Journal of the Korean Institute of Industrial Engineers
Vol. 17, No. 2, Dec., 1991

An Optimal Threshold Control in an Open Network of Queues

Kim, Sung Chul

개방대기 네트웍에서의 최적 Threshold 제어

김 성 철

Abstract

This article develops a control model for an open queueing network in terms of both the input and the output processes with stochastic intensities. The input and the output intensities are subject to some capacity limits and optimum control is characterized by a threshold type with a finite upper barrier. A discounted profit is used as a decision criteria, which is revenue minus operating and holding cost.

1. Introduction

The threshold type control is widely applied in practice. As an illustrative example, consider a production system. The input process is characterized by the output process of an upstream stage that feeds semi-products into the system and the output process models the production of the system. If the number of jobs in the system reaches the upper limits of the job population, the arriving job is blocked and is either assumed to proceed to a different production system or simply be turned off. Also the number of jobs in the system reaches the lower limit, the production of the system is temporarily suspended and the output process is turned off. And the number of jobs in the system is between the lower and the upper limits, then both the input and the output processes run at their maximum allowed capacity limits. This blocking type of control is characterized by the threshold control.

This article develops a control model for an open queueing network in terms of both the input and the output processes with stochastic intensities.

Duk Sung Women's University

The input and the output intensities are subject to some capacity limits and the optimum control is characterized by a threshold type : the input intensity drops to zero whenever the job population level reaches the upper barrier, otherwise both the input and the output go at their full capacity limits, that is, the lower control limit is zero.

As related work, Robertazzi and Lazar(1985) studied the control of the input process in Jackson (1963) network. The objective is the maximization of the average throughput of the network subject to a bounded average time delay and they showed that there exists an optimal control of the threshold type. Li(1988) studied both the input and the output control problem in a simple queueing system where the input and the output processes are conditional poisson. The optimality of the threshold type control was established under the condition of the constant capacity limits on the birth-death rates. The objective was the maximization of the long-run total discounted profit, which is the revenue minus the sum of the operating cost and the holding cost. In Chen and Yao(1990), the results are more general and specific. They extended Li (1988)' s result by allowing the input and the output processes to be general point processes with their stochastic intensities and state-dependent capacity limits. And they established conditions and restrictions on the capacity limits for the existence of the optimal and finite threshold.

In this work, we provide theoretical justification for the existence of the optimal threshold control for an open queueing network. A discounted profit is used as a decision criteria, which is revenue minus operating and holding cost.

In section 2, we start with the formulation of an intensity control poblem in an open queueing network. In section 3, we focus on a single-stage queueing system and characterize the conditions for the existence of the optimal threshold control, which will serve as a prelude to the main results in section 4. In section 4, we establish the existence of an optimal threshold control and identify a finite optimal upper barrier b^* in open queueing networks. Finally, Section 5 concludes the paper with some brief remarks.

2. Model Formulation

Consider an open queueing network(OQN) where the input and the output processes are general point processes with their stochastic intensities and state-dependent capacity limits. There are M stations($\mathcal{U}=\{1, \cdots, M\}$) with first-come-first-served queue discipline in the network. For each node in the OQN, define the queue length at node i, n_i ($i \in \mathcal{U}$), as the number of jobs in it (including both jobs in queue and jobs in service). Let the state of the system be the number of jobs in the network(N), which is the sum of the queue lengths at nodes, i. e., $N=(\underline{n})=\Sigma_{i=1}^{M} n_i$. External arrivai follows a poisson process with the state-dependent arrival rates, $\lambda(N)(N=0, 1, 2, \cdots)$. The arrival rate is controllable and satisfies the restrictions, $\lambda(N) \leq \hat{\lambda}(N)$ $(N=0, 1, 2, \cdots)$, where $\hat{\lambda}(N)$ represents the given state-dependent capacity limits on the arrival rates at the system.

An arrived job will first join station j with probaility $\gamma_{0j}(j=1, \cdots, M)$. Any job after being processed at station i will leave the system with probability $\gamma_{i0}(i \in \mathcal{U})$. Job routing within the system is governed by the transition matrix $[\gamma_{ij}]$ $(i \in \mathcal{U}, j \in \mathcal{U})$ which is sub-stochastic. The traffic equations are

$v_i = \gamma_{0i} + \Sigma_{j=1}^{M} v_j \gamma_{ji}$ $(i \;\; \mathcal{U})$, and $v_i (i \in \mathcal{U})$ can be interpreted as the average number of visits to station i.

