논문 14-39C-01-12

한국통신학회논문지(J-KICS) '14-01 Vol.39C No.01
http://dx.doi.org/10.7840/kics.2014.39C.1.106

Link Dimensioning for Multiplexed M2M Traffic

Hoon Lee◆

ABSTRACT

This work proposes an analytic framework to estimate the capacity of the link in an M2M gateway that multiplexes a massive number of machine type communication devices. First, we specify the network architecture for M2M communications. Next, we summarize the characteristics of M2M traffic, from which we extract an area about the problem that has to be resolved in the dimensioning of the link capacity. After that we propose a new method to determine the capacity of the link in an M2M gateway. Via numerical experiment reflecting the realistic environment of M2M traffic, we illustrate the validity of the proposed model.

Key Words : M2M communication, M2M gateway, M2M traffic, QoS, Link dimensioning

Ⅰ. Introduction

In the near future M2M (Machine-to-Machine) communication will become a popular service in the Internet[1]. The basic step for the provision of a new service in Internet is to design the capacity of the network resource.

A method to determine the network resource of the access point for ordinary Internet services in 3GPP network is relatively well-established, where mobile traffic is characterized by typical subscriber behavior such as call holding time for voice and volume of data for data applications.

As we shall show in Section II, it is known that M2M traffic has a specific pattern compared to conventional H2H (Human-to-Human) traffic: M2M devices (MD) generate traffic in bursts with very long interval of data transmission[2]. Likewise, call holding time has no meaning in M2M communications. Also, M2M devices generate small amount of data, so that one may argue that the volume of traffic is not significant at the initial stage of M2M communication.

However, as we have argued in [3], it is not clear whether conventional network can withstand a large-scale M2M communication. Therefore, there must be a way to estimate the bandwidth for the emerging M2M service.

Nevertheless, we could find no work that seriously handles this problem, yet. Even more, we also argue that new methods for modeling the traffic and dimensioning the link capacity are needed when M2M services are added to the ordinary data application (ODA).

Based on this finding, this work proposes a new theoretical framework to determine the network resource of the access point (AP) that is needed to support the M2M applications.

Many works have been reported to determine the capacity of network resources for the conventional IP and wireless networks that take into account the guarantee of quality of service (QoS) to the customers[4]. As we shall show later in Table 1, packet delay and loss is not relevant to almost all M2M applications except emergency alert, rather guarantee of bandwidth is more relevant.

Note that QoS can be defined in mean value or probability. This work focuses on the statistical

※ The author wishes to thank the anonymous reviewers for their constructive comments on this work.

※ This work is supported by Changwon National University 2012-2014 research project.

◆ First Author : Dept. of Information & Communications Engineering, Changwon National University, hoony@cwnu.ac.kr, 종신회원

논문번호 : KICS2013-10-462, 접수일자 : 2013년 10월 24일, 심사일자 : 2013년 12월 10일, 최종논문접수일자 : 2014년 1월 6일

approach, because this is much realistic in Internet traffic[5].

Based upon these two facts in the current state of the art in network design, this work proposes an analytic framework for the dimensioning of link capacity in the network that accommodates M2M traffic.

This work is composed as follows: In Section II, we present attributes of M2M communications. In Section III, we describe a method to classify the QoS class for M2M application. In Section IV, we present a traffic model for M2M traffic. In Section V, we present a method to dimension the capacity of the link. In Section VI, we show the result of the numerical experiment and discuss the implication of this work. Finally, in Section VII, we summarize this work.

Ⅱ. Attributes of M2M communications

Usually M2M network is composed of three parts: area network, access network, and core network.

M2M area network is a small-scale network that accommodates MDs and supports direct communication between MDs and provides an access to the outside network. M2M area network is implemented by either wired or wireless technology. Ethernet and power-line are used for wired networking and WiFi, Zigbee, and Bluetooth are used for wireless networking.

Core network is a conventional IP network operated by ISP (Internet Service Provider) that provides users (humans and machines) with Internet service as well as a connection between the M2M area networks and M2M servers.

M2M area network and core network is connected by an access network (alias, distribution network), the connection of which is realized by using either wired or wireless network. DSL (Digital Subscriber Line) or optical networks are used for wired connection and 3G/4G networks are used for wireless connection. The main trend is a 4G LTE (Long Term Evolution) network. At the entrance of the LTE network an eNodeB (eNB) is located and it accommodates the traffic between the M2M area networks and the core network.

2.1 Architecture for M2M communication

Data transmission in M2M communication usually has three types such as direct, multi-hop, and peer-to-peer transmission depending on the way of transmission between MDs[6].

