Proceedings of the 2005 International Conference on Simulation and Modeling
V. Kachitvichyanukul, U. Purintrapiban, P. Utayopas, eds.

PERFORMANCE EVALUATION OF VIDEO-ON-DEMAND SERVICE FOR E-LEARNING IN A CAMPUS NETWORK

Thanadech Thanakornworakij and Peerapon Siripongwutikorn

Department of Computer Engineering,
King Mongkut's University of Technology Thonburi
Bangkok, Thailand, 10140

ABSTRACT

This paper evaluates the possibility of VoD service deployment in the KMUTT campus network. Particularly, we are interested in the number of video flows that can be supported along the paths from the video server to end-user workstations located around the campus. Our evaluation approach is based on the Maximum Asymptotic Variance (MVA) technique, which predicts the number of admissible flows at a single node for a given delay bound requirement. We then apply a simple probabilistic bound to deal with multiple nodes. Network measurement using a Distributed Benchmark System (DBS) is performed to validate the model accuracy. In our experiments, MPEG-coded video traffic recorded from class lectures is used. Effects of delay bound requirements and the network paths are investigated. Based on the results, we found that the link capacity may have to be upgraded or some mechanism to prioritize the video traffic is needed.

1 INTRODUCTION

The King Mongkut's University of Technology Thonburi (KMUTT) is in its initial phase of launching the video-on-demand (VoD) service for E-learning to campus users. Class lectures are recorded and stored in our large video databases for later access. Satisfactory video playback quality will depend strongly on the ability of the campus network to guarantee stringent Quality-of-Service (QoS) in terms of bounded packet delay and loss. To this end, we need to determine if the network is capable of supporting a sufficient number of users (or video flows) around the campus. Particularly, we are interested in the number of video flows that can be supported along the paths from the video server to end-user workstations around the campus for a given end-to-end delay bound requirement. Additionally, we attempt to identify important factors that impacts on the network performance. These results can then be used to justify if the existing network requires the capacity upgrade in order to support the VoD service for E-learning.

Since it is inconvenient to generate a large number of video requests or changing network factors to perform actual network measurements, we have instead developed an analytical model for evaluating the network queueing delay. Our modeling approach is based on the use of Maximum Asymptotic Variance (MVA) technique to predict the maximum number of video flows that can be supported at a single node for a given delay bound requirement. Then, we extend the prediction from the single queue case to the multiple queue case. Network measurement is then performed to validate the model accuracy. With the analytical model, we predict the number of admissible flows and investigate the effects of delay bound requirements and the network paths on the network performance.

The paper is organized as follows. In Section II, related research on analytical modeling of queueing performance techniques as well as the MVA approach are reviewed. In Section III, we describe the topology of our campus network, the network factors, and characteristics of the MPEG traces under study. In Section IV, we present a simple extension to the MVA approach to deal with the multiple queue case. The model accuracy is validated by comparing to the measurement results obtained from Distributed Benchmark System (DBS). In Section V, the results from the analytical model are then reported. The conclusion is given in Section VI.

2 BACKGROUND AND RELATED WORK

In the past decade, a great deal of analytical models to evaluate the queueing performance have been developed (see e.g., [1,2,3,5,6]). They differ on how the input traffic is characterized, which in turn affects the model accuracy. Essentially, the number of packet arrivals during a time unit or the packet interarrival time is modeled by a stochastic process. Most dominant ones are Markovian, Autoregressive, and Gaussian processes. In general, simple traffic models are easy to parameterize and evaluate but inaccurate, while complex ones yield more accurate

results but analytically intractable. Among them, a traffic model based on a Gaussian process [2,6] is simple, as traffic can be represented only by the mean and the autocovariance function. Also, it has been shown to be analytically tractable, and provide most accurate results compared to other approaches, especially when the input traffic is constructed from a large number of flows. In contrast, when a large number of sources is multiplexed, traditional Markovian models results in state-space explosion, whereby the states are so large that it is computationally infeasible to obtain the results. The Gaussian model is quite accurate even if individual traffic flows are non-Gaussian and/or heterogeneous, which makes the model highly flexible. The Gaussian model allows us to obtain the buffer behavior for such various sources such as multiplexed homogeneous and heterogeneous Markov modulated sources, sources that are correlated at multiple time scales, sources whose autocorrelation function exhibits heavy (sub-exponential) tail behavior, and sources generated from real MPEG-encoded video sequences [4,7].