For each node, the service rates are conditional poisson, Let $\mu_i(n_i)$ be the service rate given that the queue length is n_i; let $\mu_i(0)=0$ $(i \in \mathcal{U})$. For $n_i > 0$, the service rates satisfy : $0 \leq \mu_i\ (n_i) \leq \hat{\mu}_i(n_i)$ $(i \in \mathcal{U}$ and $n_i = 1, 2, \cdots)$, where $\hat{\mu}_i(n_i)$ is given upper limit (positive and finite) on the service rates at node i. That is, the service rates at each node is controllable in a decentralized manner. Assume that all nodes have ample waiting room, so that no job will be blocked at any node. And a control v is said to be admissible, when the service rates satisfy : $0 \leq \mu_i\ (n_i) < \hat{\mu}_i(n_i)$ $(i \in \mathcal{U}$ and $n_i = 1, 2, \cdots)$ and the arrival rates satisfy : $0 \leq \lambda(N) \leq \hat{\lambda}\ (N)$ $(N = 0, 1, 2, \cdots)$.

Let $N(t)(t \geq 0)$ be the state of a network at time t, taking values on non-negative integers, $A(t)$ $(t \geq 0)$ be the cumulative number of arrivals into the network in $[0, t]$ and $D(t)(t \geq 0)$ be the cumulative number of departures (throughputs) from the network in $[0, t]$, where both $A(t)$ and $D(t)$ are represented by the increasing nonnegative integers in t. Then

$$N(t) = N(0) + A(t) - D(t) \quad \cdots\cdots\cdots\cdots\cdots\cdots \ (2.\ 1)$$

where $N(0) \geq 0$ is the initial state.

Let p be the revenue due to output, c be the operating cost for input, and h be the holding work-in-process inventory cost for an unit. Then the objective is to maximize the following discounted value function given a control v

$$V^v(x) = E_x^v \int_0^\infty e^{-rt}\ (pd\ D\ (t) - cdA(t) - hN(t)dt) \quad \cdots\cdots\cdots\cdots\cdots\cdots \ (2.\ 2)$$

where E_x^v denotes the conditional expectation under control v given $N(0) = x$, and $\gamma > 0$ is the discounted factor.

Substituting equation(2, 1) and integrating by parts the equation (2, 2) simplifies to

$$V^v(x) = E_x^v \int_0^\infty e^{-rt}\ [\hat{p}d\ D\ (t) - \hat{c}dA(t)] - hx/\gamma \quad \cdots\cdots\cdots\cdots\cdots\cdots \ (2.\ 3)$$

where $\dot{p} = p + h/r$ and $\hat{c} = c + h/r$.

A control v^* is optimal if $V^{v^*}(x) \geq V^v(x)$ for all admissible control v and for every initial state x. The focus of this paper is to show the existence of such an optimal control.

3. Optimal Threshold Control in a Queueing System

From the definition of stochastic intensities (Brémaud 1980), let N(t) be the state of a queueing system at time t and $\mathcal{F}_t^N$be the internal history generated by N(t). Suppose A(t) and D(t) admit $\mathcal{F}_t^N$-intensities α_t, and β_t, respectively. Then

$$0 \leq \alpha_t \leq \hat{\lambda}\ (N(t)),\ \ 0 \leq \beta_t \leq \hat{\mu}\ (N(t)) \quad \cdots\cdots\cdots \ (3.\ 1)$$

And for any non-negative $\mathcal{F}_t^N$-predictable process Ct,

$$E_x \int_0^\infty C_t\, dA(t) = E_x \int_0^\infty C_t \alpha_t dt,$$

$$E_x \int_0^\infty C_t\, dD(t) = E_x \int_0^\infty C_t \alpha_t dt. \quad \cdots\cdots\cdots\cdots \ (3.\ 2)$$

Consider a single-stage queueing system, in which the arrival and the departure processes are modelled as point processes with stochastic intensities with the arrival intensity α_t and the departure intensity β_t. Then the objective to be maximized is :

$$E_x \int_0^\infty e^{-rt}\ (\hat{p}\beta_t - \hat{c}\alpha_t)dt - hx/r. \quad \cdots\cdots\cdots\cdots \ (3.\ 3)$$

For a given non-negative integer b, let

$$\lambda^{b}(N)=\hat{\lambda}(N), \quad \text{if} \quad N\leq b-1,$$
$$\qquad 0, \qquad \text{if} \quad N>b.$$
$$\mu^{b}(N)=\hat{\mu}(N), \ \text{for } N\geq 0,$$

then the threshold control $\theta^{b}=\{(\alpha_{t}, \beta_{t}), t>0\}$ is defined by

$$\alpha_{t}=\lambda^{b}(N(t)) \ \text{ and } \ \beta_{t}=\mu^{b}(N \ (t)), \ \cdots\cdots (3.5)$$

where the integer b will be referred to as the upper barrier of the threshold control θ^{b}. The value function under the threshold control θ^{b} will be denoted by v^{b}.

