, as the monopolist, can dominate the cooperation process. Some works assumed that the PU has complete information of SUs and formulated the problem into Stackelberg games (Simeone et al. 2008; Zhang and Zhang 2009; Wang et al. 2010), while others considered that the PU has limited information about SUs and introduced auction theory and contract theory to elicit the private information of SUs (Duan et al. 2011, 2014). However, these models and approaches are difficult to be extended to the scenario of multiple PUs and multiple SUs, which is more general and practical.

This entry focuses on discussing the general CSS model with multiple PUs and SUs. Although the multi-PU multi-SU scenario is more general and practical, it is also much more challenging to analyze the interactions between multiple PUs and multiple SUs, as not only SUs compete with each other for the PUs' licensed spectrums, but PUs also compete with each other for the SUs' collaborations.

Foundations

In a multi-PU multi-SU model, the first key challenging problem is to decide *which PU cooperates with which SU*. Note that multiple SUs may be interested in cooperating with the same PU, and so are PUs. Thus, competition inherently exists not only among PUs but also among SUs. This is a typical *two-sided matching* problem (Gale and Shapley 1962). The second challenging problem is to decide *how each pair of matched PU and SU cooperate with each other*. That is, how much spectrum resource that the PU would like to share with the SU, and how much effort (e.g., energy) that the SU would like to put to help the PU. In this sense, the PU contributes spectrum resource in exchange for the SU's energy resource, or equivalently, the SU contributes energy resource in exchange for the PU's spectrum resource. Thus, this can also be viewed as *resource exchange* between them.

Combining the above two problems, the interaction between multiple PUs and SUs can be formulated as an extended two-sided matching problem with varying resource exchange. In this section, the general CSS model, the extended two-sided matching formulation, and the conditions for stable matching outcome (equilibrium) will be discussed in detail.

General CSS Model

A general CSS model consists of multiple PUs and multiple SUs. Let $\mathscr{M} \triangleq \{1, \ldots, M\}$ denote the set of PUs and $\mathscr{N} \triangleq \{1, \ldots, N\}$ denote the set of SUs. Each PU owns a dedicated licensed

spectrum band, and the spectrum bands of different Pus are *nonoverlapping* (i.e., orthogonal). Thus, there is no interference on different PUs' spectrum bands. Each SU is equipped with one radio frequency module and hence can access one PU's spectrum band at a particular time.

By cooperative spectrum sharing, SUs act as relays to help the transmissions of PUs in exchange for accessing the PUs' spectrum bands. Assume that each PU can employ *only one* SU as cooperative relay and each SU can help *only one* PU at a particular time. Note that in practice, a PU can possibly employ multiple SUs as cooperative relays. However, existing study Yi et al. (2010) has shown that choosing one relay (i.e., the most appropriate one) is usually sufficient to achieve all or almost all cooperative gain. Similarly, in practice, an SU can possibly help multiple PUs by using a proper multiple access technique. From the economic perspective, however, there is no incentive for an SU to help multiple PUs, as he can only access one PU's frequency band at a particular time.

With the above assumption, the assignment between PUs and SUs is essentially a *one-to-one matching* between them, where one PU (or SU) can only be matched with one SU (or PU). Let μ_n denote the PU matched with SU n and μ_m denote the SU matched with PU m. Note that $\mu_n = \emptyset$ or $\mu_m = \emptyset$ denotes that SU n or PU m is not matched with any PU or SU. Clearly, $\mu_n = m$ if and only if $\mu_m = n$.

For a pair of matched PU and SU, their cooperation frame is divided into 2 phases: in Phase I, the PU transmits his own data with the cooperative relay of the SU; in Phase II, the SU transmits his data on the PU's frequency band (and the PU stops his own data transmission). The cooperative relay in Phase I can use any cooperative protocol such as amplified-and-forward (AF) and decode-and-forward (DF). Obviously, the cooperative gain for the PU greatly depends on the relay effort of the SU, e.g., the SU's transmission power for relaying. Moreover, the cooperative gain for the SU depends on the length of the spectrum access time (Phase II) shared by the PU. As discussed before, such a cooperation process can be viewed as *resource exchange* between them, where the SU contributes his energy resource (in Phase I) in exchange for using the PU's spectrum resource (in Phase II).

Clearly, a larger relay effort of the SU can bring more benefit for the PU but will also incur higher cost on the SU himself, as he needs to consume more resource for helping the PU. Similarly, a larger spectrum access time shared by the PU in Phase II can bring more benefit for the SU but will reduce the PU's benefit due to the reduced time for his own transmission in Phase I. Thus, for a pair of matched PU and SU, increasing the benefit for one user will accordingly reduce the benefit for the other user. This process can be viewed as the *division of cooperative gain* between them. Let $f_n^m(x)$ denote the *gain division function* (GDF) between PU m and SU n, which characterizes the maximum gain that a PU m can achieve when cooperating with SU n and meanwhile leaving a gain of x for SU n. Let $g_n^m(y)$ denote the *inverse gain division function* (IGDF) between PU m and SU n, which characterizes the maximum gain that an SU n can achieve when cooperating with PU m and meanwhile leaving a gain of y for PU m. Obviously, IGDF is the inverse function of GDF, i.e., $y = f_n^m(x)$ if and only if $x = g_n^m(y)$.

Note that the GDF or IGDF in this entry is quite general and can be arbitrary decreasing function. An explicit example of GDF is provided in Gao et al. (2017), which is defined as the maximum data rate that PU m can achieve by jointly optimizing the SU's relay power in Phase I and spectrum access time in Phase II.

Matching Formulation

Based on the above discussion, the interactions between PUs and SUs can be formulated as an extended *two-sided matching problem*, with PUs on one side and SUs on the other side. A PU is matched with an SU (or equivalently, an SU is matched with a PU) means that the PU cooperates with the SU under certain cooperative gain division (or resource exchange) agreement between them. Accordingly, a *matching outcome* defines not only the cooperative relationships of all PUs and SUs (i.e., who cooperates with whom), but also the cooperative gain division

(or resource exchange) between each pair of matched PU and SU (i.e., how they cooperate with each other). Formally, such a matching outcome can be expressed as

$$\{(\mu_n, \delta_n), \forall n \in \mathcal{N}\}, \tag{1}$$

where $(\mu n, \delta n)$ denotes the matching result for each SU $n \in \mathcal{N}$, with μ_n denoting his matched PU and δ_n denoting his achieved cooperative gain. Obviously, in such a matching outcome, the (maximum) cooperative gain that PU μ_n can achieve is $f_n^{\mu_n}(\delta_n)$. Note that if SU n is not matched with any PU, i.e., $\mu_n = \emptyset$, he *cannot* achieve any cooperative gain, i.e., $\delta_n = 0$.

Note that a matching outcome may *not* be stable, as some users may discard their matched partners or change the gain division (resource exchange) agreement with their partners. For example, a matching outcome is *not* stable, if there exist a PU and an SU who are *not* matched with each other, but both prefer to be (i.e., both can increase their gains by matching with each other). In such a case, they would match with each other by discarding their respective partners. On the other hand, a matching outcome is also *not* stable, if there exist a PU and an SU who are matched with each other, but one or both prefer not to be (i.e., one can increase his gain by not matching with the other). In such a case, the user would discard his partner and match with another user or remain single.

Therefore, a natural question arising in such an extended two-sided matching problem is *what are the stable matching outcomes, from which none of PUs and SUs has the incentive to deviate*. A stable matching outcome is often referred to as an *equilibrium* of the matching. Formally,

Definition 1 (Equilibrium) An equilibrium of a matching is a stable matching outcome, where no user can improve his benefit via the unilateral deviation (i.e., choosing another partner or changing the gain division in agreement with his partner).

In the next section, the necessary and sufficient conditions for stable matching outcome (equilibrium) will be derived systematically.

Stable Matching Outcome (Equilibrium)
Suppose $\{(\mu n, \delta n), \forall n \in \mathcal{N}\}$ is an equilibrium. Then, for each pair of matched SU and PU (say, SU n and PU μn), the following conditions must hold: (i) both SU n and PU μ_n have no incentive to break the current matching, by pairing with another user or remaining single, and (ii) all PUs other than μ_n (or all SUs other than n) have no incentives to discard their respective partners and pair with SU n (or PU μ_n). This leads to the following necessary conditions for equilibrium.

Lemma 1 (Necessary Conditions)
If $\{(\mu n, \delta n), \forall n \in \mathcal{N}\}$ *is an equilibrium, then for each PU* μ_n *(matched with SU n), the following condition holds:*

$$\begin{aligned} &\text{(IR)} \quad f_n^{\mu_n}(\delta_n) \geq 0, \\ &\text{(IC)} \quad f_n^{\mu_n}(\delta_n) \geq f_k^{\mu_n}(\delta_k), \quad \forall k \neq n, \\ &\text{(CC)} \quad f_{\mu_m}^{m}\left(\delta_{\mu_m}\right) \geq f_n^{m}(\delta_n), \ \forall m \neq \mu_n. \end{aligned} \tag{2}$$

(*where* $f_n^m(x)$ *is the GDF, denoting the (maximum) gain that PU m can achieve when matching with SU n and leaving a gain of x for SU n.*

The first condition is called the *Individual Rationality* (IR) condition, the second condition is called the *Incentive Compatibility* (IC) condition, and the third condition is called the *Competitive Compatibility* (CC) condition. The IC and IR conditions ensure that each PU μ_n (matched with SU n) has no incentive to discard his current partner SU n by remaining single or matching with another SU. The CC condition ensures that all PUs other than μ_n have no incentive to discard their current partners and compete with PU μ_n for SU n.

The IR, IC, and CC conditions lead to the following constraints on δ_n:

$$\begin{aligned} &\text{(IR)} \quad \Rightarrow \delta_n \leq g_n^{\mu_n}(0), \\ &\text{(IC)} \quad \Rightarrow \delta_n \leq \min_{k \neq n} \left\{ g_n^{\mu_n}\left(f_k^{\mu_n}(\delta_k)\right) \right\}, \\ &\text{(CC)} \Rightarrow \delta_n \geq \max_{m \neq \mu_n} \left\{ g_n^{m}\left(f_{\mu_m}^{m}\left(\delta_{\mu_m}\right)\right) \right\}, \end{aligned} \tag{3}$$

where $g_n^m(x) = f_n^{\mu_n(-1)}(x)$ denotes maximum gain that SU n can achieve when matching with PU m and leaving a gain of x for PU m. The above

constraints on δ_n further lead to the following upper-bound and lower-bound for each δ_n:

$$\overline{\delta}_n \triangleq \min_{k \neq n} \left\{ g_n^{\mu_n} \left(f_k^{\mu_n} (\delta_k) \right), g_n^{\mu_n}(0) \right\}. \quad (4)$$

$$\underline{\delta}_n \triangleq \max_{m \neq \mu_n} \left\{ g_n^{m} \left(f_{\mu_m}^{m} \left(\delta_{\mu_m} \right) \right), 0 \right\}, \quad (5)$$

Intuitively, the gain for SU n, i.e., δn, cannot be higher than the threshold $\overline{\delta}_n$, otherwise PU μ_n (matched with SU n) would be better off by matching with another SU or remaining single. On the other hand, δ_n cannot be lower than the threshold $\underline{\delta}_n$, otherwise some other PUs $m \neq \mu_n$ (matched with other SUs) will have the incentive and ability to compete for SU n (by leaving a slightly higher gain for SU n).

With the above upper-bound and lower-bound for each δn, the necessary conditions in Lemma 1 can be rewritten as $\underline{\delta}_n \leq \delta_n \leq \overline{\delta}_n, \forall n \in \mathcal{N}$.. It is important to note that these conditions are not only necessary, but also sufficient. Formally,

Lemma 2 (Necessary and Sufficient Conditions)

A matching outcome $\{(\mu_n, \delta_n), \forall n \in \mathcal{N}\}$ *is an equilibrium, if and only if:*

$$\underline{\delta}_n \leq \delta_n \leq \overline{\delta}_n, \quad \forall n \in \mathcal{N}, \quad (6)$$

where $\overline{\delta}_n$ *and* $\underline{\delta}_n$ *are defined in* Eqs. (4) and (5)*, respectively.*

Key Applications

Two-sided matching is a widely used efficient framework for analyzing the interactions between two disjoint player sets in two-sided market scenarios. In such scenarios, the matching theory can systematically capture not only the cooperative interactions between users in different sides but also the competitive interactions between users on the same side. Most early results in this area mainly focused on the matching under complete information, without considering the incentive issues under incomplete information.

The first two-sided matching model was studied in Gale and Shapley (1962), where a basic two-sided matching model with nontransferable utilities was proposed to study the marriage market. Shapley and Shubik (1971) and Thompson (1980) studied the more general models with additive and transferable utilities. Kelso and Crawford (1982) studied the two-sided labor models with nontransferable utilities. Some later works such as Kojima and Pathak (2009) studied the incentive issue in matching under incomplete information. Gao et al. (2017) applied the matching theory to analyze the dynamic spectrum access problem between PUs and SUs in cognitive radio networks.

Open Problems

In this section, some open problems for the applications of matching to communication networks are outlined.

- *One-to-Many or Many-to-Many Matching*: This entry mainly focused on the one-to-one matching model, where a PU can only employ one SU as cooperative relay and an SU can only help one PU. In practice, however, a PU can potentially employ multiple SUs as relays and an SU can potentially help multiple PUs. This will lead to a more complicated one-to-many or many-to-many matching.
- *Implementation of Equilibrium*: This entry mainly focused on characterizing the conditions of equilibrium, without considering the implementation of equilibrium under different information scenarios, especially the incomplete information scenario. That is, what equilibrium will emerge under different information scenarios is a challenging open problem.

Cross-References

- ▸ Efficiency and Pareto Optimality
- ▸ Matching Game for Load Distribution in Multi-tier Cellular Networks
- ▸ Preferences and Utility Functions

References

Akyildiz I, Lee W, Vuran M, Mohanty S (2008) A survey on spectrum management in cognitive radio networks. IEEE Commun Mag 46:40–48

Duan L, Gao L, Huang J (2011) Contract-based cooperative spectrum sharing. In: IEEE symposium on new frontiers in dynamic spectrum access networks (DySPAN), pp 399–407

Duan L, Gao L, Huang J (2014) Cooperative spectrum sharing: a contract-based approach. IEEE Trans Mob Comput 13:174–187

Gale D, Shapley L (1962) College admissions and the stability of marriage. Am Math Mon 69:9–15

Gao L, Wang X, Xu Y, Zhang Q (2011a) Spectrum trading in cognitive radio networks: a contract-theoretic modeling approach. IEEE J Sel Areas Commun 29: 843–855

Gao L, Xu Y, Wang X (2011b) MAP: multi-auctioneer progressive auction for dynamic spectrum access. IEEE Trans Mob Comput 10:1144–1161

Gao L, Huang J, Chen Y, Shou B (2013) An integrated contract and auction design for secondary spectrum trading. IEEE J Sel Areas Commun 31: 581–592

Gao L, Duan L, Huang J (2017) Two-sided matching based cooperative spectrum sharing. IEEE Trans Mobile Comput 16:538–551

Huang J (2013) Market mechanisms for cooperative spectrum trading with incomplete network information. IEEE Commun Mag 51:201–207

Huang J, Berry R, Honig M (2006) Auction-based spectrum sharing. Mobile Netw Appl 11:405–418

Kelso JA, Crawford V (1982) Job matching, coalition formation, and gross substitutes. Econometrica 50:1483–1504

Kojima F, Pathak P (2009) Incentives and stability in large two-sided matching markets. Am Econ Rev 34:383–387

Luo Y, Gao L, Huang J (2015) Business modeling for TV white space networks. IEEE Commun Mag 53:82–88

Niyato D, Hossain E (2008) Spectrum trading in cognitive radio networks: a market- equilibrium-based approach. IEEE Wirel Commun 15:71–80

Shapley L, Shubik M (1971) The assignment game I: the core. Int J Game Theory 1:111–130

Simeone O, Stanojev I, Savazzi S, Bar-Ness Y, Spagnolini U, Pickholtz R (2008) Spectrum leasing to cooperating secondary ad hoc networks. IEEE J Sel Areas Commun 26:203–213

Thompson GL (1980) Computing the core of a market game. In: Fiacco A V, Kortanek K O (eds) Extremal Methods and Systems Analysis. Lecture Notes in Economics and Mathematical Systems. Springer, Berlin, 174:312–334

Wang H, Gao L, Gan X, Wang X, Hossain E (2010) Cooperative spectrum sharing in cognitive radio networks: a game-theoretic approach. In: IEEE international conference on communications (ICC), pp 1–5

Yi Y, Zhang J, Zhang Q, Jiang T, Zhang J (2010) Cooperative communication-aware spectrum leasing in cognitive radio networks. In: IEEE symposium on new frontiers in dynamic spectrum access networks (DySPAN), pp 1–11

Zhang J, Zhang Q (2009) Stackelberg game for utility-based cooperative cognitiveradio networks. In: ACM international symposium on mobile ad hoc networking and computing (MobiHoc), pp 3412–3415

Zhao Q, Sadler M (2007) A survey of dynamic spectrum access. IEEE Signal Process Mag 24:79–89

Matching Game for Load Balancing in Wireless Cellular Networks

▸ Matching Game for Load Distribution in Multi-tier Cellular Networks

Matching Game for Load Distribution in Multi-tier Cellular Networks

Nima Namvar[1] and Behrouz Maham[2]
[1]Department of Electrical and Computer Engineering, North Carolina A&T State University, Greensboro, NC, USA
[2]Electrical and Electronic Engineering Department, Nazarbayev University, Astana, Kazakhstan

Synonyms

Matching game for load balancing in wireless cellular networks

Definitions

Load distribution in wireless cellular networks refers to assigning the mobile users (MUs) to the base stations (BSs) throughout the network. Multi-tier cellular networks (MTCNs) are a kind of wireless cellular networks where the access points come in various size, capacity,

and transmit power. Load distribution in such networks should take into account the disparity among the access points so as to optimize the network performance.

Introduction

The concept of multi-tier cellular network (MTCN) is seen as a promising technology to cope with the unprecedented proliferation in wireless data traffic by increasing the network capacity (Damnjanovic et al. 2011). In MTCN, traditional macrocells coexist with other smaller cells, such as microcells and femtocells, which are the catalyst behind improving wireless networks in densely populated areas. In comparison to their macrocell counterparts, small cell transmissions range between a mere 0.25 and 6 W and extend coverage up to several hundred meters. Hence, by reducing the distance between the transmitter and the receiver, small cells provide higher data rates as well as more energy-efficient way of communication for their users. However, the deployment of MTCN comes with several new technical challenges, such as interference management, network planning, and load distribution (Ghosh et al. 2012). In particular, load distribution strategy in MTCN is of crucial importance for the overall network performance (Ye et al. 2013), as it determines which user should be served by which BS and when. In this article, we propose a novel strategy for load distribution in the downlink of MTCNs by exploiting a combination of new context information related to the users trajectory profile, their QoS demands, and the current load of each cell.

Problem Formulation

Consider the downlink of a two-tier MTCN consisting of macrocells and picocells. Let $\mathscr{M}$,$\mathscr{P}$, and $\mathscr{N}$ denote the set of M macro base stations (MBSs), the set of P pico base stations (PBSs), and the set of N users, respectively. For each user $n \in \mathscr{N}$, we associate a unique profile $P_n(\tau, \theta, V)$ defined by three variables τ,θ, and V representing the urgency of the service in use, the direction of its motion, and its velocity, respectively. Next, we study these parameters, and we show how such information can improve the load distribution strategy.

Service Urgency

Depending on their application in use, the users may have varying needs for the wireless resources and tolerate different amounts of latency. To capture the dependency of the users' quality of service (QoS) to the service latency, we define the users' QoS as a decreasing function of the delivery time similar to Proebster et al. (2011):

$$Q_n(t) = \frac{1}{1 + \exp(t - \tau_n)}, \tag{1}$$

where τ_n is the urgency coefficient, which accounts for the urgency of the data. The smaller is τ_n, the more urgent is the data in use.

Users' Trajectory

Users travel through the network, and active communication sessions must be transferred among the cells. To guarantee the QoS, the network should avoid risky handovers that may lead to a signal loss or erroneous communication. A handover fails when the received SINR drops under a certain threshold. To capture the effect of the users' trajectory on the user-cell association problem, we derive the probability of handover failure when a user moves between different cells.

Without loss of generality, we assume circular coverage area for tractability (Lopez-Perez et al. 2012). Consider two picocells in the vicinity of each other and assume that a user is traveling through these two cells. When a user enters a cell, the total possible time of interaction is $T_i = \frac{D}{V}$, where D is the length of coverage circle chord traversed by the user and V is its speed. Also, assume that a successful handover process needs a certain preparation time of duration T_p. If $T_i > T_p$, the user is considered as a candidate to be served; otherwise, no handover would be initiated. Note

that $D = 2R\cos(\theta)$ where R is the radius of the coverage circle and θ is the trajectory angle measured with respect to the horizontal line in the polar coordinate system. Since the directions which the user takes into the picocell are equally likely, θ is modeled as a uniform random variable $\theta \sim \text{Uniform}\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Hence, the cumulative distribution function (CDF) of D is given by:

$$\begin{aligned}\Pr(D < d) &= 2\Pr\left(\theta > \cos^{-1}\left(\frac{d}{2R}\right)\right) \\ &= 1 - \frac{2}{\pi}\cos^{-1}\left(\frac{d}{2R}\right). \end{aligned} \tag{2}$$

We assume that for having a successful handover, the handover process must be done before the user reaches the distance $r < R$ of the picocell. When the user's path tangent to the circle with radius r around the picocell with coverage radius R, D is equal to $2\sqrt{R^2 - r^2}$. Handover fails when the user's path intersects the circle r; thus the probability of handover failure is:

$$\begin{aligned}&\Pr\left(D \geq 2\sqrt{R^2 - r^2}\right) \\ &= \frac{2}{\pi}\cos^{-1}\left(\sqrt{1 - \left(\frac{r}{R}\right)^2}\right). \end{aligned} \tag{3}$$

The expression in (3) shows that the handover process becomes more reliable as r becomes smaller relative to R. The ratio of r to R varies for each cell, and thus, the different cells guarantee different levels of reliability in handover process.

User-Cell Association as a Matching Game

Matching theory is a mathematical framework for modeling the combinatorial problem of matching players in two distinct sets, depending on the individual information and preference of each player (Gu et al. 2015). Here, we model the problem of user-cell association in MTCNs as a many-to-one matching game between the cells and the users.

Definition 1 A matching μ is a function from $\mathcal{N} \cup \mathcal{P}$ to $2^{\mathcal{N} \cup \mathcal{P}}$ such that $\forall n \in \mathcal{N}$ and $\forall p \in \mathcal{P}$: (i) $\mu(n) \in \mathcal{P}$ and $|\mu(n)| \leq 1$, (ii) $\mu(p) \in 2^{\mathcal{N}}$ and $|\mu(p)| \leq q_p$ where q_p is the quota of p, and (iii) $\mu(n) = p$ if and only if n is in $\mu(p)$.

The users, who are not assigned to any member of $\mathcal{P}$, will be assigned to the nearest macro BS. Members of $\mathcal{N}$ and $\mathcal{P}$ must have strict, reflexive, and transitive preferences over the agents in the opposite set. Next, exploiting the context information about the users' trajectory and service urgency, we introduce some properly defined utility functions to effectively capture the preferences of each set.

Users' Preferences

Users demand for reliable and high-quality communication. Therefore, they prefer the SCBSs which can guarantee the safety of communication during the handover and are able to deliver the data in an acceptable time. The transmission rate a user receives from a picocell depends on the current cell load and the interference caused by the neighboring cells. Therefore, the rate of transmission is a function of current matching μ. We define the rate over load for all user-cell pairs as:

$$\begin{aligned}\nabla_{ij} &= \frac{1}{\max(1, K_j)} \\ &\quad \log_2\left(1 + \frac{P_j c_{ij}}{\sum_{k \neq j} P_k c_{ik} + \sigma^2}\right), \end{aligned} \tag{4}$$

where K_j is the total users being served by picocell j and P_j denotes its power. c_{ij} represents the channel coefficient between user i and picocell j. In addition, σ^2 is the power of additive noise. We define the following utility function for user i when admitted by picocell j:

$$U_i(\mu, R_j, r_j, \nabla_{ij}) = \frac{R_j}{r_j}\nabla_{ij}. \tag{5}$$

The expression in (5) shows that the users prefer lightly loaded picocells to maximize their utility. Moreover, note that this utility function could

help to offload the heavily loaded macrocells by pushing the users to more lightly loaded cells.

Picocells' Preferences

The main goal of picocells is to increase the network-wide capacity by offloading traffic from the macrocells while providing satisfactory QoS for the users. Each picocell p could serve a limited number of users, called its quota q_p. Thus, by prioritizing the users coming from congested cells, the picocells could offload the heavily loaded cells. On the one hand, each candidate user is carrying a potential utility as a function of the pervious cell p' load, i.e., $f\left(\frac{K_{p'}}{q_{p'}}\right)$. This utility depends on the current matching which determines the number of users in neighboring cells. On the other hand, picocell p would also prioritize the candidate users based on their service requirement, i.e., service urgency defined in (1). We define the following utility that picocell p obtains by serving a user n in its coverage area:

$$U_p(\mu, \tau_n, k_{p'}, q_{p'}) = \left[1 + \log\left(\frac{\max(1, k_{p'})}{q_{p'}}\right)\right]\frac{1}{\tau_n}. \quad (6)$$

The first term in (6) accounts for the offloading concept, and the second term is the utility achieved by the picocells p for its service to a specific application. Notice that the small cell p gains more utility by giving service to the users having more urgent data. Using these utility functions, users build an ordered list of the picocells based on their own preferences and vice versa. Here we define the preference relationship for the users as follows:

Definition 2 The preference relation $\succ_n$ of the user $n \in \mathcal{N}$ over the set $\mathcal{P}$ is a function that compare two picocells p and p' such that:

$$\mu \succ_n \mu' \Leftrightarrow U_n(\mu) > U_n(\mu'). \quad (7)$$

The preference relation for the picocells is defined similarly. Users and picocells rank the members of the opposite set based on the defined preference relations. Our purpose is to match the users to the small cells so that the preferences of both sides will be satisfied as much as possible; thereby the network-wide efficiency would be optimized. To solve a matching game, one suitable concept is that of a stable matching (Gu et al. 2015). A matching is said to be two-sided stable, if and only if there is no blocking pair, i.e., there is no user-picocell pair (n, p) where n prefers p to its currently matched user picocell and p prefers n to its currently matched user n.

Next, an efficient algorithm for solving the matching problem is presented, which reaches to a stable matching between users and picocells.

Proposed Algorithm

The deferred acceptance algorithm is a well-known approach to solving the standard matching games (Roth and Sotomayo 1992). However, in our game, the preferences of agents as shown in (5) and (6) depend on externalities through the entire matching, unlike classical matching problems in which preferences are static and independent of the matching. Therefore, the classical approaches such as the deferred acceptance cannot be used here because of the presence of externalities. To solve the formulated game, we propose a novel algorithm shown in Table 1.

Suppose that all the users are initially associated to the nearest macro BS. Each user sends its profile information to the neighboring picocells. Each picocell, on the other side, ranks the candidate users based on its utility and feeds back the awaiting users with its own context information including its rate over load defined in (4) and its corresponding coverage and handover circle radii R and r. Each user makes a ranking list of the available picocells and applies to the most preferred one of them. The picocells keep the most preferred ones up to their quota and reject the others. The users who have been rejected in the former phase would apply to their next

Matching Game for Load Distribution in Multi-tier Cellular Networks, Table 1
Proposed algorithm for the user-cell association game

Input: context-aware utilities and the preferences of each user
Output: Stable matching between the D2D pairs

Initializing: All the UEs are assigned to the nearest BS

Stage I: **Preference Lists Composition**

- Neighboring D2D users exchange their context information
- Users sort the set of acceptable candidate based on their preference functions

Stage II: **Matching Evaluation**
while: $\mu^{(n+1)} \neq \mu^{(n)}$

- Update the utilities based on the current matching μ
- Construct the preference lists using preference relations
- Each user n applies to its most preferred partner
- Each user accept the most preferred applicants and create a waiting list while rejecting the others

Repeat

- Each rejected user applies to its next preferred partner
- Each user update its waiting list considering the new applicants

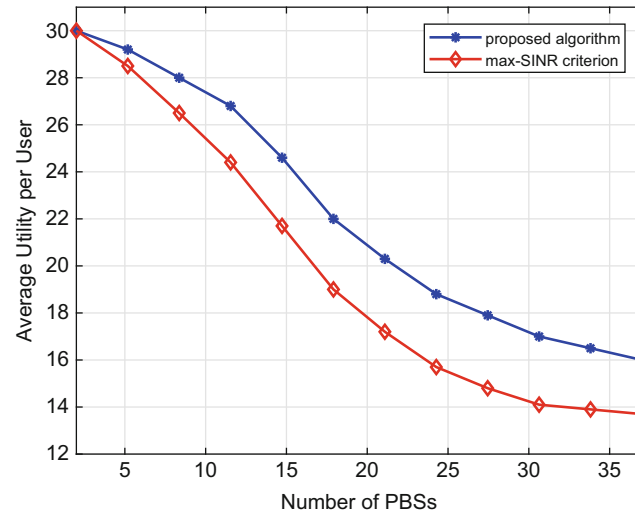
Until: all the users assigned to a waiting list
end

favorite picocell, and the picocells modify their waiting list accordingly. This procedure continues until all the users are assigned to a waiting list. However, since the preferences depend on the current matching μ, an iterative approach should be employed. In each step, the utilities would be updated based on the current matching. Once the utilities are updated, the preference lists would be updated accordingly as well. Therefore, in each iteration, a new temporal matching arises, and based on this matching, the interdependent utilities are updated as well. The algorithm initiates the next iteration based on the modified preferences. The iterations run on until two subsequent temporal matchings are the same and algorithm converges. We note that the proposed algorithm will lead to a stable matching when it converges. Indeed, the deferred acceptance in stage II would not converge if the matching is not stable. Hence, by contradiction, whenever the algorithm converges, the matching would be stable.

Simulation Results

For our simulations, we consider a single MBS with radius 1 km and overlaid by P uniformly deployed picocells. The channels suffer from a Rayleigh fading with parameter $\sigma = 2$. Noise level is assumed to be $\sigma^2 = -121$ dBm, and the minimum acceptable SINR for the UEs is 9.56 dB. There are N users distributed uniformly in the network. The QoS parameter τ_n is chosen randomly from the interval [0.5, 5] ms. The users have low mobility and can be assumed approximately static during the process time required for a matching. All the statistical results are averaged by 1000+ runs over random location of users and SCBSs, the channel fading coefficients, and other random parameters. The performance is compared with the max-SINR algorithm which is a well-known context-unaware approach exploited in wireless cellular networks for user-cell association. In this approach, each user is associated to the SCBS providing the strongest SINR.

Matching Game for Load Distribution in Multi-tier Cellular Networks, Fig. 1
Average utility per user for different number of PBSs with $N = 60$ users



Matching Game for Load Distribution in Multi-tier Cellular Networks, Fig. 2
Average utility per PBS for different number of users with $P = 20$ PBSs

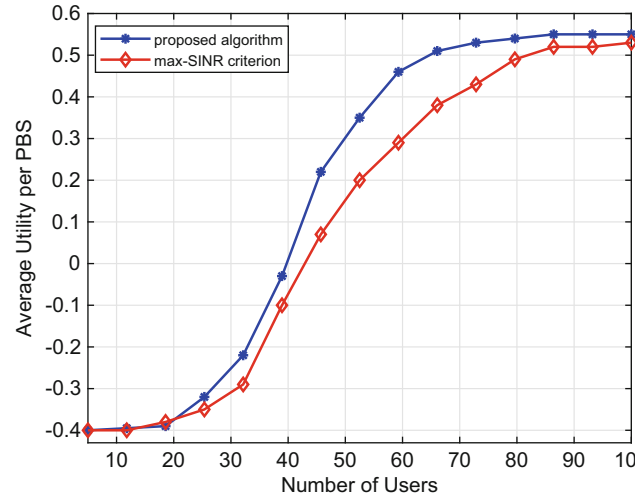


Figure 1 shows the average utility per user as a function of the number of PBSs for $N = 60$ users. As the number of PBSs increases, the average load in each cell decreases. Consequently, the available resources for each user would increase. However, increasing the number of PBSs leads to an increase in the interference between the neighboring cells and, consequently, decreases the rate. Hence, as shown Fig. 1, the average utility per user is a decreasing function with respect to the number of PBSs. We can also see that, when the number of PBSs exceeds 30, the slope of the curves becomes less steep due to the fact that the matchings would not be affected by these increments as much as before. Indeed, beyond this point, there are enough PBSs to serve all the users, and deploying more PBSs will not change the current matching dramatically. Figure 1 demonstrates that the proposed context-aware approach has a considerable performance advantage compared to the max-SINR approach. This performance advantage reaches up to 20.4% gain over to max-SINR criterion for a network with 30 PBSs.

Figure 2 shows the average utility achieved by each PBS as a function of the number of users for $P = 20$ PBSs. As the number of users N increases, the network becomes more congested, and the probability that a new user who applies for a PBS is coming from a

congested BS increases. Therefore, it is more likely for the PBSs to gain more utility by offloading the network. However, when the network is considerably congested, the new users that arrive to the network would be mostly assigned to the MBS, since many of PBSs are already servicing their maximum capacity. In this respect, Fig. 2 shows that, once the number of users exceeds 80, the average utility of PBSs remains constant. The proposed algorithm achieves up to 24.9% gain over the max-SINR approach when the number of users is 50.

Conclusions

In this work, we have proposed a new context-aware user association algorithm for the downlink of the multi-tier cellular networks. By introducing well-designed utility functions, our approach accounts for the trajectory and speed of the users as well as for their heterogeneous QoS requirements. We have modeled the problem as a many-to-one matching game with externalities, where the preferences of the players are interdependent and contingent on the current matching. To solve the game, we have proposed a novel algorithm that converges to a stable matching in a reasonable number of iterations. Simulation results have shown that the proposed approach yields considerable gains compared to max-SINR approach.

Cross-References

- ▸ Heterogeneous Wireless Networks Based on Cognitive Radio
- ▸ Interference-Aware Distributed Resource Allocation

References

Damnjanovic A, Montojo J, Wei Y, Ji T, Luo T, Vajapeyam M, Yoo T, Song O, Malladi D (2011) A survey on 3GPP heterogeneous networks. IEEE Wirel Commun 18:1021

Ghosh A, Mangalvedhe N, Ratasuk R, Mondal B, Cudak M, Visotsky E, Thomas TA, Andrews JG, Jo PXHS, Dhillon HS, Novlan TD (2012) Heterogeneous cellular networks: from theory to practice. IEEE Commun Mag 50(6):54–64

Gu Y, Saad W, Bennis M, Debbah M, Han Z (2015) Matching theory for future wireless networks: fundamentals and applications. IEEE Commun Mag 53(5):52–59

Lopez-Perez D, Guvenc I, Xiaoli C (2012) Theoretical analysis of handover failure and ping-pong rates for heterogeneous networks. In: IEEE international conference on communications (ICC), pp 6774–6774

Proebster M, Kaschub M, Valentin S (2011) Context-aware resource allocation to improve the quality of service of heterogeneous traffic. In: IEEE international conference on communications (ICC), pp 1–6

Roth AE, Sotomayo MAO (1992) Two-sided matching: a study in game-theoretic modeling and analysis. Cambridge University Press, Cambridge

Ye Q, Rong B, Chen Y, Al-Shalash M, Caramanis C, Andrews JG (2013) User association for load balancing in heterogeneous cellular networks. IEEE Trans Wirel Commun 12(6):2706–2716

Mathematical Model

- ▸ Reaction-Diffusion Channels

Maximum Likelihood Sequence Detection

- ▸ Equalization Techniques for Single-Carrier Modulations

Media Access Control for Narrowband Internet of Things: A Survey

Yuyi Sun, Shibo He, and Fei Tong
The State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China

Introduction

Low-power wide-area network (LPWAN) has emerged as an attractive communication platform

for Internet of Things (IoT) (Eletreby et al. 2017). Among a variety of communication technologies in LPWAN, narrowband IoT (NB-IoT), standardized in the third generation partnership project in Release 13 (3GPP 2015), is a promising technology that attracts attention from both industry and academia. NB-IoT inherits long term evolution (LTE) and has its distinctive advantages, including (1) massive connections, (2) wide-area coverage, (3) low power consumption, and (4) low cost (Wang et al. 2017; Beyene et al. 2017; Zayas and Merino 2017).

Media access control (MAC) protocol in NB-IoT resolves channel collisions among multiple user equipments (UEs) (3GPP TS 36.211; 3GPP TS 36.331; 3GPP TS 36.321), by efficiently allocating channel resources of the evolved Node Bs (eNBs) to UEs. When there is a data transmission demand, a UE initiates random access procedure to occupy a channel and transmits data to an eNB. Clearly, random access protocol suffers from severe collisions when there are a large number of UEs in NB-IoT, leading to more power consumption. Therefore, it is significant to efficiently analyze the performance of random access in NB-IoT.

In LTE, MAC protocols can be divided into contention-free and contention-based MAC protocols, and here we focus on contention-based MAC protocols, i.e., contention-based random access procedure. Based on random access procedure, there are some existing models designed for LTE, which cannot be applied to NB-IoT directly since there are a large number of UEs in NB-IoT and meanwhile, wide-area is required. Having this in mind, we provide a survey of the existing MAC models for LTE and NB-IoT in this paper.

MAC for LTE

Random Access Procedure in LTE

Random access procedure in LTE consists of four steps: random access preamble, random access response (RAR), scheduled transmission, and contention resolution.

- *Random Access Preamble:* a UE selects one of the preamble sequences from two groups and transmits it to an eNB.
- *RAR:* after the eNB receives the preamble sequence, it transmits an RAR to the UE. If more than one UE selects the same preamble sequence, they will receive the same RAR.
- *Scheduled Transmission:* after receiving the RAR, the UE transmits/UEs transmit a message that contains a unique identity.
- *Contention Resolution:* if the eNB can decode one of the messages from scheduled transmission, it replies to the UE to transmit data.

MAC Models for LTE

Data transmission, power mechanism, and energy consumption are crucial to MAC for LTE. Previous studies have proposed different models to analyze such problems.

Markov chain was adopted to model random access procedure and analyze data transmission in LTE (Seo and Leung 2012). Based on semi-persistent scheduling (SPS) in random access for LTE, UEs are assumed to have an infinite buffer to store packets. Markov chain is used to model the SPS in LTE, including the states ON and OFF of the traffic source, the contention and transmission states of UEs. Equilibrium point analysis is investigated, obtaining the number of UEs in steady state, through which the probability of packet dropping is analyzed.

Power ramping mechanism was modeled during random access in LTE in terms of preamble signal to interference plus noise ratio (SINR) and preamble collisions (Misic et al. 2017). The ratio of external interference to received signal power is assumed as a Gaussian distribution, and the probability of missed preamble detection and the failure caused by eNB outage can be analyzed. Performance evaluation reveals that the downlink resources bring more success probability.

Zhao et al. (2017) reduced power consumption in a random access algorithm based on statistic waiting. After investigating random access procedure in LTE, they used slotted ALOHA to simulate packet transmission. A UE initializing random access procedure receives the success

rate of the last time slot, which helps the UE decide whether to transmit a packet or not and reduces energy consumption.

Previous MAC models in LTE cannot be directly applied to NB-IoT due to the following reasons: (a) there are three coverage levels defined by minimum coupling loss (MCL) in NB-IoT (Ng et al. 2009), which covers a wider range than that in LTE, requiring more retransmission times to achieve wide-area coverage. (b) Due to massive connected UEs in NB-IoT, a more efficient MAC model is needed to resolve a great deal of collisions. (c) Due to low-traffic rate and low-power consumption requirements, UEs in NB-IoT have a finite buffer to cache data packets, while LTE requires higher data rate. Thus, the length of buffer and larger retransmission number need to be considered simultaneously in NB-IoT to assign channel resources. We introduce the MAC models for NB-IoT in the next section.

MAC for NB-IoT

Random Access Procedure in NB-IoT

Random access procedure in NB-IoT also consists of four steps: random access preamble, RAR, scheduled transmission, and contention resolution.

- *Random Access Preamble:* a UE transmits a preamble to an eNB, which has a cyclic prefix and five symbols 1. Physical resources are configured based on MCL.
- *RAR:* after the eNB receives the preamble sequence, it transmits an RAR to the UE. If more than one UE transmits the same preamble sequence, they will receive the same RAR.
- *Scheduled Transmission:* after receiving the RAR, the UE transmits/UEs transmit a message that contains a unique identity, power headroom report, buffer states, and data volume.
- *Contention Resolution:* the eNB randomly selects a UE to decode its information received from scheduled transmission, then it replies to the UE to transmit data.

MAC Models for NB-IoT

Previous MAC models for NB-IoT mainly focus on coverage requirement and system throughput.

Lin et al. (2016) designed a frequency hopping pattern for NB-IoT with single tone random access channel signal. Based on the rationale of random access channel frequency hopping, the authors proposed corresponding receiver algorithms and time-of-arrival (ToA) estimation. Comparing the statistic value of ToA and a predefined threshold can help the eNB determine the presence of the preamble sequence. Simulation results reveal that the detection probabilities exceed 99%, which can meet the coverage requirement.

Based on probabilities characterization, first-in-first-out (FIFO) queue model, and Markov chain, a system throughput model was proposed and analyzed (Sun et al. 2017). Based on random access procedure, the UEs' buffer is denoted as a FIFO queue, and the probability that a FIFO queue is empty, the probability that a packet is transmitted successfully, and the probability that a channel is busy are characterized based on backoff mechanism. Markov chain is adopted to model retransmission number and queue length simultaneously to achieve the three probabilities in steady state. The expression of the system throughput can be obtained and calculated due to the three probabilities above. Extensive simulation results show that the number of UEs and the arrival rate of packets have obvious effects on system throughput, while the maximum retransmission number and the packet length have slight influences on the system throughput.

Harwahyu et al. (2018) investigated a promising optimized random access model for NB-IoT. NB-IoT systems have three coverage enhancement (CE) levels due to different MCL. A UE initializes random access to occupy a channel from its initial CE level, which depends on the reference signal received power. When collisions happen, if the UE does not transmit successfully till the maximum retransmission number in this CE level is reached, it will switch to the next higher CE level to retransmit. When the global maximum retransmission number is reached, the UE will initialize random access again from its

initial CE level. Thus, in this entry, multiband multichannel slotted ALOHA is used to model this procedure. The first transmission and the retransmission at the same CE level and in the next higher CE level are analyzed. The optimization of random access for NB-IoT in this entry includes that: (a) more CE levels bring larger access success probability, (b) when the number of remaining retransmissions in a certain band is equal to the number of global retransmissions, access success probability increases, and (c) comparing number of successful UEs in each CE level and in their first transmission by exhaustive search can achieve the largest access success probability. Analytical model is supported by extensive simulations. The influences of parameters such as the number of UEs, the repartition ratios of subchannels, and retransmission times for different levels on the access success probability are analyzed.

In Li et al. (2018), an optimization to maximize the system throughput in random access narrowband cognitive radio IoT was introduced. This entry used slotted ALOHA to model random access procedure, including autonomous and collaborative sensing. With unique optimal detection probabilities in network level characterized, the throughput optimization of random access with autonomous sensing and collaborative sensing can be maximized. From the simulations, the system throughput with high signal noise ratio (SNR) and system interference tolerance is higher than that with low SNR and system interference tolerance. Autonomous sensing is better under high SNR, while collaborative sensing has a better performance under low SNR. The arrival rate of data is a Poisson process. Smaller arrival rate of data increases the maximum throughput rapidly and then, the maximum throughput reaches a peak, while larger arrival rate of data makes the maximum throughput decrease. As the arrival rate of data tends to infinity, the probability of detection converges to one, and the optimal probability of false alarm becomes a global optimal problem.

The NB-IoT development system in Chen et al. (2017) includes IoT cloud platform, NB-IoT development board, mobile phones, and application server. The development board embeds the entire protocol of NB-IoT, including MAC and physical protocols. Based on the applicable system, practical experiments about MAC for NB-IoT can be done, which helps design efficient MAC models for NB-IoT.

Media Access Control for Narrowband Internet of Things: A Survey, Table 1 Summary of MAC models for NB-IoT

Literature	Design	Analysis
Lin et al. (2016)	Frequency hopping pattern	ToA estimation
Sun et al. (2017)	Finite FIFO queue for UE buffer	System throughput
Harwahyu et al. (2018)	Multiband multichannel slotted ALOHA	Access success probability
Li et al. (2018)	Slotted ALOHA for MAC	Throughput optimization
Chen et al. (2017)	Development system	Practical application

Here we summarize the MAC models for NB-IoT in Table 1.

Conclusion

In this entry, we make a survey about the MAC for NB-IoT. Based on random access procedure, there are a lot of MAC models existing for LTE. However, such models cannot be applied to NB-IoT systems due to massive connections and wide-area coverage in NB-IoT. Thus, based on the features of NB-IoT, some works design new models for MAC in NB-IoT. In the future, some more parameters, such as the system latency, packet loss can be analyzed for the system performance of MAC for NB-IoT.

Acknowledgments This work was supported by NSFC under grant 61672458.

References

3GPP (2015) New Work Item: NarrowBand IoT (NB-IoT). The 69th SG RAN meeting. Available: https://www.3gpp.org/FTP/tsgran/TSGRAN/TSGR69/Docs/RP-151621.zip

3GPP TS 36.211, V13.2.0 2016 Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation; Protocol specification (Release 13)
3GPP TS 36.321, V13.2.0 2016 Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC); Protocol specification (Release 13)
3GPP TS 36.331, V13.2.0 2016 Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 13)
Beyene YD, Jantti R, Tirkkonen O, Ruttik K, Iraji S, Larmo A, Tirronen T, Torsner J (2017) NB-IoT technology overview and experience from cloud-RAN implementation. IEEE Wirel Commun 24(3):26–32
Chen J, Hu K, Wang Q, Sun Y, Shi Z, He S (2017) Narrowband internet of things: implementations and applications. IEEE Internet Things J 4(6):2309–2314
Eletreby R, Zhang D, Kumar S, Yağan O (2017) Empowering low-power wide area networks in urban settings. In: Proceedings of the conference of the ACM special interest group on data communication. ACM, Los Angeles, pp 309–321
Harwahyu R, Cheng RG, Wei CH, Sari RF (2018) Optimization of random Access Channel in NB-IoT. IEEE Internet Things J 5(1):391–402
Li T, Yuan J, Torlak M (2018) Network throughput optimization for random access narrowband cognitive radio internet of things (NB-CR-IoT). IEEE Internet Things J 5:1436. https://doi.org/10.1109/JIOT.2017.2789217
Lin X, Adhikary A, Wang YPE (2016) Random access preamble design and detection for 3GPP narrowband IoT systems. IEEE Wireless Commun Lett 5(6):640–643
Misic J, Misic, VB, Ali MZ (2017) Explicit power ramping during random access in LTE/LTE-A. In: Proceedings of IEEE wireless communications and networking conference, IEEE, San Francisco, pp 1–6
Ng MH, Lin SD, Li J, Tatesh S (2009) Coexistence studies for 3GPP LTE with other mobile systems. IEEE Commun Mag 47(4):60–65
Seo JB, Leung VCM (2012) Performance modeling and stability of semi-persistent scheduling with initial random access in LTE. IEEE Trans Wirel Commun 11(12):4446–4456
Sun Y, Tong F, Zhang Z, He S (2017) Throughput modeling and analysis of random access in narrowband internet of things. IEEE Internet Things J 5:1485. https://doi.org/10.1109/JIOT.2017.2782318
Wang YPE, Lin X, Adhikary A, Grovlen A, Sui Y, Blankenship Y, Bergman J, Razaghi HS (2017) A primer on 3GPP narrowband internet of things. IEEE Commun Mag 55(3):117–123
Zayas AD, Merino P (2017) The 3GPP NB-IoT system architecture for the internet of things. In: Proceedings of IEEE communications workshops, IEEE, Paris, pp 277–282
Zhao Y, Yang H, Liu K, Huang L, Shi M, Zhang M (2017) A random-access algorithm based on statistics waiting in LTE-M system. In: Proceedings of international conference on computer science and education, IEEE, Houston, pp 214–218

Media Access Control Protocol for Cognitive Radio Networks

▸ MAC in Cognitive Radio Networks

Medical Internet of Things

▸ Internet of Medical Things

Medium Access Control

▸ Interference-Aware Distributed Resource Allocation

Medium Access Control in Millimeter-Wave Wireless Communications

▸ Millimeter Wave MAC Layer

Message-Driven Frequency Hopping

▸ Index Modulation for OFDM

Middleware for Wireless Sensor Networks

Flavia C. Delicato and Paulo F. Pires
DCC-IM/PESC/COPPE – Federal University of Rio de Janeiro, Cidade Universitária, Rio de Janeiro, Brazil

Synonyms

Software platforms for WSN; WSN frameworks; WSN management

Definitions

Wireless Sensor Networks (WSNs) are composed of the interconnection, often through wireless links, of devices equipped with sensing, processing, storage, and communication capabilities. Such devices are called sensor nodes, and a WSN can contain tens to thousands of them. Sensor nodes are tiny in size, have reduced computational resources, and are usually battery-powered, thus making energy efficiency a crucial issue to extend the WSN lifetime. Each node contains one or more sensing units, capable of performing measurements of physical variables such as vibrations, acceleration, pressure, temperature, light, among others, converting them to digital signals. The sensors act in a collaborative way, extracting environmental data, performing simple processing, and transmitting them to one or more exit points, called sink nodes, gateways or base stations, to be analyzed and further processed. The data measured by the sensors and transformed into digital signals are translated into information about a phenomenon of interest to humans.

The term middleware, although very popular in computer science, can be a bit confusing to many people. One reason for this is the myriad of different types of middleware systems in use in the everyday software development tasks. In distributed systems, middleware is usually defined as an infrastructure software that lies between the underlying operating system, networks, and hardware and the applications running on each node of the system. In general, middleware is supposed to hide the lower-level hardware and internal operation and heterogeneity of the underlying system, providing standard interfaces, abstractions, and a set of services. Another role of middleware is to provide reusable set of services that can be combined and customized to create distributed systems faster and more reliably by integrating components that may be developed by different software infrastructure suppliers (Schantz and Schmidt 2001). By hiding the inherent complexity of distribution, the use of middleware in traditional distributed computing facilitates the work of application developers. Tasks such as concurrency control, transactions, data replication, security, and other infrastructure services are examples of services performed by a middleware. In the context of WSN, the main purposes of a middleware are supporting the development and execution of sensing-based applications and managing the utilization of the network and sensor node's resources while meeting applications' requests.

Historical Background

Despite the well-known advantages of using middleware platforms in traditional distributed systems, only from the early 2000s the researchers began to consider its adoption in the WSN design (Bonnet et al. 2001; Capra et al. 2003; Delicato et al. 2003; Murphy et al. 2001; Shen et al. 2001). Considering the simplicity of the first applications and the application-specific nature of early WSNs, the overhead of adding a middleware layer in the network design was often not worth. Therefore, the first generation of WSNs was traditionally designed from scratch in a domain-specific and task-oriented fashion. WSN systems were built as monolithic software for a specific target platform and operating system and addressing the requirements of a single target application, with little or no possibility of reusing them for newer applications. However, to accommodate the new scenarios of shared and Internet-scale WSN systems, novel systematic design approaches based on high level and preferable standardized abstractions and services were required. Thus, the support of middleware for WSNs became a crucial need to provide: (i) appropriate **system abstractions**, so that application developers can focus on the application logic and not dealing with lower level implementation details, (ii) **standardized and reusable services**, so that developers can deploy and execute application without worrying about complex, error-prone, and tedious infrastructure-level functions, such as network and resource management, context-aware adaptation, among others, (iii) **runtime environment** able to manage the execution of

multiple applications. Middleware support is also necessary to provide interoperability between different WSN systems, with external networks, as the Internet, or with enterprise systems.

To meet such demands, from the 2000s several proposals of middleware systems specially tailored for WSN emerged, each with different goals and approaches. It can be said that database-based approaches (Bonnet et al. 2001; Madden et al. 2005; Shen et al. 2001) were the first attempt to propose a rudimentary middleware layer for the project of WSNs. In the COUGAR system (Bonnet et al. 2001), the processing capabilities of sensor nodes are used to perform part of the processing of queries within the network, instead of centralizing such processing only at the sink nodes. In Bonnet et al. (2000), an SQL-based declarative query language is proposed for users to submit their queries to the WSN. Also following this line, in Shen et al. (2001), an architecture for sensor networks called SINA is presented, which provides mechanisms for query, monitoring, and submission of tasks. SINA plays the role of a middleware that abstracts WSN nodes of an RSSF as a collection of distributed objects. Users access and extract information from a WSN either by issuing declarative queries or requesting tasks using programming scripts. SINA adopts the proprietary SQTL procedural scripting language as the interface between applications and the SINA middleware. An event-driven approach is adopted as a programming model.

From these early initiatives, most of which adopted approaches with little flexibility and generally tied to specific communication protocols, a myriad of new proposals for WSN middleware began to emerge, highly diversified regarding their design approaches, abstractions and programming models. Examples of approaches are (i) event-driven, (ii) service-oriented (Mohamed and Al-Jaroodi 2011), (iii) virtual machine-based, (iv) agent-based (Chen et al. 2006; Fok et al. 2005), (v) tuple-spaces (Costa et al. 2006), (vi) component-based, and (vii) application-specific (Heinzelman et al. 2004), with the first three being the most prominent ones. Some middleware platforms use a combination of approaches. For example, many service-oriented middleware also employs VMs in their design and development.

Virtual Machine-based approach allows developers to write application code in separate, small modules that are injected and distributed throughout the network using specialized algorithms. The goal is minimizing the overall energy consumption and resource usage. The Virtual Machine (VM) in each node then interprets the injected modules. Examples of this approach include Maté (Levis and Culler 2002), ASVM (Levis et al. 2005), and DAViM (Michiels et al. 2006).

Another widely adopted programming approach to WSN middleware is based on the concept of events (Boonma and Suzuki 2012; Silva et al. 2014; Souto et al. 2004). In such approach, components, applications, and all the other participants interact through events. Applications specify their interest in given changes of state in the monitored environment (basic events). Upon detecting an event, a sensor node sends an event notification towards interested applications. The application can also specify patterns of events (composite events). Events are propagated from the event producers (sensor nodes), to the event consumers (applications).

Among the existing approaches, the service-oriented approach has been highlighted by providing innumerable advantages in the context of WSN middleware. The service-oriented design paradigm builds software or applications as services. Service-oriented computing (SOC) is based on the service-oriented architecture (SOA) and has been traditionally used in corporate information systems. The features of SOC, such as technology neutrality, loose coupling, service reusability, composability, and discoverability, are also potentially beneficial to WSN systems. Among its advantages, it can be mentioned that this approach offers a generic and flexible programming model that favors the interoperability between different applications and the network. By exposing the functionalities of the sensors as services, it offers a more flexible architecture in comparison, for example, with

databases approaches. SQL is basically a query language, not having a suitable semantics to represent the submission of tasks and actuation activities in the WSN.

Finally, Component-Based Software Engineering (CBSE) is a modern methodology that proposes software construction by plugging software components. Based on component interoperability, this programming approach aims at building more flexible and adaptable software. Recently, some middleware proposals based on CBSE have emerged in the WSN field (Costa et al. 2007; Man et al. 2016).

Foundations

WSN technology has made tremendous progress in the last decade, drawing the attention of the scientific community and industry. WSNs are the key components of the emerging Internet-of-Things (IoT) paradigm (Man et al. 2016). With their ability to instrument the physical environment, WSNs naturally constitute the core infrastructure from which IoT systems can be built for a myriad of domains. With the introduction of the IoT, it is envisioned that future WSN deployments will have to support multiple applications simultaneously. To enable this scenario, WSNs will have to evolve from application-specific systems to networks capable of sharing data and resources among several applications, in an efficient, fair, and transparent way. Of course, such sharing creates new challenges in promoting the interoperability of different sensor nodes and WSNs while achieving the desired energy efficiency and satisfying the requirements of multiple applications. Middleware is a technology that can potentially enable WSN sharing and provide solutions to these challenges.

In short, a middleware for WSN must provide support for the development, management, deployment, and execution of sensing-and-actuation-based application. Among other functions, the middleware can decide the best protocols to be used according to the application requirements, coordinate the operation of sensors to achieve the application goals, and intelligently manage the use of device and network resources. To efficiently provide the quality of service required by applications, it is often necessary to interact with the lower levels of protocols or even with hardware components. Middleware services can be developed to perform this interaction on behalf of applications. The provision of all these functions by a middleware system facilitates the tasks of both the application developers and of the network managers.

Since WSNs have several distinct features and constraints, conventional middleware technologies are not suitable for these networks. Therefore, the development of new middleware systems, specifically tailored to WSN is required. The following paragraphs present a set of requirements that WSN domain poses on middleware system development.

Resource Constraint: WSN nodes are constrained devices regarding their processing power, memory capacity, bandwidth, and energy. The processing and memory constraints mean that any software solution for WSNs must be computationally light. Any protocol or algorithm for such networks needs to be designed aiming at the efficient use of available resources. The execution of application tasks must be performed in a way that balances the usage of resources and the meeting of QoS requirements. Moreover, since the battery life is limited, and the exchange or even recharge of batteries in a large-scale network is undesirable or even unfeasible, a key requirement of WSN is to have mechanisms to manage the energy consumption in sensor nodes in an intelligent way.

Scale: A single WSN may consist of hundreds to thousands of nodes and to efficiently manage this kind of computational environment is challenging. The traditional concept of network management captures methods and tools related to the operation, administration, maintenance, and provisioning of networked systems. However, to manage WSN systems, composed of heterogeneous nodes with different requirements and operational properties, a paradigm shift is necessary. Upper layers need to efficiently capture dynamic system changes, and lower layers need to transform that information into

appropriate, autonomous, and scalable actions. In recent years, several extensions have been proposed for the traditional definition of systems and network management, which are specifically designed to address the increasing complexity and dynamism of these environments. Besides controlling the network operation, the high number of nodes will require mechanism to coordinate them in the execution of the required sensing tasks.

Heterogeneity: The WSN environment is composed of devices, communication links, protocols, and software components that are heterogeneous in terms of capabilities, programming abstractions, models, and development tools. Such heterogeneous nature of WSNs emerges not only from differences in capacity and features but also for other reasons including multivendor products and application requirements. Such heterogeneous environment with many sensors of different hardware platforms running applications developed by different teams will require using an abstraction layer that provides a common way to program, access the networks, and extract the sensing data.

Data Processing: Sensors generate a huge amount of data, typically consisting of time-series values, which are sampled over a specific period of time, thus characterizing a data stream. The input rate of a data stream ranges from a few bytes per second to a few gigabits per second. Such rate can be irregular, unpredictable, and bursty in nature. In addition, sensor data are subject to different degrees of accuracy. The network, by its very nature, has only the ability to sample physical processes, which in turn implies that the obtained result approximates the real parameters observed. Therefore, the data in WSNs have an extra dimension, namely, the accuracy. In this context, accuracy can be defined as the degree to which the value provided by the sensors approaches the "real" value of the monitored physical phenomenon. Some factors that contribute to the accuracy of the data measured by a sensor are its nominal accuracy (from the manufacturer) and its distance from the phenomenon, in addition to environmental noise.

Diverse application: WSN-based systems can offer its services to many applications in numerous domains and environments. Different applications are likely to need different deployment architectures (e.g., event-driven and time-driven), have different requirements, and often must be served with different priorities. For instance, time critical and mission critical applications demand some level of priority in the access to the shared resources, in detriment of non-critical applications.

Context-awareness: Context is key in WSNs and their applications. A large number of sensors will generate large amounts of data, which will not have any value unless they are analyzed, interpreted, and understood. Context-aware computing stores context information related to sensor data, easing its interpretation. More specifically, location or spatial information about sensors is critical, as location plays a vital role in context-aware computing. In a large-scale WSN, interactions are highly dependent on their locations, their surroundings, and presence of other entities (e.g., other systems, neighboring nodes, and people).

Dynamicity: WSNs are subject to unpredictable changes in their execution context. Besides the intrinsic variations in the network connectivity, typical of wireless links, the availability of sensors in the network is also highly variable due to node failures, energy depletion and mobility. The network topology is also frequently changed by topology control protocols that decide when and which sensors should be turned off to save power. Within such an ad hoc environment, where there is limited or no connection to a fixed infrastructure, it will be difficult to maintain a stable network to support many application scenarios. Nodes will need to cooperate to keep the network connected and active. Moreover, the required sensing tasks, that dictate the network behavior, can also change since the emergent WSN are assumed to be used by different applications (concurrently or not). These characteristics demonstrate the high degree of dynamism in the execution context of WSNs. Therefore, it is crucial that such networks are endowed with mechanisms that provide context-awareness and adaptation.

Data-Centricity: Sensor nodes are used to collect data from its surroundings. Since

applications are interested in the data collected by a sensor and express their interests in terms of type of data and QoS requirements, the scheme for node addressing in a WSN should be based on attributes that describe the node capabilities to provide data, instead of using a scheme based on any global and unique network address. Moreover, individual readings of a node are often not relevant to the application, which is more interested in the merged or synthesized information gathered from different nodes. Considering the typical spatial density in a WSN, there is often a redundancy in the data collected for different nodes inside a given region of interest. Since communication is in general more energy consuming than data processing, individual measurements should be aggregated as near to the data source as possible so that only the resulting, relevant information is transmitted to sink nodes. Therefore, in-network, data-oriented processing should be performed in every node of the WSN, aiming at reducing data redundancy, increasing the relevance of the information reported back to the application and saving energy.

Security Issues: Since WSN can be deployed for sensitive applications like military surveillance and forecasting systems, security requirements such as confidentiality, authentication, integrity, freshness, and availability should be considered. Since most existing algorithms and security models are not suitable for WSNs, solutions specifically tailored for these networks must be designed that consider the node resource constraints.

WSN Middleware Services

All services provided by the middleware must respect the intrinsic characteristics of WSNs, especially scalability and energy efficiency. The middleware should provide robustness/fault tolerance and adaptability features to deal with the dynamic execution context of WSNs and the intermittence of wireless connections. Given the data-driven nature of these networks, middleware must provide data-centric mechanisms for processing and querying data within the network. In addition, application-specific knowledge can be used to optimize network operation, while sharing resources across multiple concurrent applications. Due to the restricted capabilities of nodes, the middleware must be light in terms of its communication and processing demands. The main services to be provided by a WSN middleware are described as follows.

Resource Management: WSN middleware should provide services to manage and optimize the network resource usage in an efficient way while meeting the application requirements. One example of such services is to select the set of nodes that will participate in a given sensing task, always considering the tradeoff between meeting the application requirements and optimizing energy consumption. Another useful service is to implement data fusion techniques to merge sensor readings of individual sensors into a high-level result to be sent to the application, thus saving transmission energy while increasing the data accuracy.

Context Management and Dynamic Adaptation: Considering the high dynamism of WSN environments, in which devices may become unavailable for a variety of reasons, middleware platforms must provide strategies for dynamic adaptation, thus ensuring the availability and quality of the applications during their execution. This is particularly important for applications in critical domains, since flaws or degradation of quality parameters in such applications can be a threat to human life and health. Thus, WSN middleware must remain available and functioning properly in this dynamic environment. To do so, it needs to provide mechanisms for collecting and analyzing contextual data, and responding to changes in the execution context, without decreasing the WSN system overall performance. Adaptive measures must be applied whenever the target operational performance metrics are not met. Examples of adaptive actions are turning on/off sensors, changing the data sensing/sending rates, enabling/disabling data aggregation functions, and changing the routing protocol in use on the network. QoS metrics and adaptation plans are managed by the middleware via policies and rules predefined by applications.

Resource and Service Discovery: WSN resources are very heterogeneous, including from low-level processing, communication, and sensing capabilities of individual sensor nodes to high-level services such as processed data and complex events detected from analyses of raw sensing data. For applications to make use of these resources, middleware must provide mechanisms for their discovery. In addition, in the absence of network infrastructure (since WSN are often ad hoc in nature), the middleware must provide two levels of service discovery. The internal level of discovery is used so that the sink nodes are aware of all the capacities of the sensor nodes that make up the network. Likewise, neighbors must know each other's resources. This is done by advertising mechanism that may rely on protocols and data formats common within a same WSN. The external level of service discovery is used by an application to find out which WSN provide the services it requires and how to access those services. In this case, the middleware should provide a standardized interface to access the WSN resources, abstracting its specific formats and protocols. Considering the dynamic environment of WSNs, assumptions related to global and deterministic knowledge of the resources' availability do not hold. Moreover, considering the high scale and geographical features of the WSN deployment, human intervention for resource discovery is infeasible. Therefore, an important requirement for WSN middleware is providing automatic resource discovery, based on high-level descriptions of application functional and QoS requirements.

Code management: Deploying code in a WSN environment is challenging and should be directly supported by the middleware (Razzaque et al. 2016). In particular, code allocation and code migration services are required. Code allocation selects the set of devices or sensor nodes to be used to accomplish a user or application-level task. Code migration transfers one node's code to another, potentially reprogramming nodes in the network, for instance to address requirements of arriving applications in a dynamic way. Approaches such as VM and agent-based help provide this type of service, since they offer a common execution environment to run migrated code.

In addition to these, there are several generic services that a WSN middleware can provide, such as: naming, location, security, and time synchronization.

Key Applications

Military applications such as target detection, battlefield surveillance, and counterterrorism originally motivated WSN applications. However, sensor nodes are very flexible and can be used for continuous data sensing, event detection, and local actuator control, as well as other tasks. The advantages of WSN over traditional networks resulted in many other potential applications, ranging from infrastructure security to industrial/factory control, environment and habitat monitoring, healthcare applications, home automation, and traffic control. The design and development of a successful middleware must address challenges posed by WSN features on one hand and the application demands on the other hand.

