MMAP 이산시간 큐잉 시스템의 속산 시뮬레이션

국광호*, 강성열**

An Efficient Simulation of Discrete Time Queueing Systems with Markov-modulated Arrival Processes

Kwang-Ho Kook, Sungyeol Kang

Abstract

The cell loss probability required in the ATM network is in the range of 10^{-9}~ 10^{-12}. If Monte Carlo simulation is used to analyze the performance of the ATM node, an enormous amount of computer time is required. To obtain large speed-up factors, importance sampling may be used. Since the Markov-modulated processes have been used to model various high-speed network traffic sources, we consider discrete time single server queueing systems with Markov-modulated arrival processes which can be used to model an ATM node. We apply importance sampling based on the Large Deviation Theory for the performance evaluation of, $MMBP/D/1/K$, $\sum MMBP/D/1/K$, and two stage tandem queueing networks with Markov-modulated arrival processes and deterministic service times. The simulation results show that the buffer overflow probabilities obtained by the importance sampling are very close to those obtained by the Monte Carlo simulation and the computer time can be reduced drastically.

Key Words: simulation, important sampling, ATM, buffer overflow, Markov-modulated process

* Department of Industrial Engineering, Seoul National University of Technology
** Department of Management Information Systems, Hongik University

1. Introduction

In the high-speed networks, a wide variety of multimedia services such as voice, data, video, and image, may be supported by ATM nodes. When various sources with different traffic characteristics enter an ATM node, it is very hard to evaluate the performance of the ATM node by an analytical method. Although we can resort to simulation, it has the disadvantage of requiring large amount of computer time to obtain results with a sufficiently small confidence interval. In ATM networks this problem is very important since the cell loss probabilities of interest are extremely small, less than 10^{-9}.

In recent years, importance sampling has been widely used for the efficient simulation of rare events in stochastic processes. The basic idea of importance sampling is to simulate using a *biased* simulation distribution which can make the rare events under consideration occur more frequently and then weight the simulation data by the likelihood ratio. The optimal biased simulation distributions which minimize the variance of estimates obtained by the simulation can be derived based on the Large Deviation Theory (LDT) or on the Stochastic Gradient Decent (SGD) algorithm.

There have been many works on the fast simulation based on the LDT. Corttrell, Fort, and Malgouyres [1] proposed a fast simulation method based on the LDT and applied them to estimate the large exit times of the Aloha protocol. Parekh and Walrand [2] and Frater, Lennon, and Anderson [3] made use of the quick simulation method based on the LDT to estimate the large exit times of the cumulative backlog process of the Jackson networks of queues. Sadowsky [4] proposed an efficient simulation method based on the LDT to estimate the average time to buffer overflow in a stable $GI/GI/m$ queue. Parekh [5] made use of the quick simulation method based on the LDT to estimate tail probabilities of stationary waiting time in GI/D/1 and PCP/D/1 queues. Hiedelberger and Simba [6] considered the problem of extremely low packet loss rates in a voice-data multiplexer, via fast simulation based on LDP. There have also been works on the fast simulation based on the SGD. Devetsikiotis, Al-Qaq, Freebersyser, and Townsend [7], [8] proposed an efficient simulation based on the SGD to estimate the blocking probability for a queue with two arrival streams (M-IBP $+ \text{MMBBP}/D/1/K$) and for tandem networks of M-IBP $+\text{MMBBP}/\text{Geo}/1/K$ queues.

Important Sampling has been used to estimate rare event probabilities in communication network simulations. Townsend, Haraszti, Freebersyser, and Devetsikiotis [9] presented an overview of the important sampling applications in communication networks. Gallardo, Makrakes and Barbosa [10] proposed a method based on regenerative Important Sampling in solving long-range dependent bursty traffic in broadband telecommunication networks.

The Markov-modulated arrival processes like IBP (Interrupted Bernoulli Process) and MMBP(Markov Modulated Bernoulli Process) have been extensively used to model various broadband traffic sources, such as voice and video, and the superposed traffic.

