Proceedings of International Conference
on Computer Architecture

Delay Analysis in Computer Communication Systems

Jang Hyun Baek*, Chang Hoon Lie**
Electronics and Telecommunications Research Institute*
161 Kajong-Dong, Yusong-Gu, Taejon, KOREA, 305-350
Seoul National University**

Abstract

In this paper, we propose a seemingly efficient method computing the delay distributions in computer communication systems including switching system. The proposed method is based on GPH(Generalized Phase-Type) distributions which are the distributions of random sum of i.i.d. exponential random variables. The class of GPH distributions has several properties which can be applied conveniently for computational purposes. The basic methodology adopted in delay analysis for switching system is explained step by step. Application results are given to demonstrate the accuracy of the method and to verify that the system meets the requirements of delay. The proposed method seems to be very useful to analyze delay performance in computer communication systems.

1. Introduction

There have been several approaches for the delay performance in general computer communication networks. But most of them are focused only on the first and second moments of the delay distributions, and furthermore the distribution itself is not obtained directly. Apparently, the distribution is more informative than first two moments and there are some situations that we need to get the distribution itself. For example, to compute the performance criteria like 95 percentile of the call-setup delay in the switching system of the telecommunication network, we need the distribution information. In this paper we propose a convenient and seemingly accurate approximation method to derive delay distributions in general computer communication networks including switching system, which is based on GPH(Generalized PHase type) distributions.

GPH distributions proposed by Shanthikumar[4] are very broad class of distributions which include PH distribution[1] as a subset and have several useful properties applicable for computational purposes. Thus, GPH distributions have a potential to be conveniently used to develop approximate computational methods in performance analysis. In this paper, we utilize GPH method to derive delay distributions of switching system.

After this introduction, the GPH distribution is defined in section 2 and the CDMA(Code Division Multiple Access) mobile switching system(CMS) is introduced in section 3. In section 4, overall procedure to derive delay distributions in switching system is explained. In section 5, application results are given. Finally, concluding remarks are given in section 6.

2. GPH distribution

Let $E_0=0$ and E_n $(n=1,2,\ldots)$ be i.i.d. exponential r.v.s with rate λ, and L be a discrete r.v. with probability mass function g. Define

$$V=\sum_{n=0}^{L} E_n.$$

Then, its cumulative distribution function $F(x)=P\{V\leq x\}$ is

$$F(x)=\sum_{n=0}^{\infty} G(n)\frac{e^{-\lambda x}(\lambda x)^n}{n!},\quad x\geq 0, \qquad (1)$$

where $G(n)=\sum_{i=0}^{n} g(i)$ is the cumulative probability distribution of L. $F(x)$ given by (1) is called a *generalized phase-type*(GPH) distribution[4] and represented by $GPH(\lambda, g)$.

GPH has many promising properties. As given in [4], the class of GPH distributions are closed under finite convolutions and mixtures. And, any continuous distributions with support $[0, \infty)$ can be approximated to GPH as following.

For a bounded function F, define

$$F_\lambda(t) = \sum_{n=0}^{\infty} F(\frac{n}{\lambda}) \frac{e^{-\lambda t}(\lambda t)^n}{n!}.$$

Then, it is known that $F_\lambda(t) \to F(t)$ as $\lambda \to \infty$ [7].

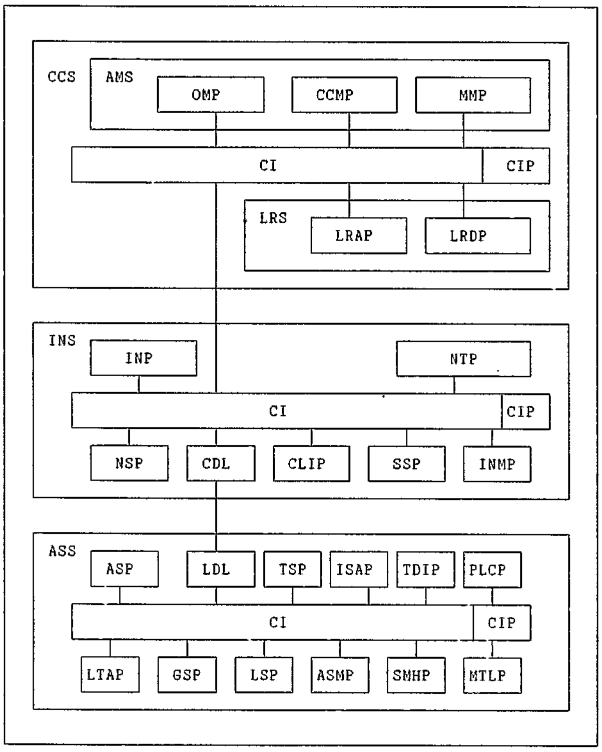
And, in the empirical study of Yoon et al[7], it is experimentally shown that with relatively small λ (<50), the approximation is very accurate.