Cross-References

- Data Gathering in Wireless Sensor Networks
- Information-Centric Wireless Sensor Networks
- Protocol Stack Architecture for Wireless Sensor Networks

References

Bonnet P, Gehrke J, Seshadri P (2000) Querying the physical world. IEEE Pers Commun 7:10–15

Bonnet P, Gehrke J, Seshadri P (2001) Towards sensor database systems. In: International conference on mobile data management. Springer, London, pp 3–14

Boonma P, Suzuki J (2012) Chapter 41, TinyDDS: an interoperable and configurable publish/subscribe middleware for wireless sensor networks. In: Information Resources Management Association (ed) Wireless technologies: concepts, methodologies, tools and applications. IGI Global, Hershey, pp 819–846

Capra L, Emmerich W, Mascolo C (2003) Carisma: context-aware reflective middleware system for mobile applications. IEEE Trans Softw Eng 29:929–945
Chen M, Kwon T, Yuan Y, Leung VC (2006) Mobile agent based wireless sensor networks. J Comput 1:14–21
Costa P, Mottola L, Murphy AL, Picco GP (2006) TeenyLIME: transiently shared tuple space middleware for wireless sensor networks. In: Proceedings of the international workshop on middleware for sensor networks. ACM, New York, pp 43–48
Costa P et al (2007) The RUNES middleware for networked embedded systems and its application in a disaster management scenario. In: Fifth annual IEEE international conference on pervasive computing and communications, PerCom '07. IEEE, White Plains, pp 69–78
Delicato FC, Pires PF, Pirmez L, da Costa Carmo LFR (2003) A flexible middleware system for wireless sensor networks. In: Endler M (ed) Proceedings of the ACM/IFIP/USENIX 2003 international conference on middleware (Middleware '03). Springer, New York, pp 474–492
Fok C-L, Roman G-C, Lu C (2005) Rapid development and flexible deployment of adaptive wireless sensor network applications. In: Proceedings of 25th IEEE international conference on distributed computing systems, ICDCS 2005. IEEE, Columbus, pp 653–662
Heinzelman WB, Murphy AL, Carvalho HS, Perillo MA (2004) Middleware to support sensor network applications. IEEE Netw 18:6–14
Levis P, Culler D (2002) Maté: a tiny virtual machine for sensor networks. ACM SIGPLAN Not 10:85–95. ACM
Levis P, Gay D, Culler D (2005) Active sensor networks. In: Proceedings of the 2nd conference on symposium on networked systems design & implementation (NSDI '05), vol 2. USENIX Association, Berkeley, pp 343–356
Madden SR, Franklin MJ, Hellerstein JM, Hong W (2005) TinyDB: an acquisitional query processing system for sensor networks. ACM Trans Database Syst (TODS) 30:122–173
Man KL, Hughes D, Guan S-U, Wong PW (2016) Middleware support for dynamic sensing applications. In: 2016 International conference on platform technology and service (PlatCon). IEEE, Jeju, pp 1–4
Michiels S, Horré W, Joosen W, Verbaeten P (2006) DAViM: a dynamically adaptable virtual machine for sensor networks. In: Proceedings of the international workshop on middleware for sensor networks (MidSens '06). ACM, New York, pp 7–12
Mohamed N, Al-Jaroodi J (2011) A survey on service-oriented middleware for wireless sensor networks. SOCA 5:71–85
Murphy AL, Picco GP, Roman G-C (2001) Lime: a middleware for physical and logical mobility. In: 21st International conference on distributed computing systems. IEEE, Mesa, pp 524–533
Razzaque MA, Milojevic-Jevric M, Palade A, Clarke S (2016) Middleware for internet of things: a survey. IEEE Internet Things J 3:70–95
Schantz RE, Schmidt DC (2001) Middleware for distributed systems: evolving the common structure for network-centric applications. In: Encyclopedia of software engineering, vol 1. Wiley, New York, pp 1–9
Shen C-C, Srisathapornphat C, Jaikaeo C (2001) Sensor information networking architecture and applications. IEEE Pers Commun 8:52–59
Silva JR, Delicato FC, Pirmez L, Pires PF, Portocarrero JM, Rodrigues TC, Batista TV (2014) PRISMA: a publish-subscribe and resource-oriented middleware for wireless sensor networks. In: Proceedings of the tenth advanced international conference on Telecommunications. Citeseer, Paris, p 8797
Souto E, Guimarães G, Vasconcelos G, Vieira M, Rosa N, Ferraz C (2004) A message-oriented middleware for sensor networks. In: Proceedings of the 2nd workshop on middleware for pervasive and ad-hoc computing (MPAC '04). ACM, New York, pp 127–134

Millimeter Wave Beam Alignment and Adjustment

▸ Millimeter Wave Beam Training and Tracking

Millimeter Wave Beam Training and Tracking

Yongming Huang[1], Jianjun Zhang[4], Chen Zhang[2,3,4], and Ming Xiao[5]
[1]School of Information Science and Engineering, Southeast University, Nanjing, China
[2]Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada
[3]The George Washington University, Washington, DC, USA
[4]Southeast University, Nanjing, China
[5]Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Millimeter Wave Beam Alignment and Adjustment

Definition

Beam training is the search procedure that the transmitter and/or the receiver selects one or multiple beams from its beam training codebook (a set of predefined beams), so as to achieve the necessary link budget for subsequent communications via transmitting or receiving signals with these beams. Beam tracking is the adjustment procedure that the transmitter and/or the receiver adjusts its beams to guarantee the link stability for time-varying propagation environment. Beam training and tracking are usually used in the millimeter wave communications to combat the severe path-loss of signal propagation and the user mobility, respectively.

Historical Background

Millimeter wave (mmwave) communications operating in the band of 30–300 GHz has attracted much attention recently and also been recognized as a promising technology for future mobile networks, owing to its abundant spectrum resources (Xiao et al. 2017). However, the propagation of mmwave signal usually suffers a large path-loss, which severely limits the coverage range of mmwave communications. To address this issue, large antenna arrays are usually equipped in the mmwave communication system, so that the array gain can be exploited to combat the path-loss and increase the communication distance (Heath et al. 2016). Thanks to the short wave length of mmwave signals, it is possible to pack a large number of antennas into a compact size.

Due to the large path-loss of mmwave signals, the necessary link budget usually cannot be satisfied before achieving the array gain provided by the large antenna arrays. Therefore, at the beginning of communications, it is essential for the transmitter and the receiver to determine their respective beams, so as to achieve the array gain and build reliable communications links (Zhang et al. 2017a). To achieve this goal, beam training is designed to select the preferred transmit beam and/or the preferred receive beam from their respective beam training codebook, i.e., a set of predefined beams. Beam training plays a key role in mmwave communications due to its importance in building reliable mmwave communication links, acquiring channel state information (CSI), achieving desired rate of data transmission, and enlarging the cell coverage area.

As a viable approach, beam training was adopted in mmwave multi-gigabit wireless local area network (WLAN) (Tsang et al. 2011) and wireless personal area network (WPAN) (Wang et al. 2009), where analog antenna array architecture and analog beams were considered. Compared with the straightforward method, i.e., to exhaustively train all possible beams in the codebook, then to find the best one or several beams for transmission, a hierarchical-search-based beam training scheme was proposed in Hur et al. (2013) to reduce the training overhead based on multi-resolution or variant beamwidth analog beams. Specifically, all possible wide beams are first trained and the best is selected. Then the beams with narrower beamwidth are trained within the coverage range of this selected best wide beam and the best is chosen. This procedure is repeated until the optimal beam with acceptable beamwidth is found. Later, to improve the accuracy of the hierarchical-search-based training scheme while still maintain its low training overhead, a beam training codebook for the analog architecture was devised in Xiao et al. (2016) by exploiting subarray and deactivation antenna processing techniques. To address the problem of error propagation in hierarchical-search-based training schemes, a novel design of training codebook was proposed in Zhang et al. (2017a) based on the hybrid antenna array architecture, which can provide more flexibility in designing training beams compared to the analog architecture (Alkhateeb et al. 2014). To further reduce the training overhead, a beam training approach was proposed in (Kokshoorn et al. 2017), based on overlapped beam patterns. The beam training approaches aforementioned are mainly designed for single data stream transmission. To enable multiple data stream transmission, a cooperative multi-subarray beam training approach was proposed in Zhang

et al. (2017b) recently for a codebook-based beamforming system.

All the abovementioned beam training schemes correspond to decoupled transmission design (called non-interleaved training) since the complete effective CSI should be obtained first for the full or selected partial beam codebook. When the size of the training beam codebook is comparative to the BS antenna number for satisfactory performance, heavy training overhead is unavoidable. The decoupled nature of non-interleaved schemes thus imposes limitations on the trade-off between training overhead and performance. In Zhang et al. (2018), an interleaved training design was proposed to achieve favorable trade-off between the outage performance and the training overhead where the BS and/or users monitors the training procedure to avoid training redundancy, i.e., to find just enough beams to avoid an outage; thus, the training length depends on the channel realization, and the termination of the training process depends on previous training results, leading to smaller average training overhead.

To achieve a large array gain, the beamwidth of the training beams shall be narrow, and thus the size of the beam training codebook could be large, which leads to a significant beam training overhead. This problem becomes even more serious for the mobile mmwave scenario, since there may be no enough time to perform beam training repeatedly. To address this issue, it is necessary to resort to beam tracking, which is realized by appropriately adjusting the beam directions. Beam tracking has been studied for a long time. The first beam tracking approach in mmwave communication (Wang et al. 2009; Kim and Lee 2015) also finds the candidate beams besides the desired ones in each training and then checks the feasibility of candidate beams in the next training. If some time-varying models are assumed, other beam training approaches can be used for beam tracking as well. Recently, beam tracking solutions based on conventional Kalman filter have drawn considerable attention, by exploiting the temporal correlation of time-varying channels (Jayaprakasam et al. 2017; Zhang et al. 2016; Va et al. 2016).

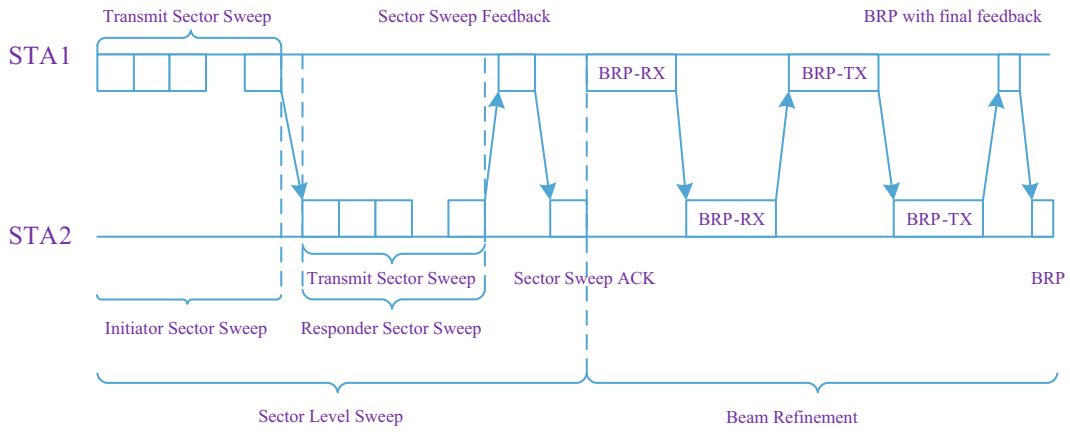
Foundations

Beam training is a viable approach to achieve the goals of building reliable mmwave communication links, enlarging the coverage area and so on in mmwave communications. In fact, beam training has been widely used in practice and adopted by some communication standards, e.g., IEEE 802.11ad and IEEE 802.15.3c. To better illustrate the principle of beam training and tracking, we take the scheme of single-beam training in IEEE 802.11ad as an example.

Beam training, termed as beamforming (BF) training in the IEEE 802.11ad, is a mechanism used by a pair of stations (STAs), e.g., one transmitter and one receiver, to achieve the necessary link budget for subsequent communication. It consists of three phases, i.e., sector-level sweep (SLS), beam refinement protocol (BRP), and beam tracking. An example of BF training procedure is provided in Fig. 1 for clarity. The STA that initiates BF training is referred to as the initiator, and the recipient STA that participates in BF training with the initiator is referred to as the responder. BF training starts with a sector-level sweep (SLS) from the initiator. The purpose of the SLS phase is to enable communications between the two participating STAs. To accelerate the speed of beam training and further reduce the training overhead, in the stage of SLS, the beams with wide beamwidth are considered at first, as shown in Fig. 2a. The SLS procedure of the BF training is presented below in details.

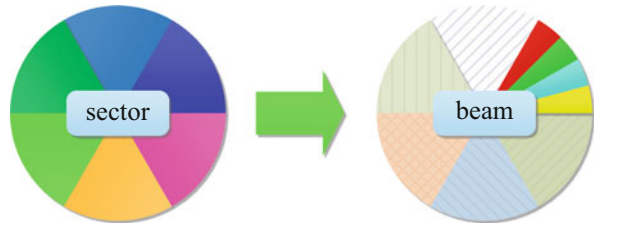
Sub-phase 1: In this sub-phase of SLS, the initiator uses sector-wide beams in turn to sweep its beam space, while the responser receives the transmitted signals using omnidirectional or quasi-omnidirectional beam, as shown in Fig. 3. Based on the received signals, the responder can determine the best sectored transmit beam of the initiator. Moreover, the index of the best sectored transmit beam should be fed back to the initiator.

Sub-phase 2: As shown in Fig. 4, sector-wide beams are used in turn by the responder to sweep its beam space, while omnidirectional or quasi-omnidirectional beam is used by the initiator to receive signals. Based on the received signals,

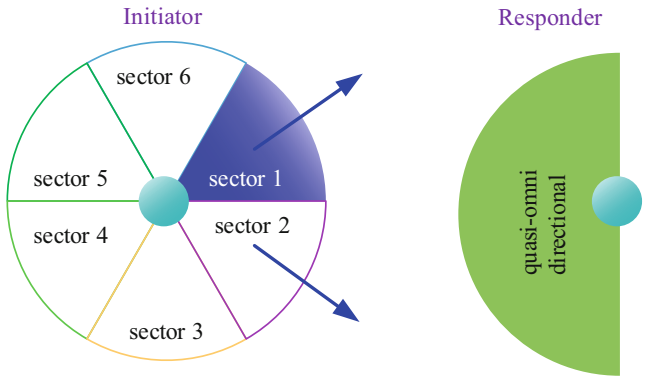


Millimeter Wave Beam Training and Tracking, Fig. 1 An example of beamforming training in IEEE 802.11ad

Millimeter Wave Beam Training and Tracking, Fig. 2 Beam training in IEEE 802.11ad: (**a**) sector-level sweep; (**b**) beam refinement protocol



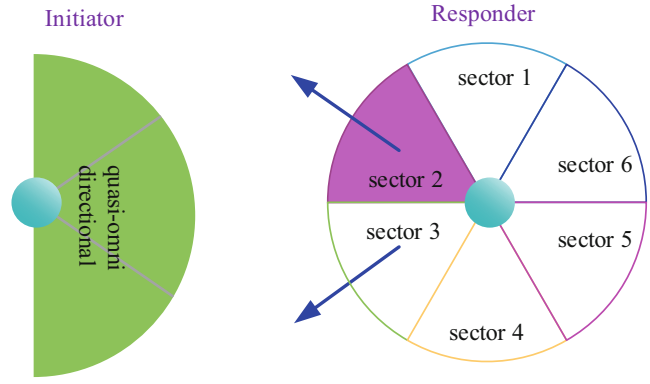
Millimeter Wave Beam Training and Tracking, Fig. 3 Beam training in IEEE 802.11ad – sub-phase 1 of SLS



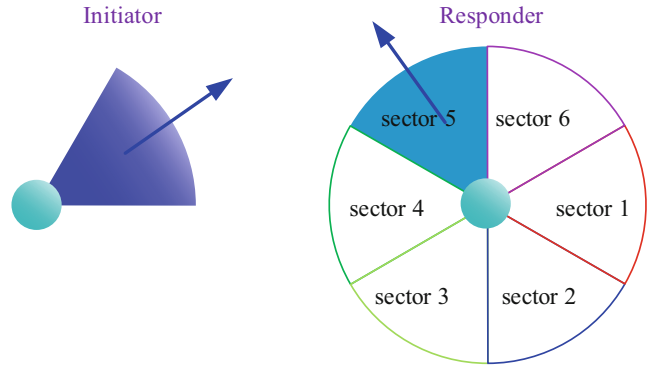
the initiator can obtain its best sector transmit beam. Moreover, the initiator can determine the best sector transmit beam of the responder and should feedback the index to the responder.

Sub-phase 3: The initiator uses the best sector transmit beam obtained in sub-phase 2 to transmit signals, as shown in Fig. 5. The responser uses sector-wide beams in turn to receive signals. Based on the received signals, the responder can obtain its best sector transmit beam and also determine its best sector receive beam.

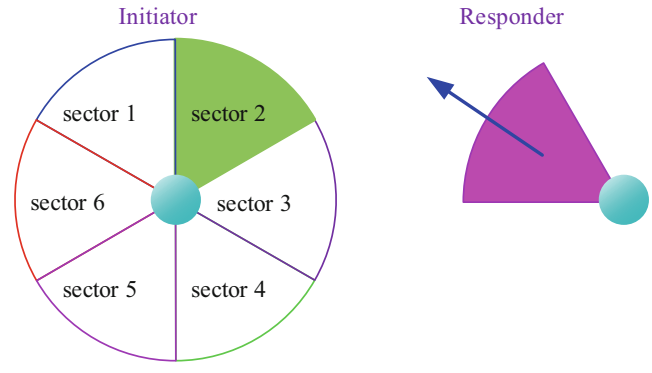
Sub-phase 4: In this sub-phase, the responder uses its best sector transmit beam to transmit signals, while the initiator uses sector-wide beams



Millimeter Wave Beam Training and Tracking, Fig. 4 Beam training in IEEE 802.11ad – sub-phase 2 of SLS



Millimeter Wave Beam Training and Tracking, Fig. 5 Beam training in IEEE 802.11ad – sub-phase 3 of SLS



Millimeter Wave Beam Training and Tracking, Fig. 6 Beam training in IEEE 802.11ad – sub-phase 4 of SLS

to receive signals, as shown in Fig. 6. Based on the received signals, the initiator can determine its best sector receive beam.

When the four sub-phases of SLS aforementioned are finished, both the initiator and the responder can determine their best sector transmit and receive beams. Next, a beam refinement protocol (BRP) may follow, if requested by either the initiator or the responder. The purpose of the BRP phase is to enable iterative refinement of the antenna weight vector of both transmitter and receiver at both participating STAs. In the phase of BRP, the beams with narrow beamwidth are considered, as shown in Fig. 2b. The training procedure in this phase is similar to that in the SLS.

Beam tracking can effectively reduce the training overhead, especially for mmwave mobile

environment. Similarly, a relatively simple beam tracking approach, which is a simplified and improved version of the beam training approach in (Wang et al. 2009), is used to show the basic principle and operational procedure. Due to its quite low complexity and easy implementation, the beam tracking approach has been adopted in current mmwave communication standard, e.g., IEEE 802.11ad. The key idea is to select several candidate beams (or beam pairs) instead of only the optimal beam (or beam pair), during the beam training procedure. For example, when the beam pair with the highest SNR is selected, the beam pairs having the second and third highest SNRs are also retained. When the channel varies, only the candidate beam pairs are tested and switched on to keep the SNR above a certain threshold. The complete beam training procedure is performed, only when all the candidate beam pairs fail. In IEEE 802.11ad, the beam refinement is performed periodically to track beam directions in the third phase.

It should be pointed out that in the literature, there also exist other beam training schemes, which can further improve the performance of beam training. However, most of these beam training schemes are extended from the schemes standardized in IEEE 802.11ad or IEEE 802.15.3c. Moreover, these beam training schemes require carefully designed training codebooks. We would like to recommend readers to refer to Zhang et al. (2017a), Tsang et al. (2011), Wang et al. (2009), Hur et al. (2013), Xiao et al. (2016), Alkhateeb et al. (2014), and Kokshoorn et al. (2017) for details.

The beam training and tracking approaches aforementioned are mainly applied to the case of single data stream. It is very desired to transmit multiple data streams simultaneously in practice. In this case, if conventional bean training schemes are still used, the resulting complexity and training overhead may be prohibitive. To address this issue, the idea of multi-subarray cooperation was introduced in Zhang et al. (2017b), where a codebook-based beamforming mmwave system was considered. It is recommended to refer to Zhang et al. (2017b) for more details.

Key Applications

Millimeter wave wireless local area networks (WLAN), 802.11ad, 802.11aj, 802.15.3c; wireless display; virtual reality or enhanced reality; automatic sync applications (e.g., uploading images from a camera to a PC); millimeter wave cellular networks.

Cross-References

- Millimeter Wave Channel Access
- Millimeter Wave Channel Modeling
- Millimeter Wave Massive MIMO

References

Alkhateeb A, El Ayach O, Leus G, Heath R (2014) Channel estimation and hybrid precoding for millimeter wave cellular systems. IEEE J Sel Top Sign Process 8(5):831–846

Heath RW, Gonza'lez-Prelcic N, Rangan S, Roh W, Sayeed AM (2016) An overview of signal processing techniques for millimeter wave mimo systems. IEEE J Sel Top Sign Process 10(3):436–453

Hur S, Kim T, Love D, Krogmeier J, Thomas T, Ghosh A (2013) Millimeter wave beamforming for wireless backhaul and access in small cell networks. IEEE Trans Commun 61(10):4391–4403

Jayaprakasam S, Ma X, Choi JW, Kim S (2017) Robust beam-tracking for mmwave mobile communications. IEEE Commun Lett 21(12):2654–2657

Kim J, Lee I (2015) 802.11 wlan: history and new enabling mimo techniques for next generation standards. IEEE Commun Mag 53(3):134–140

Kokshoorn M, Chen H, Wang P, Li Y, Vucetic B (2017) Millimeter wave mimo channel estimation using overlapped beam patterns and rate adaptation. IEEE Trans Signal Process 65(3):601–616

Tsang YM, Poon ASY, Addepalli S (2011) Coding the beams: improving beamforming training in mmwave communication system. In: 2011 IEEE global telecommunications conference – GLOBECOM 2011, pp 1–6

Va V, Vikalo H, Heath RW (2016) Beam tracking for mobile millimeter wave communication systems. In: 2016 IEEE global conference on signal and information processing (GlobalSIP), pp 743–747

Wang J, Lan Z, Pyo C-W, Baykas T, Sum C-S, Rahman M, Gao J, Funada R, Kojima F, Harada H, Kato S (2009) Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems. IEEE J Sel Areas Commun 27(8):1390–1399

Xiao Z, He T, Xia P, Xia XG (2016) Hierarchical codebook design for beamforming training in millimeter-wave communication. IEEE Trans Wirel Commun 15(5):3380–3392
Xiao M, Mumtaz S, Huang Y, Dai L, Li Y, Matthaiou M, Karagiannidis GK, Björnson E, Yang K, Chih-Lin I, Ghosh A (2017) Millimeter wave communications for future mobile networks. IEEE J Sel Top Sign Process 35(9):1909–1935
Zhang C, Guo D, Fan P (2016) Tracking angles of departure and arrival in a mobile millimeter wave channel. In: 2016 IEEE international conference on communications (ICC), pp 1–6
Zhang J, Huang Y, Shi Q, Wang J, Yang L (2017a) Codebook design for beam alignment in millimeter wave communication systems. IEEE Trans Commun 65(11):4980–4955
Zhang J, Huang Y, Zhang C, He S, Xiao M, Yang L (2017b) Cooperative multi-subarray beam training in millimeter wave communication systems. In: GLOBECOM 2017–2017 IEEE global communications conference, pp 1–6
Zhang C, Jing Y, Huang Y, Yang L (2018) Interleaved training and training-based transmission design for hybrid massive antenna downlink. IEEE J Sel Top Sign Process 12(2):541–556

Millimeter Wave Channel Access

Chen Zhang[1,2,3], Yongming Huang[4], and Yili Xia[3]
[1]Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada
[2]The George Washington University, Washington, DC, USA
[3]Southeast University, Nanjing, China
[4]School of Information Science and Engineering, Southeast University, Nanjing, China

Synonyms

Millimeter wave multiple access

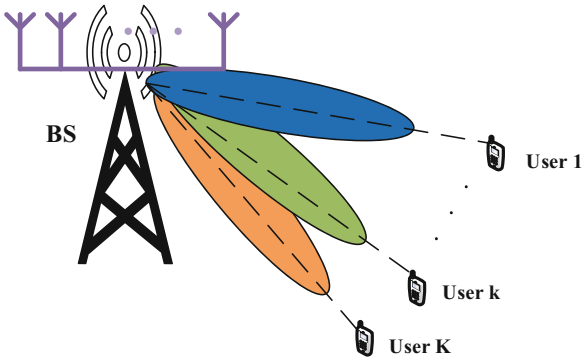
Definition

Millimeter wave (mmWave) channel access is the technology that enables more than one users connected to the same system resource block for concurrent transmission in wireless communications networks operating at mmWave band.

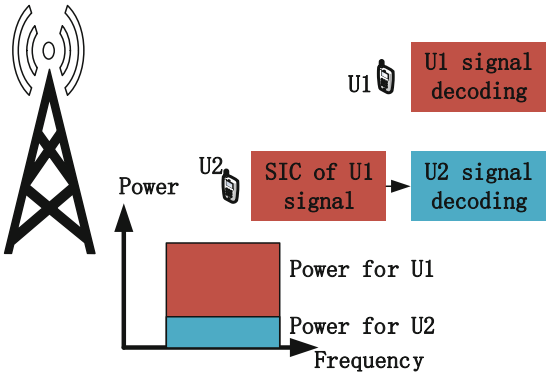
Historical Background

Channel-access technologies are used for supporting multiple-user wireless communication over networks. Several fundamental types of channel access schemes have been widely investigated in Sub-6 GHz microwave wireless communications. For example, in frequency division multiple access (FDMA), time division multiple access (TDMA), and code division multiple access (CDMA), multiple users utilize different frequency bands, time slots, and codes, to establish their own links with the base station (BS), respectively (Goldsmith 2005). These orthogonal multiple access (OMA) technologies have been already utilized in Sub-6 GHz communications systems such as 3G/4G mobile communications systems. They are also applicable to mmWave wireless communication systems. For example, the 802.11ad standard operating at 60 GHz mmWave spectrum adopts the TDMA Scheme (802.11ad 2012).

Compared with these OMA schemes, nonorthogonal multiple access (NOMA) schemes are less used in existing wireless communications standards. As a typical spatial-domain NOMA technology, spatial division multiple access (SDMA) was introduced in 1990s (Roy and Ottersten 1991) (Fig. 1). In SDMA, different beams are used to transmit messages of different users within the same resource block (e.g., time-frequency block), and each beam is optimized to make sure a large signal gain for its dedicated user and small gains for other users. The SDMA scheme has been already supported in the LTE standards since LTE Release 8 (Technical Specifications n.d.), and IEEE 802.11 ac is the first major wireless communications standard to integrate the SDMA scheme (Wireless LAN Medium Access Control 2011). Massive MIMO is the latest technology along this direction (Rusek et al. 2013), in which the BS is equipped with large-scale antenna arrays, e.g., with hundreds of antennas, to serve tens of users via SDMA. In massive MIMO systems operating



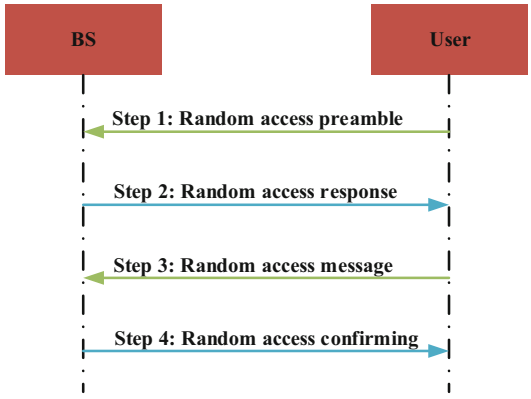
Millimeter Wave Channel Access, Fig. 1 SDMA for K users



Millimeter Wave Channel Access, Fig. 2 Two-user SISO-NOMA systems

at Sub-6 GHz, due to massive antennas at the BS and relatively rich scattering in the propagation, users' antenna-domain channels become less nonorthogonal. This enables high user multiplexing gains, even with simple linear precodings, e.g., maximum ratio transmit (MRT) and zero-forcing (ZF), at the BS (Zhang et al. 2017a).

Another NOMA technology under the spotlight in academic is the power-domain NOMA (usually called NOMA for short), which has been considered as one of the candidates for improving the spectral efficiency and connectivity density in 5G communications (Ding et al. 2014; Dai et al. 2015). The 3rd Generation Partnership Project (3GPP) (a collaboration between groups of telecommunications standards associations) also considers NOMA as a candidate channel-access technology in 5G communications and discussed it at RAN plenary #66 in the name of Multi-User Superposition Transmission (MUST) (https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2912). In conventional single-input-single-output (SISO) NOMA systems with single-antenna BS and users, different signals from different users are superimposed on each other at the BS. Multiple users share the same resource block and detect their signals via successive interference cancellation (SIC). Specifically, as shown in Fig. 2, for a two-user SISO-NOMA system, the BS serves a far-end user U1 and a near-end user U2, both users share the same resource block with a certain transmit power allocation P1 > P2. Under this scheme, user U1 decodes its own message by directly treating the message of user U2 as the background noise, while user U2 first decodes the message of user U1 and then decodes its own message after removing the message of user U1. It has been shown that the spectral efficiency can be enhanced at the cost of an increased receiver complexity compared to conventional OMA systems, and larger gaps among users' channel gains and better



Millimeter Wave Channel Access, Fig. 3 Random access procedure in Sub-6 GHz cellular networks

power allocation policy are both beneficial to the performance of NOMA systems (Zhu et al. 2017; Wang et al. 2017a).

The above-introduced channel-access technologies are mainly used in the data transmission stage. In other communication control stages, e.g., the initialization of user access and cell handover, the random access technology plays an important role. As shown in Fig. 3, four steps in the random access procedure in Sub-6 GHz cellular networks are as follows (Jeong et al. 2015): In step 1, the user transmits one selected preamble signature to the BS by using the resource block indicated in the system information broadcast by the BS. In step 2, if the BS successfully detects the user's preamble, it transmits the random-access response to the user, including the index of the detected preamble sequence, the uplink timing information, and the indication of the resource allocation for the next step. In step 3, the identity of the user is transmitted to the BS. Finally, in step 4, the identity of the user is transmitted back to the user from the BS to confirm that the random access procedure is successfully completed for the user.

Foundations

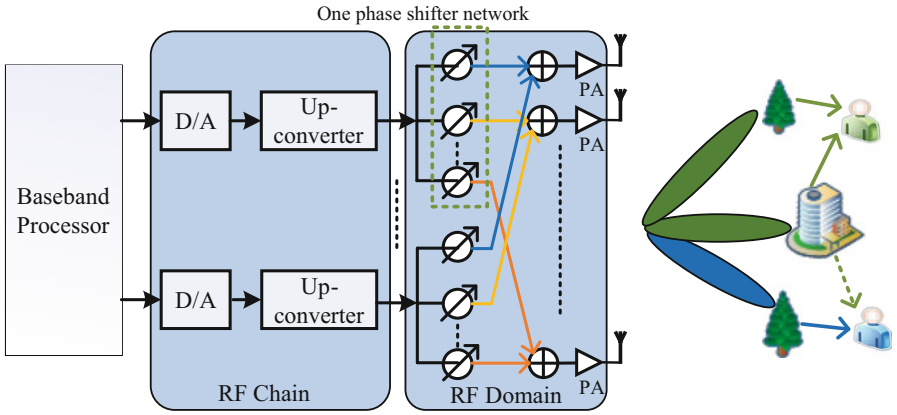
Since shorter wavelengths at the mmWave band make it easier to equip massive antennas in the same physical space, massive MIMO becomes the representative technology to build up mmWave communication systems. Due to the large spatial degrees of freedom in massive MIMO, the spatial resolution-based channel access technologies are of great importance. In what follows, the new features of channel and system structures in mmWave massive MIMO systems which affect the design of channel access will be first introduced. Then, the application of SDMA and its combination with NOMA and random access will be subsequently discussed for mmWave massive MIMO systems.

Features of mmWave Channel and System Structures

Compared to the Sub-6 GHz systems, in mmWave massive MIMO systems, channels become spatially sparse and highly directional due to limited scattering. This, however, results two adverse effects. On one hand, different users' antenna-domain channels become less orthogonal due to the reduced effective channel dimensions, especially for users with similar locations. On the other hand, channel powers are distributed in fewer scatterers, which indicates that using partial channel scatterers, e.g., via directional beamforming, for transmission results in smaller performance losses. In addition, the mmWave channel is typically with a large path loss, and this undesirable physic demands a large array gain. Moreover, due to their hardware costs and power consumptions, as shown in Fig. 4, one typical structure of mmWave massive MIMO systems is the hybrid analog/digital one with limited RF chains and phase shifter networks. Under this structure, fewer degrees of freedom are available for beamforming design, especially for the practical low-resolution phase shifters, leading to a coarser beamforming capability.

SDMA in mmWave Systems

Under the hybrid beamforming structure, its coarser beamforming capability results in a weaker capability of user interference suppression. Meanwhile, due to the channel nonorthogonality in antenna-domain, popularly used linearly precoding methods in Sub-6 GHz massive MIMO systems may have nonnegli-



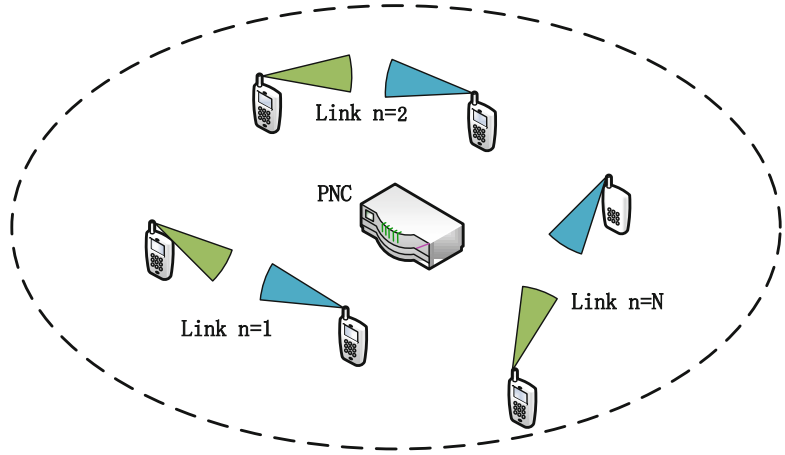
Millimeter Wave Channel Access, Fig. 4 MmWave massive MIMO system with hybrid beamforming and limited scattering channels. D/A is digital to analog converter. Up-converter moves the frequency of the input signal to higher frequency. One phase shifter network includes the same number of phase shifters as the number of antennas. PA is the power amplifier

gible performance degradation. Furthermore, another factor making designs of SDMA more challenge is the incompleteness of channel state information (CSI) at the BS. Specifically, due to the possible low prebeamforming signal-to-noise-ratio (SNR), massive antennas, and limited RF chains, the mmWave massive system tends to possess only partial or long-term CSI at the BS. Practically, the beamforming design should be based on long-term channel statistics (channel direction and spatial profile of average channel power) and the beamspace CSI, i.e., instantaneous effective CSI for limited analog beams (each analog beam corresponds to one phase setting of one phase shifter network). Therefore, the mmWave channel characteristics, the less flexible hybrid structure, and the CSI incompleteness at the BS independently or jointly make the transmit precoding/beamforming and receive combining (when users also have multiple antennas) more challenging (Sun et al. 2016).

Some effective hybrid beamforming algorithms have been proposed to guarantee a desirable SDMA performance. For example, in Alkhateeb et al. (2015), a decoupled baseband/analog design was proposed where by ignoring the user interference, the BS first finds the best beam for each user via beam training (Zhang et al. n.d.) and then conducts the baseband digital beamforming to reduce the residual user interference based on beamspace CSI. In Li et al. (2017), the capacity of beamspace channel is used as the optimization object for beamforming design, while the analog beamforming is achieved based on channel statistics only. In He et al. (2017a), the codebook (with finite analog beams)-based hybrid beamforming optimization is simplified to a joint codeword selection and beamforming design problem where both the baseband and analog designs are achieved based on complete beamspace CSI.

Spatial Sharing Multiple Access

The spatial sharing-based multiple access has been adopted into the wireless local area networks (WLAN) standard 802.11 ad (Hany and Widmer 2017). Similarly, for wireless personal area network (WPAN) standard 802.15.3c operating at mmWave bands (Baykas et al. 2011), many researchers are also interested in applying this spatial sharing idea to increase system throughput. Specifically, via utilizing the high propagation loss and antenna directionality, multiple links (each link is made up of one transmitter and one receiver) are scheduled to transmit concurrently in the same time slot. This combination of spatial sharing-/reusing-based multiple access with the original TDMA is named the spatial time division multiple access STDMA (Qiao et al. 2012) (Figs. 5).



Millimeter Wave Channel Access, Fig. 5 STDMA in WPAN. The piconet coordinator (PNC) provides the basic timing for the piconet with the beacon and manages QoS requirements, power save modes, and access control to the piconet

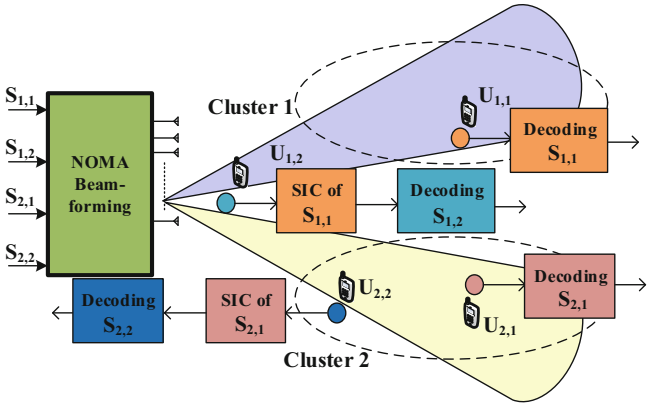
User Scheduling

The performance of SDMA is highly dependent on channel characteristics. For example, in typical mmWave network situations where users are in similar locations and near line-of-sight (LoS) channels exist, SDMA can only avoid large user interference via extremely fine beamforming. These mmWave channel characteristics make the user scheduling necessary, i.e., it is more practical to serve some users via SDMA and other users in orthogonal system resources. Practically, due to the aforementioned CSI incompleteness at the BS, the user scheduling should be based on the long-term channel statistics (Zhang et al. 2017b) and/or partial beamspace CSI (He et al. 2017b). Therefore, the optimal user scheduling for mmWave hybrid massive MIMO systems involves the joint design of user selection, analog beam assignment/design, and baseband digital beamforming, which is more complicated than its counterpart of Sub-6 GHz systems. In He et al. (2017b), the limited beamspace CSI-based joint analog beam selection and user scheduling task is formulated as a nonconvex and combinatorial optimization problem, and two codebook-based low-complexity methods are proposed to address this problem.

For the STDMA transmission, the optimal concurrent transmission scheduling problem can be converted to a Knapsack problem which is NP-complete (Pisinger 2005). In Sum et al. (2009), multiple communication links are scheduled within the same time slot if the accumulated interference in this slot is below a specific threshold. In Qiao et al. (2012), the scheduling design is optimized to maximize the number of concurrent flows such that the quality of service (QoS) requirement of each flow is satisfied. This optimization problem is decomposed and solved via a heuristic scheduling algorithm.

MIMO-NOMA in mmWave Systems

With multiple antennas at the BS in mmWave systems, the combination of NOMA and MIMO becomes a natural choice. It has been shown that (1) NOMA can achieve a higher channel capacity than OMA in both mmWave uplink and downlink systems (Naqvi and Hassan 2016); and (2) Enormous capacity improvement can be achieved by mmWave MIMO-NOMA, compared to existing LTE systems (Zhang et al. 2017c). In mmWave MIMO-NOMA, more than one users can be simultaneously supported in each beam via superposition coding and SIC operations. This is different from the conventional SDMA, where the signal for one user is only transmitted in each beam (Ding et al. 2017; Wang et al. 2017b). On the other hand, the small amount of RF chains in mmWave hybrid massive MIMO systems further limits the number of available beams and supported SDMA users. Consequently, much more users can be served within the same system resource block for any given RF chains via mmWave MIMO-NOMA,



Millimeter Wave Channel Access, Fig. 6 MIMO-NOMA system with two clusters: Each cluster includes two users covered by the same beam

and the system achievable sum-rate can be also improved Fig. 6.

The main challenges in mmWave MIMO-NOMA are given as follows (Xiao et al. n.d.; Huang et al. n.d.). Similar to SISO-NOMA, SIC in mmWave MIMO-NOMA is implemented at the user side to suppress interference. While power allocation among users is the main focus in SISO-NOMA, it is possible to eliminate the user interference via beamforming and SIC in both the power and spatial domains in mmWave MIMO-NOMA, resulting in a more complicated design problem. For example, in SISO-NOMA, the SIC order generally depends directly on channel gains, e.g., a user with a smaller channel gain has a higher SIC order (i.e., decoded first). However, in mmWave MIMO-NOMA, effective channel gains are affected by beamforming designs. This results in a coupled beamforming and SIC ordering optimization. Moreover, some features in mmWave massive MIMO systems, e.g., massive antenna arrays, channels with limited scattering, and hybrid beamforming structures, make existing designs for the conventional MIMO-NOMA less applicable.

Due to the coarse hybrid beamforming capability and limited RF chains in mmWave systems, it is hard to serve many users only based on spatial resolution, especially for users with near directions. Introducing NOMA into mmWave MIMO can provide possibility of simultaneously serving these users. This advantage is of significant importance for those scenarios where both high data rate and massive connectivity, e.g., dense urban scenarios, are highly desired. Furthermore, the increased path-loss component (Naqvi and Hassan 2016) and channel blocking effect in mmWave communications are both beneficial for NOMA performance since they result in a larger channel gain difference for users covered by one beam with near locations. Considering the abovementioned two harmonies between the mmWave systems and the principle of NOMA, the use of NOMA for mmWave communications is a promising research direction.

Random Access in mmWave Systems

The basic procedure of random access in mmWave systems is similar to that of the traditional Sub-6 GHz cellular network, as shown in Fig. 3. However, in an mmWave cellular network, highly directional beamforming should be used at both the BS and users to compensate the large channel path loss. Since random access cannot benefit from the full beamforming gain due to the lack of information on the best transmit-receive beam pair (i.e., the location of strongest channel scatterer), the design of the random access becomes more challenging in mmWave communication systems (Jeong et al. 2015; Polese et al. 2017). It has been shown that the exhaustive directional beam pair search is able to achieve a better preamble detection than the omnidirectional beam at the same cost of time overhead (Jeong et al. 2015). Furthermore, reducing the time overhead of random access can be achieved by, e.g., (1) using enhanced preamble

detection method; (2) steering multiple receive beams in the same direction simultaneously to increase power of the received signals via noncoherently accumulation; (3) exploiting beam reciprocity to utilize the downlink beam information for the uplink; and (4) having better cell deployments for stronger channels (Jeong et al. 2015). Some other approaches directly aim to reduce the time overhead of the naive exhaustive search in random access for mmWave systems. For example, the angular space can be scanned in time-varying random directions using synchronization signals periodically transmitted by the BSs (Barati et al. 2015), the two-stage hierarchical procedure can be employed to speed up user discovery (Desai et al. 2014), and the context information on users and/or BS positions provided by a separate control plane can be also used to accelerate the random access procedure (Shokri-Ghadikolaei et al. 2015).

Key Applications

TDMA has been adopted in 802.11ad WLAN standard and 802.15.3c WPAN standard; spatial sharing multiple access has been used in 802.11ad WLAN standard; SDMA (multiple-user MIMO) has been added into 802.11aj WLAN standard (Complete Proposal for IEEE 802.11aj (45 GHz) n.d.; Hong et al. 2018) and is likely to be used in 802.11ay WLAN standard (an improved version of 802.11ad) (http://www.ieee802.org/11/Reports/tgay_update.htm).

Cross-References

- Millimeter Wave Beam Training and Tracking
- Millimeter Wave Channel Modeling
- Millimeter Wave Massive MIMO

References

802.11ad 2012 – IEEE Standard for Information technology – Telecommunications and information exchange between systems–Local and metropolitan area networks–Specific requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band. [Online] Available: http://ieeexplore.ieee.org/document/6178212/

Alkhateeb A, Leus G, Heath RW (2015) Limited feedback hybrid precoding for multi-user millimeter wave systems. IEEE Trans Wirel Commun 14(11):6481–6494

Barati CN et al (2015) Directional cell discovery in millimeter wave cellular networks. IEEE Trans Wirel Commun 14(12):6664–6678

Baykas T et al (2011) IEEE 802.15. 3c: the first IEEE wireless standard for data rates over 1 Gb/s. IEEE Commun Mag 49(7):114–121

Complete Proposal for IEEE 802.11aj (45 GHz). [Online] Available: https://mentor.ieee.org/802.11/documents?is_dcn=0707

Dai L, Wang B, Yuan Y, Han S, L IC, Wang Z (2015) Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. IEEE Commun Mag 53(9):74–81

Desai V, Krzymien L, Sartori P, Xiao W, Soong A, Alkhateeb A (2014) Initial beamforming for mmWave communications. In: Proceedings of 48th Asilomar conference signals, system and computers. IEEE, pp 1926–1930

Ding Z, Yang Z, Fan P, Poor HV (2014) On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users. IEEE Signal Process Lett 21(12):1501–1505

Ding Z, Fan P, Poor HV (2017) Random beamforming in millimeter-wave NOMA networks. IEEE Access 5:7667–7681

Goldsmith A (2005) Wireless communications. Cambridge University Press, New York

Hany A, Widmer J (2017) Extending the IEEE 802.11 ad model: scheduled access, spatial reuse, clustering, and relaying In: Proceedings of workshop on ns-3 (2017 WNS3), Porto, 13–14 June 2017, 8 pp. https://doi.org/10.1145/3067665.3067667

He S, Wang J, Huang Y, Ottersten B, Hong W (2017a) Codebook based hybrid precoding for millimeter wave multiuser systems. IEEE Trans Signal Process 65(20):5289–5304

He S, Wu Y et al (2017b) Joint optimization of analog beam and user scheduling for millimeter wave communications. IEEE Commun Lett 21(12):2638–2641

Hong W, He S, Wang H, Yang G, Huang Y, Chen J, Yang L (2018) An overview of China millimeter-wave multiple gigabit wireless local area network system. IEICE Trans Commun 101(2):262–276

Huang Y, Zhang C, Wang J, Jing Y, Yang L, You X Signal processing for MIMO-NOMA: present and future challenges. IEEE Wireless Commun Mag, 25(2): 32–38

Jeong C, Park J, Yu H (2015) Random access in millimeter-wave beamforming cellular networks: issues and approaches. IEEE Commun Mag 53(1): 180–185

Li Z, Han S, Molisch AF (2017) Optimizing channel-statistics-based analog beamforming for millimeter-wave multi-user massive MIMO downlink. IEEE Trans Wirel Commun 16(7):4288–4303
Naqvi SAR, Hassan SA (2016) Combining NOMA and mmWave technology for cellular communication. In: Proceedings of the IEEE vehicular technology conference (VTC Fall), Montreal. IEEE
Pisinger D (2005) Where are the hard knapsack problems? Comput Oper Res 32:2271–2284
Polese M, Giordani M, Mezzavilla M, Rangan S, Zorzi M (2017) Improved handover through dual connectivity in 5G mmWave mobile networks. IEEE J Sel Areas Commun 35(9):2069–2084
Qiao J, Cai LX, Shen X, Mark JW (2012) STDMA-based scheduling algorithm for concurrent transmissions in directional millimeter wave networks. In: 2012 IEEE International Conference on Communications (ICC), Ottawa. IEEE
Roy RH III Ottersten B Spatial division multiple access wireless communication systems, U.S. Patent 5515378, 7 May 1991
Rusek F et al (2013) Scaling up MIMO: opportunities and challenges with very large arrays. IEEE Signal Process Mag 30(1):40–60
Shokri-Ghadikolaei H, Fischione C, Fodor G, Popovski P, Zorzi M (2015) Millimeter wave cellular networks: a MAC layer perspective. IEEE Trans Commun 63(10):3437–3458
Sum C, Lan Z, Funada R, Wang J, Baykas T, Rahman MA, Harada H (2009) Virtual time-slot allocation scheme for throughput enhancement in a millimeter-wave multi-Gbps WPAN system. IEEE J Sel Areas Commun 27(8):1379–1389
Sun S, Rappaport TS, Heath RW, Nix A, Rangan S (2016) MIMO for millimeter-wave wireless communications: beamforming, spatial multiplexing, or both? IEEE Commun Mag 52(12):110–121
Technical Specifications and Technical Reports for a UTRAN-based 3GPP system, 3GPP TR 21.101. https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=544
Wang J, Peng Q, Huang Y, Wang H, You X (2017a) Convexity of weighted sum rate maximization in NOMA systems. IEEE Signal Process Lett 24(9):1323–1327
Wang B, Dai L, Wang Z, Ge N, Zhou S (2017b) Spectrum and energy efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array. IEEE J Sel Areas Commun 35(10):2370–2382
Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: enhancements for very high throughput for operation in bands below 6GHz, standard IEEE P802.11ac, Draft 0.1, IEEE Computer Society, Jan 2011
Xiao Z, Dai L, Xia P, Choi J, Xia X Millimeter-wave communication with non-orthogonal multiple access for 5G, IEEE Wireless Commun Mag, to be published. [Online] Available: https://arxiv.org/abs/1709.07980
Zhang C, Jing Y, Huang Y, Yang L (2017a) Performance scaling law for multi-cell multi-user massive MIMO. IEEE Trans Veh Technol 66(11):9890–9903
Zhang C, Huang Y, Jing Y, Jin S, Yang L (2017b) Sum-rate analysis for massive MIMO downlink with joint statistical beamforming and user scheduling. IEEE Trans Wirel Commun 16:2181–2194
Zhang D, Zhou Z, Xu C, Zhang Y, Rodriguez J, Sato T (2017c) Capacity analysis of non-orthogonal multiple access with mmwave massive MIMO systems. IEEE J Sel Areas Commun 35(7):1606–1618
Zhang C, Jing Y, Huang Y, Yang L (2018) Interleaved training and training-based transmission design for hybrid massive antenna downlink. IEEE J Select Topics Signal Process https://doi.org/10.1109/JSTSP.2018.2818648
Zhu J, Wang J, Huang Y, He S, You X, Yang L (2017) On optimal power allocation for downlink non-orthogonal multiple access systems. IEEE J Sel Areas Commun 35(12):2744–2757

Millimeter Wave Channel Measure

Sherif Adeshina Busari[1], Shahid Mumtaz[1], Kazi Mohammed Saidul Huq[1], and Jonathan Rodriguez[1,2]
[1]Instituto de Telecomunicações, Aveiro, Portugal
[2]University of South Wales, Pontypridd, UK

Synonyms

mmWave channel campaigns; mmWave channel measurement; mmWave channel sounding

Definitions

Channel measurement refers to the techniques widely used in telecommunications for studying the properties of wireless channels. It is done by carrying out real-world measurements using a set of specialized equipment known as channel sounders (transmitter(s) and receiver(s)). Typically, the field measurements are undertaken at different times, cities, environments, and under

different scenarios, resulting in what is referred to as channel measurement campaigns. The results from the measurements are used to develop channel models for the characterization and performance evaluation of wireless systems and networks. Millimeter wave (mmWave) channel measurements thus refer to measurements carried out at mmWave frequencies (30–300 GHz) (Busari et al. 2017).

Key Points

The mmWave frequency bands fall in the extremely high frequency (EHF) range of the electromagnetic spectrum. The bands have frequencies (f) spanning [30, 300] GHz, which corresponds to wavelengths (λ) of [10, 1] mm, respectively, as $f = c/\lambda$ and $c = 3 \times 10^8$ m/s is the speed of light (Busari et al. 2017). While this is the general range for mmWave, some authors in the literature limit the mmWave frequencies to the 30–90/100 GHz bands, preferring to refer to the 90/100–300 GHz frequencies as part of the terahertz (THz) bands (Mumtaz et al. 2017a). In addition, frequencies above 6 GHz but lower than 30 GHz, such as frequencies between 10 and 28 GHz, are also referred to as mmWave (Xiao et al. 2017). In each case, the bands have wavelengths up to a few millimeters, hence the name mmWave.

Relative to the sub-6 GHz microwave (μWave) frequencies used for many common applications, the mmWave bands offer several advantages. The larger bandwidths translate to higher capacity, the smaller wavelengths enable more (i.e., massive) antennas in same physical space, and the relatively closer spectral allocations lead to a more homogeneous propagation (Busari et al. 2017).

On the other hand, mmWave signals are prone to higher path loss, higher penetration loss, severe atmospheric absorption, higher susceptibility to shadowing and diffraction, more attenuation due to rain, and higher vulnerability to blockages by objects, as their wavelengths are typically less than the physical dimensions of the obstacles (Busari et al. 2017; Rappaport et al. 2013).

To counter the effects of the propagation losses, reduce interference, increase system throughput, and maximize network's energy efficiency, directional communication is mandatory for practical mmWave systems. Thus, beamforming can be employed to increase the antenna array gains, suppress interference, and improve the signal to noise ratio (SNR) of the links (Huang et al. 2018; Cui et al. 2018).

In addition, relay stations can be used to circumvent obstacles, thereby avoiding blockages and minimizing outages (Yang et al. 2018). More so, for the 50–200 m cell size envisaged for mmWave small cells (and other mmWave short-range applications), the expected attenuation (about 7 dB/km), due to heavy rainfalls, has minimal effect within such distances. Also, the high mmWave path loss limits inter-cell interference (ICI) and allows for more frequency reuse, which by extension improves the overall system capacity (Rappaport et al. 2013).

Historical Background

The study of mmWave predates the twentieth century. In the 1890s, scientists performed experiments at mmWave frequencies. However, mmWave channel measurements and system designs targeted at applications for radio communication started in the 1990s and early 2000s (Xiao et al. 2017). The spectrum crunch at μWave and the projected explosive traffic demands with the coming of fifth-generation (5G) networks and the Internet of Things (IoT)/Internet of Everything (IoE) have again motivated rigorous research and development (R&D) for mmWave systems and networks (Busari et al. 2017).

Foundations

MmWave channel measurement has received tremendous attention in recent years. Researchers in the academia and the industry have carried

out extensive mmWave measurement campaigns in different cities, countries, and continents of the world. A summary of mmWave channel measurements in the major cities of Europe, Asia, and the United States of America between 2012 and 2017 can be found in Xiao et al. (2017).

Since 2011, the New York University (NYU) WIRELESS researchers have been conducting extensive measurements at 28, 38, 60, and 73 GHz (Rappaport et al. 2013, 2015). The EU mmMAGIC project team also carried out mmWave measurements for mmWave frequencies 10–100 GHz (Jaeckel et al. 2014). Others include Zhao et al. at 32 GHz Zhao et al. (2017), Ai et al. at 26 GHz [8], METIS at 50–70 GHz (Ai et al. 2017). and the 5GCM at 6–100 GHz (Ghosh 2015), among others.

While most existing mmWave channel measurements were done for static radio channels, the world's first demonstration for the dynamic channel was carried out by Samsung at 28 GHz in 2014, reaching 1.2 Gbps transmission rate at 110 kmph. Contrary to earlier thoughts, this dynamic channel measurement points to the suitability of mmWave frequencies for mobile applications. It is also particularly useful to buttress the viability of mmWave small cells which typically do not involve high-velocity users (Mumtaz et al. 2017b).

For good mmWave channel measurements, the sounding equipment (both hardware and software) must have high performance and reliability. This is essential for the accuracy of the measurements and the corresponding models derived from them. High-frequency (mmWave) bands place high demands on the sounder components, as they must ensure wide frequency coverage, high dynamic range, and high output signal (in terms of power, stability, and quality) as well as guarantee the least possible distortion and harmonic content (Busari et al. 2017; Mumtaz et al. 2017b).

Efficient channel sounders must therefore be capable of measurements with operating frequencies up to several tens/hundreds of GHz and several GHz of mmWave bandwidth as well as multiple antennas at the transceivers. Similarly, sounding equipment should be capable of measurements in static and dynamic conditions and indoor and outdoor environments. They should as well have capabilities for efficient signal generation, data acquisition, and storage. The equipment should also guarantee accurate synchronization, calibration, sensitivity, resolution, and flexibility for extension while maintaining reasonable system cost (Busari et al. 2017; Mumtaz et al. 2017b).

The extensive mmWave channel measurements have revealed marked differences in the propagation characteristics of mmWave relative to μWave. MmWave signals exhibit line-of-sight (LOS) or near-LOS propagation with absolute increase in path loss with increasing frequency, specular reflection attenuation, diffuse scattering, very high diffraction attenuation, and frequency dispersion effect, which, with the prospect of the huge bandwidth, allows the propagation to be considered as frequency-dependent. The mmWave signals also exhibit quasi-optical properties as the short wavelengths lead to weak diffractions. With highly directional mmWave communications, delay spreads of less than 20 ns are typical, and the number of scattering clusters becomes much less than at μWave (Rappaport et al. 2015; Xiao et al. 2017; Busari et al. 2017).

Under static conditions, mmWave channels exhibit frequency-selective fading which would need to be addressed by modulation or equalization. At mmWave frequencies, the assumption of asymptotic pair-wise orthogonality between channel vectors under independent and identically distributed (i.i.d) Rayleigh fading channel which is valid for μWave bands no longer holds. The number of independent multipath components (MPCs) becomes limited, and so channel vectors exhibit correlated fading. The channels are non-orthogonal and the waves are spherical. Spatial non-stationarity also becomes severe. Extensive indoor and outdoor channel measurement campaigns and simulations have been carried out by Rappaport et al. (2013), Rappaport et al. (2015), Jaeckel et al. (2014), Zhao et al. (2017), Ai et al. (2017), METIS (2015), and Ghosh (2015), among others, to deduce these outcomes.

Following the mmWave channel measurement activities, channel parameters (such as path loss, penetration loss, power delay profile (PDP), delay spread, shadowing, angular spreads, coherence bandwidth, etc.) are estimated from the obtained data or field results to develop channel models. Considering all necessary factors (such as frequency, propagation environment, scenario, etc.), the developed channel models provide statistical, mathematical, and analytical frameworks for simulation studies and performance evaluation of the communication systems. Such frameworks are also used to compare and/or validate empirical data from field deployment and operation tests (Mumtaz et al. 2017b).

Some of the general mmWave channel models that have resulted from mmWave measurement efforts include MiWEBA, COST 2100, QuaDRiGa, mmMAGIC, METIS, NYUSIM, 3GPP TR 38.900, 5GCMSIG, IMT-2020, and MG5GCM channel models, among others (3GPP 2016).

Key Applications

Due to the attractive capabilities and high commercial potentials of mmWave communications, spectrum allocation, regulatory, and standardization activities are being undertaken and rigorously pursued by relevant bodies such as the ITU, IEEE, ETSI, and the 3GPP. In fact, the 60 GHz mmWave WiFi (IEEE 802.11ad) that supports transmission rates up to 4–7 Gbps has already been standardized and has hit the market. It is expected to be followed by the 60 GHz mmWave IEEE 802.11ay capable of higher data rates up to 20–40 Gbps (Xiao et al. 2017).

5G mmWave cellular networks are also anticipated to be deployed by year 2020 to support many high data rate applications such as short-range communications, vehicular networks, and wireless in-band fronthauling/backhauling, among others. As a result, measurement campaigns, channel modeling, demos, prototyping, and field tests for mmWave technologies and applications have intensified. Standardization, holistic deployment, performance evaluation, and business models will follow these efforts (Busari et al. 2017).

Future Directions

The mmWave channel is intrinsically an ultra-broadband channel with huge bandwidth and high spatial multiplexing capabilities to significantly enhance wireless access and improve cell and user throughputs. The mmWave spectrum has abundant available bandwidth (up to 270 GHz). With a reasonable assumption of 40% availability over time, the bands will possibly open up about 100 GHz new spectrum for different applications and services (such as cellular and fixed wireless communications) (Busari et al. 2017; Liu et al. 2018).

Since there is spectrum shortage at the congested sub-6 GHz μWave bands, relative to the explosive data rate demands projected for 2020 and beyond, mmWave wireless offers promising and exciting potentials to support next-generation systems. As a result, it features in the list of key enablers for the multi-gigabit-per-second (Gbps) wireless access for the fifth-generation (5G) and beyond-5G (B5G) mobile networks, as well as for high-rate wireless personal and local area networks (WPAN and WLAN) (Yang et al. 2018).

Though there are many challenges to address, the mmWave technology shows amazing prospects and potentials that will undoubtedly usher in a new era for services and applications not hitherto possible. Certainly, mmWave channel measurement is a critical component of the exciting journey.

Cross-References

- ▸ Resource Allocation in SDN/NFV-Enabled 5G Networks
- ▸ Millimeter-wave Communications

References

3GPP (2016) Study on channel model for frequency spectrum above 6 GHz (3GPPP TR38.900), v14.2.0. Technical report, 3GPP. http://www.3gpp.org/ftp//Specs/archive/38_series/38.900/

Ai B, Guan K, He R, Li J, Li G, He D, Zhong Z, Huq KMS (2017) On indoor millimeter wave massive MIMO channels: measurement and simulation. IEEE J Sel Areas Commun 35(7):1678–1690. https://doi.org/10.1109/JSAC.2017.2698780

Busari SA, Huq KMS, Mumtaz S, Dai L, Rodriguez J (2017) Millimeter-Wave Massive MIMO Communication for Future Wireless Systems: A Survey. IEEE Communications Surveys & Tutorials, 20(2):836–869, Secondquarter 2018. https://doi.org/10.1109/COMST.2017.2787460

Cui Y, Fang X, Fang Y, Xiao M (2018) Optimal Nonuniform Steady mmWave Beamforming for High-Speed Railway. IEEE Transactions on Vehicular Technology, 67(5):4350–4358. https://doi.org/10.1109/TVT.2018.2796621

Ghosh A (2015) 5G channel model for bands up to 100 GHz. In: IEEE Globecom 2015. IEEE Globecom, San Diego. http://www.5gworkshops.com/5GCM.html

Huang Y, Zhang J, Xiao M (2018) Constant Envelope Hybrid Precoding for Directional Millimeter-Wave Communications. IEEE Journal on Selected Areas in Communications. https://doi.org/10.1109/JSAC.2018.2825820

Jaeckel S, Raschkowski L, Borner K, Thiele L (2014) QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials. IEEE Trans Antennas Propag 62(6):3242–3256. https://doi.org/10.1109/TAP.2014.2310220

Liu Y, Fang X, Xiao M, Mumtaz S (2018) Decentralized beam pair selection in multi-beam millimeter-wave networks. IEEE Trans Commun PP(99):1–1. https://doi.org/10.1109/TCOMM.2018.2800756

METIS (2015) Channel models deliverable 1.4 version 3 document ICT-317669 METIS/D14. Technical report, METIS

Mumtaz S, Jornet JM, Aulin J, Gerstacker WH, Dong X, Ai B (2017a) Terahertz communication for vehicular networks. IEEE Trans Veh Technol 66(7):5617–5625

Mumtaz S, Rodriguez J, Dai L (eds) (2017b) mmWave massive MIMO: a paradigm for 5G, 1st edn. Academic Press, London/San Diego

Rappaport TS, Sun S, Mayzus R, Zhao H, Azar Y, Wang K, Wong GN, Schulz JK, Samimi M, Gutierrez F (2013) Millimeter wave mobile communications for 5G cellular: it will work! IEEE Access 1:335–349. https://doi.org/10.1109/ACCESS.2013.2260813

Rappaport TS, MacCartney GR, Samimi MK, Sun S (2015) Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design. IEEE Trans Commun 63(9):3029–3056. https://doi.org/10.1109/TCOMM.2015.2434384

Xiao M, Mumtaz S, Huang Y, Dai L, Li Y, Matthaiou M, Karagiannidis GK, Bjornson E, Yang K, Chih-Lin I, Ghosh A (2017) Millimeter wave communications for future mobile networks. IEEE J Sel Areas Commun 35(9):1909–1935. https://doi.org/10.1109/JSAC.2017.2719924

Yang G, Xiao M, Al-Zubaidy H, Huang Y, Gross J (2018) Analysis of millimeter-wave multi-hop networks with full-duplex buffered relays. IEEE/ACM Trans Netw 26(1):576–590. https://doi.org/10.1109/TNET.2017.2786341

Yang G, Xiao M, Alam M, Huang Y (2018) Low-latency heterogeneous networks with millimeter-wave communications. CoRR. arXiv preprint arXiv:180109286

Zhao X, Li S, Wang Q, Wang M, Sun S, Hong W (2017) Channel measurements, modeling, simulation and validation at 32 GHz in outdoor microcells for 5G radio systems. IEEE Access 5:1062–1072. https://doi.org/10.1109/ACCESS.2017.2650261

Millimeter Wave Channel Modeling

Wei Huang[1], Jie Huang[2], Yongming Huang[3], and Cheng-Xiang Wang[1]

[1]Southeast University, Nanjing, China

[2]Shandong University, Qingdao, China

[3]School of Information Science and Engineering, Southeast University, Nanjing, China

M

Synonyms

Millimeter wave channel modeling

Definition

Millimeter wave channel modeling is to use various deterministic or stochastic channel modeling approaches to describe the channel propagation characteristics at millimeter wave frequency bands and verify them by channel measurements.

Historical Background

Channel modeling is important for wireless communication system design, test, and performance

evaluation. It has been the foundation of wireless communications for years. As wireless communication systems evolve from the first generation (1G) to the forthcoming fifth generation (5G), many channel models have been proposed, among which the series of COST-, WINNER-, and 3GPP-family models are the most widely investigated ones. The description of wireless channels has also evolved from time domain to time, delay, and spatial domains. Another obvious distinction between different generations of wireless communication systems lies in the carrier frequency.

Due to the congestion of sub-6 GHz bands, millimeter wave (mmWave) technology has been proposed as a key technology for 5G to enable high data rate transmissions (Wang et al. 2014). In general, it denotes the frequency range of 30–300 GHz, but sometimes 10–30 GHz bands are also included as they share some similar propagation characteristics (Huang et al. 2017). MmWave channel measurements and modeling have drawn much attention since T. S. Rappaport's breakthrough work in Rappaport (1991). It demonstrated that mmWave can work for 5G mobile cellular networks.

In the early stage, research of mmWave concentrated on 60 GHz bands as there are at least 5 GHz bandwidths worldwide (Wu et al. 2017a). Recently, other frequency bands such as 11, 15, 26, 28, 32, 38, and 73 GHz have widely been investigated (Rappaport et al. 2015). Various channel measurements and models at these frequency bands have pushed the application of mmWave communications to 5G wireless networks.

Moreover, channel models are required for simulating propagation in a reproducible and cost-effective way and are used to accurately design and compare radio air interfaces and system deployments. Common wireless channel model parameters include carrier frequency, bandwidth, two-dimensional (2-D) or three-dimensional
(3-D) distance between the transmitter (Tx) and receiver (Rx), Doppler effects, delays, and angles. The definitive challenge for a 5G channel model is to provide a fundamental physical basis while being flexible and accurate, especially across a wide frequency range such as 0.5–100 GHz. Recently, a lot of research works aiming at understanding the propagation mechanisms and channel behavior at the frequency bands above 6GHz have been published.

