DOI: 10.3969/j. issn.1673-5188.2017.S2.006
http://kns.cnki.net/kcms/detail/34.1294.TN.20180117.1109.002.html, published online January 17, 2017

Measuring QoE of Web Service with Mining DNS Resolution Data

LIU Yongsheng[1], GU Yu[2], WEN Xiangjiang[1], WANG Xiaoyan[3], and FU Yufei[4]
(1. Network Technology Research Institute, China United Network Communications Corporation Limited, Beijing 100048, China;
2. Department of Computer Science, Hefei University of Technology, Hefei 230009, China;
3. Ibaraki University, Mito 310-8512, Japan;
4. Personal Computer and Intelligent Equipment Group, Lenovo, Beijing 100094, China)

Abstract

Internet service providers (ISPs) are paying more attention to the Quality of Experience (QoE) of the web service that is one of the most widely used Internet services. Measuring it with existing systems deployed in the network so far may save investment for ISPs since no additional QoE system is required. In this paper, with Domain Name System (DNS) resolution data that are available in the ISP's network, we propose the First Webpage Time (FWT) algorithm in order to measure the QoE of the web service. The proposed FWT algorithm is analyzed in theory, which shows that its precision is guaranteed. Experiments based on the ISP's DNS resolution data are carried out to evaluate the proposed FWT algorithm.

Keywords

FWT algorithm; web service; QoE; ISP

1 Introduction

Quality of Experience (QoE) refers to the degree of delight or annoyance of the user of an application or service [1]. The QoE of web service affects users to choose an Internet service provider (ISP) or an Internet content provider (ICP) to a certain extent. Hence, it attracts more and more attentions from ISPs and ICPs [2]–[4].

The probe system [5] is a widely used tool of obtaining QoE data. A probe deployed at the user side simulates a user to browse webpages, records the departure time and arrival time of every packet, and then summarizes the Key Quality Indicator (KQI). However, the probe system faces many intrinsic deflects. It is difficult to deploy probes at the user side in a large scale due to the user information protection, the compensation of traffic charges, the maintenance of equipment, and so on. This results in the limitation in the number of the QoE data. Moreover, the simulation results deviate from the real user experience to some extent since the probe system has to seek the idle interval of users to launch a task.

Using existing systems deployed in the network to measure the QoE of web service can overcome deflects of the probe system and also save investment for ISPs. Many systems, such as the Domain Name System (DNS) [6], [7], Network Management System (NMS), and Deep Packet Inspection (DPI) system, have been deployed in the ISP's network already, and mining data from those systems has become one of the research hot spots.

Some research related to mining DNS data has been carried out. Keralapura et al. [8] and Giordano et al. [9] studied the user browsing behaviors and built the user profiles. Hours et al. [10] exploited a causal method that is for analyzing parameters of the DNS service. Jang et al. [11] proposed to append the refresh expired cached records to solicited DNS queries. Ma et al. [12] and Gao et al. [13] investigated the DNS query patterns of domain names and detected malicious agents that may launch attacks to the DNS. Rijswijk-Deij et al. [14] created a unique active measurement infrastructure for DNS.

In this paper, we first propose the First Webpage Time (FWT) system, a novel KQI, to measure the QoE of web service. Next,we propose the FWT algorithm with mining DNS resolution data.

The rest of this paper is organized as follows. The system model and symbols used in the paper are described in Section 2. The proposed algorithm and its corresponding analysis is presented in Section 3. The experiments are presented in Section 4. Finally, the paper is concluded in Section 5.

2 System Description

We assume the webpage used in the system is W. The domain name of W is denoted by dom_{head}. Note that W just contains the contextual contents like the Hyper Text Markup

Language (HTML) codes and Cascading Style Sheet (CCS) codes. Any embedded file such as an image is not included in W. The displayed webpage ordinarily contains embedded contents like images, frame pages, audio files, and so on. In W, the embedded contents are represented by hyperlinks. The hyperlinks are discovered by parsing the contextual contents of W. To display W, the browser resolves the domain name of the hyperlink of the embedded file from the DNS server firstly, and then retrieves the embedded file from the web server. Suppose the domain name of the last hyperlink in W is denoted by dom_{tail}. The last hyperlink means it locates in the ending part of W. Furthermore, the last hyperlink in W also implies that dom_{tail} does not appear in W before. In other words, it is the first time for dom_{tail} to emerge in W. How to select dom_{tail} is described in the next section.