3. CDMA switching system

In this section, CMS is introduced. CMS architecture and call processing scenario in CMS are explained.

3.1. CMS architecture

The CMS is a very large capacity digital mobile switching system developed by ETRI(Electronics and Telecommunications Research Institute), along with other Korean manufacturers. It performs MSC (Mobile Switching Center) function and LR(Location Registration) function.



<Fig. 1 : CMS architecture>

As shown in Fig. 1, CMS is composed of three subsystems[8]. ASS(Access Switching Subsystem) performs distributed call processing function and local M&A(Maintenance and Administration) function. INS(Interconnection Network Subsystem) performs centralized call processing function and control function for interconnection of processors. CCS(Central Control Subsystem) consists of AMS(Administration and Maintenance Subsystem) and LRS(Location Registration Subsystem). AMS performs centralized M&A function and LRS performs management for location information of mobile subscribers.

The processors in which messages are processed are shown in Fig. 1. The functions of main processors related to call processing function are as follows.

- ASP(Access Switching Processor) performs call processing function and local M&A function in ASS
- TSP(Time Switch Processor) controls time switch
- SSP(Space Switch Processor) controls space switch
- LSP(Local Service Processor) controls tone generators
- SMHP(Signaling Message Handling Processor) performs SS(Signaling System) No. 7 signaling
- LRDP(Local Register Data Processor) and LRAP (Local Register Application Processor) store and manage location information of mobile subscriber
- INP(Interconnection Network Processor) performs connection and release of swich,
- NTP(Number Translation Processor) performs number translation

3.2. Call processing scenario

For call connection, the switching system must process various messages in many processors according to call processing scenario. Fig. 2 shows a call processing scenario for M-M(from mobile subscriber to mobile subscriber) internal call and indicates message flow among blocks included in processors. In Fig. 2, MSA, LSL, MCC, MSL and CDGM are included in ASP, TSL in TSP, SNC in INP, SSW in SSP, NTR, RCO and MPC in NTP, LCC and LLU in LRAP, respectively. And subscriptions a, b of block name indicate that the block is related to originating subscriber(a) or terminating subscriber(b).

The delay items we will evaluate are represented with thick lines in the figure and explained in detail later.

4. Overall procedure to derive delay distributions in CDMA switching system

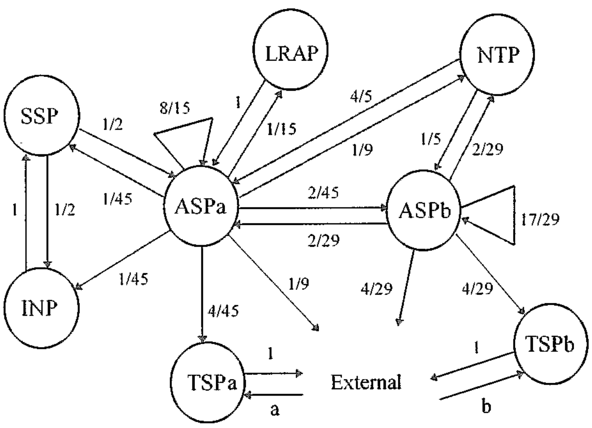
Overall procedure to derive delay distributions in CMS is explained step by step in detail.

4.1. Queueing network modeling for call processing in CMS

To analyze the switching system numerically, we model it as queueing network. Processors in the switching system are considered as nodes and transitions among processors are considered as links among nodes. From the call processing scenario, transition probabilities are obtained and assigned to the corresponding links.