Foundations

For mmWave channel modeling, some fundamental knowledge will be useful. MmWave frequency allocations, propagation characterizations, channel model standardizations, channel modeling methods, and proposed mmWave channel models will be discussed here.

Frequency Allocations

The 57–64 GHz band was allocated in North America and Korea. In Japan, Europe, China, and Australia, the 59–66, 57–66, 59–64, and 59.4–62.9 GHz bands were allocated, respectively. At the World Radio Communication Conference 2015 (WRC-15), 11 frequency bands between 24.25 GHz and 86 GHz are allocated for 5G by ITU. The 24.25–27.5, 37–40.5, 42.5–43.5, 45.5–47, 47.2–50.2, 50.4–52.6, 66–76, and 81–86 GHz bands were allocated, and the 31.8–33.4, 40.5–42.5, and 47–47.2 GHz bands require additional allocations. Those frequency allocations have accelerated the developments of mmWave communications around the world.

In the United States, the 27.5–28.35 (28), 37–38.6 (37), 38.6–40 (39), and 64–71 GHz bands were allocated for flexible wireless use. Additional allocated bands include 24.25–24.45, 24.75–25.25 (24), and 47.2–48.2 GHz. Europe limits its considerations to the bands listed byWRC-15, in particular the 24.25–27.5 (26), 31.8–33.4 (32), and 40.5–43.5 GHz bands. In Korea, the 26.5–29.5 GHz (28 GHz) band was allocated. In Japan, the 27.5–29.5GHz (28 GHz) band was allocated in accordance with America and Korea. In China, the focused bands are 24.25–27.5 GHz (26GHz) and 37–43.5 GHz bands.

Propagation Characterizations

MmWave has some unique propagation characteristics, which should be carefully studied (Hemadeh et al. 2018). MmWave has high path loss and high penetration loss, especially for outdoor-to-indoor (O2I) scenarios. Atmospheric attenuation becomes severe at mmWave bands. The attenuation caused by oxygen absorption can be 15 dB/km and 1.4 dB/km for 60 GHz and 120 GHz, respectively. For water vapor, the attenuation would be 0.18 dB/km and 28.35 dB/km at 23 GHz and 183 GHz, respectively. Rain can also cause 2.55–20 dB/km attenuation depending on the rainfall rate. Foliage also adds additional attenuation. As the size of many objects is at the level of the wavelength, diffraction becomes an important propagation mechanism, and the first Fresnel zone is narrow. The channel also shows sparsity in spatial domain. Blockage caused by human movement in indoor scenarios and buildings in outdoor scenarios will make the line-of-sight (LOS) component unfeasible, thus changing the channel from LOS to non-LOS (NLOS) scenarios (Qi et al. 2017).

Channel Model Standardizations

Several standards have dedicated to 60 GHz bands. IEEE 802.15.3c and 802.11ad have been developed for wireless personal area network (WPAN) in March 2007 and wireless local area network (WLAN) in May 2010, respectively, while 802.11ay is being developed for next-generation wireless fidelity (WiFi). MiWEBA channel model was released in June 2014 for 60 GHz (57–66 GHz) outdoor scenarios, including open area (campus), street canyon, hotel lobby, fronthaul/backhaul, and device-to-device (D2D).

The METIS channel model was released in February 2015. It can support up to 86 GHz with high bandwidth. Supported scenarios include dense urban, urban, rural, indoor (office and shopping mall), and highway.

The mmMAGIC channel model was released in May 2017. Its frequency range is from 6 GHz to 100 GHz. Supported scenarios include urban micro (UMi) (street canyon and open square), indoor (office, shopping mall, and airport), O2I, stadium, and metro station.

The 5GCM white paper was first released in December 2015 in IEEE Globecom'15 and updated in October 2016. It intends to develop channel models for frequency bands up to 100 GHz using the geometry-based stochastic channel modeling (GBSM) approach and has conducted many channel measurements in various scenarios including UMi, urban macro (UMa), and indoor. The majority of the modeling ideas and parameter values were adopted in 3GPP TR 38.900 named "Study on Channel Model for Frequency Spectrum Above 6 GHz", which was first released in February 2016 and updated in July 2017. Its alternative version is 3GPP TR 38.901 named "Study on Channel Model for Frequencies From 0.5 to 100 GHz", in which the sub-6 GHz bands are also covered. It was first released in February 2017 and updated in December 2017.

Channel Modeling Methods

In general, channel modeling methods can be classified into deterministic and stochastic ones or the combination of them (Yang et al. 2017). Deterministic channel models are site-specific and can be verified by comparison of detailed path parameters with channel measurements. Stochastic channel models aim to model general environments and can be verified by comparison of channel statistical properties with channel measurements.

Deterministic channel models include ray tracing model, map-based model, and point cloud model. Ray tracing is based on geometry optics (GO), geometrical theory of diffraction (GTD), and uniform theory of diffraction (UTD), which are approximation and simplification of high-frequency electromagnetic propagations mechanisms. The METIS map-based model is based on ray tracing using a simplified geometric description of the environment. Point cloud is a ray tracing-like prediction tool to characterize the environment with higher precision (Järveläinen et al. 2016). Quasi-deterministic (Q-D) model, which is adopted by MiWEBA and IEEE 802.11ay, is a semi-deterministic channel model.

Stochastic channel models include SV-based model, propagation graph model, and GBSM. SV-based model assumes rays arrive in clusters and is used in IEEE 802.15.3c and 802.11admodels. Propagation graph model is based on graph theory. It models Txs, Rxs, and scatterers as vertices and propagation paths as edges. GBSM has widely been used for channel modeling and adopted in many standard channel models like WINNER, 3GPP TR 38.900/901, NYU WIRELESS, mmMAGIC QuaDRiGa, and METIS stochastic models.

Proposed mmWave Channel Models

In order to meet the requirements of mmWave channel models, many major organizations, such as 3GPP, METIS, mmMAGIC, and 5GCM, have proposed the models for LOS probability, path loss, and building penetration. In the LOS probability model, the LOS probability is often modeled as a function of the 2-D distance between the Tx and Rx, and it includes UMi LOS probability, UMa LOS probability, and InH LoS probability. Furthermore, the large-scale path loss and O2I penetration loss models were also proposed in Rappaport et al. (2017).

Key Applications

As we know, 5G will be used in three scenarios, i.e., enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultrareliable low-latency communications (uRLLC). MmWave will be a key technology to enable these applications. Typical and potential application scenarios include fronthaul/backhaul, indoor hotspots, cellular networks, high-speed train (HST), vehicle-to-vehicle (V2V), D2D, unmanned aerial vehicle (UAV) communications, etc.

Future Directions

mmWave Channel Measurements in New Scenarios

As mmWave will be applied to various new scenarios, channel measurement campaigns in these challenging scenarios are indispensable, which include but are not limited to massive multiple-input multiple-output (MIMO) (Busari et al. 2017), HST, V2V, D2D, UAV communications, etc. MmWave channel measurements in these new scenarios will lead to a better understanding of mmWave propagation characteristics and more accurate mmWave channel models.

General 5G Channel Models

Though many channel models have been proposed, most of them cannot fulfill all the requirements of 5G, covering the vast mmWave frequency bands and new propagation characteristics. A preliminary general 5G channel model was proposed in Wu et al. (2017b). It aimed at capturing small-scale fading channel characteristics of key 5G communication scenarios, such as massive MIMO, HST, V2V, and mmWave. It will serve as a good guideline for future more general and standard 5G models.

Big Data-Enabled mmWave Channel Modeling

Wireless big data has now been investigated both in academia and industry to handle massive datasets generated in wireless networks using artificial intelligence (AI) technologies and machine learning (ML) algorithms (Bi et al. 2015). Big data analytics have been applied to different layers of wireless networks for channel estimation, positioning/localization, network optimization and management, etc. Some preliminary works also apply big data analytics to wireless channel modeling (Ma et al. 2017). Big data will be a powerful tool for mmWave channel modeling by training and learning big channel measurement or simulation datasets.

References

Bi S, Zhang R, Ding Z, Cui S (2015) Wireless communications in the era of big data. IEEE Commun Mag 53(10):190–199

Busari SA, Huq KMS, Mumtaz S, Dai L, Rodriguez J (2017) Millimeter-wave massive MIMO communi-

cation for future wireless systems: a survey. IEEE Commun Surv Tutorials PP(99):1–1
Hemadeh I, Satyanarayana K, El-Hajjar M, Hanzo L (2018) Millimeter-wave communications: physical channel models, design considerations, antenna constructions and link-budget. IEEE Commun Surv Tutorials PP(99):1–1
Huang J, Wang CX, Feng R, Sun J, Zhang W, Yang Y (2017) Multi-frequency mmWave massive MIMO channel measurements and characterization for 5G wireless communication systems. IEEE J Sel Areas Commun 35(7):1591–1605
Järveläinen J, Haneda K, Karttunen A (2016) Indoor propagation channel simulations at 60 GHz using point cloud data. IEEE Trans Antennas Propag 64(10): 4457–4467
Ma X, Zhang J, Zhang Y, Ma Z, Zhang Y (2017) A PCA-based modeling method for wireless MIMO channel. In: 2017 IEEE conference on computer communications workshops (INFOCOMWKSHPS), Atlanta, pp 874–879
Qi W, Huang J, Sun J, Tan Y, Wang CX, Ge X (2017) Measurements and modeling of human blockage effects for multiple millimeter wave bands. In: 2017 13th international wireless communications and mobile computing conference (IWCMC), Valencia, pp 1604–1609
Rappaport TS (1991) The wireless revolution. IEEE Commun Mag 29(11):52–71
Rappaport TS et al (2015) Millimeter wave wireless communications. Prentice-Hall, upper saddle river
Rappaport TS et al (2017) Overview of millimeter wave communications for fifth-generation (5G) wireless networks with a focus on propagation models. IEEE Trans Antennas Propag 65(12):6213–6230
Wang CX, Haider F, Gao X, You XH, Yang Y, Yuan D, Aggoune HM, Haas H, Fletcher S, Hepsaydir E (2014) Cellular architecture and key technologies for 5G wireless communication networks. IEEE Commun Mag 52:122–130
Wu X, Wang CX, Sun J, Huang J, Feng R, Yang Y, Ge X (2017a) 60-GHz millimeter-wave channel measurements and modeling for indoor office environments. IEEE Trans Antennas Propag 65(4):1912–1924
Wu S, Wang CX, Aggoune EHM, Alwakeel MM, You XH (2017b) A general 3D non-stationary 5G wireless channel model. IEEE Trans Commun PP(99):1–1
Yang Y, Xu J, Shi G, Wang CX (2017) 5G wireless systems: simulation and evaluation techniques, Wireless networks. Springer, Cham. URL http://cds.cern.ch/record/2300921

Millimeter Wave Communication Security

▸ Millimeter Wave Security

Millimeter Wave Large-Scale MIMO

▸ Millimeter Wave Massive MIMO

Millimeter Wave MAC Layer

Hossein S. Ghadikolaei[1] and Carlo Fischione[2]
[1]Department of Networks and Systems Engineering, KTH Royal Institute of Technology, Stockholm, Sweden
[2]KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Medium access control in millimeter-wave wireless communications

Definition

The millimeter-wave (mmWave) medium access control (MAC) is a set of routines that regulate access to a common wireless communications medium in the mmWave frequency bands. MAC layer is part of layer 2 of the OSI model (Bertsekas and Gallager 1992).The primary functions of this layer for mmWave wireless communications are the following:

- Data encapsulation, including the frame assembly and error control;
- Media access control, including initial frame transmission, retransmission in the case of failure, and flow control; and
- Initial access, mobility management, and handover.

The first two items are common to MAC of any communication systems, and the last one is more specific for mmWave networks.

The choice of MAC protocol directly influences most performance measures, including

end-to-end latency, network throughput, reliability, and energy efficiency, among others.

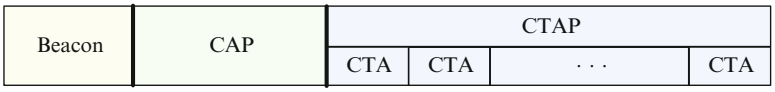
Historical Background

Millimeter-wave communications use the part of the electromagnetic spectrum in the range 30–300 GHz, which corresponds to wavelengths from 10 to 1 mm. In the literature, however, mmWave frequencies casually refer to the frequency band between 6 and 300 GHz. The mmWave wireless communications exhibit the following fundamental physical characteristics: short wavelengths, large bandwidth, severe propagation loss, high attenuation through most solid materials, sparse scattering environments, and large-scale antennas (Rappaport et al. 2013). The special characteristics of mmWave systems demand a thorough reconsideration of traditional protocol design principles, especially at the MAC layer with respect to the MAC protocols of traditional wireless communications.

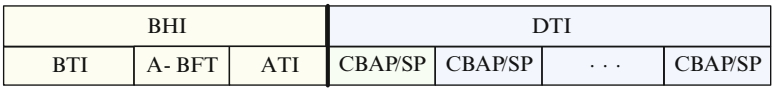
When sending data over a network, the MAC layer encapsulates its incoming payload (data frames from higher layers) into frames appropriate for the transmission medium, adds a frame check sequence to identify transmission errors, and then forwards the data to the physical layer after running a multiple access procedure (Bertsekas and Gallager 1992). This procedure is necessary to resolve or avoid collisions, retransmit the collided or lost frames, and control the transmission rate. At the receiver side, the MAC layer verifies the sender's frame check sequences, removes preambles and frame headers, and delivers the payload to the higher layers (Bertsekas and Gallager 1992). In wireless networks, the design of the MAC layer is tightly integrated with the design of the physical layer. The multiple access protocol of short-range wireless networks (such as wireless personal and local area networks) may rely on carrier sensing among devices and multihop communications. The multiple access protocols of the cellular networks, like Long-Term Evolution (LTE), are usually optimized and run by the base stations (3GPP TS 36.321 2017).

At the mmWave frequency band, there are several standardization activities within cellular networks, wireless personal area networks (WPANs), and wireless local area networks (WLANs), such as 3GPP NR 3GPP TS 38.321 (2017), IEEE 802.15.3 Task Group 3c (TG3c) (IEEE 802.15.3c 2009), IEEE 802.11ad standardization task group (IEEE 802.11ad 2012), WirelessHD Consortium, and Wireless Gigabit Alliance (WiGig). The existing established standards are very suboptimal, and efficient MAC layer design for mmWave networks is a topic of intense research. Ghadikolaei et al. (2015) and Ghadikolaei et al. (2016) discussed the main MAC layer issues of infrastructure-based (e.g., cellular) and infrastructure-free (e.g., ad hoc) mmWave networks, respectively. For short-range networks, IEEE 802.11ay is the most recent study group within IEEE, formed in May 2015, which aims to modify IEEE 802.11ad to enhance the throughput, range, and most importantly the use cases, while ensuring backward compatibility and coexistence with legacy mmWave standards. Broadly speaking, IEEE standards define a network with one coordinator and several mmWave devices. The coordinator, which can be a device itself, is responsible for broadcasting, synchronization beacons, running initial access beacons, and managing radio resources. Figure 1 illustrates generic timing segmentation of IEEE 802.15.3c and IEEE 802.11ad.

The IEEE 802.15.3c MAC procedure was standardized in 2009 (IEEE 802.15.3c 2009). It selects one device, among a group of devices in the network, as the piconet coordinator (PNC), broadcasting beacon messages. Time is divided into successive super-frames, each consisting of three portions, beacon, contention access period (CAP), and channel time allocation period (CTAP), as shown in Fig. 1a. In the beacon, the coordinator transmits omnidirectional or multiple quasi-omnidirectional beacons to facilitate the device discovery procedure. In the CAP, devices with low QoS requirements start their data transmissions. Devices with high QoS requirements contend to register their channel access requests at the PNC, based on carrier



(a) Superframe of IEEE 802.15.3c



(b) Beacon interval of IEEE 802.11ad

Millimeter Wave MAC Layer, Fig. 1 Network timing structure of existing IEEE mmWave standards. In IEEE 802.15.3c, beacon messages are transmitted in the Beacon phase. Channel access requests are made in CAP and served in CTAP using TDMA. Similar procedures are adopted in IEEE 802.11ad. (This picture is reproduced from Ghadikolaei et al. 2016). (**a**) Superframe of IEEE 802.15.3c. (**b**) Beacon interval of IEEE 802.11ad

Millimeter Wave MAC Layer, Table 1 Application scenarios for short-/medium-range mmWave networks. This table is deduced from ongoing discussions inside IEEE 802.11ay study group. "NS" means not specified yet

Usage models	Delay (s)	Availability	Range (m)	Rate (Gbps)	Application scenarios
Ultrashort-range communications	<1	NS	<10	10	Wireless tollgate to transfer e-magazine, picture library, 4K movie trailers, 4K movies
8K Video transfer at smart home	<0.005	NS	<5	28	8K video stream between a source device (e.g., setup box) and a sink device (e.g. smart TV), replacement of wired interface
Augmented reality	<0.005	NS	<10	20	Interface between a mobile wearable device and its managing device to deliver 3D video
Data center	<0.1	99.99%	<5	40	Inter-rack connectivity
Vehicular networks	<0.1	NS	<1000	ns	Inra- and inter-car connectivity, intersection collision avoidance, public safety
Video on-demand	<0.1	NS	<100	NS	Broadcast in crowd public places (e.g., classroom, in flight, train, bus, exhibitions)
Mobile offloading	<0.1	99.99%	<100	20	Offload video traffic from cellular interface to the mmWave interface
Mobile fronthauling	<0.035	99.99%	<200	20	Wireless connections between remote radio heads and base band unit
Mobile backhauling	<0.035	99.99%	<1000	20	Small cell backhauling, mutihop backhauling, inter-building communications

sense multiple access with collision avoidance (CSMA/CA). The PNC serves the registered devices during CTAP. Resource allocation in CTAP is based on time division multiple access (TDMA), in which dedicated channel time allocations (CTAs) are used for different data traffic.

The IEEE 802.11ad was standardized in 2012. It adds modifications to the IEEE 802.11 physical and MAC layers to enable mmWave communications at 60 GHz (IEEE 802.11ad 2012). It defines a network as a personal basic service set (PBSS) with one coordinator, called PBSS control point (PCP), and several stations. A super-frame, called beacon interval, is divided into a beacon header interval (BHI) and a data transfer interval (DTI); see Fig. 1b. BHI consists of a beacon transmission interval (BTI), an association beamforming training (A-BFT), and an announcement transmission interval (ATI). DTI consists of several contention-based access periods (CBAPs) and service periods (SPs). In BTI, PCP transmits directional beacon frames that contain basic timing for the personal BSS, fol-

lowed by beamforming training and association to PCP in the A-BFT period. ATI is allocated for request-response services where PCP sends information to the stations. Depending on the required QoS level, a device will be scheduled in the CBAP to transmit data using CSMA/CA or in the SP for contention-free access using TDMA. This schedule is announced to the participating stations prior to the start of DTI.

Key Applications

Currently, the mmWave spectrum is primarily used for satellite communications, long-range point-to-point communications, military applications, and local multipoint distribution service (Rappaport et al. 2014). There are growing interests in using mmWave systems to support extremely high data rate and low latency services for short- and medium-range networks. Table 1 shows some of these potential use cases.

Cross-References

- Interference Management
- MAC in Cognitive Radio Networks
- Millimeter-wave Communications
- Multiple Access Technique for Cellular Wireless Networks
- QoS-Aware MAC
- Resource Allocation in SDN/NFV-Enabled 5G Networks

References

3GPP TS 36321 (2017) Evolved universal terrestrial radio access (E-UTRA); medium access control (MAC) protocol specification

3GPP TS 38321 (2017) Nr; medium access control (MAC) protocol specification evolved universal terrestrial radio access (E-UTRA); medium access control (MAC) protocol specification

Bertsekas DP, Gallager RG (1992) Data networks, vol 2. Prentice-Hall, Englewood Cliffs

Ghadikolaei HS, Fischione C, Fodor G, Popovski P, Zorzi M (2015) Millimeter wave cellular networks: a MAC layer perspective. IEEE Trans Commun 63(10): 3437–3458

Ghadikolaei HS, Fischione C, Popovski P, Zorzi M (2016) Design aspects of short range millimeter wave wireless networks: a MAC layer perspective. IEEE Netw 30(3):88–96

IEEE 80211ad (2012) IEEE 802.11ad. Part 11: wireless LAN medium access control (MAC) and physical layer (PHY) specifications – amendment 3: enhancements for very high throughput in the 60 GHz band

IEEE 802153c (2009) IEEE 802.15.3c Part 15.3: wireless medium access control (MAC) and physical layer (PHY) specifications for high rate wireless personal area networks (WPANs) amendment 2: millimeter-wave-based alternative physical layer extension

Rappaport TS, Sun S, Mayzus R, Zhao H, Azar Y, Wang K, Wong GN, Schulz JK, Samimi M, Gutierrez F (2013) Millimeter wave mobile communications for 5G cellular: it will work! IEEE Access 1:335–349

Rappaport TS, Heath R, Daniels RC, Murdock JN (2014) Millimeter wave wireless communications. Pearson Education, Upper Saddle River

Millimeter Wave Massive MIMO

Wei Xu[1], Yongming Huang[2], and Ming Xiao[3]
[1]Southeast University, Nanjing, China
[2]School of Information Science and Engineering, Southeast University, Nanjing, China
[3]Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Millimeter wave large-scale MIMO

Definition

Millimeter wave massive multiple-input multiple-output (MIMO) is a wireless communication technique that transmits and receives millimeter-length electromagnetic wave signals through massive antenna arrays at transmitters and receivers. Multiantenna arrays in MIMO enable simultaneous transmissions of multiple data

streams through wireless channels. In most cases, massive MIMO implicitly implies that multiple users are served simultaneously, that is, a multiuser massive MIMO. If only the transmitter deploys a multiantenna array while each receiver equips a single antenna, the terminology MIMO refers to, more formally, multiple-input single-output (MISO). If only the receiver deploys an antenna array while the transmitter equips a single antenna, the terminology MIMO refers to, more formally, single-input multiple-output (SIMO). In literature, millimeter wave massive MIMO can also represent millimeter wave massive MISO or millimeter wave massive SIMO communications.

Historical Background

The development of millimeter wave massive MIMO traces back to a natural combination of two lines of technology evolutions, i.e., millimeter wave communication and massive MIMO, in wireless communication. Millimeter wave refers to the radio spectrum with wavelength between 1 and 10 mm, corresponding to the spectrum between 30 and 300 GHz that is sometimes called the extremely high frequency (EHF) range. Millimeter wave communication was firstly experimented by J. C. Bose in 1895 realizing transmission and reception of electromagnetic waves at 60 GHz over 23 m distance (Emerson 1997). Early applications, however, emerged in radio astronomy and military nearly half a century later. Commercial applications of millimeter wave radar and communication devices started to become popular with the development of millimeter wave integrated circuits after the 1980s. Communications in millimeter wave usually operate with a much larger bandwidth that can be tens of times of the available bandwidth for wireless communications in microwave, which is able to support very high date rate communications. On the other hand, attenuation of radio waves increases exponentially in both propagation distance and the radio frequency. Millimeter wave communications thus face the challenge of high propagation loss due to the high frequency. Moreover, compared to microwave propagation, attenuation of millimeter wave propagation also depends heavily on many other factors, e.g., atmospheric humidity, fog, and rain. Studies were conducted in investigating propagation characteristics of millimeter waves for communication system design (Rappaport et al. 2013). In 2013, millimeter wave communication was specified in the Wi-Fi standard IEEE 802.11ad in the 60 GHz spectrum to achieve high data rate transmission up to 7 Gbit/s (Wi-Fi Alliance 2016).

The invention of MIMO in wireless communication traces back to the early 1990s. The well-known MIMO transmission strategy, named BLAST, was firstly reported in Foschini (1996). Theoretical channel capacity of the MIMO channel was analyzed in Foschini and Gans (1998) in 1998. It revealed that MIMO is able to boost the channel capacity multiple times by deploying multiantenna arrays at both transmitter and receiver. Commercial applications of MIMO were successful in both Wi-Fi and 3G/4G cellular networks, e.g., by standardizing multiantenna stations and terminals with multiple antennas up to eight.

In 2010, a milestone in the development of MIMO technology was the initial introduction of massive MIMO by Dr. T. Marzetta from Bell lab (Marzetta 2010). Massive MIMO proposes to equip the device with an antenna array using a very large number of antenna elements, which is able to support many users to transmit data simultaneously and efficiently without interfering each other too much. The study in Ngo et al. (2013) discovered that the transmit power can be effectively scaled down with the number of antennas while still guaranteeing a desired transmission rate in massive MIMO. Due to its advantages, massive MIMO has been recognized as one of the most essential technology evolutions in the next-generation cellular network, namely 5G. With the deployment of a massive antenna array, however, implementation of massive MIMO in practice also faces many challenges. For instance, estimation of the large-dimensional MIMO channel is resource-hungry; deployment of a massive antenna array occupies a huge space when using

microwave spectrum; and the cost of constructing and driving a massive antenna array could be expensive (Rusek et al. 2013).

Millimeter wave massive MIMO, more than a natural combination of massive MIMO communication with millimeter wave spectrum, exhibits multiple unique features compared to conventional massive MIMO using the microwave spectrum. Besides the large amount of spectrum resource at millimeter wave band, the much shorter wavelength facilitates packing massive antennas into an array of compact size, which endues millimeter wave massive MIMO with great potential in providing ultra-high data rate communications (Xiao et al. 2017). Because millimeter wave channels are sparse in propagation paths and suffer large path-loss and blockage, the MIMO transmission and reception design of millimeter massive MIMO is substantially different from that for conventional microwave massive MIMO. In order to combat the large path-loss, power efficiency has been one of the principle considerations in system design. It becomes particularly important to form directional beams in millimeter wave MIIMO by using massive antenna array (Huang et al. 2018). Transmit power can thus be mostly assembled in the directional and narrow beams to combat large path-loss and enhance communication links.

Understanding propagation characteristics of millimeter wave massive MIMO channels is crucial for both optimization of antenna array topology and design of directional beam patterns. A thorough survey was reported in Rappaport et al. (2017) on existing measurement results and modeling of millimeter wave MIMO propagations, especially in the context of their applications in 5G. A common agreement is that the channel is sparse in the angular domain, and in a large portion of application scenarios, line-of-sight propagation components play a significant part of all propagation effects. The sparsity nature, in one way, provides great potential in solving the channel estimation challenge in massive MIMO, while in the other way it limits the number of degree of freedoms (DoF) that can be achieved through the large dimensional MIMO channel.

In recent years, new signal processing techniques (Heath et al. 2016) as well as hardware-friendly architectures (Liang et al. 2014; Ayach et al. 2014) are emerging to exploit these unique features and meet corresponding challenges in millimeter wave massive MIMO. One of the well-acknowledged new solutions refers to as hybrid analog-digital precoding/combing that balances the sparsity channel feature and hardware constraints in millimeter wave MIMO circuit design (Brady et al. 2013; He et al. 2017). Design parameters in hybrid analog-digital millimeter wave massive MIMO system were analyzed in Xu et al. (2017) and Xue et al. (2017), proposing an effective DoF in performance characterization (Xue et al. 2017). Another design trend is to replace high-resolution analog-to-digital converters (ADC) in conventional massive MIMO by low-resolution ADCs in order to achieve reduction in both circuit power consumption and hardware cost for implementations. Studies, e.g., (Liu et al. 2016; Saxena et al. 2017), discovered that low-resolution ADCs cause negligible performance loss compensable by a massive antenna array in millimeter wave communications.

Key Applications

Millimeter wave massive MIMO enables ultra-high data rate communications in applications like wireless backhaul and ultra-dense small cell coverage in cellular networks. It is also becoming a highly attractive solution for realizing low latency and/or wideband communications in unmanned aerial vehicular communications and the coming intelligent vehicular communication networks (Xiao et al. 2016). Vehicular radar communication and detection is also a growing application of millimeter wave massive MIMO.

Cross-References

- ▸ Millimeter Wave Beam Training and Tracking
- ▸ Millimeter Wave Channel Access
- ▸ Millimeter Wave Channel Modeling

References

Ayach O et al (2014) Spatially sparse precoding in millimeter wave MIMO systems. IEEE Trans Wirel Commun 13(3):1499–1513

Brady J, Behdad N, Sayeed A (2013) Beamspace MIMO for millimeter-wave communications: system architecture, modeling, analysis, and measurements. IEEE Trans Antennas Propag 61(7):3814–3827

Emerson D (1997) The work of Jagadis Chandra Bose: 100 years of mm-wave research. In: IEEE MTT-S international microwave symposium digest, Denver

Foschini G (1996) Layered space–time architecture for wireless communication in a fading environment when using multiple antennas. Labs Syst Tech J Bell 1:41–59

Foschini G, Gans M (1998) On limits of wireless communications in a fading environment when using multiple antennas. Wirel Pers Commun 6(3):311–335

He S, Wang J, Huang Y, Ottersten B, Hong W (2017) Codebook based hybrid precoding for millimeter wave multiuser systems. IEEE Trans Signal Process 65(7):5289–5304

Heath R et al (2016) An overview of signal processing techniques for millimeter wave MIMO systems. IEEE J Sel Top Sign Proces 10(3):436–453

Huang Y, Zhang J, Xiao M (2018) Constant envelope hybrid precoding for directional millimeter-wave communications. IEEE J Sel Areas Commun

Liang L, Xu W, Dong X (2014) Low-complexity hybrid precoding in massive multiuser MIMO systems. IEEE Wireless Commun Lett 3(6):653–656

Liu J, Xu J, Xu W, Jin S, Dong X (2016) Multiuser massive MIMO relaying with mixed-ADC receiver. IEEE Signal Process Lett 24(1):76–80

Marzetta T (2010) Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans Wirel Commun 9(11):3590–3600

Ngo H, Larsson E, Marzetta T (2013) Energy and spectral efficiency of very large multiuser MIMO systems. IEEE Trans Commun 61(4):1436–1449

Rappaport T et al (2013) Millimeter wave mobile communications for 5G cellular: it will work! IEEE Access 1:335–349

Rappaport T et al (2017) Overview of millimeter wave communications for fifth-generation (5G) wireless networks-with a focus on propagation models. IEEE Trans Antennas Propag 65(12):6213–6230

Rusek F et al (2013) Scaling up MIMO: opportunities and challenges with very large arrays. IEEE Signal Process Mag 30(1):40–60

Saxena A, Fijalkow I, Swindlehurst A (2017) Analysis of one-bit quantized precoding for the multiuser massive MIMO downlink. IEEE Trans Signal Process 65(17):4624–4463

Wi-Fi Alliance (2016) Wi-Fi Certificated WiGig: Wi-Fi expands to 60 GHz. https://www.wi-fi.org/

Xiao Z, Xia P, Xia X (2016) Enabling UAV cellular with millimeter-wave communication: potentials and approaches. IEEE Commun Mag 54(5):66–73

Xiao M et al (2017) Millimeter wave communications for future mobile networks. IEEE J Sel Areas Commun 35(9):1909–1935

Xu W, Liu J, Jin S, Dong X (2017) Spectral and energy efficiency of multi-pair massive MIMO relay network with hybrid processing. IEEE Trans Commun 65(9):3794–3809

Xue C, He S, Huang Y, Wu Y, Yang L (2017) An efficient beam-training scheme for the optimally designed subarray structure in mmWave LoS MIMO systems. EURASIP J Wirel Commun Netw 2017:31

Millimeter Wave Multiple Access

▸ Millimeter Wave Channel Access

Millimeter Wave NOMA

Bichai Wang[1], Linglong Dai[1], and Ming Xiao[2]
[1]Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

Multiple access in high frequency

Definitions

Millimeter-wave NOMA refers to using non-orthogonal multiple access (NOMA) in millimeter-wave frequencies. On the one hand, millimeter-wave communications, operating from 30 to 300 GHz, provide an opportunity to meet explosive capacity demand for wireless

communications. On the other hand, multiple access technologies are necessary for supporting multiple users in wireless communication systems. Particularly, NOMA can significantly improve the spectral efficiency and connectivity density. In contrast to the conventional orthogonal multiple access (OMA) schemes realized in the time-, frequency-, code-domain or their combinations, NOMA can be realized in a new domain, i.e., the power domain. By performing superposition coding at the transmitter and successive interference cancelation (SIC) at the receiver, multiple users can be simultaneously supported at the same time-frequency-space resources, and the channel gain difference among users can be translated into multiplexing gain by superposition coding. By integrating NOMA into millimeter-wave communications, potential performance gain can be achieved.

Key Points

Multiple access technologies are necessary for supporting multiple users in mobile networks, and they have been widely investigated in the lower frequency bands. Different multiple access technologies have been utilized in practical systems, including frequency division multiple access (FDMA), time division multiple access (TDMA), code division multiple access (CDMA), and orthogonal frequency division multiple access (OFDMA). These multiple access technologies are also applicable to millimeter-wave band, but in different flavors than in lower frequency bands due to the increased complexity caused by the greatly increased bandwidth, and the different channel characteristics in millimeter-wave bands, e.g., highly directional transmissions (Mumtaz et al. 2016).

In addition to orders-of-magnitude lager bandwidths, the smaller wavelengths at millimeter-wave allow more antennas in a same physical space, which enables massive multiple input multiple output (MIMO) to provide more multiplexing gain and beamforming gain (Swindlehurst et al. 2014). In fact, beamforming with a large antenna array is a key characteristic of millimeter-wave communication, which is used to compensate for the high pass loss of millimeter-wave signals (Xiao et al. 2017a).

To further increase the spectral efficiency, NOMA has been recently considered in millimeter-wave communications. It has been shown that NOMA can significantly improve the spectral efficiency compared to the conventional OMA schemes (Dai et al. 2015). In Ding et al. (2017), the integration of NOMA with millimeter-wave communications has be investigated, where random steering single-beamforming was adopted, which can work only in a special case that the NOMA users are close to each other. In Xiao et al. (2017b), joint power allocation and beamforming to maximize the sum rate of a two-user millimeter-wave NOMA system was proposed, where an analog beamforming structure with a phased array was considered.

Furthermore, it is well known that in conventional MIMO systems, each antenna usually requires one dedicated radio-frequency (RF) chain to realize the fully digital signal processing (Rusek et al. 2013). In this way, the use of a very large number of antennas in millimeter-wave massive MIMO systems leads to an equally large number of RF chains, which will result in unaffordable hardware cost and energy consumption (Heath et al. 2016). To address this issue, hybrid precoding (HP) has been proposed to significantly reduce the number of required RF chains in millimeter-wave massive MIMO systems without an obvious performance loss (Rusek et al. 2013).

By using NOMA in HP-based millimeter-wave massive MIMO systems, more than one user can be supported in each beam with the aid of intra-beam superposition coding and SIC, which is essentially different from conventional millimeter-wave massive MIMO using one beam to serve only one user at the same time-frequency resources. Particularly, NOMA was applied to beamspace MIMO for the first time in Wang et al. (2017), which can be regarded as a low-complexity realization of HP, and power allocation was optimized to maximize

the achievable sum rate. In addition, NOMA was also utilized in fully connected HP architecture in Yuan et al. (2017), and digital precoding was designed by modifying the conventional block diagonalization (BD) precoding scheme. Furthermore, a more sophisticated digital precoding design, i.e., minorization-maximization (MM)-based precoding, was proposed in Zhao et al. (2017) to maximize the achievable sum rate. Besides, power allocation was optimized in Hao et al. (2017) to maximize the energy efficiency of millimeter-wave massive MIMO-NOMA systems, and an iterative algorithm was proposed to obtain the optimal power allocation.

Note that the highly direction feature of millimeter-wave propagation makes the users' channels (along the same or similar direction) highly correlated, which facilitates the integration of NOMA in millimeter-wave communication. Therefore, considering the harmony between the millimeter-wave channel characteristics and the principle of NOMA, the use of NOMA for millimeter-wave communications is a very promising research direction deserves further investigations.

Cross-References

- ▸ Lens Antenna Array
- ▸ Millimeter Wave Massive MIMO

References

Dai L, Wang B, Yuan Y, Han S, Chih-Lin I, Wang Z (2015) Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. IEEE Commun Mag 53(9):74–81

Ding Z, Fan P, Poor HV (2017) Random beamforming in millimeter-wave NOMA networks. IEEE Access 5:7667–7681

Hao W, Zeng M, Chu Z, Yang S (2017) Energy-efficient power allocation in millimeter wave massive MIMO with non-orthogonal multiple access. IEEE Wireless Commun Lett 6(6):782–785

Heath RW, Gonzalez-Prelcic N, Rangan S, Roh W, Sayeed AM (2016) An overview of signal processing techniques for millimeter wave MIMO systems. IEEE J Sel Top Sign Proces 10(3):436–453

Mumtaz S, Rodriguez J, Dai L (2016) MmWave massive MIMO: a paradigm for 5G. Academic Press, London/San Diego

Rusek F, Persson D, Lau BK, Larsson EG, Marzetta TL, Edfors O, Tufvesson F (2013) Scaling up MIMO: opportunities and challenges with very large arrays. IEEE Signal Process Mag 30(1):40–60

Swindlehurst AL, Ayanoglu E, Heydari P, Capolino F (2014) Millimeter-wave massive MIMO: the next wireless revolution? IEEE Commun Mag 52(9): 56–62

Wang B, Dai L, Wang Z, Ge N, Zhou S (2017) Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array. IEEE J Sel Areas Commun 35(10): 2370–2382

Xiao Z, Dai L, Ding Z, Choi J, Xia P (2017a) Millimeter-wave communication with non-orthogonal multiple access for 5G. arXiv preprint arXiv:170907980

Xiao Z, Zhu L, Choi J, Xia P, Xia XG (2017b) Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter-wave communications. arXiv preprint arXiv:171101380

Yuan W, Kalokidou V, Armour SMD, Doufexi A, Beach A (2017) Application of non-orthogonal multiplexing to mmWave multi-user systems. In: IEEE vehicular technology conference (IEEE VTC'17 Spring), IEEE, pp 1–5

Zhao Y, Xu W, Jin S (2017) An minorization-maximization based hybrid precoding in NOMA-mMIMO. In: IEEE international conference on wireless communications and signal processing (IEEE WCSP'17), IEEE, pp 1–6

Millimeter Wave Physical Layer Security

- ▸ Millimeter Wave Security

Millimeter Wave Security

Pei Zhou, Qing Xue, and Xuming Fang
Southwest Jiaotong University, Chengdu, China

Synonyms

Millimeter Wave Communication Security; Millimeter Wave Physical Layer Security

Definition

Different from the traditional communication security methods which provide data security by cryptographic techniques, physical layer security methods guarantee secure wireless transmissions by exploiting the imperfections of the communications medium (Yang et al. 2015). With physical layer security methods, communication security can be achieved by degrading the quality of signal reception at potential eavesdroppers and therefore preventing them from acquiring confidential information from the received signal. However, compared to microwave communications, severe path loss in millimeter wave (mmWave) makes it hard to support long-distance transmissions. Thanks to the short wavelength of mmWave, it is possible to employ a large number of antennas in a small size terminal. With the help of the large number of antennas, directional beamforming and multiple-antenna technologies, etc. become the powerful tools for achieving long-distance communications and enhancing the physical layer security in mmWave networks. With the degrees of freedom provided by multiple antennas (Wang and Wang 2016), better communication security can be achieved in mmWave communications.

Key Points

The conventional communication security can be achieved by ensuring all involved entities load proper and authenticated cryptographic information. However, the physical layer security does not consider about how those security protocols are executed and does not require to implement any extra security schemes or algorithms on other layers above the physical layer (Liu et al. 2017). There are many techniques that can be used for improving the security of mmWave, such as directional transmission, antenna subset modulation, and directional modulation.

Directional Transmission: Different from microwave networks, directional transmission techniques should be used in mmWave networks to achieve better performance. Thus, the signal or interference power in mmWave communications is highly directional and closely related to critical parameters, such as transmission distance, transmit power, offset angle of departure/arrival, and beamwidth. They all have impact on the performance of mmWave communications due to the inherent directivity. If one or more of these parameters do not reach the predefined thresholds, the directional communication in mmWave networks will be damaged. Therefore, it is hard for eavesdroppers to acquire confidential information from the received signal. Meanwhile, physical layer security could naturally gain the benefit from the directivity.

Antenna Subset Modulation: With multi-antenna techniques, by using random antenna subsets and performing analog precoding with antenna selection, instead of using all antennas for beamforming, a random set of antennas are co-phased to transmit the information symbol to the desired receiver, while the rest of the antennas are used to randomize the far field radiation pattern at non-desired receiver directions. The indices of these antennas are randomized in every symbol transmission. Thus, coherent transmission to the desired receiver and a noise-like signal that jams potential eavesdroppers can be achieved (Eltayeb et al. 2017). Similarly, another antenna subset modulation method to realize physical layer security of mmWave is to modulate the radiation pattern at the symbol rate by driving only a subset of antennas in the array. This results in a directional radiation pattern that projects a sharply defined constellation in the desired direction and expanded further randomized constellation in other directions (Valliappan et al. 2013). The way to implement antenna subset modulation is randomly selecting an antenna subset for every symbol. The symbol modulation for a desired receiver along the main lobe will not be affected by randomly switching antenna subsets.

Directional Modulation: Several directional modulation approaches have been proposed that leverage multiple transmit antennas including near field antenna-level modulation,

switched antenna phased array transmitters, and spatial keying transmission techniques such as spatial modulation and space shift keying to achieve enhanced security (Valliappan et al. 2013). A hybrid multiple-input multiple-output (MIMO) phased array time-modulated directional modulation scheme was proposed to improve the physical layer security of mmWave communications (Wang and Zheng 2018). Because the hybrid MIMO phased array can take advantage of the spatial diversity of MIMO, the main advantage of phased array in coherent directional gain will not be sacrificed. Then, the transmit antenna arrays will be divided into multiple subarrays, and each subarray can be used to form a directional beam. All subarrays are jointly combined to work as an MIMO for higher angular resolution or multiuser communications. Therefore, even the eavesdropper's position is unknown, physical layer security can also be realized by applying a time-modulated directional modulation scheme.

In addition, heterogeneous networks (HetNets) architecture has attracted a lot of research interests recently. To provide physical layer security in the new HetNet architecture, dense small cells are deployed to bring ubiquitous inter-tier and intra-tier interference (Zhu et al. 2017). Such interference can be utilized for confounding the eavesdroppers and achieving secure communications at the physical layer as mentioned in *antenna subset modulation*.

We can see from the above physical layer security methods that they do not rely on upper-layer data encryption or secret keys. Thus, the processing overhead and additional communication overhead can be reduced. Physical layer security of mmWave communications is a promising research topic for mmWave security.

Key Applications

In order to provide ultrahigh-speed wireless transmissions for future communication systems, mmWave will play an important role in wireless local area networks (WLANs), fifth generation (5G) mobile communication system, vehicular communication system, and so on. With the increasing demand of privacy and security of every user, communication security is critical to all the above communication systems. mmWave security technologies can be used in any mmWave networks to provide the desired users secure communications and prevent the information from being obtained by potential eavesdroppers.

Cross-References

- ▸ Artificial Noise Schemes Based on MIMO Technology in Secure Cellular Networks
- ▸ Millimeter Wave Massive MIMO

References

Eltayeb ME, Choi J, Al-Naffouri TY, Heath RW (2017) Enhancing secrecy with multiantenna transmission in millimeter wave vehicular communication systems. IEEE Trans Veh Technol 66(9):8139–8151

Liu Y, Chen HH, Wang L (2017) Physical layer security for next generation wireless networks: theories, technologies, and challenges. IEEE Commun Surv Tut 19(1):347–376

Valliappan N, Lozano A, Heath RW (2013) Antenna subset modulation for secure millimeter-wave wireless communication. IEEE Trans Commun 61(8):3231–3245

Wang C, Wang HM (2016) Physical layer security in millimeter wave cellular networks. IEEE Trans Wirel Commun 15(8):5569–5585

Wang WQ, Zheng Z (2018) Hybrid MIMO and phased-array directional modulation for physical layer security in mmWave wireless communications. IEEE J Sel Areas Commun. https://doi.org/10.1109/JSAC.2018.2825138

Yang N, Wang L, Geraci G, Elkashlan M, Yuan J, Di Renzo M (2015) Safeguarding 5G wireless communication networks using physical layer security. IEEE Commun Mag 53(4):20–27

Zhu Y, Wang L, Wong KK, Heath RW (2017) Secure communications in millimeter wave ad hoc networks. IEEE Trans Wirel Commun 16(5):3205–3217

Millimeter-Wave (mmWave) Multiple-Input Multiple-Output (MIMO) Technique

Xinyu Gao[1], Linglong Dai[2], Yongming Huang[3], and Ming Xiao[4]
[1]Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Electronic Engineering, Tsinghua University, Beijing, China
[3]School of Information Science and Engineering, Southeast University, Nanjing, China
[4]Department of Information Science and Engineering (ISE), School of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Synonyms

High-frequency MIMO; Millimeter-wave massive MIMO

Definitions

Multiple-input multiple-output (MIMO) is a technique that employs multiple transmit and receive antennas to send and receive more than one data stream simultaneously over the same time/frequency resource. Millimeter-wave (mmWave) MIMO is the MIMO technique operated at mmWave frequencies (i.e., 30–300 GHz).

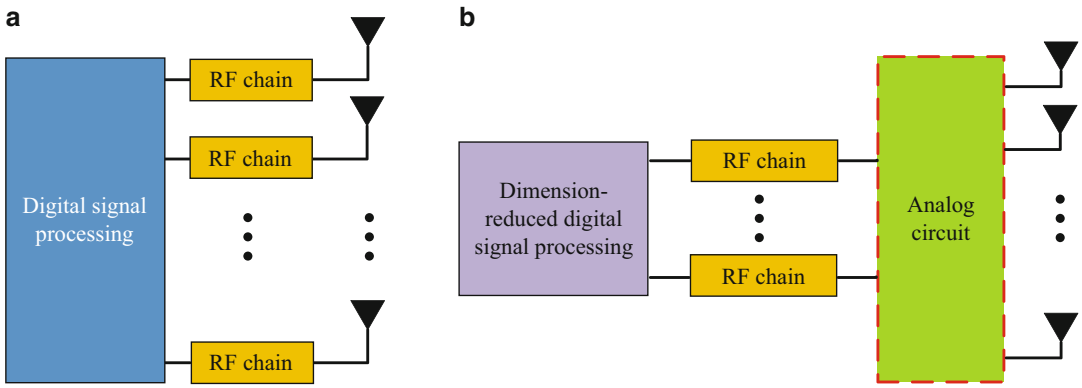
Historical Background

The history of MIMO can be traced back to 1993, when Arogyaswami Paulraj and Thomas Kailath proposed to "split a high-rate signal into several low-rate signals to be transmitted from spatially separated transmitters and recovered by the receive antenna array based on differences in directions-of-arrival" (Li et al. 2010). After that, this technique has attracted great interest from both the academic and industry communities and obtained a huge development. Currently, MIMO technique has been widely used in wireless communication systems working at sub-6 GHz frequencies, such as 4G cellular systems (Li et al. 2010). By employing multiple antennas, MIMO can achieve multiplexing gain with reduced interference to improve the spectrum efficiency.

However, the MIMO technique in current 4G cellular systems only employs a small number of transmit antennas and receive antennas due to the limited physical space at the transmitter and receiver. As a result, the improvement in the spectrum efficiency is still limited, which cannot meet the one thousand times increase in data traffic predicted for further 5G cellular systems (Mumtaz et al. 2016). To solve this problem, one possible way is to extend the MIMO technique at sub-6 GHz frequencies to mmWave frequencies (Mumtaz et al. 2016). On one hand, mmWave band can provide more than 2 GHz bandwidth for communication, which is much wider than the 20 MHz bandwidth in current 4G cellular systems. On the other hand, the short wavelengths associated with mmWave frequencies enable a large antenna array to be packed in a small physical space, which means that MIMO with a large antenna array can be easily employed at mmWave frequencies to considerably improve the spectrum efficiency.

Foundations

The mmWave MIMO technique will be quite different from the one at sub-6 GHz frequencies due to the different channel characteristics and additional hardware constraints found at mmWave frequencies (Gao et al. 2018). The conventional MIMO architecture is fully digital as shown in Fig. 1a, where each antenna requires one dedicated RF chain (including power amplifier, low-noise amplifier, data converter, mixer, and so on) to perform digital signal processing in the baseband (Li et al. 2010). For mmWave MIMO,



Millimeter-Wave (mmWave) Multiple-Input Multiple-Output (MIMO) Technique, Fig. 1 MIMO architectures: (**a**) fully digital architecture; (**b**) hybrid analog and digital architecture

this architecture will suffer from unaffordable hardware cost and power consumption, since (1) the number of antennas is usually large due to the short wavelengths (e.g., 256 antennas at mmWave frequencies instead of 8 antennas at sub-6 GHz frequencies) (Mumtaz et al. 2016) and (2) the power consumption of RF chain is high due to the increased sampling rate (e.g., 250 mW/RF chain at mmWave frequencies, compared with 30 mW/RF chain at sub-6 GHz frequencies) (Gao et al. 2018).

To this end, the hybrid analog and digital architecture is proposed for mmWave MIMO (Gao et al. 2018), as shown in Fig. 1b. This architecture divides the conventional digital signal processing of large size into two parts, i.e., a large-size analog signal processing (realized by analog circuit) and a dimension-reduced digital signal processing (requiring a small number of RF chains). The smallest number of required RF chains in hybrid architecture can equal the number of data streams for communication, which is usually much smaller than the number of antennas. In this way, the number of RF chains can be significantly reduced, leading to quite low energy consumption. On the other hand, due to the limited number of effective scatterers at mmWave frequencies, the mmWave MIMO channel matrix is usually low-rank (Gao et al. 2018), and the maximum number of data streams that can be simultaneously transmitted in such low-rank channel is small. Therefore, as long as the number of RF chains is larger than the rank of channel matrix, the dimension-reduced digital signal processing is still able to fully achieve the multiplexing gain. As a result, hybrid architecture can achieve a better trade-off between the hardware cost/energy consumption and the sum-rate performance (Gao et al. 2018).

M

Architecture for mmWave MIMO

The analog circuit of mmWave MIMO with hybrid architecture can be implemented by different hardware networks (Gao et al. 2018; Méndez-Rial et al. 2016) as listed below.

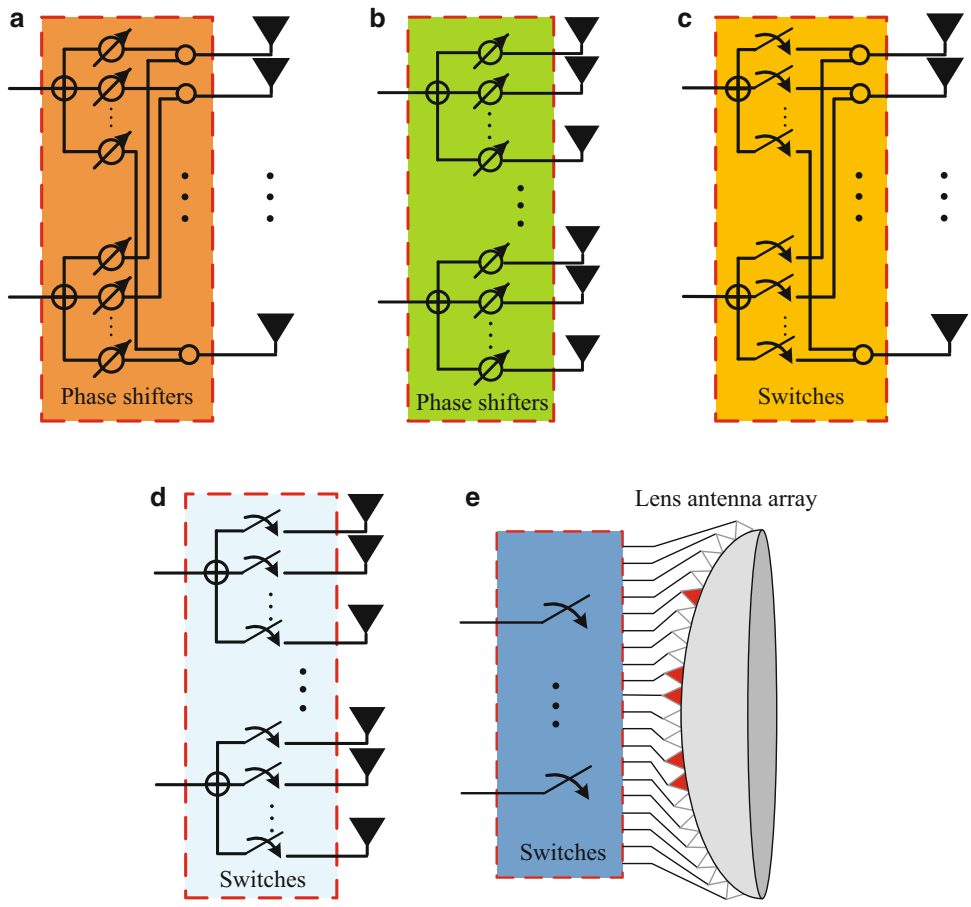
(N1) *Fully connected network with phase shifters:* In this network, each RF chain is connected to all antennas via phase shifters, as shown in Fig. 2a. For each RF chain, full array gain can be achieved by adjusting the phases of transmitted signals on all antennas to provide high array gain. By employing such network, all elements of the analog precoder/combiner have the same fixed amplitude.

(N2) *Sub-connected network with phase shifters:* In this network, each RF chain is only connected to a sub-antenna array via phase shifters, as shown in Fig. 2b. Due to this limitation, for each RF chain, only the transmitted signals on a subset

of antennas can be adjusted. Therefore, compared with N1, the achieved array gain in this network is reduced. However, this network is preferred in practice, since the number of phase shifters required in this network is significantly reduced to save energy consumption. This network imposes two hardware constraints: (i) the analog precoder/combiner should be a block diagonal matrix; (ii) all the nonzero elements of the analog precoder/combiner have the same fixed amplitude.

(N3) *Fully connected network with switches:* In this network, each RF chain is connected to all antennas similar to N1 but via low-cost switches instead of phase shifters as shown in Fig. 2c. Each RF chain selects several antennas for communication, and the transmitted signals cannot be adjusted. Therefore, this network will suffer from serious degradation of array gain. However, as the phase shifters are replaced by simple switches, the hardware cost and energy consumption can be considerably reduced. In this network, all the elements of the analog precoder/combiner can be only selected from the set {0, 1}.

(N4) *Sub-connected network with switches:* In this network, each RF chain is connected to a sub-antenna array like N2 with phase



Millimeter-Wave (mmWave) Multiple-Input Multiple-Output (MIMO) Technique, Fig. 2 Analog circuit with different networks: (**a**) Fully connected network with phase shifters (N1); (**b**) Sub-connected network with phase shifters (N2); (**c**) Fully connected network with switches; (**d**) Sub-connected network with switches; (**e**) Lens antenna array

shifters being replaced by switches, as shown in Fig. 2d. Each RF chain can only select a subset of antennas instead of the whole antenna array to transmit signals. Obviously, the achieved array gain is further reduced compared with N3. However, it also enjoys the lowest hardware cost and energy consumption, as it requires only a small number of switches without the need of splitters. The hardware constraints induced by this network are similar to those of N2, i.e., the analog precoder/combiner should be a block diagonal matrix, but its nonzero elements can be only selected from {0, 1}.

(N5) *Lens antenna array:* An alternative quite different from the four networks discussed above is to utilize lens antenna array, as shown in Fig. 2e. The lens antenna array (a feed antenna array placed beneath the lens) (Brady et al. 2013) can realize the functions of signal emitting and phase shifting simultaneously. It can concentrate the signals from different propagation directions (beams) on different feed antennas. As the scattering at mmWave frequencies is not rich (Gao et al. 2018), the channel power will be concerted on only a small number of beams. Therefore, the simple selecting network with switches can be used to significantly reduce the MIMO dimension as well as the number of RF chains without obvious performance loss. Moreover, since all the phase shifters and splitters are replaced by one simple lens, the hardware cost and energy consumption of this network is also considerably low. Mathematically, the lens antenna array plays the role of a discrete Fourier transform (DFT). Therefore, in this network, each column of the analog precoder/combiner restricts to an DFT column.

Signal Processing for mmWave MIMO

The signal processing (including precoding/combing, channel estimation, and so on) for mmWave MIMO is also different from the one for conventional MIMO since the fully digital architecture is replaced by the hybrid architecture.

We first discuss how to design the hybrid precoding (includes digital precoder and analog precoder) and combining (includes digital combiner and analog combiner) for mmWave MIMO with hybrid architecture. In general, the design target is to maximize the achievable sum rate. However, obtaining the optimal solution is not a trivial task, and the main difficulties are twofold. Firstly, the designs of digital precoder/combiner and analog precoder/combiner are coupled, which makes the optimization problem non-convex. Secondly, as described above, the analog circuit of hybrid architecture imposes non-convex hardware constraints on the analog precoder/combiner. To alleviate these two challenges, one feasible way is to decompose the original optimization problem into several subproblems, and each subproblem is approximated as a convex one and then solved by efficient convex optimization algorithms.

For example, in El Ayach et al. (2014), a spatially sparse precoding scheme is proposed for N1. It first assumes that the hybrid combiner at the receiver is ideal without any hardware constraint. Then, it approximates the sum-rate optimization problem as the one minimizing the distance between the optimal fully digital precoder without any constraint and the hybrid precoder. Then, a variant of the orthogonal matching pursuit (OMP) algorithm is developed to obtain the near-optimal hybrid precoder. Finally, the hybrid combiner can be designed in a similar way based on the effective channel. For N2, a successive interference cancelation (SIC)-based precoding scheme is proposed (Gao et al. 2016). Similar to the spatially sparse precoding scheme, it first decouples the design of precoder and combiner. Then, it decomposes the total achievable sum-rate optimization problem with non-convex constraints into a series of simple sub-rate optimization problems, each of which only considers one sub-antenna array. After that, it obtains the optimal hybrid precoding vector for each sub-antenna array, which is sufficiently close to the unconstrained optimal solution. Then, borrowing the concept of SIC for multiuser detec-

tion, the achievable sub-rate of each sub-antenna array is optimized one by one until the last sub-antenna array is considered. After that, the hybrid combiner can be obtained in a similar way based on the effective channel. Besides these two schemes, some other hybrid precoding and combing schemes designed for N1-N5 can be further found in Méndez-Rial et al. (2016).

The optimal performance of hybrid precoding and combining can be only achieved with perfect channel state information (CSI). However, for mmWave MIMO with hybrid architecture, estimating the complete CSI is not a trivial task (Alkhateeb et al. 2014). Firstly, due to the lack of antenna gain before the establishment of the transmission link, the SNR for channel estimation in mmWave MIMO is quite low. Secondly, the number of RF chains in hybrid architecture is usually much smaller than the number of antennas, so we cannot simultaneously obtain the sampled signals on all antennas. As a result, the traditional channel estimation schemes requiring the sampled signals on all antennas will involve unaffordable pilot overhead in mmWave MIMO. To solve this problem, two dominant categories of channel estimation schemes have been proposed.

The key idea of the first category is to reduce the dimension of channel estimation problem. For example, in Hogan and Sayeed (2016), a two-step channel estimation scheme is proposed for N5. In the first step, it performs the beam training between the transmitter and receiver to obtain the analog precoder and analog combiner. In the second step, the effective channel matrix in the analog domain is estimated by classical algorithms, such as least squares (LS). Note that the size of the effective channel matrix is much smaller than that of the original channel matrix. Therefore, the pilot overhead in the second step is quite low. The second category of channel estimation schemes is to exploit the sparsity of mmWave MIMO channels. Instead of estimating the effective channel matrix of small size, it can directly obtain the complete channel matrix with low pilot overhead. For example, in Alkhateeb et al. (2014), an adaptive channel estimation scheme is proposed for N1. It divides the total channel estimation problem into several subproblems, each of which only considers one channel path. For each channel path, it first starts with a coarse angle of arrival (AoA)/angle of departure (AoD) grids and determines the AoA and AoD of this path belonging to which angle range by employing OMP algorithm. Then, the narrowed direction grids are used, and the direction of this path is further refined. Besides these two schemes, some other channel estimation schemes designed for N1–N5 can be further found in Méndez-Rial et al. (2016).

Key Applications

MmWave MIMO has been considered as a promising technique for further wireless communication systems. The applications of mmWave MIMO are immense. For example, mmWave MIMO could be used to enable high-rate low-latency data transmission for 5G cellular systems. In addition, with the recent excitement related to connected and autonomous vehicles, mmWave MIMO can play an important role in providing high data rate connections between cars. Finally, mmWave MIMO is also of interest for high-speed wearable networks that connect cell phone, smart watch, augmented reality glasses, and virtual reality headsets.

Future Directions

There are still some open issues on mmWave MIMO. For example, most of the existing signal processing for mmWave MIMO is designed under the narrowband and time-invariant channels. However, due to the large bandwidth and the high frequency, the mmWave MIMO channels are more likely to be broadband and time-varying, which incurs new challenges. Take the hybrid precoding for example. Broadband means that the analog precoder cannot be adaptively adjusted according to the frequency, leading to more difficulties in signal processing design, while time-varying means that we need to re-estimate the channel and re-compute the hybrid precoding frequently, leading to high pilot

overhead and computational complexity. Therefore, designing signal processing for mmWave under broadband time-varying channels will be an urging problem to solve.

Cross-References

- ▸ Hybrid Precoding
- ▸ Massive MIMO Channel Estimation

References

Alkhateeb A, El Ayach O, Leus G, Heath RW (2014) Channel estimation and hybrid precoding for millimeter wave cellular systems. IEEE J Sel Topics Signal Process 8(5):831–846

Brady J, Behdad N, Sayeed A (2013) Beamspace MIMO for millimeter-wave communications: system architecture, modeling, analysis, and measurements. IEEE Trans Antennas Propag 61(7):3814–3827

El Ayach O, Rajagopal S, Abu-Surra S, Pi Z, Heath RW (2014) Spatially sparse precoding in millimeter wave MIMO systems. IEEE Trans Wireless Commun 13(3):1499–1513

Gao X, Dai L, Han S, I CL, Heath RW (2016) Energy-efficient hybrid analog and digital precoding for mmWave MIMO systems with large antenna arrays. IEEE J Sel Areas Commun 34(4):998–1009

Gao X, Dai L, Sayeed A (2018) Low RF-complexity technologies to enable millimeter-wave mimo with large antenna array for 5G wireless communications. IEEE Commun Mag 56(4):211–217

Hogan J, Sayeed A (2016) Beam selection for performance-complexity optimization in high-dimension MIMO systems. In: Conference on information science and systems, pp 337–342

Li Q, Li G, Lee W, Lee MI, Mazzarese D, Clerckx B, Li Z (2010) MIMO techniques in WiMAX and LTE: a feature overview. IEEE Commun Mag 48(5):86–92

Méndez-Rial R, Rusu C, González-Prelcic N, Alkhateeb A, Heath RW (2016) Hybrid MIMO architectures for millimeter wave communications: phase shifters or switches? IEEE Access 4:247–267

Mumtaz S, Rodriguez J, Dai L (2016) MmWave massive MIMO: a paradigm for 5G. Academic Press, Elsevier

Millimeter-Wave Communications

- ▸ Per-Beam Synchronization for Millimeter-Wave Massive MIMO

Millimeter-Wave Massive MIMO

- ▸ Millimeter-Wave (mmWave) Multiple-Input Multiple-Output (MIMO) Technique

MIMO

- ▸ Index Modulation

MIMO Detection

Lin Bai[1], Tian Li[2], and Quan Yu[3]
[1]School of Electronic and Information Engineering, Beijing Laboratory for General Aviation Technology (Beihang University), Beijing, China
[2]The 54th Research Institute of CETC, Shijiazhuang Hebei, China
[3]School of Electronic and Information Engineering, Beihang University, Beijing, China

Synonyms

MIMO signal demodulation; MIMO signal equalization

Definitions

MIMO detection is to jointly detect multiple signals at the receiver, where the signals are transmitted through a wireless channel.

Historical Background

MIMO is an important technology in B3G/4G wireless communication systems, where both transmitters and receivers are equipped with antenna arrays. By exploiting the spatial diversity

benefitted from the multiple antennas, MIMO was first developed by Marconi in 1908 to combat channel fading. Then, the staff in Bell Laboratory extended this technology in different transmission environment. In 1995, Telatar studied the channel capacity of an $N_r \times N_t$ point-to-point MIMO system under Rayleigh fading condition and proofed that the achievable rate increases linearly with $\min(N_r, N_t)$ (Telatar 1999). This improved capacity is then regarded as the spatial multiplexing gain. In order to fully exploit the spatial multiplexing gain provided by MIMO systems, different Bell-Labs layered space time (BLAST) architectures were developed from 1996 to 1998 (Foschini 1996, 1999; Wolniansky 1998). As an opposite performance metric, spatial diversity gain is used to judge the transmission quality of a MIMO system. To maximize the diversity gain, Alamouti proposed a space-time block code (STBC) in 1998 (Alamouti 1998). With the STBC method, a full diversity gain, i.e., $N_t N_r$, can be obtained. In general, BLAST architectures using different MIMO detection schemes are always applied to achieve a high spectral efficiency without losing too much diversity gain. Thus, MIMO detection has been widely considered a key technology in designing MIMO systems.

Foundations

A typical MIMO system is shown in Fig. 1, where the transmission model is given by:

$$\mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{n}. \tag{1}$$

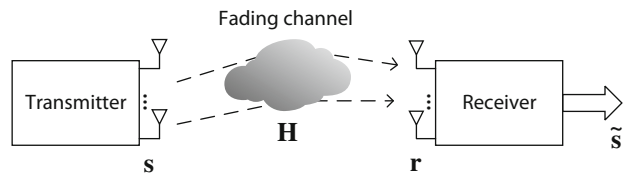
Here, $\mathbf{H}$, $\mathbf{s}$, and $\mathbf{n}$ denote the channel matrix, transmitted signal vector, and background noise, respectively. MIMO detection is to recover the transmitted signal at the receiver, where the estimated signal is denoted by $\tilde{\mathbf{s}}$ in Fig. 1.