In this paper, we are interested in the buffer overflow probabilities in the discrete time single server queueing systems with Markov-modulated arrival processes which can be used to model ATM node. Townsend[8] showed that the cell loss probabilities in the tandem network with Markov-modulated arrival processes can be obtained by the SGD. On the other hand, we apply important sampling based on the LDT for the performance evaluation of, $MMBP/D/1/K$, $\sum MMBP/D/1/K$, and two stage tandem queueing networks with Markov-modulated arrival processes and deterministic service times as in [11]. The simulation results show that the buffer overflow probabilities obtained by the importance sampling are very close to those obtained by the Monte Carlo simulations and the computer time can be reduced drastically by the importance sampling.

2. Preliminaries

Suppose that we are interested in the buffer overflow probability of a $GI/GI/1/K$ queue. Let A and B represent interarrival and service time distribution functions and let M_a and M_b be moment generating functions of A and B. It follows from [4] that the optimal biased interarrival and service time distributions for the importance sampling can be obtained by the following exponential change of interarrival and service time distributions

$$dA^*(z) = \frac{e^{\theta_a^* z} dA(z)}{M_a(\theta_a^*)}$$

$$dB^*(z) = \frac{e^{\theta_b^* z} dB(z)}{M_b(\theta_b^*)}, \tag{1}$$

where $\theta_a^* = -\theta^*$, $\theta_b^* = \theta^*$ for θ^* satisfying

$$M_a(-\theta^*)M_b(\theta^*) = 1 . \tag{2}$$

Let $N(t)$ denote the number of jobs in queue at time t. Then the buffer overflow probability ξ is the probability that with $N(0)=0$, $N(t)$ reaches K before reaching zero again. We define a cycle as the duration starting with an empty system and ending at the instant the system, for the first time, either becomes empty again or reaches K. Define

$$V_m = \begin{cases} 1 : N(t) \text{ reaches } K \text{ during cycle } m \\ 0 : \text{otherwise} \end{cases}$$

Then, after n cycles of simulation under A^*, B^*, ξ can be estimated by

$$\xi = \frac{V_1L_1 + V_2L_2 + \cdots + V_nL_n}{n}, \tag{3}$$

where L_m is given by

$$L_m = \prod_{k=1}^{K_a} \frac{dA(A_k)}{dA^*(A_k)} \prod_{k=1}^{K_b} \frac{dB(B_k)}{dB^*(B_k)} . \tag{4}$$

In (4), $A_k \geq 0$ is the interarrival time between the arrivals of jobs $k-1$ and k, $B_k \geq 0$ is the service time of job $k-1$, K_a is the number of jobs arrived, and K_b is the number of jobs served in the m th cycle respectively.

3. Simulation of Buffer Overflow Probabilities

3.1 MMBP/D/1/K Queue

We consider a MMBP/D/1/K queue with two state MMBP arrival process and the deterministic service time, $1/\mu$. In MMBP arrival process, a slot is either in state 1 or in state 2. A slot in state 1 contains a cell with probability α and no cell with probability $1-\alpha$, and a slot in state 2 contains a cell with probability β and no cell with probability $1-\beta$. Given that the slot is in state 1, the next slot is also in state 1 with probability p and changes to state 2 with probability $(1-p)$. Similarly, given that the slot is in state 2, the next slot is also in state 2 with probability q and changes to state 1 with probability $(1-q)$.

Since the moment generating function $M_a(-\theta)$ of the interarrival time of cells following MMBP process is given by [12]

$$M_a(-\theta)=\frac{c_2e^{-2\theta}+c_1e^{-\theta}}{d_2e^{-2\theta}+d_1e^{-\theta}+d_0}\ , \quad (5)$$

where

$c_2=(1-p-q)[(1-q)(1-\beta)\alpha^2+(1-p)(1-\alpha)\beta^2]$

$c_1=(1-q)\alpha[p\alpha+(1-p)\beta]+(1-p)\beta[q\beta+\alpha(1-q)]$

$d_2=(1-\alpha)(1-\beta)(p+q-1)[(1-q)\alpha+(1-p)\beta]$

$d_1=-[(1-q)\alpha+(1-p)\beta][q(1-\beta)+p(1-\alpha)]$

$d_0=(1-q)\alpha+(1-p)\beta,$

and the moment generating function $M_b(\theta)$ of the constant service time of $1/\mu$ slots is

$$M_b(\theta)=e^{\frac{1}{\mu}\theta}, \quad (6)$$

we can obtain optimal biased interarrival time distribution by (1) and obtain ξ by (3).