With Fig. 2, we describe transition diagram of call processing in CMS as queueing network model(see Fig. 3). In Fig. 3, nodes and links stand for processors and transition of messages, respectively.

The resultant queueing network has eight nodes, of which only two nodes are related to subscriber access. These two exogenous arrival processes are known to be Poisson processes and their exact service time distributions are known in general switching system.



<Fig. 3 : Queueing network model>

4.2. Calculation of total arrival parameters at each node

Let the exogenous arrival process at node i have arrival rate λ_{0i}, and let coefficient of variation c_{0i} and service process have service rate μ_i and coefficient of variation c_{si}, respectively. Because the stochastic processes in the queueing network are very complicated and the arrival processes at nodes are closely related one another, it is impossible to obtain exact relations. Thus, the relations of first two moments are used to approximately describe the interactions within the network.

More specifically, we calculate total arrival rate and squared coefficient of variation for each node using the following relations[5].

$$\lambda_j = \lambda_{0j} + \sum_{i=1}^{N} \lambda_i q_{ij}, \quad j = 1, 2, \ldots N \qquad (2)$$

$$c_{aj}^2 = a_j + \sum_{i=1}^{N} c_{ai}^2 b_{ij}, \quad j = 1, 2, \ldots, N \qquad (3)$$

where

λ_{0j} : external arrival rate to node j

q_{ij} : transition probability from node i to j

$$a_j = 1 + w_j[(p_{0j} c_{0j}^2 - 1) + \sum_{i=1}^{N} p_{ij}((1 - q_{ij}) + q_{ij}\rho_i^2 x_i)]$$

$$b_{ij} = w_j p_{ij} q_{ij} (1 - \rho_i^2)$$

$$x_i = 1 + [\max(c_{si}^2, 0.2) - 1]$$

$$w_j = [1 + 4(1 - \rho_j)^2 (\nu_j - 1)]^{-1}.$$

4.3. Approximation of arrival distribution to a GPH distribution at each node

If a general distribution with mean m and variance v is approximated to a GPH distribution, i. e. if

$X \sim GPH(\lambda, g)$

where $L \sim G(n)$, then, followings are satisfied.

$E(L) = \lambda m$

$V(L) = \lambda^2 v - \lambda m$

Therefore if the distribution of L is approximated, a general distribution is represented as a GPH distribution.

Negative binomial distribution is selected as an adequate distribution of L, because it has two parameters and the squared coefficient of variation may be less than 1 or equal to 1 or greater than 1.

When $E(L)$ and $V(L)$ is known, n and p can be obtained as

$$p = \frac{\lambda v}{m} - 2 \qquad (4)$$

$$n = \frac{\lambda m}{p} \qquad (5)$$

As n is an integer, n obtained from (5) is rounded off and then p is calculated from (4) with integer n. Finally, an unknown distribution with known mean and variance is approximated to a GPH distribution.

4.4. Approximation of service time

distribution to a GPH distribution at each node

Using the results in section 2, the service time distribution can be approximated to a GPH distribution.

4.5. Calculation of delay distribution at each node

Since interarrival time distribution and service time distribution at each node are approximated to GPH distributions respectively, each node is modeled to GPH/GPH/1 queue[4]. Therefore.delay distribution at each node is calculated using the following result.

Let A_n and B_n be the interarrival time between the nth and $(n+1)$th customer and the service time of the nth customer. And let W^n and W_q^n be the time spent in the system and the waiting time in the queue, respectively, for the nth customer.

Suppose A_n and B_n are both GPH r.v.s with representations (λ_1, g_a) and (λ_2, g_s), respectively. Without loss of generality we will assume that $\lambda_1=\lambda_2=\lambda$. And suppose W_q^0 has a GPH distribution with representation (λ, g_q^0). Then W^n and W_q^n have GPH distribution with representation (λ, g_w^n) and (λ, g_q^n), respectively, in which

$$g_w^n = g_q^n * g_s, \quad g_q^{n+1} = g_w^n \oplus g_a^*, \quad n=0,1,\ldots$$

where $g_a^*(0)=\sum_{k=0}^{\infty}(\frac{1}{2})^{k+1}g_a(k+1)+g_a(0)$

$$g_a^*(n)=\sum_{k=0}^{\infty}\binom{n+k}{k}(\frac{1}{2})^{n+k+1}g_a(k+1)$$

and $\oplus$ stands for negative convolution with nonnegativity condition.