When MDs transmit data or access to servers (servers are usually located at core network) via connecting directly to an eNB that is located at the edge of an M2M access network, it is called a direct transmission (DT).

However, when a large number of MDs try to access an eNB, the link between MDs and eNB may be congested due to the large number of flows. This is not favorable for the reliable provision of M2M communications.

In a peer-to-peer transmission (P2PT) between MDs, there are two ways of data exchange: When MDs at the same location exchange data, it is transmitted directly. When MDs access to a server, it has to go to eNB. So, P2PT has the same congestion problem of DT when the number of MD increases.

In a multi-hop transmission (MHT), MDs transmit data or access to eNB via M2M-GW (gateway), in which case M2M-GW acts like an AP to the eNB. It is usual that MDs at the same location have the same features about traffic volume, QoS requirements, and transmission path to the server. Therefore, it is favorable that they are grouped together for control, management or charging of data[6].

From the above discussion, it is envisioned that MHT is most favorable for M2M communications, and M2M-GW is the bottleneck for the performance of M2M communication. From now on we focus on the MHT architecture.

2.2 Attributes of M2M traffic

Lee argued that there are seven typical characteristics of M2M applications[3], which is summarized in Table 1 (For a detailed discussion, see [3]):

Table 1. Characteristics of M2M traffic

Characteristics	Problems / resolutions
Massive number of devices (up to several millions)	Heavy processing overhead / Aggregated handling of data
Small-sized data (a few hundred bytes)	Heavy processing overhead and low resource utilization / Aggregated handling of data
Asymmetric traffic volume (mostly from device to network)	Unbalance in traffic volume / Accurate estimation of up & down link capacity
Infrequent & bursty transmission (inter-session time up to a few hours)	Difficulty in exact traffic modeling / Aggregated traffic modeling from each cluster
Time flexibility (most of messages are time-controllable)	Extreme difference in time requirements / Differentiated handling of time-sensitive and time-tolerant data
Group-based access (almost the same location)	Group diverse types of devices / Sophisticated mapping and translation of traffic types and QoS
Diverse QoS requirements (from emergent to time-agnostic)	Difficulty in guaranteeing every QoS requirement / Standard-aware classification of QoS

As one can see from Table 1, one has to take into account at least the following two points when modeling M2M traffic: First, packets from diverse applications at the same location have to be classified into different QoS classes. Next, many flows with small-sized packets with the same QoS class as well as location have to be aggregated into a group flow.

In the next two sections we will discuss the QoS classification of M2M traffic and multiplexing of packets.

Ⅲ. QoS class of M2M traffic

It is known that M2M applications require different level of QoS. The conventional multimedia services are broadly classified into two categories[7]: resilient and non-resilient. Resilient services are flexible in QoS requirements, examples of which are low-fidelity video and data services[8]. Non-resilient services have hard QoS requirements, example of which is data for emergency alert.

In this work we assume that bandwidth is the main measure for the QoS, so that M2M applications can be classified into GB (Guaranteed Bandwidth) or NGB (Non-Guaranteed Bandwidth), where (resilient,nonresilient) traffic is mapped into (NGB,GB) class.

For M2M applications with NGB, there is no necessity to guarantee bandwidth. The M2M traffic from NGB class uses the bandwidth that is not used by ODAs. Nevertheless, it is necessary to prepare a certain amount of bandwidth for NGB class, otherwise the performance of M2M traffic will deteriorate due to dynamic interference of ODAs' traffic.

Ⅳ. Traffic model for multiplexed M2M flows

Most M2M devices are connected to the network via an AP, in which case the bottleneck is the M2M-GW where multiplexing is carried out.

In [9], Suznjevic et al. have described multiplexing policies for the TCMTF (Tunneling Compressed Multiplexed Traffic Flows), where it is argued that M2M traffic have specific properties suitable for TCMTF, because of the following three reasons: First, packets generated by M2M devices have low payload to header size ratio. Second, acceptable delays for those services are very loose, which can go up to an hour. Finally, packet transmission occurs in one-way (from devices to servers). Note that all those properties are included in Table 1.

In order to quantize the traffic volume, let us assume as follows: The M2M devices connected to a multiplexer are homogeneous in traffic characteristics, the number of which is N. The probability of packet generation from each device at an observation period follows a Bernoulli process with probability p.

For a multiplexer accommodating N homogeneous sources with Bernoulli arrival, the multiplexed packet arrival constitutes a binomial process X, which is denoted by X~B(N,p), that has mean and variance (Np,Npq), where q=1-p[10].

