STRONG ASYMPTOTIC OPTIMALITY OF FOCUSED FACTORY

J. GEORGE SHANTHIKUMAR
Department of Industrial Engineering and Operations Research
and Walter A. Haas School of Business
University of California at Berkeley
Berkeley, CA 94720

and

SUSAN H. XU
College of Business Administration
The Pennsylvania State University
University Park, PA 16802

March 1999

ABSTRACT: In this paper we consider a production enterprise that has several factories. This enterprise manufactures several different types of products on a produce to order basis. The level of operational control available within each factory with respect to scheduling is limited. Hence, it is of interest to find the appropriate allocation of the different types of products to the different factories. In a focused factory only similar products will be produced within the same factory. We formulate a multiple server, multiple job class queueing model for the production enterprise. Using this model, we show that the traditional wisdom of forming focused factories to process similar tasks that reduce fluctuations in processing times is strongly asymptotically optimal. Suggestions for further work on this topic are summarized as well.

Key Words: Focused factory, $G/GI/c$ queueing model, Multiple job classes, Strong asymptotic optimality.

1. INTRODUCTION

Traditional wisdom is that when there is a lack of or limited control available, group similar entities together. This is a principle that is universally followed. In a production environment, this suggests focused factories. There are several empirical studies of real systems and numerical studies of their models to verify this (e.g., see Souza, Wagner and Whybark (1998)). The purpose of this paper is to provide an analytical justification for this.

In this paper we consider a production enterprise that has several factories that manufactures several different types of products on a produce to order basis. The level of operational control available within each factory with respect to scheduling is limited. Hence, it is of interest to find the appropriate allocation of the different types of products to the different factories. We formulate a multiple server, multiple job class queueing model for the production enterprise. Using this model, we show that the above traditional wisdom of forming focused factories to process similar tasks that reduce fluctuations in processing times is strongly asymptotically optimal in a sense described below (for more details of this and other applications see Shanthikumar and Xu (1997, 1998)).

Suppose $C(\mathbf{x}, \theta)$ is the performance measure of a queueing system, where $\mathbf{x}$, $(\mathbf{x} \in \mathcal{X})$ is the design variable(s) or the control policy (such as the number of servers at different service centers in a production or service system and/or the workload allocated there or the scheduling control policy) and θ, $(\theta \in \Theta)$ is a system parameter (such as the customer arrival rate to the service system and/or the population size of a closed queueing network). Let $\mathbf{x}^*(\theta)$ be an optimal value of $\mathbf{x}$ that maximizes $C(\mathbf{x}, \theta)$. If our approximation results in a solution $\hat{\mathbf{x}}(\theta)$, we are interested in evaluating the degree of sub-optimality $\hat{\delta}(\theta) = C(\mathbf{x}^*(\theta), \theta) - C(\hat{\mathbf{x}}(\theta), \theta)$. It is customary only to evaluate this numerically. But in cases, where it is possible to identify a θ^* such that $C(\mathbf{x}^*(\theta), \theta) \to \infty$ as $\theta \to \theta^*$ (such as the utilization of a server going to one, or the population size in a closed system going to infinity), one is often interested in showing that

$$\frac{\hat{\delta}(\theta)}{C(\mathbf{x}^*(\theta),\theta)} \to 0 \text{ as } \theta \to \theta^*. \tag{1.1}$$

If this is achieved we say that $\hat{\mathbf{x}}(\theta)$ is an **asymptotically optimal** design or control policy **in** θ. Such an approach is common in heavy traffic analysis. But even when this is achieved we may have

$$\hat{\delta}(\theta) \to \infty \text{ as } \theta \to \theta^*.$$

Therefore just showing (1.1) alone is not sufficient to verify that $\hat{\mathbf{x}}(\theta)$ is a good design or a good control policy. Suppose we can show that there exists a finite constant K independent of θ such that

$$\hat{\delta}(\theta) \le K,\ \theta \in \Theta \text{ and as } \theta \to \theta^*. \tag{1.2}$$

If the above can be verified, we will say that the design or the control policy ($\hat{\mathbf{x}}(\theta)$) is **strongly asymptotically optimal in** θ.

The organization of this paper is as follows: In Section 2, we describe the multiple server, multiple job class queueing model for the production enterprise. Strong asymptotic optimality of the focused factory is established in Section 3. Suggestions for further work on this topic are summarized in Section 4.