MIMO detection in uncoded systems In general, to obtain an optimal bit-error rate (BER) performance with a full diversity gain in uncoded MIMO systems, maximum likelihood (ML) detection method can be carried out by employing exhaustive search, where the estimated signal vector is derived as $\tilde{\mathbf{s}} = \arg\min_{\mathbf{s}\in S^{N_t}} ||\mathbf{r} - \mathbf{H}\mathbf{s}||^2$. Here, S denotes the symbol alphabet. Due to the involved constellation searching, the computational complexity of the ML method grows exponentially with the number of transmit antennas, which makes it infeasible to be applied in a real MIMO system. Then, several low-complexity detection schemes have been widely studied in literatures (Wolniansky 1998; Bai 2012; Yao 2002; Liu 2011). Among the proposed sub-optimal methods, linear detectors are extensively applied in practice, such as zero-forcing (ZF) and minimum mean square error (MMSE) linear filters (Bai 2012):

- *ZF detectors.* The signal received by a linear detector is filtered by a well-designed matrix which enables the transmitted multiple signal to be detected separately without involving interferences. For the ZF method, the filter matrix can be designed by inversing the channel matrix, i.e., $\mathbf{W}_{ZF} = (\mathbf{H}^{\mathrm{H}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{H}}$. Then, the estimated signal vector is given by:

$$\begin{aligned}\tilde{\mathbf{s}} &= \mathbf{W}_{ZF}\mathbf{r} \\ &= \mathbf{s} + (\mathbf{H}^{\mathrm{H}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{H}}\mathbf{n}.\end{aligned} \tag{2}$$

However, as the equivalent noise becomes $(\mathbf{H}^{\mathrm{H}}\mathbf{H})^{-1}\mathbf{H}^{\mathrm{H}}\mathbf{n}$ in (2), the noise power would be amplified if the channel matrix is not orthogonal. Furthermore, the ZF detector



MIMO Detection, Fig. 1 An illustration of a MIMO system

cannot offer a satisfied BER performance when the signal-to-noise ratio is low.

- *MMSE detectors.* In order to alleviate the co-channel interference while restraining the impact of background noise at the same time, the MMSE method was proposed. In an MMSE detector, the filter matrix is developed on the criterion of minimizing the estimation error, i.e., $\mathbf{W}_{MMSE} = \arg\min \mathbb{E}[||\mathbf{s} - \mathbf{W}\mathbf{r}||^2] = \mathbf{H}(\mathbf{H}^{\mathrm{H}}\mathbf{H} + N_0/E_s\mathbf{I})^{-1}$. Here, N_0 and E_s denote the power of the noise and average signal, respectively. Compared with the ZF method, the filter matrix in MMSE approach can be designed by finding a balance between interference cancelation and noise amplification without incurring too much complexity.

Although the MMSE detector can improve the BER performance of ZF method, the noise will still be enhanced when filtering the co-channel interference. In Wolniansky (1998), a QR factorization-based successive interference cancelation (SIC) was proposed, where the detection of each signal can be carried out after removing the other estimated signals. Specifically, by multiplying $\mathbf{Q}^{\mathrm{H}}$, where $\mathbf{Q}$ is the unitary component in the QR factorization of $\mathbf{H}$, the signal vector at the receiver is derived as

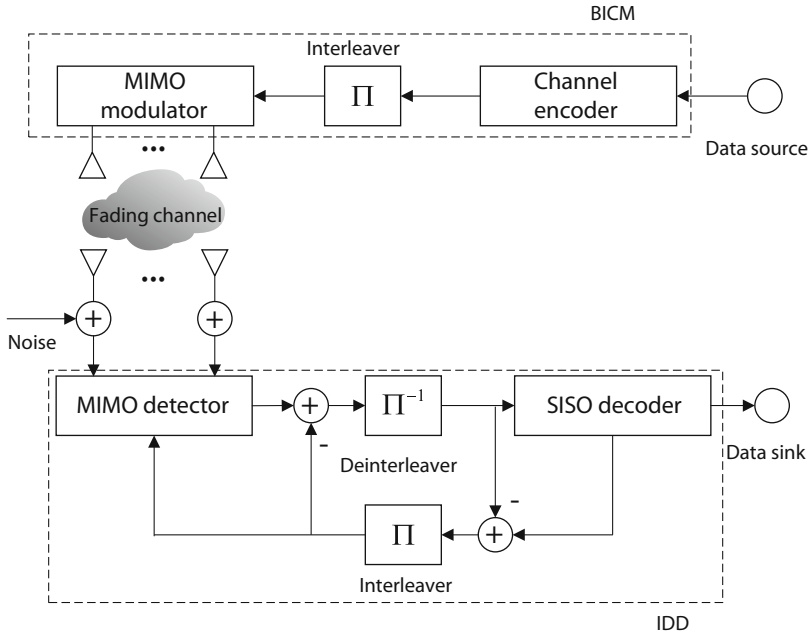
$$\begin{aligned}\mathbf{v} &= \mathbf{Q}^{\mathrm{H}}\mathbf{r} \\ &= \mathbf{Q}^{\mathrm{H}}\mathbf{Q}\mathbf{R}\mathbf{s} + \mathbf{Q}^{\mathrm{H}}\mathbf{n} \\ &= \mathbf{R}\mathbf{s} + \mathbf{Q}^{\mathrm{H}}\mathbf{n}. \end{aligned} \tag{3}$$

In (3), $\mathbf{R} = \mathbf{Q}^{\mathrm{H}}\mathbf{H}$. Since $\mathbf{R}$ is upper triangular, signals in $\mathbf{s}$ can be estimated from the N_tth layer to the first layer in a successive manner. In summary, the SIC detector outperforms the above linear detectors because the noise would not be amplified by multiplying a unitary matrix.

In order to achieve the same diversity provided by the ML method, Yao (2002) proposed a lattice reduction-aided MIMO detection method. From the point of view of lattice decoding, MIMO detection is to find the closest vector point in a lattice generated by the basis vectors of the channel matrix. Applying the Lenstra-Lenstra-Lovász algorithm (Lenstra 1982), a basis can be transformed into a nearly orthogonal one. Thus, the decoding radius is increased which helps to find the correct transmitted signal vector. Inspired by lattice decoding, Liu (2011) applied the Klein sampling algorithm Klein (2000) in MIMO detections. In this scheme, the searching area can be further extended by involving a randomized rounding instead of the conventional one.

MIMO detection in coded systems To improve both the channel capacity and BER performance, coded MIMO systems are widely considered, where the bit-interleaved coded modulation (BICM) is usually adopted at the transmitter (Caire 1998). At the receiver side, the iterative detection and decoding (IDD) architecture (Hochwald 2003) based on turbo principle is employed, which is shown in Fig. 2. With a soft-input soft-output (SISO) detector and a SISO decoder, the extrinsic information, i.e., log-likelihood ratio of each bit, is exchanged iteratively between them to obtain a good trade-off of the BER performance and complexity.

For MIMO detectors in IDD, the maximum a posteriori probability (MAP) method can provide an optimal BER performance. Unfortunately, since the MAP detector adopts the same exhaustive search with the ML method, the computational complexity becomes $O(2^{N_t}N_t^2)$ which also makes it infeasible to be implemented in a real MIMO system. In Wang (1999), an MMSE-based soft cancelation (SC) method was proposed to provide a sub-optimal performance with relatively low computational complexity. Since the MMSE-SC is developed on a symbol-wise linear filter, the offered BER performance is always taken as a benchmark when designing low-complexity sub-optimal methods. To improve the performance without incurring too much complexity, a bit-wise MMSE filter was studied in Li (2013). By exploring the distribution of the a posteriori probability of the transmitted signal vector, (Bai 2013) applied the Klein sampling in IDD and proposed an a priori information-based sampler. After that, Bai



MIMO Detection, Fig. 2 An illustration of a MIMO-BICM system

(2016) considered a large-scale antenna scenario and developed a Klein sampling-aided Markov chain Monte Carlo (MCMC) detector. In this scheme, the detection is carried out in a block-wise manner using the MCMC approach, while each block can be drawn with the Klein sampler.

Key Applications

MIMO detection is fundamental in wireless communication systems where the transmitter and the receiver are equipped with multiple antennas. Another typical application scenario can be found in multiuser detection.

Cross-References

- 5G Wireless
- Key Technologies in 4G/LTE Network
- Millimeter Wave Massive MIMO

References

Alamouti SM (1998) A simple transmit diversity technique for wireless communications. IEEE J Sel Areas Commun 16(8):1451–1458

Bai L, Choi J (2012) Low complexity MIMO detection. Springer, New York

Bai L, Choi J (2013) Lattice reduction-based MIMO iterative receiver using randomized sampling. IEEE Trans Wireless Commun 12(5):2160–2170

Bai L, Li T, Liu J et al (2016) Large-Scale MIMO detection using MCMC approach with blockwise sampling. IEEE Trans Commun 64(9):3697–3707

Caire G, Taricco G, Biglieri E (1998) Bit-interleaved coded modulation. IEEE Trans Inf Theory 44(3):927–946

Foschini GJ (1996) Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas. Bell Labs Tech J 1(2):41–59

Foschini GJ, Golden GD, Valenzuela RA et al (1999) Simplified processing for high spectral efficiency wireless communication employing multi-element arrays. IEEE J Sel Areas Commun 17(4):1841–1853

Hochwald BM, Brink ST (2003) Achieving near-capacity on a multiple-antenna channel. IEEE Trans Commun 51(3):389–399

Klein P (2000) Finding the closest lattice vector when it's unusually close. In: Proceedings of ACM-SIAM symposium on discrete algorithms, San Francisco, pp 937–941
Li Q, Zhang J, Bai L et al (2013) Lattice reduction-based approximate MAP detection with bit-wise combining and integer perturbed list generation. IEEE Trans Commun 61(8):3259–3269
Liu S, Ling C, Stehle D (2011) Decoding by sampling: a randomized lattice algorithm for bounded distance decoding. IEEE Trans Inf Theory 57(9):5933–5945
Lenstra AK, Lenstra HWJ, Lovász L (1982) Factoring polynomials with rational coefficients. Math Ann 261(4):515–534
Telatar E (1999) Capacity of multi-antenna Gaussian channels. Europ Trans Telecommun 10(6):585–595
Wang X, Poor HV (1999) Iterative (turbo) soft interference cancellation and decoding for coded CDMA. IEEE Trans Commun 47(7):1046–1061
Wolniansky PW, Foschini GJ, Golden GD et al (1998) V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel. In: Proceedings of 1998 IEEE ISSSE, Pisa, pp 295–300
Yao H, Wornell GW (2002) Lattice-reduction-aided detectors for MIMO communication systems. In: Proceedings of 2002 IEEE GLOBECOM, pp 424–428

MIMO Relay

Yue Rong
School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley, WA, Australia

Synonyms

MIMO relay channel; MIMO relay network

Definitions

Multiple-input multiple-output (MIMO) relay refers to a cooperative communication technology employing relay nodes, when one or more nodes of a relay system have multiple transmit/receive dimensions. MIMO relay communications can improve the reliability and extend the coverage of communication systems.

Historical Background

A relay channel, also known as three-terminal communication channel, was first investigated in van der Meulen (1971). The capacity of this channel was studied in 1979 (Cover and El Gamal 1979). During 1980s and 1990s, further information theoretic results on the relay channel were discovered (Vanroose and van der Meulen 1992). Since the turn of the century, the relay channel has attracted a renewed interest due to the boom of applications of wireless communications (Sendonaris et al. 2003).

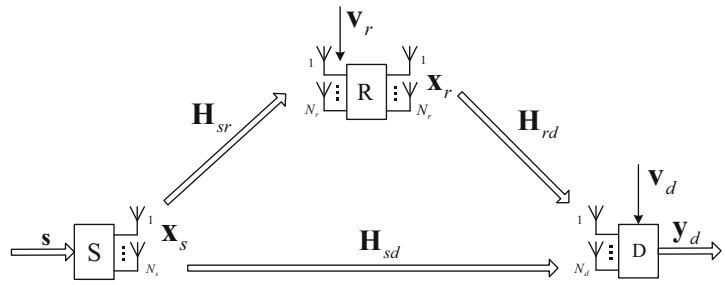
When multiple antennas are deployed at one or more nodes of the relay system, we call such relay system a MIMO relay channel. The achievable rate and capacity upper bound of a MIMO relay channel have been studied in Wang et al. (2005). A diversity-multiplexing trade-off of multi-antenna cooperative systems has been studied in Yuksel and Erkip (2007).

More applied issues of wireless relay system have been discussed in Sendonaris et al. (2003) in terms of user cooperation diversity, where the key idea is to enable multiple terminals (source and/or relays) to cooperate with each other to create a virtual antenna array that provides some form of spatial diversity. The authors of Sendonaris et al. (2003) showed that user cooperation is beneficial in terms of increasing system throughput and cell coverage, as well as decreasing sensitivity to channel variations.

Recent surveys on topics in *amplify-and-forward* (AF) MIMO relay systems can be found in Sanguinetti et al. (2012). A testbed implementation of MIMO relay system was reported in Maltsev et al. (2010). A general framework of optimizing the source and relay precoding matrices has been developed in Rong et al. (2009). There are special issues in *IEEE Journal on Selected Areas in Communications* on the topic of MIMO relay (Hua et al. 2012).

Foundations

The simplest MIMO relay system consists of three nodes as illustrated in Fig. 1, where one



MIMO Relay, Fig. 1 Block diagram of a dual-hop MIMO relay system

source node (S) transmits information to one destination node (D) with the aid of a relay node (R). Each node has multiple transmit/receive dimensions. Let us denote the number of dimensions at node i as N_i, $i \in \{s, r, d\}$. Mathematically the channel response from one node to another is represented through a matrix. As an example, the source-relay, relay-destination, and source-destination channels are represented by matrices $\mathbf{H}_{sr}$, $\mathbf{H}_{rd}$, and $\mathbf{H}_{sd}$, respectively. MIMO relay channel is an important extension of single-hop MIMO channel, which has been extensively studied in the last decade. Similar to a single-hop MIMO channel, the multiple transceiving dimensions (degrees of freedom) in a MIMO relay channel can be used to achieve various tradeoffs between diversity gain and multiplexing gain (Yuksel and Erkip 2007), where the diversity gain helps improving the system reliability, while the multiplexing gain is mainly used to increase the system throughput.

Since each hop of the relay system in Fig. 1 forms a MIMO channel, the input-output relationship from node $j \in \{s, r\}$ to node $i \in \{r, d\}$ is given by

$$\mathbf{y}_i = \mathbf{H}_{ji}\mathbf{x}_j + \mathbf{v}_i \tag{1}$$

where $\mathbf{x}_j$ is the transmitted signal vector at node j with a dimension of N_j and $\mathbf{y}_i$ and $\mathbf{v}_i$ are $N_i \times 1$ vectors of received signal and noise at node i, respectively.

Compared with single-hop MIMO systems, dual-hop or multihop MIMO systems provide more freedoms to system designers. In particular, for an optimal dual-hop relay system performance, both of the transmitted signal vectors $\mathbf{x}_s$ and $\mathbf{x}_r$ from two different nodes need to be optimized, and both of the received signal vectors $\mathbf{y}_r$ and $\mathbf{y}_d$ at two different nodes need to be optimally processed. On the other hand, such a flexibility also makes the optimization problems much more challenging than those for single-hop MIMO systems. In fact, most of the optimization problems in MIMO relay systems are highly nonconvex involving multiple matrix/vector variables. As a matter of fact, except for some upper bounds (Wang et al. 2005), the exact capacity of a MIMO relay channel is still not available.

A relay can be half-duplex or full-duplex. A half-duplex relay receives and transmits in two separate time/frequency channels. A full-duplex relay receives and transmits at the same time and same frequency (Cirik et al. 2014).

Relay Strategies

An important part of a relay system is known as relay strategy, which determines how the relay processes $\mathbf{y}_r$ to generate $\mathbf{x}_r$. Mathematically, the relay strategy can be represented by a function of $\mathbf{x}_r = \mathbf{f}(\mathbf{y}_r)$, where $\mathbf{f}$ represents the relay strategy. Interestingly, so far, there is no relay strategy that works best under all scenarios. Generally speaking, there are two main categories of relay strategies: *regenerative* relay and *non-regenerative* relay. In a regenerative strategy, the relay node first extracts out (decodes) the information from $\mathbf{y}_r$. Then, the relay node generates $\mathbf{x}_r$ by re-encoding the information. This strategy is also known as *decode-and-forward* (DF). A relaxed form of this strategy is called *compress-and-forward* where the signal received at the relay node is partially compressed before it is

forwarded. There are also other variations of this strategy (Kramer et al. 2005).

For a *non-regenerative* strategy, the relay node only amplifies (including a possible linear transformation) and retransmits its received signals, without attempting to decode the information-carrying symbols. Thus, a non-regenerative relay is also referred to as AF relay. The authors of Fan and Thompson (2007) compared the performance-complexity trade-offs of non-regenerative and other MIMO relay techniques. Both regenerative and non-regenerative strategies are affected by the noise at the relay node, although in different ways. The complexity and the processing delay of a non-regenerative strategy are generally much smaller than those of a regenerative strategy. The non-regenerative strategy is also believed by many to provide a better trade-off between benefits and implementation costs.

With a linear AF relay, we can in general write $\mathbf{x}_r = \mathbf{F}_r \mathbf{y}_r$, where $\mathbf{F}_r$ is the relay amplifying (precoding) matrix.

Key Applications

Relay nodes are needed in situations where the path loss between source and destination is too high, and/or the transmission power from the source is too limited by regulation and/or hardware constraints. Relay nodes are easy to install to extend the reach of a backbone network. Special relay nodes can be equipped with multiple antennas to compensate the limitation of most user terminals such as handsets. Cooperative relay nodes can form a virtual multi-antenna relay (Sendonaris et al. 2003), which increases the spatial diversity order of the source-relay-destination channel and thus improves the reliability of communication.

MIMO relay has applications in many physical forms of communication channels, such as:

- Multi-antenna multihop wireless networks, which are perhaps the most commonly referred to in the context of MIMO relays. Multihop wireless backhaul networks are being considered in several industry standards such as IEEE802.16j (Genc et al. 2008).
- Underwater acoustic relay channel. Since the bandwidth of underwater acoustic channel is inversely proportional to the transmission distance, MIMO relay techniques can enhance the capacity of underwater acoustic communication systems over a long distance (Al-Dharrab et al. 2013).
- Wireline DSL relay channel, which can be modeled as MIMO relay channel due to the crosstalk arising from the electromagnetic coupling between neighboring twisted-pairs.
- Powerline relay channel. Recently, cooperative relay communication technology has been adopted into the indoor powerline communication environment (Wu and Rong 2015).

Cross-References

▸ Network MIMO
▸ Relaying in LTE-Advanced

References

Al-Dharrab S, Uysal M, Duman TM (2013) Cooperative underwater acoustic communications. IEEE Commun Mag 51:146–153

Cirik AC, Rong Y, Hua Y (2014) Achievable rates of full-duplex MIMO radios in fast fading channels with imperfect channel estimation. IEEE Trans Signal Process 62:3874–3886

Cover TM, El Gamal AA (1979) Capacity theorems for the relay channel. IEEE Trans Inf Theory 25:572–584

Fan Y, Thompson J (2007) MIMO configurations for relay channels: theory and practice. IEEE Trans Wireless Commun 6:1774–1786

Genc V, Murphy S, Yu Y, Murphy J (2008) IEEE 802.16j relay-based wireless access networks: an overview. IEEE Wireless Commun 15:56–63

Hua Y, Bliss DW, Gazor S, Rong Y, Sung Y (2012) Guest editorial: theories and methods for advanced wireless relays – issue I. IEEE J Sel Areas Commun 30:1361–1367

Kramer G, Gastpar M, Gupta P (2005) Cooperative strategies and capacity theorems for relay networks. IEEE Trans Inf Theory 51:3037–3063

Maltsev A, Khoryaev A, Lomayev A, Maslennikov R, Antonopoulos C, Avgeropoulos K, Alexiou A, Boccardi F, Hou Y, Leung KK (2010) MIMO and

multihop cross-layer design for wireless backhaul: a testbed implementation. IEEE Commun Mag 48:172–179
Rong Y, Tang X, Hua Y (2009) A unified framework for optimizing linear nonregenerative multicarrier MIMO relay communication systems. IEEE Trans Signal Process 57:4837–4851
Sanguinetti L, D'Amico AA, Rong Y (2012) A tutorial on the optimization of amplify-and-forward MIMO relay systems. IEEE J Sel Areas Commun 30:1331–1346
Sendonaris A, Erkip E, Aazhang B (2003) User cooperation diversity – Part I and Part II. IEEE Trans Commun 51:1927–1948
van der Meulen EC (1971) Three-terminal communication channels. Adv Appl Prob 3:120–154
Vanroose P, van der Meulen EC (1992) Uniquely decodable codes for deterministic relay channels. IEEE Trans Inf Theory 38:1203–1212
Wang B, Zhang J, Høst-Madsen A (2005) On the capacity of MIMO relay channels. IEEE Trans Inf Theory 51:29–43
Wu X, Rong Y (2015) Joint terminals and relay optimization for two-way power line information exchange systems with QoS constraints. EURASIP J Adv Signal Process v2015:84
Yuksel M, Erkip E (2007) Multi-antenna cooperative wireless systems: a diversity multiplexing tradeoff perspective. IEEE Trans Inf Theory 53:3371–3393

MIMO Relay Channel

▸ MIMO Relay

MIMO Relay Network

▸ MIMO Relay

MIMO Satellite

▸ Satellite MIMO

MIMO Signal Demodulation

▸ MIMO Detection

MIMO Signal Equalization

▸ MIMO Detection

Mining Task Offloading in Wireless Blockchain Networks

Mengting Liu[1], Richard Fei Yu[2,3], Yinglei Teng[1], Victor C. M. Leung[4], and Mei Song[1]
[1]School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[2]Carleton University, Ottawa, Canada
[3]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada
[4]The University of British Columbia, Vancouver, BC, Canada

Synonyms

Computation offloading; Wireless blockchain networks

Definitions

Mining task offloading is a promising solution to address the computation-intensive mining tasks by offloading them to edge computing nodes using mobile edge computing technology.

Historical Background

Blockchain, the technology underpinning cryptocurrency, has been widely applied to many fields, such as security, smart cities, and supply chain UK Government (2016) and Iansiti and Lakhani (2017). However, the application of blockchain technology into wireless mobile networks is hindered by a main challenge brought by a computational process called

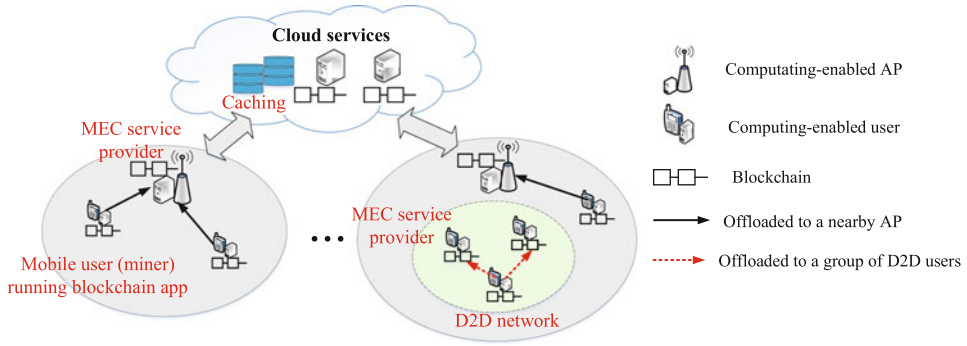
mining. Specifically, in order to confirm and secure the integrity and validity of transactions, the participants (a.k.a. miners) need to solve a computational difficult problem, i.e., the proof-of-work puzzle Beccuti (2017) and Kiayias et al. (2016), which sets an extremely high demand for the computational capabilities and storage availability of the miners.

To release the burden of the miners, *Mobile Edge Computing* (MEC) is considered as a promising solution that brings network resources closer to the end users, which can effectively reduce the energy consumption and shorten the transmission delay Stanciu (2013) and Xiong et al. (2017). As a key enabler of MEC, nearby resource-rich access points (APs) can process the users' requests as a substitute of clouds, which is also called cloudlet and able to reduce the transmission delay due to its proximity to users Mao et al. (2017b), Wang et al. (2016), and Yu et al. (2016). Nonetheless, due to its limited computation capacity and storage space, it is difficult for each AP to handle a large amount of requests from the users at a time. Moreover, AP computation offloading may lose its advantages when the wireless channel is in deep fading or the user is far away from the AP. To avoid the global network bottlenecks, device-to-device (D2D) computation offloading acts as an additional type of edge computing is playing an increasing important role in enhancing MEC performance by exploiting surplus computing resources in neighboring user devices Mao et al. (2017a).

Therefore, there are mainly two offloading modes. One is "offloaded to the nearby AP" where the powerful computing capability enabled by the APs can be leveraged to enhance the task offloading performance. The other is "offloaded to a group of nearby users" (a.k.a. multiuser cooperative edge computing) where a group of users can share their computational resources to aid the computation process through D2D computation offloading. Accordingly, in MEC-enabled wireless blockchain networks, the mobile miners can resort to the nearby edge nodes (APs or mobile users) to perform the computation-intensive tasks.

Foundations

Since blockchain in general are currently employed in wired networks, a framework of wireless blockchain network with MEC is necessary to perform the mining task offloading. Figure 1 gives a brief framework of MEC-enabled wireless blockchain networks Liu et al. (2018), where both the APs and mobile users have a computation capacity to provide task-execution services. Accordingly, the computational difficult mining task can be offloaded in two ways: (1) The miners can offload the full mining task to a nearby AP. (2) The miners can divide the whole mining task into several parts and separately forward them to a group of nearby mobile users through D2D links.



Mining Task Offloading in Wireless Blockchain Networks, Fig. 1 An illustration of MEC-enabled wireless blockchain networks

Applying the blockchain technology to wireless networks induces several issues related to the resource allocation and management, which is an ever-popular topic in wireless networks. The most representative ones include:

- Offloading scheduling Liu et al. (2018): For the miners in wireless blockchain networks, offloading scheduling scheme including the offloading mode selection and offloading task allocation among two offloading modes is crucial to guarantee the miners' demands as well as balance the computation load of edge computing nodes.
- Resource pricing Xiong et al. (2017): In blockchain-based networks, the price setting of resources, an indispensable part in the incentive mechanism, plays an important role in coordinating the wireless resource allocation.
- Content caching: For the miners, some information related with the block (like the cryptographic hashs of blocks) and computation results need to be stored for future request. The caching strategy of miners (where to cache and which part to cache) coupled with the storage space allocation is a key factor for wireless blockchain networks.

Key Applications

Mining task offloading provides an effective way to address the computational intensive mining tasks for miners, which can facilitate the application of blockchain technology into wireless networks.

Cross-References

- ▸ Integrated System of Networking, Caching, and Computing
- ▸ Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks
- ▸ Mobile Edge Computing: Low Latency and High Reliability

References

Beccuti J (2017) The bitcoin mining game: on the optimality of honesty in proof-of-work consensus mechanism. Tech. rep

Iansiti M, Lakhani KR (2017) The truth about blockchain the truth about blockchain. Tech. rep

Kiayias A, Koutsoupias E, Kyropoulou M, Tselekounis Y (2016) Blockchain mining games. In: Proceedings of the 2016 ACM conference on economics and computation, Maastricht, The Netherlands

Liu M, Yu FR, Teng Y, Leung VCM, Song M (2018) Joint computation offloading and content caching for wireless blockchain networks. In: Infocom'18 workshops, IEEE publication, Honolulu, HI

Mao Y, You C, Zhang J, Huang K, Letaief KB (2017a) A Survey on mobile edge computing: the communication perspective. IEEE Commun Surv Tutorials 19(4):2322–2358

Mao Y, Zhang J, SS H, LK B (2017b) Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems. IEEE Trans Wirel Commun 16(9):5994–5600

Stanciu A (2013) Blockchain based distributed control system for edge computing. In: 21st international conference on Control Systems and Computer Science (CSCS), IEEE publication, Bucharest, Romania

UK Government (2016) Distributed ledger technology beyond block chain. Tech. rep

Wang Y, Sheng M, Wang X, Wang L, Li J (2016) Mobile-edge computing: partial computation offloading using dynamic voltage scaling. IEEE Trans Commun 64(10):4268–4282

Xiong ZH, Zhang Y, Niyato D, Wang P, Han Z (2017) When mobile blockchain meets edge computing: challenges and applications. Comput Sci

Yu Y, Zhang J, Letaief KB (2016) Joint subcarrier and cpu time allocation for mobile edge computing. In: IEEE global communications conference, IEEE publication, Washington, DC

MIP

▸ Mobile IP

MIPv4

▸ Mobile IP

MIPv6

▸ Mobile IP

Mixed Network

Sanaa Taha[1] and Xuemin (Sherman) Shen[2]
[1]Information Technology Department, Faculty of Computers and Information, Cairo university, Cairo, Egypt
[2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

Synonyms

Digital pseudonyms; Mixnet; Return address; Unreachable electronic mail

Definitions

Mixed network, is a message routing protocol supporting sender and receiver anonymity by using a group of proxy servers, called mixes or stages. At each mix, a batch of inputs is mixing together and reordering to exit in a different order.

Introduction

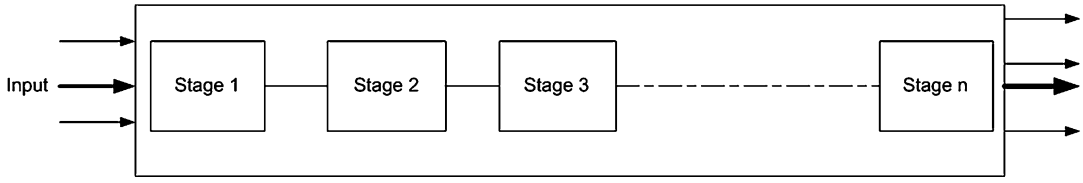
Mixed network, or mixnet, is a message routing protocol supporting sender and receiver anonymity by using a group of proxy servers, called mixes or stages. At each mix, a batch of inputs is mixing together and reordering to exit in a different order. The first mixnet protocol was proposed by David Chaum in 1981 to achieve anonymity in electronic voting (e-voting) application. Consequently, many anonymity networks have been proposed, such as Crowds (Reiter and Rubin 1998), MixMaster (Cornelius et al. 2008), Freenet (Clarke et al. 2000), Tarzan (Freedman and Morris 2002), Tor (Dingeldine et al. 2004), GNUnet (https://gnunet.org/), and I2P (https://geti2p.net/). Additionally, the mixnet also provides message integrity and verifiability. From performance perspective, the mixnet should acquire less network latency and high throughput. The goal of mixnet is either to support one (sender) or two-way (sender and receiver) anonymity.

In the context of wireless networks, many types of applications use mixnet routing, including e-voting, anonymous e-mails, and location privacy in wireless networks (Lu et al. 2009). For secure and error-free election, the e-voting system requires verification of voter eligibility, ballot integrity, and tally accuracy. For anonymous e-mails, mixnet should apply reliability to anonymous Internet communications to assure detecting any misuse of anonymous emails senders. Also, location privacy applications can be efficiently used with anonymous sender, such as Radio frequency identification (RFID) tags anonymity (Lin et al. 2010).

Applications employing mixnet routing protocol face passive and active attacks (Duddu and Samanta 2018). Passive attacks, also called traffic analysis, have the goal of correlating inputs to corresponding outputs at each mix stage, while active attacks, also called traffic manipulation, have the goal of controlling one or more mix stages to attack message integrity.

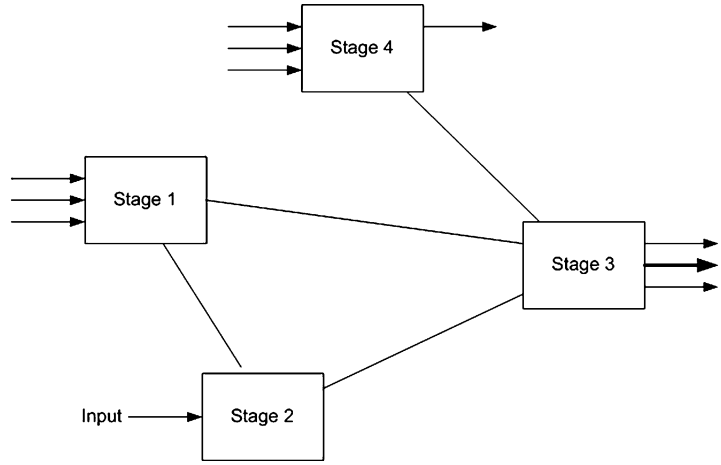
Mixnet Types

The topology of the mixnet can be cascade or free-route (Taha and Shen 2013), as illustrated in Figs. 1 and 2, respectively. On one hand, the mixes in the cascade mixnet are ordered in sequential and every batch of inputs should go through all mixes in the network. On the other hand, the free-route mixnet is not restricted to a fixed group of mixes and all inputs may traverse different paths in the network. Depending on mixing operation types, the mixnet implements one (or the two) of those topology types. Additionally, those mixing operations are categorized into decryption, hybrid, re-encryption,



Mixed Network, Fig. 1 Cascade mixnet topology

Mixed Network, Fig. 2
Free route mixnet topology



and universal re-encryption mixnet (Sampigethaya and Poovendran 2006).

In decryption mixnet, the sender uses each mix's public key to repeatedly encrypt the message, starting with the first mix to the receiver and ending with the first mix to the sender. In each encryption, the sender adds the next mix's address; hence, each mix knows the next mix's address only when decrypting the message. Only the last mix knows the receiver, x, of the message when decrypting the message. Equation (1) shows the encrypted message transferred from the sender, s.

$$E_K(m,r) = A_1 \| E_{K_1}\left(A_2 \| E_{K_2}\left(A_3 \| \ldots . E_{K_{n-1}}\left(A_n \| E_{K_n} \| \left(A_x \| E_{K_x}(m) \| r_n\right) \| r_{n-1}\right) \ldots . \| r_2\right) \| r_1\right) \quad (1)$$

where $E_K(m,r)$ is an encryption operation, El Gamal or RSA encryption, using a public key K, A_i is the address of a mix i, and r_i is a random number generated by a mix i. This type of mixing is not only used for sender anonymity, but it also can be used for two-way anonymous communications to achieve receiver anonymity. This can be achieved by adding to the previous equation the Return Path Information (RPI), which includes the return information encrypted using sender's public key, K_s. However, using many public key operations may limit the mixnet high-latency applications.

For hybrid mixing, the public key is used only to encrypt the shared keys, sk, between the sender and every mix in the network. The shared keys are used in encrypting the transmitted messages as illustrated in Eq. (2)

$$E_K(m,r,sk) = A_1 \| E_{K_1}(sk_1) \| E_{sk_1}\left[A_2 \| E_{K_2}(sk_2) \ldots . \| E_{sk_n}\left[A_n \| E_{K_n}\left(A_x \| E_{K_x}(m) \| r\right)\right] \ldots \ldots . .\right]\right] \quad (2)$$

For re-encryption mixnet type, the encrypted message does not have to go through all mixes, i.e., free-route mixnet is more suitable for this type. Additionally, and unlike other mixnet operation types, the length of the message does not

changed as the message goes through mixes. The reason of this fixed length message is that the sender uses an encryption key, which is a combination of all mixes' keys. In other words, Eq. (3) shows the encryption message produced at the sender

$$E_K(m, r) = \left(g^r \| (A_x \| m)\, K^r\right) \tag{3}$$

where g is the generator of a finite group Z_p, and K is the public key of the mixnet, calculated as follows:

$$K = \prod_{j=1}^{n} K_j = \prod_{j=1}^{n} g^{d_j} = g^{\sum_{j=1}^{n} d_j} \tag{4}$$

However, to achieve multicasting transmission, (i.e., transmitting to multiple receivers), all mixes in the re-encryption mixnet must share their key and perform decryption after mixing to get the address of the receiver. Such dependency can be avoided when using the universal re-encryption mixnet, where the sender transmits two encrypted messages: one for the original message and the other for the receiver's public key used to encrypt this original message, as illustrated in Eq. (5).

$$E_K(m, r) \| E_K\left(1, r'\right) = \left(g^r \| m K^r\right) \| \left(g^{r'} \| K^{r'}\right). \tag{5}$$

Table 1 shows a comparison between the different mixing types of mixnet

Degree of Anonymity

The degree of anonymity, d, is a measure of anonymity level achieved when applying the mixnet protocol, and this level varies depending on the supported applications. For instance, in e-voting, the voter's identity should be totally hidden, while in electronic payment system, seller's identity can be revealed by an authorized party in order to solve any dispute.

According to Reiter and Rubin (1998), the degree of anonymity is defined as: $d = 1 - p$, where p is an attacker probability about a potential sender. Additionally, Entropy is employed as a measure of the degree of anonymity as follows (Diaz 2006):

$$H(x) = \sum_{i=1}^{N} \left[p_i \log \frac{1}{p_i} \right],$$

where N is number of nodes inside the network and p_i is the probability associated with a node i. For a uniform distribution, all nodes have an equally likely probability of $\frac{1}{N}$ and Entropy records a maximum, H_M. Therefore, the degree of anonymity can be calculated as follows: $= \frac{H(x)}{H_M}$.

The effective anonymity set size, N, is another factor in measuring the degree of anonymity, which increases as the set size increases.

Mixnet-Based Wireless Applications

Shuffling protocol (Peng 2011) is employed to implement the mixnet, in which re-encryption followed by shuffling proof operations is applied on the group of ciphertext. Shuffling proof is a complex process, used to prove that the permutation of the input results the correct output. Shuffling protocol should support correctness, soundness, and zero knowledge privacy. Correctness means that shuffling inputs specific batches to get intended output, soundness means to verify the correctness of shuffling, and zero knowledge privacy means that it is impossible to reveal the permutation used in shuffling operation.

Another mixnet-based project is the Participatory sensing system (PSS) (Peng 2011), which is a data collection, processing, and dissemination system that is cost-effective and reliable. The system monitors a set of assets, called point of interest (POIs), by a group of entities, such as mobile nodes, which are participating to observe the assets and send their observed information to the application server. Based on the idea of mixnet, if the observer senses the attribute of POI, it sends an anonymous observation report to the application server. In order to anonymously send the report, the observer entity transmits the report

Mixed Network, Table 1 Mixnet types comparison

	Decryption	Hybrid	Re-encryption	Universal re-encryption
Mix type	Cascade	Cascade	Cascade and free route	Free route
Multiple receiver	Not applied	Not applied	Could be applied	Could be applied
Anonymity	One-way, could be two-way by adding RPI	One-way, could be two-way by adding RPI	Two-way is possible	One-way only
Message Integrity	Using public-key encryption	Using public and symmetric keys	Using El Gamal encryption	Using El Gamal encryption
Verifiability (Puiggali 2010; Chen 2007)	Using public key verification techniques	Using public key verification techniques	Using El Gamal verifications	Using El Gamal verifications
Fault Tolerance	Not applied	Not applied	applied	applied
Latency	High	Higher	Medium	Low
Throughput	lower	Low	Medium	High
Scalability	Not applied	Not applied	Could be enhanced	Could be enhanced
Efficiency	Low	Low	Medium	Medium

to another entity in the network and identifies specific number of hops for the message to circulate before receiving to the application server.

The Onion Routing (Tor) (Burke et al. 2006) is another predecessor for mixnet to achieve anonymous communications, in which the message has layers of encryption, and one layer is pealed at a time as the message traverses through Tor volunteering servers. However, Tor faces obfuscations of limited capacity and performance due to relaying on volunteering servers. In (The Tor project 2003), a Software-Defined Networking (SDN)-based Onion Routing (SOR) is proposed as a novel Cyber anonymity using Tor (Elgzil et al. 2017). The goal is to leverage the large capacities, strong connectivity, and economies of scale inherent to commercial data centers, by building onion routing tunnels over anonymity service providers.

Cross-References

▸ Proxy Mobile IPv6

References

Burke J, Estrin D, Hansen M, Parker A, Ramanathan N, Reddy S, Srivastava MB (2006) Participatory sensing. In: Proceedings of ACM conference on embedded networked sensor systems, Boulder, Colorado, USA

Chaum D (1981) Untraceable electronic mail, return addresses, and digital pseudonyms. Commun ACM 4(2):84–88

Chen J (2007) Verifiable mixnets techniques and prototype implementation. Master's thesis, Darmstadt University of Technology

Clarke I, Sanberg O, Wiley B, Hong TW (2000) Freenet: a distributed anonymous information storage and retrieval systems. In: Proceeding of the ICSI workshop on design issues in anonymity and unobservability, Springer

Cornelius C, Kapadia A, Kotz D, Peebles D, Shin M, Triandopoulos N (2008) Anonysense: privacy-aware people-centric sensing. In: Proceeding of the 6th international conference on mobile systems, applications, and services, Breckenridge, Colorado, pp 211–224

Diaz C (2006) Anonymity metrics revisited. In: Dagstuhl seminar proceedings, IBFI, Schloss Dagstuhl, Germany. http://drops.dagstuhl.de/opus/volltexte/2006/483

Dingeldine R, Mathewson N, Syverson P (2004) Tor: the second-generation onion router. Naval Research Lab, Washington, DC

Duddu V, Samanta D (2018) Network and security analysis of anonymous communication networks. arXiv preprint arXiv:1803.11377. Available online, https://arxiv.org/abs/1803.11377

Elgzil A, Chow CE, Aljaedi A, Alamri N (2017) Cyber Anonymity based on software-defined networking and onion routing. In: Proceeding of IEEE conference on dependable and secure computing, Taipei, pp 358–365

Freedman MJ, Morris R (2002) Tarzan: a peer-to-peer anonymizing network layer. In: Proceeding of the 9th ACM conference on computer and communications security, Washington, DC, USA

GNUnet. Available online, https://gnunet.org/
I2P. Available online, https://geti2p.net/
Lin X, Lu R, Kwana D, Shen X (2010) REACT: an RFID-based privacy-preserving children tracking scheme for large amusement parks. Comput Networks (Elsevier) 54(15):2744–2755
Lu R, Lin X, Zhu H, Ho PH, Shen X (2009) A novel anonymous mutual authentication protocol with provable link-layer location privacy. IEEE Trans Veh Technol 58(3):1454–1466
Peng K (2011) Survey, analysis and re-evaluation – how efficient and secure a mix network can be. In: Proceeding of IEEE 11th international conference on Computer and Information Technology (CIT), Paphos, Cypru, pp 249–254
Puiggali J (2010) Universally verifiable efficient re-encryption mixnet. Krimmer and Grimm [32], pp 241–254
Reiter M, Rubin A (1998) Crowds: anonymity for web transactions. ACM Trans Inf Syst Secur 1(1):66–92
Sampigethaya K, Poovendran R (2006) A survey on mix networks and their secure applications. In: Proceeding of the IEEE, vol 94, No. 12, Dec 2006
Taha S, Shen X (2013) ALPP: anonymous and location privacy preserving scheme for mobile IPv6 heterogeneous networks. Secur Commun Networks 6(4):401–419
The Tor project (2003.) https://www.torproject.org/. Accessed Nov 2015

Mixnet

▸ Mixed Network

mmWave Channel Campaigns

▸ Millimeter Wave Channel Measure

mmWave Channel Measurement

▸ Millimeter Wave Channel Measure

mmWave Channel Sounding

▸ Millimeter Wave Channel Measure

Mobile Access Gateway

▸ Proxy Mobile IPv6

Mobile Big Data

▸ Data-Driven Mobile Social Networks

Mobile Caching

▸ Wireless Edge Caching

Mobile Cloud Computing

▸ Mobile Edge Computing: Low Latency and High Reliability

Mobile Content Delivery

▸ Wireless Edge Caching

Mobile Content Delivery Architecture

▸ Mobile Content Distribution Architecture

Mobile Content Distribution Architecture

Junfeng Xie
School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China

Synonyms

Content delivery architecture in mobile cellular network; Content distribution architecture in

mobile cellular network; Mobile content delivery architecture

Definition

Mobile content distribution architecture is a promising technique to optimize the mobile content delivery and decrease duplicate transmission by offering flexible storage capability in mobile network, which can utilize network resources in a more efficient manner and improve end users' Quality of Experience (QoE).

Foundations

In recent years, to reduce mobile core network traffic and deliver content efficiently over the cellular network, many content distribution architectures have been proposed. Mobile content delivery network (CDN), mobile edge computing (MEC), and information-centric networking (ICN) are the representative architectures. So in this section, we briefly present the research on the three emerging content distribution techniques.

Mobile CDN

In the past decade, CDN has become the most popular and widely used content distribution architecture in the current wired Internet. Today, a large fraction of Internet traffic is delivered by CDN. Although CDN works very well in wired networks, it cannot be deployed in mobile networks directly. CDN is not suitable for mobile networks for two reasons. Firstly, CDN systems just help clients to select the "best" server at the initial stage and do not consider clients' mobility. Secondly, DNS-based schemes are used in traditional CDN systems to select cache servers and perform load balancing. Nevertheless, in a highly volatile mobile environment, user equipment (UE)'s location and IP address can change over time rapidly, which makes the address not always truly reflect the position of a UE. Thus, how to deploy CDN in mobile network becomes an interesting research direction.

One feasible mobile CDN implementation solution to address the scalability and mobility issues of the mobile networks is jointing distributed mobility management (DMM) and CDN (Liebsch and Yousaf 2013; Munaretto et al. 2014). DMM has been proposed by the Internet Engineering Task Force (IETF) as a new paradigm for flat networks in 5G mobile architectures to enable IP address continuity in DMM. Other mobile CDN implementation solutions have also been studied. Literatures (Dramitinos et al. 2013; Pentikousis et al. 2013) utilize the emerging software defined networking (SDN) approach to increase the operators' innovation potential and support video services more effectively.

MEC

With the evolution of mobile base stations and the emergence of cloud computing technology, edge computing as a novel paradigm has attracted widespread attention. In recent years, a few edge computing architectures have been presented, such as Cloudlet, Edge Computing, Fog Computing, Mobile Cloud Computing (MCC), Mist Computing, and MEC (Mach and Becvar 2017). Among all these edge computing architectures, MEC is the most popular architecture. MEC has been proposed by the standards institute ETSI (European Telecommunications Standards Institute) as a promising technology to reduce the load of mobile core network by shifting computational capability from the core network to the mobile edge (i.e., the radio access network). The key characteristics of MEC are proximity, lower latency, and location awareness. The MEC white paper has been issued in 2014 (Patel et al. 2014). MEC is an emerging technology, which is very likely to be applied in 5G mobile networks. Thus, there is a trend of integrating content distribution with MEC to improve content delivery efficiency and reduce content delivery delay.

ICN

Over the past several years, the research on the content distribution architecture is from two approaches: "clean slate" and "evolutionary".

Mobile CDN and MEC all belong to the "evolutionary" approaches. However, ICN (Xylomenos et al. 2014; Xie et al. 2017c) is the representative of "clean slate" approaches. The key concept of ICN is the content-centric communication model instead of the end-to-end communication model in traditional networks. So far a few ICN architectures have been proposed, such as Data-Oriented Network Architecture (DONA), Publish Subscribe Internet Technology (PURSUIT), Architecture and Design for the Future Internet (4WARD), Scalable and Adaptive Internet Solutions (SAIL), and Named-Data Networking (NDN). Although the concept of ICN has already attracted great attention in academia and industry, the deployment of ICN needs to update all network elements, user devices, as well as origin servers to be ICN-aware. In this case, the pure commercial ICN deployment would not be done in the near future. So literatures (Veltri et al. 2012; Vahlenkamp et al. 2013; Chanda and Westphal 2013; Chang et al. 2014) have studied the issue on how to deploy the ICN architectures in the legacy IP network.

Future Directions

Despite the potential vision of these mobile content distribution architectures, many significant research challenges remain to be addressed by future research efforts. In this section, we discuss some of these challenges and present some future research directions.

Adaptive Video Content Distribution over Cellular Networks

In mobile cellular networks, due to the varying wireless network conditions and the diversity of end-user devices, different users should be served with different data rates to enhance the users' QoE. Researchers have proposed various techniques to provide better QoE for mobile users by adapting multimedia content over wireless channels real-timely. Recently, the Scalable Video Coding (SVC) (Xie et al. 2017b) and HTTP Adaptive Bit Rate (ABR) streaming (Xie et al. 2017a) have been considered as the main solutions. In SVC, each video is encoded into one mandatory base layer and several optional enhancement layers. The enhancement layers build upon the base layer and provide multiple qualities. In HTTP ABR streaming, a video is divided into a series of small video chunks with a typical time duration of 2–10 s, and each video chunk is encoded into multiple bitrates (i.e., qualities). A client can adaptively request a video content with different qualities according to its wireless network condition. Caching adaptive video content in Radio Access Network (RAN) is a promising way to provide better content delivery service and to improve content distribution efficiency (Xie et al. 2016). In this scenario, how to allocate limited cache resource for adaptive video contents has become an important issue. Therefore, it is necessary to design effective cache resource allocation schemes to optimize the adaptive video content placement.

HTTPS Traffic Processing Strategies

Most of the current works on content distribution solutions are studied especially for HTTP traffic. With increased emphasis on privacy, security, and regulatory compliance, organizations and companies are more and more dependent on encryption technologies to protect information, among which TLS/SSL have become the standard of choice for providing secure applications. For instance, TLS/SSL are used by HTTPS to ensure secured transactions. On the other hand, with the rapid development of processing technologies, the cost to encrypt data is falling rapidly. Unfortunately, once an encrypted connection between a client and a server is established, it is difficult to manage, accelerate, or audit activity due to the private nature of the TLS/SSL tunnel. Therefore, it is necessary to study how to solve the HTTPS traffic processing problem in content distribution framework in the future.

Combining with Network Function Virtualization and Cloud Computing

With the rapid development of processing and virtualization technologies, computing resources

have become cheaper, more powerful, and more ubiquitously available than ever before. Network Function Virtualization (NFV) is a promising technology which decouples network functions from the underlying specialized hardware to make the network architecture more flexible and open. The NFV and cloud computing technologies draw ISPs' attention because they can reduce operators' capital expenditures for scaling up the network and make network reconfiguration quick and adaptive. Therefore, the deployment of content distribution architecture combining with NFV and cloud computing technologies is another future research direction.

Key Applications

Mobile content distribution architectures can be deployed in mobile networks to provide better content delivery service and improve end users' QoE.

Cross-References

- Enhancing QoE-aware Wireless Edge Caching with Software-defined Wireless Networks
- Integrated System of Networking, Caching, and Computing
- Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks
- Quality of Experience for Wireless Video Streaming
- Wireless Video Delivery
- Wireless Video Streaming

References

Chanda A, Westphal C (2013) ContentFlow: mapping content to flows in software defined networks. CoRR abs/1302.1493

Chang D, Kwak M, Choi N, Kwon T, Choi Y (2014) C-flow: an efficient content delivery framework with OpenFlow. In: Proceedings of the IEEE ICOIN'14, Phuket

Dramitinos M, Zhang N, Kantor M, Costa-Requena J, Papafili I (2013) Video delivery over next generation cellular networks. In: Proceedings of the IEEE CNSM'13, Zurich

Liebsch M, Yousaf FZ (2013) Runtime relocation of CDN serving points enabler for low costs mobile content delivery. In: Proceedings of the IEEE WCNC'13, Shanghai

Mach P, Becvar Z (2017) Mobile edge computing: a survey on architecture and computation offloading. IEEE Commun Surv Tutorials 19(3):1628–1656

Munaretto D, Giust F, Kunzmann G, Zorzi M (2014) Performance analysis of dynamic adaptive video streaming over mobile content delivery networks. In: Proceedings of the IEEE ICC'14, Sydney

Patel M, Naughton B, Chan C, Sprecher N, Abeta S, Neal A (2014) Mobile-edge computing introductory technical white paper. White paper, Mobile-edge Computing (MEC) industry initiative

Pentikousis K, Wang Y, Hu W (2013) Mobileflow: toward software-defined mobile networks. IEEE Commun Mag 51(7):44–53

Vahlenkamp M, Schneider F, Kutscher D, Seedorf J (2013) Enabling information centric networking in IP networks using SDN. In: Proceedings of the IEEE SDN4FNS'13, Trento

Veltri L, Morabito G, Salsano S, Blefari-Melazzi N, Detti A (2012) Supporting information-centric functionality in software defined networks. In: Proceedings of the IEEE ICC'12, Ottawa

Xie J, Xie R, Huang T, Liu J, Yu FR, Liu Y (2016) Caching resource sharing in radio access networks: a game theoretic approach. Front Inf Technol Electron Eng 17(12):1253–1265

Xie J, Xie R, Huang T, Liu J, Liu Y (2017a) Energy-efficient cache resource allocation and QoE optimization for HTTP adaptive bit rate streaming over cellular networks. In: Proceedings of the IEEE ICC'17, Paris, France, pp 1–6

Xie J, Xie R, Huang T, Liu J, Liu Y (2017b) Energy-efficient content placement for layered video content delivery over cellular networks. In: Proceedings of the IEEE GLOBECOM'17, Singapore, pp 1–6

Xie J, Xie R, Huang T, Liu J, Liu Y (2017c) ICICD: an efficient content distribution architecture in mobile cellular network. IEEE Access 5:3205–3215

Xylomenos G, Ververidis C, Siris V, Fotiou N, Tsilopoulos C, Vasilakos X, Katsaros K, Polyzos G (2014) A survey of information-centric networking research. IEEE Commun Surv Tutorials 16(2):1024–1049

Mobile Crowdsensing

- Private Truth Discovery in Cloud-Empowered Crowdsensing Systems

Mobile Crowdsensing: Issues and Challenges

Yang Liu
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

Synonyms

Crowdsensing; Crowd sensing

Definitions

Mobile crowdsensing can be referred to as crowd sensing or crowdsensing, which is from the idea of crowdsourcing. Crowdsourcing comes from "wired" magazine which describes a novel distributed problem-solving and working model in which companies allocate tasks through the Internet, find ideas, or solve technological problems (Ganti et al. 2011). Recently, the idea of crowdsourcing has been combined with mobile sensing and uses the mobile devices as basic sensing units. The network forms a crowdsensing network, which realizes the distribution of sensing tasks and the collection of data.

Historical Background

At present, the Internet of Things has developed deeply, and the need for physical environment sensing is very thorough. With the development of wireless communication, smartphones, tablets, wearable devices, and mobile terminals such as in-vehicle sensing devices integrate more and more sensors. With the increase number of wireless mobile devices, more than a dozen people have been developing how to use specific consciously deployed sensors to provide sensing services. Also, the Internet of Things will provide more large-scale, complex, and comprehensive sensing services, thus entering a new era of development.

Foundations

In traditional networks, people usually consume sensing data. In crowdsensing, however, people serve both as a "consumer" of sensing data and as a "producer" of sensing data, using a popular new word, which is called "prosumer." This basic human-centered feature brings unprecedented opportunities for IoT sensing and transmission. The specific performance is as follows:

- The costs of network deployment are lower. First, mobile devices do not require special deployment. Second, people help data sensing and transmission. As users who carry the devices arrive at different places, they can sense anytime and anywhere; on the other hand, due to the contacts between mobile users, they can use "storage-carry-forward" opportunistic transmission mode in an intermittently connected environment (Liu et al. 2018).
- Network maintenance is simpler. First, the users usually have better energy supply, computing, storage, and communication capabilities. Second, the mobile devices are usually kept by their holders, so they are usually in a better working state. For example, people can always charge mobile devices when their mobile phones are needed.
- The crowdsensing system is large scale. We need to recruit more users to participate in the system when there are more tasks.

Due to the above advantages, the crowdsensing network has become a new and important way to sense for the Internet of Things.

A typical crowdsensing is usually composed of a sensing platform and multiple users. Multiple servers are in the platform; the users can utilize various sensors embedded in the smart phones, e.g., GPS, accelerometer, gravity sensor, gyroscope, electronic compass, light distance sensor, microphone, cameras, etc. , to collect various data and move through cellular networks and report the data. The workflow is as follows:

- Task publishing: data demanders publish tasks on the platform. There are two methods to assign tasks: pull-based method and push-based method. Pull-based method refers to that participants browse and retrieve tasks from the platform and actively register as workers of a task; push-based method refers to that the platform finds suitable candidates from all participants. No matter which task assignment method is adopted, when the participants accept the task, they will become a worker of the task;
- Task execution: one worker arrives at the designated place according to the task requirements and uses a specific application which automatically saves sensor data according to the task requirements. For real-time sensing tasks, workers must upload data immediately; for nonreal-time sensing tasks, workers are allowed to choose the most economical communication mode to upload data before the task deadline;
- Data aggregation: affected by the distributed sensing mode, there are a lot of redundant, i.e., low quality or repetitive data. In the data aggregation stage, the platform filters and optimizes the data according to the task requirements and selects the high-quality data set that meets the task requirements;
- Result handover: data demanders can download data aggregation results from the platform before or after the end of the task. Most mobile crowdsensing tasks are difficult to define data collection constraints accurately at the very beginning. Therefore, the platform allows data demanders to see data aggregation results before the end of the task, so that the data demander has the opportunity to adjust the task requirements before the end of the task.

Key Applications

Mobile crowdsensing can be applied in many important fields, such as intelligent transportation, public transportation, tourism encyclopedia, indoor positioning, information dissemination, news reporting, and event sensing. Some typical applications are listed below:

- Intelligent transportation. Gazetiki (Popescu et al. 2008) is an application that uses Wikipedia and web search to build a geographic encyclopedia. Gazetiki first identifies all geographical names and their coordinates through Wikipedia and classifies and sorts them. When users search geographic information, Gazetiki shows them detailed geographic information and relevant photos. ParkNet (Mathur et al. 2009) uses GPS and the ultrasonic sensor installed on the right door to detect the empty parking space and share the detection results; the literature in (Zhou et al. 2012) designs the bus arrival time prediction system under Android platform.
- Shopping. Mobishop (Sehgal et al. 2008) collects people's shop ticket photos, extracts commodity prices through OCR character recognition technology, and pushes the collected commodity price information to different customers, so that people can share commodity price information.
- Entomology. Lostadybug (Losey et al. 2012) is a project to collect Ladybug photos, which is used to study Ladybug species, living conditions, and habits. As of March 2017, the website has collected 38,000 data from all over the world.
- Environmental monitoring. Creekwatch (Kim et al. 2011) is developed by IBM to monitor river water quality. When people pass a stream or river, the creekwatch application will submit the status of the river in its current location, the status includes the amount of water and waste, and it will also submit some photos. Creekwatch shows people the status of rivers around the world through a website.
- Emergency aid. Wreckwatch (White et al. 2011) can detect the accident through the sensor of mobile phone in the car and then find passersby to take photos of the scene of the accident and send them to the rescue center, so that the rescue workers can judge the location of the accident and the emergency degree of the rescue timely and accurately.

- News reporting. When there is a special event, Mediascope (Jiang et al. 2013) will retrieve all kinds of photos from the mobile photo album shared by the witnesses on the scene for the purpose of news reporting. The retrieval conditions include image similarity and geographic information.
- Information dissemination. Fliermeet (Guo et al. 2014) collects photos of public posters in the city; calculates the relationship among people, posters, and places; constructs the preferences of different individuals for different types of posters; pushes the posters they like or need to different people; and improves the communication efficiency of urban information.
- Enjoying flowers. SakuraSensor (Morishita et al. 2015) judges the sakura blossom situation through the video clips people take in the car during the sakura blossom season and calculates the optimal route to visit the cherry blossom beauty according to people's driving path.
- Indoor locating and navigating. An image-aided positioning system called Argus (Xu et al. 2015) is based on mobile terminals. Argus extracts geometric constraints from the image data of crowdsourcing and maps these geometric constraints to RSS fingerprint space to distinguish the location information of similar fingerprints and reduce the fuzziness of fingerprints.

Problems and Challenges Faced by Crowdsensing

Crowdsensing provides a new way to realize in-depth sensing but also brings new challenges to the research of theory, technology, and application. It can be summarized as the following aspects:

- The efficient transmission of crowdsensing data. Many crowdsensing applications need to continuously collect sensory data and transmit it to the data center. Reporting the sensing data based on the connection between the mobile cellular network and the Internet will consume too much user equipment power and data traffic, resulting in greater pressure for the mobile cellular network. Therefore, it is necessary to design energy- efficient data transmission methods, such as based on short-range wireless communication.
- Resource optimization of crowdsensing networks. Overcoming the resource constraints of mobile nodes in terms of energy, bandwidth, computing, etc. is the key to the practical application of crowdsensing networks. First, because the number of users and the availability of sensors can change dynamically over time, it is difficult to accurately model and predict energy and bandwidth requirements to complete a specified task. Second, it is necessary to consider how to select a valid subset of users and reasonably schedule the sensing and communication resources under resource constraints.
- Incentive mechanisms of crowdsensing networks. A crowdsensing application relies on a large number of ordinary users to participate, and users will consume their own equipment power, computing, storage, communication, and other resources and bear the threat of privacy leakage when participating in sensing. Therefore, a reasonable incentive mechanism is needed to pay for user participation. The cost is compensated to attract enough users to ensure the quality of the data collection required.
- User privacy protection. By collecting the sensing data which is related to a user's position, the accurate position information of the user can be obtained (Liu et al. 2019b). Through long-term monitoring and analysis of location information, the user's home and work addresses, daily activity scope, and common traffic routes can be found. Mining the sensing data of motion state sensor can obtain the sensitive information of a user's daily life habit, health status, and so on; combining with the environment sensing data, it can also get the user's situation at any time. Although the large-scale collection

of sensitive sensing data of mobile intelligent terminals can analyze and mine many valuable information, such as urban traffic congestion, road conditions, air quality, environmental pollution, infrastructure conditions, public services, etc., once these sensing data is leaked, it will seriously threaten the privacy of users, so it is necessary to take effective measures to protect users' privacy.

- Data and platform security. Since the mobile crowdsensing network gathers a large number of sensitive and private data and can mine a large number of information with great application value, it will greatly increase the risk of hacker attack and confidential data disclosure (Liu et al. 2019a). How to better protect the security of data and platform has become a prominent and urgent key issue.
- Data authenticity and integrity. To build an efficient and valuable mobile crowdsensing network depends on users to provide authentic, complete, and reliable sensing data. However, for collecting sensing data, malicious users may submit false sensing data; or, for sensing data transmission, they may lose some data, which may lead to mining out wrong information and making wrong decisions. Therefore, it is very important to take corresponding technical measures to ensure the authenticity and integrity of data.
- The balance between the improvement of sensing quality and efficiency and the optimal utilization of resources. Due to the large number of sensors and the dynamic change of performance, the task allocation and resource consumption prediction under resource constraints become more complex and difficult (Liu et al. 2017). We can collect the sensing data of different types of sensing devices to achieve the same sensing goal, but the sensing quality and resource consumption are different. For example, location information can be mined by collecting sensing data from GPS, Wi-Fi, and mobile cellular networks. GPS has the highest location accuracy and resource consumption. Therefore, the balance between sensing quality and resource consumption needs to be solved. In addition, it is possible to execute multiple priorities and different resource consumption sensing tasks on the same sensing device at the same time.

Reference

Ganti RK, Ye F, Lei H (2011) Mobile crowdsensing: current state and future challenges. IEEE Commun Mag 49(11):32–39

Guo B, Chen H, Yu Z, Xie X, Huangfu S, Zhang D (2014) FlierMeet: a mobile crowdsensing system for cross-space public information reposting, tagging, and sharing. IEEE Trans Mob Comput 14(10):2020–2033

Jiang Y, Xu X, Terlecky P, Abdelzaher T, Bar-Noy A, Govindan R (2013) Mediascope: selective on-demand media retrieval from mobile devices. In: Proceedings of the IPSN

Kim S, Robson C, Zimmerman T, Pierce J, Haber EM (2011) Creek watch: pairing usefulness and usability for successful citizen science. In: Proceedings of the CHI

Liu Y, Wu H, Xia Y, Wang Y, Li F, Yang P (2017) Optimal online data dissemination for resource constrained mobile opportunistic networks. IEEE Trans Veh Technol 66(6):5301–5315

Liu Y, Quan W, Wang T, Wang Y (2018) Delay-constrained utility maximization for video Ads push in mobile opportunistic D2D networks. IEEE Internet Things J 5(5):4088–4099

Liu Y, Hao L, Liu Z, Sharif K, Wang Y, Das SK (2019a) Mitigating interference via power control for two-tier femtocell networks: a hierarchical game approach. IEEE Trans Veh Technol 68(7):7194–7198

Liu Y, Wang H, Peng M, Guan J, Xu J, Wang Y (2019b) DeePGA: a privacy-preserving data aggregation game in crowdsensing via deep reinforcement learning. IEEE Internet Things J. https://doi.org/10.1109/JIOT.2019.2957400

Losey J, Allee L, Smyth R (2012) The lost ladybug project: citizen spotting surpasses scientist's surveys. Am Entomol 58(1):22–24

Mathur S, Kaul S, Gruteser M, Trappe W (2009) ParkNet: a mobile sensor network for harvesting real time vehicular parking information. In: Proceedings of the MobiHoc Workshop

Morishita S, Maenaka S, Nagata D, Tamai M, Yasumoto K, Fukukura T, Sato K (2015) SakuraSensor: quasi-realtime cherry-lined roads detection through participatory video sensing by cars. In: Proceedings of the UbiComp

Popescu A, Grefenstette G, Moëllic PA (2008) Gazetiki: automatic creation of a geographical gazetteer. In: Proceedings of the 8th ACM/IEEE-CS joint conference on digital libraries

Sehgal S, Kanhere SS, Chou CT (2008) Mobishop: using mobile phones for sharing consumer pricing informa-

tion. In: Proceedings of the international conference on distributed computing in sensor systems
White J, Thompson C, Turner H, Dougherty B, Schmidt DC (2011) Wreckwatch: automatic traffic accident detection and notification with smartphones. Mob Netw Appl 16(3):285
Xu H, Yang Z, Zhou Z, Shangguan L, Yi K, Liu Y (2015) Enhancing Wifi-based localization with visual clues. In: Proceedings of the UbiComp
Zhou P, Zheng Y, Li M (2012) How long to wait? Predicting bus arrival time with mobile phone based participatory sensing. In: Proceedings of the MobiSys

Mobile Data Offloading Techniques in WiFi-Cellular Networks

▸ WiFi Offloading in WiFi-Cellular Networks

Mobile Data Offloading via D2D-Assisted Communications in Wireless Networks

Yuan Wu[1,2], Kejie Ni[1], Xiaowei Yang[1], and Liping Qian[1]
[1]College of Information Engineering, Zhejiang University of Technology, Hangzhou, China
[2]State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao SAR, China

Technical Background

The past decades have witnessed the proliferation of smart wireless devices and mobile Internet services, yielding a rapid growth of the data-hungry mobile applications such as video surveillance, online gaming, and augmented/virtual reality. These applications have resulted in a tremendous traffic pressure in radio access networks (RANs) of wireless networks due to the limited radio resources. Conventional wireless infrastructures in RANs require to deliver the mobile users' data directly through the infrastructures, e.g., marco/micro base stations (BSs) and Wi-Fi access points (APs), which thus lead to a tremendous traffic pressure in RANs. For instance, due to the limited radio resources, mobile users might suffer from traffic congestion when delivering traffic through a same BS/AP at a hot spot and during the peak hours. Moreover, the interference/congestion may also lead to excessive energy consumptions of mobile users. Device-to-device (D2D) communication, an emerging paradigm in the fifth generation (5G) cellular networks, provides a promising efficient approach to accommodate the traffic demands. Instead of requiring the traffic to traverse through the cellular BS, D2D communications allow two nearby mobile stations (MSs) to perform a direct communication by using either the licensed inband spectrums or the out-band spectrums. Exploiting the close proximity between the MSs, D2D communications can bring lots of benefits, e.g., enhancing throughput, reducing traffic delay, and saving energy consumption. In particular, exploiting D2D communications to actively offload users' traffic has been envisioned as an efficient approach to relieve the traffic pressure in RANs and improve the efficiency in radio resource utilization (Li et al. 2014; Han et al. 2010; Wu et al. 2017, 2018a; Chuang and Lin 2012; Al-Kanj et al. 2014; Gao et al. 2014).