For the purpose of illustration, we consider a queue with an MMBP arrival process having $\alpha=0.4$, $\beta=0.2$, $p=0.8$, $q=0.8$ and with constant service time of 2.5 slots ($\mu=0.4$). From (5) and (6), we can obtain $\theta^*=0.31889$ satisfying

$$M_a(-\theta)\times M_b(\theta)=\frac{-0.01824e^{-2\theta}+0.0384e^{-\theta}}{0.03456e^{-2\theta}-0.1344e^{-\theta}+0.12}\times e^{2.5\theta}=1.$$

According to Eq.(1), the moment generating function of the biased interarrival time distribution becomes

$$M_a^*(-\theta)=\frac{-0.021392e^{-2\theta}+0.061954e^{-\theta}}{0.018264e^{-2\theta}+0.09777e^{-\theta}+0.12}. \quad (7)$$

From (5) and (7), we see that the biased arrival process also follows an MMBP process whose parameters satisfy $c_2=-0.021392$, $c_2=0.061954, d_2=0.018264$, $d_1=0.09777$, $d_0=0.12$

The optimal biased arrival process, MMBP process with parameters $\alpha^*=0.63595$, $\beta^*=0.44975$, $p^*=0.83389$, $q^*=0.92855$, can be obtained by the stochastic gradient technique of the Robbins-Monro type which minimizes

$$f(\alpha^*,\beta^*,p^*,q^*)=(c_2+0.021392)^2+(c_1-0.061954)^2+(d_2-0.018264)^2+(d_1-0.09777)^2+(d_0-0.12)^2$$

Similarly, we can obtain the biased service time distribution which is the same as the original service time distribution.

Now the buffer overflow probability ξ can be estimated by (3) after n cycles of simulation under biased distributions. Since the service time is not changed, we can obtain L_m by updating L_m in each slot based on the state of the biased arrival process during mth cycle of simulation as shown in Table 1.

Table 2 shows ξ obtained by the Monte Carlo simulation and importance sampling when simulation was done until buffer overflows occur 200 times. We see that ξ can be estimated very accurately by importance sampling and the number of cells generated under the Monte Carlo simulation increases exponentially but the number of cells generated under the importance sampling increases almost linearly.

Table 1: Computation procedure of the L_m

IBP state of previous slot	IBP state of current slot	L_m
active	active, cell arrives	$L_m = L_m \times \frac{p\alpha}{p^*\alpha^*}$
	active, cell doesn't arrive	$L_m = L_m \times \frac{p(1-\alpha)}{p^*(1-\alpha^*)}$
	silent	$L_m = L_m \times \frac{(1-p)}{(1-p^*)}$
silent	silent	$L_m = L_m \times \frac{q}{q^*}$
	active, cell arrives	$L_m = L_m \times \frac{(1-q)\alpha}{(1-q^*)\alpha^*}$
	active, cell doesn't arrive	$L_m = L_m \times \frac{(1-q)(1-\alpha)}{(1-q^*)(1-\alpha^*)}$

Table 2: Buffer overflow probabilities of MMBP/D/1/K queue

buffer size K	Monte Carlo $\alpha=0.4$, $\beta=0.2$, $p=0.8$, $q=0.8$, $\mu=0.4$		Importance Sampling $\alpha^*=0.63595$, $\beta^*=0.44975$, $p^*=0.83389$, $q^*=0.92855$, $\mu^*=0.4$	
	$E[\xi]$	# of cells	$E[\xi]$	# of cells
3	0.28058	4564	0.27650	624
4	0.12538	15198	0.12516	1325
5	0.05687	39430	0.05161	2047
6	0.02769	88967	0.02670	2850
7	0.00878	207258	0.00994	3901

3.2 $\sum$MMBP/D/1/K Queue

In this section, we consider discrete time single server queues with multiple arrival streams and deterministic service time, $1/\mu$. As shown in (1), we need the moment generating function of the superposed arrival process in order to apply importance sampling based on the LDT. Since it is very hard to get the moment generating function of the superposed arrival process, we adopt a heuristic approach to obtain approximate biased arrival process. If there are N arrival streams with mean arrival rates λ_i, $i=1,\cdots N$, the arrival stream i will be serviced with rate proportional to their relative arrival rates under the heavy traffic. So we can derive the approximate biased interarrival time distribution of the stream i by considering an imaginary queue with arrival stream i and deterministic service time $1/\widehat{\mu}_i$, where $\widehat{\mu}_i=\mu\times\lambda_i/\sum\lambda_i$, $i=1,\cdots,N$.

