연세경제연구 제 Ⅲ 권, 제 2호
Yonsei Economic Studies, Vol. Ⅲ, No. 2 (September 1996)

Optimal Number of Customers and Buffer Allocation in a Two-Node Cyclic Network

Byung Chun Park · Robert D. Foley

We deal with network design problems for a two-node cyclic queueing network with blocking. The network we consider has multiple general servers at one node and multiple exponential servers at the other node. We show that there is a single queue which is stochastically equivalent to the network and hence the throughput of the network is determined as that of this equivalent queue. Under certain conditions, we show there are very simple rules to determine the optimal allocation of buffer capacity and the optimal number of customers which maximize throughput of the network.

Ⅰ. Introduction

Suppose that we have a cyclic queueing network with a single exponential server with rate λ, and s_g general servers with i.i.d. service times with distribution $G(\cdot)$. If we have a fixed amount of buffer capacity, what is the best way to allocate the buffer capacity between the two nodes to maximize throughput? Similarly, what number of customers would give the highest throughput?

At first glance, the answers to these questions would seem to depend on the relative speeds of the servers. However, the best buffer allocation and the best number of customers are unaffected by the speeds of the servers. We will look at a slightly more general system, but for the system described above, our approach would reduce to showing that the throughput in this network is the same as the throughput of an $M/GI/s/c(k,b_e,b_g)$,

Department of Industrial Engineering, Keimyung University, Taegu, 704-701, Korea. School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

i.e., a queue with arrival rate λ, service time distribution $G(\cdot)$, s servers where s is the minimum of S_g and the number of customers in the network k, and buffer capacity $c(k,b_e,b_g)$ where b_e and b_g are the buffer capacities at the exponential and general nodes, respectively. So to maximize the throughput of the network, we need only maximize the throughput of the equivalent $M/GI/s/c(k,b_e,b_g)$ queue. Since, the throughput of this equivalent queue is an increasing function of the buffer capacity, the buffer allocation problem and the number of customers problem reduce to selecting k, b_e, b_g to maximize $c(k,b_e,b_g)$.

A two-node cyclic queueing network model has widely been used in the performance evaluation of computer systems, production lines, and so on. In particular, this model has been used for evaluating multiprogramming systems consisting of a central processing unit and a secondary memory or input-output device (Gelenbe [4]).

Boxma [3] derived an expression for the cycle time distribution in two-node cyclic queueing systems with one general and one exponential server, assuming an infinite buffer. Akyildiz [1] studied two-node cyclic queueing systems with blocking, where both stations have generally distributed service times. In Akyildiz [1], a station's server is represented by a Cox-model with 2 phases, and then iterative procedures are used to obtain an approximate solution. Bocharov [2] studied a two-node closed queueing network with finite capacity, where the service times of customers at both nodes have phase type distributions. Sparaggis and Gong [11] studied optimal buffer allocation problems in a two-node queueing network with blocking, where both nodes have a general server.

All the above models assumed a single server at both nodes. Instead, we assume multiple servers. Given a network, we show there is a stochastically equivalent queue and the throughput of the network is determined by the throughput of this queue. Using the equivalence relation between the network and the queue, we show, under certain conditions, there are very simple rules to determine the optimal number of customers and the optimal buffer allocation in the network.

For convenience we will use $x \wedge y$ to denote the minimum of x and y, $x \vee y$ to denote the maximum of x and y, and $\lfloor x \rfloor$ and $\lceil x \rceil$, to denote the floor and ceiling functions of x.

Ⅱ. Network Description

The network that we will analyze is a two node, closed queueing network with k customers. Since we assume finite buffers, customers may be blocked. In particular, we assume "production blocking." That is, if a customer completes service at a node and there is no available space at the next node, the customer remains at the service position and prevents the server from working on other customers until space becomes available. A customer is considered to be in the queue at a node, if its next service completion will be at that node.

One of the two nodes is a $\cdot/G/s_g/b_g$ queue with s_g servers and a buffer space of b_g consisting of the S_g servers and $b_g - s_g$ waiting spaces. The service times at this node form a sequence of non-negative variables $S_1, S_2, \cdots$, and we do not make any assumptions about the probabilistic structure of this sequence at this time. S_n will be the service time of the nth customer to start service after time 0. We will refer to this node as the general node. The other node is a $\cdot/M_Q/s_e/b_e$ queue with s_e servers, buffer size b_e, state dependent service rates, i.e., the departure rate from this node is conditionally independent of the past given the number of customers awaiting service at this node. This node will be referred to as the exponential node.