When one of the following conditions are satisfied, one can approximate the binomial process X~B(N,p) into a normal distribution X~N(Np,Npq)[10]:

C1: Min(Np,Nq) > 10
C2: 0.1<p<0.9 and Npq>5
C3: Npq >25

It is not likely that M2M traffic satisfies C2, because p is very small for the most of M2M applications. Let us investigate the conditions C1 and C3. If N=3,000 and p=0.01, then Min(Np,Nq)=30 and Npq=29.7. Therefore, the above approximation is strong, and it can be applied to the multiplexed source of the M2M traffic.

Now let us define the traffic for multiplexed M2M applications. The total input to the M2M-GW is defined by a normal distribution with (mean,variance)= $(\lambda_T, \sigma_T^2) = (Np, Npq)$

Let us assume that the ratio of GB and NGB class in the multiplexed M2M traffic to the M2M-GW is α:1-α.

Ⅴ. Link Dimensioning

Now let us present a method to determine the capacity of the link for the AP of M2M area network (M2M-GW) that takes into account the characteristics of the M2M traffic with GB and NGB classes. First, let us summarize the state of the art in link dimensioning. After that, let us propose a new method to determine the link capacity of the AP that is designed to support M2M devices with robustness about the QoS.

A. State of the art in link dimensioning

There exist three steps for the determination of the link capacity, which is given as follows[4]: First, the characteristics of the traffic are defined, which can be represented by mean or higher moments for the traffic volume. Second, the desired degree of robustness in QoS is defined, which specifies the margin for the bandwidth. Finally, the capacity of the link is determined, which takes into account the former two specifications.

Now let us present a method to define the desired degree of robustness and introduce a few approaches for link dimensioning in IP network, from which we propose an appropriate method to apply in M2M traffic.

There exist three typical algorithms for link dimensioning: non-robust, strictly-robust, and statistically-robust[11], details of which are described below.

(1) Non-robust algorithm

Non-robust algorithm (NRA) does not take into account variation of the traffic volume in the network. Link capacity for NRA is usually based on the mean traffic demand. For example, when the estimated mean traffic rate is normally distributed with μ and variance σ^2, NRA determines the link capacity, C_{NRA}, in the following manner:

$$C_{NRA} = \mu \tag{1}$$

As we can see from (1), NRA does not guarantee the link speed other than the mean rate. In this case, the probability that the offered load exceeds the link capacity is 0.5. Therefore, NRA is not robust against the variation of the offered load.

(2) Strictly robust algorithm

Strictly robust algorithm (ScRA) guarantees the maximum traffic volume to the flows. For example, when the peak value of the estimated traffic rate is χ, then ScRA determine the link capacity of χ. Therefore, ScRA is robust against the worst case traffic variation. Instead, it requires extremely high capacity to the network especially when the variance is high.

(3) Statistically robust algorithm

Statistically robust algorithm (StRA) offers a statistical guarantee of the link rate to the flow by allocating the bandwidth in a statistical way, via which it achieves a tradeoff between the economy and guarantee of QoS. The main idea of StRA is to prepare a margin to the link capacity of NRA. When the estimated mean traffic rate is μ, StRA determines that the link capacity is (μ+ν), where ν is the margin and it is determined by the policy of ISP. StRA is presented as follows:

$$P(C_{StRA} > \mu+\nu) \leq \epsilon \tag{2}$$

where C_{StRA} is the capacity of the link for StRA and ε is the target value for the robustness. It is usual that μ and ε are given as conditions, from which the margin ν is determined. In the following discussion, we will propose a method to determine ν.

B. Link dimensioning for M2M traffic

When applications from MDs are combined with the ODAs such as VoIP and multimedia applications over the wireless cellular network, the problem of resource management for the high-density users in high-load network becomes a very important issue [2].

It is recommended that a certain amount of the radio channel has to be prepared for the MDs, otherwise they are overwhelmed by ODAs or ODAs are interfered by MDs.

It is usual that the amount of resource is determined by taking into account two factors: the volume of the traffic and the required level of QoS.

The volume of the traffic is assumed to be normally distributed with parameters $(\lambda_{GB},\sigma^2_{GB})$ and $(\lambda_{NGB},\sigma^2_{NGB})$, which is the mean and variance of the GB and NGB traffic, respectively. Then, the mean and variance of the total traffic are defined as follows:

$$\begin{aligned}\lambda_T &= \lambda_{GB}+\lambda_{NGB}\\ \sigma^2_T &= \sigma^2_{GB}+\sigma^2_{NGB}\end{aligned} \tag{3}$$

Now let us describe a way to provide bandwidth for each class of traffic. For an NGB class, we prepare bandwidth based on the mean rate, which is equal to λ_T. For a GB class, we prepare bandwidth in the statistical way, which is described as follows: Let U be the arrival rate of traffic from GB class to the M2M-GW, which has mean and variance $(\lambda_{GB},\sigma^2_{GB})$. Let Y be the capacity of an output link at M2M-GW, which is defined by the statistical value of the traffic arrival rate, which is given as follows:

$$Y=\lambda_{GB}+k\sigma_{GB} \tag{4}$$

where *k* is determined by the requirement of the robustness in QoS.