An immediate consequence is the following. Suppose g_w and g_q are the steady-state probability distribution functions of the time spent in the system and the waiting time in the queue, respectively, in a discrete GI/G/1 queueing system with interarrival time distribution g_a^* and service time distribution g_s. Then for the GPH/GPH/1 queue,

$$\lim_{n\to\infty} P\{\ W_q^n \le x\ \} = \sum_{k=0}^{\infty} G_q(k)\frac{e^{-\lambda x}(\lambda x)^k}{k!}$$

$$\lim_{n\to\infty} P\{\ W^n \le x\ \} = \sum_{k=0}^{\infty} G(k)\frac{e^{-\lambda x}(\lambda x)^k}{k!}$$

One may now use the computational approaches developed for the discrete queueing systems to obtain the above equations.

4.6. Calculation of delay distribution

Overall procedure to derive delay distribution in a switching system is composed of two steps. First step is to model message transition process in switching system to a semi Markov process, and second step is to obtain the distribution of time to enter a specific absorbing state in this semi Markov chain.

General Semi Markov kernel $R_{ij}(t)$ of semi Markov process X is

$$R_{ij}(t) = P_{ij}\ H_{ij}(t) \qquad (6)$$

where P_{ij} is transition probability and $H_{ij}(t)$ is distribution of time to transition, and $R(t)$ has derivative $r(t)$[3]. Furthermore, let $R^{(n)}$ and $r^{(n)}$ be the n-fold matrix convolutions of R and r with themselves respectively.

If distribution of time to transition H is represented in GPH distribution, then we can use the results in section 4.5 and obtain numerical results for related functions without calculation of integration.

Now assume that for $i\neq j$, H_{ij} is the $GPH(\lambda,\ \hat{h}_{ij})$ distribution of the form

$$H_{ij}(t)=\sum_{n=0}^{\infty}\hat{H}_{ij}(n)e^{-\lambda t}\frac{(\lambda t)^n}{n!}$$

where $\hat{H}_{ij}(n)=0,\ i\neq j$. Then, from (6) we have the semi Markov kernel $R(t)$

$$R(t) = \sum_{n=0}^{\infty}\hat{R}(n)\,e^{-\lambda t}\frac{(\lambda t)^n}{n!} \qquad (7)$$

where $\hat{R}_{ij}(t) = P_{ij}\ \hat{H}_{ij}(t)$.

A matrix of functions of the form (7) is said to be of the *generalized matrix phase type*(GMPH). We use $GMPH(\lambda, \hat{r})$ to represent the GMPH function like $R(t)$ in (7).

With this representation, we consider the transition probabilities of X. Let $Q_{ij}(t)=P\{X(t)=j\,|\,X(0)=i\}$ then

$$Q(t)=\sum_{n=0}^{\infty}\hat{Q}(n)\frac{e^{-\lambda t}(\lambda t)^n}{n!}$$

where $\hat{Q}(0) = I$

$$\hat{Q}(n)=\sum_{k=0}^{n}\hat{m}(k)\,[I-diag(\hat{R}(n-k)1)]$$

$$\hat{m}(0) = I$$

$$\hat{m}(k)=\sum_{v=0}^{k-1}\hat{m}(v)\ \hat{r}(k-v),\ \ k=1,2,\ldots$$

$$\hat{r}(l)=\hat{R}(l)-\hat{R}(l-1),\quad l=1,2,\ldots$$

Now consider the probability distribution of T_{iB}, the first transition time from a state in A to any states in B for exclusive state space A and B. Then, $P\{T_{iB} > t\} = (Q_A(t)1)_i$ where $Q_A(t)$ is the transition probability matrix of X_A. Finally when R is of $GMPH(\lambda, d)$, we have

$$P\{T_{iB} > t\} = \left(\sum_{n=0}^{\infty} \hat{Q}_A(n) 1 \frac{e^{-\lambda t}(\lambda t)^n}{n!} \right)_i .$$