A concrete example of the D2D-assisted data offloading for content distribution in wireless network is shown in Fig. 1. Specifically, there exist a group of MSs, denoted by $\mathcal{U} = \{1, 2, \ldots, U\}$, coexisting in a given area, and there exist a group of content blocks (CBs), denoted by $\mathcal{K} = \{1, 2, \ldots, K\}$, available at the BS. We use L^k to denote the size of CB k. Each MS i is interested in receiving a subset of $\mathcal{K}$, and we use a U -by-K matrix $\mathbf{A}$ to denote the relationship between the MSs and the CBs, i.e., $A_{i,k} = 1$ if MS i is interested in obtaining CB k, and $A_{i,k} = 0$ otherwise. Set $\mathcal{C}_i$ denotes the set of the interested CBs which MS i wants to receive, i.e., $\mathcal{C}_i = \{k \in \mathcal{K} | A_{i,k} = 1, i \in \mathcal{U}\}$, and set Ω^k denotes the set of MSs which are interested in obtaining CB k, i.e., $\Omega^k = \{i \in \mathcal{U} | A_{i,k} = 1, k \in \mathcal{K}\}$. The group of MSs are in close proximity. To deliver CB k to all MSs in Ω^k, the D2D-assisted offloading mecha-

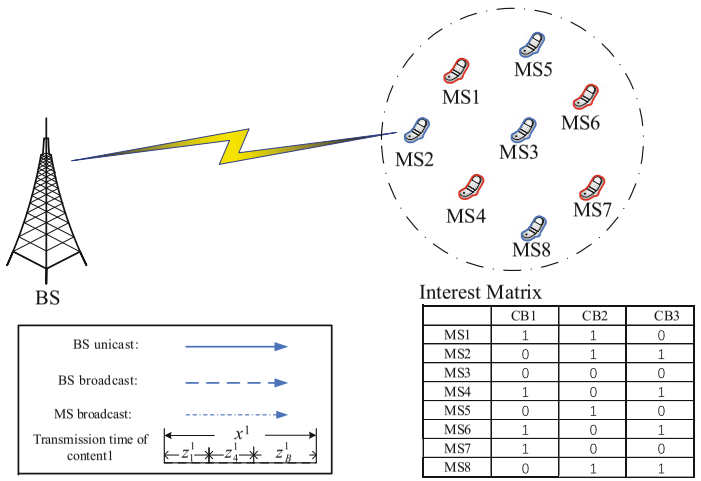
nism works as follows: The BS determines the transmission rate r^k of CB k, which is equivalent to its transmission duration $x^k = \frac{L^k}{r^k}$. The BS first sends part of data of CB k to some selected MS i, and meanwhile, MS i relays its received data via broadcasting to the other MSs in Ω^k. Let z_i^k denote the relaying duration of MS $i \in \Omega^k$ for CB k. There exists $\sum_{i\in\Omega^k} z_i^k \leq x^k$, the network scenario with $\mathcal{U} = \{1, 2, \ldots, 8\}$, $\mathcal{K} = \{1, 2, 3\}$, and the detailed matrix $\mathbf{A}$. Figure 1b, c, and d illustrate the process of distributing CB 1 to MSs 1, 4, 6, and 7 in a step-by-step manner, i.e., including Phase-I, Phase-II, and Phase-III, respectively. Figure 1b and c illustrate Phase-I and Phase-II of relaying through MS 1 and MS 4, respectively. In Phase-I, MS 1 sends its received data to MSs 4, 6, and 7. In Phase-II, MS 4 sends its received data to MSs 1, 6, and 7. Finally, taking into account the MSs' limited transmit powers and energy capacities, in Phase-III, the BS directly broadcasts the rest data of CB 1 to the MSs in Ω^1 for finishing the delivery of CB 1. Let z_{B}^1 denote the duration for the BS to perform this broadcasting. Then, there exists $z_{\mathrm{B}}^1 + \sum_{i\in\Omega^1} z_i^1 \leq x^1$. Figure 1d shows the case of BS's broadcasting to all MSs in Phase-III.

The D2D-assisted data offloading has attracted lots of interests, which can be in general categorized into two groups. The first group of studies focused on exploiting the D2D-assisted offloading for delay-tolerance applications. For instance, in (Li et al. 2014), a delay-tolerant network (DTN)-based offloading problem has been studied, which jointly takes into account the heterogeneous data traffics, users' different interests, and the limited storage resources. In (Han et al. 2010), accounting for the feature of mobile social networks, a target set selection problem for D2D-assisted information distribution has been investigated. In (Chuang and Lin 2012), exploiting the features of the encounter frequency and social communities, a community-based opportunistic D2D information dissemination scheme has been proposed, with the objectives of minimizing cellular traffic load and delivery time. The second group of studies focused on exploiting the D2D-assisted offloading for delay-sensitive (real-time) applications. For instance, in (Al-Kanj et al. 2014), the authors investigated how to separate the mobile users into different groups and to select a leader of each group for distributing the contents.

Key Applications

The above mobile data offloading via D2D-assisted communications can be used for lots of applications. The following are two typical examples.

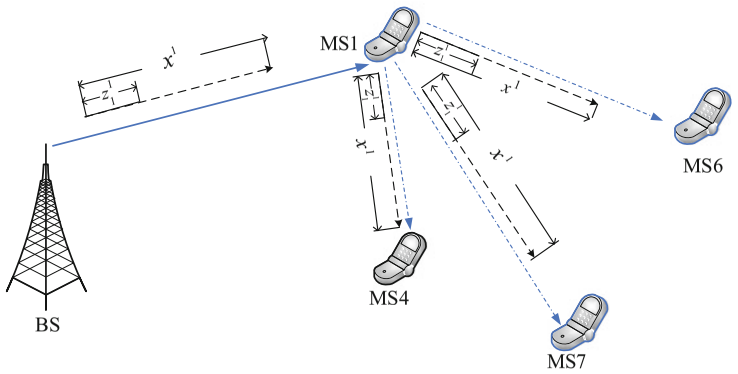
- *Vehicular networks*: The emerging vehicular communications/networks require data and message transmissions between the moving vehicles and infrastructures (e.g., roadside units (RSUs)) as well as the transmissions between the moving vehicles. Taking into account the mobility of vehicles and the delay requirement of data transmissions (e.g., for the safety-oriented applications), exploiting the D2D- assisted data offloading among the vehicular users provides an effective approach to improve the throughput and reliability of vehicular data transmissions.
- *Mobile edge computing*: Mobile edge computing (MEC), which enables mobile terminals to offload the computation tasks to nearby edge servers/terminals with sufficient computation resources, has been considered as a promising approach to implement the emerging computation-intensive mobile internet applications (e.g., augmented/virtual reality). However, MEC involves intensive data transmissions between the mobile terminals and the edge servers (e.g., a mobile user offloads its computation task to an edge server), which thus necessitate resource-efficient data transmissions with low latency and high throughput. The D2D-assisted data offloading has been envisioned as a promising approach to achieve this objective, and the recent advanced non-orthogonal multiple



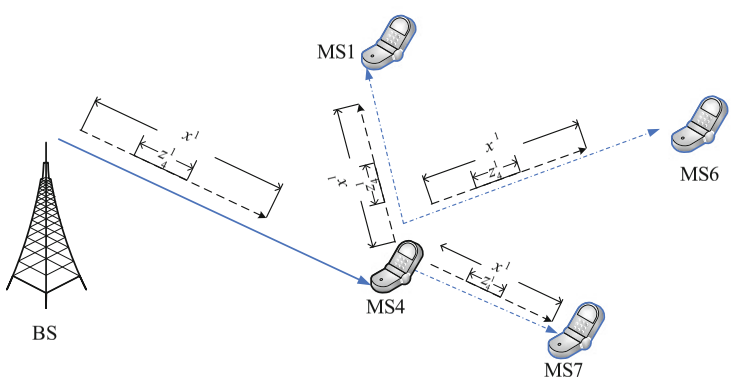
Interest Matrix

	CB1	CB2	CB3
MS1	1	1	0
MS2	0	1	1
MS3	0	0	0
MS4	1	0	1
MS5	0	1	0
MS6	1	0	1
MS7	1	0	0
MS8	0	1	1

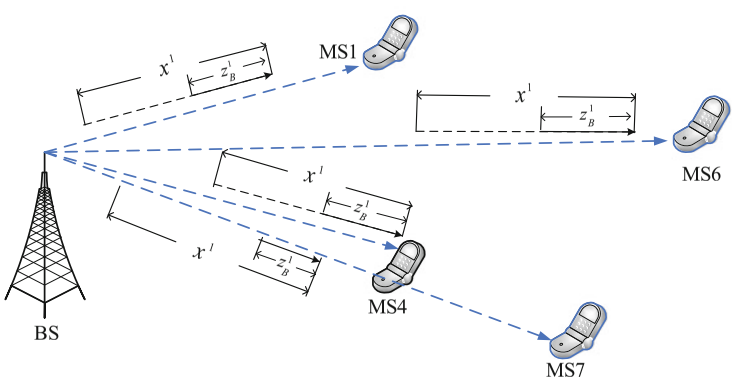
(a) Network scenario.



(b) Delivery of CB1:PhaseI.



(c) Delivery of CB1:PhaseII.



(d) Delivery of CB1:PhaseIII.

Mobile Data Offloading via D2D-Assisted Communications in Wireless Networks, Fig. 1 An example of the D2D-assisted data offloading for content distribution

access (NOMA) (Wu et al. 2018b) also facilitates the massive D2D communications for data offloading.

References

Al-Kanj L, Poor V, Dawy Z (2014) Optimal cellular offloading via device-to-device communication networks with fairness constraints. IEEE Trans Wirel Commun 13(8):4628–4643

Chuang YJ, Lin KCJ (2012) Cellular traffic offloading through community based opportunistic dissemination. In: Proceedings of 2012 IEEE wireless communications and networking conference. pp 3188–3193

Gao L, Iosifidis G, Huang J, Tassiulas L (2014) Hybrid data pricing for network-assisted user-provided connectivity. In: Proceedings of 2014 IEEE INFOCOM. pp 682–690

Han B, Hui P, Kumar VS, Marathe M, Pei G, Srinivasan A (2010) Cellular traffic offloading through opportunistic communications: a case study. In: Proceedings of the 5th ACM workshop on challenged networks. pp 31–38

Li Y, Qian M, Jin D, Hui P (2014) Multiple mobile data offloading through disruption tolerant networks. IEEE Trans Mob Comput 13(7):1579–1596

Wu Y, Chen J, Qian L, Huang J, Shen X (2017) Energy-aware cooperative traffic offloading via device-to-device cooperations: an analytical approach. IEEE Trans Mob Comput 16(1):97–114

Wu Y, Qian L, Mao H, Yang X, Shen X (2018b) Optimal power allocation and scheduling for non-orthogonal multiple access relay-assisted networks. IEEE Trans Mob Comput. https://doi.org/10.1109/TMC.2018.2812722

Wu Y, Qian L, Zheng J, Zhou H, Shen X (2018a) Green-oriented traffic offloading through dual-connectivity in future heterogeneous small-cell networks. IEEE Commun Mag 56(5):140–147

Mobile Device Security and Privacy

▸ Mobile Security and Privacy

Mobile Edge Caching

▸ Mobile Edge Caching in HetNets

Mobile Edge Caching in HetNets

Xiuhua Li[1,3], Xiaofei Wang[2], and Victor C. M. Leung[3]

[1]Chongqing University, Chongqing, P. R. China

[2]School of Computer Science and Technology, Tianjin University, Jinnan, Tianjin, China

[3]The University of British Columbia, Vancouver, BC, Canada

Synonyms

HetNet; Mobile edge caching

Definition

Mobile edge caching refers to distributing content files (e.g., videos, audios, photos, application programs, and so on) from service providers over the Internet to caches that are deployed at the edges (e.g., mobile devices and base stations) of mobile networks, aiming at bringing content closer to mobile users in the distance of the network topology to deal with the challenge of the explosive growth in content requests from users in mobile networks. HetNet is short for heterogeneous network and is a form of radio access networks (RANs) with complex interoperation between macro cells and small cells, which consists of different types of base stations (BSs) such as femto BSs, pico BSs, micro BSs, and macro BSs.

Mobile edge caching in HetNets is the combination between the technique of mobile edge caching and the network architecture of HetNets, aiming to achieve their joint benefits in enhancing network performances. Due to the high disparity of content popularity (i.e., a small number of content files actually may attract a large amount of downloads), by deploying hierarchical collaborative caching at the edges of HetNets, the content delivery cascades can be optimized during the intermediate transmissions, while reduced content delivery delays can be also provided.

Historical Background

With the rapid advancement of mobile networks from 2G and 3G to current 4G, people's daily life has changed significantly, and people are increasingly enjoying online social activities on mobile devices (e.g., smartphones and electronic tablets). As a result, requests for various content files (e.g., videos, audios, photos, application programs, and so on) from mobile users are increasing at an explosive speed, which has become a serious issue of mobile network operators (MNOs). However, current RANs and backhaul networks cannot support these content requests effectively due to the scarcity of network resources and the limit of their network architectures (Wang et al. 2015). Thus, to deal with those issues, it is necessary to employ revolutionary schemes in network architectures and data transmissions toward the fifth-generation (i.e., 5G) mobile networks.

Mobile edge caching is regarded as an effective technique to reduce the reduplicated network traffic load in mobile networks. In particular, studies in (Cha et al. 2007; Chen et al. 2002) have shown that a large of portion of the network traffic is caused by massive duplicated downloads of the same popular content. For instance, top 10% of videos in YouTube account for about 80% of all the views (Chen et al. 2002). Thus, with this technique, by caching popular content at the edges (e.g., mobile devices and BSs) of mobile networks, the requested content can be closer to mobile users in the distance of the network topology, and mobile users can directly access the content cached at the edges instead of downloading the content from SPs over the Internet via backhaul networks. Reduplicated transmissions from servers to clients are avoided, and most of the requests from users can be satisfied intermediately in mobile networks. Consequently, mobile edge caching can effectively enhance the network performances, especially on offloading network traffic (Chen and Yang 2016; Li et al. 2016, 2017), reducing system costs (Zhi et al. 2016; Gregori et al. 2016), and improving the quality of service (QoS) or quality of experience (QoE) of mobile users (Golrezaei et al. 2012; Zhao et al. 2016; Ao and Psounis 2015; Hong and Choi 2016; Li et al. 2015).

Another effective approach is to introduce the architecture of HetNets, which consists of different types of BSs (such as femto BSs, pico BSs, micro BSs, and macro BSs) with different wireless coverages (Li et al. 2016, 2017). HetNets can greatly enhance wireless link quality between mobile users and BSs and thus improve network capacity.

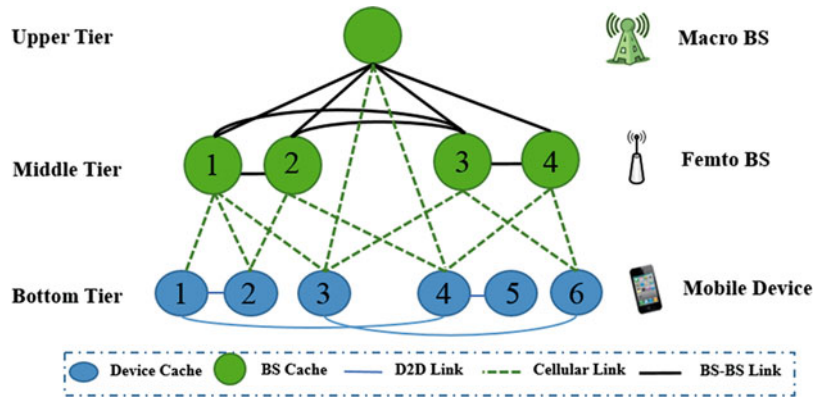
Considering the great potentials of the above two techniques, it is beneficial to combine them together, i.e., mobile edge caching in HetNets, which can effectively address the above discussed issues of MNOs. Studies in (Wang et al. 2015; Chen and Yang 2016; Zhi et al. 2016; Gregori et al. 2016; Golrezaei et al. 2012; Zhao et al. 2016; Ao and Psounis 2015; Hong and Choi 2016; Li et al. 2015) focused on single-tier caching in either mobile devices or BSs in mobile networks. Studies in (Li et al. 2016, 2017; Yang et al. 2016; Jiang et al. 2017; Xu and Tao 2017) focused on hierarchical BS caching in HetNets, where the edge caching in mobile devices is not considered. Studies in (Rao et al. 2016; Wang et al. 2017) explored the mobile edge caching in both mobile devices and BSs in mobile networks.

Foundations

To satisfy content requests from mobile users, mobile edge caching in HetNets needs to consider two phases, i.e., content placement phase and content delivery phase.

In the content placement phase, mobile edge caching in HetNets mainly deals with the following four issues:

- **Caching Topology** – To bring content closer to mobile users, it is important to decide where to cache in mobile networks. In HetNets, content can be cached at the edges that consist of mobile devices and BSs, which can form an either single-tier or hierarchical edge caching topology. Firstly, MNOs need to design the network architecture of HetNets,



Mobile Edge Caching in HetNets, Fig. 1 An illustration of caching topology

especially about how many tiers or types of BSs there are in HetNets as well as how different BSs connect. Secondly, MNOs need to consider whether mobile devices are able to cache content or not. Thirdly, MNOs need to consider which tiers of BSs are able to cache content or not. As a result, the mobile edge caching topology can be achieved, denoting the deployment of caches at the edges of HetNets.

As illustrated in Fig. 1, the caching topology of mobile edge caching in a HetNet is hierarchical and consists of three tiers of caching, i.e., bottom tier with six mobile device caches, middle tier with four femto BS caches, and upper tier with one marco BS cache. Here, the HetNet consists of two tiers of BSs, i.e., four femto BSs and one macro BS. All the BSs and mobile devices are able to cache content. Besides, the connection among BSs, among mobile devices, and between BSs and mobile devices is via BS-BS links, D2D links, and cellular links, respectively. In particular, mobile devices are possible to be connected with any tier of BSs. For instance, mobile device 3 is connected with femto BS 1 and femto BS 3 at the middle tier as well as the macro BS at the upper tier.

- **Content cacheability and storage format** – Mobile edge caching in HetNets aims to achieve a trade-off between network performances (e.g., network traffic load, system costs, and QoS/QoE) that are usually expensive to be improved and storage costs that are becoming much cheaper. However, the practical scale of content owned by SPs is growing rapidly, and thus it is impossible to cache all the content. Thus, it is important to decide what content to cache (i.e., content cacheability) taking content popularity into account. As practically captured in (Cha et al. 2007; Chen et al. 2002), only a small amount of popular content accounts for a large portion of content requests from mobile users, while a long tail of content remains unpopular. Besides, different types of content have different cacheability (Wang et al. 2015). For instance, among the content types, videos and photos have the highest revisit rate. Moreover, a large content file can be chunked into a series of original segments, and original segments can be encoded into an arbitrary number of packets to explore the diversity of cached content, while a given number of different coded packets can be collected together and decoded into the original segments for recovering the content file (Golrezaei et al. 2012; Yang et al. 2016). Thus, a content file can be cached by storing in three formats: (1) entire original content in a neither unchunked nor uncoded case, (2) original segments in a chunked but uncoded case, and (3) coded packets from the original segments in a chunked and coded case.
- **Caching policy** – Caching policies, deciding what to cache, how to cache, and when to release caches for what network objectives, are crucial to achieve the performance gains

of mobile edge caching. In particular, the optimization objectives of mobile edge caching in HetNets can be various for enhancing network performances, such as offloading network traffic, reducing system costs, improving mobile users' QoS/QoE, and so on. It is also important to estimate the gain of caching a content file by evaluating its current popularity, potential popularity, storage size, and locations of existing replicas in the network topology based on the system learning and analysis from mobile users' social behaviors and preferences (Li et al. 2016, 2017). Mobile edge caching policies can be operated in offline/online manners based on the practical network requirements. Offline caching (i.e., proactive) is relatively static with some prior network knowledge such as content requests and content popularity, while online caching (i.e., reactive caching) is relatively flexible according to the dynamics of network knowledge. Rather than employing traditional caching policies (e.g., least recently used (LRU), least frequently used (LFU), and first-in first-out (FIFO)), it is important and challenging to design proper cooperative mobile edge caching policies in HetNets to improve the network performances.

- **Operation time of caching** – According to the change of content popularity and operated manners of mobile edge caching, different types of content can be cached in different time periods. Online caching needs to cache content immediately or in a short time, while offline caching does not. For instance, in offline caching, short-lifetime popular news with short videos are updated every a few hours, while long-lifetime new movies and new music videos are, respectively, posted weekly and monthly (Li et al. 2017). In order to reduce the traffic load and avoid possible network traffic congestion especially in busy hours, content can be cached in off-peak hours (e.g., late night).

In the content delivery phase, each mobile user requests content based on its own preference. In order to satisfy a user's content request, mobile edge caching in HetNets mainly deals with the following two issues:

- **Content request routing –** Content request routing shows the possible routes for delivering the requested content in the derived caching topology to a user before operating practical wireless transmissions of content in HetNets. Specifically, in the network, the whole process of content request routing can be summarized as:
  - If a user's requested content is locally cached in its mobile device, then the request can be satisfied locally.
  - Otherwise, the user can first find the content in caches of other users in close proximity and then establishes a device-to-device (D2D) link with a user where the content is available and finally fetches the content in a D2D manner (e.g., Wi-Fi Direct, or Bluetooth) (Gregori et al. 2016; Rao et al. 2016).
  - If not yet satisfied, the user has to be served by the associated BSs via cellular links. If the requested content is locally cached, then the associated BSs satisfy the request directly; otherwise, the associated BSs need to explore the cooperation possibility for fetching the content from other BSs where the content is available.
  - If not yet satisfied, then downloading the content directly from SPs over the Internet via backhaul networks is the last resort.
- **Wireless transmission of content –** After the content request routing is known, the requested content will be delivered to mobile users with wireless transmissions via either D2D links or cellular links. In terms of D2D links, they are established only when a pair of mobile users are in close proximity and willing to share the cached content in a D2D manner. Here, social behaviors and content preference of mobile users need to be analyzed and estimated in advance by system learning. In terms of cellular links, the associated BSs with noncooperation or

cooperation can transmit the requested content to the user by unicasting/multicasting based on the practical data transmission schemes in HetNets. In particular, resource allocation and scheduling for wireless transmissions of content via cellular links is necessary to enhance the network performances while satisfying the content requests.

Key Applications

The technique of mobile edge caching in HetNets can be utilized for general MNOs to deal with the explosive growth in content requests from mobile users, and thus there are also venders that are manufacturing cache-enabled base station products, e.g., cache-enabled femtocell products and cache-enabled Wi-Fi routers. For companies of content delivery networks (CDNs), a new key trend is also the extension of their services into mobile edge networks, particularly for BSs with opened interfaces for MNOs and third-party content providers. Mobile edge caching in HetNets can also be utilized in 4G networks, but will be widely deployed in 5G networks.

Future Directions

From the evolution of mobile edge caching in HetNets, it appears that, in the future, emphasis will be given on the following research directions as follows:

- Big data-based large-scale online/offline optimization of content caching for MNOs
- Machine learning-based analysis and prediction on mobile users' social behaviors and preference for caching and prefetching optimization
- More rapid and secure collaborations among mobile devices and BSs in HetNets
- Pricing methodology for mobile edge caching in HetNets

Cross-References

- ▸ Content-Centric Wireless Sensor Networks
- ▸ Resource Allocation
- ▸ Mobile Edge Computing

References

Ao WC, Psounis K (2015) Distributed caching and small cell cooperation for fast content delivery. In: Proceedings of theACM MobiHoc, June 2015, pp 127–136

Cha M, Kwak H, Rodriguez P, Ahn YY, Moon S (2007) I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the ACM IMC, Oct 2007, pp 1–14

Chen B, Yang C (2016) Caching policy optimization for D2D communications by learning user preference. In: Proceedings of the IEEE WCNC, May 2016, pp 1–6

Chen Y, Qiu L, Chen W, Nguyen L, Katz R (2002) Clustering web content for efficient replication. In: Proceedings of the IEEE ICNP, Nov 2002, pp 165–174

Golrezaei N, Shanmugam K, Dimakis AG, Molisch AF, Caire G (2012) FemtoCaching: wireless video content delivery through distributed caching helpers. In: Proceedings of the IEEE INFOCOM, Mar 2012, pp 1107–1115

Gregori M, Vilardebó JG, Matamoros J, Gündüz D (2016) Wireless content caching for small cell and D2D networks. IEEE J Sel Areas Commun 34(5):1222–1234

Hong JP, Choi W (2016) User prefix caching for average playback delay reduction in wireless video streaming. IEEE Trans Wirel Commun 15(1):377–388

Jiang W, Feng G, Qin S (2017) Optimal cooperative content caching and delivery policy for heterogeneous cellular networks. IEEE Trans Mobile Comput 16(5):1382–1393

Li X, Wang X, Xiao S, Leung VCM (2015) Delay performance analysis of cooperative cell caching in future mobile networks. In: Proceedings of the IEEE ICC, June 2015, pp 5652–5657

Li X, Wang X, Leung VCM (2016) Weighted network traffic offloading in cache-enabled heterogeneous networks. In: Proceedings of the IEEE ICC, May 2016, pp 1–6

Li X, Wang X, Li K, Han Z, Leung VCM (2017) Collaborative multi-tier caching in heterogeneous networks: modeling, analysis, and design. IEEE Trans Wirel Commun 16(10):6926–6939

Rao J, Feng H, Yang C, Chen Z, Xia B (2016) Optimal caching placement for D2D assisted wireless caching networks. In: Proceedings of the IEEE ICC, May 2016, pp 1–6

Wang X, Li X, Leung VCM, Nasiopoulos P (2015) A framework of cooperative cell caching for the future mobile networks. In: Proceedings of the HICSS, Jan 2015, pp 5404–5413

Wang W, Lan R, Gu J, Huang A, Shan H, Zhang Z (2017) Edge caching at base stations with device-to-device offloading. IEEE Access 5:6399–6410
Xu X, Tao M (2017) Modeling, analysis, and optimization of coded caching in small-cell networks. IEEE Trans Commun 65(8):3415–3428
Yang C, Yao Y, Chen Z, Xia B (2016) Analysis on cache-enabled wireless heterogeneous networks. IEEE Trans Wirel Commun 15(1):131–145
Zhao Z, Peng M, Ding Z, Wang W, Poor HV (2016) Cluster content caching: an energy-efficient approach to improve quality of service in cloud radio access networks. IEEE J Sel Areas Commun 34(5):1207–1221
Zhi W, Zhu K, Zhang Y, Zhang L (2016) Hierarchically social-aware incentivized caching for D2D communications. In: Proceedings of the IEEE ICPDS, Dec 2016, pp 316–323

Mobile Edge Computing

▸ Computation Offloading in Mobile Edge Computing
▸ Data-Driven Edge Computing
▸ Joint Caching, Computing, and Routing for Video Transcoding in Wireless Networks

Mobile Edge Computing: Low Latency and High Reliability

Zhiyuan Ren[1], Chen Chen[1], and Jun Fu[2]
[1]Xidian University, Xi'an, Shaanxi, China
[2]Future Forum, Beijing, China

Synonyms

Edge computing; Mobile cloud computing

Definition

Mobile edge computing (MEC) has been proposed in recent years for offloading computation tasks from user equipment (UE) to the network edge to break hardware limitations and resource constraints at UE. Through migrating the computation, storage, and servicing capability to the edge of network, MEC can deploy application, service, and content locally in a distributed way. To a certain extent, it will satisfy the critical low latency and high reliability requirements for future eMBB, uRLLC, mMTC scenarios.

Historical Background

The European Telecommunications Standards Institute (ETSI) proposed mobile edge computing (MEC) in 2014, which is a technology based on 5G evolution architecture and deeply integrates the base station (BS) and Internet services. This technology is highly concerned by operators and equipment vendors, i.e., Huawei conducted joint tests with major operators and believes that MEC is one of the key technologies for coordinating user needs and network capabilities, which will help customers to achieve the goal of intelligent manufacturing, such as production, maintenance, management, inspection, and security. In 2016, the Edge Computing Consortium (ECC) was formally established, which would further promote the development of MEC technology.

At present, researches on MEC mainly focus on the computing offloading, energy efficiency, edge storage, low-latency guarantee, etc.

Computing offloading means that the terminal device with limited computing capability transmits the heavy computing task to the nearby MEC servers, and the MEC server replaces the terminal to complete the computing task. After the computing task is over, the MEC server returns the result to the terminal device. Chen et al. (2016) proposed a game theoretic approach for the computation offloading decision making problem among multiple-users in multichannel environments. To solve the problem of computing offloading in the Internet of Vehicles, Zhang et al. (2016) proposed a cloud-based mobile edge computing offloading architecture and designed an efficient allocation schedule of computing resources. Under the latency constraints of computing tasks, the resource utilization can be improved. Swaroop Nunna et al. (2015) proposed the concept of constructing next-generation collaborative architecture based on

MEC server under 5G network. It is located at the edge of mobile network and cooperates with the underlying communication network, which provides a potential architecture solution for the environment-aware collaboration of critical mission. Jararweh et al. (2016) proposed a software-defined system based on edge computing (SDMEC). Through the collaborative work of SD-Network, SD-Compute, SD-Storage, and SD-Security, the capabilities of cloud are expanded to the edge of the network, which satisfied the requirements of some applications, i.e., traffic regulation, content sharing, mobile gaming, etc.

In the view of the bright prospects of MEC, academics and industries have already carried out research on edge computing and have made some progress. However, the MEC technology is still immature, and specific implementation plans and related technical standards are still evolving.

Foundations

MEC offers application developers and content providers cloud-computing capabilities and an IT service environment at the edge of the network, as Fig. 1 shows. Through migrating the services which requires high complexity and high-energy consumption to the MEC server, the smart mobile terminals could solve the problem of limited capability in computation, storage, and battery. In additional, MEC can effectively reduce the demand of network bandwidth and the increasing transmission pressure of the core network with local offloading, caching, and deployment. To satisfy the future requirements of eMBB (enhanced mobile broadband), uRLLC (ultra-reliable low-latency communication), and mMTC (massive machine type communication), the following key technologies under MEC have been studied:

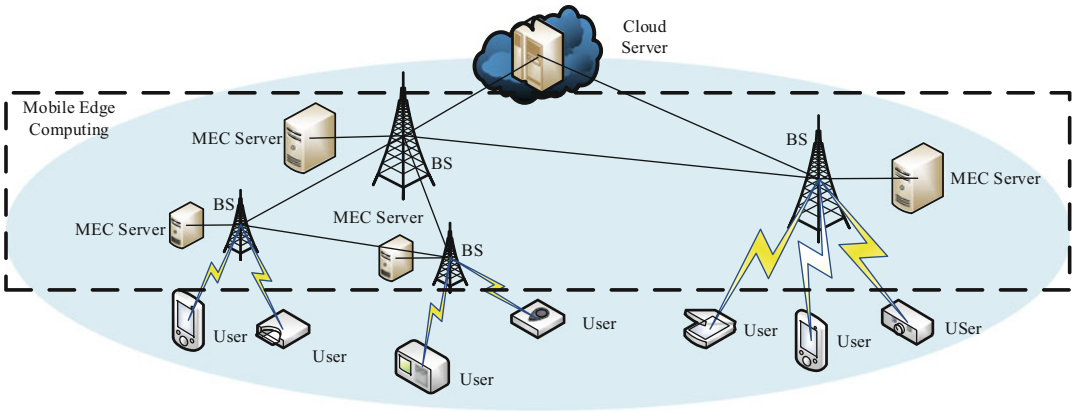
Traffic offloading. Traffic offloading is one of the basic features in the MEC. By leading a part of traffic to local network (campus network or enterprise network), traffic offloading can greatly reduce the stress of core network bandwidth and decrease the transmission delay. The future network will be a merging network, which includes 4G, 5G, Wi-Fi, etc. Therefore, the future deployment of applications need to consider how to realize the local traffic offloading technology based on MEC in 4G networks. In addition, the development of the local traffic offloading technology in MEC needs to further clarify the commercial model.

Cache and acceleration. MEC can deploy the content closer to the end user. Thus, data can be directly obtained from local server, which would greatly reduce the service access delay. Besides, several questions need to be considered when caching is used in MEC:

1. According to the MEC typical service scenarios, and the existing caching models (local DNS, redirection, transparent proxy), MEC needs to choose appropriate caching model.
2. MEC needs to design efficient algorithm to optimize the cache performance or make a trade-off between its deployment place and the hitting ratio.
3. Cache channel from MEC to remote server needs to be optimized to support cache acceleration scheme based on MEC.
4. To avoid the same video content repeatedly download due to different resolution (different bit rate), MEC needs to consider how to cache the same video content (HD) version and regenerate the content for different resolution (rate) to provide service for different kinds of terminals.

Mobility management. Under the MEC environment, to provide service continuity, new mobility management needs to be deployed at the edge of the network. The mobility management is expected to handle three scenarios including the change of data forwarding path caused by UE mobility, application data migration due to load balance, and interaction between MEC system and other systems when the UE moves in/out MEC service area.

MEC-based computing offloading. MEC-based computation offloading is a technology which leverages the edge devices near terminals to process task. By using a series of network



Mobile Edge Computing: Low Latency and High Reliability, Fig. 1 The architecture of MEC

equipment near the BS, the delay sensitive service should be offloaded to the edge devices for rapidly response and lower latency. Moreover, besides eMBB scenario, the 5G network contains lots of low-latency high-reliability services, such as V2X, intelligent manufacturing, etc. These applications require many sensors or terminals for information collection, then transmit the information to the network side for cooperative process. Although the computing capability of single edge device is finite, the network has large number of edge devices which means the distributed computing can be used in this scenario. Therefore, the computing capacity of all the edge devices could be combined, which could form a distributed edge computing pool to remarkably reduce the computing pressure of cloud platform, ensure the QoS of communication, and improve the stability and reliability of the network.

MEC security. The MEC server is a brand-new type entity in the mobile network. While connecting to several mobile network entities including Operation Administration and Maintenance (OAM) system or Lawful Interception, etc., it also connects to the third-party application server and even accommodates the third-party application. It is noted that current security techniques including IP sec, TLS, firewall, etc. could secure these connections to block the attack to the MEC server, mobile system, and the third-party server. However, these solutions are not safe enough for the new security issues with the introduction of MEC. It is necessary to setup a unified, credible security evaluation system to evaluate the security of the applications running in the MEC system, which will only guarantee the trustable third-party application in the server. Meanwhile, the security aspects of the interface between the application and the network should also be considered for safely communication, e.g., the unauthorized invocation. There are also new business opportunities for operators to provide the security service. The mobile operator could provide the service to shield the application in the MEC server from the malicious attack and check the security bugs of the third-party apps.

Key Applications

According to the definition of the ETSI Industry Specification Group (ISG), MEC offers application developers and content providers cloud-computing capabilities and an IT service environment at the edge of the network. This makes it possible to deploy the application, service, and content closer to the UE in a distributed manner. The MEC application focuses on the scenarios of low latency and high bandwidth. According to different service characteristics, the MEC applications can be divided into the following two categories: local service and vertical industry. The local service

applications mainly consist of campus network and AR/VR service, etc. And the vertical industry applications mainly include the V2X and industrial Internet scenarios. The following are some typical applications of the local service and vertical industry applications.

Local Service

Typical Application 1: Local Video Surveillance
In the scenario of local video surveillance, the traffic volume of video surveillance especially HD video is very huge. For example, the uplink bandwidth is up to 16Mbps for 4K HD video. If all data must be sent to the central cloud, it will consume large backhaul bandwidth. On the other hand, the video needs to be processed and analyzed timely to handle the emergency situation. The MEC can satisfy the user requirement by offloading the traffic and processing the data locally to provide a better user experience. Besides video surveillance, many applications like sports event live require low latency and high bandwidth, and MEC can significantly enhance the user experience and reduce the operator network load in such environment.

Typical Application 2: Augmented Reality
There exist many scenarios for AR, such as museum, city memorial, sports event, and concert, which can enhance the user experience remarkably. MEC can achieve efficient office automation assistance, local resource access, and internal communication in these applications and can provide the consumer better experience and the ability to access local service.

Vertical Industry

Typical Application 1: Autonomous Driving
The autonomous driving vehicles generate large amounts of data. For example, one autonomous driving vehicle can generate and process TB level data per hour, which requires high bandwidth for data delivery and needs immediate processing the data and delivers the result to the vehicles with ultra-low latency. The latency requirement for assistant driving is 20-100 ms, but autonomous driving requires 3 ms. MEC is one of the technologies for realizing low-latency traffic in mobile networks. There are two types of MEC deployments for Internet of Vehicles (IoV). The first one needs building an edge cloud system at the base station side, and the other one uses strong computing capability BS to provide low-latency services for mobile terminals.

Typical Application 2: Industrial Internet
Industry Internet of Things (IIoT) enables intelligent manufacturing for future industry, whose production data would be generated locally. The data processing and the result feedback need ultra-low latency which may be lower than 1 ms. Thus, the processing and computing can only be finished locally due to the privacy requirement. MEC can satisfy this requirement, which can also complete the coordination of different production processes, such as Drone and Robot.

Further Directions

MEC is a new network architecture concept, which has many new features comparing to the existing 3G/4G cellular systems with better latency and reliability performance. Therefore, there are many research challenges and opportunities in the future research works. The security of the MEC, the integration of the computing, storage and communications in the MEC architecture, the online caching in the MEC, etc. should be further researched.

Cross-References

- Cloud Computing
- Computation Offloading in Mobile Edge Computing
- Mobile Edge Computing
- Mobility Management

References

Chen X, Jiao L, Li WZ, Fu XM (2016) Efficient multi-user computation offloading for mobile-edge cloud computing. IEEE/ACM Trans Netw 24(5):2795–2808

Jararweh Y, Doulat A, Darabseh A et al (2016) SDMEC: Software defined system for mobile edge computing. Paper presented at IEEE international conference on cloud engineering workshop, Berlin, 4–8 Apr 2016

Nunna S, Kousaridas A, Ibrahim M (2015) Enabling real-time context-aware collaboration through 5G and mobile edge computing. Paper presented at 12th international conference on information technology – new generations, Las Vegas, 13–15 Apr 2015

Zhang K, Mao YM, Leng SP, Vinel A et al (2016) Delay constrained offloading for Mobile Edge Computing in cloud-enabled vehicular networks. Paper presented at 8th international workshop on resilient networks design and modeling, Halmstad, 13–15 Sept 2016

Mobile IP

Charles Perkins
Huawei Technologies, Santa Clara, CA, USA

Synonyms

MIP; MIPv4; MIPv6

Definition

Mobile IP (MIP) is a mobility management protocol developed within the IETF (Perkins 2002; Perkins et al. 2011). It is often also used to refer to one of, or a suite of, related protocols such as Proxy MIP, Mobile IPv6, Hierarchical Mobile IP, Fast Mobile IP, and numerous extensions for these.

Historical Background

The development of Mobile IP started in the early 1990s following on work done at Columbia University, IBM T.J. Watson Research, Carnegie Mellon University, and a number of other research groups. Further development has continued in research departments around the world for many years, investigating the effects of Mobile IP on TCP, performance under high packet loss, hierarchical routing schemes, distributed mobility, interactions with BGP, etc.

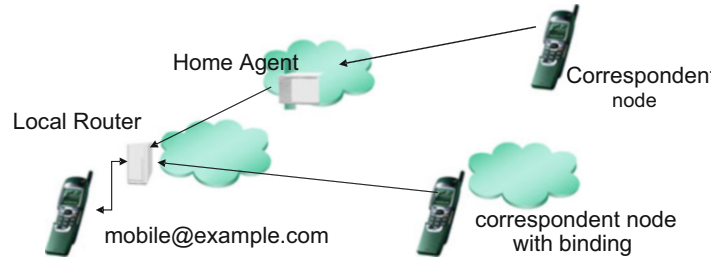
The main problem solved by Mobile IP was to loosen IP's tight binding between the identity and the location of a mobile device. Since the beginning of the Internet, network-addressable devices had been known as stationary computers, and so routing data to the computer running an application really amounted to using routing prefix of the IP address to locate the network on which the computer was located. Put another way, once data arrived at the location determined by the IP address, there was not any question about the identity of the networked computer that would receive it. With the arrival of smaller mobile computers in the late 1980s, this assumption was no longer valid.

Protocol Operation

Mobile IP solves this problem by associating two IP addresses with a mobile device. One is the locator, and this IP address changes when the device moves to a new network. The other IP address is the device's identity, which does not change. The protocol itself is designed as an extension to IP, which binds (associates) the two IP addresses at a router on the device's "home network." The home network is the network to which data will be routed based on the IP address which identifies the mobile device. Once data arrive there, the router on the home network instead routes the data to the IP address which is the locator of the mobile device. The router that does this is called the "home agent." The locator IP address is routable to the network (called the visited network) currently visited by the mobile device. The association between the identifying IP address and the locating IP address is called a "binding," and as the locator IP address changes, the binding changes to reflect the updated address.

The process of updating the binding is the central design point for Mobile IP – and, indeed,

Mobile IP, Fig. 1



a similar requirement exists for practically every mobility management protocol whether or not developed in the IETF. Naturally, there are many details. Binding updates have to be done securely, to prevent unauthorized network devices from binding the MN's identifier IP address to a false locator IP address. There are many ways to prevent unauthorized tampering with the true locator IP address, and this is an area of protocol design that is still evolving in the Internet today. Another design area is related to the mechanism by which the home agent redirects traffic from the home network to the current location of the mobile device. This usually involves tunneling (encapsulating) the data once it arrives at the home network by putting on a new IP header that will cause the data to be routed to the locator IP address.

Many tunneling protocols exist in the IETF, and probably every one of them could be used for the purpose of redirecting traffic from the identifier IP address to the locator IP address, which in Mobile IP documents is typically called the "care-of address." Moreover, it is also possible to insert into a data packet a "source route" which supplies a list of intermediate routing points that can be followed to ensure the ultimate delivery of the packet. Tunneling and source routing are more or less equivalent, but tunneling has had more design work in the IETF and so has generally been favored for redirecting traffic as with Mobile IP. Lately, however, source routing is receiving renewed interest in the IETF (check the segment routing WG Fig. 1).

In the figure, a mobile device (mobile@example.com) has established a point of attachment with the Internet by way of a local router. For the rest of this entry, the mobile device will be called a mobile node, abbreviated MN, as is usually done in IETF documents. Any other Internet device that communicates with MN will be called a Correspondent Node (the name often used in Mobile IP documents for the other communication endpoint for some traffic to/from MN). Traffic from the Internet can reach the mobile node (MN) by arriving at the Home Network, where the packets will be redirected (tunneled) to the local router currently serving the mobile node.

As a general observation, mobility management systems universally require some way for packets to be redirected to the current point of attachment of a mobile node. Determining the address for the redirection requires a lookup operation somewhere in the Internet. For business reasons, wireless operators have demanded that this lookup operation occurs within the operator's network, but this is not mandated by Mobile IP. The exact architectural placement of the lookup operation has a crucial effect on the performance of mobility management. Mobile IP is unique among mobility management systems in that the lookup operation occurs naturally on the routing path of the traffic to the mobile node, as determined by the mobile node's home address. All other systems have to insert some architectural augmentation for this to happen, which can easily be the source of bugs, performance loss, and other problems.

Packets routed to and from MN, which are redirected from the home network, experience additional delay because of the longer routing path. Depending upon local requirements in the network visited by MN, it is also likely that packets from the mobile node will have to be tunneled back to the home network. However, if a

correspondent node is able to receive information about the IP address of the mobile node in the visited network, then that correspondent node can use the information to route packets directly to and from the visited network. Any method by which the correspondent node is able to obtain the mobile node's care-of address is a kind of route optimization; various such methods have been proposed for MIPv4 and MIPv6.

Mobile IP for IPv4 (MIPv4) (Perkins 2002) and Mobile IP for IPv6 (MIPv6) (Perkins et al. 2011) are very similar in general operation, but there are important differences. The first distinction is that MIPv4 specifies the behavior of the local router, which for MIPv4 is called a Foreign Agent (FA). The Foreign Agent can issue periodic beacons to notify mobile devices that the FA is offering mobility services, and also provide a local IP address (i.e., a locator IP address). Once the mobile device MN identifies a Foreign Agent, then MN can use a (topologically correct) locator IP address as the local endpoint for traffic which will be tunneled from the Home Network. Thus, equipped MN supplies the locator IP address to the Home Agent and instructs the Home Agent to intercept traffic destined for the MN. If the tunnel is established between the Home Agent and the Foreign Agent, then the FA decapsulates packets and delivers them to the MN. Otherwise, the tunnel can be established between the Home Agent directly with the MN, and the FA has only to route the encapsulated packets as usual.

MIPv6, on the other hand, does not need to specify a Foreign Agent. By the design of IPv6, MN can get a locator IP address either by way of Router Advertisement or by way of DHCP (depending on local network administrative requirements); local IPv6 routers can be expected to support DHCP. Also, MIPv6 enables a special kind of routing header that causes packets to be delivered to MN's care-of address, avoiding the extra overhead of IPv6-within-IPv6 encapsulation.

As noted above, binding updates have to be performed securely. When Mobile IPv4 was designed, there was no IPsec. As a result, various Mobile IPv4-specific mechanisms were designed. These are still in use with MIPv4 and have not been reported as broken or suffering from vulnerabilities that have plagued many other security protocols. The mip4 working group of the IETF even designed a key management and allocation protocol which enabled establishment of session keys for mobility management, based on an underlying security association between the home agent and the mobile node.

The security design of Mobile IPv6 was largely designed to utilize IPv6's security design (IPsec). IPsec offers authentication by insertion of a specially encrypted signature to guarantee that tampering can be detected. For further privacy, IPsec also offers payload encryption although this is not demanded for the safety of Mobile IPv6 signaling. Key management and refresh is also handled by standard IPsec methods.

Related Protocols

PMIP: At the time that Mobile IP was designed (late 1980s and early 1990s), IPv6 did not exist, and telecommunications was often considered synonymous with the international telephone system. Internet communication systems were mainly built on Unix systems running TCP/IP, and voice-over-IP (VOIP) did not exist for all practical purposes. In fact, VOIP was made illegal in some countries. As a consequence, the Internet was a small fraction of the total volume of communications. Internet applications typically did not have real-time requirements and ran on devices that did not move. As a further consequence, Mobile IP was not offered by the operating systems of the time that were dominant on laptop computers, which were the only ones that could move (albeit slowly). As the importance of Internet communications grew, network operators preferred to manage mobile devices without updating them to run Mobile IP. Instead, a network-proxy approach was desired which would restrict mobility management to the domain of a particular operator and still allow devices to be mobile within that domain even though the mobile devices did not have Mobile

IP available. This proxy approach is called Proxy Mobile IP (PMIP) (Gundavelli et al. 2008).

HMIP: An important variation of Mobile IP is "Hierarchical Mobile IP" (HMIP) (Soliman et al. 2008). HMIP was designed to separate the handling of local mobility (within one operator domain) from mobility between domains. The basic idea is to run local mobility management signaling to a local home agent as long as possible, and only transmit control messages to the (possibly distant) home network when the mobile device makes a new point of attachment outside of the domain which it had been visiting.

FMIP: A more sophisticated approach for reducing the latency penalty association with mobility signaling across the larger Internet during handover has been designed and is known by the name of Fast Mobile IP (FMIP) (Koodli 2009). The basic idea is to enable neighboring IP access points to cooperate with each other by sharing information about the mobile node (MN). As MN moves from the network of one router (say, R1) to the network of another router (say, R2), R1 can send information to R2 about the mobile node, including its home address and its current locator IP address, temporary security credentials, multicast subscriptions, and other information (called "context") about the mobile node to speed up the handover. Then, while packets in flight are still arriving at router R1, those packets can be further redirected to R2, the router serving MN at its new point of attachment. The signaling between the routers exchanges the context and sets up the temporary tunnel between the routers. The tunnel is configured to remain active for a long enough time so that the home agent can update its binding for MN to reflect the locator IP address associated with R2's network.

Route optimization: As mentioned previously, route optimization techniques have been investigated for Mobile IP, so that a mobile node's packets could be exchanged directly with a correspondent node without traveling to and from the home agent. One of the important contributions of Mobile IPv6 was to integrate a route optimization mechanism directly with the base protocol (Arkko et al. 2007). The route optimization protocol messages have the simple goal of establishing a binding (as before, an association between the MN's home address and its care-of address) at the correspondent node that can be used for tunneling. Then the correspondent node can use that binding to deliver packets directly to the mobile node's care-of address. This was ground-breaking because it was the first time that a secure operation was designed to work across such a potentially large population of Internet devices (namely, the many possible correspondent nodes that might interact with a mobile node). The observation that enabled such a scalable approach to security was that the binding update can be authorized by assuring that the new care-of address belongs to the same mobile node with which a correspondent node had already been exchanging packets. This is a significantly weaker requirement than verifying the identity of a mobile node, and yet perfectly strong enough to prevent hijacking the mobile node's traffic with its correspondent node.

Future Directions

Mobile IP and its related protocols have provided many innovations and a scalable approach to wide-area mobility management in the Internet. For various reasons, Mobile IP as specified within the IETF remains sparsely deployed in today's wireless operator networks, but it has received a great deal of attention and has been influential in the evolution of modern telecommunication networks. With the advent of 5G networks, Mobile IP and its descendants within the [dmm] working group are seen as viable candidates for some 5G mobility management needs. With multigigabit wireless technologies already available, the high speed and architectural simplicity of Mobile IP become more attractive, as well as its inherent ability to serve all applications without requiring per-application mobility support.

Cross-References

▸ Distributed IP Mobility Management
▸ Network-layer Mobility Management
▸ Proxy Mobile IPv6

References

Arkko J, Vogt C, Haddad W (2007) Enhanced route optimization for Mobile IPv6. RFC Editor
Gundavelli S, Leung K, Devarapalli V et al (2008) Proxy Mobile IPv6. RFC Editor
Koodli R (2009) Mobile IPv6 fast handovers. RFC Editor
Perkins C (2002) IP mobility support for IPv4. RFC Editor
Perkins C, Johnson D, Arkko J (2011) Mobility support in IPv6. RFC Editor
Soliman H, Castelluccia C, ElMalki K, Bellier L (2008) Hierarchical Mobile IPv6 (HMIPv6) mobility management. RFC Editor

Mobile Networks

▸ Capacity of Wireless Ad Hoc Networks

Mobile Node

▸ Proxy Mobile IPv6

Mobile Security

▸ Mobile Security and Privacy

Mobile Security and Privacy

Feng Lin
Department of Computer Science and Engineering, University of Colorado Denver, Denver, CO, USA

Synonyms

Mobile device security and privacy; Mobile security; Smartphone security and privacy

Definitions

The protection of smartphones, tablets, other portable computing devices, and the networks they connect to, from threats and vulnerabilities associated with wireless computing

Historical Background

Mobile security and privacy research has brought public attentions since the prevalence of cellphone and cellular networks. Notare et al. (1999) proposed a distributed security management system for telecommunications networks against cloned cellphones (same number and series of a genuine phone) in 1999. Thereafter, research in this area can be divided into two eras. The first era focuses on normal cellphones and voice services, such as cellphone authentication (Manabe et al. 2009), cellphone identification (Celiktutan et al. 2007), and cellphone cloning issue (Singh et al. 2007).

Since the debut of iPhone in 2007 and later the Android phones, research in mobile security and privacy is gradually shifting from cellphones to smartphones, which enriches more diverse application scenarios concentrating on mobile apps and mobile networks. The potential threats in mobiles may be caused by social engineering, compromised devices, malware, web browser or OS vulnerability, vulnerable application, and data interception. Especially, repackaging is one type of common attacks for smartphones targeting mobile apps, in which an attacker could modify a legitimate app to include malicious code and publish on third-party app stores. Correspondingly, the countermeasures have been developed to protect the mobile security and privacy, for example, multifactor authentication scheme that could avoid attacker easily guessing password or brute-force attack; detecting malware behaviors could play a significant role in reducing attack risks.

Different from traditional computer systems, the primary design objectives for mobile devices are low power consumption and portability. Hence, there are five key aspects of device- and application-level security features of mobile platform that distinguish mobile security from conventional computer security (La Polla et al. 2013): mobility, strong personalization, strong

connectivity, technology convergence, and reduced capabilities.

Mobile Vulnerability

- **Web-based services**
  The mobile web is the most used platform other than mobile operating system (OS) platform. Similar to traditional web services on personal computers, mobile web applications that employ lightweight pages also suffer from security threats, such as phishing scams, drive-by downloads, browser exploits, cross-site scripting (XSS), SQL injection, etc. On the other hand, cyberattacks can exploit software vulnerabilities via mobile web browser to compromise mobile devices (Coursen 2007), such as PDF reader or image viewer.

- **Wireless networks**
  Mobile devices can provide Internet connection and device-to-device communications at any time and in any place thanks to the wireless networks technologies, including Wi-Fi, cellular networks, Bluetooth, and near-field communication (NFC). Though these wireless networks provide convenience for users, using them in a public setting may cause potential risks. For example, Wi-Fi sniffing happens frequently with free Wi-Fi networks in airports, malls, and other public places when the mobile device is connecting to a malicious wireless access point (Beyah and Venkataraman 2011) and, hence, causes mobile devices' data exposure to attackers including account and personal information (Sujithra and Padmavathi 2012). Mobile networks may also suffer from overbilling attacks in which additional fees, such as fees of data traffic a user never used, are charged to the victim's accounts and transferred to the attacker's wallet.

- **Mobile malware**
  As of the first quarter of 2018, more than 7 million of mobile apps are on the markets, where Android users were able to choose between 3.8 million apps and Apple users can choose 2 million available apps in Apple App Store (Statista 2018). Among such huge mobile apps, there are many individual-developed third-party apps with malicious intent. Once those malicious apps, which may contain virus, Trojan, or botnet, are installed and gain access to the mobile devices, they could cause serious security breaches and incidents. As reported in Bradley (2011), more than 50 apps have been found to be infected by an Android malware called DroidDream in the official Google Play Store, which can stealthily gain root access to the device and further download additional malicious programs without user's knowledge and permission. As an example of smart malware, a multifarious malware for iOS devices, called iSAM (Damopoulos et al. 2011), has recently been designed, which integrates several typical malicious features of malware such as collecting confidential data stealthily and denial of application and network services, sending a large number of malicious SMS.

- **Weak authentication**
  Most contemporary mobile devices use PIN, graphical pattern, or biometrics-based authentication such as fingerprint scan or face recognition. Those authentication methods are either vulnerable to malicious hacking, or the credentials are easily to be disclosed to others intentionally or voluntarily. PIN and pattern can suffer from *smudge attack* (Aviv et al. 2010), in which the password can be inferred by the smudge left on the screen surface when your fingertip touch on the screen. The brute-force attacks could be another threat to the PIN-based password. Though the biometrics-based authentication is superior because of the uniqueness of individuals, the biometric credentials are easy to be duplicated or counterfeited as anyone can obtain your fingerprint from a cup that you have hold.

What is even worse is that an attacker can obtain such biometric information from a distance. Chaos Computer Club announced that one of its members had been able to replicate the fingerprint of German Minister of Defense Ursula von der Leyen, using only photographs taken of her finger (CCC 2014).

Goal of Attacks

- **Denial of service**
  Due to the portability and small-sized design, mobile devices usually have capability-limited hardware and battery, which restrict the computing, transmission, and power supply of the mobile device. Therefore, denial of service (DoS) attack is prone to target these shortages of mobile devices to disable one service, reduce the capability, or even make the whole device unusable by utilizing different attack techniques such as broadcasting highly malicious traffic stream, sending huge messages, or increasing power consumption. Battery power exhaustion attack is one of those DoS attacks that drains the battery faster than usual to forcefully shut down the device. The power of the device runs out up to 22 times faster than its normal condition by exploiting the wireless networks protocol vulnerabilities (Racic et al. 2006). The *water torture* attack (Johnston and Walker 2004) is another example of battery exhaustion attack carried out at the physical (PHY) layer that forces the device to send bogus frames. In addition, low-end mobile devices can be forced to shut down by receiving huge amount of SMS.

- **Privacy leakage**
  Many mobile apps collect user data without users' permission such as address book or location information. A social media app was found to download users' full address books including names, phone numbers, and email addresses without users' knowledge or consent when enabling "find friends" feature. It is reported that mobile apps on iPhone and iPad send personal information to advertising networks without the user consent too (Whitney 2010). Another most concerned privacy leakage is the location-related information. With the social networks apps becoming prevalent in our daily life, location-specific content have become more accessible and personalized to users' own context. Location tracking attack refers to attacker attempts to reveal the mobile device's location over time through examining their communications or hacking GPS information. Furthermore, if the business information is stored on the device, such invasions not only hurt the user's privacy but also increase the likelihood of security compromise to enterprise security.

- **Sniffing sensor information**
  Comparing to traditional cellphones that are mainly used for communications, modern smartphone is a sensor-rich device that can provide extended functions equipped with multiple sensors, e.g., camera, microphone, GPS, inertial motion unit (IMU), compass, step recorder, etc. Once the attacker can access those sensors without authorization, all the user's actions will be sniffed and recorded. The stealthy video capture spyware can secretly start the built-in camera to record the private video, and it consumes little power that will not draw any attention from the user (Xu et al. 2009). Another example of compromised sensor is the microphone; *Soundcomber* (Schlegel et al. 2011) managed to extract private data, such as the touch sound of PIN or phone number, from the microphone.

Secure Protection Guidelines

- **Advanced authentication**
  One-time validation of a user's identity, referred to as static authentication, has shown its vulnerability to attacks. Specifically, malicious adversaries may access the mobile

device that has been logged in by an authentic user when the authentic user is not nearby. Continuous authentication represents a new security mechanism which continuously monitors the user's trait and uses it as a basis to reauthenticate periodically throughout the login session. Hence, it has been adopted to overcome the limitations of traditional static authentication. These continuously monitored traits include typing habit and hand morphology (Feng et al. 2013), graphic touch gesture feature (Zhao et al. 2013), and cardiac motion and geometric information (Lin et al. 2017). Multifactor authentication is another alternative approach to enhance the security of the single trait authentication, which offers multifactor protections in case one credential has been compromised. For example, a set of behavioral biometrics of micro-movements and orientation patterns during hand movement, orientation, and grasp has been proposed for mobile authentication (Sitová et al. 2016). Controls such as one-time passwords, grid-based authentication, and digital-certificate-based authentication schemes can also help augment existing security solutions.

- **Malware detection**
  To protect our mobile devices, it is essential to detect the mobile malware and illegal activities including anomaly, misuse, or specification-based system. Basically, there are two kinds of detection techniques as static analysis and dynamic analysis. In static analysis, malicious codes are detected by unpacking and decompiling the application. "DroidMOSS" developed by Zhou et al. (2012) could generate fingerprint for an application and perform similarity test for two "same" apps; then DroidMOSS made sure the application is not affected or repacked with malware to avoid application-based threats on mobile devices, while the dynamic analysis identified the malicious behaviors by running the application on an emulator or a device. For example, monitoring power consumption is an easy way to detect whether malware are on your smartphone. Since the detection and analysis tasks utilize heavy resources of mobile devices, some of these tasks can be moved to the cloud for processing in the future.

- **Application development guidelines**
  The awareness of secure programming guidelines is vital to the secure application development by preventing the occurrence of common errors in programming. Some useful guidelines include performing secure logging and error handling; following the principle of least privilege; validating input data; implementing secure data storage; and avoiding insecure mobile OS features, etc. (Jain and Shanbhag 2012).

- **Hardware-associated protection**
  Hardware-associated protection implements a trusted execution environment by designing a hardware-enforced isolated execution environment for security-critical code, which provides protective mechanism in storing and processing sensitive data (Zhang et al. 2016). Techniques such as secure and authenticated boot (Ekberg et al. 2013) are employed to maintain a root of trust on the device. The most popular approach for mobiles is TrustOTP (trust one-time passwords) (Sun et al. 2015) that builds upon the ARM's TrustZone (Brasser et al. 2016) technology. Because it is a hardware-based protection, TrustOTP can prevent attacks that are in the mobile OS and even works when the mobile OS crashes.

References

Aviv AJ, Gibson KL, Mossop E, Blaze M, Smith JM (2010) Smudge attacks on smartphone touch screens. Woot 10:1–7

Beyah R, Venkataraman A (2011) Rogue-access-point detection: challenges, solutions, and future directions. IEEE Secur Priv 9(5):56–61

Bradley T (2011) Droiddream becomes android market nightmare. PCWorld
Brasser F, Kim D, Liebchen C, Ganapathy V, Iftode L, Sadeghi AR (2016) Regulating arm trustzone devices in restricted spaces. In: Proceedings of the 14th annual international conference on mobile systems, applications, and services. ACM, pp 413–425
CCC (2014) Fingerprint biometrics hacked again. http://www.ccc.de/en/updates/2014/ursel. Accessed by 13 May 2017
Celiktutan O, Avcibas I, Sankur B (2007) Blind identification of cellular phone cameras. In: Security, steganography, and watermarking of multimedia contents IX, international society for optics and photonics, vol 6505, p 65051H
Coursen S (2007) The future of mobile malware. Netw Secur 2007(8):7–11
Damopoulos D, Kambourakis G, Gritzalis S (2011) iSAM: an iPhone stealth airborne malware. In: IFIP international information security conference. Springer, pp 17–28
Ekberg JE, Kostiainen K, Asokan N (2013) Trusted execution environments on mobile devices. In: Proceedings of the 2013 ACM SIGSAC conference on computer & communications security. ACM, pp 1497–1498
Feng T, Zhao X, Carbunar B, Shi W (2013) Continuous mobile authentication using virtual key typing biometrics. In: 2013 12th IEEE international conference on trust, security and privacy in computing and communications (TrustCom). IEEE, pp 1547–1552
Jain AK, Shanbhag D (2012) Addressing security and privacy risks in mobile applications. IT Prof 14(5):28–33
Johnston D, Walker J (2004) Overview of ieee 802.16 security. IEEE Secur Priv 2(3):40–48
La Polla M, Martinelli F, Sgandurra D (2013) A survey on security for mobile devices. IEEE Commun Surv Tutor 15(1):446–471
Lin F, Song C, Zhuang Y, Xu W, Li C, Ren K (2017) Cardiac scan: a non-contact and continuous heart-based user authentication system. In: Proceedings of the 23rd annual international conference on mobile computing and networking (MobiCom 17), Snowbird, pp 315–328
Manabe H, Yamakawa Y, Sasamoto T, Sasaki R (2009) Security evaluation of biometrics authentications for cellular phones. In: Fifth international conference on intelligent information hiding and multimedia signal processing, 2009 (IIH-MSP'09). IEEE, pp 34–39
Notare MA, da Silva Cruz FA, Riso BG, Westphall CB (1999) Wireless communications: security management against cloned cellular phones. In: Wireless communications and networking conference, 1999 (WCNC 1999), vol 3. IEEE, pp 1412–1416
Racic R, Ma D, Chen H (2006) Exploiting mms vulnerabilities to stealthily exhaust mobile phone's battery. In: SecureComm, vol 6. Citeseer, pp 1–10
Schlegel R, Zhang K, Zhou Xy, Intwala M, Kapadia A, Wang X (2011) Soundcomber: a stealthy and context-aware sound trojan for smartphones. In: NDSS, vol 11, pp 17–33
Singh R, Bhargava P, Kain S (2007) Cell phone cloning: a perspective on gsm security. Ubiquity 2007:1
Sitová Z, Šeděnka J, Yang Q, Peng G, Zhou G, Gasti P, Balagani KS (2016) Hmog: new behavioral biometric features for continuous authentication of smartphone users. IEEE Trans Inf Forensics Secur 11(5):877–892
Statista (2018) Number of apps available in leading app stores as of 1st quarter 2018. https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/
Sujithra M Padmavathi G (2012) Mobile device security: a survey on mobile device threats, vulnerabilities and their defensive mechanism. Int J Comput Appl 56(14):24–29
Sun H, Sun K, Wang Y, Jing J (2015) Trustotp: transforming smartphones into secure one-time password tokens. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. ACM, pp 976–988
Whitney L (2010) Apple sued over privacy in iPhone, iPad apps. https://www.cnet.com/news/apple-sued-over-privacy-in-iphone-ipad-apps/
Xu N, Zhang F, Luo Y, Jia W, Xuan D, Teng J (2009) Stealthy video capturer: a new video-based spyware in 3G smartphones. In: Proceedings of the second ACM conference on wireless network security. ACM, pp 69–78
Zhang L, Zhu D, Yang Z, Sun L, Yang M (2016) A survey of privacy protection techniques for mobile devices. J Commun Inf Netw 1(4):86–92
Zhao X, Feng T, Shi W (2013) Continuous mobile authentication using a novel graphic touch gesture feature. In: 2013 IEEE sixth international conference on biometrics: theory, applications and systems (BTAS). IEEE, pp 1–6
Zhou W, Zhou Y, Jiang X, Ning P (2012) Detecting repackaged smartphone applications in third-party android marketplaces. In: Proceedings of the second ACM conference on data and application security and privacy. ACM, pp 317–326

Mobile Social Network

▸ Data-Driven Mobile Social Networks

Mobility

▸ Delay-Tolerant Network Routing

Mobility Management

▸ Mobility Management in NDN

Mobility Management in 5G

Dongmyoung Kim
AIX Center, SK Telecom, Jung-gu, Seoul, Korea

Synonyms

Location management; Reachability management

Definitions

Mobility management in cellular networks is a set of functions that the network uses to keep track of the location and status of the mobile device and to provide proper service to the mobile users.

