A Practical Evaluation Method of Network Traffic Load for Capacity Planning

Takeshi Kitahara(✉), Shuichi Nawata, Masaki Suzuki, Norihiro Fukumoto, and Shigehiro Ano

KDDI R&D Laboratories Inc., 2-1-15 Ohara, Fujimino-shi, Saitama, Japan
kitahara@kddilabs.jp

Abstract. Communications network operators are supposed to provide high quality network service at low cost. Operators always monitor the amount of traffic and decide equipment investment when the amount exceeds a certain threshold considering trade-offs between link capacity and its utilization. To find the proper threshold efficiently, this paper proposes a practical threshold definition method which consists of fine grained data collection and computer simulation. We evaluate the proposed method using commercial traffic data-set. The results show the proper timing for the equipment investment.

Keywords: Capacity planning · Traffic monitoring · Traffic load testing · Queuing simulation

1 Introduction

The largest mission for communications network operator is to provide high-quality network service at low cost, but there always exists trade-off between quality and cost. To ensure quality of network service, it is crucial to keep traffic load of the link below a certain threshold level and to upgrade the link capacity immediately after the load exceeds the threshold. From the viewpoint of the capital expenditure, on the other hand, this threshold value should be set to high so that traffic can be accommodated into the link as much as possible. Thus finding the appropriate threshold value is the main effort for capacity planning.

Traffic load is typically observed with a monitoring tool such as MRTG [1] (Multi Router Traffic Grapher) in terms of average volume during several minutes, because watching traffic load with finer granularity requires computing resources. Considering practical use, the threshold mentioned above should be represented as the ratio of average traffic volume in minutes to the original link capacity. For example, the threshold would be 80 % if five-minute average of 800 Mbit/s is the maximum load for a certain GbE (Gigabit Ethernet) link so as not to degrade quality. However, averaging through minutes masks information about burstiness of traffic on each link. In the case of bursty traffic where the difference between the instantaneous peak and several-minute average is large, probability of packet loss would be high compared to the case of traffic where such the difference is small even though average loads in minutes are equivalent. Thus capacity for links with bursty traffic should be upgraded earlier than links with not-bursty traffic. A lot of past work addressed characterizing traffic

K. El-Fakih et al. (Eds.): ICTSS 2015, LNCS 9447, pp. 263–268, 2015.
DOI: 10.1007/978-3-319-25945-1_18

burstiness [2–6], but there are no generalized models which can be applied to individual network operating and/or planning task because network structure, network usage, and applications on network are quite diverse nowadays.

For the purpose of obtaining an appropriate threshold of upgrading capacity for each link, we propose a practical method to evaluate traffic load considering its burstiness while satisfying the target level of packet loss ratio. Furthermore, we also evaluate this method using commercial traffic data-set in an operator's network.

2 Proposed Method

There are two challenges to be addressed in this study. One is to comprehend the burstiness about the traffic on the target link and the other is to calculate the threshold for upgrading capacity considering burstiness. Thus the proposed method consists of (1) collecting data about inter-packet arrival time and packet size for several minutes on the target link and (2) evaluating the burstiness of the collected data and to calculate the threshold value using computer simulation.

(1) Data collection

Burstiness of traffic cannot be estimated from five-minute average load where operators usually observe. To show that, we select four data-sets which were observed at two different points and on two different dates. Figure 1 shows the data-sets. They are hereinafter referred to as Data 1, Data 2, Data 3 and Data 4. All data-sets describe traffic load on Link A or Link B in terms of bit/s for five minutes. Note that X-axes of all figures indicate 0 to 300 s. Each of Link A and Link B belongs to different commercial network. Traffic load during five minutes of Data 1 and Data 3 are nearly equal to that of Data 2 and Data 4, respectively. The normalized five-minute volumes of Data 1 through Data 4 are 1.003, 1.000, 1.210 and 1.223, respectively. Note that Y-axes of all figures are in the same scale. We also show three kinds of granularity of each data-set, which are one-millisecond, one-second and five-minutes. For evaluation of burstiness, the proposed method obtains packet-by-packet data. Note that this is only for deciding the threshold and observing fine grained data is not required for daily operation.