In [2], the tail of the steady-state queue length distribution in an infinite buffer queue with a stationary Gaussian input process is derived by using the technique called "Maximum Variance Asymptotic" (MVA). The result is extended to the case of finite buffer queue in [5], which is more difficult to evaluate compared to that presented in [2]. Since in our network, the routers have a large buffer size, it is safe to assume that the buffer size is infinite so that the result in [2] can be applied. Consider an aggregate traffic flow that consists of a large number of flows. Such an aggregate flow can be characterized by a Gaussian process with mean rate at time n equal to λ_n. Let $\overline{\lambda} := E\{\lambda_n\}$ and $\sigma^2 := Var\{\lambda_n\}$. The queue length Q_n (or workload) at time n in a single server queue with an infinite buffer can be expressed by Lindley's equation:

$$Q_n = (Q_{n-1} + \lambda_n - c)^+ \tag{1}$$

where c is the link capacity. Define a stochastic process X_n as

$$X_n := \sum_{k=1}^{n} \lambda_k - cn \tag{2}$$

We assume that λ_n is stationary and ergodic, and that the system is stable, i.e., $E\{\lambda_n\} < c$. Then, it has been shown in [4] that the distribution of Q_n converges to the steady
state distribution as $n \to \infty$ and that the supremum distribution of X_n is the steady state queue distribution:

$$P\{Q > x\} = P\left\{\sup_{n>1} X_n > x\right\} \tag{3}$$

Let $C_\lambda(l)$ be the autocovariance function of λ_n. Then, thevariance of X_n can be expressed in terms of $C_\lambda(l)$. For each x > 0, define the normalized variance $\sigma^2_{x,n}$ of X_n as

$$\sigma^2_{x,n} := \frac{Var\{X_n\}}{(x - E\{X_n\})^2} = \frac{nC_\lambda + 2\sum_{l=1}^{n-1}(n-l)C_\lambda(l)}{(x+kn)^2} \tag{4}$$

, where $\kappa := c - \overline{\lambda}$. Let m_x be the reciprocal of the maximum of $\sigma^2_{x,n}$ for a given x,

$$m_x := \frac{1}{\max_{n\geq 1} \sigma^2_{x,n}} = \min_{n\geq 1} \frac{(x+\kappa n)^2}{Var\{X_n\}} \tag{5}$$

To obtain m_x, one must determine the time n at which the normalized variance $\{Var\{X_n\}/(x+\kappa n)^2\}$ is maximized. The queue length distribution is then given by

$$\mathrm{P}(\mathrm{Q} > \mathrm{x}) = e^{-(m_x/2)} \tag{6}$$

This result is called the Maximum Variance Asymptotic (MVA) approximation for an infinite buffer.

3 NETWORK MODEL AND TRAFFIC CHARAC-TER ISTICS

Table 1: Statistical of MPEG video flows

Frame Characteristics	Flow 1	Flow 2
Mean frame size, μ (bytes)	4597.5	3733.7
Max. frame size (bytes)	12926	17954
Min. frame size (bytes)	842.8	407.6
Peak/mean rate	2.8116	4.8086
Coef. of variation, σ/μ	0.4463	0.9894

Figure 1 shows the network topology in the KMUTT main campus. The video server for the VoD service is located at the central library. From the topology, the backbone network can be divided into two parts that are connected together via a 200-Mbps link. Two gigabit routers are connected to international channels. A local network in the central library building is connected to the switch in the