2. MODEL: MULTIPLE SERVER, MULTIPLE JOB CLASS QUEUEING SYSTEM

In this section we will formulate a queueing model for the production enterprise. Consider a single-stage queueing system that has s stations (each representing a factory) of equal capabilities. Jobs (customer orders) arrive at the system according to a renewal stream with inter-arrival times $\{A_n, n = 1, 2, \ldots\}$ with finite mean $1/\lambda$ and squared coefficient of variation (scv) C_a^2. The probability that a new arrival belongs to class j (product type j) is q_j, $j = 1, 2, \ldots, r$, $\sum_{j=1}^{r} q_j = 1$. Upon arrival, the class of the job is identified. Then a class j job is routed to server i with probability θ_{ij}, $\sum_{i=1}^{s} \theta_{ij} = 1$, $j = 1, 2, \ldots, r$. Let $\Theta = \{\theta_{ij}, i = 1, \ldots, s, j = 1, \ldots, r\}$.

Note that in a real production enterprise, one will partition the set of product types and allocate these partitions to the different factories. Then the θ_{ij}'s will be either zero or one. But we will formulate it in this general setting and show that the strongly asymptotically optimal allocation probabilities will not only be mostly zero's or one's, but will also satisfy the concept of focused factory.

The generic processing time of a class j job is S_j, $j = 1, 2, \ldots, r$, with finite mean $1/\mu_j$ and scv $C_{S_j}^2$. Let $\lambda_j = q_j\lambda$ be the arrival rate of class j jobs to the system and $\lambda_{ij} = \theta_{ij}\lambda_j$ the rate of class j jobs routed to station i, $i = 1, 2, \ldots, s$. Suppose that the service discipline at each station is first-come first-served (this reflects the limited scheduling control capabilities available at the factories).

We require that the allocation is equitable to all s stations. That is, the average workload allocated to each station is identical. Therefore, our problem is

$$\begin{aligned} &\min_{\Theta} \sum_{i=1}^{s} E[N_i(\Theta)] \\ \text{subject to} &\sum_{i=1}^{s} \theta_{ij} = 1,\ j = 1, 2, \ldots, r \\ &\sum_{j=1}^{r} \lambda_{ij} E[S_j] = \rho,\ i = 1, 2, \ldots, s, \end{aligned} \tag{2.1}$$

where $E[N_i(\Theta)]$ is the number of jobs in station i, $\rho = (\lambda/s)\sum_{j=1}^{r} q_j E[S_j]$ is the workload per unit time routed to each station.

The problem of allocating multiple-classes of jobs in a multiple cell system with Poisson arrival process has been studied by Buzacott and Shanthikumar (1992, 1993). The approach adopted in the next section is similar to theirs.

3. STRONG ASYMPTOTIC OPTIMALITY OF FOCUSED FACTORIES

In this section we will, using the model developed in the previous section, show that the focused factory is strongly asymptotically optimal in the sense described in the introduction. For this, we first consider a simplified version of the problem whose solution will lead to a solution to the preceding problem. Suppose that we have two stations ($s = 2$) and $r = 2n$ classes, for some $n = 1, 2, \dots,$ so that

$$\lambda_j E[S_j] = \text{ constant } = \frac{\rho s}{r} = \frac{\rho}{n}, \ j = 1, 2, \dots, r. \tag{3.1}$$

We wish to group these classes into two sets T and $\bar{T}$ such that $|T| = n = |\bar{T}|$ and classes in T are assigned to station 1 and classes in $\bar{T}$ are assigned to station 2. The problem is then

$$\min_T \{E[N_1(T)] + E[N_2(\bar{T})]\}, \tag{3.2}$$

where $E[N_1(T)]$ ($E[N_2(\bar{T})]$) is the expected number of jobs in station 1 (station 2) under allocation T ($\bar{T}$). without loss of generality we assume that the classes are ordered so that

$$E[S_1] \leq E[S_2] \leq \cdots \leq E[S_r]. \tag{3.3}$$

Then, from (3.1),

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r. \tag{3.4}$$

We also assume that the service times of different classes satisfy the following agreeability condition for the mean remaining service times

$$\frac{E[S_1^2]}{2E[S_1]} \leq \frac{E[S_2^2]}{2E[S_2]} \leq \cdots \leq \frac{E[S_r^2]}{2E[S_r]}. \tag{3.5}$$