For the above-mentioned arrival process and the requirement for the server, we have the following formula for the statistical QoS:

$$P(U\geq Y) \leq d \tag{5}$$

where d is the target probability.

We can rewrite the formula (5) as follows:

$$P(U\geq \lambda_{GB}+k\sigma_{GB}) \leq d \tag{6}$$

In (6), higher k incurs higher requirement for the bandwidth to the M2M-GW, whereas higher d requires smaller bandwidth to the M2M-GW, which is a trade-off problem. Therefore, k has to be controlled to a certain degree, which is affected by an upper bound d.

In order to simplify the presentation, let us normalize the normal distribution into a standard normal distribution such that $N(\lambda_{GB},\sigma^2_{GB}) \rightarrow N(0,1)$.

It is known that, for a random variable Z with N(0,1), there exists a Williams approximation for the probability that Z is in the range [-k,k], which is defined as follows[10]:

$$P(-k\leq Z\leq k)=\sqrt{1-e^{(-\frac{2k^2}{\pi})}} \tag{7}$$

In (7), Z can be regarded as the normalized value for the volume of the traffic at the observation period.

Let us rewrite the right side of the equation (7) as follows:

$$\zeta = \sqrt{1-e^{(-\frac{2k^2}{\pi})}} \qquad (8)$$

Note that the following equation holds for equations (6) and (7):

$$d = \frac{1-\zeta}{2} \qquad (9)$$

Let k$_\zeta$ be a k that Z has a cumulative 100ζ% value in the distribution, then, the following formula for k_ζ holds:

$$k_\zeta \geq \sqrt{-\frac{\pi}{2}log_e(1-\zeta^2)} \qquad (10)$$

Then, we can obtain the following formula for the required capacity of the link that satisfies (5):

$$C \geq \lambda_{GB} + \sigma_{GB}\sqrt{-\frac{\pi}{2}log_e(1-(1-2d)^2)} \qquad (11)$$

where C represents the lower bound for Y (minimum value of the link capacity) that is required to satisfy the formula (5). Note that C depends on two factors: the statistics of the offered load and the degree of robustness.

Ⅵ. Numerical experiment

Now let us carry out a numerical experiment, via which we show the implication of the work. Consider the uplink of M2M applications (we can apply this framework to the downlink, too. But, it is trivial). As to the parameters associated with the M2M devices, we referred some of the parameters presented in [8], which is summarized in Table 2.

Table 2. Parameters used in numerical experiments

Parameter	Value
d (Statistical QoS target)	Used as an input parameter
N (number of M2M device)	3,000
p (probability of message generation per device per second)	0.01
Mixing ratio between GB and NGB class	0.5:0.5
Payload size per message (bytes)	200/500/1,000 (each with probability of 1/3)

We assume that the proportion of messages for the small, medium, and large payload is evenly distributed with probability 1/3. In this work we investigated the performance of the system for d varying widely from 0.1% to 50%.

In our experiment we carried out three scenario for the purpose of comparison between the conventional method (Scenario 1 and 2) and our method (Scenario 3), which is given as follows:

Scenario 1: Packets from GB and NGB classes are treated equally with NRA

Scenario 2: Packets from GB and NGB classes are treated equally with ScRA. For the input traffic with (mean, variance)=(λ_T, σ_T^2) and peak χ, where the server provides the customer with the rate χ

Scenario 3: Packets from NGB class are served by mean rate λ_{NGB}, whereas those from GB class are served by statistical rate given by (11)

Fig.1 illustrates the capacity of the link for M2M-GW. For the scenarios 1 and 2, the link capacity is constant, because they are independent from d. As one can find from the figure, scenario 1 requires the smallest bandwidth at the cost of non-robustness about the variation of the traffic volume, whereas scenario 2 requires the highest bandwidth at the expense of strict-robustness about the variation of the traffic volume. Scenario 3 is located between those two scenarios, which shows a

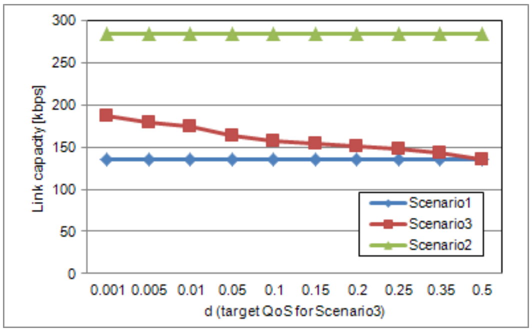

Fig. 1. Link capacity for M2M-GW

trade-off between the cost and performance.

