CONFIDENCE INTERVALS ASSOCIATED WITH PERFORMANCE ANALYSIS OF SYMMETRIC LARGE CLOSED CLIENT/SERVER COMPUTER NETWORKS

Vyacheslav Abramov

School of Mathematical Sciences, Monash University,
Clayton Campus, Building 28, Level 4, Wellington road,
Victoria 3800, Australia.

e-mail: vyacheslav.abramov@sci.monash.edu.au

Abstract

The paper studies a closed queueing network containing a server station and k identical client stations. The client stations are subject to breakdowns, and a lifetime of each client station is assumed to be a random variable independent of all other ones having the probability distribution $G(x)$. The server station is an infinite server queueing system, and client stations are single server queueing systems with autonomous service, i.e. every client station serves customers (units) only at random instants generated by strictly stationary and ergodic sequence of random variables. The total number of units in the network is N. The service times of units in the server station are independent exponentially distributed with parameter λ. The expected times between departures in client stations are $(N\mu)^{-1}$. After a service completion in the server station a unit is transmitted to the j th client station with equal probability $1/l$, where $l \le k$ is the number of currently available (i.e. not failure) client stations, and being processed in the j th client station the unit returns to server station. The parameter N is assumed to be large. The aim of this paper is to study the behaviour of bottleneck queues and to find confidence intervals associated with increasing a given high level of queue proportional to N in client stations.

Key words: Closed networks, Performance analysis, Normalized queue-length process, Confidence intervals

2000 Mathematical Subject Classification: 60K30, 60K25

1. Introduction

Consider a large closed queueing network containing a server station (infinite-server queueing system) and k identical single-server client stations. The total number of customers (units) is N, where N is assumed to be a large parameter. The departure process from client stations is assumed to be autonomous. Queueing systems with autonomous service mechanism have been introduced and originally studied by Borovkov [6, 7]. The formal definition of these systems in the simplest case of single arrivals and departures is as follows. Let $A(t)$ denote an arrival point process, let $S(t)$ denote a departure point process, and let $Q(t)$ be a queue-length process, and all these processes are started at zero $(A(0) = S(0) = Q(0) = 0)$. Then the autonomous service mechanism is defined by the equation:

$$Q(t) = A(t) - \int_0^{\infty} \mathrm{I}\{Q(s-) > 0\} dS(t).$$

The queueing systems with autonomous service mechanism have been studied in many papers (e.g. Abramov [1, 3, 4], Fricker [8, 9], Gelenbe and Iasnogorodski [10]). In the present paper we study a closed client/server network (see Figure 1).

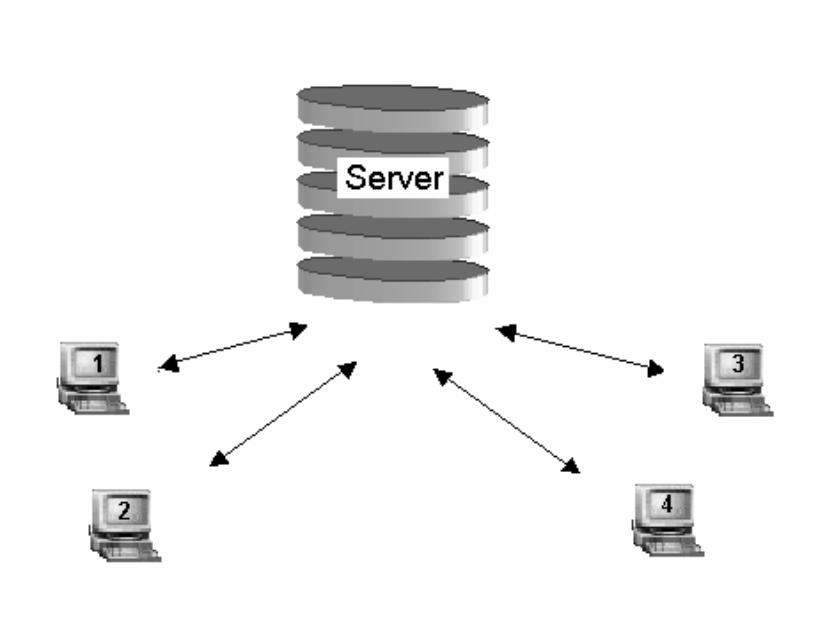

Figure 1. An example of client/server network

The arrival process from the server to the ith client station is denoted $A_{i,N}(t)$. The service time of each unit in the server station is exponentially distributed with parameter λ. Therefore, the rate of arrival to client stations depends on the number of units in the server station. If there is N_t units in the server station in time t, then the rate of departure of customers from the server in time t is λN_t. There are k client stations in total, and each of client station is a subject to breakdown. The lifetime of each client station is a continuous random variable independent of lifetimes of other client stations and has the probability distribution $G(x)$.