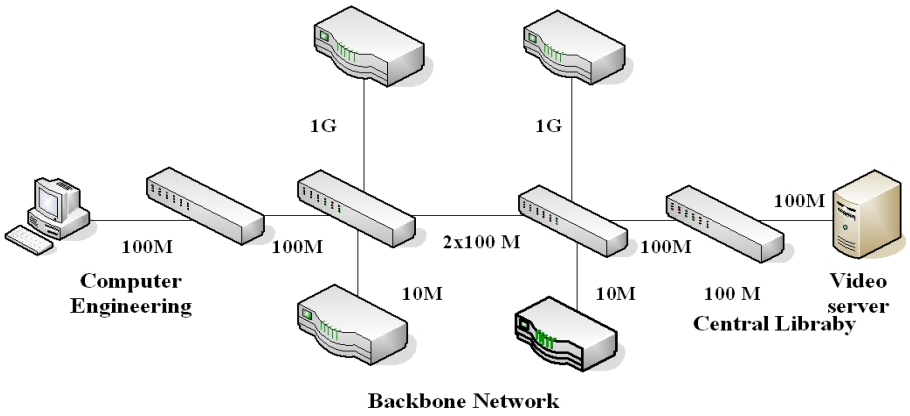


Figure 1: Topology of the Campus Network.

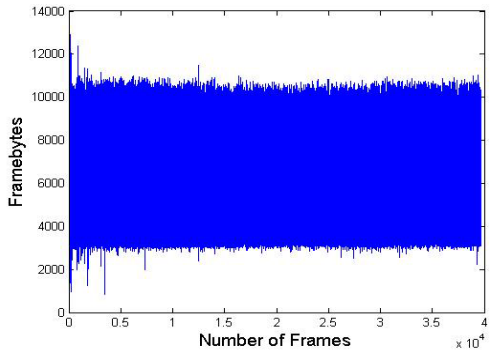


Figure 2: Time series of entire VBR video sequence (flow1)

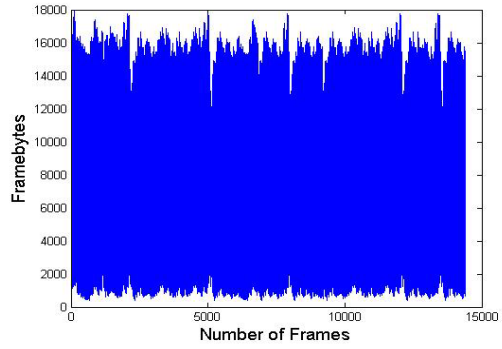


Figure 3: Time series of entire VBR video sequence (flow2)

hand side. The department of computer engineering, which is where we set up a computer for experiments, is connected to the switch in the left hand side. The path from the video server to the end host therefore has four hops. From the network topology, it was found that the longest path length from the video server to any host in the campus has four hops.

3.1 MPEG Video Traces

Figures 2 and 3 show the frame sizes of the entire MPEG-coded video traces under investigation. Those MPEG flows are sampled from a large video database of MPEG video recorded from class lectures. Most of the video flows stored in the server have more or less the same average rate but have different statistical characteristics. Several basic statistics for the traces are given in Table 1, including the mean, the minimum, and the maximum frame sizes (in bytes), and the peak to mean frame rates.

3.2 Network Factors

Many factors can affect the number of flows that can be carried by the network, including video traffic characteristics, delay bound requirement, and the network paths. The number of admitted flows reduces as the traffic becomes more bursty (higher peak-to-mean rate ratios). The delay bound requirement includes both the maximum tolerable packet delay as well as the probability of delay bound violation. The network path corresponds to the number of hops. In a longer path, less number of video flows can be carried.