In other words, we require that if the mean service time of a class is smaller than that of another class, then the same relationship holds with respect to the mean remaining service time. Then from (3.1) and (3.5) one has

$$\lambda_1 E[S_1^2] \leq \lambda_2 E[S_2^2] \leq \cdots \leq \lambda_r E[S_r^2]. \tag{3.6}$$

Next we present the expression of $E[N_1(T)]$ (for the derivation of this result, see Shanthikumar and Xu (1998)). Since the service discipline is FCFS, we can aggregate all classes in T into a single job class, say class T-jobs, and find the means and scv's of inter-arrival times and service times of class T jobs. Let A_T be the generic inter-arrival time of class T-jobs. The mean and scv of A_T, denoted by $1/\lambda_T$ and $C_a^2(T)$, are

$$\frac{1}{\lambda_T} := E[A_T] = \frac{1}{\sum_{j \in T} \lambda_j}$$

and

$$C_a^2(T) := (\frac{\lambda_T}{\lambda})(C_a^2 - 1) + 1, \tag{3.7}$$

respectively. The mean service time of a class-T job is

$$\frac{1}{\mu_T} = \sum_{j\in T} \frac{\lambda_j}{\lambda_T}\frac{1}{\mu_j} = \frac{\rho}{\lambda_T} \tag{3.8}$$

where $\rho = \frac{1}{2}\sum_{j=1}^{r} \frac{\lambda_j}{\mu_j}$, and the last equality of (3.8) is due to (3.1). Note that the fraction of time that station 1 is busy is $\frac{\lambda_T}{\mu_T} = \rho$. The scv of the service time of the class-T job is found to be

$$C_S^2(T) = (\mu_T)^2 \sum_{j\in T} \frac{\lambda_j E[S_j^2]}{\lambda_T} - 1 = \frac{\lambda_T \sum_{j\in T}\lambda_j E[S_j^2]}{\rho^2} - 1 \tag{3.9}$$

With the above notation, it is known (e.g., see. equation 3.112 of Buzacott and Shanthikumar 1993) that

$$\begin{aligned} E[N_1(T)] =& \frac{(C_a^2(T)+1) - 2\rho + \rho^2(C_S^2(T)+1)}{2(1-\rho)} - \frac{\lambda_T E[I_T^2]}{2E[I_T]} + \rho \\ =& \frac{C_a^2(T) + C_s^2(T)}{2(1-\rho)} + \frac{1+\rho}{2} - \frac{(1+\rho)C_S^2(T)}{2} + \frac{\lambda_T E[I_T^2]}{2E[I_T]} \end{aligned} \tag{3.10}$$

where

$$I_T = \max\{\hat{A}_T - W_T, 0\} \tag{3.11}$$

and W_T is the stationary sojourn time of a T-job. Similarly,

$$E[N_2(\bar{T})] = \frac{C_a^2(\bar{T}) + C_s^2(\bar{T})}{2(1-\rho)} + \frac{1+\rho}{2} - \frac{(1+\rho)C_S^2(\bar{T})}{2} + \frac{\lambda_{\bar{T}} E[I_{\bar{T}}^2]}{2E[I_{\bar{T}}]} \tag{3.12}$$

The proof of the next lemma can be found in Shanthikumar and Xu (1997).

Lemma 3.1: Let T be a subset of $\{1, 2, \ldots, r\}$ and $|T| = n$.

$$\alpha_T := \sup_x\{\lambda_T E[A_T - x | A_T > x]\} \text{ and } \beta_T := \sup_x\{\lambda_T \frac{E[(A_T - x)^2 | A_T > x]}{2E[A_T - x | A_T > x]}\}. \tag{3.13}$$

Then

$$\lambda_T E[I_T^2] \over E[I_T] \le \frac{\lambda_T}{\lambda}(C_a^2 - 1 + 2\alpha_T + 2\beta_T) + 2. \tag{3.14}$$

The next lemma gives the bounds for $E[N_1(T)] + E[N_2(\bar{T})]$.