When there are i customers in the exponential node, service completions in the exponential node occur at rate λ_i for $[0 \vee (k-b_g-s_e)] \leq i \leq [k \wedge (b_e+s_g)]$. While the exponential node is empty or blocked, the exponential server is not working so $\lambda_{0 \vee (k-b_g-s_e)} = 0$. Otherwise, we assume $\lambda_i > 0$.

Note that in some cases it is possible to have up to $b_g + s_e$ customers at the general node since there may be S_e blocked customers at the exponential node. Similarly, there may be up to $b_e + s_g$ customers in some cases at the exponential node. Also note that 3 identical exponential servers each with service rate λ and no waiting space is not the same as a single server queue with two waiting spaces and state dependent service rate $\lambda_i = \lambda i$ since up to 3 customers may be blocked in the first system, but only one in the second.

Due to physical considerations and to avoid degeneracies, we assume throughout;

(1) s_e, s_g, b_e, b_g and k are positive integers,

(2) s_e, $s_g \geq 1$

(3) $b_g \geq s_g$ and $b_e \geq s_e$,

(4) $0 < k < b_e + b_g$.

Ⅲ. The $M_Q/G/s/c$ Queue

Consider a queue with s servers and $c-s$ waiting spaces. Hence, there is at most c customers in the system. Assume that we have a work conserving queueing discipline. For simplicity, assume that customers are not pre-emptied. Let $Y(t)$ denote the queue length process, and we assume that $Y(0) = 0$. To construct the queue length process, we use the same sequence of service times $S_1, S_2, \cdots$ as in the previous section, and let S_n be the service time of the nth arrival.

Assume that arrival process to the system is a doubly stochastic Poisson process where the arrival rate at time t is $\lambda_{k-r^{-1}(Y(t))}$, i.e., determined by the number in the system and the function $r(\,.\,)$ which will be specified later. Thus, given $Y(t) = i$ customers, the time until the next potential arrival after time t is exponentially distributed with parameter λ_{k-i} and conditionally independent of $\{Y(s);\ s \leq t\}$. Of course, if there is a departure before this potential arrival occurs, the arrival rate changes to λ_{k-i+1}. Using the service time sequence and this property of the arrival process, the queue length process Y for the $M_Q/G/s/c$ queue can be constructed.

We will let $N(t)$ denote the number of departures during $(0,t]$. Let $TPT(c)$ denote the system throughput of the $M_Q/G/s/c$ queue provided it exists. More precisely,

$$TPT(c) = \lim_{t \to \infty} N(t)/t$$

provided the limit exists. Note that we have not made any assumptions about the probabilistic structure of the service times so it is easy to construct examples where the $TPT(c)$ is undefined. If we assume that the service times are independent and identically distributed, we will denote this by $M_Q/GI/s/c$. If we also assume that the arrival process is a Poisson process, we will denote this by $M/GI/s/c$.

Ⅳ. Equivalence between the Network and Queue

We will exploit the fact that the Network and the Queue are stochastically equivalent. We start with the definition of "r-equivalent."

Definition 1 *Let* $X = \{X(t);\ t \geq 0\}$ *and* $Y = \{Y(t);\ t \geq 0\}$ *be two stochastic processes. We will call Y r-reducible to X if there exists a function* $r(\,.\,)$ *such that* $\{r(X(t));\ t \geq 0\}$ *is stochastically indistinguishable from Y. If the function r is invertible, then X is also r-reducible to Y and we call this r-equivalent.*

Remark 1 *The states of X are merely a relabelling of the states of Y via the function r, and results for one process can be translated into results for the other process.*

Now we state the main results.