Note from Fig.1 that the capacity of the link in the M2M-GW with scenario 3 decreases as the required value of the QoS is relaxed. When d is equal to 0.5, where the violation probability of the statistical QoS reaches 50%, the capacity of the link for the scenario 3 is the same as that of the scenario 1.

Before closing the numerical experiment, one has to note the following fact: The link capacity linearly increases with the number of device or the probability of packet generation, so that it is straightforward to compute the link capacity for higher number of device or higher value of probability. Therefore, we did not carry out further numerical evaluation for various mixing of the source traffic parameters (N,p).

Finally, we argue that the most important intuition that we can obtain from the above result is that the network operator can fine-tune the amount of bandwidth of the M2M-GW by referring to the relationship between the performance and network resource. This is the most important implication and novelty of this work.

Ⅶ. Conclusion

In this work we have proposed a practical method to estimate the link capacity of the output link for the M2M gateway in the MHT framework of M2M network. To that purpose, we investigated the generic architecture of M2M communications, attributes of M2M traffic, and a QoS class for the M2M traffic, via which we could extract a theoretic framework for the source traffic and link capacity.

There are a number of implications that we have obtained from this work, which is summarized as follows: First, as to the source model of the M2M traffic, we proposed a Gaussian statistical model, which can cover every kind of applications generated by M2M devices. Second, the proposed model is based on the 2-class statistical QoS with guaranteed and nonguaranteed bandwidth, via which one can cover two distinct attributes of M2M traffic with different QoS requirements.

From our numerical experiment, we illustrated the implication of the proposed method by showing that the link capacity determined by the proposed algorithm lies between the two extreme cases of non-robust and strictly robust algorithms.

This work can be applied to a wide spectrum of network dimensioning at the access part of the M2M network that accommodates a wide range of applications such as smart farming, smart grid, smart city, e-health, anti-disaster system, etc.

In the future, we will accumulate real-field data for the M2M traffic, via which we can fine-tune the parameters for the source traffic model.

References

[1] G. Wu et al., "M2M: From mobile to embedded internet," *IEEE Commun. Mag.*, pp. 36-43, Apr. 2011.

[2] K. Raymond, *Summary of the LOLA Project*, http://www.eurecom.fr/~nikaeinn/files/lola/LOLA_Summary_YEAR1.pdf.

[3] H. Lee, "The seven characteristics of M2M traffic: Observations and implications," *J. CWNU*, Vol. 2, pp. 341-349, Dec. 2013.

[4] R. Cahn, *Wide area network design*, Morgan Kaufmann, pp. 279-333, 1998.

[5] H. Lee and Y. Nemoto, "Providing the statistical quality of service objectives in high speed networks," *Computer Networks* (Former Computer Networks and ISDN Systems) 29(1997), pp. 1919-1931, Dec. 1997.

[6] K. Zheng et al., "Radio resource allocation in LTE-advanced cellular networks with M2M communications," *IEEE Commun. Mag.*, pp. 184-192, Jul. 2012.

[7] Y. Zhang et al., "Home M2M networks: architectures, standards, and QoS improvement," *IEEE Commun. Mag.*, pp. 44-52, Apr. 2011.

[8] D. Boswarthick et al. ed., *M2M Communications: A systems approach*, Wiley, pp. 79-83, 2012.

[9] M. Suznjevic and J. Saldana, "*Delay limits and multiplexing policies to be employed with Tunneling Compressed Multiplexed Traffic*

Flows," Jun. 14, 2013, http://tools.ietf.org/html/draft-suznjevic-tsvwg-mtd-tcmtf-01.

[10] H. Lee, *Professor Lee's lectures on probability*, HongPub, p. 102, Aug. 2009.

[11] S. Sharafeddine and Z. Dawy, "*Robust network dimensioning for realtime services over IP networks with traffic variation*," Computer Communications, vol. 33, pp. 976-983, May 2010.

Hoon Lee

1984 B. E. from Kyungpook Nat'l Univ. (KNU)

1986 M. E. from KNU

1996 Ph.D. from Tohoku University, Japan.

1986~2001 KT R&D Center

2001~Changwon Nat'l Univ.

2005~2006 Visiting@U. of Missouri-Kansas City

2011~2012 Visiting@Marquette University

<Research fields> Design and performance evaluation of Internet