All client stations are assumed to be identical, and a unit transmitted from the server chooses each one with equal probability. (For this reason the network is called symmetric.) Therefore, if there are l available client stations in time t, then the rate of arrival to each of these client stations is $\lambda N_t / l$.

The departure instants from the jth client station $(j = 1,2,...,k)$ are $\xi_{j,N,1}, \xi_{j,N,1} + \xi_{j,N,2}, \xi_{j,N,1} + \xi_{j,N,2} + \xi_{j,N,3}, ...$ where each sequence $\xi_{j,N,1}, \xi_{j,N,2,}, ...$ forms a strictly stationary and ergodic sequence of random variables (N is the series parameter). The corresponding point process associated with departures from the client station j is denoted

$$S_{j,N}(t) = \sum_{i=1}^{\infty} \mathrm{I}\left(\sum_{l=1}^{i} \xi_{j,N,l} \le t \right),$$

and satisfies the condition

$$\mathrm{P}\left(\lim_{t\to\infty} \frac{S_{j,N}(t)}{t} = \mu N \right) = 1.$$

Then, the relations between parameters λ, μ and k are assumed to be

(1.1) $$\frac{\lambda}{k\mu} < 1,$$

and

(1.2) $$\frac{\lambda}{\mu} > 1.$$

Condition (1.1) means that all of the client stations are initially non-bottleneck, i.e. the service rate is greater than arrival rate. Condition (1.2) means that after one or other breakdown all of the client stations become bottleneck. Denote

$$l_0 = \max\left(l : \frac{\lambda}{l\mu} > 1 \right).$$

the maximum number of available client stations when the client stations all are bottleneck. Then for all $l \le l_0$ the rest l client stations will be bottleneck as well.

The queue-length process in the j th client station is defined as

$$Q_{j,N}(t) = A_{j,N}(t) - \int_0^t \mathrm{I}\{Q_{j,N}(s-) > 0\} dS_{j,N}(s),$$

where $A_{j,N}(t)$ is the arrival process to the client station j.

Let $\alpha < 1$ be a given positive number. We say that the network is at risk if the total number of units in client stations exceeds the value αN. Assuming that at the initial time moment all of the units are in the server station, the aim of this paper is to find a confidence interval $[0, \theta)$ such that with given high probability P (say $P = 0.95$) the network will not be at risk during that time interval $[0, \theta)$. For networks with an arbitrary number of client stations this problem is hard, because the behaviour of bottleneck queues is very complicated (see next section for explicit results). Therefore in the present paper we study this problem for the case of network with two client stations only.

A large closed client/server queueing network with bottlenecks has been studied in many papers. The bottleneck analysis of Markovian networks has been provided by Kogan and Liptser [11]. Abramov [1, 2, 3] has extended the results of [11] to non-Markovian networks. Specifically, [1] has studied the variant of network with autonomous service mechanism in client stations. The results of [1] have been then extended to networks with two types of node and multiple unit classes in [4]. However, the contribution of the aforementioned papers is purely theoretical. All of them provide the bottleneck analysis for the particular case of one bottleneck station and under the assumption that at the initial time moment all of the units are concentrated at the server station.

The detailed bottleneck analysis of the network including all cases related to bottleneck stations as well as initial conditions has been recently done in Abramov [5]. The results of [5] are promising for the solution of many applied problems. Specifically, the analysis of [5] is devoted to closed client/server networks in semi-Markov environment requiring the study of these networks under most general assumptions. The asymptotic solution of the problem of the present paper, as N increases indefinitely, is based on the study of [5].

It is worth noting that reliability of computer systems themselves has been studied in many papers. We refer the book of Xie, Dai and Poh [12], where a reader can find the detailed information related to this subject. The confidence intervals that are studied in the present paper are related to reliability of information, which heavily depends on reliability of the network.