Theorem 1 *For any fixed k,* b_e, b_g, s_e *and* s_g, *the stochastic process X representing the number of customers at the general node of the network is r-equivalent to Y representing the queue length process in an* $M_Q/G/s/c$ *queueing system, where*

$$s = s_g \wedge k$$

and

$$c = k \wedge (b_e + s_g) \wedge (b_g + s_e) \wedge (b_e + b_g + s_e + s_g - k). \qquad (1)$$

Furthermore, the number of service completion process at the general node of the network is stochastically identical to $\{N(t);\ t \geq 0\}$, *the departure process from the* $M_Q/G/s/c$ *queue.*

Remark 2 *This equivalence was proven in Lavenberg* [6], *when both nodes had single servers, one with exponential service times, the other with i.i.d. service times, and* $k \leq b_e$. *In this case, the re-labelling function* $r(\,.\,)$ *is just the identity function.*

Proof Let $X(t)$ be the number of customers in the general node at time t. As mentioned

before, $X(t)$ includes any blocked customers at the exponential node.

We will construct a sample space where the network process X and the queue length process Y have their specified distributions and for all $t \geq 0$, $r(X(t)) = Y(t)$. Thus, the two processes will be r-equivalent. The constructed processes will be right-continuous with left-hand limits.

Fix k, b_e, b_g, s_e and s_g. Define

$$r(i) = i - m_g \text{ for } i = m_g, \cdots, m_g + c$$

where $m_g = [\,0 \vee (k - b_e - s_g)]$ which is the least number of customers in the general node.

Now define $Y(0) = 0$, and $X(0) = m_g$; hence, neither has a customer in service. In the queue, $Y(0)$ jumps up to state one at the time of the next arrival which is exponentially distributed with parameter $\lambda_{k - r(0)}$. Similarly, in the network, if there are m_g customers waiting for the general server, the other $k - m_g$ are waiting for the exponential server, so customers are leaving the exponential server at rate λ_{k-m_g}. Since, $r^{-1}(0) = m_g$, we can have the first arrival occur at the same time A_1 in each process. Each system will have a customer begin service at time A_1 and their service time will be S_1 (If $m_g > 0$, the customer that begins service is not necessarily the arrival, since one of the blocked customers may now begin service). Thus, the general node will have a service completion and the queue will have a departure at exactly the same time $A_1 + S_1$. So assuming $S_1 > 0$, we have $X(A_1) = m_g + 1$, $Y(A_1) = 1$, and $r(m_g + 1) = 1$. Furthermore, in this state, both processes have arrivals occurring at the same rate λ_1. Hence, whichever occurs first, an arrival or departure, can be constructed to occur at the same time in each process. If an arrival occurs first, the arrival will occur to each system simultaneously and the second customer to start service in each system receives the same service time S_2. Since the number of servers in the general node and the queue are the same, either both wait or both have a customer start service. Continuing in this fashion, the processes can be constructed as claimed.

Note that a service completion at the general node occurs at time t if $X(t) - X(t-) = 1$. In the r-equivalent queue, we have the state transition $r(X(t)) - r(X(t-))$ which is also 1, i.e., a departure from the queue. So the service completion process from the node on this sample space is identical to the departure process from the queue. ■

Ⅴ. Network Design to Maximize Throughput

In this section we look at the problem of allocating buffers and determining the optimal number of customers in the network. The following corollaries will hold when $TPT(c)$, the throughput of the queue, is non-decreasing in c. Sonderman [10] has shown that $TPT(c)$ is non-decreasing in c for $\cdot$ */GI/s/c* queues with i.i.d. service times and an arrival process which is independent of the service times but otherwise arbitrary. We will call a cyclic queueing network with a single server in the exponential node and multiple servers with i.i.d. service times in the general node, a *Sonderman network*.

First, we consider the problem of allocating the buffer capacity. Assume that the number of servers at each node, the number of customers, and the combined buffer capacity b are given. The following corollary determines an optimal way of allocating the buffer capacity to maximize the network throughput. Let $w_g = b_g - s_g$ be the waiting spaces at the general node, $w_e = b_e - s_e$ at the exponential node, and $w = w_g + w_e$ be the combined number of waiting spaces.

Corollary 1 *If $TPT(c)$ is a non-decreasing function of c, then an optimal solution to maximize the throughput of the network is to split the waiting spaces as evenly as possible between the exponential and general nodes; i.e., $w_e = \lfloor W/2 \rfloor$ and $w_g = w - w_e$. In particular, the above holds for Sonderman networks.*

Proof To prove this, note that the claimed values for w_e and w_g maximize (1). To show that the results hold for the Sonderman network, simply apply Theorem 1 and Corollary 2(a) of Sonderman [10] to note that the throughput of an *M/GI/s/c* queue is a non-decreasing in c. ■

Remark 3 *In a sense, the service positions are also buffer positions that are shared by both nodes. The optimal allocation of the waiting spaces is a minimax allocation to maximize the minimum number of customers that can be awaiting service at the two nodes.*

Corollary 2 *Even if the number of waiting positions at the two nodes of the network are interchanged, the throughput does not change.*

Proof Note that (1) is unchanged even if the values of w_g and w_e are interchanged. ■

Next we determine the number of customers k in the network which maximizes system throughput given the number of servers and the buffer allocation. The following corollary gives the optimal number of customers under the monotonicity hypothesis.