A browser or operating system (OS) has DNS cache for decreasing the waiting time of users. That is, an IP address of a domain name would be locally stored for a moment. Usually, the memorized duration is short. Beyond the duration, the browser has to resolve the domain name from the DNS server again. Let p denote the probability that a domain name has to be resolved from the DNS server when used. Assume that the DNS server records the resolution when a domain name is requested by the user. The record R has at least four elements, denoted by a 4-tuple as shown in (1) in which t represents the resolved time of the DNS server, IP_u represents the IP address of the user, dom represents the requested domain name, and IP_r represents the IP address corresponding to dom.

$$R=<t,IP_u,dom,IP_r> . \tag{1}$$

Considering the Content Delivery Network (CDN) and the dispatch among web servers, IP_r may be different for two requests of the same domain name. For simplicity, we assume there is only one web server for each domain name. This can be achieved by classifying records by dom and IP_r. This assumption does not affect the proposed algorithm at all.

Let the set g in (2) includes all records when W is browsed by a user in one time. It is possible that g contains all domain names in W. Or, some records of domain names in W may be missed in g due to the DNS cache.

$$g=\{R_2,R_3,...,R_{tail}\} . \tag{2}$$

The critical records in this system are R_{head} and R_{tail} where $R_{head}=<t_{head},IP_u,dom_{head},IP_{head}>$ and $R_{tail}=<t_{tail},IP_u,dom_{tail},IP_{tail}>$. Let g', the subset of g, contains both R_{head} and R_{tail} and other possible records, as shown in (3).

$$g'=\{R_{head},R_3,...,R_{tail}\} . \tag{3}$$

Table 1 lists notions of all symbols used in the system.

▼Table 1. Symbols in the system

Symbol	Notion
W	Webpage
dom	A domain name
dom_{head}	The domain name of W
dom_{tail}	The last domain name in W
R	A record of a domain name resolution
R_{head}	The record of dom_{head}
R_{tail}	The record of dom_{tail}
T	The FWT interval
p	Probability that R exists in DNS resolution data
t	Time of an event
G	All records of one webpage
G_W	All records of W
g	A set comprised of R of one browse of W
g'	The subset of g
n	The number of g in DNS
k	The number of g' in DNS

DNS: domain name system FWT: First Webpage Time

3 Proposed FWT Algorithm

This section describes the proposed FWT algorithm in detail. Firstly, we give the definition of the FWT and the method to compute it with DNS resolution data. Then, the novel FWT algorithm is introduced. Finally, the proposed algorithm is analyzed in theory.

3.1 Definition of FWT

As mentioned in the Introduction Section, a KQI is required to measure the web service. We define a new KQI called FWT (Definition 1). It reflects the waiting time before a user could see the textual contents of the requested webpage after entering dom_{head}. One of the main advantages of FWT is that it is easily computed with the DNS records, which is introduced in the next subsection. Therefore, no extra devices are needed to be deployed in the network. The QoE data of all users in the region are obtained since the DNS server covers the whole region. Moreover, they are real QoE rather than simulation values. Besides, the process of gathering QoE does not interfere with the use of the Internet service of any user.

Definition 1 (FWT): An FWT is the loading time of the textual contents of a specified webpage, starting with the time that a user enters the domain name of it and ending with the time that the browser displays its contextual contents.

In short, using FWT to evaluate web service enjoys the merits of the coverage of all users, the real QoE, no deployment and maintenance of new devices, and no interference to users.

3.2 Computation of FWT

Firstly, we introduce the process of browsing a webpage in brief, as shown in **Fig. 1**. To get W, a user has to resolve dom_{head} from the DNS in advance. Here, suppose the time of

sending dom_{head} at the user side is t_1. At time t_{head} at the DNS side, the DNS replies with the IP address of the web server, denoted by IP_{head}, to the user. Therefore, the DNS keeps a record of R_{head}. After receiving IP_{head}, the user sends the request of getting W to the web server and then receives one or several packets comprising W. Next, the browser parses W. As soon as the dom_{tail} is met, the user sends the request of resolving it to the DNS at t_2. The DNS responds with IP_{tail} at time t_{tail} to the user. Meanwhile, the other 4-tuple R_{tail} is recorded. Considering that an ordinary webpage of HTML is in the magnitude of KB and that the maximum size of an Ethernet frame is about 1.5 KB, it is very probable that dom_{tail} is contained in the last frame that is transmitted to the user.