Historical Background

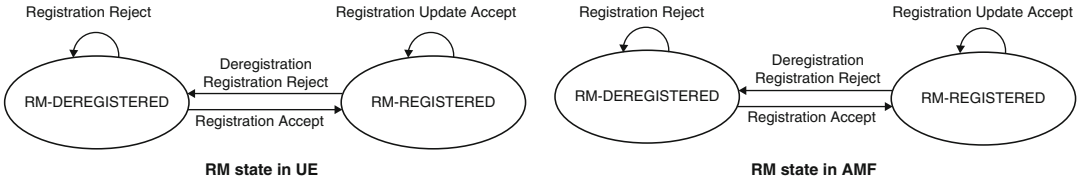
Mobility management in cellular networks is a set of functions that the network uses to keep track of the location and status of the mobile device and to provide proper service to the mobile users. In the cellular networks, mobile user equipment (UE) may move around whole network coverage, and it is expected that both Mobile Terminated (MT) and Mobile Oriented (MO) services are provided regardless of UE location and movement. Therefore, mobility management is one of the major functions in the cellular networks. Mobility management in cellular networks generally includes user equipment (UE) registration to the network, UE location tracking, paging to enable mobile terminated communication, support of handover, and so on. Additionally, in a wide sense, mobility management also includes the functions to support session continuity during user mobility, e.g., IP address allocation, traffic anchoring, tunneling, and so on. The earlier generations of 3GPP cellular networks, including 3GPP LTE system, have supported the mobility management function (3GPP 2017a). 3GPP 5G network also supports the mobility management functions similar to the ones supported in 3GPP LTE system with some enhancements as specified in 3GPP (2017b,c,d,e). This article summarizes the mobility management in 3GPP 5G network.

Key Mobility Management Functions in 5G Network

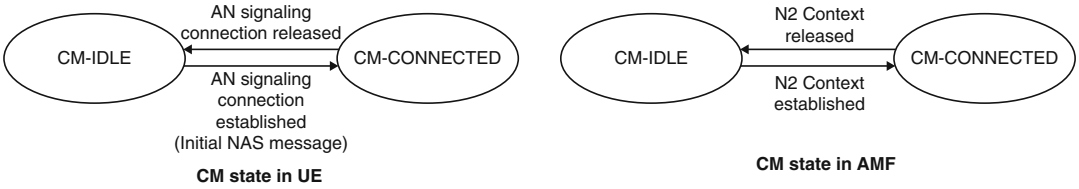
User State Management

In the cellular network, which mobility management functions and procedures are applied to a mobile user depend on state of the user. Three types of states are defined in the 5G system, i.e., Registration Management (RM) state, Connection Management (CM) state, and Radio Resource Control (RRC) state. RM and CM states are managed by the core network. The RM state represents whether a UE is registered in the 5G Core Network (CN), and the CM state shows whether Non-Access Stratum (NAS) signaling connection between UE and 5G CN is established. On the other hand, RRC state is managed by Radio Access Network (RAN), and it represents whether a connection between UE and 5G RAN exists or not. A UE in the RRC-CONNECTED state is in CM-CONNECTED state, and a UE in the RRC-IDLE state is in CM-IDLE state except during the transient phase, because radio link connection is required to establish NAS signaling connection to the core network (Figs. 1 and 2).

RRC-INACTIVE state is newly introduced in the RRC state model. The new state is proposed to be used as a primary sleeping state prior to RRC-IDLE state. When a UE moves to the new state, both the UE and RAN keep the context information of the UE's RRC connection, such as UE capabilities and security context, though the active radio link between the UE and RAN is released. The new state enables a lightweight transition from inactive to active



Mobility Management in 5G, Fig. 1 RM state model in 5G system



Mobility Management in 5G, Fig. 2 CM state model in 5G system

data transmission. A UE in the RRC-INACTIVE state is in CM-CONNECTED state except during transient phase. The 5G CN is not aware of the UE transitions between CM-CONNECTED with RRC-CONNECTED and CM-CONNECTED with RRC-INACTIVE state.

Handover and Cell Reselection

The choice of mobility procedure when the user moves to the coverage of other cell depends on RRC state of the user. The mobile UE in RRC-CONNECTED state uses a network-triggered procedure called handover for cell change, while mobile users in RRC-IDLE state or RRC-INACTIVE state uses a mobile-triggered procedure called cell reselection for cell change.

For the UE in RRC-CONNECTED state, Radio Access Network monitors the signal strength/quality from multiple cells to a user based on measurement report from the user and decides to trigger handover to serve the mobile user in a better cell. RAN and core network is aware of the UE mobility event and UE location after handover.

In the RRC-IDLE state or RRC-INACTIVE state where the user does not maintain active connection with RAN, the UE decides whether to camp on the current cell or to reselect a neighboring target cell based on signal strength measurements without interaction with the network. Therefore, the network may not be aware of the mobility event and exact location of the UE after cell reselection. Therefore, specific UE location tracking mechanisms are needed for the UEs in RRC-IDLE state or RRC-INACTIVE state, and they are described in the next section.

UE Location Tracking and Paging

UE location tracking is responsible for detecting whether the UE is reachable and providing UE location for the network to reach the UE. In most cellular networks, the location update procedure, which allows a mobile device to inform the cellular network when it moves from one location area to the next area, is used as the basic mechanism for location tracking of idle UEs, e.g., Tracking Area Update procedure in 3GPP LTE network. 5G network also supports such location update procedures.

In the 5G network, such functions can be located either at 5G CN (in case of RRC-IDLE state) or 5G RAN (in case of RRC-INACTIVE state). When a UE registers with the network, the network allocates a set of Tracking Areas (TAs) as a registration area of the UE. Each TA is composed of one or multiple cells, and each cell broadcasts the TA which the cell is

contained in. The UE in RRC-IDLE state moving to a new cell checks whether the new cell is still contained in the current registration area based on the broadcasted information. The idle UE triggers a registration update procedure only if the new cell is not contained in the current registration area. Accordingly, 5G (core) network keeps track of the location of UE in RRC-IDLE state in a registration area level.

In 5G system, RAN also needs to support the location tracking for the UE in RRC-INACTIVE state. In that state, RAN needs a new location tracking functionality to detect the location of the UE because the connection between the UE and the RAN is not active. RAN allocates RAN notification area, which is composed of multiple cells, to a UE, and the UE in RRC-INACTIVE state triggers a RAN notification area update procedure only if the new cell is not contained in the current RAN notification area. 5G (radio access) network keeps track of the location of UE in RRC-INACTIVE state in a RAN notification area level.

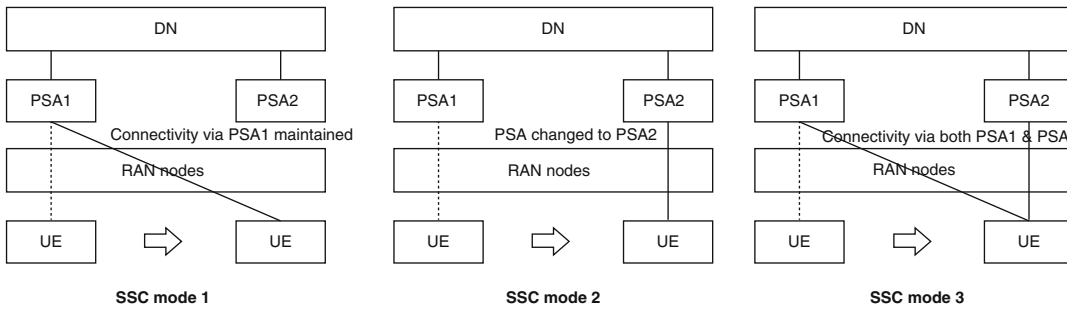
When downlink traffic arrives and the UE is in RRC-IDLE state or in RRC-INACTIVE state, the network (CN or RAN) triggers paging procedures with the consideration of the UE's expected location obtained by the UE location tracking procedures. Then, the UE which detected the paging message sends response to the network, and network gets to know the exact location of the UE, i.e., the current serving cell. The network provides the mobile terminated service to the detected UE location.

Dynamic Mobility Policy Enforcement

Policy Control and Charging (PCC) framework defined in the earlier generation of 3GPP system enables dynamic policy enforcement. In PCC framework of 3GPP LTE system, only session and QoS-related policies are supported. On the other hand, 5G PCC framework also supports mobility policy enforcement. Policy Control Function (PCF), which is the network function mainly managing the policy, can direct interact with AMF, which is the network function managing UE mobility, to support mobility-related policy enforcement. The mobility policy includes Service Area Restrictions and RAT/Frequency Selection Priority.

Support of Session and Service Continuity

Providing session and service continuity of an active session for mobile user is one of the most important tasks of cellular networks. To support session continuity, 5G system provides session anchoring in the gateway and tunneling (tunnel generation and update to the new RAN node on UE mobility) between the anchoring gateway and RAN node like the earlier generation of cellular networks. In 5G system the anchoring gateway (such as P-GW in 4G) is referred to as PDU session anchor (PSA) User Plane Function (UPF). The PSA terminates the user plane in the 5G core network and interfaces with the data network. In case of IP type of session, IP address preservation is an important aspect of session and service continuity, so that the PDU session anchor



Mobility Management in 5G, Fig. 3 SSC mode in 5G system

is responsible for IP anchoring as P-GW does in the EPC.

The main enhancement on session and service continuity in 5G system is that the different levels of session and service continuity procedures can be used according to the characteristics of each session. In the LTE system, continuity of IP session for all UEs is guaranteed in the whole system area. That is, the P-GW and the IP address of UE's PDU session are maintained regardless of the location of the UE such that the session continuity is maintained. On the other hand, 5G system aims to provide various levels of session continuity depending on the type of UE and type of service by allocating different Session and Service Continuity mode (SSC mode) per PDU session.

Three types of SSC modes are defined in 3GPP 5G system, namely, SSC mode 1, SSC mode 2, and SSC mode 3. Separate SSC modes are closely related to how the PDU session anchor is allocated and managed. In SSC mode 1, session continuity is guaranteed in all areas by keeping the PSA identical regardless of access network type and UE location as in the LTE system. In SSC mode 2, the same PSA is maintained across only a subset of the whole network. The PSA of a PDU session may be released and reallocated based on some criteria, e.g., to be served by the better UPF considering the UE's new point of attachment to the network. Lastly, SSC mode 3 enables the UE to connect to a new PSA before the connection with the existing PSA is released, and hence, the UE can receive services from two PSAs at the same time. By using this mode, various optimization can be supported, e.g., a newly created flow is served by a new PSA located in a preferred location while maintaining the existing flows in the previous PSA to provide session continuity (Fig. 3).

Cross-References

- 5G Wireless
- Future Wireless Network Architecture and Network Slicing
- Handoffs, Modeling in IP Networks
- Network-Layer Mobility Management

Acknowledgments This work was supported by the ICT R&D program of MSIP/IITP. [R7116-16-1001, "Development of 5G Core Network Technologies Standards"]

References

3GPP (2017a) 3GPP TS 23.401, general packet radio service (GPRS) enhancements for evolved universal terrestrial radio access network (E-UTRAN) access, Rel.15
3GPP (2017b) 3GPP TS 23.501, system architecture for the 5G system; stage 2, Rel.15
3GPP (2017c) 3GPP TS 23.502, procedures for the 5G system; stage 2, Rel.15
3GPP (2017d) 3GPP TS 38.300, NR; NR and NG-RAN overall description; stage 2, Rel.15
3GPP (2017e) 3GPP TS 38.331, NR; NR and NG-RAN overall description; stage 2, Rel.15

Mobility Management in LTE-U/LAA

- Handover in LTE-U/LAA

Mobility Management in NDN

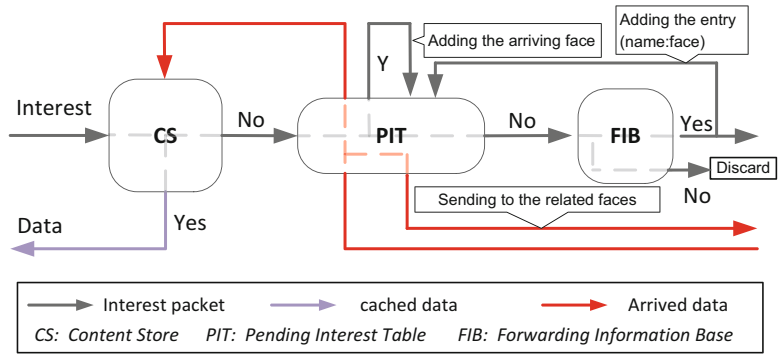
Zhiwei Yan
CNNIC, Beijing, China

Synonyms

Mobility management; NDN

Definition

Named Data Networking, as a clean-slate network model, is designed to support direct routing with content name in order to improve the security and efficiency for the future Internet. In NDN (Zhang et al. 2010), content such as a movie is divided into a set of individually named smaller content objects. Content object names are hierarchical and human-readable similar to the



Mobility Management in NDN, Fig. 1 NDN communication model

domain names in the current Internet, which can be arbitrarily long. All communications in NDN are consumer-initiated as shown in Fig. 1.

Thus, a consumer can retrieve an individual content object by sending a signaling message called an Interest packet which specifies the name of the desired content object. When a router receives an Interest packet and has a copy of the content object in its local content store (CS), the router sends back the copy without further propagating the Interest packet. If the router has no copy, it looks up the next-hop neighbor(s) in the forwarding information base (FIB) to forward the Interest packet to perform the longest prefix match of the name against its forwarding table. Whenever the Interest packet is relayed, a NDN router keeps the Interest packet in its Pending Interest Table (PIT) so that the response can be routed to the requester (consumer) along the reverse path.

Mobility management is the scheme to support the continued communication for both consumer and publisher during their movements. Because NDN does not need to establish the communication session with location-based information (such as IP address in the TCP/IP model), consumer mobility can be naturally supported. Then the mobility management scheme for mobile consumer mainly aims to reduce the handover latency and let the mobile consumer receive the packet as soon as possible from its new location. For mobile publisher who may announce the routing information to the NDN network, the mobility management scheme needs to guarantee the routing of Interest packets to the publisher's new location but avoid the large-scale route aggregations caused by the update of prefix location.

Historical Background

In basic NDN, after a publisher changes its location, the consumer may not access content from the publisher. This is due to the fact that Interest packets cannot be forwarded to the publisher by looking up the current FIB entry. Then the publisher has to update the FIB entries in all the NDN routers. Frequent updating of the FIB table will cause the NDN core network unstable and result in huge cost for the routing synchronization, especially when many publishers move frequently. Several approaches have been proposed to solve the problems related to the NDN producer mobility. They are mainly divided into three categories:

Tunnel-Based Redirection Scheme: The TBR scheme (Lee et al. 2012) is similar to Mobile IPv6 (MIPv6) (Johnson et al. 2011), which uses a tunnel to redirect incoming Interest packets between the mobile publisher (MP) and its home access router (AR-H). Specially speaking, after the MP moves, it sends a Prefix Update (PU) message to its AR-H, which contains the MP prefix and the MP's foreign access router (AR-F) prefix. Upon receiving the PU message, the AR-H records the binding information of the MP and AR-F prefixes. When the AR-H receives an Interest packet from a consumer that requires the MP's content, it is encapsulated with the MP's AR-F prefix and forwarded to the MP.

Routing-Based Approach: In the routing-based approach (Han et al. 2014), two faces corresponding to before and after handover exist in each entry of the FIB. When a router receives an Interest packet, the mobility face (corresponding to the state after handover) of the FIB is firstly checked. After the MP moves, it sends an informing control message to its AR-H. Upon receiving the informing control message, the AR-H sends back an updating control message to update the FIBs along the path from the AR-H to the MP. In this way, the Interest packet can be forwarded to the current MP.

Locator/Identifier Split Approach: In the locator/identifier split approach (Hermans et al. 2012), a new field called Location Name is added to the Interest packets. When an NDN router receives an Interest packet, the Location Name is firstly searched in the FIB. After the MP moves, it sends the binding information to the AR-H, and then the binding information with the MP prefix and its AR-H prefix are recorded in the AR-H. When the AR-H receives an Interest packet that requires the MP's content, it puts the AR-F prefix in the Location Name of the Interest packet. In this way, the Interest packet with the Location Name is forwarded to the MP. The content can be cached with its original content name in the routers.

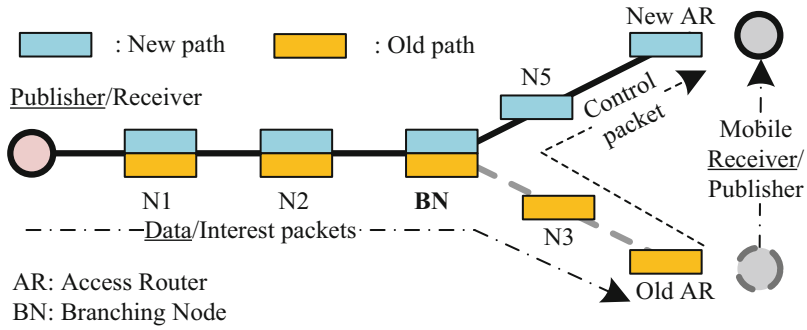
As stated above, the content name-based routing has no relationship with its location in NDN. Then the mobile consumer can reissue the Interests from the new location to support its mobility. However, how to use this "loose" routing scheme and on-path caching for efficient mobility management has a huge research space (Kim et al. 2012). Do-hyung Kim differentiated the real-time service and unreal-time service of NDN. Then a rendezvous point is deployed for the necessary location management for the mobile consumer (Lee et al. 2011). J. Lee and some other researchers proposed the proxy-based mobility management schemes for mobile consumer in NDN (Lee and Kim 2011; Oh et al. 2010; Wang et al. 2013), mainly to reduce the packet loss during the handover. The main operation can be illustrated as (1) after the consumer detects its movement, it immediately sends "Hold request" message to its proxy. (2) When receiving the "Hold request" message, the proxy stops delivering data packets and only stores the data packets in its local repository for subsequent retransmissions. (3) When the mobile consumer acquires new location information, it notifies the location information to its proxy node with "Handover notification" message. Then the proxy node transmits the stored data packets to the new location of the mobile consumer.

However, in reality, it's normally required to support mobility management for both consumer and publisher with the same logic because a node may act as consumer and publisher simultaneously. Z. W. Yan proposed a distributed mobility management solution (Yan et al. 2016), which can well support both mobile consumer and mobile publisher and even their simultaneous mobility as shown in Fig. 2.

In this solution, a new singling message named as "Control" message is designed in order to probe the branching node on the paths before and after handover, and it contains necessary information to indicate the new access router and the pending content. Then the on-path routers from the branching node to the new access router will adjust the routing information (PIT for mobile consumer and FIB for mobile publisher) accordingly. In this way, the mobile consumer can get the data as soon as possible after the handover. Besides, the established states can be well used during the handover and then the protocol cost can be reduced. Similarly, the publisher mobility is also based on the branching node probing and hop-by-hop state update, and then the signaling cost caused by the routing flooding can be reduced significantly.

Key Applications

With the development of mobile Internet, more and more traditional networking services will be deployed in the mobile environment. For NDN, the mobile applications supported by the above-mentioned mobility management schemes may be deployed in everywhere, for example, smart



Mobility Management in NDN, Fig. 2 Distributed mobility management in NDN

transport, data dissemination in partially connected environments, collaborative sensing, and presence-aware services (Siris et al. 2012).

Future Directions

With the development of wireless communication technologies, such as 5G, more and more content will be produced and consumed in the high-speed wireless networks. NDN provides a solution to support efficient routing which totally de-coupes the content identifier and its location. Then this distributed communication model asks for distributed mobility management scheme (Chan et al. 2014) to guarantee its efficiency and scalability.

Cross-References

- Distributed IP Mobility Management
- Handoffs, Modeling in IP Networks
- Mobile IP
- Mobility Management with ID/Locator Separation
- Proxy Mobile IPv6

References

Chan H, Liu D, Seite P, Yokota H, Korhonen J (2014) Requirements of distributed mobility management. RFC 7333

Han D, Lee M, Cho K, Kwon T, Choi Y (2014) Publisher mobility support in content centric networks. In: Proceedings of conference on information networking, Phuket, pp 214–219

Hermans F, Ngai E, Gunningberg P (2012) Global source mobility in the content-centric networking architecture. In: Proceedings of the 1st ACM workshop on emerging name-oriented mobile networking design-architecture, algorithms, and applications, South Carolina, USA

Johnson D, Perkins C, Arkko J (2011) IP mobility support for IPv6. RFC 6275

Kim D et al (2012) Mobility support in content centric networks. In: Proceedings of the ICN workshop on Information-centric networking, Helsinki, Finland

Lee J, Kim D (2011) Proxy-assisted content sharing using content centric networking (CCN) for resource-limited mobile consumer devices. IEEE Trans Consum Electron 57(2):477–483

Lee J, Kim D, Jang MW, Lee BJ (2011) Proxy-based mobility management scheme in mobile content centric networking (CCN) environments. In: Proceedings of the 29th International Conference on Consumer Electronics (ICCE), Las Vegas, USA

Lee J, Cho S, Kim D (2012) Device mobility management in content-centric networking. IEEE Commun Mag 50(12):28–34

Oh SY, Lau D, Gerla M (2010) Content centric networking in tactical and emergency MANETs. In: Proceedings of the 3rd IFIP, Venice, Italy

Siris V et al (2012) Content-centric architectures for moving objects. COST Action IC0906 WiNeMO white paper

Wang L, Waltari O, Kangasharju J (2013) MobiCCN: mobility support with greedy routing in content-centric networks. Proceedings of IEEE Globecom

Yan Z, Zeadally S, Zhang S, Guo R, Park Y (2016) Distributed mobility management in named data networking. Wirel Commun Mob Comput 16(13):1773–1783

Zhang L, Estrin D, Burke J, Jacobson V, Thornton JD, Smetters DK, Zhang B, Tsudik G, Claffy KC, Krioukov D, Massey D, Papadopoulos C, Abdelzaher T, Wang L, Crowley P, Yeh E (2010) Named data networking (NDN) project. NDN Technical Report NDN-0001

Mobility Management with ID/Locator Separation

Ved P. Kafle
Network System Research Institute, National Institute of Information and Communications Technology, Koganei, Tokyo, Japan

Synonyms

ID/locator separation-based mobility management; ID/locator split-based mobility management; Mobility management with ID/locator split

Definition

Mobility management includes network functions that make mobile devices remain connected to the network despite their movement from one place to another, and making them reachable irrespective of their location. Mobility management includes two types of functions: location update and handover. Location update function maintains reachability of mobile devices irrespective of their location, and handover function maintains connectivity during the time when mobile devices detach from one network and attach to another.

ID/locator separation is a concept of using two distinct values (i.e., identifier and locator) for the representation of identity and location of a mobile device. IDs are used in the application and transport layers to identify the communication endpoints or session states, while locators are used in the network layer to denote the location of mobile devices by the routing system in the network topology. This concept is different from the original concept of the Internet, where an IP address is used as both ID and locator. Mobility management with ID/locator separation becomes easier than in conventional IP networks because the change of locators during mobility does not require changing IDs being used in the transport session states.

Historical Background

Mobility was not considered when designing the Internet about 40 years ago. It was assumed that the IP address of a device, host, or node (Note: device, host, and node are used interchangeably in this article) remained static during a communication session. The IP address has been used in the role of both identifiers and locators. Namely, the IP address is used in the network layer protocols as a locator to locate the device in the routing system and forward packets towards it. The same IP address is also used in the transport and application layers to identify the communication sessions or endpoints. Actually, this dual semantic of IP address is hindering the Internet from supporting mobility natively because when the mobile device moves, it has to change its IP address, resulting in terminating the session identified by the IP address.

To enable a mobile device continue its communication session even after moving from one network to another in the Internet, Mobile IP protocols have been developed and standardized by the Internet Engineering Task Force (IETF) in 2002 (for IP version 4, or IPv4) and in 2004 (for IP version 6, or IPv6). Mobile IP protocols allow mobile devices to possess two types of addresses: home address and care-of address. Home address is a persistent address obtained from the home network prefix and used in application and transport layers, whereas care-of address is obtained from a visiting or foreign network and used temporarily in the network layer as long as the mobile device resides in the visiting network. Mobile IP protocols are based on the concept of "map and encapsulate," where the IP address, i.e., home address, existing in the IP header of packets destined to the home network is mapped with the care-of address by searching in a mapping table, and IP packets are encapsulated by another IP header containing the care-of address in the destination address field and forwarded to the mobile device's current location.

Mobile IP protocols are complicated due to the necessity of maintaining home agents for managing up-to-date mapping between home addresses and care-of addresses and tunneling

packets. Mobility management with ID/locator separation can be simpler and more efficient because it allows a single ID to dynamically associate with various locators at the same time or different instances of time. It has the potential to natively support heterogeneous types of network-layer protocols and multihoming. These features enable to perform make-before-break or seamless handover easily without requiring any network-side anchor points and tunnel management.

Key Proposals

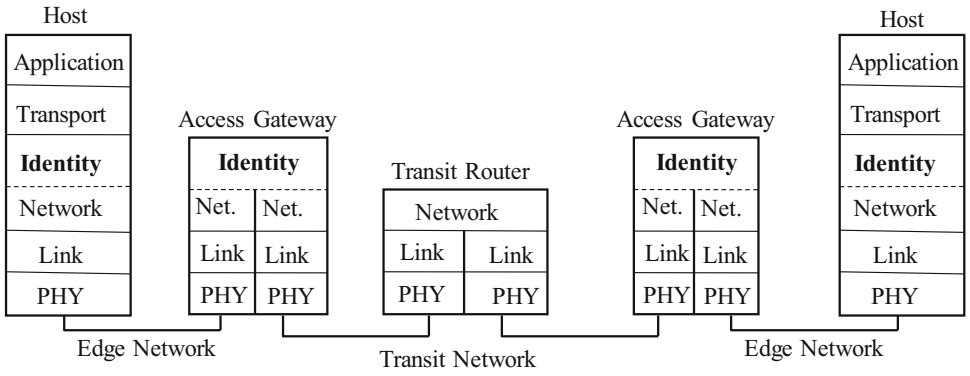
ID/locator separation-based network technologies and standards have recently been developed by IETF, which has published several RFCs on Host Identity Protocol (HIP), Locator ID Separation Protocol (LISP), and Identifier-Locator Network Protocol (ILNP). Similarly, research projects in various countries have also proposed ID/locator separation-based networking technologies, such as Heterogeneity Inclusion and Mobility Adaptation through Location ID Separation (HIMALIS) in Japan, Mobile Oriented Future Internet (MOFI) in Korea, and MobilityFirst in the USA. Since these proposals are based on the same concept of ID/locator separation, they have many common features, which are discussed next.

ID/locator separation-based architectures use nontopological values as the identifiers of mobile hosts. They use topologically significant values as locators to represent the location of mobile hosts in the network. The set of locator values associated with a host change when the host's connectivity in the Internet topology changes due to mobility; however, the identifier does not change. Dynamic bindings exist between identifiers and locators values, which are updated as locators change. The bindings are maintained inside the host as well as stored in a mapping service system such as the Domain Name System (DNS).

Figure 1 depicts a union of protocol stacks of all ID/locator split architectures mentioned above. This figure shows the existence of a new layer, which is often known as identity layer, between the transport and network layers in the host and access gateway protocol stacks. The identity layer isolates the applications and transport layers from the underlying network layer and performs ID-to-locator mapping. Some proposals (such as ILNP and MOFI) do not explicitly mention about the existence of an independent identity layer, but assume similar functions existing in the network layer. In host-based ID/locator separation architectures (such as HIP), the identity layer exists only in the host protocol stack and in the network-based architectures (such as LISP), the identity layer functions exist in the access gateways only, while in the hybrid architectures (such as HIMALIS and MOFI) these functions do exist in both hosts and access gateways.

In ID/locator separation architectures, the application and transport layer protocols use IDs to represent communication endpoints and session states. IDs have end-to-end significance and they do not change due to mobility. On the other hand, locators are used in the network layer for routing and forwarding of packets. The locator values appear in the network header (i.e., IP header) of packets and may have only local significance and change as the packets pass through the access gateways. The set of locator values associated with a node also change when the node's connectivity in the Internet topology changes due to mobility; however, the identifier does not change, thus the transport and transport layers session states remain unchanged after mobility.

Mobility management with ID/locator separation requires involvement of the identity layer, the network layer, as well as the lower layers, when a mobile device moves from one network to another. The lower layers perform the lower-layer handoff functions such as detecting that the radio signal coming from a new network is getting stronger than that from the currently connected network and setting up radio channels and parameters for a new connection. Similarly, the network layer configures a new locator based on the addressing scheme used in the new network and notifies the identify layer about the new locator. The identity layer performs the location update signaling function to update the ID-to-locator mapping cached in the correspondent host



Mobility Management with ID/Locator Separation, Fig. 1 ID/locator protocol stack generic representation

and stored in the mapping server system. After the location update, ongoing communication sessions of the mobile host continue from the new network. Moreover, since ID/locator separation architectures naturally support multihoming by associating the same ID with multiple locators, the mobile host can momentarily be multihomed during the handover time by setting a new link and locator in the new network while still holding the link and locator from the old network. Seamless handover is easily achieved by having the mobile host perform location update in the correspondent host from the new network while still being connected with the old network.

These ID/locator separation-based architectures have some subtle differences in terms of the ID type and configuration methods, ID-to-locator mapping database system, and components handing mobility management functions. These are discussed next.

HIP

HIP architecture and protocol have been specified in RFCs 4423 and 7401, respectively. HIP prescribes to use public keys (and their hash values) as host IDs and IP addresses as locators. HIP extends the Domain Name System (DNS) records to store host IDs as a new resource record type. A host acquires its peer host's ID and locator by sending a domain name resolution request to a DNS server. After obtaining the ID and locator of a mobile host, the correspondent host initializes communication with the mobile host by exchanging few control packets (known as Base Exchange). Both the source and destination hosts' IDs appear in the identity header and locators in the network header of control packets, while the subsequent data packets include IPsec header instead of the identity header.

For mobility management, HIP introduces rendezvous servers as the anchor points that store the current locators of mobile hosts. Rendezvous server's locators are registered in the DNS server as the mobile host's global locators. The first control packet of an incoming communication request from a correspondent host to the mobile host arrives at the rendezvous server and gets forwarded to the mobile host. The mobile host notifies the correspondent host by sending information about its current locator in the subsequent control packet so that the next packets sent from the correspondent host reach the mobile hosts directly, i.e., without requiring the rendezvous server in the path. On changing its locator due to mobility, the mobile host sends a location update signaling message to update the ID/locator mapping cache of the correspondent host as well as the rendezvous server. In this way, the mobile host can communicate from the new network as well as maintain its reachability by updating the ID/locator mapping stored in the rendezvous server.

ILNP

ILNP architecture has been specified in RFC 6740. It divides 128 bits long IPv6 address into

two 64 bits long parts. The lower 64 bits have been used as identifiers, while the higher 64 bits are used as locators. ILNP assumes that "well-behaved" applications that use FQDN or other nontopological identifiers at the application layer, rather than IP addresses or lower layer identifiers, will perceive no architectural difference between IP and ILNP. Moreover, the ILNP packet format in wire is similar to IP packets; thus, ILNP can operate without requiring ID/locator separation supporting access gateways. However, it assumes the existence of routers that make the routing and forwarding decision on the basis of only the upper 64-bits values of the destination IP address present in the packet header. ILNP also proposes to add new resource record types to DNS server to store the NID (i.e., 64 bits node identifier), L32 (i.e., 32 bits locator), and L64 (i.e., 64 bits locators).

ILNP does not introduce any anchor point for the mobility management. It rather makes the mobile host to perform locator update signaling with the correspondent host as soon as it detects that its locator value has changed. The mobile host is also expected to update its locator values stored in the DNS server. However, since the DNS updates take a longer time to propagate, the mobile host may be unreachable for a while from new correspondent hosts.

LISP

LISP has been specified in RFC 6830. It uses prefix aggregatable endpoint IDs (EIDs), which are also used as locators in the edge network. In the transit network, routing locators (RLOC) are used as locators. EIDs to RLOCs mapping takes place in Ingress and Egress Tunneling Routers (ITR/ETR) located at the border between the edge and transit networks. Both EIDs and RLOCs are syntactically identical to IP addresses; they are different in semantics of how they are used. EID-to-RLOC mapping records are maintained in an overlay network of mapping database, called LISP Alternate Logical Topology (LISP-ALT).

The application and transport layers use EIDs, which also appear in the IP header of packets dispatched from the host. When the packets reach the ITR, the destination EID value present in the IP header is mapped with a related RLOC obtain from the LISP map server. The IP packets are then encapsulated with LISP header, UDP header, and outer IP header containing the RLOCs in source and destination address fields. When the LISP packets reach at the ETR, their encapsulating outer IP header, UDP header, and LISP header are removed and simply forwarded in the destination edge network by the inner IP header containing EIDs.

Since it also uses EIDs as local locators in edge networks, LISP does not provide host mobility support natively. To provide host-mobility, it proposes to make the LISP mobile host look like a LISP site and possess a lightweight version of the ITR/ETR functions. When the mobile host moves from one network to another, it receives a new RLOC and updates the ITRs that are encapsulating packets from correspondent hosts by using the previous mapping, as well as the map server. The mobile host may use one of the various methods such as by sending a solicit-map-request message or piggybacking mapping data to update the mapping. Nonetheless, LISP lacks the procedure to complete the mapping update process in a short time, thus does not guarantee smooth handover.

HIMALIS

HIMALIS architecture (Kafle and Inoue 2010; Kafle et al. 2014) has the protocol stack exactly as shown in Fig. 1. The identity layer exists in the end hosts as well as in the access gateways that connect the edge network with the transit network. The transit routers are simply IP routers. HIMALIS IDs are 128 bits values containing organizational prefix, scope, and version bits in the prefix.

In HIMALIS, the identity layer obtains mapping between IDs and locators from the mapping system consisting of Domain Name Registry (DNR) and Host Name Registry (HNR) through a name resolution process. That is, these registries are used to resolve hostnames to IDs and locators during a communication initialization phase. The HIMALIS packets contain the identity header along with the transport and network headers. The identity header contains the source

and destination IDs which do not change end-to-end, while the network layer contains the source and destination locators (i.e., IP address) that do change as the packet passes through the access gateway. Having said so, the network header containing the local IP addresses is valid only within the edge network. Thus, a HIMALIS host knows only the local IP addresses of its own and uses the gateway's local IP address as the destination IP address in the network header of outgoing packets. The packets are routed by using local IP addresses in the edge network. As the packets reach the access gateway, their network header is removed and ID present in the identity header is searched in the mapping table to find the corresponding global locators. These global locators are used in the new network header to be attached to the packet. These global locators are used by the core router to route packets in the transit network.

In the destination access gateway, the global IP header is removed and a new local IP header, containing the destination host's and the access gateway's local IP addresses in the destination and source addresses, respectively, is attached to the packet. The packets reach the destination host.

HIMALIS supports mobility across various types of network protocols and differently scoped IP addressing in the edge networks. For example, one edge network can use IPv4 local addressing scheme, while the other edge network can use IPv6 addressing scheme, yet the hosts located in these networks can communicate to each other as the local network protocols are translated by the access gateways. In this way, the application and transport layer functions and session states are completely isolated from the network layer so that changes of network layer protocols and IP address as the host moves from one network to another due to mobility would have no adverse impact on the application sessions if the network layer readdressing is performed instantly during movement. As the host moves to a new network and obtains a new locator from the new access gateway, it immediately performs location update with the correspondent host by sending a signaling message containing its global locator (i.e., locator of the new access gateway) so that the correspondent host starts forwarding packets to the new network. At the meantime, the mobile host also informs its previous access gateway about its new locator so that the previous access gateway forwards packets destined to the mobile host to the new access gateway. In this way, HIMALIS attempts to achieve lossless handover.

MOFI

The MOFI architecture (Kim et al. 2013) is designed with three functional features in consideration: global identifier and local locator, query-first data delivery, and distributed control of identifier-locator mapping. MOFI uses global host IDs and local or private IP addresses as locators, and mapping is stored in distributed hash-based ID-locator mapping registers. MOFI does not mention about an explicit identity layer, but divides the network layer into two sublayers: communication and delivery. The communication sublayer is responsible for end-to-end communication, and the delivery sublayer is responsible for routing and forwarding in the access and backbone networks. The communication sublayer performs signaling operation for the locator search from the mapping registers collocated with access gateways.

When the mobile host moves from an old access gateway to a new access gateway, it obtains a new locator. As in HIMALIS, the mobile host informs the new access gateway about the information of the old access gateway so that they establish a handover tunnel to forward packets from the old access gateway to the new access gateway. Then the mobile host sends a location update message to the correspondent host. The mapping registers are also updated with the new locator.

MobilityFirst

The MobilityFirst architecture (Raychaudhuri et al. 2012) focuses on making the network centered on mobility and trustworthiness. It has proposed a globally unique ID (GUID) namespace for network attached objects such as smartphones, people, a group of devices/people, content, or even context. GUIDs are formed from

public keys assigned by a name certification service to the objects. It also introduces a name-based service layer that uses the GUIDs to enable mobility-centric services while ensuring security and trustworthiness. It prescribes a hybrid name/address based routing scheme, employing a fast global name resolution service (GNRS) to dynamically bind the destination GUID to a set of network addresses or locators. Every packet includes GUID in its header so that network nodes can offer GUID-based redirection or late binding to network addresses. On mobility, network addresses do change while GUID remains the same. The change in GUID to address mapping is reflected in GNRS by a mapping update signaling process.

Future Directions

There are several issues related to the deployment of ID/locator separation-based mobility architectures. Among them, security, ID namespace, ID-to-locator mapping directory service, and interoperability with existing IP networks are briefly discussed below.

Security: ID/locator separation-based architectures are considered to enhance security features of mobile networks because they allow mobile devices to use topology-independent names or identifiers to be used in the representation of security context which remain valid despite changing the network layer addresses and locators. HIP, MobilityFirst, and HIMALIS have indicated security as the prominent feature of the new architectures.

ID namespace: Introduction of a network topology-independent IDs and their management is an important issue. Different proposals have considered different types of IDs such as public keys and their hash values (in HIP and MobilityFirst), lower 64-bits of IPv6 addresses (in ILNP), IP address blocks not advertised in BGP routers (in LISP and MOFI), and new IDs (in HIMALIS). A harmonized approach is yet to be initiated.

ID-to-locator mapping directory service: Secured and scalable ID-to-locator mapping directory service that can provide the mappings in a very low latency while being able to keep the mapping records up-to-date despite frequent changes in locators triggered by mobility is a challenging issue. Different proposals have considered different approaches to maintain the mapping directory service, such as HIP extends DNS records and introduces rendezvous servers, LISP specifies the LISP-ALT, and HIMALIS proposes to use HNRs.

Interoperability with IP network: Due to introduction of new ID handling functions, the ID/locator separation-based mobility management networks need additional proxy functions for interoperability with each other and with the traditional IP networks. For example, there are proxy functions being proposed for HIP, LISP, HIMALIS, and MOFI. The proxy functions can exist either in host or in gateway or both.

Besides providing better mobility management, ID/locator separation-based mobility architectures are also favorable in supporting heterogeneous types of network layer protocols, enhancing security, alleviating BGP routing overhead, and making architecture extendible. These issues have been researched and standardized in various standard development organizations (SDOs) such as IETF and ITU. Although we listed only IETF standards on ID/locator separation networking protocols in the previous sections, there are similar work in ITU (for detail, please see ITU-T Recommendations Y.2015, Y.2057, Y.3032, and Y.3034).

Cross-References

- ▸ Distributed IP Mobility Management
- ▸ Mobile IP
- ▸ Mobility Management in 5G
- ▸ Network-layer Mobility Management

References

Kafle VP, Inoue M (2010) HIMALIS: heterogeneity inclusion and mobility adaptation through locator id separation in new generation networks. IEICE Trans Commun E93-B(3):478–489

Kafle VP, Fukushima Y, Harai H (2014) ID/locator split-based distributed mobility management mechanism. Wirel Pers Commun 76(4):693–712
Kim J-I, Jung H, Koh S-J (2013) Mobility Oriented Future Internet (MOFI): architecture design and implementation. ETRI J 35(4):666–676
Raychaudhuri D, Nagaraja K, Venkataramani A (2012) MobilityFirst: a robust and trustworthy mobility-centric architecture for the future internet. ACM SIGMOBILE Mobile Comput Commun Rev 16(3):2–13

Mobility Management with ID/Locator Split

▸ Mobility Management with ID/Locator Separation

Mobility Session

▸ Proxy Mobile IPv6

Mobility Support in IP Multicast

▸ Mobility in Multicast

Mobilty in Multicast

Seil Jeon
Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea

Synonyms

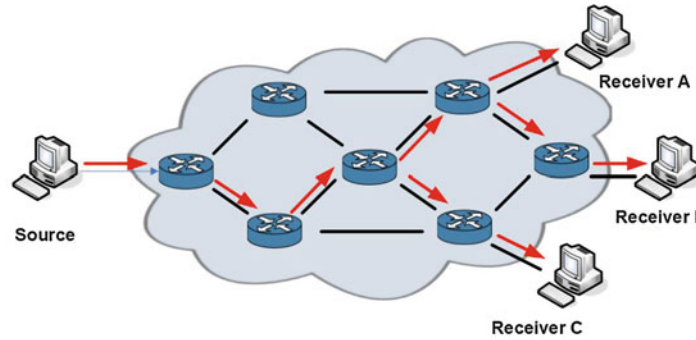
Mobility support in IP multicast; Multicast mobility

Definitions

Mobility in multicast means the technical ability to provide mobility support in IP multicast communications. IP multicast session can be resumed after the change of point of attachment by a mobile terminal, without mobility management consideration. But the traditional IP multicasting has two technical issues in a mobility event; one is that the multicast session will be disrupted. Second, a join procedure for the multicast session takes a considerable time for the multicast membership update and multicast routing path update after the IP connection is established. To avoid such time-consuming procedure and session disruption by mobility in IP multicast communication, mobility support mechanisms for IP multicast are required. Depending on mobility management protocol such as Mobile IP or Proxy Mobile IPv6, or other mobility protocols over which IP multicast works, design criteria for mobility management extension are considered.

Historical Background

IP multicast is a well-known transmission technique that efficiently can deliver an IP packet from a sender to multiple receivers by duplicating the packet over IP multicast-enabled networks, without creation and management of multiple IP sessions per content request. Setting up a multicast session is powered by two main technologies: (i) multicast membership group management and (ii) multicast routing. A multicast group is a set of network devices identified by a common multicast address. For the receivers interested in receiving the IP packets of a certain multicast group, it is required to subscribe the membership group using Internet Group Management Protocol (IGMP) for IPv4 (Cain et al. 2002) or Multicast Listener Discovery (MLD) for IPv6 (Vida and Costa 2004). After the interested multicast group information is gathered by the local multicast router, it should send a join message with the multicast group information to the upstream multicast router for establishing a multicast branch between the two multicast routers on the requested multicast group, using multicast routing protocol such as PIM-SM and PIM-DM (Fenner et al. 2016; Adams et al. 2005) or IGMP/MLD Proxy (Fenner et al. 2006) (Fig. 1).



Mobilty in Multicast, Fig. 1 IP multicast transmission to multiple receivers

As wireless/mobile network is evolving and growing with the demand of IP multimedia service, mobility in multicast began to study in earnest with the advent of Mobile IP for the IP session continuity support of a mobile terminal. There are a variety of solutions for mobility in multicast, addressing and/or mitigating the tunnel convergence issue, join latency, point of failure, and so on. In Romdhani et al. (2004), IP multicast support mechanisms over mobile networks were comprehensively investigated, dealing with technical challenges and solutions for mobile source and mobile receiver over Mobile IP (MIP) and its variants such as Fast Mobile IPv6 (FMIPv6) and Hierarchical Mobile IPv6 (HMIPv6).

As Proxy Mobile IPv6 (PMIPv6) addressing the limitations of the host-based mobility protocols (i.e., Mobile IP) appeared (Gundavelli et al. 2008), the design approach for IP multicast support over PMIPv6 has been investigated in IETF Multicast Mobility (MULTIMOB) WG, with the aim of specifying IP multicast support and optimization solutions aligned with PMIPv6 networks (Schmidt et al. 2011; Zuniga et al. 2013) and handling the behavior of IGMPv3/MLD for wireless networks and mobility (Asaeda et al. 2012). All the produced RFCs in IETF MULTIMOB WG are as follows:

- RFC 6224: Base Deployment for Multicast Listener Support in Proxy Mobile IPv6 (PMIPv6) Domains
- RFC 6636: Tuning the Behavior of the Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) for Routers in Mobile and Wireless Networks
- RFC 7028: Multicast Mobility Routing Optimization for Proxy Mobile IPv6
- RFC 7261: Proxy Mobile IPv6 (PMIPv6) Multicast Handover Optimization by the Subscription Information Acquisition through the LMA (SIAL)
- RFC 7287: Mobile Multicast Sender Support in Proxy Mobile Ipv6 (PMIPv6) Domains
- RFC 7411: Multicast Listener Extensions for Mobile IPv6 (MIPv6) and Proxy Mobile IPv6 (PMIPv6) Fast Handovers

For tackling the problems and limitations of the centralized mobility management, i.e., MIP and PMIPv6, Distributed Mobility Management (DMM) has appeared in IETF DMM WG, analyzing the issues of centralized mobility management and extracting the requirements (Chan et al. 2014). In the requirements, IP multicast issues and problems in the centralized mobility management are analyzed, and design considerations for IP multicast over DMM are given.

Key Points

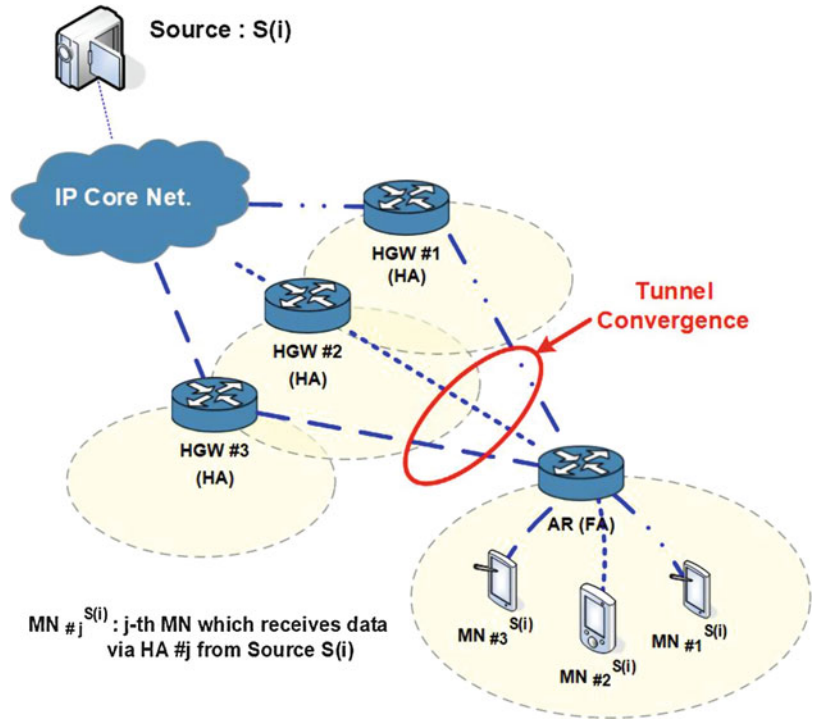
For IP multicast deployment over the mobile networks, it is essentially required to install one of the two IP multicast agents or both into the mobility management entities. The one is an Internet Group Management Protocol (IGMP)/Multicast Listener Discovery (MLD) Forwarding Proxy (in short IGMP/MLD Proxy) (Fenner et al. 2006); the other is a

multicast routing protocol. Both or each of them can be used for providing multicast service over wireless/mobile networks. Though the mobile network can be deployed with both multicast membership protocol and multicast routing protocol, the multicast membership protocol can solely be used and installed in the mobility management entities, and the multicast traffic distribution in the core network can be implemented and deployed by an operator-specific solution. On the other hand, the multicast routing infrastructure can solely be used for multicast packet distribution over the multicast core network, while the subscription of a certain multicast group can be implemented by an operator-specific solution. The IGMP/MLD Proxy provides an available option to associate the mobile network with a multicast infrastructure. It is lightweight compared to the multicast routing protocol and does not require the multicast routing protocol support. Due to the reason, it has been tried to use IGMP/MLD Proxy in PMIPv6 networks (Schmidt et al. 2011; Zuniga et al. 2013). In any cases chosen from the deployment options, mobility management in multicast should consider the following design issues and KPIs.

- *Multicast Traffic Duplication*: IP multicast works per group, not per terminal, while IP mobility management is provided and handled per terminal with a tunneling method. IP multicast aims to provide an efficient transmission of the same packet for multiple terminals. Suppose that there are multiple terminals attached at the same mobility access router and they are supported by the IP tunnels from their home networks in an IP mobility architecture. IP multicast subscription on the same multicast channel for individual mobile terminal will be delivered through each tunnel and multicast traffic will be duplicate. This problem was described in PMIPv6-based networks with a proposed solution to tackle the problem in Jeon et al. (2009), which has been contributed for an optimized IP mobile multicast solution (Zuniga et al. 2013). This problem has been identified in a distributed mobility environment, as analyzed in Jeon et al. (2012) (Fig. 2).
- *Join Latency*: To receive a copy of IP multicast packet, a receiver should join a group by initiating a multicast membership request when it learns of the group using the membership subscription protocol (IGMP or MLD). The IP multicast router sends a join message to the upstream multicast router to join the group requested from the receiver. Then, an IP multicast branch is established between the two multicast routers. This process takes a considerable time, so reducing the join latency is of importance for seamless service continuity in IP mobile multicast. For addressing the issue, tuning the behavior of IGMP and MLD protocols is required, since the behaviors and parameters of the protocols are originally specified for wired networks (Asaeda et al. 2012). Sending the IP multicast context information to a predicted mobility access router was suggested by extending mobility protocol (Jeon et al. 2009; Figueiredo et al. 2015) or using context transfer protocol (Von Hugo and Asaeda 2013).

Key Applications

- Mobile IPTV: Mobile IPTV is a popular application where mobility in multicast can be provided. Nowadays, for receiving multimedia streaming such as drama, movie, and news, IP multicast with mobility can be considered for watching streaming contents, particularly in commute time.
- Personal Broadcasting Service (PBS): personal broadcasting service where a mobile sender broadcasts the real-time streaming in a street or moving car is going to be popular. The multicast mobility architecture for PBS was designed and demonstrated in Medieval (2012).
- Multimedia Broadcast/Multicast Service (MBMS): MBMS is the 3GPP-based multicast/broadcast platform, which enables multicast/broadcast services for mobile



Mobilty in Multicast, Fig. 2 Multicast traffic duplication problem in a mobile IP-based network

terminals. MBMS aims for efficiency of multimedia traffic transmission on the same region covered by a base station (BS). However, due to the economies of channel popularity per BS region and dedicated MBMS radio frequency, it is somewhat deprecated.

Cross-References

- Distributed IP Mobility Management
- Network-Layer Mobility Management

References

Adams A, Nicholas J, Siadak W (2005) Protocol independent multicast-dense mode (PIM-DM): protocol specification (Revised). IETF RFC 3973

Asaeda H, Wu Q, Liu H (2012) Tuning the behavior of the internet group management protocol (IGMP) and multicast listener discovery (MLD) for routers in mobile and wireless networks. IETF RFC 6636

Cain B, Deering S, Kouvelas I, Fenner B, Thyagarajan A (2002) Internet group management protocol, version 3. IETF RFC 7761

Chan HA et al. (2014) Requirements for distributed mobility management. IETF RFC 7333

Fenner B, He H, Haberman B, Sandick H (2006) Internet group management protocol (IGMP)/multicast listener discovery (MLD)-based multicast forwarding (ÏGMP/MLD Proxying"). IETF RFC 4605

Fenner B, Handley M, Holbrook H, Kouvelas I, Parekh R, Zhang Z (2016) Protocol independent multicast-sparse mode (PIM-SM): protocol specification (Revised). IETF RFC 7761

Figueiredo S, Jeon S, Gomes D, Aguiar RL (2015) D3M: multicast listener mobility support mechanisms over distributed mobility anchoring architectures. J Netw Comput Appl 53:24–38

Gundavelli S, Leung K, Devarapalli V, Chowdhury K, Patil B (2008) Proxy mobile IPv6. IETF RFC 5213

Jeon S, Kang N, Kim Y (2009) Mobility management based on proxy mobile IPv6 for multicasting services in home networks. IEEE Trans Consum Electron 55(3):1227–1232

Jeon S, Figueiredo S, Aguiar RL (2012) A channel-manageable IP multicast support framework for distributed mobility management. In: IFIP wireless days 2012, Dublin

Medieval F (2012) MEDIEVAL D1.3: final architecture design

Romdhani I, Kellil M, Lach HY, Bouabdallah A, Bettahar H (2004) IP mobile multicast: challenges and solutions. IEEE Commun Surv Tutorials 6(1):18–41
Schmidt T, Waehlisch M, Krishnan S (2011) Base deployment for multicast listener support in proxy mobile IPv6 (PMIPv6) domains. IETF RFC 5213
Vida R, Costa L (2004) Multicast listener discovery version 2 (MLDv2) for IPv6. IETF RFC 3810
Von Hugo D, Asaeda H (2013) Context transfer protocol extension for multicast. Draft-vonhugo-multimob-cxtp-extension-03
Zuniga JC, Contreras LM, Bernardos CJ, Jeon S, Kim Y (2013) Multicast mobility routing optimizations for proxy mobile IPv6. IETF RFC 7028

Modeling Approaches for Simulating Molecular Communications

L. Felicetti, M. Femminella, and G. Reali
Department of Engineering, University of Perugia, Perugia, Italy

Synonyms

Molecular communication simulators; Simulation of diffusion-based molecular communications

Definition

Molecular communication involves biological entities that are able to transmit and receive molecules, which represent the information signal.

Historical Background

The molecular communications (MolCom) represent an emerging research area that consists of transmission of information by means of exchange of molecules, carried out by either natural or artificial nanomachines. The physical mechanisms that allow transferring information at such small scales are typically inspired by the biological mechanisms that exist in living bodies to exchange many types of signaling molecules, such as proteins, pheromones, and immune system activation signals, both within and between different cells. The diffusion-based MolCom are the most studied mechanisms; they are based on the molecule propagation in the fluid medium according to the laws of diffusion (Philibert 2006), without the presence of any flow or drift.

In recent years, a multitude of significant breakthroughs occurred in some strategic socioeconomic fields (Akyildiz et al. 2008). They encompassed multidisciplinary research, bringing together information and communication technologies, molecular biology, physics, chemistry, biotechnology, environmental control, and material science. Together, they will allow realizing the vision of transferring information within biological environments at extremely small scales, down to a size comparable to that of molecules. In this case MolCom are considered an alternative approach to electromagnetic communications due to their unique features of biocompatibility and minimal invasiveness, which are essential for them to be used in living bodies.

Key Applications

Up to now, most interesting potential applications have been identified in the biomedical field (Felicetti et al. 2016). The emerging applications of MolCom techniques focus on both emulations and the development of body area nanonetworks oriented to the diagnosis and treatment of diseases.

As for the first one, the emulation of nanoscale biological processes allows achieving personalized predictions of the evolution of diseases, starting from a limited number of biomarkers, avoiding any traditional in vivo tests on patients (Hood et al. 2011; Bragazzi 2013) helping medical personnel to identify optimal treatments. The computational models are based on a detailed characterization of the molecular communication parameters done by using results of in vitro and in vivo experiments.

For the latter one, the design of nano-sensors/actuators could allow the detection and treatment of a large set of diseases (e.g., cardiovascular diseases, tumors, etc.). In fact, these nano-devices could activate the immune system and/or trigger drug delivery systems in small specific areas without affecting the rest of the body (i.e., smart drugs).

However, other fields, such as food production, functionalized materials and fabrics, as well as environmental and military applications, can be envisioned (Akyildiz et al. 2008).

Communication Models

The most simple and energy-efficient propagation model, without any external cause, is the diffusion-based molecular communication system. In the last years, several researches have been focused on the theoretical analysis and mathematical modeling of the propagation medium, on the receiver modeling (Yilmaz et al. 2014), and on the reception techniques.

In diffusion-based systems, the transmit signal is encoded on the physical characteristics of information molecules (e.g., cytokines, hormones, DNA). Those are able to randomly propagate in the fluid medium via diffusion spreading in the environment according to the negative gradient of the concentration with a velocity that depends on the thermal energy and on the particle size.

The information can be encoded on the concentration of emitted particles, on the frequency of emission, and on the type of released molecules.

The information molecules can interact with the receiver node in several ways, changing or not their concentration around the receiver. The simplest reception model assumes the ideal passive (or transparent) receiver which is permeable to the hitting molecules and is capable only to count the number of the molecules inside its volume. This means that the hitting molecules may unavoidably contribute to the received signal more than once in different symbol intervals and the receiver may encounter high intersymbol interference (ISI) (Deng et al. 2017).

In a more realistic molecular communication system, each molecule is removed from the environment after the adsorption, and it contributes only once to the received signal. The adsorption takes place through the surface receptors that cover the external membrane of the receiver node. These receptors are able to react only with specific types of information molecules. In more detail, these receptors may adsorb or bind with these molecules. In the first case, the chemical complex formed by the molecule (known also as ligand) and the receptor is internalized by the nanomachine and follows a process known as trafficking (Lauffenburger and Linderman 1996). On the latter case, it may happen that some molecules desorb from their receptors and may contribute several times to the received signal.

In what follows, the main receiver models for a 3D diffusion-based molecular communication system in a fluid environment are analyzed, from the more complex and general to the simplest one. As already mentioned, in the most general case, the information molecules released from the transmitter can be adsorbed and desorb from the surface of the receiver, and the net number of adsorptions is counted for information decoding. This general model will be introduced first, and then the other models will be derived from it.

The overall system consists of a point transmitter and a spherical receiver in an unbounded homogeneous environment. The point transmitter is located at a distance r_0 from the center of the receiver and is at a distance $d = r_0 - r_r$ from the nearest point on the surface of the receiver with radius r_r. The surface of the receiver node is assumed to be covered by a number of compliant receptors that can adsorb at most one molecule at a time. It is assumed that the spherical receiver has no physical limitation on the number or placement of receptors on its surface. Thus, there is no limit on the number of molecules adsorbed to the receiver surface. This is an appropriate assumption for a sufficiently low number of adsorbed molecules, or for a sufficiently high concentration of receptors.

Emission Process

The point transmitter emits the information molecules at $t = 0$. Fick's diffusion equation describes the propagation of the information molecules in the environment:

$$C(r, t \to 0 | r_0) = \frac{1}{4\pi {r_0}^2} \delta(r - r_0), \quad (1)$$

where $C(r, t \to 0 | r_0)$ is the molecule distribution function at time $t \to 0$ and distance r from the center of the receiver with initial distance r_0. The first boundary condition is:

$$\lim_{r \to \infty} C(r, t | r_0) = 0, \quad (2)$$

such that at arbitrary time, the molecule distribution function equals zero when r goes to infinity.

Propagation via Diffusion

The released information molecules randomly diffuse in the medium by means of Brownian motion (Berg 1993) colliding with other molecules. In the case in which the concentration of information molecules could be assumed to be sufficiently low, the collisions between information molecules could be ignored (Berg 1993). This means that each information molecule diffuses independently with constant diffusion coefficient D.

Fick's second law describes the propagation model in a 3D environment (Yilmaz et al. 2014; Philibert 2006):

$$\frac{\partial C(r, t | r_0)}{\partial t} = \nabla^2 (D \cdot C(r, t | r_0)) - k_d C(r, t | r_0). \quad (3)$$

In (3), we have also included the negative contribution of molecule degradation. In fact, signaling molecules released in the environment may degrade with a certain probability and transform into molecules not recognized by the receiver. k_d is the degradation reaction coefficient ($time^{-1}$), which is the time constant of this chemical phenomenon. In the following, we assume no molecule degradation, i.e. $k_d = 0$.

A more general case considers also the presence of a non-negligible flow in the environment, which often is also constrained in a limited space. This case, that will be not treated in this chapter for space limitation, is well modeled by the advection-diffusion equation (Gentile et al. 2008).

Reception

A generic receiver node should be capable of adsorbing incoming molecules that hit its surface or reflect them back into the fluid environment, based on the adsorption rate k_1 ($length \times time^{-1}$), which depends on the chemical affinity of each molecule with the surface receptors of the receiver node. The adsorbed molecules either desorb or remain stationary at the surface of the receiver, based on the desorption rate k_{-1} ($time^{-1}$).

The adsorption and desorption reactions that can occur on its surface give the second boundary condition of the information molecules:

$$D \frac{\partial (C(r, t | r_0))}{\partial r}\bigg|_{r=r_r^+} = k_1 C(r_r, t | r_0) - k_{-1} C_a(t | r_0), \quad (4)$$

where $C_a(t | r_0)$ is the average concentration of molecules that are adsorbed to the receiver surface at time t. The change in the adsorbed concentration over time is equal to the flux of diffusion molecules toward the surface:

$$\frac{\partial C_a(t | r_0)}{\partial t} = D \frac{\partial (C(r, t | r_0))}{\partial r}\bigg|_{r=r_r^+}. \quad (5)$$

Combining (4) and (5), the radiation boundary condition is obtained, and it shows that the equivalent adsorption rate is proportional to the surface molecule concentration:

$$\frac{\partial C_a(t | r_0)}{\partial t} = k_1 C(r_r, t | r_0) - k_{-1} C_a(t | r_0). \quad (6)$$

At $t = 0$, there are no information molecules at the receiver surface, so the second initial condition is:

$$C\left(r_r,\, 0 \middle|\, r_0\right) = 0,$$
$$C_a\left(0 \middle|\, r_0\right) = 0. \tag{7}$$

Depending on the values assumed by both k_1 and k_{-1}, it is possible to define how the receiver node works, as defined below:

- Both k_1 and k_{-1} are non-zero finite constants: (4) is the boundary condition for the adsorption/desorption receiver. A complete transient and steady-state solution for this quite general case can be found in Deng et al. (2015). A variation of this model, taking into account also the receptor occupancy status, is illustrated in Pierobon and Akyildiz (2011).
- k_1 is a non-zero finite constant and $k_{-1} = 0$, and (4) is the boundary condition for the partial absorbing receiver (or receiver with finite receptors). The partially absorbing receiver with finite receptors fits with the realistic case in which the complex formed by the ligand and the receptor is internalized. Clearly, this may happen only when a molecule binds to one of the compliant surface receptors; otherwise it will be reflected away. A quite complete treatment of the solution of this specific case can be found in Akkaya et al. (2015) and references therein.
- $k_1 \to \infty$ and $k_{-1} = 0$: (4) is the boundary condition for the fully adsorbing receiver. In this configuration, each colliding molecule with the surface of the receiver will be adsorbed without any further desorption. This configuration models the case in which the surface of the receiver node is largely covered by receptors compliant with ligand molecules. A quite complete treatment of the solution of this model can be found in Yilmaz et al. (2014).
- Finally, a further popular model, which is slightly different from those presented above, is the transparent (or virtual) receiver. In this model, a receiver is able to simply count the molecules crossing its volume, *without any interaction* with them (i.e., adsorption, desorption, or even bounces). Clearly, in this case the coefficients k_1 and k_{-1} are not used, and the only boundary condition is (2). For this type of receiver, an interesting analysis is presented in Llatser et al. (2013).

The probability of finding a molecule at distance r and time t is given by the time-varying spherically symmetric spatial distribution $C\left(r,\, t \middle|\, r_0\right)$. The cumulative fraction of adsorbed molecules at time t can be expressed as:

$$F\left(\Omega_{r_r}, t \middle| r_0\right) = \int_0^t 4\pi r_r^2 D \left.\frac{\partial C\left(r, \tau \middle| r_0\right)}{\partial r}\right|_{r=r_r} d\tau \tag{8}$$

where Ω_{r_r} represents the surface of the spherical receiver with radius r_r. Upon an emission of Q molecules at time $t_0 = 0$, the average number of adsorbed molecules after t seconds will be $N\left(\Omega_{r_r}, t \middle| r_0\right) = QF\left(\Omega_{r_r}, t \middle| r_0\right)$. The asymptotic concentration of adsorbed molecules can be found as $t \to \infty$ for the different formulations of $N\left(\Omega_{r_r}, t \middle| r_0\right)$, depending on the values of k_1 and k_{-1}. As for the transparent receiver, the solution is simpler, and the average number of "counted" molecules is given by $N\left(\Omega_{r_r}, t \middle| r_0\right) = C\left(0, t | r_0\right) V_r$, where V_r is the volume of the spherical transparent receiver.

For a treatment more focused on the chemical species involved into the adsorption and desorption reaction, that is the molecules or ligands, the receptors, and the complex obtained by their binding, the interested reader can refer to Lauffenburger and Linderman (1996).

Mapping Models on Simulation Platforms

BiNS Biological and Nanoscale Communication Simulator

General description BiNS is a simulation package for MolCom systems developed at the University of Perugia (Felicetti et al. 2012).

Its customizable design provides a set of tools which allow creating objects modeling the behavior of biological entities. They can be regarded as either nodes (transmitters, receivers), carriers, or

surrounding objects (e.g., vessel walls). In addition, BiNS permits to configure the properties of the simulated communication channel (e.g., the blood stream or the environment of an in vitro experiment) with the desired accuracy.

The simulator has been implemented in Java and contains generic type of software object, named Nano Object. Nodes and carriers are specific implementations of the Nano Object, and, although they share its general features, they can exhibit very different functions.

The simulation is organized in discrete time steps. Each step consists of a number of phases, in which software objects are triggered in order to execute the operations associated with their specific behavior. The main phases are:

- transmission phase
- reception phase
- information processing phase,
- motion phase
- object destruction phase (during which objects are removed due to lifetime expiration or because they exited from the area of interest)
- collision management phase, which implements the elementary interaction between particles.

BiNS has been validated through wet-lab experiments related to cardiovascular medicine.

Technical requirements and scalability The list of simulated nano-objects is split into smaller lists, which can be handled in parallel by a thread pool.

The simulator uses a fine-grained approach for handling collisions between objects. A collision can produce either a bounce or an assimilation. The latter happens only when a carrier collides with a compliant receptor on the surface of a node. The receptors can implement a non-negligible absorption time; thus a compliant particle could find a receptor in either busy or free state. The simulated environment can be either unbounded or bounded by a surface of custom shape, referred to as simulation domain. For example, in order to simulate communications within a blood vessel, a cylindrical volume was used. Within a domain, the octree paradigm allows distributing the workload associated with the management of the objects in the domain volume to different threads. There is an ongoing effort for the GPU porting of the collision handling algorithm in order to reduce the computational time significantly.