Then the following conclusion is derived in chen and Yao(1990).

Theorem 1 : If the capacity limit $\hat{\lambda}(N)$ is non-increasing, the capacity $\hat{\mu}(N)$ is non-decreasing and concave, and $\hat{\mu}(\infty)\langle\infty$, then there exists an optimal threshold control with a finite upper barrier b^{*} where b^{*} is the value of b which satisfies.

$$b^{*}=\min\{b : V^{b+1}(b+1)-V^{b+1}(b)<\hat{c}\}<\infty \qquad \cdots\cdots (3.6)$$

And the optimal upper barrier b^{*} obtained is independent of the initial stat x.

Indeed, in the context of production systems, as the work-in-process level increases, the arrival process often slows down and eventually drops to zero, while the departure process gets faster but gradually flattens out.

Therefore, the optimal threshold control in a general queueing system with the above conditions can be summarized as follows : if the state is $N(t)<b^{*}$, then the input and the output are at full blasts, and if the state $N(t)$ reaches the upper barrier b^{*}, the input intensity drops to zero while the output intensity is still at its full blast. And if the initial state $x\in[0, b^{*}]$, then $N(t)\in[0, b^{*}]$ for any $t\geq 0$, And the optimal threshold b^{*} is the value which satisfies equation(3, 6).

4. Optimal Threshold Control in a Generalized Jackson Network

The queueing network introduced in section 2 is known to have the following product form equilibrium distribution :

$$P(\underline{x}=\underline{n})=P(\underline{x}=\underline{0})^{-1} \prod_{k=1}^{|n|-1} \lambda(k)\prod_{i=1}^{M} F_{i}(n_{i}) \qquad \cdots\cdots (4.1)$$

where

$$F_{i}(n_{i})=v_{i}^{ni}\prod_{k=1}^{ni} 1/\mu_{i}(k).$$

Many conditions can be altered without changing the product-form result (Kelly 1979), which is essentially what we need for the main result in this paper.

Let $\lambda(N)=0$ if $N\geq L$, and $\lambda(N)=\infty$ otherwise, where L is a predetermined number, such that $\prod_{N=0}^{k} \lambda(N)=0$ for all $K\geq L$. This asymptotic special case of the generalized Jackson model (1963) is the closed queueing network (CQN) model of Gorden and Newell(1967), where the total number of jobs within the network remains constant. This can be interpreted as maintaining a fixed number of jobs within the network.

Therefore, the state space, &, in CQN becomes

$$\& =\{\underline{n}\in Z_{+}^{M} \mid |\underline{n}| =N\}, \qquad \cdots\cdots (4.2)$$

where z_{+}^{M} denotes the M-vector whose components are non-negative integers ; $|\underline{n}| =\Sigma_{i=1}^{M} n_{i}$. And the product-form equilibrium distribution is:

$$P(\underline{x}=\underline{n})=G^{-1}(N) \prod_{i=1}^{M} F_{i}(n_{i}) \qquad \cdots\cdots (4.3)$$

where

$$F_i(n_i) = v_i^{n_i} \prod_{k=1}^{n_i} / \mu_i(k)$$
$$G(N) = \sum_{|n|=N} \prod_{i=1}^{M} F_i(n_i)$$

The throughput function of the CQN is known to be :

$$TH(N) = G(N-1)/G(N) \quad \cdots\cdots\cdots\cdots\cdots\cdots (4.\ 4)$$

Another special case of the Jackson model is the open queueing network with finite buffer capacity (OQNF) where $\lambda(N)=0$ if $N \geq L$. That is, whenever the total number of jobs within the network reaches this threshold (L), arriving jobs are blocked from entry and lost, and L can be interpreted as buffer limit.

Index two CQNs by the superscripts 1 and 2, where only the service rates are differ each other, while all other things being the same. Then the following lemmas are of interest.