(2) Computer simulation

The simulation solves the least required link capacity under the given conditions regarding target packet loss ratio and buffer size when traffic pattern is provided as input. The input data in this study is time-series data describing a pair of inter-arrival time and length of each packet, which are extracted from Data 1 through Data 4. The size of each input data for the simulation ranges from around 70 Mbytes to 130 Mbytes. The least required capacity corresponds to the value of the threshold discussed so far. The notations of R, B, X, m and p are output rate from the queue, buffer size, input traffic to the queue, input mean rate to the queue and target packet loss ratio, respectively. In the simulation, an event will be processed each time the packet #n (n = 1,…, N) arrives at the queue. Note that X consists of N packets. We define t_n, l_n and $Q[n]$ as the time when the packet #n arrives, the length of the packet #n and the queue length at the time t_n, respectively. Figure 2 represents the relationship among the notations.

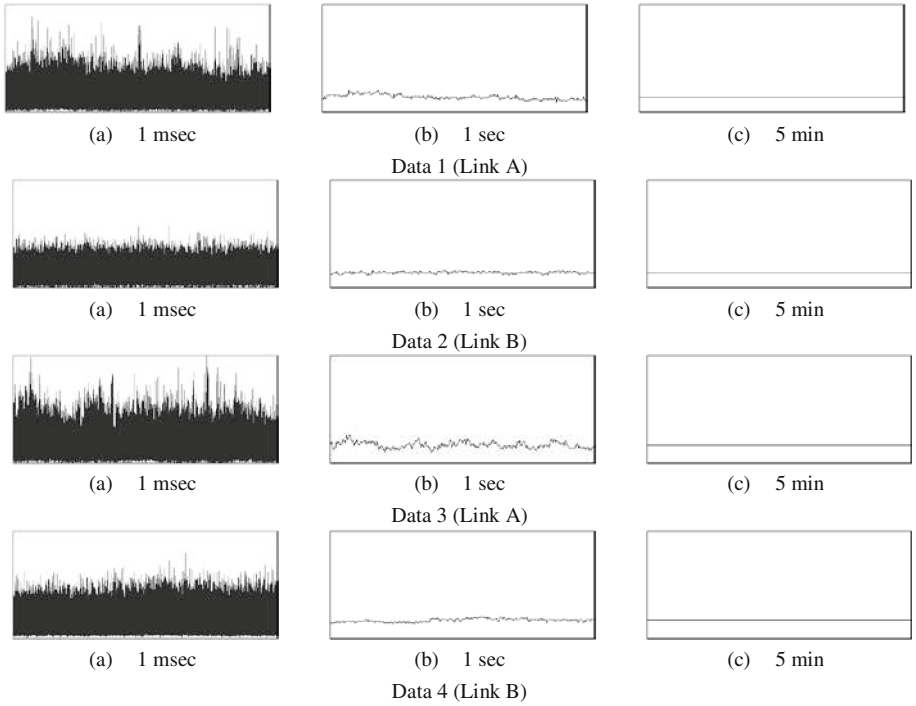


Fig. 1. Traffic volume during five minutes

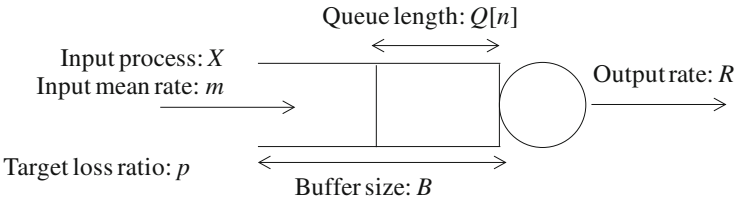


Fig. 2. Notations of the simulation.