We now consider a queue with multiple MMBP arrival process and a deterministic service time $1/\mu$. Suppose that 5 MMBP processes with the same parameter values ($\alpha_i=0.4$, $\beta_i=0.2$, p_i=0.8, q_i=0.8, $i=1,\cdots,5$) arrive to a queue with the constant service time 0.4 slot ($\mu=2.5$). Since the mean arrival rate of each arrival stream is 0.3/slot, under the heavy traffic, the service rate of stream i will become

$$\mu_i=2.5\times\frac{0.3}{5\times0.3} = 0.5.$$

Then as in Section 3.1, we can derive biased MMBP arrival streams with parameter values

$$\alpha_i^*=0.78666,\ \beta_i^*=0.65830,\ p_i^*=0.88009,$$
$$q_i^*=0.94781,\ i=1,\cdots,5$$

by considering a queue with MMBP arrival stream having parameters $\alpha_i=0.4$, $\beta_i=0.2$, $p_i=0.8$, $q_i=0.8$ and a constant service time, 2 slots. The biased service time for importance sampling is not changed ($\mu^*=\widehat{\mu}_1+\cdots+\widehat{\mu}_5=2.5$).

Table 3 shows ξ obtained by the Monte Carlo simulation and importance sampling when applied to stream 1 only, to stream 1 and 2, to stream 1, 2, and 3, to stream 1, 2, 3, and 4, and to all streams. From Table 3, we can see that importance sampling gives very accurate values of buffer overflow probabilities even though applied to the subset of the arrival streams.

3.3 Two Stage Tandem Queueing Network

In this section, we consider a two-stage tandem queueing network with multiple 2 state MMBP arrival streams and deterministic service time (Figure 1). Suppose that 5 arrival streams following MMBP processes arrive to the first queue and there is no external arrivals to the second queue. When a service completion occurs at the first queue, we assume that the cells belonging to stream 1 proceed to the second queue and the cells belonging to stream 2, 3, 4, and 5 leave the network. Then the arrival process to the second queue becomes the departure process of stream 1 at the first queue.

Table 3: Buffer overflow probabilities of $\sum_{i=1}^{5}$MMBP/D/1/K queue

Buffer Size K	Monte Carlo $\alpha=0.4,\ \beta=0.2,\ p=0.8,\ q=0.8,\ \mu=2.5$		Importance Sampling (stream 1) $\alpha^*=0.78666,\ \beta^*=0.65830,\ p^*=0.88009,\ q^*=0.9478,\ \mu^*=2.5$		Importance Sampling (stream1, stream2)	
	$E[\xi]$	# of cells	$E[\xi]$	# of cells	$E[\xi]$	# of cells
3	0.53879	2707	0.52737	2540	0.52189	2185
4	0.28902	7906	0.27604	6391	0.27125	4805
5	0.10557	22070	0.10453	17501	0.10341	10613
6	0.03700	91811	0.03729	45389	0.03482	21156
7	0.01217	177980	0.00871	61532	0.01081	40088

Buffer Size K	Importance Sampling (stream1, stream2, stream3)		Importance Sampling (stream1, stream2, stream3, stream4)		Importance Sampling (all streams)	
	$E[\xi]$	# of cells	$E[\xi]$	# of cells	$E[\xi]$	# of cells
3	0.52187	2183	0.52615	2081	0.50367	2036
4	0.29015	4558	0.28320	3746	0.28105	3389
5	0.11726	8368	0.11324	6621	0.11612	5622
6	0.03961	14077	0.03257	10090	0.03313	8209
7	0.00866	20501	0.01032	13770	0.01072	10890

Since the departure process of m-MMBP/Geo/1/K queue with m-state MMBP arrival process and geometric service time can be approximated by the m-MMBP process [13], we assume that the departure process of stream 1 from the first queue can be approximated by the 2 state MMBP process. A process can be approximated by a 2-state MMBP process if we know the average number of cells per slot, ρ, the squared coefficient of variation of the interarrival time, c^2, the lag 1