4 MODELING APPROACH

A flow can be admitted to the network if the network can guarantee the end-to-end (queueing) delay bound of D seconds, with the probability of delay violation less than ε, which is set to 0.05. Write $P(Q > x)$ as $P(Q/c > x/c) = P(\text{packet delay} > x/c)$. Therefore, we have

$$P\{\text{packet delay} > D\} < \varepsilon$$

where $D = x/c$, which relates D to x. In our experiments, we specify the default delay bound D to 10 msec, and link capacity c to 100 Mbps. This results in a bound on the queue size x = 125,000 bytes. For each new flow, the input

traffic parameters (the mean rate and the autocovariance function of the aggregate flow) are updated to take into account the new flow. Then, P(Q > x) is evaluated from (4), (5) and (6). The new flow is admitted if the result is less than 0.05. This process is repeated until P(Q > x) is greater than 0.05, and we obtain the maximum number of flows that can be carried by the queue.

The MVA approach only applies for a single queue case.. To extend the MVA result to a multiple queue case, we apply a simple probabilistic bound. We will first analyze a case of 2 queues, and then generalized the result to a case of N queues. Let d_1 and d_2 be the delay experienced by packets in queue 1 and queue 2 respectively, and d be the end-to-end packet delay. Assume that d_1 and d_2 are IID. From

$$\begin{aligned} P(d < D) &= P(d_1 + d_2 < D) \\ &= \sum_{i+j=D} P(d_1 < i)P(d_2 < j) \\ &> P(d_1 < D/2)P(d_2 < D/2) \end{aligned}$$

Therefore, $P(d_1 < D/2)P(d_2 < D/2)$ is a lower bound for P(d < D). If we approximate P(d < D) with this lower bound, it follows that

$$\begin{aligned} P(d > D) &= 1 - P(d < D) \\ &\approx 1 - P(d_1 < D/2)P(d_2 < D/2) < \varepsilon \\ &= P(d_1 < D/2)P(d_2 < D/2) > 1 - \varepsilon \end{aligned}$$

Since d_1 and d_2 are IID,

$$P(d_i > D/2) < 1 - (1-\varepsilon)^{1/2}$$

In general,

$$P(d_i > D/N) < 1 - (1-\varepsilon)^{1/N} \qquad (8)$$

where D is the delay bound requirement and N is the number of queues (or hops). Consequently, a given end-to-end delay bound D and its probability of violation ε can be translated into a nodal delay requirement through (8). Then, the MVA result is applied to determine the number of flows that can be carried at each node.

4.1 Validation with Measurement

Distributed Benchmark System (DBS) is used to measure the network performance. The DBS is composed of 3 programs: DBS controller, DBS daemon, and DBS viewer. DBS controller (DBSC) controls TCP/UDP data transfer such as MPEG video flows. It reads commands from a command file and sends command instructions to DBS daemon (DBSD), which starts the actual data transfer. DBSD is a daemon program that is launched on the experimenting host. After that DBSC receives the result from DBSD and saves them into local files. DBS viewer (DBS_view) is a program that analyzes data from the local files and plots the transition of sequence numbers, throughput, and delay.

To establish the measurement experiment, we set up two hosts in the campus: the first one in the central library and the other one in the department of computer engineering. This path has four routers, which is a longest path length in the campus network. However, since the routers have to carry other best-effort TCP/IP traffic in addition to the MPEG video flows, we cannot directly determine the number of flows in the end-to-end path that can actually be supported. To work around this problem, we consider only the bottleneck router and assume that the best-effort traffic is approximately Gaussian. Then, we subtract the actual link utilization measured during the peak usage period from the link utilization obtained from the analytical model. The utilization difference is then used to determine the number of video flows to feed into the network.

We test the model with an MPEG flow generated by the Distributed Benchmark System (DBS). The statistics of such MPEG flow is as follows: maximum frame size = 40960 bytes, minimum frame size = 2,048 bytes, mean frame size = 7,338.7 bytes, peak per mean frame rate = 5.58. Let us first determine the number of flows predicted by the model. Since there are four hops on the path, we obtain from (8) the nodal delay requirement as $P(d_i > 2.5) < 0.0127$, which is equivalent to x = 31,250 bytes. After substituting all the parameters in the model, we obtain 15 flows. In conclusion, from the analytical model, with a 4-hop path, c = 100 Mbps, D = 10 msec, and ε = 0.05, the number of flows computed from (8) is 15, which results in 22% link utilization in each router. Therefore, a single flow contributes the utilization of 1.47%.