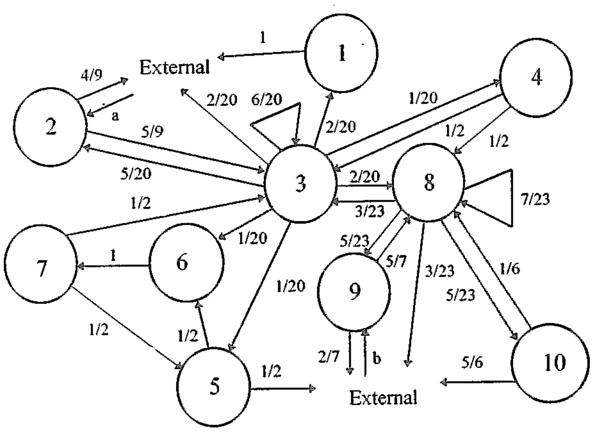
Since all the equations to derive delay distributions are represented in sum of discrete functions, we can compute them numerically.

5. Application to switching system

In this section, we apply the above procedure to an ISDN switching system and the CMS. Application to the ISDN switching system is to demonstrate the accuracy of the proposed method and application to the CMS is to check the system to meet the requirements for delay. For detailed description for application to ISDN switching system, refer to [6].

5.1. Sojourn time analysis in ISDN switching system

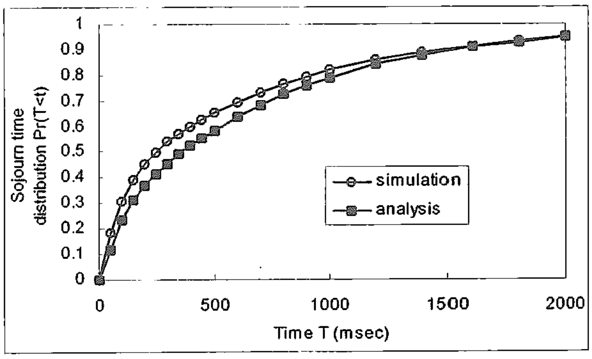
First, we apply the above procedure to the sojourn time analysis in the ISDN switching system. Sojourn time is the time from entering the system till exiting the system. Based on ISDN internal call processing scenario[6], we obtain a 10-node queueing network with a transition probability matrix as shown in Fig. 4.



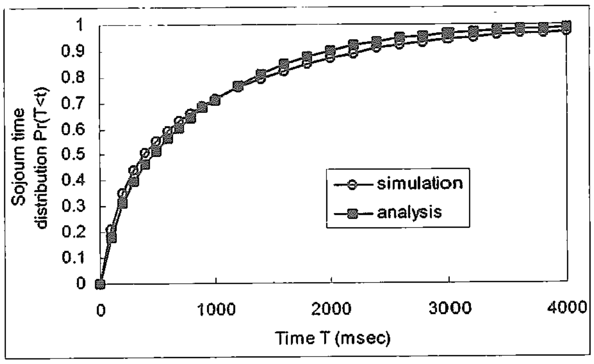
<Fig. 4 : Queueing network model for ISDN switching system>

The arrivals from outside are occurred at node 2 and node 9 following Poisson processes whose rates are given according to specific traffic condition. Service time distributions are assumed to be exponential distribution and Erlang-2 distribution. If real service time distribution is given, sojourn time distribution should be obtained according to it.

In Fig. 5 and 6 the computation results are given together with simulation results. We can see the proposed method gives the sojourn time distribution accurately for heavy traffic loads.