The paper is motivated by significant practical problems in telecommunication systems. Support and exchange of information is very expansive and often increases the related costs of the equipment itself. On the other hand, reliable support of information is derivative from high reliability of equipment and directly depends on that reliability. A special circle of practical problems is related to support of large databases. Then “units" are associated with units of information (records), and “client stations" are associated with users of a database. “Failing station" can be associated with absence of connection or very low rate of exchange. Low exchange in certain stations can result in bottleneck of entire network leading to destruction of a database.

The paper is organized as follows. In Section 2 we recall some of the results of [5] which are necessary for our purpose and then adapt them to the case of symmetric network considered here. In Section 3 we derive the distribution of the normalized queue-length processes in available client stations. In Section 4 we establish results for confidence intervals in the particular case of two client stations. In Section 5 we give a simple numerical example. In Section 6 we conclude the paper.

2. Bottleneck client stations

In this section we recall some results of bottleneck analysis of [5] corresponding to the cases considered in the present paper. We start from the elementary case of l equivalent bottleneck stations exactly, i.e. the case that at the initial time moment $t = 0$ there are l bottleneck stations is discussed. For simplicity, assume that all of these l stations are absolutely reliable, and at the initial time moment $t = 0$ there are $(1-\beta)N$ units in the server station, $0 < \beta \le 1$, and the rest βN are distributed between l client stations. So, because the network is symmetric, the assumption that there are approximately $\beta N / l$ units in each client station in time $t = 0$, according to the law of large numbers, is reasonable. The assumption that the client stations are bottleneck means that $\lambda(1-\beta) > l\mu$.

The result on asymptotic behaviour of normalized queue-length in client stations follows from Proposition 5.3 of [5] which related to an asymmetric network with bottleneck stations and arbitrary initial queue-length. Recall this result.

Lemma 2.1. *Assume that all client stations are initially bottleneck, and the initial queue-lengths in client stations are asymptotically equivalent to* $N\beta_1, N\beta_2, ..., N\beta_k$ *correspondingly* $(\beta_1 + \beta_2 + ... + \beta_k \le 1)$, *as* $N \to \infty$. *Then, the normalized queue-length in the* j *th client station in limit as* $N \to \infty$ *is determined as*

$$(2.1) \qquad q_j(t) = \beta_j + \left(1 - \sum_{j=1}^{k} \beta_j\right)\left([\lambda_j(0) - \mu_j]t - \lambda_j(0)\int_0^t r(s)ds\right),$$

$$(2.2) \qquad r(t) = \left(\frac{\sum_{j=1}^{k}(\lambda_j(0) - \mu_j)}{\sum_{j=1}^{k}\lambda_j(0)}\right)\left(1 - \exp\left[-t\sum_{j=1}^{k}\lambda_j(0)\right]\right),$$

where $q_j(t)$ denotes the normalized queue-length process in the j th client station in limit, i.e. $q_j(t)$ is the limit in probability of $Q_{j,N}(t)/N$ as N increases indefinitely.

In the notation of this lemma $\lambda_j(0)N$ denotes the instantaneous rate of units to the j th client station in time $t = 0$, and $\mu_j N$ denotes the service rate in the j th client station. In our particular case the number of nodes is l, the instantaneous rate of units to each client station is $(1-\beta)\lambda N/l$ and the service rate is μN and all $q_j(t)$ are equal, i.e. $q_j(t) \equiv g(t)$ for all $j = 1,2,...,l$. Therefore in our case from this Lemma 2.1 we have the following statement.

Proposition 3.2. *We have:*

$$(2.3) \qquad g(t) = \frac{\beta}{l} + (1-\beta)\left(\left[\frac{(1-\beta)\lambda}{l} - \mu\right]t - \frac{(1-\beta)\lambda}{l}\int_0^t r(s)ds\right),$$

where

$$(2.4) \qquad r(t) = \left(1 - \frac{\mu l}{(1-\beta)\lambda}\right)\left(1 - e^{-(1-\beta)\lambda t}\right).$$

3. Limiting normalized cumulative queue-length process

In this section we study the limiting (as $N \to \infty$) normalized cumulative queue-length process in client station. The limiting normalized cumulative queue-length process is denoted $q(t)$. At the initial time $t = 0$ there are k available client stations. Let $\tau_1, \tau_2, ..., \tau_k$ be the moments of their breakdown, $0 \le \tau_1 \le \tau_2 \le ... \le \tau_k$. The above moments of breakdown are associated with the behaviour of the time-dependent network which can be considered as a network in semi-Markov environment. Therefore one can apply Theorem 5.4 of [5].