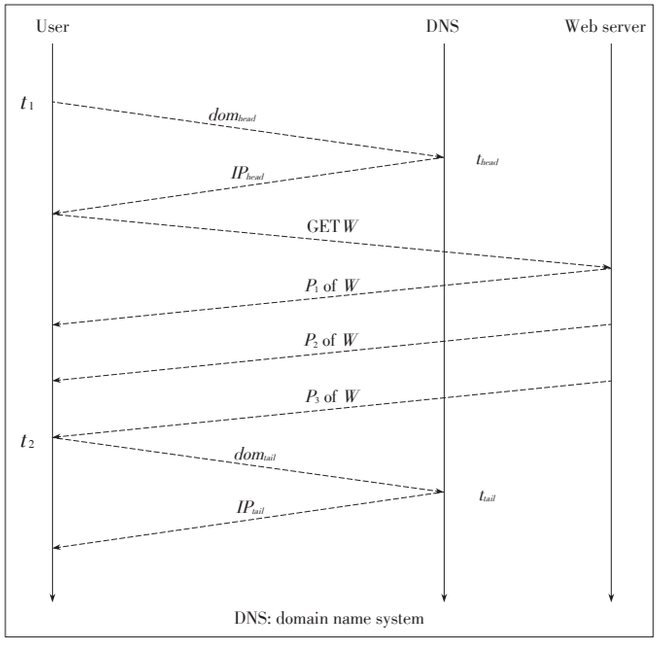
According to Definition 1, FWT equals the difference of t_2 and t_1 obviously at the user side:

$$T = t_2 - t_1 . \tag{4}$$

Considering the counteractive transmission time of dom_{head} with IP_{head} and dom_{tail} with IP_{tail}, FWT very approximately equals the difference of t_{tail} and t_{head} at the DNS side, as shown in (5). Although the transmission time is not exactly the same, the deviation would be neutralized effectively with averaging a large number of FWTs for the same webpage. In the proposed algorithm introduced in the next subsection, FWT is actually computed by averaging a large number of values. On the other hand, FWT from (5) is very useful to compare the web service for users that have different network conditions like the access bandwidth and the access mode.

$$T' = t_{tail} - t_{head} . \tag{5}$$

One of the key points to compute FWT is the selection of the dom_{tail}. It is widely seen that the webpage contains embedded contents at its bottom part and the domain names in their hyperlinks are totally new, such as an advertisement from other corporations, the contact information, and the privacy policy. All of them are options for dom_{tail}.



▲Figure 1. The process for a user to browse a webpage.

3.3 FWT Algorithm

Now, we introduce the proposed FWT algorithm in detail. After a specified period (e.g., a day), the DNS keeps millions or billions of records in chronological order comprised of domain name resolutions of all users in a region (e.g., a city or a province). These 4-tuple records in DNS are firstly classified by webpages into different groups, denoted by G. Therefore, to compute the FWT of W, the group G_W is selected. Then, G_W is partitioned into sets, each of which refers to one browse of W, as shown in (6). The partition is conducted according the resolved time t and the IP address of the user IP_u. The records belonging to an individual user are arranged in the time order. Then, R_{head} and R_{tail} are used to specify the starting position and ending position of g, respectively. If R_{head} or R_{tail} is missed in g, an adjacent record in time order will be used.

$$G_W = \{g_1, g_2, ..., g_n\} . \tag{6}$$

Let k denote the number of sets in G_W containing both dom_{head} and dom_{tail}. For the sake of clarity, we suppose they are denoted by $g'_1, g'_2, ..., g'_k$. According to (5), FWT is able to be computed by an individual g'. Therefore, there are k FWTs, denoted by $T'_1, T'_2, ..., T'_k$, respectively. The final FWT for W is the average of summarizing T'_i, as shown in (7). Although dom_{head} or dom_{tail} may be missed in some g, the value of k is big enough to ensure the precision of the averaged FWT, which is introduced in the next subsection.

$$FWT = \frac{\sum_{i=1}^{k} T'_i}{k} . \tag{7}$$

3.4 Discussion

In this part, we show the total number of sets that can be used to compute FWT, namely, the expected value of k.