N3Sim

General description N3Sim (Llatser et al. 2011) is a Java-based complete simulation framework for diffusion-based molecular communications, which allows the evaluation of the communication performance of molecular networks with several transmitters and receivers in an infinite space with a given concentration of molecules. The transmitters encode the information by releasing particles into the medium, thus varying the concentration rate in their vicinity. The diffusion of particles through the medium is modeled as Brownian motion, taking into account particle inertia and collisions among particles. Finally, the receivers decode the information by sensing the local concentration in their neighborhood. It implements a three-layer architecture. The user interface layer interacts with the user to read the input data for the simulation, while the data layer writes the simulation results to files. The domain layer contains the intelligence of the system, i.e., the molecular communication model. N3Sim allows automating the execution of multiple simulations in a simple manner by means of user-defined scripts.

Technical requirements and scalability N3Sim is designed to be executed on a single machine with Java 1.6 or higher. The simulator has been tested both on Linux and Windows. The scalability of the simulator can be improved by selecting a larger value of the simulation time step (at the expense of the accuracy) or by deactivating the collisions among molecules in scenarios with a low molecular concentration. As an example, simulations with up to 10^6 simultaneously emitted molecules have been

successfully performed. The time granularity of a simulation is defined by the user by choosing the simulation time step in the configuration file (typically a few *ms*). The simulation space can be either bounded or unbounded, and both two-dimensional and three-dimensional configurations are supported.

NS-3 Based Simulators

General description NS-3 is a discrete-event network simulator for Internet systems, which was not originally developed for MolCom simulators. Nevertheless, its flexible structure has allowed implementing some basic elements of MolCom. In particular there are two simulation tools that have been developed in NS-3: one in the framework of the IEEE P1906.1 working group (Bush et al. 2015; IEEE Std 1906.1-2015 2016) and another specific implementation for bacteria communications (Jian et al. 2017).

NS-3 is organized in different software libraries that can work together. User programs, written in C++ or Python programming languages, can adapt these libraries or be linked with them. External animators and data analysis and visualization tools are available. Nevertheless, in order to exploit the full potentials of the NS-3, the command line interface is suggested.

Technical requirements and scalability The simulator is designed to be executed on a single machine. The minimal requirements to execute basic simulations are a gcc or clang compiler and Python interpreter. The simulation granularity is defined by users. Event design can be adapted to any granularity and is synchronized in the simulator implementation processing loop.

AcCoRD

General description The Actor-based Communication via Reaction-Diffusion (AcCoRD) is an open-source generic reaction-diffusion MolCom simulator developed in C language (Noel et al. 2017). The actors can be both transmitter and receivers, and the simulation environment is defined as a collection of microscopic and mesoscopic regions that differ for the simulation accuracy and the computational cost. In the microscopic region, each molecule is individually modeled in each time step, whereas in the mesoscopic one, the total amount of molecules in each region is analyzed. Chemical reactions (e.g., molecule degradation, reversible/irreversible surface binding, etc.) can be defined in the propagation environment.

Technical requirements and scalability Microscopic and mesoscopic regions can be combined and dimensioned according to the simulation needs, balancing the local accuracy with the computational cost. Also a visualization tool developed in MATLAB is included in the simulator allowing an offline visualization of the simulated environment.

Simulators for Specific MolCom Environments

In addition to the general simulation platforms presented in the section before, there are a number of more specific simulation packages targeted to specific environments. Examples of these simulators include platforms dealing with bacteria colonies (Jian et al. 2017; Wei et al. 2013) introduced also in section "NS-3 Based Simulators", calcium signaling (Barros et al. 2015), or nervous systems (Malak et al. 2014).

MolCom Simulators Functional Comparison

A comparison of the main simulators that have been developed to characterize the broad set of MolCom systems is sketched in Table 1, where the compatibility with the different receiver models introduced above is illustrated.

However, this is a rough classification of the available simulation tools. In fact, more realistic and refined models can be simulated only by a subset of these ones. For instance, when the interactions between molecules are significant, diffusion equations may become nonlinear (Aranovich and Donohue 2005), regardless of the receiver model. Also, the partial adsorbing receiver can be modeled with a finite adsorption time for each receptor, which is more realistic.

Modeling Approaches for Simulating Molecular Communications, Table 1 Compatibility of the most popular MolCom simulators with the reception models

Simulator	Adsorption/desorption receiver	Partial adsorbing receiver	Full adsorbing receiver	Transparent receiver
BiNS[a]	Easy to upgrade[e]	Yes	Yes	Yes
N3Sim[b]	No	No	Yes	Yes
NS-3[c]	No	No	No	Yes
AcCoRD[d]	Yes	Yes	Yes	Yes

[a]Available on http://conan.diei.unipg.it/lab/index.php/product/biological-nanoscale-simulator-bins2
[b]Available on http://www.n3cat.upc.edu/tools/n3sim/download
[c]Available on https://www.nsnam.org/releases/
[d]Available on https://warwick.ac.uk/fac/sci/eng/staff/ajgn/software/accord/downloads/
[e]A minor upgrade is necessary to implement desorption by applying the Algorithm 1 illustrated in Deng et al. (2015)

These more complex but also realistic systems can be simulated by BiNS2. At the same time, the desorption functionality is already implemented in AcCoRD, whereas BiNS2 needs an upgrade to implement it.

Cross-References

- Applications of Molecular Communication Systems
- Drug Delivery via Nanomachines
- Molecular Communication for Wireless Body Area Networks
- Nanonetworks
- Reaction-Diffusion Channels
- Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion

References

Akkaya A et al (2015) Effect of receptor density and size on signal reception in molecular communication via diffusion with an absorbing receiver. IEEE Commun Lett 19(2):155–158. https://doi.org/10.1109/LCOMM.2014.2375214

Akyildiz IF et al (2008) Nanonetworks: a new communication paradigm. Comput Netw 52(12):2260–2279. https://doi.org/10.1016/j.comnet.2008.04.001

Aranovich GL, Donohue MD (2005) Diffusion equation for interacting particles. J Phys Chem B 109(33):16062–16069

Barros T et al (2015) Comparative end-to-end analysis of ca2+-signaling-based molecular communication in biological tissues. IEEE Trans Commun 63(12):5128–5142. https://doi.org/10.1109/TCOMM.2015.2487349

Berg H (1993) Random walks in biology. Princeton University Press, Princeton

Bragazzi NL (2013) From p0 to p6 medicine, a model of highly participatory, narrative, interactive, and "augmented" medicine: some considerations on Salvatore Iaconesi's clinical story. Patient Prefer Adherence 7:353–359. https://doi.org/10.2147/PPA.S38578, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3640773/, ppa-7-353[PII]

Bush SF et al (2015) Defining communication at the bottom. IEEE Trans Mol Biol Multi-Scale Commun 1(1):90–96. https://doi.org/10.1109/TMBMC.2015.2465513

Deng Y et al (2015) Modeling and simulation of molecular communication systems with a reversible adsorption receiver. IEEE Trans Mol Biol Multi-Scale Commun 1(4):347–362. https://doi.org/10.1109/TMBMC.2016.2589239

Deng Y et al (2017) Analyzing large-scale multiuser molecular communication via 3-d stochastic geometry. IEEE Trans Mol Biol Multi-Scale Commun 3(2):118–133. https://doi.org/10.1109/TMBMC.2017.2750145

Felicetti L et al (2012) A simulation tool for nanoscale biological networks. Nano Commun Netw 3(1):2–18

Felicetti L, Femminella M, Reali G, Lio P (2016) Applications of molecular communications to medicine: a survey. Nano Commun Netw 7:27–45

Gentile F et al (2008) The transport of nanoparticles in blood vessels: the effect of vessel permeability and blood rheology. Ann Biomed Eng 36(2):254–61

Hood L et al (2011) Predictive, personalized, preventive, participatory (p4) cancer medicine. Nat Rev Clin Oncol 8:184 EP https://doi.org/10.1038/nrclinonc.2010.227, perspective

IEEE Std 19061-2015 (2016) IEEE recommended practice for nanoscale and molecular communication framework. IEEE Std 19061-2015, pp 1–64. https://doi.org/10.1109/IEEESTD.2016.7378262

Jian Y et al (2017) nanoNS3: a network simulator for bacterial nanonetworks based on molecular communication. Nano Commun Netw 12:1–11. https://doi.org/10.1016/j.nancom.2017.01.004, http://www.sciencedirect.com/science/article/pii/S1878778916300941

Lauffenburger D, Linderman J (1996) Receptors: models for binding, trafficking, and signalling. Oxford University Press, New York

Llatser I et al (2011) Exploring the physical channel of diffusion-based molecular communication by simulation. In: IEEE GLOBECOM 2011. https://doi.org/10.1109/GLOCOM.2011.6134028

Llatser I et al (2013) Detection techniques for diffusion-based molecular communication. IEEE J Sel Areas Commun 31(12, supplement): 726–734

Malak D et al (2014) Communication theoretical understanding of intra-body nervous nanonetworks. IEEE Commun Mag 52(4):129–135. https://doi.org/10.1109/MCOM.2014.6807957

Noel A et al (2017) Simulating with accord: actor-based communication via reaction diffusion. Nano Commun Netw 11:44–75. https://doi.org/10.1016/j.nancom.2017.02.002, http://www.sciencedirect.com/science/article/pii/S1878778916300618

Philibert J (2006) One and a half century of diffusion: Fick, Einstein, before and beyond. Diffus Fundam 4:6.1–6.19

Pierobon M, Akyildiz I (2011) Noise analysis in ligand-binding reception for molecular communication in nanonetworks. IEEE Trans Signal Process 59(9):4168–4182

Wei G et al (2013) Efficient modeling and simulation of bacteria-based nanonetworks with BNSim. IEEE J Sel Areas Commun 31(12):868–878. https://doi.org/10.1109/JSAC.2013.SUP2.12130019

Yilmaz HB et al (2014) Three-dimensional channel characteristics for molecular communications with an absorbing receiver. IEEE Commun Lett 18(6):929–932. https://doi.org/10.1109/LCOMM.2014.2320917

Modulation in Molecular Signaling

H. Birkan Yilmaz[1], M. Şükrü Kuran[2], and Ilker Demirkol[1]
[1]Department of Telematics Engineering, Universitat Politecnica de Catalunya, Barcelona, Spain
[2]Department of Computer Engineering, Abdullah Gul University, Kayseri, Turkey

Synonyms

Modulation techniques for *synthetic* molecular communications

Definitions

Molecular signaling is a way of communicating via use of molecules. There are still many applications and environments where the classical technologies are not convenient or appropriate. Inspired by nature, one possible solution to these problems is to use chemical signals as carriers of information, which is called molecular communication (MC). Modulation in *synthetic* MC is the way of embedding the information on one or more properties of molecule emission process.

Historical Background

In 1959, Richard Feynman gave a lecture at American Physical Society meeting in Caltech that is titled "There is Plenty of Room at the Bottom" (Feynman 1959). The talk was pointing out the manipulation and controlling things at a small scale with a huge potential to advance the knowledge frontier. Over the past couple of decades there have been considerable advancements in the fields of nanotechnology, biotechnology, and microrobotics, where designing and engineering of microscale or nanoscale devices begin to take shape. On the other hand, the cooperation and coordination of these tiny devices necessitates the miniaturization of communication networks, in which one of the techniques is MC.

Molecular signaling/communication is at its infancy phase having been first proposed in 2005 from an engineered communication perspective (Nakano et al. 2005). Many theoretical and simulation-based studies are published in the molecular communications (MC) domain. In this entry, the works related to MC modulation techniques are given in chronological order and from simple to complex systems. Readers may feel the chronological order while reading the entry.

Background on Molecular Communication

In MC, molecules are utilized to convey information among communication nodes at micro-

and macroscales (Farsad et al. 2016b). Some of the microscale MC systems can be listed as quorum sensing in bacteria (Cobo and Akyildiz 2010; Abadal and Akyildiz 2011), neurotransmitter signaling in neuromuscular junctions (Kuran et al. 2013), and inter-cell calcium signaling (Nakano et al. 2005; Kuran et al. 2012). Examples of macroscale MC systems in nature include pheromone signaling among plants, chemical signaling among animals, and odor tracking of blue crabs in the ocean (Zimmer and Butman 2000).

In MC, information molecules propagate in a fluid environment via various processes (such as diffusion), and they arrive at the receiver node in a probabilistic manner in which the received molecules constitute the received molecular signal (Farsad et al. 2016b). In its basic form, an MC system mainly consists of a transmitter node, fluid environment, information molecules, and a receiver node (Nakano et al. 2013; Farsad et al. 2016b). In the following subsections, first the common performance metrics for MC channels are given followed by the types and dynamics of the modulation techniques.

Common Metrics for Molecular Communication

One of the first metrics to evaluate the modulation techniques that have been used in MC literature is the system response. This metric mainly focuses on the number of received molecules with respect to time. It is generally used to show the effects of different channel types (e.g., free diffusion, vessel-like environment, constrained diffusion) and transmitted molecular signal waveforms (e.g., sinusoidal, square) on the received signal.

Other common metrics can be listed as bit error rate (BER), symbol error rate (SER), and the channel capacity. In the MC literature, as a performance metric, BER or SER is usually evaluated against several system parameters such as signal-to-noise ratio (SNR), transmitted power, symbol duration, communication distance, and diffusion coefficient. As for SNR, the key challenge is to clearly define what is the definition of noise in this domain and which noise sources to consider (Nakano et al. 2013). Channel capacity metric is evaluated based on the given error rates and the channel model and provides the achievable data rate of the communication system in question. A key point regarding the correctness of this metric is the selection of the appropriate channel model. Currently, most of the MC literature assume a memoryless MC channel to evaluate the channel capacity. However, as described by Genc et al., due to the diffusion-based dynamics of the MC channel, the memoryless assumption is far from appropriate in the MC domain, and instead a model with memory is required (Genc et al. 2016).

After briefly mentioning these common performance metrics, the rest of the entry will be focused on the types of modulation techniques used in MC systems. At the end of this entry, a table of studies are given in a table with also mentioning the metrics used in each study.

Modulation Techniques

All MC modulation techniques depend on releasing special molecules called messenger molecules (MM) from the transmitter and modulating the bit values of a given message upon the features of these MMs. There are works in the MC literature on the selection of feature(s) to be used to represent the information. These works can be broadly classified into four groups:

Concentration-based Techniques Information is embedded upon the varying concentration level of the transmitted signal.

Type-based Techniques Information is embedded upon the type of the MMs of the transmitted signal.

Timing-based Techniques Information is embedded upon the MM release time instances.

Hybrid Techniques Techniques utilizing more than one of the three features (concentration,

type, timing) of the transmitted signal to represent information.

Concentration-Based Techniques

The main idea of concentration-based techniques is carrying information on varying released MM concentration over fixed period discrete time slots (i.e., symbol durations, t_s). In its simplest form, each symbol represents one bit value and is called on-off keying (OOK). In OOK, the transmitter releases a fixed number of MMs (i.e., n_1) if the corresponding bit value (k-th symbol $S[k]$) is bit-1 or releases no molecules at all if it is bit-0 Mahfuz et al. (2010). At the receiver side, the receiver counts the number of MMs arrive within each symbol duration (i.e., $N^{Rx}[k]$) and makes a threshold-based decision to decode the bit value of the given symbol duration ($\hat{S}[k]$).

A more generalized version of OOK is called concentration shift keying (CSK) where each symbol, depending on the system design, represents m-bits of information (Kuran et al. 2011). From a communication point of view, CSK is akin to amplitude modulation (AM) method in analog modulation and amplitude shift keying (ASK) in digital modulation. In CSK, for the kth symbol in the message, the transmitter releases $N^{Tx}[k]$ number of MMs depending on the current symbol value as

$$N^{Tx}[k] = n_{S[k]}, \quad S[k] \in (0, 1, \ldots, 2^m - 1), \tag{1}$$

where $S[k]$ can take one of the 2^m symbol values (e.g., sym_0, sym_1) in the modulation alphabet and $n_{S[k]}$ denotes the number of molecules to emit for the symbol $S[k]$. In order to decode $\hat{S}[k]$ from the received signal, the receiver uses $2^m - 1$ thresholds (i.e., $\lambda_0, \lambda_1, \ldots, \lambda_{2^m-2}$) as

$$\hat{S}[k] = \begin{cases} \text{sym}_0, & N^{Rx}[k] < \lambda_0 \\ \text{sym}_i, & \lambda_i \leq N^{Rx}[k] < \lambda_{i+1} \\ \text{sym}_{2^m-1}, & \lambda_{2^m-2} \leq N^{Rx}[k] \end{cases}. \tag{2}$$

Type-Based Techniques

A second group of modulation techniques, called type-based techniques, focuses on using multiple types of MMs in the communication system as a basis of a modulation technique. In these techniques, the transmitter has the capability of releasing different types of MMs (mm_{type}) that can only be received by a particular type of receptors at the receiver surface.

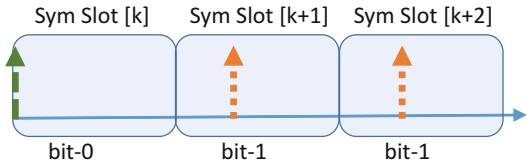
In the first type-based modulation technique, molecular shift keying (MoSK), each different symbol value (sym_i) is represented by a specific type of MM (Kuran et al. 2011; Aminian et al. 2015; Galmés and Atakan 2016). During each symbol duration, the receiver counts the number of arriving MMs for each MM type and decodes $\hat{S}[k]$ based on thresholding or the majority decision:

$$\hat{S}[k] = \text{sym}_i, \quad \text{where } i = \arg\max_r N^{Rx}_{mm_r}[k]. \tag{3}$$

Compared to CSK, MoSK considerably reduces the intersymbol interference (ISI) element of the signal due to the fact that each bit value is represented by a different type of molecule. On the other hand, it increases the system complexity since the transmitter is required to synthesize different types of molecules and the receiver is required to have multiple receptors on its surface. Moreover, molecule types should be similar in terms of diffusion coefficient due to propagation dynamics and single symbol duration.

Timing-Based Techniques

The main idea of timing-based modulations is to encode the information on the release time of the information molecules. In its simplest form, emission of MMs is done at an instance in the symbol slot where the time instance can be between the start and the end of the symbol slot. The time instance represents the intended symbol, and the receiver aims to detect the emission time hence the intended symbol. Capacity limit of timing-based techniques is derived and analyzed in Farsad et al. (2016a) and Murin et al. (2016).



Modulation in Molecular Signaling, Fig. 1 Example scenario showing the emission times with the bit sequence 011

Timing-based approach is also called pulse position modulation (PPM), which is similar to PPM in optical communications (Garralda et al. 2011). In the binary case of timing-based techniques, the transmitter simply emits molecules at t_{sym_0} for sym_0 (which is usually at the start of the symbol duration) and emits molecules at t_{sym_1} for sym_1 (which is usually around the mid symbol duration) (Fig. 1).

Another release timing-based modulation scheme is called time-elapse communication (TEC) and is proposed for slow networks such as on-chip bacterial communication (Krishnaswamy et al. 2013). In TEC, information is encoded in the time interval between two consecutive pulses (i.e., two consecutive pulses of information particle transmission), which is similar to pulse-width modulation (PWM). At the receiver side, the information is decoded by measuring the duration between two pulses. The authors showed that TEC can outperform on-off keying when techniques such as differential coding are used.

Hybrid Techniques

The three modulation techniques addressed above constitute the main categories of modulation techniques for MC. These techniques, which utilize the molecular concentration, type, or release timing, are simple and easy to implement. They show low system complexity unless the modulation order goes up to very high levels. However, in practice more efficient ways to enhance performance depending on the system conditions are required. This subsection introduces more advanced techniques most of which are operated by combining two or more basic principles.

In a molecular concentration-based modulation system, one of the main issues to be solved is ISI. Since the same type of MM is used over multiple symbol duration, previously transmitted molecules affect the current symbol, which act as interference. Thus, there have been many studies aiming at effectively eliminating the ISI. The following techniques merge the advantages of CSK and MoSK: molecular concentration shift keying (MCSK) (Arjmandi et al. 2013), molecular transition shift keying (MTSK) (Tepekule et al. 2014), and zebra-CSK (Pudasaini et al. 2014).

Arjmandi et al. proposed a hybrid technique called MCSK, where if $S[k] = \text{sym}_1$, the transmitter releases (mm_a) type of MMs if it is an odd-numbered symbol or releases (mm_b) type of MMs if it is an even-numbered symbol Arjmandi et al. (2013). Similar to binary CSK, in case $S[k] = \text{sym}_0$, the transmitter does not release any MMs. The receiver decodes the signal by checking the concentration of (mm_a) or (mm_b) depending on the symbol parity. In the binary case, MCSK outperforms MoSK in terms of BER due to the fact that the ISI component of the molecular signal after a sequence of sym_1's grows too large so that it is highly likely that the next symbol with sym_0 will be mis-decoded as sym_1. The alternating approach of MCSK remedies such mis-decodings. However, the advantage of MCSK diminishes in the quadruple case consequently, quadruple MoSK outperforms quadruple MCSK.

Tepekule et al. proposed a hybrid technique called MTSK to eliminate the destructive ISI while utilizing the constructive ISI (Tepekule et al. 2014). The MTSK is also similar to MCSK, but it determines the molecule type to be transmitted considering the current and the previous symbols. MTSK changes the molecule type at the symbol transitions, which results in reduction of the destructive ISI.

A short comparison of all the modulation techniques mentioned in this entry is given in the following table based on their modulation technique group, receiver type, environment considered, and the performance metric used in the evaluation of the described technique

Modulation in Molecular Signaling, Table 1 Comparison of modulation techniques for molecular communication

Technique name	References	Group	Receiver type	Metrics used
CSK	Kuran et al. (2010, 2011), Mahfuz et al. (2011), and Lin et al. (2012)	Concentration	Abso-FPP	System Resp., Ch. capacity, SER
MoSK	Kuran et al. (2011), Aminian et al. (2015), and Galmés and Atakan (2016)	Type	Abso-FPP	Ch. capacity
Timing	Farsad et al. (2016a) and Murin et al. (2016)	Timing	Abso-FPP	Ch. capacity, BER
PPM	Garralda et al. (2011)	Timing	Passive	System Resp.
TEC	Krishnaswamy et al. (2013)	Timing	Microf'[=]luidic	Ch. capacity
MCSK	Arjmandi et al. (2013)	Hybrid	Abso-FPP	BER
MTSK	Tepekule et al. (2014)	Hybrid	Abso-FPP	BER
Zebra-CSK	Pudasaini et al. (2014)	Hybrid	Abso-FPP	Ch. capacity

(Table 1). As for the receiver types, Abso-FPP refers to an absorbing receiver design which considers the first passage probability (FPP) of the MMs, whereas passive refers to a passive receiver in which upon coming into contact with the receiver, the MMs are not absorbed but are still allowed to move around in the environment.

Key Applications

The potential applications of MC include medical applications, control of chemical reactions, coordination among nanorobots in microscales, underground communications, underwater localization and communications, environmental monitoring, communication applications in pipes or duct systems, and coordination among search and rescue robots for harsh environments in macroscales.

Cross-References

- Molecular Communication for Wireless Body Area Networks
- Nanonetworks
- Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion

References

Abadal S, Akyildiz IF (2011) Automata modeling of quorum sensing for nanocommunication networks. Elsevier Nano Commun Netw 2(1):74–83

Aminian G, Mirmohseni M, Kenari MN, Fekri F (2015) On the capacity of level and type modulations in molecular communication with ligand receptors. In: IEEE international symposium on information theory (ISIT), pp 1951–1955

Arjmandi H, Gohari A, Kenari MN, Bateni F (2013) Diffusion-based nanonetworking: A new modulation technique and performance analysis. IEEE Commun Lett 17(4):645–648

Cobo LC, Akyildiz IF (2010) Bacteria-based communication in nanonetworks. Elsevier Nano Commun Netw 1(4):244–256

Farsad N, Murin Y, Eckford A, Goldsmith A (2016a) On the capacity of diffusion-based molecular timing channels. In: IEEE international symposium on information theory (ISIT), pp 1023–1027

Farsad N, Yilmaz HB, Eckford A, Chae CB, Guo W (2016b) A comprehensive survey of recent advancements in molecular communication. IEEE Commun Surv Tut 18(3):1887–1919

Feynman RP(1959) Plenty of room at the bottom. Eng Sci (Caltech) 23(5):22–36

Galmés S, Atakan B (2016) Performance analysis of diffusion-based molecular communications with memory. IEEE Trans Commun 64(9):3786–3793

Garralda N, Llatser I, Cabellos-Aparicio A, Alarcón E, Pierobon M (2011) Diffusion-based physical channel identification in molecular nanonetworks. Elsevier Nano Commun Netw 2(4):196–204

Genc G, Kara YE, Yilmaz HB, Tugcu T (2016) ISI-Aware modeling and achievable rate analysis of the diffusion channel. IEEE Commun Lett 20(9):1729–1732

Krishnaswamy B, Austin CM, Bardill JP, Russakow D, Holst GL, Hammer BK, Forest CR, Sivakumar R (2013) Time-elapse communication: bacterial commu-

nication on a microfluidic chip. IEEE Trans Commun 61(12):5139–5151
Kuran MS, Yilmaz HB, Tugcu T, Ozerman B (2010) Energy model for communication via diffusion in nanonetworks. Elsevier Nano Commun Netw 1(2): 86–95
Kuran MS, Yilmaz HB, Tugcu T, Akyildiz IF (2011) Modulation techniques for communication via diffusion in nanonetworks. In: IEEE international conference on communication (ICC), Kyoto, pp 1–5
Kuran MS, Tugcu T, Edis B (2012) Calcium signaling: overview and research directions of a molecular communication paradigm. IEEE Wirel Commun 19(5): 20–27
Kuran MS, Yilmaz HB, Tugcu T (2013) A tunnel-based approach for signal shaping in molecular communication. In: IEEE international conference on communication (ICC), Budapest, pp 776–781
Lin WA, Lee YC, Yeh PC, Lee Ch (2012) Signal detection and ISI cancellation for quantity-based amplitude modulation in diffusion-based molecular communications. In: 2012 IEEE global communications conference (GLOBECOM), Anaheim, pp 4362–4367
Mahfuz MU, Makrakis D, Mouftah HT (2010) On the characterization of binary concentration-encoded molecular communication in nanonetworks. Elsevier Nano Commun Netw 1(4):289–300
Mahfuz MU, Makrakis D, Mouftah HT (2011) On the characteristics of concentration-encoded multi-level amplitude modulated unicast molecular communication. In: 2011 24th Canadian conference on electrical and computer engineering (CCECE), Niagara Falls, pp 312–316
Murin Y, Farsad N, Chowdhury M, Goldsmith A (2016) On time-slotted communication over molecular timing channels. In: Proceedings ACM international conference on nanoscale computer and communication, New York, p 9
Nakano T, Suda T, Moore M, Egashira R, Enomoto A, Arima K (2005) Molecular communication for nanomachines using intercellular calcium signaling. In: Proceedings of IEEE international conference on nanotechnology (NANO), pp 478–481
Nakano T, Eckford AW, Haraguchi T (2013) Molecular communication. Cambridge University Press, Cambridge
Pudasaini S, Shin S, Kwak KS (2014) Robust modulation technique for diffusion-based molecular communication in nanonetworks. arXiv preprint arXiv:1401.3938
Tepekule B, Pusane AE, Yilmaz HB, Tugcu T (2014) Energy efficient ISI mitigation for communication via diffusion. In: IEEE international Black Sea conference on communication and networking (BlackSeaCom), Chişinău, pp 33–37
Zimmer RK, Butman CA (2000) Chemical signaling processes in the marine environment. Biol Bull 198(2):168–187

Modulation Techniques for Synthetic Molecular Communications

▸ Modulation in Molecular Signaling

Molecular Abnormality Detection

▸ Molecular Event Detection

Molecular Anomaly Detection

▸ Molecular Event Detection

Molecular Bit Decisions

▸ Molecular Bit Detection

Molecular Bit Detection

Mark Leeson
University of Warwick, Coventry, UK

Synonyms

Molecular bit decisions; Molecular bit determination; Molecular receiver decoding

Definition

Molecular bit detection is the process at a molecular communications (MC) receiver of distinguishing between ones and zeros that have been sent by a transmitter.

Historical Background

Although Harry Nyquist is most often remembered for establishing 2 W pulses per second as the maximum signaling rate that can be supported over a baseband channel of bandwidth W, he also provided the axioms on which digital communications are built (Nyquist 1928). These are the division of time into defined intervals that are now referred to as symbol intervals and the conveying of information by changing a signal property in these intervals. The use of one bit of information per symbol leads to one binary digit (bit) being conveyed by each interval and to a bit rate that is purely the number of intervals per second. The binary digits must be put into a form that is matched to the transmission capabilities of the channel to be used. For most communication systems currently deployed, this entails modulating electromagnetic (EM) signals with the data to be transmitted leading to tractable analysis based on Maxwell's equations. Moreover, many established wired and wireless communication systems experience noise from thermal effects or other users that may be modeled using a Gaussian probability density function (PDF). This noise can cause problems in distinguishing a one bit from a zero bit causing bit errors. MC systems differ significantly from traditional communication systems in that they do not employ EM waves but rely on chemical carriers to convey their information. The natural world makes use of chemical signals for short- and long-range communications (Farsad et al. 2016). The biocompatibility and low energy nature of MC signals make them suitable for applications where EM signals cannot be employed. In general, MC may utilize molecular rails using carrier substances, may employ fluid flow, or may be based on diffusion (Deng et al. 2015). Here, the focus is on MC via Diffusion (MCvD) where molecules propagate randomly, making the process different to the spreading of EM waves. Despite the long evolutionary history of MCvD, it was only recently considered for microscale (Hiyama et al. 2005) and macroscale (Farsad et al. 2013) communication systems.

Foundations

The reliable recovery of data transmitted over a communication channel depends critically on the detection algorithm employed at the receiver. This estimates the signal from the noisy, distorted version that arrives after traversing a communication channel. In the case of bit detection, this will customarily entail the identification of a received bit as a "one" or a "zero." The most fundamental modulation scheme is referred to as on-off keying or OOK, where the presence of a signal in a bit period represents a "one" and signal absence a "zero." In the analysis that follows, this scheme is assumed for simplicity to illustrate the principle of bit detection. More sophisticated schemes have been developed for MCvD (Wang et al. 2017), and the analysis given below may be adapted accordingly. To achieve the goal of a communication system and transfer information from one location to another, a suitable signal must first be generated by the transmitter. This must then propagate to the receiver to be decoded. Thus, the system comprises transmitter, receiver, and channel. Here, the second two of these are particularly relevant since the receiver must make a decision on the received bits after they have suffered potential distortion and noise from the channel, which is the environment in which the signal propagates. In traditional communication systems, the channel is typically a wired or wireless connection carrying appropriate EM waves. In MCvD, particles act as chemical carriers to convey the information. These carriers are small, nanometers to micrometers in size, and travel in an aqueous or gaseous environment where they are free to move and diffuse. Noise may be introduced mainly by the channel through the diffusive nature of the carriers and the presence of other transmitters.

The received signal in MCvD is thus extremely random, and it is common to take the distribution of the received signal for a single molecule as the mean channel impulse response $\overline{h}(t)$. The number of molecules successfully received during the current bit interval or timeslot, N_0, thus has mean given by:

$$\overline{N_0} = N_{\text{tx}}\overline{h}(t) \tag{1}$$

when the transmitter releases N_{tx} molecules to represent a one. There are several receiver models that have been proposed for different MCvD scenarios and each of these produces a different $\overline{h}(t)$. The principal categories are passive (Mahfuz et al. 2014), fully absorbing (Yilmaz et al. 2014), and reactive (Ahmadzadeh et al. 2016) depending of the behavior of the molecules and the receiver. The simplest form of $\overline{h}(t)$ results from a full absorbing receiver so that will be used as the basis for illustration here. The others produce results with similar behaviors but different optimum values and performance details.

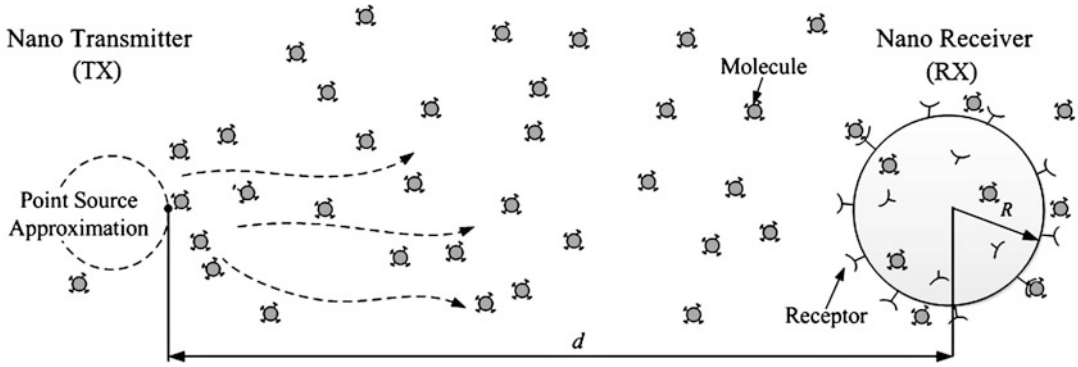
Although diffusion is a well-established topic in Physics, the solution of its defining equation in realistic scenarios for communicating information is very challenging. Therefore, it is common to make the approximation illustrated in Fig. 1 to the scenario where a point source transmitter of molecules (TX) releases N_{tx} of these at time zero and they travel to the receiver (RX) via 3D Brownian motion. A fraction of the molecules arrives at RX within the timeslot.

As introduced above, the simplest case (assumed here) is that RX is fully absorbing, and molecules are captured as soon as they collide with the surface of RX. The major impact of the different receiver types is to alter the probability $P_{\text{ca}}(d, t)$ that RX captures a molecule at a time t and distance d as defined in Fig. 1.

For full absorption and a counting receiver, the capture probability may be written as (Yilmaz et al. 2014):

$$P_{\text{ca}}(d, t) = \frac{R}{d}\,\text{erfc}\left(\frac{d - R}{\sqrt{4Dt}}\right) \tag{2}$$

where R is the receiver radius and D is the diffusion coefficient in the medium in which communication is taking place. The fundamental representation of information is via a binary form with the transmitted information represented by a sequence of binary symbols. Each consecutive time slot contains either N_{tx} or no molecules. Each timeslot is of duration t_{s} and molecules arriving at RX are captured and removed from the environment. RX can either measure the number of molecules arriving at a certain time instant (amplitude detection) or the accumulated total of molecules arriving during a certain time period (energy detection) (Llatser et al. 2013). The latter is a better choice, particularly for long distance transmission, since it requires far fewer molecules to be transmitted for the same bit error rate (BER) (Aijaz and Aghvami 2015). Thus, it is assumed that RX determines whether a one or a zero was sent by counting the number of absorbed molecules. There have been developments in recent years to add further sophistication to the detection process within the constraints of the capabilities of nano-machines. These have included the detection of sequences of bits using both the maximum a posteriori (MAP) and maximum likelihood (ML) criteria (Kilinc and Akan



Molecular Bit Detection, Fig. 1 Molecular communication channel

2013; Meng et al. 2014). ML produces optimum receiver performance and offers viable suboptimum methods for nano-machines in a range of scenarios (Noel et al. 2014; Fang et al. 2018; Kuscu and Akan 2018).

To avoid detailed discussion of the ML method, for which readers are referred to the references above, single bit threshold detection is assumed. Thus, if the number of molecules received in an intended time slot exceeds a predesigned threshold, the symbol is interpreted as "one," otherwise, it is interpreted as "zero" (Kuran et al. 2011). It is possible to develop an adaptive threshold (Damrath and Hoeher 2016), but again for simplicity a fixed value is assumed here. The approximation is made that molecules are released as an impulse at the beginning of the timeslot and diffuse toward RX to be either registered or not registered by RX within the timeslot. Thus, the distribution is derived from a series of trials with the outcome received or not received producing a binomial distribution for N_0:

$$N_0 \sim \text{Binom}\left(N_{\text{tx}}, P_{\text{ca},0}\right) \tag{3}$$

where $P_{\text{ca},\,0} = P_{\text{ca}}(d, t_{\text{s}})$. For large numbers of molecules, this may be conveniently approximated by a normal distribution

$$N_{0\_\text{Norm}} \sim \mathcal{N}\left(N_{\text{tx}} P_{\text{ca},0}, N_{\text{tx}} P_{\text{ca},0}\left\{1 - P_{\text{ca},0}\right\}\right) \tag{4}$$

but this can be inaccurate when the number of molecules is small (Lu et al. 2015, 2016), leading to an asymmetric binomial distribution. Thus, for N_{tx} values in the hundreds of molecules, the Poisson approximation may be utilized:

$$N_{0\_\text{Pois}} \sim \text{Pois}\left(N_{\text{tx}} P_{\text{ca},0}\right). \tag{5}$$

However, the molecules released at TX are not guaranteed to reach RX in one timeslot, so the remaining molecules may arrive later causing intersymbol interference (ISI). The number of molecules received from the previous i^{th} symbol in the current time slot is represented by N_i, $i \in \{1, \ldots I\}$, where I is the ISI length (Lu et al. 2015).

$$N_{\text{i}} \sim \text{Binom}\left(N_{\text{tx}}, P_{\text{ca,i}} - P_{\text{ca,i}-1}\right) \tag{6}$$

$$P_{\text{ca,i}} = P_{\text{ca}}\left(d, \{i+1\}\, t_{\text{s}}\right). \tag{7}$$

With corresponding approximations

$$N_{\text{i}\_norm} \sim \mathcal{N}\left(N_{\text{tx}} \Delta P_{\text{i}}, N_{\text{tx}} \Delta P_{\text{i}}\left\{1 - \Delta P_{\text{i}}\right\}\right) \tag{8}$$

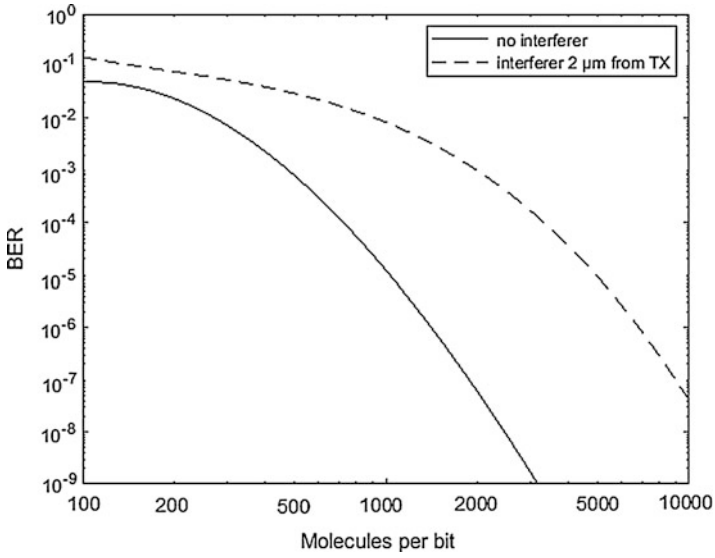
and

$$N_{\text{i\_Pois}} \sim \text{Pois}\left(N_{\text{tx}} \Delta P_{\text{i}}\right) \tag{9}$$

where $\Delta P_{\text{i}} = P_{\text{ca, i}} - P_{\text{ca, i}\,-\,1}$.

From these expressions, as shown in (Lu et al. 2017), it is possible to write down the effect of the number of previous slots that comprise the ISI length. Based on this analysis, a prediction of the MCvD system BER may be made by determining the probabilities that the noise sources will cause the signal at RX to be the wrong side of the decision level. Thus, in Fig. 2 the solid line indicates the BER obtained when the distance d is 7 μm and the diffusion coefficient D is 79.4 $\mu\text{m}^2\text{s}^{-1}$(Lu et al. 2016).

In common with many other wireless systems, MCvD is also susceptible to one further source of error, namely interference from other users. This may be quantified by including an interfering receiver in the system that may absorb molecules destined for RX considered above (Lu et al. 2016). To illustrate a particularly problematic case, the dashed line in Fig. 2 illustrates the BER obtained over the original path when there is an interferer on the same side as the RX but at a distance of just 2 μm from TX (Lu et al. 2016). Although this is a particularly bad case that produces a power penalty of almost 7 dB at a BER level of 10^{-6}, it does graphically make the point that care must be taken when transmission is over more than just a point to point link. The use of different molecules for transmission to different receivers is possible (Kuran et al. 2011) but adds considerable complexity to MCvD systems.



Molecular Bit Detection, Fig. 2 BER results for a MC system with 7 µm between TX and RX; solid line with no interferer; dotted line with an interferer 2 µm from TX

Improving Decisions

The reliability of information transmission can be increased by the employment of error correcting codes (ECCs) as is the case to improve the BER performance of conventional communication systems (Costello and Forney 2007). The small size and limited capabilities of nano-machines means that codes must be chosen for energy efficiency (Bai et al. 2014) and simplicity of implementation (Leeson and Higgins 2012). In recent years, considerable research efforts have been made in the development of nanoscale logic gates (Pilarczyk et al. 2018), making ECC implementation conceivable in the near future. ECC performance has thus been investigated numerically for MCvD in anticipation of future developments (Lu et al. 2015).

Key Applications

In areas where EM communication is not suitable, MC offers a solution for the near future. At the macroscale, this includes tunnel or pipe networks where experimental results have shown that EM waves fail to propagate but molecules get through but with a long delay (Qiu et al. 2014). This would facilitate, for example, the recovery of data from embedded sensors or communications between search-and-rescue robots. Applications at the micro- or even nanoscale have driven the applications agenda to date with the promise of nanorobot drug delivery and tissue engineering (Nakano et al. 2013). In this regime, EM communication is difficult given the necessary ratio of the antenna size to wavelength ratio and the need for a line of sight using optical communications (Akyildiz et al. 2008).

Cross-References

- Brownian Motion
- Modeling Approaches for Simulating Molecular Communications
- Modulation in Molecular Signaling
- Molecular Event Detection
- Receiver Mechanisms for Synthetic Molecular Communication Systems with Diffusion

References

Ahmadzadeh A, Arjmandi H, Burkovski A, Schober R (2016) Comprehensive reactive receiver modeling for diffusive molecular communication systems: reversible binding, molecule degradation, and finite number of receptors. IEEE Trans Nanobioscience 15(7):713–727

Aijaz A, Aghvami A-H (2015) Error performance of diffusion-based molecular communication using pulse-based modulation. IEEE Trans Nanobioscience 14(1):146–151

Akyildiz IF, Brunetti F, Blazquez C (2008) Nanonetworks: a new communication paradigm. Comput Netw 52(12):2260–2279

Bai C, Leeson MS, Higgins MD (2014) Minimum energy channel codes for molecular communications. Electron Lett 50(23):1669–1671

Costello DJ, Forney GD (2007) Channel coding: the road to channel capacity. Proc IEEE 95(6):1150–1177

Damrath M, Hoeher PA (2016) Low-complexity adaptive threshold detection for molecular communication. IEEE Trans Nanobioscience 15(3):200–208

Deng Y, Noel A, Elkashlan M, Nallanathan A, Cheung KC (2015) Modeling and simulation of molecular communication systems with a reversible adsorption receiver. IEEE Trans Mol Biol Multi-Scale Commun 1(4):347–362

Fang Y, Noel A, Yang N, Eckford AW, Kennedy RA (2018) Maximum likelihood detection for cooperative molecular communication. In: Proceedings of the IEEE international conference on communications (ICC), Piscataway, NJ pp 1–7

Farsad N, Guo W, Eckford AW (2013) Tabletop molecular communication: text messages through chemical signals. PLoS One 8(12):e82935

Farsad N, Yilmaz HB, Eckford A, Chae C-B, Guo W (2016) A comprehensive survey of recent advancements in molecular communication. IEEE Commun Surv Tutorials 18(3):1887–1919

Hiyama S, Moritani Y, Suda T, Egashira R, Enomoto A, Moore M, Nakano T (2005) Molecular communication. In: Proceedings of the NSTI nanotechnology conference and trade show, Austin, TX, 3, pp 391–394

Kilinc D, Akan OB (2013) Receiver design for molecular communication. IEEE J Sel Areas Commun 31(12):705–714

Kuran MS, Yilmaz HB, Tugcu, T, Akyildiz AF (2011) Modulation techniques for communication via diffusion in nanonetworks. In: Proceedings of the IEEE international conference on communications (ICC), Piscataway, NJ pp 1–5

Kuscu M, Akan OB (2018) Maximum likelihood detection with ligand receptors for diffusion-based molecular communications in internet of bio-nano things. IEEE Trans Nanobioscience 17(1):44–54

Leeson MS, Higgins MD (2012) Forward error correction for molecular communications. Nano Commun Netw 3(3):161–167

Llatser I, Cabellos-Aparicio A, Pierobon M, Alarcon E (2013) Detection techniques for diffusion-based molecular communication. IEEE J Sel Areas Commun 31(12):726–734

Lu Y, Higgins MD, Leeson MS (2015) Comparison of channel coding schemes for molecular communications systems. IEEE Trans Commun 63(11):3991–4001

Lu Y, Higgins MD, Noel A, Leeson MS, Chen Y (2016) The effect of two receivers on molecular communications systems. IEEE Trans Nanobioscience 15(8):891–900

Lu Y, Higgins MD, Leeson MS, Chen Y, Jennings P (2017) A revised look at the effects of the channel model in molecular communication systems. IET Micro Nano Lett 12(2):136–139

Mahfuz M, Makrakis D, Mouftah H (2014) A comprehensive study of sampling-based optimum signal detection in concentration-encoded molecular communication. IEEE Trans Nanobioscience 13(3):208–222

Meng LS, Yeh P-C, Chen K-C, Akyildiz IF (2014) On receiver design for diffusion-based molecular communication. IEEE Trans Signal Process 62(22):6032–6044

Nakano T, Eckford AW, Haraguchi T (2013) Molecular communications. Cambridge University Press. New York, NY

Noel A, Cheung KC, Schober R (2014) Optimal receiver design for diffusive molecular communication with flow and additive noise. IEEE Trans Nanobioscience 13(3):350–362

Nyquist H (1928) Certain topics in telegraph transmission theory. Trans AIEE 47(2):617–644

Pilarczyk K, Wlaźlaka E, Przyczyna D, Blachecki A, Podborska A, Anathasiou V, Konkoli Z, Szaciłowski K (2018) Molecules, semiconductors, light and information: towards future sensing and computing paradigms. Coord Chem Rev 365:23–40

Qiu S, Guo W, Wang S, Farsad N, Eckford A (2014) A molecular communication link for monitoring in confined environments. In: Proceedings of the IEEE international conference on communications (ICC), Piscataway, NJ pp 718–723

Wang J, Yin B, Peng M (2017) Diffusion based molecular communication: principle, key technologies, and challenges. China Commun 14(2):1–18

Yilmaz HB, Heren AC, Tugcu T, Chae C-B (2014) Three-dimensional channel characteristics for molecular communications with an absorbing receiver. IEEE Commun Lett 18(6):929–932

Molecular Bit Determination

▸ Molecular Bit Detection

Molecular Communication for Wireless Body Area Networks

M. D. Nashid Anjum and Honggang Wang
Department of Electrical Engineering, University of Massachusetts Dartmouth, North Dartmouth, MA, USA

Acronyms

BAN	Body Area Network
BANN	Body Area Nanonetwork
BAN^2	Body Area Nanonetwork
MC	Molecular Communication
IoNT	Internet of Nano Things
WBAN	Wireless Body Area Network

Definition

A wireless body area network (i.e., WBAN) generally consists of a central hub (i.e., base station) and a number of lightweight, low-powered, miniature, wearable sensors that operate in the proximity of a human body. Typically, it is located on the body, garment, underneath the skin, or deep into the tissue (Cavallari et al. 2014; Anjum and Wang 2016). WBAN is also known as a wireless body sensor network (WSN) or simply a body area network (BAN). On the contrary, a molecular communication (i.e., MC) system is an in-body BAN which consists of bio-nanomechanics such as biosensors and bio-actuators where the data transmission between the transmitter and receiver nanomachines is carried on by the encoded molecules (Cavallari et al. 2014; Nakano et al. 2012; Akyildiz et al. 2008). MC for WBAN is also termed as body area nanonetworks, i.e., BANN or BAN^2 (Suzuki et al. 2014; Atakan et al. 2012). The integration of MC-based in-body BANN and wearable WBAN is known as the Internet of Nano Things (IoNT) Dressler et al. (2015).

Historical Background

The use of the term "molecular communication" can be traced back to the 1970s where a number of researchers individually studied the various modes of molecular transmission among neuron cells (Dismukes and Key 1979). Later on, MC in host-parasite interaction and MC between plant-pathogen interaction are studied in Dixon and Lamb (1990) and Hadwiger et al. (1981), respectively. However, in the field of communication engineering, the concept of MC was developed in the early 2000s to create a communication network for bio-nanomachines (Nakano et al. 2005, 2012).

System Architecture

The system architecture of a MC typically consists of four key components – the sender, receiver, information bearing molecules (also known as signaling molecule), and propagation medium. To support the movement of the information molecule (i.e., IM), a MC may include an interface, transport, guide, and addressing molecules. All of these components are various forms of bio-nanomachines which are biochemically reactive, micrometer to nanometer range devices. The transfer of information in a typical MC takes the following five steps: encoding, transmitting, propagation, receiving, and decoding.

- *Encoding and Transmitting:* The encoding and transmitting of IM are done by the transmitter bio-nanomachine which is also known as the sender. The sender encodes the information in various forms within the molecules, such as the type of molecule used, their 3D structure (e.g., protein structure), sequential structure (e.g., DNA sequence), concentration (e.g., calcium concentration), number of molecule, release time of molecules, etc. The release of IM into the propagation medium is done

by budding vesicles (if the sender is a cell), opening the molecular gate (e.g., ion channel) of the sender membrane, or catalyzing a chemical reaction (Nakano et al. 2013).

- *Propagation:* The scales of propagation of the CM are divided in three categories – intracellular (<= 100 μm), intercellular (<= 100 mm), and inter-organ (up to few meters). The mode of propagation can be two types – passive and active. In the passive mode, a large number of molecules diffuse in all available directions. This is a slower process and does not require any infrastructure hence suitable for intracellular and intercellular communication. In addition, the propagation time t over a distance L is $t \approx \frac{L^2}{D}$, where D is the diffusion coefficient which depends on the molecular size and structure, viscosity of the medium, and the temperature. If $D = 100\,\mu m^2 s$, a molecule takes about 2.78 h to propagate over a distance of 1 mm. Unlike a passive mode, the active mode of propagation directs the movements of molecules toward the targeted location. It is comparatively a faster process and takes fewer molecules thus suitable for inter-organ propagation scale. However, unlike passive modes, it requires propagation infrastructure and a constant supply of energy. Different forms of active transportation provide different propagation speeds such as motor proteins (e.g., kinesin, 2–4 μm/s), bacterial chemotaxis (e.g., E. coli, few μm/s), and diffusion-reaction mechanism (e.g., calcium signaling, 20 μm/s), etc. (Nakano et al. 2013). However, to support the information molecules, the propagation medium may contain some other types of molecules, such as transport molecules, to move information molecules, guide molecules to direct the movement of the transport molecules, interface molecules to selectively transport the information molecules, and address molecules to specify the receiver (Nakano et al. 2013).
- *Receiving and Decoding:* Receiving and decoding processes are done by the receiver nanomachine. Reception of information molecules can be done through permeable plasma membrane, surface receptors, or chemically gated surface channels. The decoding process is essentially the reaction with the information molecule upon the reception at the receiving end. Furthermore, the consequences of the chemical reaction during decoding may generate movement, morphological changes, change in chemical functionality, or new molecules. The receiver may initiate multihop communication by releasing the produced molecules into the propagation medium (Nakano et al. 2013).

Communication Engineering Aspects

Although MC exists in nature for billions of years, the advancement in MC engineering has started primarily from the early 2000s. However, compared to traditional wireless communication, the advancement of MC is still in infancy (Farsad et al. 2016). This section sheds light on some of the engineering aspects of MC.

Physical Layer Aspects

The physical layer deals with the biophysical mechanism of data transmission, propagation, and reception. In essence, the data is transmitted in the form of IMs, also known as signal molecules. Moreover, the physical layer also provides the biophysical basis of hardware and interface (Farsad et al. 2016; Nakano et al. 2013).

- *Hardware:* The key hardware for MC are the sender, propagation channel, receiver, IM, and supporting molecules, as mentioned in Section "Applications." Essentially, these are all various kinds of bio-nanomachines.
- *Data/Message:* In a traditional WBAN, the physical layer data is transmitted into the form of 2.4 or 60 GHz mmWave (Anjum et al. 2017). Unlike traditional WBAN, the physical layer data or messages of MC are transmitted in the form of information molecules which

can be various kinds of protein molecules, vesicles, or bacteria.

- *Modulation Techniques:* Modulation in MC is the process of encoding data/messages in the IMs. The modulation or encoding of IMs can be done in various ways as we discussed in section "Applications."
- *Propagation Channel Models:* Propagation channel models for intercellular and intracellular MC are known as microscale models. On the other hand, propagation channel models for inter-organisms are known as macroscale models. Major passive propagation models for microscale MC are free diffusion propagation and diffusion with first fitting. Major active propagation models are flow-assisted propagation, bacteria-assisted propagation, motor protein moving over microtubule tracks, microtubule filament motility over stationary kinesin, neurochemical propagation, and propagation through gap junction. The major macroscale propagation models are diffusion and flow-based propagation. Some other macroscale propagation models are advection, convection, mechanical dispersion, and turbulent flows. Macroscale models require comparatively larger amount of molecules and external source of power; hence, these models are active in nature (Farsad et al. 2016).
- *Noise Sources:* The sources of noise are primarily transmitter emission noise, random propagation, i.e., diffusion noise, reception noise, environmental noise such as degradation and reaction, multiple transmitters, etc.
- *Channel Capacity:* Practically, MC channels are not memoryless; thus, the channel capacity is different for different models of propagation. If x^n is a sequence of n successive transmission symbols, and y^n is the corresponding received symbols, then the channel capacity can be presented by Farsad et al. (2016):

$$\mathscr{C} = \lim_{n\to\infty} \inf \sup_{x^n} I(X^n; Y^n) \qquad (1)$$

Link Layer Aspects

- *Error Handling:* To detect the transmitted message, the number of received molecules must be larger than a threshold. Hence, transmitting larger numbers of molecules increases the signal to interference plus noise ratio. Another approach of avoiding error is optimizing the transmission rate. Error detection and correction can be done by adding redundant information to the information molecule. For instance, adding an additional DNA sequence to a DNA molecule allows error detection and forward error correction at the receiving end (Nakano et al. 2012).
- *Addressing:* Addressing is the process of specifying the target receiver. Usually, addressing of the receiver is done by the type of molecule transmitted by the sender.
- *Synchronization:* Synchronization of the sender and receiver in MC is a complicated process because of the immense propagation delay and jitter. No significant research regarding synchronization is reported so far.
- *Media Access Control:* The interference among multiple pairs of senders and receivers can be avoided by transmitting different types of molecules for each sender-receiver pair. This technique is not sufficient for a large number of sender-receiver pairs. In that case, some other techniques could be employed such as channel reservation, switching or time-division multiplexing, etc. However, such approaches are not implemented yet.
- *Flow Control:* If the chemical reaction on the receiver side is significantly slower compared to the sender side, then the flow control is required to avoid buffer overflow on the receiver side. No such work has been reported yet in the literature.

Network Layer Aspects

The routing of the existing MC is limited to the static routing tables which fails to address dynamic locations of the network nodes. In case of a static routing, a sender transmits the data using a bacterium with addressing

molecules which indicates the target receiver. The router node receives the bacterium and employs statically defined chemical processes to retransmit the data using a bacterium which follows the chemical gradients to the following target router. Advancements regarding other network layer issues such as congestion control, topology management, network scalability, etc. are still limited (Nakano et al. 2013).

Mathematical Models

The simplest mathematical model to represent the movement of the information molecule is a random walk model which does not consider the directional drift or chemical reaction. Hence, a random walk model is suitable for representing passive propagation models. A more complex mathematical model is a random walk with drift which is suitable for representing the active propagation models where the movement of the IM is drifted directionally. Another kind of mathematical model for MC is the random walk with chemical reactions. This model considers the chemical reactions of IMs induced by the amplifiers. Amplifiers are located in the environment and react with molecules that can increase the reliability of the molecular propagation by multiplying the number of propagating molecules.

Applications

Although the primary application of molecular communication evolves around biomedical fields, its application in environmental and manufacturing fields is also reported (Nakano et al. 2012).

- *Biomedical Applications:* In the biomedical field, the applications of MC are health monitoring within an organism, disease diagnosis known as lab-on-a-chip, drug delivery within an organism, regenerative medicine, etc.
- *Environmental Applications:* MC can be used for monitoring and controlling environmental pollution.
- *Manufacturing Applications:* MC can be exploited to control the transport of bio-nanomachines or molecules and can also be modified to develop novel patterns and structures of molecules (Nakano et al. 2013).

Conclusions

Although the concept of BAN^2 has been introduced in the early 2000s, the advancement of this research field is still at its infancy level, after two decades. However, BAN^2 has an immense potential and prospect in the biomedical and environmental applications. As a result, IEEE P1906.1 Standards Working Group for Nanonetworking has been formed in 2011 for standardizing the MC protocols.

References

Akyildiz IF, Brunetti F, Blázquez C (2008) Nanonetworks: a new communication paradigm. Comput Netw 52(12):2260–2279

Anjum MN, Wang H (2016) Optimal resource allocation for deeply overlapped self-coexisting WBANs. In: 2016 IEEE global communications conference (GLOBECOM). IEEE

Anjum MDN, Fang H (2017) Coexistence in millimeter-wave WBAN: a game theoretic approach. In: 2017 international conference on computing, networking and communications (ICNC). IEEE

Atakan B, Akan OB, Balasubramaniam S (2012) Body area nanonetworks with molecular communications in nanomedicine. IEEE Commun Mag 50(1):28–34

Cavallari R et al (2014) A survey on wireless body area networks: technologies and design challenges. IEEE Commun Surv Tutorials 16(3):1635–1657

Dismukes RK (1979) New concepts of molecular communication among neurons. Behav Brain Sci 2(3): 409–416

Dixon RA, Lamb CJ (1990) Molecular communication in interactions between plants and microbial pathogens. Annu Rev Plant Biol 41(1):339–367

Dressler F, Fischer S (2015) Connecting in-body nano communication with body area networks: challenges and opportunities of the internet of nano things. Nano Commun Netw 6(2):29–38

Farsad N et al (2016) A comprehensive survey of recent advancements in molecular communication. IEEE Commun Surv Tutorials 18(3):1887–1919

Hadwiger LA, Loschke DC (1981) Molecular communication in host-parasite interactions: hexosamine polymers(Chitosan) as regulator compounds in race-specific and other interactions. Phytopathology 71(7):756–762

Nakano T et al (2005) Molecular communication for nanomachines using intercellular calcium signaling. In: 5th IEEE conference on nanotechnology. IEEE
Nakano T et al (2012) Molecular communication and networking: opportunities and challenges. IEEE Trans Nanobioscience 11(2):135–148
Nakano T, Eckford AW, Haraguchi T (2013) Molecular communication. Cambridge University Press, Cambridge
Suzuki J et al (2014) A service-oriented architecture for body area nanonetworks with neuron-based molecular communication. Mob Netw Appl 19(6):707–717

Molecular Communication Simulators

▸ Modeling Approaches for Simulating Molecular Communications

Molecular Event Detection

Reza Mosayebi
Sharif University of Technology, Tehran, Iran

Synonyms

Molecular abnormality detection; Molecular anomaly detection; Molecular target detection

Definitions

Molecular event detection is one of the key challenges in many microscale revolutionary applications, such as environmental and health monitoring and disease diagnosis, which deals with detecting undesired signals at the nano- and microscales.

Historical Background

One of the key challenges in disease diagnosis and health monitoring applications is the problem of event detection, e.g., early tumor detection, which has received significant attention in medicine and other related fields (Chen et al. 2016; Wuab and Qu 2015). Upon detecting tumor (event), a drug can be released at an appropriate rate to mitigate the effect of tumor (Chude-Okonkwo et al. 2017). This motivates the investigation of event detection at microscale using molecular communications.

Event detection has been extensively studied in different fields (Chandola et al. 2009). In this context, event is also referred to as abnormality, anomaly, and outlier which has several applications including failure detection in computer science, segmentation of signals in biomedical applications, and fraud detection for credit cards. However, in molecular communication systems, due to diffusion of molecules, the number of molecules received at a receiver (received signal) has Poisson distribution which imposes inherent randomness to the molecular channels. Due to this specific characteristic of molecular communication, classical methods cannot be directly applied for detecting molecular events. Therefore, several approaches have been proposed for molecular event detection (Felicetti et al. 2014; Okaie et al. 2016; Ghavami and Lahouti 2017; Nakano et al. 2017; Mai et al. 2017; Mosayebi et al. 2017). In all of the proposed works for molecular event detection, multiple nanosensors (NSs) for initial detection of event are employed. Upon detection, the NSs send their decisions via releasing an amount of molecules to a fusion center (FC) where the final decision regarding the presence of event is made. In particular, in Felicetti et al. (2014), it is assumed that mobile NSs move through the vasculature and gather at the target location by binding to a tumor. Next, in Okaie et al. (2016), a similar idea is used where two types of NSs are used, named leader and follower NSs. Here, leader NSs create an attractant gradient around the target. Then, follower NSs move according to the attractant gradient, approach the target, and perform necessary actions such as releasing drug. Employing static NSs is proposed for anomaly detection in body tissue in Ghavami and Lahouti (2017), Mai et al. (2017), and Mosayebi et al. (2017). In particular, in Ghavami and Lahouti (2017) a macroscale noise channel is considered between

the NSs and the FC, while in Mai et al. (2017) and Mosayebi et al. (2017), a microscale Poisson signal-dependent noise channel is considered. Finally, an abstract modeling and simulation of both static and mobile NS network for target detection is proposed in Nakano et al. (2017).

System Model

For molecular event detection, one can consider a network of M NSs that observe a part of a tissue or move inside the vasculature and sense their environment (sensing scheme), along with an FC. Upon detecting an event (target or abnormality), NSs will release an amount of molecules into the medium, where some of them may be collected by the FC (reporting scheme) and where the final decision regarding the presence of an event will be made. The absence and presence of event are denoted by hypotheses $\mathscr{H}_0$ and $\mathscr{H}_1$, respectively. In the following, the sensing and reporting schemes will be briefly described. Before going into the details of sensing and reporting schemes, it is noteworthy to mention that the approaches that will be described in the following are examples of what has been done in the molecular communication literature and are non-exhaustive for what might be practical.

Sensing Model

Each NS can measure one or more sensing variables, referred to as inputs. For instance, for tumor detection, examples of inputs include the concentration of specific types of molecules, referred to as *biomarkers*, such as nucleic acids and proteins, and a lack of oxygen (Wuab and Qu 2015; Chude-Okonkwo et al. 2017). Then, based on the measured inputs, the NS determines a numerical value which reflects the presence or absence of the event. For the sensing scheme, there are two different models in the literature, referred to as *hard* and *soft* decision schemes. For the hard decision scheme (Nakano et al. 2017; Ghavami and Lahouti 2017; Mai et al. 2017), the decision made by each NS $k \in \{1, \ldots, M\}$ is a Bernoulli random variable (RV) C_k with the realization values of $c_k = 0$ and $c_k = 1$ upon detecting hypotheses $\mathscr{H}_0$ and $\mathscr{H}_1$, respectively. However, for the soft decision scheme, one of the $L > 2$ values in $\mathscr{X} = \{0, 1/(L-1), 2/(L-1), \ldots, 1\}$ can be assumed as a soft decision for the k-th NS, where small and large values of c_k indicate that NS k leans toward hypotheses $\mathscr{H}_0$ and $\mathscr{H}_1$, respectively (Mosayebi et al. 2017). For RV C_k under hypotheses $\mathscr{H}_0$ and $\mathscr{H}_1$, general probability mass functions (PMFs) $\mathrm{g}_0(\cdot)$ and $\mathrm{g}_1(\cdot)$, respectively, can be assumed, where under hard decision, i.e., $L = 2$, they have only two nonzero values for $c_k = 0$ and $c_k = 1$. It is noteworthy to indicate that the functionality of $\mathrm{g}_i(\cdot), i = 0, 1$ with respect to its argument is a function of several factors such as the NS structure and its sensing accuracy.

Reporting Model

For each sensor k, if $c_k \neq 0$, an instantaneously number (Ghavami and Lahouti 2017; Mai et al. 2017; Mosayebi et al. 2017) or a constant rate of molecules which can be degraded in the medium (Okaie et al. 2016; Nakano et al. 2017) will be released toward the FC. For each of the released schemes, a mean value of $\mathrm{g}_i(c_k)\theta_k$ for the number of molecules released by the k-th NS and collected at the FC can be considered. Here, θ_k is a function of the number/rate of the released molecules, distance between the NS and the FC, the diffusion coefficient of the released molecules, and the shape and size of the FC. In line with the references for molecular event detection and for tractability of analysis, in the following it is assumed that $\theta_k = \theta, \forall k \in \{1, \ldots, M\}$. Here, two cases can be imagined:

1. (Case I) Each NS sends a unique type of molecules, different from the molecules released by other NSs.
2. (Case II) The same type of molecule is used by all NSs.

In the following, for each case, the PMF of the signal received at the FC is modeled. Derived PMFs for the signal received at the FC will be further used to design optimal decision rule for the FC.

Case I

For Case I, since the types of molecules released by different NSs are different, the FC can distinguish between the messages sent by different NSs. Furthermore, an environmental noise for each type of molecules can be considered which is an independent source of molecules. Let Y_k denote the RV which models the number of released molecules by the k-th NS and the corresponding environmental noise present at the FC. Therefore, conditioned on hypothesis $\mathscr{H}_i, i = 0, 1$, the PMF of Y_k can be modeled as follows (Mosayebi et al. 2017, Eq. (8)):

$$\mathsf{P}(Y_k = y_k|\mathscr{H}_i) = \sum_{c_k \in \mathscr{X}} \mathrm{g}_i(c_k) \times \frac{\exp\left(-\mathrm{g}_i(c_k)\theta - \lambda_k\right)\left(\mathrm{g}_i(d_k)\theta + \lambda_k\right)_k^y}{y_k!}, \quad i = 0, 1, \tag{1}$$

(2)

where $\mathsf{P}(\cdot)$ denotes probability measure, y_k is a realization of Y_k, and λ_i is the mean number of the environmental noise molecules with the same type as the molecules released by NS k. Since the types of molecules are different, the RVs Y_k become independent. Now, by defining $\mathbf{Y} = [Y_1, \ldots, Y_M]^T$ as the vector containing all RVs $Y_k, k \in \{1, \ldots, M\}$, the PMF of the signal received at the FC can be characterized as follows:

$$\mathsf{P}(\mathbf{Y} = \mathbf{y}|\mathscr{H}_i) = \prod_{k=1}^{M} \mathsf{P}(Y_k = y_k|\mathscr{H}_i) = \prod_{k=1}^{M} \sum_{c_k \in \mathscr{X}} \mathrm{g}_i(c_k) \times \frac{\exp\left(-\mathrm{g}_i(c_k)\theta - \lambda_k\right)\left(\mathrm{g}_i(c_k)\theta + \lambda_k\right)^{y_k}}{y_k!}, \quad i = 0, 1, \tag{3}$$

where $\mathbf{y} = [y_1, \ldots, y_M]^T$ is a realization of $\mathbf{Y}$.