Corollary 3 *If TPT(c) is a non-decreasing function of c, then any k satisfying*

$$s_e + s_g + (w_e \wedge w_g) \leq k \leq s_e + s_g + (w_e \vee w_g) \qquad (2)$$

maximizes the network throughput. In particular, the above holds for Sonderman networks.

Proof To prove this, simply note that any k satisfying (2) maximizes (1) and that $TPT(c)$ is non-decreasing for Sonderman networks. ■

Corollary 4 *The throughput of the network is symmetric in k around $s_e + s_g + w/2$ in the region where the number of customers in the network satisfy the non-degeneracy conditions.*

Proof Note that (1) is symmetric in k around $s_e + s_g + w/2$. ■

Lastly, let us simultaneously determine the optimal number of customers in the network and the optimal buffer allocation between the two nodes assuming a fixed capacity $b = b_e + b_g$, s_e servers at the exponential node, and s_g servers at the general node. Since the optimal solution to the buffer allocation problem did not depend on the number of customers, the next corollary follows immediately from the previous corollaries.

Corollary 5 *If TPT(c) is a non-decreasing function of c, then $w_g = \lfloor W/2 \rfloor$ or $w_g = \lceil W/2 \rceil$, and $k = s_e + s_g + \lfloor W/2 \rfloor$ or $k = s_e + s_g + \lceil W/2 \rceil$ are solutions that maximize the network throughput. In particular, the above holds for Sonderman networks.*

Remark 4 *If TPT(c) is strictly increasing in c, the above three corollary gives the only optimal solutions.*

Remark 5 *The results of Sonderman* [10] *cannot be used directly on M_Q/GI/s/c queues since Sonderman* [10] *assumes that the sequence of potential arrivals is independent of the service process.*

❖ REFERENCES ❖

1. Akyildiz, I. F., "General Closed Queueing Networks with Blocking," *Performance '87*, 1988, North-Holland.
2. Bocharov, P. P., "On Two-Node Queueing Networks with Finite Capacity," *Proceedings of First International Workshop on Queueing Networks with Blocking*, 1988, North Carolina State University, pp. 239~257.
3. Boxma, O. J., "The Cyclic Queue with One General and One Exponential Server," *Advances in Applied Probabilty*, 15, 1983, pp. 857~873.
4. Gelenbe, E., "On Approximate Computer System Models," *Journal of ACM*, 22, 1975.
5. Gordon, W. J. and Newell, G. F., "Cyclic Queueing System with Restricted Length Queues," *Operations Research*, 15, 1967.
6. Lavenberg, S. S., "The Steady-State Queueing Time Distribution for the M/G/1 Finite Capacity Queue," *Management Science*, 21, 1975, pp. 501~506.
7. Onvural, R. O. and Perros, H. G.(1986), "On Equivalencies of Blocking Mechanisms in Queueing Networks with Blocking," *Operations Research Letters*, 5, 1986, pp. 293~298.
8. Shanthikumar, J. G., "Monotonicity Properties in Cyclic Queueing Networks with Finite Buffers," *Performance '87*, 1988, North-Holland.
9. _____ , and Yao, D. D., "Monotonicity and Concavity Properties in Cyclic Queueing Networks with Finite Buffers," *Queueing Networks with Blocking*, H. G. Perros and T. Altiok, Elsevier (eds.), 1989, North-Holland.
10. Sonderman, D., "Comparing Multi-Server Queues with Finite Waiting Rooms, I: Same Number of Servers," *Advances in Applied Probability*, 11, 1979, pp. 439~447.
11. Sparaggis, P. D. and Gong, W., "Optimal Buffer Allocation in a Two-Stage Queueing System," *Journal of Applied Probability*, 30, 1993, pp. 478~482.