According to Eq. (1), $Q[n]$ will be updated each time n is increased, which means a packet arrives at the queue. This process will continue until n reaches N.

$$
\begin{aligned}
Q_{tmp} &= Q[n-1] - R \cdot (t_n - t_{n-1}) \ (n = 1, 2, \ldots, N) \\
Q[n] &= \begin{cases} Q_{tmp} + l_n & (if \ Q_{tmp} + l_n < B) \\ Q_{tmp} & (otherwise) \end{cases} \\
N_{loss} &= \begin{cases} N_{loss} & (if \ Q_{tmp} + l_n < B) \\ N_{loss} + 1 & (otherwise) \end{cases}
\end{aligned} \tag{1}
$$

where N_{loss} describes the number of packet losses during the simulation. Figure 3 shows an example of queuing process. In this example, packet #n enters into the queue but packet #(n + 1) is dropped. To calculate the least required capacity meeting the packet loss ratio p, the following steps are performed. Such the capacity is denoted as C_{opt} and we employ a bisection method to find the value of C_{opt} for this study.

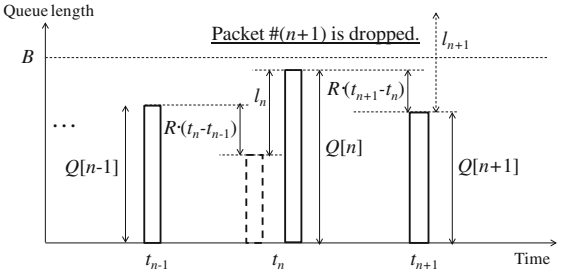


Fig. 3. Example of simulation process.

Step 1

Determine p and B based on the operation policy. Calculate the value of m of a data-set X we focus on. Determine C_{low} and C_{high} for initial values of R. We used $C_{low} = m$ and $C_{high} = 10\ m$ as the initial values, respectively. In addition, dC should be determined as the terminal condition. In this simulation, we employ 100 kbit/s as the value of dC.

Step 2

Perform the queuing simulation based on Eq. (1) in the both cases of C_{low} and C_{high}. The values of N_{loss} for both cases are obtained through the simulation. Note that the simulation will be terminated, if N_{loss} exceeds $p{\cdot}N$ during the simulation.

Step 3

Update C_{low} and C_{high} by using a bisection method [6]. If $N_{loss}/N > p$, C_{low} is set to $(C_{low} + C_{high})/2$. In the same manner, if $N_{loss}/N < p$, C_{high} is set to $(C_{low} + C_{high})/2$. Figure 4 shows the conceptual diagram of the bisection method for this simulation.

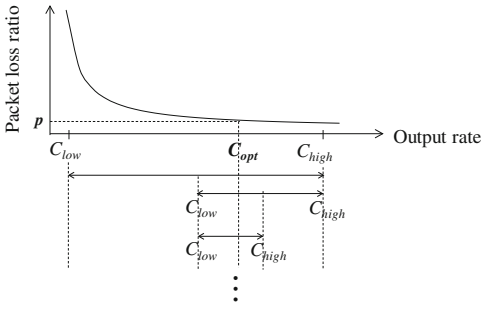


Fig. 4. Bisection method for the simulation.

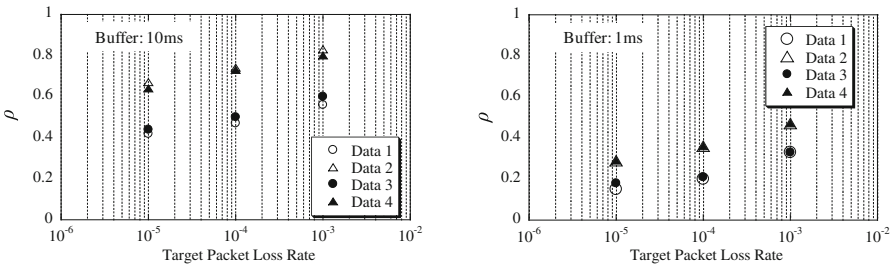
Step 4

Repeat Step 3 until $C_{high} - C_{low} < dC$. Then the value obtained is C_{opt}. We define ρ as m/C_{opt}, which is the key indicator for capacity planning. This is because the value of ρ represents maximum allowable ratio of traffic load to meet the target quality for the link.