The random time interval [0, τ_k] is the lifetime of the entire system. Therefore $q(t)$ is to be considered during the aforementioned random interval [0, τ_k]. Recall that $l_0 = \max(l : \lambda/(l\mu) > 1)$. Therefore, according to Theorem 5.4 of [5] we obtain that in the random interval [0, τ_{k-l_0}), $q(t) = 0$. Next, in the random interval [$\tau_{k-l_0}, \tau_{k-l_0+1}$) the equation for $q(t)$ is

$$(3.1)\qquad q(t)=(\lambda-l_0\mu)(t-\tau_{k-l_0})-\lambda\int_{\tau_{k-l_0}}^{t}r(s)ds,$$

where $r(t)$ is given by (2.4). Equation (3.1) follows from (2.1) and (2.2) as follows. Setting $\beta=0$, $l=l_0$ and replacing t with $(t-\tau_{k-l_0})$ from (2.1) we obtain:

$$g(t)=\left(\frac{\lambda}{l_0}-\mu\right)(t-\tau_{k-l_0})-\frac{\lambda}{l_0}\int_{\tau_{k-l_0}}^{t}r(s)ds.$$

Hence, taking into account that $q(t)=l_0g(t)$ we arrive at (3.1).

In the next interval $[\tau_{k-l_0+1},\tau_{k-l_0+2})$, $l_0>1$, we have the following equation:

$$(3.2)\qquad g(t)=\frac{q(\tau_{k-l_0+1})}{l_0-1}+[1-q(\tau_{k-l_0+1})]\left\{\left[\frac{[1-q(\tau_{k-l_0+1})]\lambda}{l_0-1}-\mu\right](t-\tau_{k-l_0+1})\right.$$
$$\left.-\frac{[1-q(\tau_{k-l_0+1})]\lambda}{l_0-1}\int_{\tau_{k-l_0+1}}^{t}r(s)ds\right\}.$$

Therefore, in the time interval $[\tau_{k-l_0+1},\tau_{k-l_0+2})$

$$(3.3)\qquad q(t)=q(\tau_{k-l_0+1})+[1-q(\tau_{k-l_0+1})]\{([1-q(\tau_{k-l_0+1})]\lambda-\mu(l_0-1))(t-\tau_{k-l_0+1})$$
$$-[1-q(\tau_{k-l_0+1})]\lambda\int_{\tau_{k-l_0+1}}^{t}r(s)ds\}.$$

In an arbitrary time interval $[\tau_i,\tau_{i+1})$, $i=k-l_0,k-l_0+1,...,k-1$, we have:

$$q(t)=q(\tau_i)+[1-q(\tau_i)]\{([1-q(\tau_i)]\lambda-\mu(k-i))(t-\tau_i)$$
$$-[1-q(\tau_i)]\lambda\int_{\tau_i}^{t}r(s)ds\}.$$

In the last endpoint τ_k we set $q(t)=1$.

4. Confidence intervals

The formulae for the limiting normalized cumulated queue-length process are complicated. Therefore in this section we obtain confidence intervals for the particular case of two client stations. In this case only simple representation (3.1) is used, which in the case of two client stations looks as follows:

$$(4.1) \qquad q(t) = (\lambda - \mu)(t - \tau_1) - \lambda \int_{\tau_1}^{t} r(s)ds,$$

where

$$(4.2) \qquad r(s) = \left(1 - \frac{\mu}{\lambda}\right)\left(1 - e^{-\lambda t}\right).$$

The confidence interval is structured from two intervals. The first one is $[0,\ \tau_1)$, where the limiting normalized cumulative queue-length is zero. The second interval is $[\tau_1,\ \theta]$, where θ is a point where $q(\theta) \le \alpha$. Equations (4.1) and \(4.2) are defined for $t < \tau_2$, where τ_2 is a random breakdown point of the second client station.

Let θ^* be a point where $q(\theta^*) = \alpha$. The point θ^* is a random point depending on τ_1. However, under the assumption that one or other client station is active, the length of the interval $[\tau_1, \theta^*]$ is fixed and uniquely defined from (4.1) and (4.2).

Let us derive probability distribution of the process $q(t)$. Clearly, that the probability that $q(t) = 0$ coincides with the probability that the length of the interval $[0, \tau_1)$ is greater than t. Therefore,

$$(4.3) \qquad \mathrm{P}\big(q(t) = 0\big) = [1 - G(t)]^2.$$

Next,

$$(4.4) \qquad \mathrm{P}\big(q(t) \le \gamma < 1\big) = [1 - G(t)][1 - G(t - t_\gamma)],$$

where t_γ is such the value of t under which

$$(4.5) \qquad (\lambda - \mu)t - \lambda \int_{0}^{t} r(s)ds = \gamma.$$

Equations \(4.3) and (4.4) are easily obtained by standard arguments of probability theory.