<Fig. 5 : Sojourn time distribution of ISDN switching system in utilization 0.8>



<Fig. 6 : Sojourn time distribution of ISDN switching system in utilization 0.9>

5.2. Delay analysis in CMS

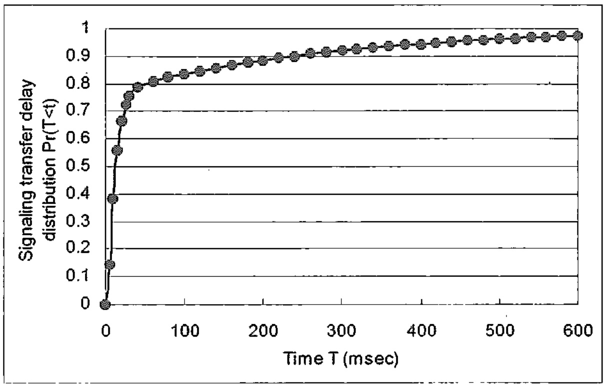
Next, we apply the procedure to the delay analysis in CMS. To check the system to meet the requirements for delay, we select two delay items, signaling transfer delay and connection release delay. Signaling transfer delay is defined as the interval from the instant that a message is received from a signaling system until the moment the corresponding message is passed to another signaling system[9]. And connection release delay is defined as the interval from the instant when RELEASE message is received from a signaling system until the instant when the connection is no longer available for use on the call and a corresponding RELEASE message is passed to the other signaling system involved in the connection[9].

In Fig. 2, delay items are represented with thick lines. Upper part corresponds to signaling transfer delay and lower part corresponds to connection release delay.

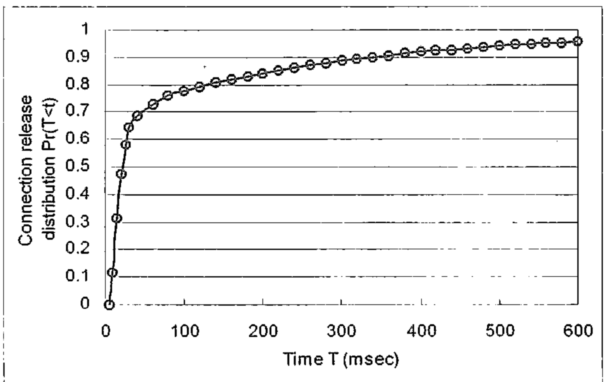
Table 1 shows the requirements for two delay items. As described in section 1, requirements for delay are recommended 95 percentile as well as mean.

Table 1. Requirements for delay (msec)

item	mean	95 percentile
signaling transfer delay	350	700
connection release delay	400	700



<Fig. 7 : Signaling transfer delay distribution>



<Fig. 8 : Connection release delay distribution>

The resultant queueing network for the system has eight nodes as shown in Fig. 3. The arrivals from outside are occurred at two nodes following Poisson processes and service time distribution at each node is given in [10]. Fig. 7 and 8 show the computational results in 500,000 BHCA(Busy Hour Call Attempt) of call processing capacity. In this case, Erlang-2 service time distributions are assumed.

From the figures, we can see the system meets the requirement for signaling transfer delay and connection release delay in 95 percentile as well as in mean.