Lemma 1 : Assume that the service rate $\mu_i(n_i)$ is nondecreasing function as a function of the local queue length n_i. If the service rates of a station are increased, that is $\mu_i^{(1)}(n_i) \geq \mu_i^{(2)}(n_i)$ ($n_i = 1, \cdots$, N), for a given N, the throughput function of a CQN, TH(N), is also increased, that is $TH^{(1)}(N) \geq TH^{(2)}(N)$.

Proof : The proof is given in Shanthikumar and Yao(1986). It was verified through the concept of "equilibrium rate," associated with the probability mass function of a discrete random variable. When there are multiple parallel servers with fixed service rate per server, the second order property with respect to the number of servers at a station can also be proved by comparing the sample paths of the service completion processes (Shanthikumar and Yao 1987).

Lemma 2 : If the service rate $\mu_i(n_i)$ is nondecreasing concave (convex) function as a function of n_i, the CQN throughput, TH(N), has also the same property as a function of the job population N.

Proof : The proof is due to Shanthikumar and Yao(1988), and is also based on the equilibrium rate representation of the throughput. And the non-decreasing properties and the second-order properties of equilibrium rates are shown to be preserved under convolution.

Note that the above lemma implies that for all CQNs with multiple parallel-server stations (including single servel stations), TH(N) is increasing and concave, since in this case $\mu_i(n_i) = \mu_i' \cdot \min(n_i,\ C_i)$ is increasing and concave in n_i where μ_i' is the constant service rate per server at station i and $C_i (\geq 1)$ is the number of servers there. On the other hand, if all stations are infinite server stations (i. e., $C_i \geq N$ for all i), then $\mu_i(n_i) = n_i \mu_i'$ is a linear function and hence the linearity of TH(N) from above.

For a given nonnegative integer, b, let

$$\lambda^b(N) = \begin{cases} \hat{\lambda}(N), & \text{if } N \leq b\text{-}1 \\ 0\ , & \text{if } N \geq b, \end{cases}$$
$$\mu_i^b(N) = \hat{\mu}_i(n_i) \text{ for } n_i \geq 1 (i=1, \cdots\cdots, M). \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (4.\ 5)$$

then the threshold control in a general queueing network $\theta^b = \{(\alpha_t,\ \beta_t),\ t \geq 0\}$ is again defined by

$$\alpha_t = \lambda^b(N(t)) \quad \text{and} \quad \beta_t = TH^b(N(t)) \quad \cdots\cdots (4.\ 6)$$

where TH^b denote the throughput of the network given θ^b with service rates $(\mu_i^b(n_i(t)))_{i=1}^{M}$, where $\Sigma_{i=1}^{M}\ n_i(t) = N(t)$.

Then the main result of the paper is the following.

Theorem 2 : If the capacity limit $\hat{\lambda}(N)$ is non-increasing, the capacity limits, $\hat{\mu}(n_i)$ $(i \in \mathcal{U})$, is non-decreasing and concave, and $\hat{\mu}_i(\infty)<\infty(i\in \mathcal{U})$, then there exists an optimal threshold control with a finite upper barier b^* where b^* is the value of b which satisfies

$$b^*=\min\{b : V^{b+1}(b+1)-V^{b+1}(b)<\hat{c}\}<\infty \qquad \cdots\cdots (4.7)$$

And the independency of the optimal upper barrier b^* of the initial state x is still preserved.

Proof : Construct a CQN with M+1 stations which consists of two subsets of stations. A subset, $\mathcal{U}$, consist of M stations, {1, ⋯, M}, which are the stations of the original OQN and the other subset, $\mathcal{B}$, consists of a fictitious station which corresponds to the outside of the original OQN. Refer this station as station 0 and $\mathcal{B}=\{0\}$, Let $\mathcal{G}=\mathcal{B}+\mathcal{U}$. Assume for a moment that the total number of jobs in this constructed CQN equals to N* and these jobs cycle between station 0 and the other subnetwork for services. Therefore when N jobs are in the subnetwork, the number of jobs at station 0 equals N*−N and the output rate from station 0 equals to λ(N) which corresponds to the arrival rate to the original OQN. And hence, when all N* jobs are in the sub-network, the output rate from station 0 drops to zero and the service rate $\lambda(k)$ $(k=0, 1, \cdots, N^*)$ is non-decreasing.