Suppose the probability that a specified domain name dom is not cached locally is p when dom is used to get contents of the corresponding webpage. Equally, the probability that the DNS has the record of dom is p for this browse of W. Let X denote the random variable of R_{head}. Thus, the frequency function of X is

$$p(x) = \begin{cases} p, & \text{if } x = 1 \\ 1 - p, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases} . \tag{8}$$

Likewise, let Y denote the random variable of R_{tail}. Thus, Y has the same frequency function.

In one browse of W, both records R_{head} and R_{tail} are kept by the DNS with the probability of p^2 apparently. In other words, the probability that g contains both R_{head} and R_{tail} is

p^2. Thus, the probability that g contains R_{head} and R_{tail} can be represented by the joint frequency function of X and Y, $p(x,y)$, as shown below.

$$p(x,y)=\begin{cases}p^2, & \text{if } x=1 \text{ and } y=1\\ 1-p^2, & \text{otherwise}\end{cases}. \quad (9)$$

Let Z be the random variable representing that there are k sets containing both R_{head} and R_{tail} among n sets in total. As mentioned before, the k sets are denoted by $g'_1, g'_2, \ldots, g'_k$. Note that Z satisfies the Binomial Distribution. Therefore, the probability that there are k sets containing both R_{head} and R_{tail} among n sets in total is

$$P(Z=k)=\binom{n}{k}(p^2)^k(1-p^2)^{n-k}. \quad (10)$$

According to the probability theory, the expected value of Z is np^2, as shown in (11). That is the weighted average value of k.

$$E(Z)=np^2. \quad (11)$$

For a popular webpage, n could be in the magnitude of millions or billions. If there is one time of resolving a domain name among ten times of using the same domain name in average, k is at least in the magnitude of thousands. That means there are enough values to ensure the precision of the averaged FWT.

4 Experiments

This section introduces the experimental results including the selection of dom_{tail} for some popular webpages, the proportion of R_{head} and R_{tail} in browsing these webpages, and the computation of FWT.

The DNS server in the experiment covers a city and collects billions of domain name resolution each day. The total number of records in the experiment is about 10 billion. These records contain the domain name resolution query and the domain name resolution response. That means there are two records for each domain name request from users. The records of the domain name resolution response are used as input data to the experiment since a user may launch multiple queries for the same domain name resolution in order to speed up the process or avoid resolution failure due to loss of packets. A response record contains the fields of the source IP address and port, the destination IP address and port, the identity, the domain name, the resolution type, the resolved IP addresses and the resolved time. As mentioned in Section 2, a webpage contains many domain names besides dom_{head} and dom_{tail}. As a result, these domain names including dom_{head} and dom_{tail} are used to compute the browse times for a webpage.

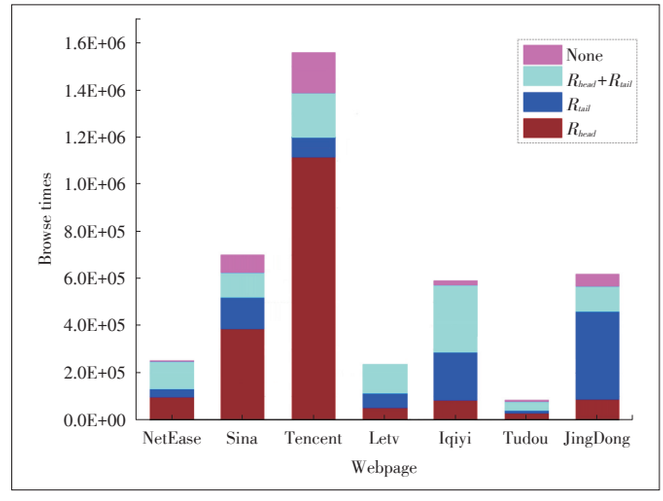
Table 2 shows the selection of dom_{tail} for some popular webpages. The domain names locate at the bottom of each webpage, and do not appear before. Since ICPs would like to gather the QoE of their own webpages, some script files like javascript are inserted at the end of the webpage. It is probable that the dom_{tail} of the NetEase, the Tencent and the Iqiyi is for this purpose. In short, dom_{tail} in Table 2 are selected by parsing the HTML code of these webpages.