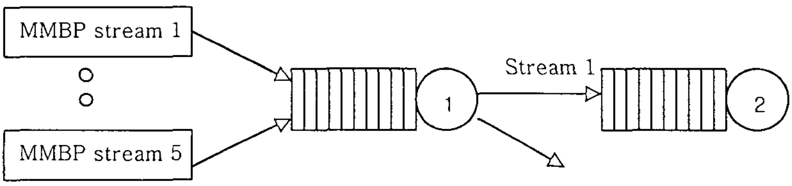

Figure 1. Two-stage tandem queueing network

autocorrelation for the interarrival time and the number of arrivals in each slot, $\psi(1)$, $\phi(1)$ [12], [13]. In order to approximate the arrival process to the second queue by a 2 state MMBP process, we assume that $\psi(1)$, $\phi(1)$ of the departure process of the stream 1 from the first queue have the same values as those of the external arrival stream 1 and estimate ρ and c^2 by applying the QNA technique [14]. (The $\psi(1)$ value of the departure process from m-MMBP/Geo $/1/K$ queue is very close to that of the arrival process [13].) After the arrival process to the second queue has been approximated by 2-state MMBP process, we can obtain biased arrival process to the second queue as in Section 3.1.

For the purpose of an illustration, suppose that stream 1, ... , stream 5 follow the same MMBP processes with parameters $\alpha=0.4$, $\beta=0.2$, $p=0.8$, $q=0.8$, the service time of the first queue is deterministic 0.5 slot, and the service time of the second queue is deterministic 2.5 slots. We can obtain $\rho=0.3$, $c^2=0.89$, $\psi(1)=0.01715$, $\phi(1)=0.02857$ for the arrival process to the second queue by the above procedure. From these, we can approximate the arrival process to the second queue by the 2-state MMBP process with parameters $\alpha=0.42524, \beta=0.22601$, $p=0.77839$, $q=0.86908$. Then we can obtain the biased arrival process to the second queue, a MMBP process with parameters $\alpha^*=0.51881, \beta^*=0.32126$, $p^*=0.97886$, $q^*=0.91004$.

Table 4 shows ξ in the second queue obtained by simulating the whole tandem queueing network by the Monte Carlo simulation, by simulating the second queue in isolation by the Monte Carlo simulation, and by simulating the second in isolation queue by importance sampling. From Table 6, we can observe that about 20% of relative error occurs by assuming that the departure process follows MMBP process with the same $\psi(1)$, $\phi(1)$ values as those of the external arrival process. This error may be reduced by approximating departure processes more accurately.

Table 4: Buffer overflow probabilities of two-stage tandem queueing network

buffer size K	Monte Carlo (queueing network) $\alpha=0.4, \beta=0.2, p=0.8$ $q=0.8, \mu_1=80, \mu_2=0.4$		Monte Carlo (second queue) $\alpha=0.42524$, $\beta=0.22601$, $p=0.77839$, $q=0.86908$, $\mu=0.4$		Importance Sampling (second queue) $\alpha^*=0.51881$, $\beta^*=0.32126$, $p^*=0.97866$, $q^*=0.91004$, $\mu=0.4$	
	$E[\xi]$	# of cells	$E[\xi]$	# of cells	$E[\xi]$	# of cells
3	0.28090	4819	0.30257	4318	0.29243	769
4	0.12516	15322	0.13004	14627	0.14420	1649
5	0.05346	36417	0.06614	35126	0.06629	2500
6	0.02480	89830	0.02969	87531	0.02912	3506
7	0.01167	209257	0.01344	203780	0.01316	4557

4. Conclusion

In this paper we have illustrated the applicability of the importance sampling to the performance studies of the ATM node which can be modeled as a discrete time queueing system with Markov-modulated arrival processes. The results show that we can obtain large speed-up factors by the importance sampling. When voice, data, video, and image services are supported by the ATM node simultaneously, it is very hard to apply importance sampling because of the difficulties in obtaining the moment generating function of the superposed arrival process. Since the importance sampling even when applied to the subset of the arrival streams gives very accurate results as shown in Section 3.2, we may be able to apply importance sampling only to the traffic sources following Markov-modulated arrival processes by the proposed heuristic method.

On the other hand, the biased arrival distributions obtained by our heuristic approach can be improved by applying Stochastic Gradient Descent algorithm to minimize the variance of the estimates obtained by the importance sampling. This will be able to reduce much effort in finding optimal biased arrival distribution based on the SGD by providing good initial starting values of the parameters. In order to apply importance sampling for the performance studies of the networks of queues, further studies are necessary for the departure processes from each queue in the network.