From the network measurement in the peak usage period, the highest link utilization from the four routers is 15%. Therefore, 22-15 = 7% utilization is left to carry MPEG video traffic. This 7% utilization can support 7/1.47 ≈ 4 flows. So, we use DBS to generate 4 MPEG flows with the above characteristics into the network during the peak usage period, measure the actual end-to-end delay, and then compare the results obtained from the analytical model (15 flows). The end-to-end delay obtained from DBS consists of processing delay, queueing delay, transmission delay, and propagation delay. It is reasonable to assume that all the delay components are fixed except the queueing delay. Subtracting the lowest delay value(which corresponds to packet with zero queueing delay) from the rest will result in the end-to-end queueing delay.

Figure 5 shows the end-to-end queuing delay distribution obtained from the measurement. From the figure, the proportion of end-to-end queuing delay that exceed 10 ms is 10% (0.1), which deviates from the specified value of 5% (ε = 0.05) much less than an order of

magnitude. This difference is due to the lower bound approximation of P(d < D) we use in (8). Therefore, we argue that the analytical model can be regarded as sufficiently accurate.

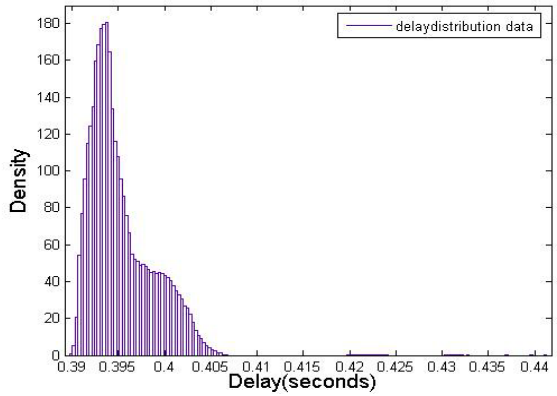


Figure 4: End-to-End Delay Distribution from Measurement

5 NUMERICAL RESULT

Based on the MPEG video flows whose characteristics shown in Table 1, we apply the analytical model to predict the number of flows that can be supported over a four-hop path, with D = 10 msec and ε = 0.05. The above end-to-end queueing delay requirement is translated into a nodal queueing delay requirement of 2.5 msec, and the probability bound of delay violation of 0.0127. Figure 5 shows the probability of nodal delay bound violation, $P(d_i > D)$ vs. the number of flows, respectively for Flow1 and

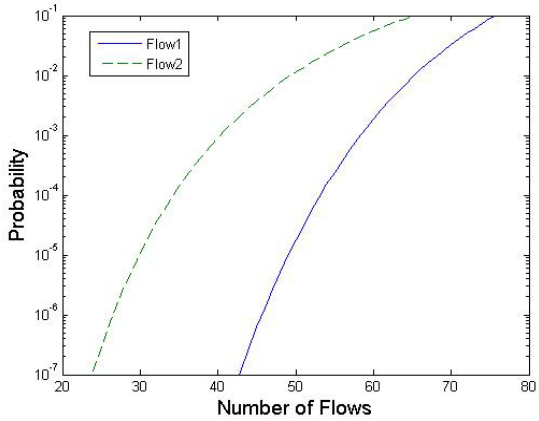
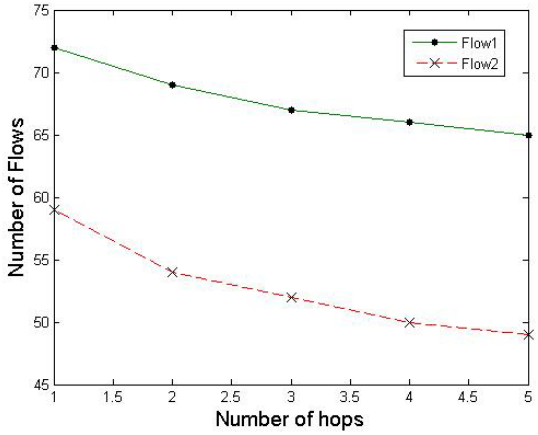