Hence, by using the subscripts $\mathcal{U}$, $\mathcal{B}$, and $\mathcal{G}$ to denote, respectively, any quantities related with the sub-network $\mathcal{U}$, $\mathcal{B}$ and CQN $\mathcal{G}$, the marginal equilibrium distribution of the number of jobs in the sub-network $\mathcal{U}$ with the total population being N equal to :

$$\begin{aligned}
&P(|\underline{n}|_{\mathcal{U}}=N)\\
&=G_\beta(N^*-N)\ G_{\mathcal{U}}(N)/G_{\mathcal{G}}(N^*)\\
&=\lambda(N-1)G_\beta(N^*-N+1)\ G_{\mathcal{U}}(N-1)/TH_{\mathcal{U}}(N)\\
&\quad *G_{\mathcal{G}}(N^*)\\
&=P(|\underline{n}|_{\mathcal{U}}=N-1)\cdot\lambda(N-1)/TH_{\mathcal{U}}(N). \qquad \cdots\cdots (4.8)
\end{aligned}$$

where

$$G_\beta(N^*-N)=1/\prod_{k=N}^{N^*-1}\ \lambda(k)=\lambda(N-1)$$
$$*G_\beta(N^*-N+1)$$

and $G_{\mathcal{U}}(N)=G_{\mathcal{U}}(N-1)/TH_{\mathcal{U}}(N)$.

Therefore the following is obvious :

$$\begin{aligned}
&\lambda(N-1)P(|\underline{n}|_{\mathcal{U}}=N-1)\\
&\quad =TH_{\mathcal{U}}(N)P(|\underline{n}|_{\mathcal{U}}=N). \qquad \cdots\cdots (4.9)
\end{aligned}$$

Assume N* converges to infinity, then the subnetwork $\mathcal{U}$ remains the same as the original OQN and the station 0 corresponds to the outside of the OQN. And equation (4.10) leads to the detailed balance conditions of the random variable representing the equilibrium number of jobs in a birth-death queue with state-dependent birth rate λ(N) and death rate $TH_{\mathcal{U}}(N)$. Therefore, the original OQN gives rise to the same behavior as a single birth-death queue in Theorem 1 with N being a state variable. The required non-decreasing and concave properties of TH(N) are obvious through Lemma 1 and Lemma 2.

Theorem 2 can be readily adapted to a system of OQNF. In this case we assume that the input capacity limit $\hat{\lambda}(N)=0$ for $N\geq L$ and the initial work-in-process inventory $x=N(0)\in[0, L]$. Then the optimal control is also threshold type and the result is summarized in the following theorem.

Theorem 3 : Under the same assumptions as in Theorem 1, the optimal control for OQNF is a threshold control, and its optimal barrier is

$$b_L^*=\min\{b\leq L-1 : V^{b+1}(b+1)-V^{b+1}(b)\leq\hat{c}\}$$

and if no such b^* exists then $b_L^*=L$.

Proof : From the proof of Theorem 2, OQN behaves as a single birth-death queue. Then the proof of Theorem 7. 1 in Chen and Yao (1990) is easily extended.

5. Conclusion

Open Queueing network proves to be effective tools in the design and control of production systems and the threshold control mechanism prevails in a wide range of production systems. This paper characterizes the conditions and the form of an optimal threshold control in terms of both the input and the output processes in an open queueing network. Though it doesn't necessarily provide an algorithm to efficiently compute the optimal threshold, the result derived here in a very general conditions is very constructive one from a theoretical point of view.

References

[1] Bremaud, P. 1981, Point Processes and Queues, Springer-Verlag, New York.

[2] Chen, H. and Yao, D. D., Optimal Intensity Control of a Queueing System with state Dependent Capacity Limits, to appear in IEEE Transactions on Automatic Control.

[3] Gorden, W. J. and Newell, G. F. 1967, Closed Queueing Networks with Exponential Servers, Operations Research 15, 252-267.

[4] Jackson, J. R. 1963, Jobshop like Queueing Systems, Management Science 10, 131-142.

[5] Robertazzi, T. G. and Lazar, A. A. 1985, On the Modeling and Optimal Flow Control of the Jacksonian Network, Performance Evaluation 5, 29-43.

[6] Shanthikumar, J. G. and Yao, D. D. 1986, The Effect of Increasing Service Rates in a Closed Queueing Network, Journal of Applied Probability 23, 474-483.

[7] Shanthikumar, J. G. and Yao, D. D. 1987, Optimal Server Allocation in a System of Multi-Server Stations, Management Science 9, 1173-1180.

[8] Shanthikumar, J. G. and Yao, D. D. 1988, Second-order Properties of the Throughput of a Closed Queueing Network, Mathematics of Operations research 13, 524-534.