Then the probability that the limiting normalized cumulated queue will reach the value γ before absorbing at 1 is

$$\frac{\int_0^\infty [1 - G(t)][1 - G(t - t_\gamma)]dt}{\int_0^\infty [1 - G(t - t_\gamma)]^2 dt}.$$

The problem is to find the value $\gamma \le \alpha$ such that

$$\frac{\int_0^\infty [1-G(t)][1-G(t-t_\gamma)]dt}{\int_0^\infty [1-G(t-t_\gamma)]^2 dt} \geq P. \quad (4.6)$$

It is written the inequality rather then equality because the exact equality can be reached for $\gamma > \alpha$, while for all $\gamma \leq \alpha$ there must be inequality (4.6).

5. Numerical example

Consider the following example. Let $\lambda = 4,\ \mu = 3,\ \alpha = 0.2,\ P = 0.95,\ G(x) = 1 - e^{-2x}$.

From (4.6) we have:

$$\frac{\int_0^\infty e^{-2(t-t_\gamma)} e^{-2t} dt}{\int_0^\infty e^{-4(t-t_\gamma)} dt} = e^{-2t_\gamma}.$$

Solution of the equation $e^{-2t_\gamma} = 0.95$ yields $t_\gamma = 0.025647$. From (4.5) we obtain:

$$\gamma = \int_0^{t_\gamma} e^{-4t} dt = \int_0^{0.025647} e^{-4t} dt = 0.25 - 0.25e^{-0.102588} = 0.024375.$$

This value of γ is less than $\alpha = 0.2,$ and therefore this value $\gamma = 0.024375$ is the required value for a confidence interval.

6. Concluding remarks

We found confidence intervals associated with increasing a given high level. The confidence intervals that established in the present paper are random. They are obtained in terms of the parameter γ, which is the value of limiting cumulative normalized queue-length under which the probability that the system will be active is not smaller than a given value *P*. Thus, the strategy is to observe the cumulative queue-length process until the total number of units in client stations reaches the value γN. As soon as the total number of units exceeds this value there is not enough confidence that the system and/or information will be available.

Acknowledgement

The research was supported by Australian Research Council, grant # DP0771338.

References

[1] **V. M. Abramov,** 2000. A large closed queueing network with autonomous service and bottleneck. *Queueing Systems*, 35, 23-54.

[2] **V. M. Abramov,** 2001. Some results for large closed queueing networks with and without bottleneck: Up- and down-crossings approach. *Queueing Systems*, 38, 149-184.

[3] **V. M. Abramov,** 2004. A large closed queueing network containing two types of node and multiple customers classes: One bottleneck station. *Queueing Systems*, 48, 45-73.

[4] **V. M. Abramov,** 2006. The effective bandwidth problem revisited. arXiv: math. 0604182.

[5] **V. M. Abramov,** 2007. Large closed queueing networks in semi-Markov environment and their application. arXiv: math. 0612224.

[6] **A. A. Borovkov,** 1976. *Stochastic Processes in Queueing Theory.* Springer-Verlag, Berlin.

[7] **A. A. Borovkov,** 1984. *Asymptotic Methods in Queueing Theory.* John Wiley, New York.

[8] **C. Fricker,** 1986. Etude d'une file GI/G/1 á service autonomé (avec vacances du serveur). *Advances in Applied Probability,* 18, 283-286.

[9] **C. Fricker,** 1987. Note sur un modele de file GI/G/1 á service autonomé (avec vacances du serveur). *Advances in Applied Probability,* 19, 289-291.

[10] **E. Gelenbe and R. Iasnogorodski,** 1980. A queue with server of walking type (autonomous service). *Ann. Inst. H. Poincare*, 16, 63-73.

[11] **Ya. Kogan and R. Sh. Liptser,** 1993. Limit non-stationary behaviour of large closed queueing networks with bottlenecks. *Queueing Systems*, 14, 33-55.

[12] **M. Xie, Y. S. Dai and K. L. Poh,** 2004. *Computer Systems Reliability: Models and Analysis.* Kluwer, Dordrecht.