References

[1] M. Cottrel, J. C. Fort and G. Malgouyres, "Large Deviations and Rare Events in the Study of Stochastic Algorithms", IEEE Trans. on Auto. Cont., Vol. AC-28, No. 9, 1983.

[2] S. Parekh and J. Walrand, "Quick Simulation Method for Excessive Backlogs in Networks and Queues", IEEE Trans. on Auto. Con., Vol. 34, No. 1, Jan. 1989. pp.887 ~ 896

[3] M. R. Frater, T. M. Lennon and B. D. O. Anderson, "Optimally Efficient Estimation of the Statistics of Rare Events in Queueing Networks", IEEE Trans. on Auto. Cont., Vol. 36, No. 12, Dec. 1991.

[4] J. S. Sadowski, "Large Deviations Theory and Efficient Simulation of Excessive Backlogs in a GI/GI/m Queue", IEEE Tran. on Auto. Con., Vol. 36, No. 12, Dec. 1991.

[5] S. P. Parekh, "Quick Simulation of Stationary Tail Probabilities at a Packet Switch," ITC 14, 1994.

[6] P. Heidelberger and R. Simha, "Fast Simulation of a voice-data multiplexer", International Journal of Modeling and simulation, Vol. 18, No. 4, 1998

[7] M. Devetsikiotis, W. A. Al-qaq, J. A. Freebersyser and J. K. Townsend, "Stochastic Gradient Techniques for the Efficient Simulation of High Speed Networks using Important Sampling", In Proc. IEEE Global Telecom. Conf., GLOBECOM '93, Houston, Dec. 1993.

[8] J. A. Freebersyser, M. Devetsikiotis, W. A. Al-qaq, and J. K. Townsend, "Fast Simulation of Tandem Networks Using Importance Sampling and Stochastic Gradient Techniques", In Proc. 1996 IEEE

International Conf. on Comm., ICC '96, Dallas, June 1996.

[9] J. K. Townsend, Z. Haraszti, J. A. Freebersyser, and M. Devetsikiotis, "Simulation of Rare Events in Communication Networks", IEEE Communications, Vol. 36, No. 8, Aug. 1998.

[10] J. R. Gallardo, D. Makrakis, and L. O. Barbosa, "Fast Simulation of Broadband Telecommunications Networks Carrying Long-Range Dependent Bursty Traffic," Winter simulation conference Proceedings, Vol. 1, 1999, pp. 374 ~ 381.

[11] K. H. Kook, C. H. Lee, S. Y. Kang, and C. H. Oh, "Performance Evaluation of ATM Node by Importance Sampling Simulation," Journal of the Korean Operations Research and Management Science Society, Vol. 22, No. 4, Dec. 1997, pp. 1 ~ 16.

[12] R. O. Onvural, "Asynchronous Transfer Mode Techniques," Artech House, Inc., 1994.

[13] D. Park and M. G. Perros, "m-MMPP Characterization of the Departure Process of an m-MMPP/Geo/1/K/Queue," ITC 14, 1994.

[14] W. Whitt, "The Queueing Network Analyzer", The Bell System Technical Journal, Vol.62, No.9, Nov.1983, pp 2779 ~

● 저자소개 ●

국광호

1979 서울대학교 공과대학 산업공학과 학사

1981 서울대학교 공과대학 산업공학과 석사

1989 미국 Georgia Institute of Technology 산업시스템공학과 박사

1982~1984 청주대학교 산업공학과 전임강사

1989~1993 한국전자통신연구원 선임연구원

1993~현재 서울산업대학교 산업정보시스템공학과 교수

관심분야: 대기행렬, 추계적과정, 통신 트래픽 제어 및 성능 평가

강성열

1981 서울대학교 공과대학 산업공학과 학사

1983 서울대학교 공과대학 산업공학과 석사

1992 미국 Georgia Institute of Technology 산업시스템공학과 박사

1983~1998 한국전자통신연구원 책임연구원

1999~2000 KAIST 테크노경영대학원 초빙교수

2000~현재 홍익대학교 경영정보학과 부교수

관심분야: 네트워킹, 인터넷, 통신망 트래픽 및 성능 분석, 추계적 모델링