3 Simulation Result and Discussion

Figure 5 shows the result of the simulation using Data 1 through Data 4. In the simulation, buffer size B for each simulation is set to m multiplied by 1 and 10 ms. This implies that maximum delay of the buffer is around 1 or 10 ms. We observe the significant difference regarding ρ between Data 1 and Data 2 although the values m are nearly equivalent as mentioned before. This is also applied to between Data 3 and Data 4. Since each of Data 1 and Data 3 is bursty compared to each of Data 2 and Data 4 respectively as intuitively confirmed in Fig. 1, the difference of ρ is found to depend on the burstiness of the traffic. Furthermore ρ of Data 1 is quite close to that of Data 3 and this relationship is applied to between Data 2 and Data 4. This suggests that the degree of the burstiness should be dependent on the link we observed. In other words, average volume do not have a large impact on the value of ρ. In addition, we confirm the target packet loss ratio has a considerable impact to the value of ρ. This targeting is left to a policy of the trade-off between quality and investment. To make a right decision for operators, the obtained values of ρ are quite beneficial. Furthermore, computing time required to obtain the results is important from the practical viewpoint. In this simulation, it takes about 10 to 20 min to complete a simulation to obtain one value of ρ corresponding to one point in Fig. 5.

There are not purely theoretical methods to calculate the value of ρ, because ρ is determined by quite many factors. The possible factors affecting ρ are the number of nodes where the traffic goes through, services and applications generating traffic, customers where the network is targeted for, and so on. This paper, however, does not focus on these factors because these factors themselves are not so important from the viewpoint of practical capacity planning. The largest interest for practitioners is "*when do I need to increase the capacity?*". The answer is when the observed mean rate



(a) Buffer size is 10ms (b) Buffer size is 1ms.

Fig. 5. Simulation results regarding maximum allowable ratio of traffic load ρ.

reaches original link capacity C multiplied by ρ. Since the mean rate is able to be predicted using conventional techniques of time-series analysis, the appropriate timing for enhancement of capacity can be easily estimated. If a practitioner collects some datasets on a certain link and calculates ρ corresponding to each dataset, the lowest value of ρ should be applied to the link against the unexpected excessive load while it depends on operator's policy. Note that we assume that the value of ρ of the future, which is the time when total traffic volume increases, should not be significantly changed. Figure 5 supports this assumption.

4 Conclusion

While research related to burstiness of traffic have been studied in these twenty years, network operators might still rely on their experienced knowledge. Since real traffic is quite diverse, there are no generalized models which can apply to all kinds of traffic. Thus we were eager to fill the gap between research and practice. This paper proposed a practical evaluation method for network capacity planning. We applied the proposed method to commercial traffic data which were observed at two different points and on two different dates. Since the proposed method is quite specific and concrete, network operators can easily apply this method to their work. We believe the contribution of this paper helps them to improve quality of their work.

References

1. Shipway, S.: Using MRTG with RRDtool and Routers2, Cheshire Cat Computing (2010)
2. Leland, W.E., et al.: On the self-similar nature of ethernet traffic (extended version). IEEE/ACM Trans. on Netw. **2**(1), 1–15 (1994)
3. Taqqu, M.S., et al.: Proof of a fundamental result in self-similar traffic modeling. ACM SIGCOMM Comput. Commun. Rev. **27**(2), 5–23 (1997)
4. Erramilli, A., et al.: Experimental queuing analysis with long-range dependent packet traffic. IEEE/ACM Trans. on Netw. **4**(2), 226–244 (1996)
5. Kandula, S., et al.: The nature of data center traffic: measurements & analysis. In: Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference (IMC 2009) (2009)
6. Benson, T., et al.: Understanding data center traffic characteristic. ACM SIGCOMM Comput. Commun. Rev. **40**(1), 92–99 (2010)
7. Burden, R.L., Faires, J.D.: 2.1 The bisection algorithm. In: Numerical Analysis. PWS Publishers, Englewood Cliffs (1985)