Figure 5: Probability of Nodal Delay Bound Violation vs. Number of flows



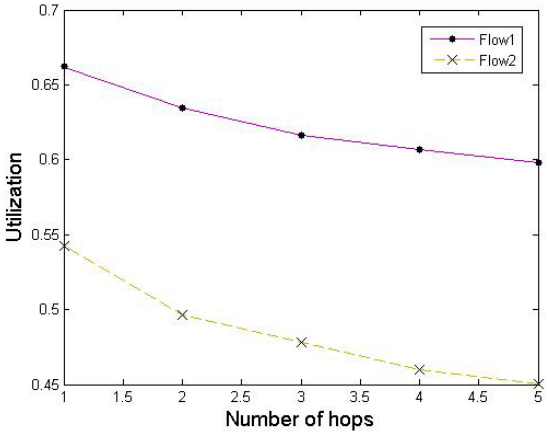
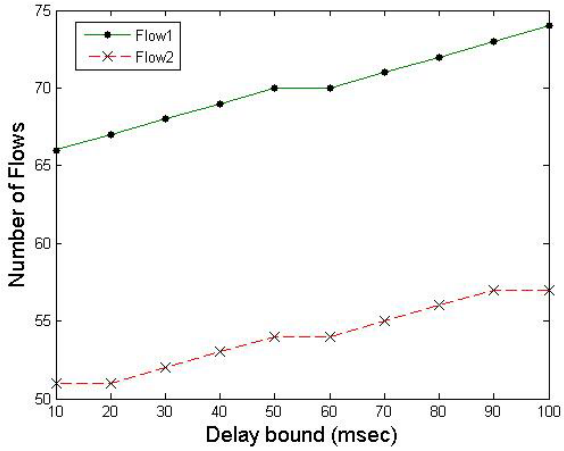


Figure 6: Effects of Path Length.



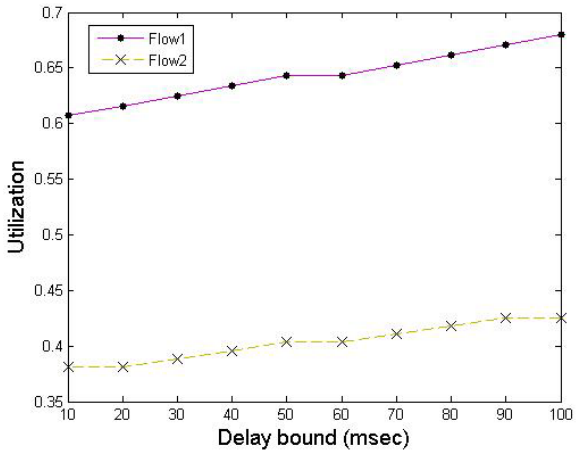


Figure 7: Effects of Delay Bound

Flow2 traffic. The maximum number of flows for Flow1 that satisfies the delay bound requirement, $P(d_i > D) < 0.0127$, is 65, and 49 flows for Flow2. The corresponding link utilization is respectively 60% and 36%. The network can support less number of Flow2 traffic due to its higher peak-to-mean rate. The results also suggest that each flow contributes about 1% of link utilization. Given that 15-20% of the link capacity carries best-effort (non-video) traffic, each node can admit only a few tens of users (assuming one flow per user). Clearly, a link capacity upgrade is needed if users from all around the campus are to be supported.