Fig. 2 shows the browse times for the selected webpages. The R_{head} in the legend refers to that in one browse of the webpage dom_{head} and other domain names excluding dom_{tail} are recorded by the DNS server. Each webpage is browsed several hundred thousand at least. The Iqiyi is the most popular webpage among the video webpages and the Tencent is the most popular webpage in the news category. It can be seen that almost all webpages have at least one hundred thousand browses generating both R_{head} and R_{tail}. Hence, there are enough data for computing FWT to ensure the precision of the result, which is consistent with the theoretical analysis in this paper. The browses of the Tencent generating only R_{head} are great larger than others. The reason may be that the software of QQ that is a very popular instant messenger in China causes the resolution of www.qq.com deliberately.

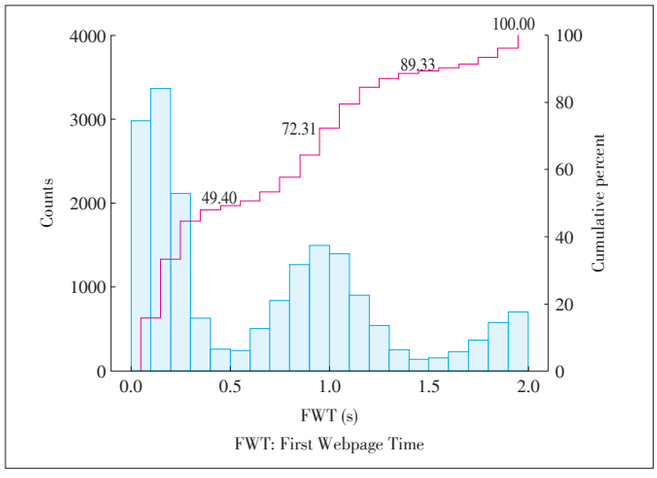
Fig. 3 shows the FWT results computed according to (7). These results are classified into different ranges by the step of 0.05 s. The values less than 0.5 s account for half of the whole results. Moreover, these values satisfy the normal distribution.

▼**Table 2. Selection of dom_{tail}**

Webpage	dom_{head}	dom_{tail}
NetEase	www.163.com	analytics.163.com
Sina	www.sina.com.cn	sina.wrating.com
Tencent	www.qq.com	qos.report.qq.com
Letv	www.letv.com	ark.letv.com
Iqiyi	www.iqiyi.com	b.scorecardresearch.com
Tudou	www.tudou.com	js.tudouui.com
Jingdong	www.jd.com	static.360buyimg.com



▲**Figure 2. Browse times of each webpage.**



▲**Figure 3. FWT distribution.**

Hence, it is better to compute the final FWT according to (7) with these values. On the other hand, the values between 0.5 s and 1.5 s accounts for 40 percent of the whole results and satisfy another normal distribution. With these values, we can get another averaged FWT. It is known that the access network is one of the most probable factors that affect the FWT. The access bandwidths of users vary due to different contracts. Therefore, FWTs of users belonging to the group with the same access bandwidth are close. As a result, when the averaged FWT is computed, we suggest considering these impact factors like the access bandwidth. The values greater than 2 s are excluded for clarity and they probably are outliers.

5 Conclusions

Mining data from the existing network systems to measure the web service has become one of the research hot spots recently. In this paper, by mining the DNS resolution data, we propose the novel FWT algorithm in order to measure QoE of the web service. The proposed algorithm is analyzed in theory that the number of FWTs of a webpage by mining DNS resolution data is np^2 in average. Therefore, there are enough FWTs to be averaged to ensure its precision. We also propose the FWT as a new KQI for the web service. The method of computing the FWT with DNS resolution data is introduced and its correctness is discussed. Finally, experiments based on the ISP's DNS resolution data are carried out in aspects of the proportion of valid DNS data, the distribution of FWT values and the computation of FWT.