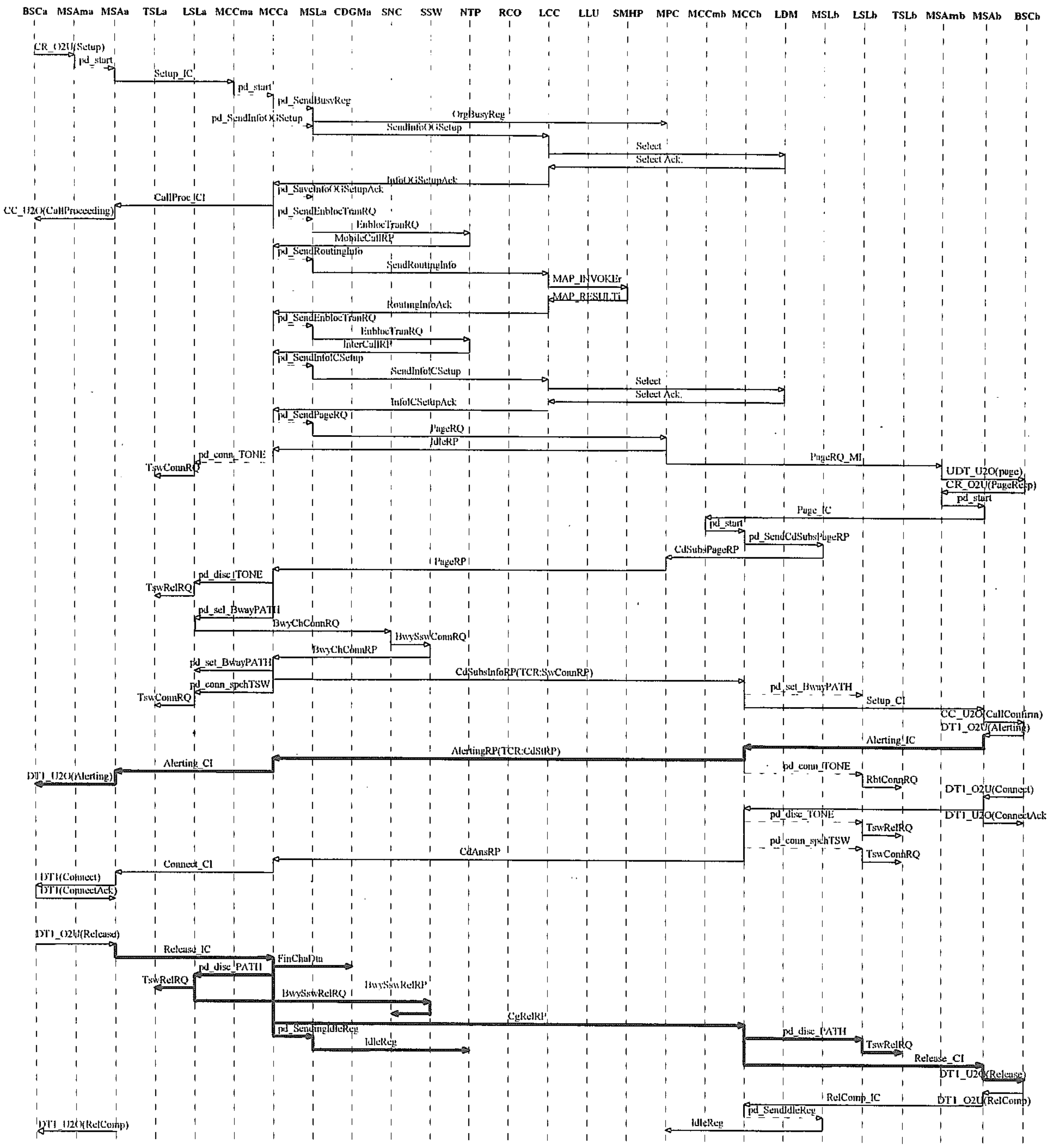
6. Conclusion

A computational method to analyze delay distributions in switching system has been proposed based on GPH semi Markov chain modeling. The methods to compute transition function and first transition time distribution in GPH semi Markov chains were established. The validity of the proposed method was confirmed by application results. And using the method, we showed that the CMS meets the requirements for delay. The proposed method seems to be conveniently applicable for delay analysis in switching system.

References

1. Neuts, M. F., *Matrix-Geometric Solutions in Stochastic Models : An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore, 1981.
2. Robertazzi, T. G., *Computer Networks and Systems : Queueing Theory and Performance Evaluation*, Springer-Verlag, 1992.
3. Ross, S. M., *Stochastic Process*, John and Sons, Inc., 1983.
4. Shanthikumar, J. G., "Bilateral Phase-Type Distributions," *Naval Research Logistics Quarterly*, vol.32, pp.119-136, 1985.
5. Whitt, W., "The Queueing Network Analyzer," *B. S. T. J.*, vol.62, pp.2779-2814, 1983.
6. Yoon, B. S., Baek, J. H. and Lie, C. H., "Analytic Method for Delay Time Distributions in Switching System," *Journal of Korean Institute of Communication Sciences*, vol.20, pp.1664-1677, 1995.
7. Yoon, B. S., Park, K. W. and Lie, C. H., "Probabilistic approximations based on GPH distributions;" *Journal of Korean Operations Reserach Society*, vol.19, pp.85-98, 1994.
8. Cho, K. S. and Sone, C. S., "TDX-10 CDMA digital mobile switching system," *KITE magazine*, Jan. 1994.

9. GSM System 03.05-Version 4.0.0.
10. Baek, J. H., Cho, K. S. and Lie, C. H., "Delay time analysis of CDMA mobile switching system by queueing network technique," *Proc. of KIE'95*, pp.444-449, 1995.



<Fig. 2 : Call processing scenario>