Figure 6 shows number of admissible flows for different number of hops, and the corresponding link utilizations. The network can support more video flows if they traverse less number of hops. In typical scenarios, a path will contain 2-3 hops, in which case the network can accommodate only a few more flows compared to the worst case (4 hops).

The effects of the delay bound requirements on the number of admissible flows and the utilization are shown in Figure 7. As expected, more number of flows can be supported when the delay bound is relaxed, and the amount of increase is linear. However, even if we increase the delay bound by 10 times (10 msec to 100 msec) only 8 more flows can be admitted. Therefore, relaxing the delay bound requirement does not significantly help improving the system capacity.

6 CONCLUSION

In this paper, we present an analytical model based on MVA approximation to predict the number of video flows that can be supported by the KMUTT campus network for E-learning service. Since the MVA approximation applies only for a single queue, we extend the method to translate the end-to-end queuing delay bound to a nodal delay bound based on a simple probabilistic bound. The results obtained from the model were validated with the measurement on the real network by using the distributed benchmark system. It is indicated that the analytical model is sufficiently accurate for the purpose of performance prediction. The model is then used to predict the number of flows that can be carried over the network for a given delay bound requirement. MPEG flows with different statistical characteristics are tested.

We have found that on a longest network path in the network, the number of admissible flows is well below a hundred, corresponding to the link utilization around 40 - 60%, depending on the flow characteristics. These results are for the best case when no other kind of traffic other than the video flows is presented. In practice, the video traffic must share the link with best-effort traffic. As such, the number of admissible flows can be much lower. Also, the number of admissible flows increases very slowly when increasing the delay bound. Therefore, in order to support users from all around the campus, we may have to upgrade the link capacity, or prioritize the video traffic to deliver satisfactory video quality for the E-learning service deployment

REFERENCES

Anick, D., Mitra, D., and Sondhi, M., Stochastic theory of a data-handling system with multiple sourcees, Bell Systems Technical Journal, vol. 61, pp. 1261–1281, 1982.

Choe J., and Shroff, N. B., A Central-Limit-Theorem-Based Approach for Analyzing Queue Behavior in High-speed Networks, IEEE/ACM Transactions on Networking, October 1998; 6; No 5: 659-671.

Courcoubetis, C., Siris, V., and Stamoulis, G., Application and evaluation of large deviation techniques for traffic engineering in broadband networks, in ACM SIGMETRICS'98, Madison, WI, USA, 1998.

Garrett, M. W., and Willinger, W., Analysis, Modeling and Generation of Self-Similar VBR Video Traffic, Proc. ACM SIGCOMM, pp. 269-280, 1994.

Gu´erin, R., Ahmadi, H., and Naghshineh, M., Equivalent capacity and its application to bandwidth allocation in high-speed networks, IEEE J. Select. Areas Commun., vol. 9, no. 7, pp. 968–981, Sept. 1991.

Kim, H., S., and Shroff, N. B., Loss Probability Calculations and Asymptotic Analysis for Finite Buffer Multiplexers, IEEE/ACM Transactions on Networking, December 2001; 9: No 6: 755-768.

Sahinoglu, Z., and Tekina, S. On Multimedia Networks: Self-Similar Traffic and Network Performance, IEEE Commun. Mag., January 1999.

AUTHOR BIOGRAPHIES

THANADECH THANAKORNWORAKIJ is a master student of Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand. He received a B.S. degree in applied mathematic from King Mongkut's Institute of Technology North Bangkok, Thailand, in 2001. His research interests include performance analysis of communication network and Quality of Service . His email address is no_on17@hotmail.com

PEERAPON SIRIPONGWUTIKORN received her B.S. degree in telecommunication from King Mongkut's Institute of Technology Ladkrabang, Thailand, in 1995 and completed his PhD in telecommunication from University of Pittsburgh, Pittsburgh, in 2003. He is currently holding a faculty position at the Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand. His research interests include Quality of Service, Ad hoc networks and Performance analysis of communication . His email address is peeropon@cpe.kmutt.ac.th