References

[1] *Definition of Quality of Experience (QoE)*, ITU-T P.10/G.100, Apr. 2007.
[2] Y. Zaki, J. Chen, T. Pötsch, T. Ahmad, and L. Subramanian, "Dissecting web latency in Ghana," in *Proc. Internet Measurement Conference (IMC'14)*, Vancouver, Canada, 2014, pp. 241–248.
[3] L. Xu, Y. Luan, X. Cheng, et al., "WCDMA data based LTE site selection scheme in LTE deployment," in *Proc. the International Congress on Signal and Information Processing, Networking and Computers (ICSINC)*, Beijing, China, Oct. 2015, pp. 249–260.
[4] L. Xu, Y. Chen, K. K. Chai, J. Schormans, and L. Cuthbert, "Selforganising cluster-based cooperative load balancing in OFDMA cellular networks," *Wiley Wireless Communications and Mobile Computing*, vol. 15, no. 7, pp. 1171–1187, Jul. 2015.
[5] Q. Wang, J. Shao, F. Deng, et al., "An online monitoring approach for web service requirements," *IEEE Transactions on Services Computing*, vol. 2, no. 4, pp. 338–351, Aug. 2009. doi: 10.1109/TSC.2009.22.
[6] A. M. Kara, H. Binsalleeh, M. Mannan, A. Youssef, and M. Debbabi, "Detection of malicious payload distribution channels in DNS," in *Proc. IEEE International Conference on Communications (ICC)*, Sydney, Australia, Jun. 2014, pp. 853–858.
[7] S. Hao, H. Wang, A. Stavrou, and E. Smirni, "On the DNS deployment of modern web services," in *Proc. IEEE International Conference on Network Protocols (ICNP'23)*, San Francisco, USA, Nov. 2015, pp. 100–110.
[8] R. Keralapura, A. Nucci, Z. L. Zhang, and L. Gao, "Profiling users in a 3G network using hourglass co-clustering," in *Proc. the Sixteenth Annual International Conference on Mobile Computing and Networking (MobiCom'10)*, Chicago, Illinois, USA, 2010, pp. 341–352.
[9] D. Giordano, S. Traverso, and M. Mellia, "Exploring browsing habits of internauts: A measurement perspective," in *Proc. Asian Internet Engineering Conference (AINTEC'15)*, Bangkok, Thailand, 2015, pp. 54–61.
[10] H. Hours, E. Biersack, P. Loiseau, A. Finamore, and M. Mellia, "A study of the impact of DNS resolvers on performance using a causal approach," in *Proc. the Twenty-seventh International Teletraffic Congress (ITC'27)*, Ghent, Belgium, Sep. 2015, pp. 10–18.
[11] B. Jang, D. Lee, K. Chon, and H. Kim, "DNS resolution with renewal using piggyback," *Journal of Communications and Networks*, vol. 11, no. 4, pp. 416–427, Aug. 2009.
[12] X. Ma, J. Li, J. Tao, and X. Guan, "Towards active measurement for DNS query behavior of botnets," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Anaheim, USA, Dec. 2012, pp. 845–849.
[13] H. Gao, V. Yegneswaran, J. Jiang, et al., "Reexamining DNS from a global recursive resolver perspective," *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 43–57, Oct. 2016. doi: 10.1145/2829988.2789996.
[14] R. van Rijswijk-Deij, M. Jonker, A. Sperotto, and A. Pras, "The Internet of names: A DNS big dataset," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 91–92, Aug. 2015. doi: 10.1145/2785956.2789996.

Manuscript received: 2017-09-01

Biographies

LIU Yongsheng (liuys170@chinaunicom.cn), Ph.D., is an engineer in the Network Technology Research Institute, China United Network Communications Corporation Limited (China Unicom), China. His research interests include big data and data mining for network, Artificial Intelligence, IP & bearer technology, and so on. He is also a member of China Communications Standards Association (CCSA) and expert of International Telecommunication Union-Telecommunication Sector (ITU-T).

GU Yu (yugu.bruce@gmail.com) received the B.Eng and Ph.D. degrees from the University of Science and Technology of China in 2004 and 2010. From February 2006 to August 2006, he interned at the Wireless Network Group, Microsoft Research Asia, Beijing. From 2007 to 2008, he was a visiting scholar in the Department of Computer Science, University of Tsukuba, Japan. From 2010 to 2012, he was a JSPS Research Fellow in the National Institute of Informatics, Japan. He is now a professor in the School of Computer and Information, Hefei University of Technology, China. His research interests include information science, pervasive computing, and wireless networking, especially wireless sensor network.

WEN Xiangjiang (wenxj9@chinaunicom.cn) is a senior engineer and works with the Network Technology Research Institute of China Unicom. His research interests include Network modeling and IP & Bearer Technology.

FU Yufei (fuyf2@lenovo.com) achieved her master degree in telecommunication system management from Northeastern University in Boston, USA. She worked with Nokia and Microsoft from 2014 to 2015, and mainly focused on wireless network protocol. Now she works with Lenovo, and mainly focuses on tablet development. Her research interests include configuration management, product development management, and telecommunication network protocol.