Case II

For Case II, since the types of molecules released by different NSs are the same, the FC cannot distinguish between the messages of different NSs. Therefore, the PMF of the number of molecules received at the FC is a function of RV $C = \sum_{k=1}^{M} C_k \in \mathscr{X} = \{0, 1/L - 1, 2/L - 1, M\}$. Since RVs C_k are statistically independent, the PMF of RV C can be obtained using convolution operator. This PMF is denoted by $\mathrm{G}_0(\cdot)$ and $\mathrm{G}_1(\cdot)$ for hypothesese $\mathscr{H}_0$ and $\mathscr{H}_1$, respectively. Furthermore, the same type of molecule can always be considered in the environment which plays the role of environmental noise. Therefore, by denoting the random variable Y as the total number of released molecules by different NSs and environmental noise that are collected at the FC, one can see that conditioned on hypothesis $\mathscr{H}_i, i = 0, 1$, Y can be modeled as follows (Mosayebi et al. 2017, Eq. (17)):

$$\mathsf{P}(Y = y|\mathscr{H}_i) = \sum_{c \in \mathscr{X}} \mathrm{G}_i(c) \times \frac{\exp\left(-\mathrm{G}_i(c)\theta - \lambda\right)\left(\mathrm{G}_i(c)\theta + \lambda\right)^y}{y!}, \quad i = 0, 1, \tag{4}$$

(5)

where y is the realization of Y, c is the realization of C, and λ is mean number of environmental noise molecules.

Detector Design for the FC

For the FC, a decision $\mathrm{d} \in \{0, 1\}$ is made, where $\mathrm{d} = 0$ and $\mathrm{d} = 1$ means that the FC chooses hypothesis $\mathscr{H}_0$ and $\mathscr{H}_1$, respectively. Due to the inherent randomness of the received signal and the presence of noise, the selected hypothesis by the FC may be incorrect. Therefore, two types of errors can be imagined: first, if the FC chooses $\mathrm{d} = 1$, while hypothesis $\mathscr{H}_0$ is correct. For this case, the conditional probability of error is referred to as *probability of false alarm* and is denoted by $\mathrm{P_{fa}} = \mathsf{P}(\mathrm{d} = 1|\mathscr{H}_0)$. Similarly, another type of error exists where the FC chooses $\mathrm{d} = 0$, while hypothesis $\mathscr{H}_1$ is correct. For this case, the conditional probability of error is referred to as *probability of missed detection* and is denoted by $\mathrm{P_m} = \mathsf{P}(\mathrm{d} = 0|\mathscr{H}_1)$.

For the FC, the goal is to design the optimal decision rule that minimizes the probability of missed detection subject to a pre-assigned upper bound ω on the probability of false alarm. That is,

$$\min_{\text{detectors}} \mathrm{P_m}, \text{ subject to } \mathrm{P_{fa}} \leq \omega. \tag{6}$$

The optimal solution of (6) is the Neyman-Pearson detector (Kay 1998), which compares the log-likelihood ratio (LLR) with the maximum threshold τ that ensures $\mathrm{P_{fa}} \leq \omega$. The LLR for each mentioned cases above can be written as

$$\mathsf{LLR} = \begin{cases} \sum_{k=1}^{M} \log\left(\frac{\mathsf{P}\left(Y_k=y_k\middle|\mathscr{H}_1\right)}{\mathsf{P}\left(Y_k=y_k\middle|\mathscr{H}_0\right)}\right); & \text{Case I,} \\ \log\left(\frac{\mathsf{P}\left(Y=y\middle|\mathscr{H}_1\right)}{\mathsf{P}\left(Y=y\middle|\mathscr{H}_0\right)}\right); & \text{Case II,} \end{cases} \tag{7}$$

where $\mathsf{P}\left(Y_k = y_k\middle|\mathscr{H}_i\right)$ and $\mathsf{P}\left(Y = y\middle|\mathscr{H}_i\right)$, $i = 0, 1$, are given in (2) and (5), respectively. Hence, the optimal decision rule can be written as

$$\mathrm{d_{Optimal}} = \begin{cases} 0, & \text{if } \mathsf{LLR} \leq \tau, \\ 1, & \text{if } \mathsf{LLR} > \tau, \end{cases} \tag{8}$$

Remark 1 Since the general form of LLR in (7) may be computationally complex to be implemented in microscale, several approximations for LLR are proposed and investigated in Mosayebi et al. (2017).

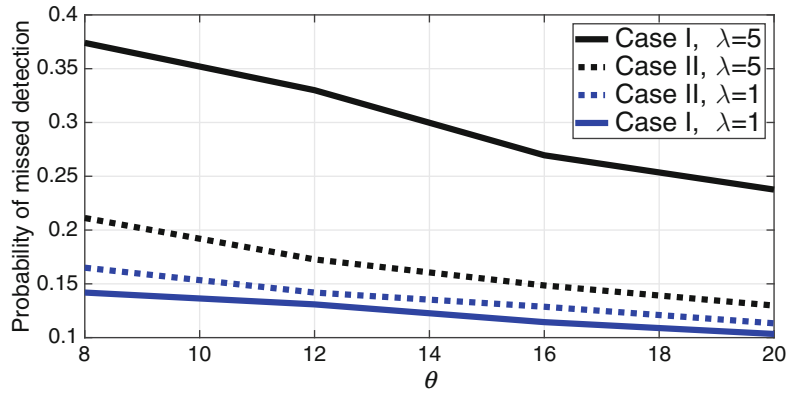
Remark 2 In practice, for event detection, some system parameters such as θ and $\mathrm{g}_i(\cdot), i = 0, 1$ are not known for the FC. As a first step to address this, in Ghavami and Lahouti (2017), one unknown system parameter is assumed. For this case, in Ghavami and Lahouti (2017), the maximum likelihood estimation of the unknown system parameter is derived for the initial decision at the NSs. Then, based on the estimation, a suboptimal hard decision rule for the NSs is proposed and investigated.

Numerical Result

For a simple comparison between the proposed schemes, i.e., Cases I and II, the performance of Neyman-Pearson detectors is compared in Fig. 1 for

$$\mathrm{g}_0(c_k) = \frac{\exp\left(-2.5c_k\right)}{\sum_{x\in\mathscr{X}}\exp\left(-2.5x\right)},$$
$$\mathrm{g}_1(c_k) = \frac{\exp\left(3.5c_k\right)}{\sum_{x\in\mathscr{X}}\exp\left(3.5x\right)}, \tag{9}$$

$L = 4$, $M = 2$, and $\mathrm{P_{fa}} = 0.05$. To this end, $\mathrm{P_m}$ is plotted versus θ for different values of λ. To have a fair comparison between Cases I and II, it is assumed that $\lambda_1 = \lambda_2 = \lambda$. From Fig. 1, it can be observed that as the value of θ increases, $\mathrm{P_m}$ decreases since the reporting scheme becomes more reliable. In addition, one can see that the relative performance of Case I and Case II depends on the value λ. For large λ,



Molecular Event Detection, Fig. 1 Probability of missed detection versus θ for Cases I and II using the respective optimal decision rules for $\mathrm{P_{fa}} = 0.05$. All curves were obtained via simulation

Case II outperforms Case I, whereas for small λ, Case I outperforms Case II. In fact, the advantage of Case II over Case I comes from the fact that the mean number of noise molecules for each type of molecule is assumed to be constant, i.e., $\lambda_1 = \lambda_2 = \lambda$, which leads to a higher overall noise for Case I compared to Case II. Alternatively, the trivial advantage of Case I over Case II is that the FC can distinguish between the molecules released by different NSs and exploit this additional knowledge for the improvement of the detection performance. Therefore, the superiority of Case I over Case II depends on the system parameters.

Future Directions

Although there are several works for molecular event detection, there exist some issues to be addressed. For instance, having a more adequate sensing model for particular applications such as event detection in blood vessels, body tissues, and air is highly recommended. The sensing model should be a function of the NS structure and the stochasticity of the event. In addition, considering more practical assumptions such as unknown sensing distribution, unknown locations for NSs, and unknown target location will result in different decision rules for the NSs and the FC. The idea of jointly estimation and detection which is introduced in the conventional communication systems can be also specialized for the molecular event detection systems.

Cross-References

- ▸ Applications of Molecular Communication Systems
- ▸ Drug Delivery via Nanomachines
- ▸ Molecular Bit Detection

References

Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41(3):1–58

Chen G, Roy C, Prasad PN (2016) Nanochemistry and nanomedicine for nanoparticle-based diagnostics and therapy. Chem Rev 116(5):2826–2885

Chude-Okonkwo U, Malekian R, Maharaj B, Vasilakos A (2017) Molecular communication and nanonetwork for targeted drug delivery: a survey. IEEE Commun Surv Tut 19(4):3046–3096

Felicetti L, Femminella M, Reali G, Lió P (2014) A molecular communication system in blood vessels for tumor detection. In: ACM 1st annual international conference nanoscale computing and communication, Atlanta

Ghavami S, Lahouti F (2017) Abnormality detection in correlated Gaussian molecular nano-networks: design and analysis. IEEE Trans NanoBiosci 16(3):189–202

Kay SM (1998) Fundamentals of statistical signal processings, vol 2. Springer/Prentice Hall PTR, Upper Saddle River

Mai TC, Egan M, Duong TQ, Di Renzo M (2017) Event detection in molecular communication networks with anomalous diffusion. IEEE Commun Lett 21(6):1249–1252

Mosayebi R, Jamali V, Ghoroghchian N, Schober R, Nasiri-Kenari M, Mehrabi M (2017) Cooperative abnormality detection via diffusive molecular communications. IEEE Trans NanoBiosci 16(8):828–842

Nakano T, Okaie Y, Kobayashi S, Koujin T, Chan CH, Hsu YH, Obuchi T, Hara T, Hiraoka Y, Haraguchi T (2017) Performance evaluation of leader-follower-based mobile molecular communication networks for target detection applications. IEEE Trans Commun 65(2):663–676

Okaie Y, Nakano T, Hara T, Nishio S (2016) Target detection and tracking by bionanosensor networks. Springer, Singapore

Wuab L, Qu X (2015) Cancer biomarker detection: recent achievements and challenges. Chem Soc Rev 44(10):2963–2997

Molecular Propagation

▸ Brownian Motion

Molecular Receiver Decoding

▸ Molecular Bit Detection

Molecular Target Detection

▸ Molecular Event Detection

Molecular, Biological, and Multiscale Communications

▸ Applications of Molecular Communication Systems

Multicarrier Index Keying with Orthogonal Frequency Division Multiplexing

▸ Index Modulation for OFDM

Multicast Mobility

▸ Mobility in Multicast

Multi-channel Modulation

▸ Principle of OFDM and Multi-carrier Modulations

Multicore Architectures

▸ Imprecise Computation Task Mapping on Multi-core Wireless Sensor Networks

Multilayer Wireless Networks

▸ Architectures, Key Techniques, and Future Trends of Heterogeneous Cellular Networks

Multimedia Security

Leo Yu Zhang[1] and Kai Zeng[2]
[1]School of Information Technology, Deakin University, Geelong, VIC, Australia
[2]George Mason University, Fairfax, VA, USA

Synonyms

Digital right management

Definitions

Multimedia is content composed of different types of digital objects such as text, audio, image, and video. Multimedia security addresses the protection of the intellectual property and the privacy of multimedia during acquisition, storage, processing, transmission, consumption, and disposal.

Historical Background

Advances in digital technologies have created significant changes in the way how multimedia is produced, distributed, marketed, and consumed. Typically, parties involved in multimedia business are *creator/owner*, *distributor*, and *consumer/user*. By preserving the security of multimedia, it is essentially protecting rights of all these parties. One notable milestone is the usage of Content Scramble System in the late 1990s, which is built on a stream cipher with 40-bit key, to protect commercially produced DVD discs by binding the content to a particular playback device (CSS 1996). It is later superseded by newer schemes such as the Content Protection for Recordable Media (CPR 2001) or the Advanced Access Content System (AAC 2005) due to low security strength. Similarly, the primary protection method for either the analog television or the digital television is encryption, such as VideoGuard (VG1 2012). The encryption system is built into the hardware of platform-supplied set-

top boxes so as to provide conditional access of the multimedia content. Additionally, it is common that the *creator* or *distributor* embeds a watermark-like logo in the broadcast content to increase brand recognition and assert ownership. The protection method used for online streaming television, such as the ones provided by Netflix and Hulu, is similar to dedicated terrestrial television, but this time the conditional access control is guaranteed by different technologies, such as the Common Encryption (CEN 2016) and the Encrypted Media Extensions (EME) (2017). For example, the EME, recommended by W3C, provides copy protection of streaming content by providing a communication channel between web browsers and agent software.

All the abovementioned protection methods fit the centralized business model, where the *creator*/*distributor* plays the core role, but, driven by different technologies, e.g., cloud computing, IoT, healthcare personnel, and social networks, the business model itself keeps evolving. For example, the patient provides the health record to a hospital's server for personnel Medicare service, and people upload photos to Flickr and share them with friends. Under this decentralized scenario, there is no (strong) monetary incentive for *owners* to deploy a complex digital rights management (DRM) system, yet they still concern about privacy leakage. Despite the lack of direct monetary gain, the ubiquitous privacy concerns and users' unwillingness to participate continuously promote collaboration among legislative institution, industry, and academia to design new mechanisms to preserve the rights of different parties. For example, HIPAA (HIP 1996) regulates transactions of electronic healthcare data, Google's Encrypted BigQuery (EBQ 2015) offers client-side encryption of user's query, and the prototype PIXEK (Pix 2018) provides an end-to-end protection mechanism for their photos on the cloud.

Foundations

From the discussion above, the protection of multimedia, especially when the media object is with low commercial value, requires the effort from law enforcement agency, coordination between different parties, and the appropriate usage of the technical measures. Here we mainly focus on technical measures, and the fundamental techniques used for multimedia security are discussed as follows.

Image/Video Source Identification

Image source identification aims to univocally link the image content to the device (i.e., camera) that captured it. Currently, the most promising technology to achieve this task exploits the detection of the sensor pattern noise (SPN) left by the acquisition device (Lukas et al. 2006). This footprint is universal (i.e., every sensor introduces one) and unique (i.e., two SPNs are different even in case of images captured by cameras of the same brand and model). Machine learning techniques, such as support vector machine (SVM), are usually used to classify different image sources (cameras) according to the extracted features (fingerprints) of SPNs. For static images, SPN has been proven to be robust to common processing operations like JPEG compression (Lukas et al. 2006). The photo response non-uniformity (PRNU) feature can be used for the case with severe compression (Milani et al. 2012). However, compression and distortion in digital videos can usually increase the false-alarm probability. In addition, when considering video streaming over wireless networks, blocking and blurring effects caused by packet loss should be tackled for effective camera source identification. Detailed discussions can be found in Chen et al. (2015).

Digital Watermarking

Digital watermarking relies on the fact that multimedia data is noise-tolerant. It works by first imperceptibly embedding a unique watermark (fingerprint) into each copy of the considered multimedia content and then detecting the existence of the unique watermark from a suspicious copy. The research focus for watermarking is to investigate the trade-offs among embedding capacity, imperceptibility, and robustness. Most, if not all, digital watermarking systems are devel-

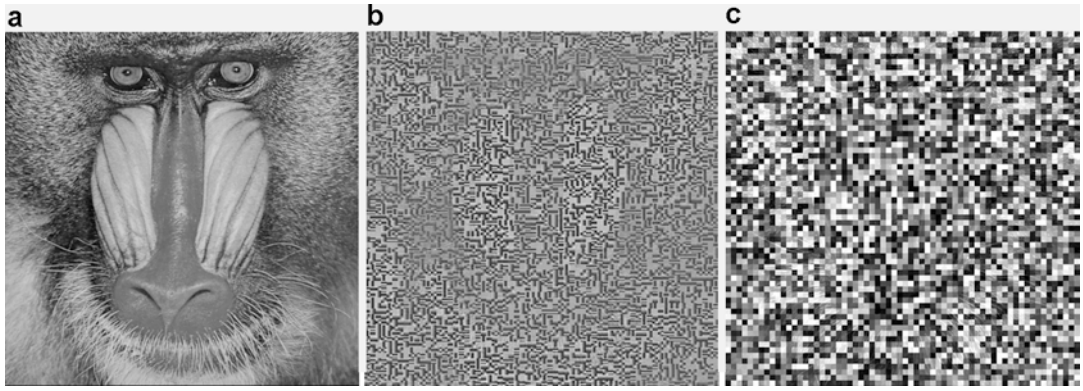
oped from the following two seminal designs: the spread spectrum system (Cox et al. 1997) and the quantization index modulation system (Chen and Wornell 2001). Detailed discussion about different digital watermarking systems can be found in Cox et al. (2007) and Barni and Bartolini (2004).

Format Compatible Encryption

Format Compatible Encryption (FCE) relies on the fact that multimedia data is always encoded and then stored under certain formats, such as WAV or MP3 for speech, JPEG or PNG for image, and WAV or MP4 for video. By modifying the behavior of encoder, FCE aims to make the syntax of encrypted data compliant with a standard decoder, i.e., it is decodable without decryption but suffered by (controllable) quality degradation. For authenticated users with the key, the joint encryption and encoding process can be removed without quality loss. For example, the JPEG compression of still images is composed by the DCT transformation, quantization, and Huffman coding processes. A FCE scheme for JPEG could be secretly shuffling or quantizing the transformation coefficients, as shown in Fig. 1. Detailed review of FCE scheme for multimedia data can be found in Stutz and Uhl (2012) and Massoudi et al. (2008).

Privacy-Aware Processing

Privacy-aware processing of multimedia data could rely on two different strategies: the one based on trusted execution environment and the one based on secure Multi-Party Computation (MPC). For the first class, representative technology includes Intel Software Guard Extensions (SGX) and AMD secure encrypted virtualization. For example, the work in Ohrimenko et al. (2016) presents a design that allows multiple hospitals to perform collaborative data analytics while guaranteeing the privacy of their patient datasets by using SGX. The MPC-based privacy-aware processing technology stems from the simple fact that all arithmetic operation can be expressed as a function of Boolean circuits, while Boolean circuit itself has a secured counterpart called Yao's Garbled circuit (Yao 1986). For example, Garbled circuit is used for server-side privacy-preserving image denoising (Zheng et al. 2017). Moreover, it is generally acknowledged that employing Garbled circuit, homomorphic encryption, and secret sharing to devise hybrid privacy-aware protocol for specific tasks will reduce cost and thus bring computation and bandwidth gain. Details about Intel Software Guard Extensions can be found in SGX (2015), and a MPC technique can be found in Lindell (2003).



Multimedia Security, Fig. 1 A FCE example on 512 × 512 image "Baboon": (**a**) Original image. (**b**) Standard JPEG decoding result when the 8 most significant DCT coefficients are randomly permuted. (**c**) Standard JPEG decoding result when all the DC coefficients are encrypted

Key Applications

Multimedia security technology is a key-enabler for many businesses and services. The applications described below cover its typical usage.

- Video Conferencing: Video conferencing is a special way to conduct meetings; it allows users to communicate in real-time remotely so as to save a ton in travel cost. It is widely used to enable discussion on the most important perspectives of business right up to executive level. Possible security threats to video conferencing could be enormous, but the typical way to harden the security is to combine an identity-based access control policy and data encryption by deploying dedicated video conferencing systems, such as Cisco Webex or Skype for Business.
- Online Streaming: It is reported that Netflix, after hitting 50.85 million subscribers in 2016, has surpassed the total number of subscribers of all US cable companies. By shifting from cabled TV to streaming service, users are not restricted to special playback devices and able to enjoy content anytime anywhere. The secure multimedia streaming is commonly achieved by two mechanisms: using encryption to enforce conditional access control and using watermarking to deter pirates distribution. Beside technical measures, law enforcement has been advocated by *creators/distributors* under this centralized business model. For example, in a recent response to a call from NTIA, MPAA suggests criminal charges against online video piracy and coordination between a broad range of online service providers (MPA 2018).
- Privacy-Preserving Multimedia Processing: Hosting or contributing personal data to a public server for personalized service becomes popular due to the maturity of artificial intelligence and cloud computing. In this case, content *owners* are likely to be individuals and the value of a single piece of data is not high but the whole dataset is the key to success of many businesses. For example, recommender systems collect and analyze user's past behavior and decisions to generate recommendations. Since privacy leakage could happen at any stage of the life cycle of multimedia data, it essentially requires imposing protect to data acquisition, storage, processing, transmission, consumption, and disposal.

Cross-References

- ▸ Access Control
- ▸ Data-Driven Security
- ▸ IoT Security
- ▸ Mobile Security and Privacy

References

(1996) Content Scramble System. http://www.dvdcca.org/css.aspx. Accessed 02 Aug 2018

(1996) Summary of the HIPAA Security Rule. https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html. Accessed 02 Aug 2018

(2001) 4C Entity CPRM Specification. http://www.4centity.com/specification.aspx. Accessed 02 Aug 2018

(2005) Advanced Access Content System Specification. https://www.aacsla.com/specifications/. Accessed 02 Aug 2018

(2012) VideoGuard DRM Design Guides. https://www.cisco.com/c/en/us/support/video/videoguard-drm/products-implementation-design-guides-list.html. Accessed 02 Aug 2018

(2015) Encrypted BigQuery Client. https://opensource.google.com/projects/encrypted-bigquery-client. Accessed 02 Aug 2018

(2015) Intel Software Guard Extensions. https://software.intel.com/en-us/sgx. Accessed 02 Aug 2018

(2016) ISO Common Encryption Protection Scheme. https://www.w3.org/TR/eme-stream-mp4/. Accessed 02 Aug 2018

(2017) W3C Recommendation Encrypted Media Extensions. https://www.w3.org/TR/encrypted-media/. Accessed 02 Aug 2018

(2018) Pixek. https://pixek.io/. Accessed 02 Aug 2018

(2018) Response of the Motion Picture Association of America to National Telecommunications and Information Administration Internet Priorities Inquiry. https://www.mpaa.org/wp-content/uploads/2018/07

/180717-MPAA-response-to-NTIA-internet-priorities-inquiry.pdf. Accessed 02 Aug 2018
Barni M, Bartolini F (2004) Watermarking systems engineering: enabling digital assets security and other applications. CRC Press, Boca Raton, FL
Chen B, Wornell GW (2001) Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. IEEE Trans Inf Theory 47(4):1423–1443
Chen S, Pande A, Zeng K, Mohapatra P (2015) Live video forensics: source identification in lossy wireless networks. IEEE Trans Inf Forensics Secur 10(1):28–39
Cox I, Kilian J, Leighton FT, Shamoon T (1997) Secure spread spectrum watermarking for multimedia. IEEE Trans Image Process 6(12):1673–1687
Cox I, Miller M, Bloom J, Fridrich J, Kalker T (2007) Digital watermarking and steganography. Morgan Kaufmann, San Francisco, CA
Lindell Y (2003) Composition of secure multi-party protocols: a comprehensive study. Springer Berlin Heidelberg, New York
Lukas J, Fridrich J, Goljan M (2006) Digital camera identification from sensor pattern noise. IEEE Trans Inf Forensics Secur 1(2):205–214. https://doi.org/10.1109/TIFS.2006.873602
Massoudi A, Lefebvre F, De Vleeschouwer C, Macq B, Quisquater J (2008) Overview on selective encryption of image and video: challenges and perspectives. Eurasip J Inf Secur 5:1–18
Milani S, Fontani M, Bestagini P, Barni M, Piva A, Tagliasacchi M, Tubaro S (2012) An overview on video forensics. APSIPA Trans Signal Inf Process 1:1–18
Ohrimenko O, Schuster F, Fournet C, Mehta A, Nowozin S, Vaswani K, Costa M (2016) Oblivious multi-party machine learning on trusted processors. In: USENIX security symposium, pp 619–636
Stutz T, Uhl A (2012) A survey of H.264 AVC/SVC encryption. IEEE Trans Circuits Syst Video Technol 22(3):325–339
Yao ACC (1986) How to generate and exchange secrets. In: 27th annual symposium on foundations of computer science (FOCS), pp 162–167
Zheng Y, Cui H, Wang C, Zhou J (2017) Privacy-preserving image denoising from external cloud databases. IEEE Trans Inf Forensics Secur 12(6):1285–1298

Multiparty Key Agreement for Wireless Networks

▸ Group Key Agreement for Wireless Networks

Multipath Transport Protocol

▸ Collaborative Multipath Transmission

Multiple Access in High Frequency

▸ Millimeter Wave NOMA

Multiple Access Methods

▸ Multiple Access Techniques

Multiple Access Technique for Cellular Wireless Networks

Qi-Yue Yu and Hong-Chi Lin
School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, Hei Longjiang, China

Synonyms

Division multiplexing

Definitions

Multiple access is the technique that multiple signals simultaneously access the same terminal, i.e., a base station, or one terminal simultaneously transmit multiple signals to different user terminals. Different resources, i.e., time, frequency, code, etc., can be assigned to different signals to avoid or reduce the multiple access interferences (MAC).

History Background

During the last decades, multiple access technique has been widely used in mobile communications and provided increasable capability and various services, i.e., from traditional voice to multimedia communications, as shown in Fig. 1. It plays a fundamental and important role in cellular networks.

A cellular always includes one base station (BS) and many users that are located in its coverage. And the coverage area of one cellular is generally determined by the transmit power of its BS and mathematically expressed as a hexagon, as shown in Fig. 2. There are two basic conceptions, one is uplink transmission and the other is downlink transmission. Uplink transmission means that many users transmit their signals to the BS, and downlink is the opposite processing that the BS simultaneously transmits signals to the multiple users. And multiple access is a technique that is used to reflect both uplink and downlink access processing.

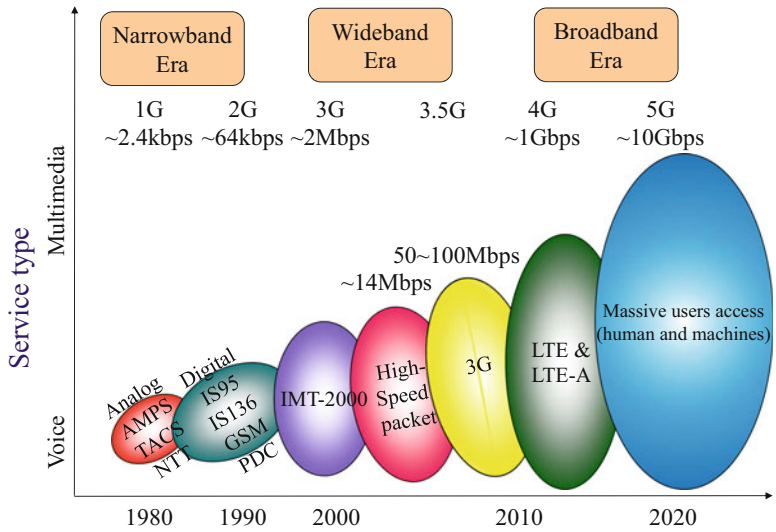
The multiple access technology is always viewed as one of the most important techniques for cellular wireless communications. With the increasing of data rates, the multiple access technology updates correspondingly.

There are lots of works that have been done on this topic, from the first-generation (1G) frequency-division multiple access (FDMA), the second-generation (2G) time-division multiple access (TDMA), the third-generation (3G) code-division multiple access (CDMA) (Adachi et al. 2005; Schulze and Luders 2005) to the fourth-generation (4G) orthogonal frequency-division multiple access (OFDMA). The evolution of multiple access techniques has pushed for a rapid advancement of wireless communications. Recently, many researchers believe that, for the coming fifth generation (5G), it will be non-orthogonal multiple access (NOMA) era.

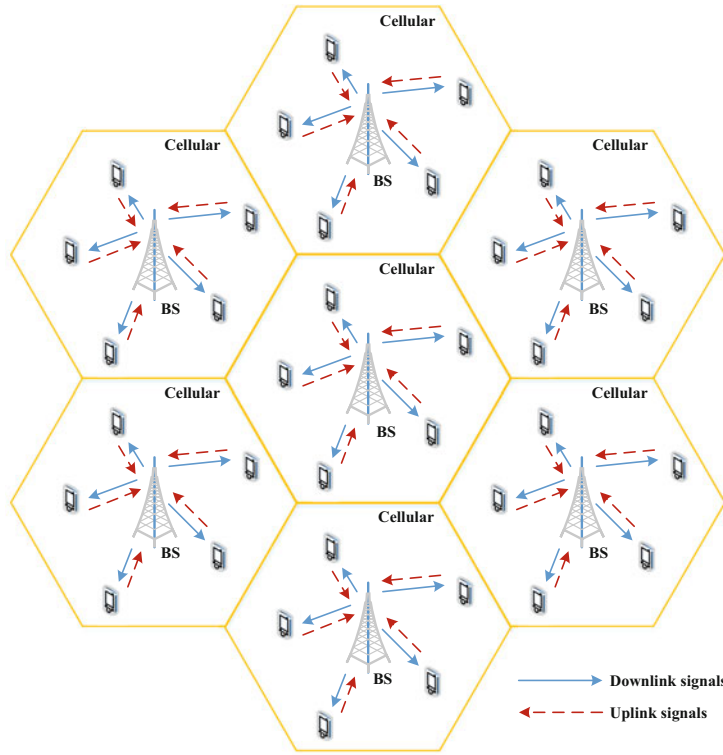
Classification

FDMA

Frequency-division multiple access (FDMA), as its name, utilizes frequencies to divide different users, and each user is allocated a carrier frequency. It is noted that an extra frequency band



Multiple Access Technique for Cellular Wireless Networks, Fig. 1 The development of wireless communications



Multiple Access Technique for Cellular Wireless Networks, Fig. 2 A diagram of cellular networks

is necessary between any adjacent frequencies, to reduce the sidelobes' interferences. For downlink, different users occupy different frequencies; therefore the BS can realize simultaneous transmission. For uplink, there is no multiple access interference (MAI) among different users, since each user occupies its respective frequency. A diagram of FDMA is shown in Fig. 3a.

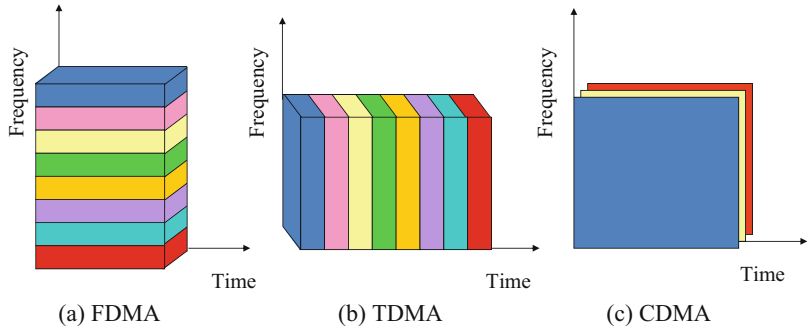
There are two major advantages of FDMA, one is that it is easy to realize and the other is there is no MAI. The disadvantages are also obvious, for example, it occupies a wider frequency bandwidth, leading to a lower spectrum efficiency (SE). It is interesting to design filters to reduce sidelobes, i.e., filter bank. If there is only a few users and there are enough frequency resources, FDMA is a promising technique.

TDMA

Time-division multiple access (TDMA) is similar to the FDMA case, which explores time instead of frequency to distinct different users, as shown in Fig. 3b. The major difference between TDMA and FDMA is that the synchronization is very important for TDMA. Without accurate time synchronization, there will be serous MAI among users, and even the performances of the whole system can be destroyed. Thus, many works have been done on the synchronization problem of TDMA.

CDMA

Code-division multiple access (CDMA) is widely used in 3G system, in which every user occupies the same time and frequency, and users are separated by different spreading sequences/vectors, as shown in Fig. 3c. Each user is assigned a spreading sequence/vector, and these spreading vectors are expected to be orthogonal to each other. The classical structure is named as direct-sequence (DS)-CDMA; besides this basically one, there are generally two approaches: single-carrier (SC)-CDMA and multicarrier (MC)-CDMA, as shown in Figs. 4 and 5, respectively (Adachi et al. 2005). Both SC- and MC-CDMA can flexibly provide variable-rate transmissions, with retaining multi-



Multiple Access Technique for Cellular Wireless Networks, Fig. 3 A diagram to show the features of FDMA, TDMA, and CDMA

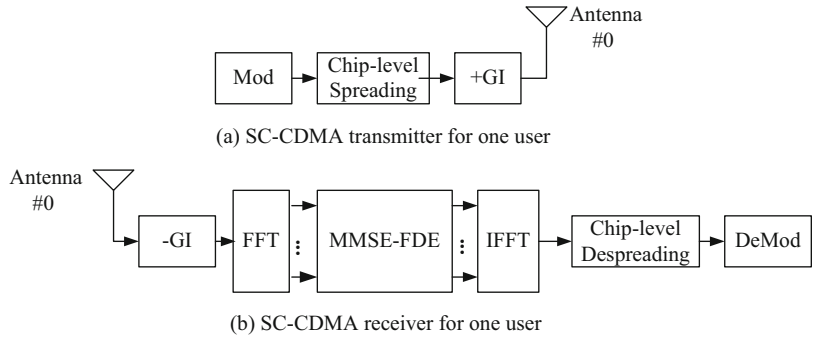
ple access capability. Actually, OFDM is a special case of MC-CDMA when the spreading factor is one.

The SC-CDMA and MC-CDMA, respectively, explore time-domain and frequency-domain spreading (chip-level spreading is used for further discussion). At the transmitter, after the binary information data is modulated, the data-modulated symbols are then spread, time domain spread for SC-CDMA, and frequency domain spread for MC-CDMA. Without/with IFFT processing for SC-/MC-CDMA, the processed signals insert GI and then transmit to the channels.

At the receiver, frequency-domain equalization (FDE) can be used to improve the BER performances. It has been proved that the use of minimum mean square error (MMSE) weight can provide the best BER performances among various FDE weights, since the MMSE weight can provide the best compromise between the noise enhancement and suppression of frequency-selective fading channels. Simulation results show that MC-CDMA with MMSE-FDE provides much better BER performances than SC-CDMA with traditional coherent rake combining. It has been derived that SC-CDMA with proper FDE, instead of the traditional rake combining, can achieve good BER performances that are comparable to MC-CDMA.

For uplink CDMA systems, different users' signals are asynchronously arrived at the BS via different fading channels, leading to serious MAIs, which limit the uplink capacity. The capability of MAI suppression achievable with pure CDMA is not sufficient. It is an important research topic to suppress the effect of MAI. There are many interesting methods, such as multiuser detection (MUD), soft-iterative multistage receiver, etc.

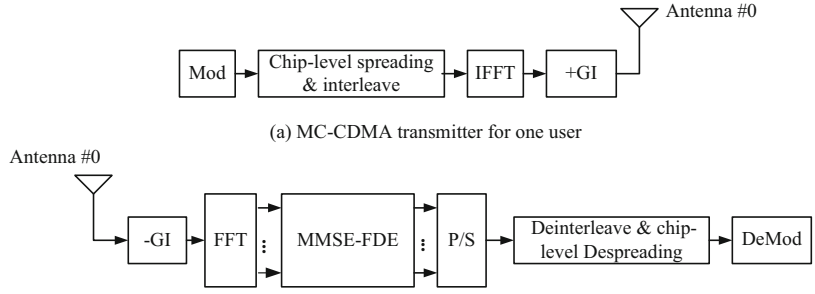
In papers (Liu and Adachi 2006; Yu et al. 2013, 2014), two-dimensional (2D) block spread CDMA is proposed, which can be applied to solve the MAI and achieve frequency diversity gain in frequency-selective fading channels. The 2D block spread CDMA consisted of chip-level and block-level spreading, and they are implemented as different roles. The chip-level spreading plays the same part as traditional SC-/MC-CDMA, which can achieve frequency diversity gain by using MMSE-FDE in the receiver. In 2D block spread CDMA, block-level spreading is performed to each block after chip-level spreading. Before the transmission, the GI, which is larger than or equal to the maximum delay among different users, is inserted. At the receiver, after removing the GI, block-level de-spreading is performed in order to remove the MAI. If the maximum timing offset among uses is within the GI length, perfect removal of MAI is possible in the case of block fading, i.e., the channel stays constant during at least one block. Both chip-level spreading codes and block-level spreading codes can be constructed using the orthogonal variable spreading factor code tree. The 2D block SC-CDMA and MC-CDMA are given in Figs. 6 and 7.



(a) SC-CDMA transmitter for one user

(b) SC-CDMA receiver for one user

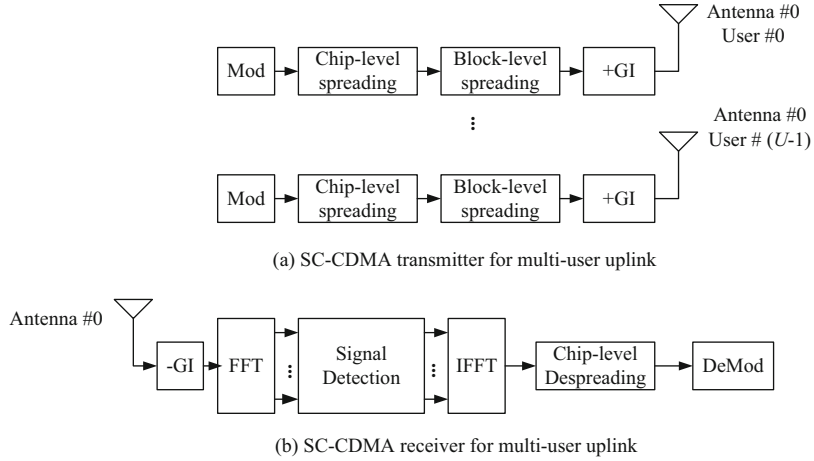
Multiple Access Technique for Cellular Wireless Networks, Fig. 4 A diagram of SC-CDMA transmitter and receiver, where "Mod" and "DeMod" indicate modulation and demodulation, respectively, and "FDE" means frequency-domain equalization



(a) MC-CDMA transmitter for one user

(b) MC-CDMA receiver for one user

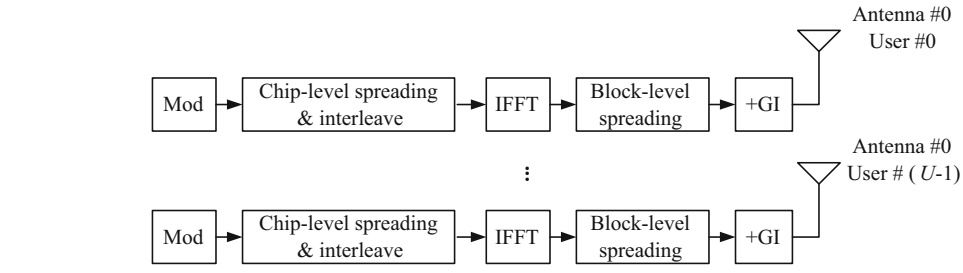
Multiple Access Technique for Cellular Wireless Networks, Fig. 5 A diagram of MC-CDMA transmitter and receiver, where "Mod" and "DeMod" indicate modulation and demodulation, respectively, and "FDE" means frequency-domain equalization



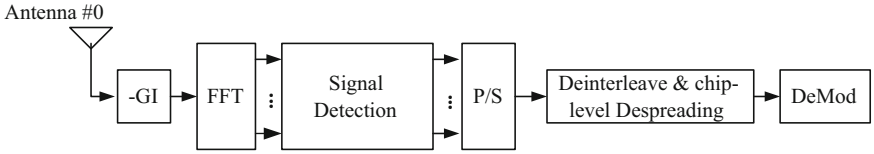
(a) SC-CDMA transmitter for multi-user uplink

(b) SC-CDMA receiver for multi-user uplink

Multiple Access Technique for Cellular Wireless Networks, Fig. 6 A diagram of 2D block SC-CDMA transmitter and receiver, where "Mod" and "DeMod" indicate modulation and demodulation, respectively, and "FDE" means frequency-domain equalization



(a) MC-CDMA transmitter for multi-user uplink



(b) MC-CDMA receiver for multi-user uplink

Multiple Access Technique for Cellular Wireless Networks, Fig. 7 A diagram of 2D block MC-CDMA transmitter and receiver, where "Mod" and "DeMod" indicate modulation and demodulation, respectively, and "FDE" means frequency-domain equalization

IDMA

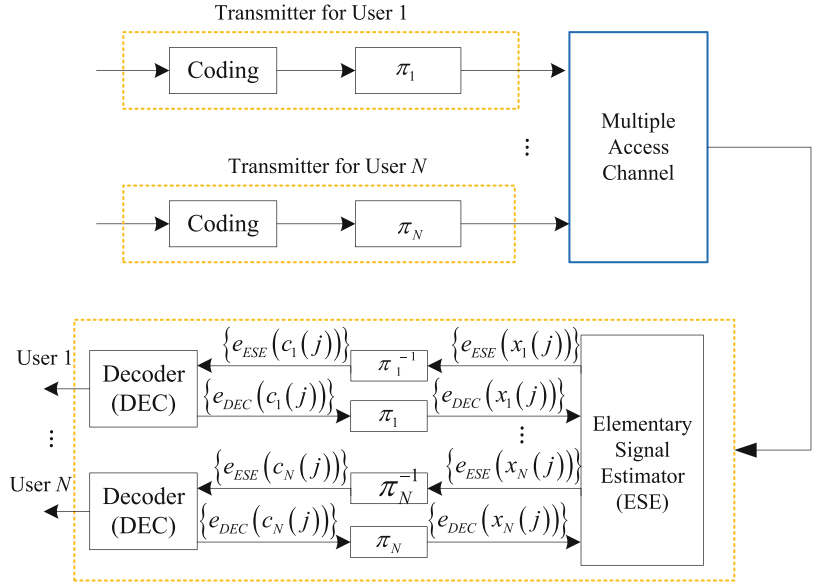
Interleave-division multiple access (IDMA) is also an interesting multiple access scheme, which was proposed in Ping Li et al. (2006) by Prof. Li, as shown in Fig. 8. Actually, IDMA and CDMA work in a similar way, both of which occupy all the same time- and frequency-domain to transmit data information. There are serveral differences between IDMA and CDMA, such as: (1) in CDMA, transmit symbols are modulated by different spreading sequences/vectors to distinguish different users, whereas IDMA exploits different interleavers to separate different signals, and (2) at a receiver, de-spreading is used to recover transmitted signals in CDMA, while iterative algorithm, just like decoding algorithms for low-density parity-check (LDPC), is used in IDMA. IDMA can be flexibly applied in different modulations and channel conditions.

OFDMA

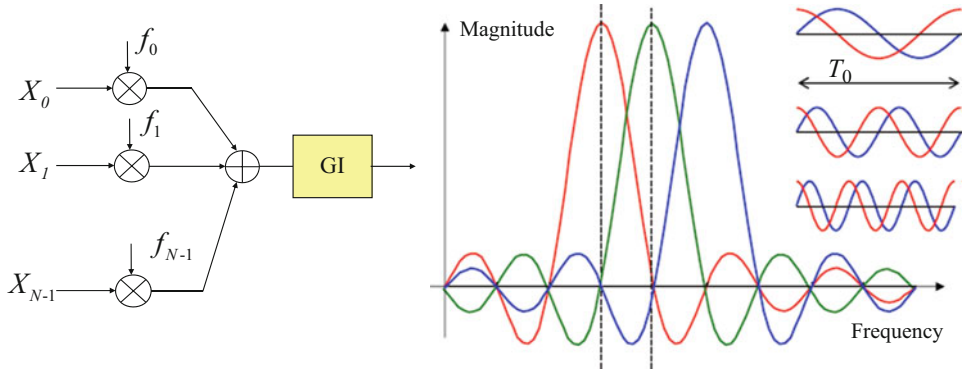
Orthogonal frequency-division multiple access (OFDMA) or orthogonal frequency-division multiplexing (OFDM) is one of the most appreciated multiple access schemes. Actually, OFDM has been widely used in 4G; WLAN, i.e., IEEE 802.11a/g; and digital audio and video broadcasting (DAB, DVB) systems. It also uses orthogonal subcarriers to separate users; however its bandwidth is much smaller than the FDMA case. And it doesn't need extra bandwidth to avoid the effect of sidelobe. Generally speaking, a subcarrier's spectrum of an OFDM symbol is the Fourier transform of a rectangular window of symbol duration T_s, which is a sinc-function with zeros at $\frac{1}{T_s}$. And the centers of other carriers are put in these zeros; thereby, the subcarriers are orthogonal to each other, as shown in Fig. 9.

DAB always explores COFDM (coded OFDM), in which the data transmitted on the subcarriers is protected by forward error correction (FEC) coding. COFDM also allows different groups of bits to be protected with a different strength code rate. Considering the subcarriers are subject to flat fading, interleaving can be used to improve the bit error rate (BER) performance of OFDM. For an OFDM system, interleaving is generally classified as time and frequency interleaving or random and regular interleaving.

An OFDM modulator mainly consists of two parts: IDFT (inverse discrete Fourier transform) and guard interval (GI) that is also called cyclic prefix (CP), as shown in Fig. 10. The IDFT makes OFDM practically feasible. And we don't need many filters and oscillators any more. In reality,



Multiple Access Technique for Cellular Wireless Networks, Fig. 8 A diagram of IDMA transmitter and receiver given in Ping Li et al. (2006)



$$\text{where } x_n = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X_k e^{j2\pi nk/N}, n = 0,1,\ldots N-1$$

Multiple Access Technique for Cellular Wireless Networks, Fig. 9 A diagram to show the spectrum of OFDM signals

the actual processing utilizes IFFT (inverse fast Fourier transform) instead of IDFT.

The GI (or CP) should be larger than, or at least equal to, the maximum multipath delay. Hence, it can avoid the inter-symbol interferences (ISI) caused by the multipath delay. The GI can efficiently avoid the effect of ISI; however, it will decrease the throughput of an OFDM system. Therefore, some papers (Liu et al. 2017a) try to remove GI, with acceptable detection complexity.

There are lots of papers focusing on OFDM system, from synchronization, nonconstant power envelope, to the channel estimation, etc. For the synchronization issue, the frequency-domain nature of OFDM makes the effect of

several synchronization errors to be explained with the properties of DFT (discrete Fourier transform). In fact, OFDM synchronization functions can be performed either in time or frequency domain, which is different from the traditional single-carrier systems. To decrease the effect of asynchronization, low sidelobes of OFDM symbols are appreciated. Filtered multitone (FMT) modulation is a useful scheme, which is a kind of filter bank implementation of multicarrier system (Proakis 2009).

Another major problem of OFDM is relatively high PAPR (peak-to-average power ratio) issue. In general, the large signal peak occurs when the signals in the sub-channels add constructively in phase. Such large signal peaks may result in clipping of the signal voltage in a DAC (digital-to-analogue converter), when the OFDM signal is synthesized digitally. It may saturate the power amplifier, leading to the intermodulation distortion. To avoid the PAPR problem, there are generally two kinds of methods. One is exploring distortion methods, i.e., clipping, filtering, peak cancellation, etc.; and the other is exploring the methods without distortion, for example, selective mapping, partial transmit sequence, etc (Proakis 2009).

Figure 10 shows the implementation of an OFDM symbol. After passing through IFFT processor, the values are fed to DACs. And the low-pass filtered signals of the real and imaginary streams are amplitude modulated onto RF carrier and then added together and sent through a band-pass filter (BPF) to the transmit antenna.

NOMA

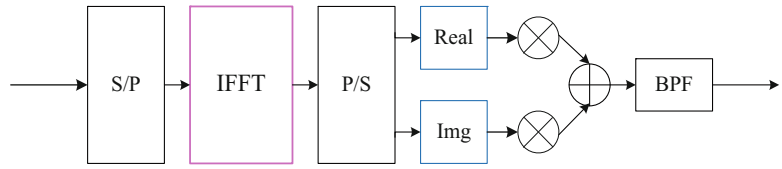
Non-orthogonal multiple access (NOMA) explores the same resource block to serve different users, thus transfers information of multiple users superposition (Ding et al. 2017; Dai et al. 2015). It is different from orthogonal multiple access (OMA), which assigns orthogonal resources to different users. Owing to the orthogonal feature, different signals can be separated without MAI. However, the spectrum efficiency (SE) of OMA schemes is low.

NOMA can improve SE by overloaded information, which can be realized in power domain, by different modulation amplitudes, and/or code domain, via multidimensional constellation codebooks. Therefore, the core of all the emerging multiple access techniques is to optimally combine modulation and coding techniques to support high spectrum efficiency.

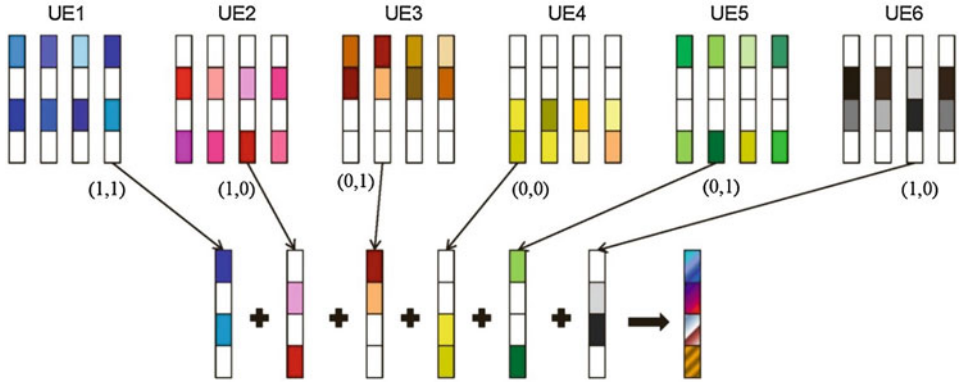
Nowadays, the multiple access schemes of 5G need to support massive users of both human and machines. High SE multiple access schemes become emergency. Many novel multiple access techniques have been proposed, such as power-domain NOMA, sparse code multiple access (SCMA), pattern division multiple access (PDMA), multiuser shared access (MUSA), etc. Compared to the well-known OMA schemes such as OFDMA, NOMA provides improved SE by transmitting overloaded information in the same time resources or frequency resources (Ding et al. 2017; Dai et al. 2015).

Power-domain NOMA allocates different power to different users according to their different channel state information (CSI). In general, a user with better CSI is assigned a small power; in contrast, a user with worse CSI is assigned a large power to keep fairness. And successive interference cancelation (SIC) algorithm is utilized at the receiver to decode the superimposed signal in sequence. Power-domain NOMA is easy to realize with low complexity, and it is applicable for Internet of Things networks.

SCMA, which was first proposed in 2013 (Nikopour and Baligh 2013), can also be viewed as a type of NOMA technique, which works based on traditional DS-CDMA technique, as shown in Fig. 11. Its input binary bits first pass through a modulator to map the binary bits to complex symbols, and then direct sequence spreading is performed to all mapped complex symbols. Compared to the DS-CDMA, SCMA combines modulation and spreading, and thus it can achieve shaping gain of multidimensional constellations due to its optimal SCMA codebook design (Nikopour and Baligh 2013; Nikopour et al. 2014; Zhou et al. 2017).



Multiple Access Technique for Cellular Wireless Networks, Fig. 10 A diagram of OFDM signals



Multiple Access Technique for Cellular Wireless Networks, Fig. 11 A diagram of SCMA system that serves six users with four resources

Key Applications

Cellular systems, Satellite communications, Internet of Things (IoT) systems, etc.

Cross-References

- Beam Division Multiple Access (BDMA) Transmission
- Cyclic Prefix-Free OFDM/OFDMA Systems, Design, and Implementation
- Non-orthogonal Multiple Access (NOMA)
- Multiple Access Techniques
- Sparse Code Multiple Access

References

Adachi F, Garg D, Takaoka S, Takeda K (2005) Broadband CDMA techniques. IEEE Wirel Commun 12(2):8–18

Dai L, Wang B, Yuan Y, Han S, Chih-Lin I, Wang Z (2015) Nonorthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. IEEE Commun Mag 53(9):74–81

Ding Z, Lei X, Karagiannidis GK, Schober R, Yuan J, Bhargava VK (2017) A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends. IEEE J Sel Areas Commun 35(10):2181–2195

Liu X, Chen H, Chen S, Meng W (2017a) Symbol cyclic shift equalization algorithm – a CP-free OFDM/OFDMA system design. IEEE Trans Veh Technol 66(1):282–294

Liu X, Chen H, Lyu B, Meng W (2017b) Symbol cyclic shift equalization PAM-OFDM – a low complexity CP-free OFDM scheme. IEEE Trans Veh Technol 66(7):5933–5946

Liu L, Adachi F (2006) 2-Dimensional OVSF spread/chip-interleaved CDMA. IEICE Trans Commun E89-B(12):3363–3375

Nikopour H, Baligh H (2013) Sparse code multiple access. In: Proceedings of the IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), pp 332–336

Nikopour H, Yi E, Bayesteh A, Au K, Hawryluck M, Baligh H, Ma J (2014) SCMA for downlink multiple access of 5G wireless networks. In: Proceedings of the IEEE GLOBECOM, pp 3940–3945

Ping L, Liu L, Wu K, Leung WK (2006) Interleave division multiple-access. IEEE Trans Wirel Commun 5(4):938–947

Proakis JG (2009) Digital communications, 5th edn. Publishing House of Electronics Industry, Beijing

Schulze H, Luders CN (2005) Theory and applications of OFDM and CDMA: wideband wireless communications. John Wiley & Sons, pp 269–272

Yu Q, Meng W, Adachi F (2013) Two-dimensional block-spread CDMA relay using virtual-four-antenna STCDTD. IEEE Trans Veh Technol 62(8):3813–3827

Yu Q, Meng W, Adachi F (2014) Combined code reuse scheme with two-dimensional OVSF codes assignment algorithm for uplink multi-user/multi-rate block spread multi-cellular CDMA. Wirel Commun Mob Comput 14(13):1314–1328
Zhou Y, Yu Q, Meng W, Li C (2017) SCMA codebook design based on constellation rotation. IEEE Int Conf Commun 1–6

Multiple Access Techniques

Fan Jiang[1], Zijun Gong[1], Kun Hao[2,3], and Yan Zhang[2,3]
[1]Department of Electrical and Computer Engineering, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL, Canada
[2]Department of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, NL, Canada
[3]Tianjin Chengjian University, Tianjin, China

Synonyms

Channel access methods; Channel access techniques; Multiple access methods

Definition

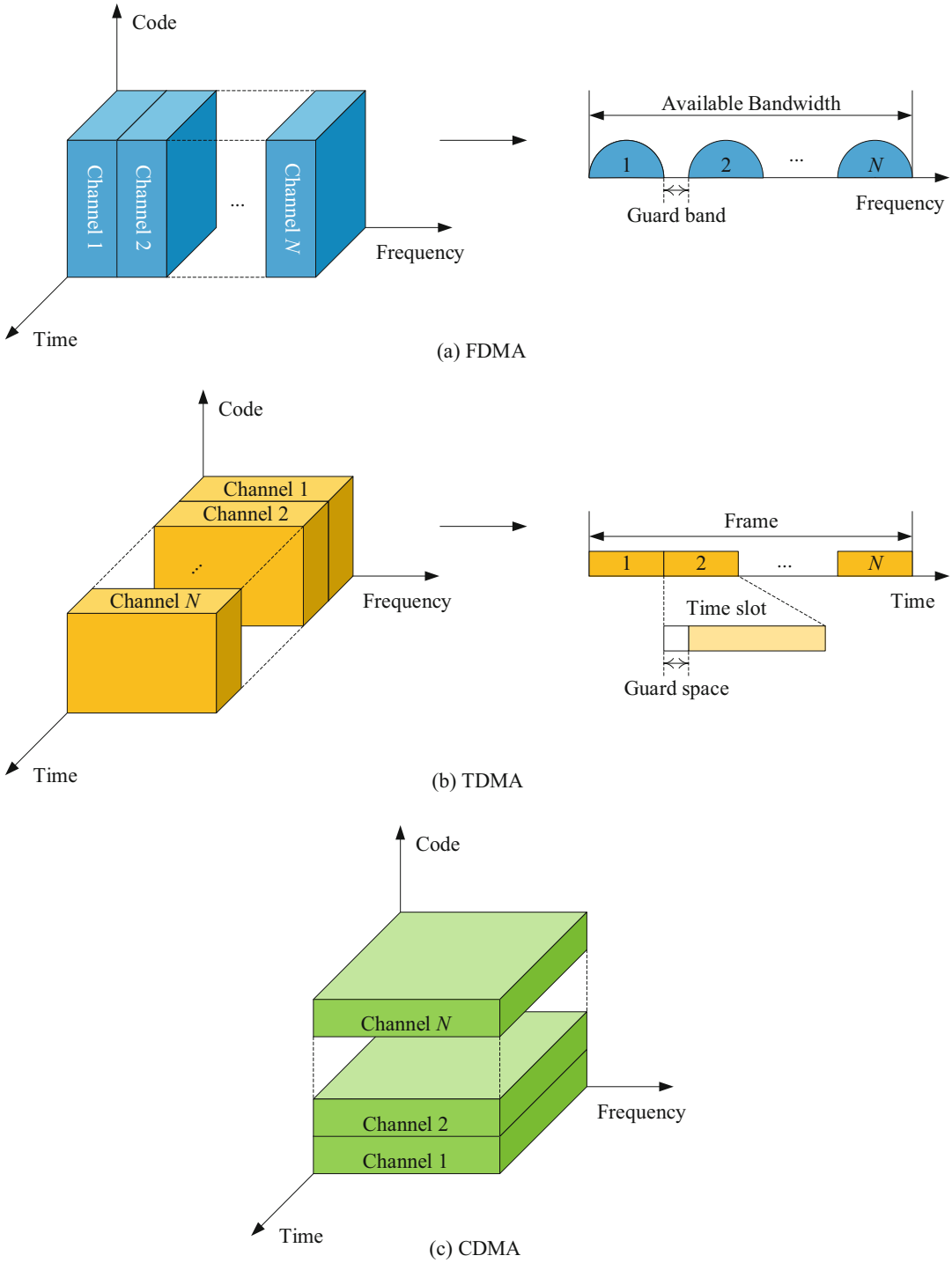
Multiple access techniques allow multiple terminals to share the common communication medium based on multiplexing. The multiplexing is provided in the physical layer, and the shared communication medium can be wired cables or wireless spectrum. The availability of the communication medium is limited, for example, limited frequency band and limited time. Therefore, multiple access techniques are essential to maintain successful communication among multiple devices. The basic well-known multiple access techniques are frequency division multiple access (FDMA), time division multiple access (TDMA), code division multiple access (CDMA), and spatial division multiple access (SDMA). A class of non-orthogonal multiple access (NOMA) techniques has recently gained much interest. The control mechanism to distribute channels to users is also known as medium access control (MAC).

Historical Background

The applications of the multiple access techniques present in every generation of wireless communication systems. In the first US analog cellular system, AMPS (Advanced Mobile Phone System) is based on FDMA. TDMA and CDMA dominate in the second-generation (2G) wireless communication systems. For example, in GSM (Global System for Mobile communication), TDMA is adopted for distinguished users, while the CDMA is adopted in IS-95 (Interim Standard 95) systems. The varieties of multiple access techniques based on TDMA and CDMA have been widely adopted in ad hoc sensor networks. CDMA becomes the dominant multiple access technique in the third-generation (3G) wireless communication systems, with the prevailing worldwide standards known as CDMA2000, WCDMA (Wideband CDMA), and TD-SCDMA (Time Division – Synchronous CDMA). In fourth generation (4G), the hybrid multiple access techniques, including orthogonal FDMA (OFDMA), TDMA, and SDMA, are preferred in multiple-input multiple-output (MIMO) systems (Miao et al. 2016). With the increase of the MIMO size, massive MIMO, also known as large-scale MIMO, together with millimeter wave communications, becomes one of the promising techniques in fifth-generation (5G) wireless communication systems (Marzetta 2010; Dai et al. 2015; Sun et al. 2015). In massive MIMO, the SDMA techniques, for example, BDMA (beam division multiple access), are recently studied in Sun et al. (2015).

Foundations

The standard FDMA, TDMA, and CDMA techniques can be further illustrated in Fig. 1. As it is shown in Fig. 1a, when FDMA is adopted,

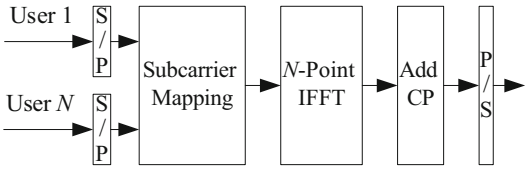


Multiple Access Techniques, Fig. 1 Illustration of FDMA/TDMA/CDMA techniques

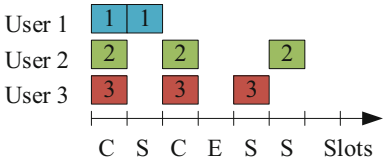
the system bandwidth is divided into several narrowbands, with nonoverlapping frequency slots, (i.e., the adjacent frequency bands are separated by a guard band). Usually, in a FDMA system, each user occupies a pair of frequency bands (i.e., frequency division duplexing (FDD)). One is for forward channel (downlink), and the other is for the reverse channel (uplink). FDMA was the initial multiple access technique for cellular systems. An advanced form of FDMA is the orthogonal frequency division multiple access (OFDMA), which is shown in Fig. 2. In OFDMA, each user is assigned a subset of subcarriers, and the low-complexity fast Fourier transform (FFT) operations can be used at both the transmitter and receiver side in the system (Miao et al. 2016).

In TDMA systems, the data frame is segmented into nonoverlapping time slots, separated by guard space. When the time slot is assigned to a user, the user takes up the entire frequency band for data transfer. An advanced form of TDMA is dynamic TDMA (DTDMA), where different time slots may be assigned to the same user according to the scheduling algorithm. For example, the user terminal may use the first time slot for the first data frame but use another time slot for the next data frame. DTDMA usually works with MAC protocol to randomly access the shared channel. Common examples for DTDMA include the slotted ALOHA in GSM systems (initial access stage for users) and carrier sense multiple access with collision avoidance (CSMA/CA) in IEEE 802.11. An example of the slotted ALOHA can be illustrated in Fig. 3 (Tanenbaum 2003). In slotted ALOHA, the time frame is divided into equal size slots. The data packet from each user is transmitted at the beginning of next slot. If collision occurs, retransmitted packet will select future slots with probability p, until successful. Specifically, in Fig. 3, collisions occur in the first and third slot, while packets 1, 2, and 3 from each user are successfully delivered in the second, sixth, and fifth slot, respectively. DTDMA schemes have been widely in ad hoc networks since the channel can be randomly accessed.

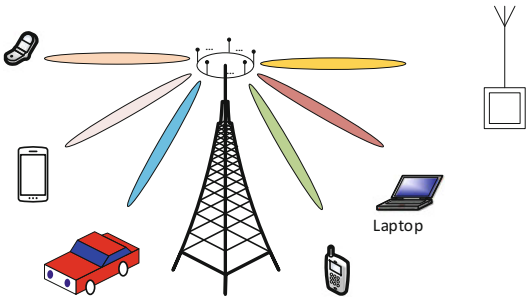
The CDMA scheme is based on spread-spectrum code. Signals from different users are



Multiple Access Techniques, Fig. 2 OFDMA block diagram



Multiple Access Techniques, Fig. 3 Slotted ALOHA, *C* collision, *S* successful, *E* empty



Multiple Access Techniques, Fig. 4 Illustration of BDMA

transferred simultaneously over the same carrier frequency. At the receiver side, the user signal is separated by cross-correlation of the received signal with each possible code sequence. Two typical forms of CDMA are the direct sequence spread spectrum (DSSS) and the frequency-hopping spread spectrum (FHSS) (Miao et al. 2016). The CDMA scheme is known to have low signal-to-noise ratio (SNR) working region, which indicates lower transmission power is required in the system. The near-far problem is a serious one in CDMA systems.

SDMA utilizes the spatial separation of the users to fully use the frequency spectrum. In Fig. 4, an example of SDMA techniques, i.e., BDMA is illustrated (Sun et al. 2015). As it is shown in Fig. 4, multiple beams are generated by using the phased array antennas. SDMA has been

one of the promising multiple access techniques for 5G.

Besides the orthogonal multiple access techniques, a class of NOMA schemes is also discussed for 5G (Dai et al. 2015). NOMA accommodates multiple users through non-orthogonal resource allocation. Generally, NOMA can be implemented via power or code domain multiplexing. Examples of NOMA schemes are power-domain NOMA, multiple access with low-density spreading code, sparse code multiple access, multiuser shared access, and so on (Dai et al. 2015).

Key Applications

Multiple access techniques are fundamental in wireless communication systems, including cellular networks, ad hoc networks, and vehicular networks, for multiple user terminals to access the shared common medium.

Cross-References

- ▸ Non-orthogonal Multiple Access
- ▸ Massive MIMO
- ▸ Medium Access Control
- ▸ Resource Allocation

References

Dai L, Wang B, Yuan Y, Han S, I C, Wang Z (2015) Non-orthogonal multiple access for 5G: solutions, challenges, opportunities and future research trends. IEEE Commun Mag 53(9):74–81

Marzetta TL (2010) Noncooperative cellular wireless with unlimited numbers of Base Station antennas. IEEE Trans Wirel Commun 9(11):3590–3600

Miao G, Zander J, Sung KW, Slimance B (2016) Fundamentals of mobile data networks. Cambridge University Press, New York

Sun C, Gao XQ, Jin S, Matthaiou M, Ding Z, Xiao CS (2015) Beam division multiple access transmission for massive MIMO communications. IEEE Trans Commun 63(6):2170–2184

Tanenbaum AS (2003) Computer networks. Prentice Hall PTR, Upper Saddle River

Multiple Network Paths

▸ Collaborative Multipath Transmission

Multi-provider Pricing

▸ Oligopoly Pricing in Wireless Networks

Multi-tone Modulation

▸ Principle of OFDM and Multi-carrier Modulations

Multiuser Information Theory

▸ Network Information Theory