Lemma 3.2: Let T be a subset of $\{1, 2, \ldots, r\}$ and $|T| = n$. Let

$$\hat{N}_1(T) = \frac{(C_a^2 - 1)(\frac{\lambda_T}{\lambda}) + \frac{1}{\rho^2}(\lambda_T \sum_{j \in T} \lambda_j E[S_j^2])}{2(1-\rho)}, \tag{3.15}$$

$$\hat{N}_2(\bar{T}) = \frac{(C_a^2 - 1)\frac{\lambda_{\bar{T}}}{\lambda}) + \frac{1}{\rho^2}(\lambda_{\bar{T}} \sum_{j \in \bar{T}} \lambda_j E[S_j^2])}{2(1-\rho)}, \tag{3.16}$$

and

$$l = 2C_S^2 + C_a^2 + 2\alpha + 2\beta + 1, \qquad \text{and} \qquad u = 2 \tag{3.17}$$

where

$$C_S^2 = \max_T \{C_S^2(T)\}, \tag{3.18}$$

$$\alpha = \max_T \{\alpha_T\}, \tag{3.19}$$

$$\beta = \max_T \{\beta_T\}. \tag{3.20}$$

Then

$$\hat{N}_1(T) + \hat{N}_2(\bar{T}) - l \le E[N_1(T)] + E[N_2(\bar{T})] \le \hat{N}_1(T) + \hat{N}_2(\bar{T}) + u. \tag{3.21}$$

For the proof of this Lemma, see Shanthikumar and Xu (1998).

Consider the lower bound

$$\begin{aligned} E[N_1(T)] + E[N_2(\bar{T})] &\ge \hat{N}_1(T) + \hat{N}_2(T) - l \\ &= \frac{(C_a^2 - 1) + \frac{1}{\rho^2}[(\lambda_T \sum_{j \in T} \lambda_j E[S_j^2]) + (\lambda_{\bar{T}} \sum_{j \in \bar{T}} \lambda_j E[S_j^2])]}{2(1-\rho)} - l. \end{aligned} \tag{3.22}$$

Let $T' = \{1, 2, \ldots, n\}$ and $\bar{T}' = \{n+1, n+2, \ldots, 2n\}$. By (4.5),

$$\lambda_T \lambda_k E[S_k^2] + \lambda_{\bar{T}} \lambda_l E[S_l^2]$$
$$\geq \lambda_{T'} \min\{\lambda_k E[S_k^2], \lambda_l E[S_l^2]\} + \lambda_{\bar{T}'} \max\{\lambda_k E[S_k^2], \lambda_l E[S_l^2]\},$$

it is immediate from (3.5) that

$$\frac{\lambda_T \sum_{j \in T} \lambda_j E[S_j^2]) + \lambda_{\bar{T}} \sum_{j \in \bar{T}} \lambda_j E[S_j^2]}{2(1-\rho)}$$
$$\geq \frac{\lambda_{T'} \sum_{j \in T'} \lambda_j E[S_j^2]) + \lambda_{\bar{T}'} \sum_{j \in \bar{T}'} \lambda_j E[S_j^2]}{2(1-\rho)}. \tag{3.23}$$

Because $C_a^2 - 1$ and ρ are fixed in (4.26), we see that allocation $(T', \bar{T}')$ minimizes the lower bound for the number of jobs in the system.

The next theorem addresses the strong optimality of $(T', \bar{T}')$. See Shanthikumar and Xu (1998) for its proof.

Theorem 3.1: Let $(T^*, \bar{T}^*)$ be the optimal grouping that minimizes equation (4.3). The grouping $(T', \bar{T}')$ is strongly asymptotically optimal under heavy traffic in the sense that

$$0 \leq (E[N_1(T')] + E[N_2(\bar{T}')]) - (E[N_1(T^*)] + E[N_2(\bar{T}^*)]) \leq l + u < \infty, \tag{3.24}$$

and

$$\frac{E[N_1(T')] + E[N_2(\bar{T}')] - E[N_1(T^*)] + E[N_2(\bar{T}^*)]}{E[N_1(T^*)] + E[N_2(\bar{T}^*)]} \to 0 \text{ as } \lambda \to \mu. \tag{3.25}$$

Now consider our original problem of probabilistic routing of multiple-classes of jobs to two stations. We assume that (3.3) and (3.5) still hold. Suppose $\rho_j = \lambda_j E[S_j]$, $j = 1, 2, \ldots, r$, are rationals. It implies that there exists a $\hat{\rho} > 0$ such that $n_j = \rho_j / \hat{\rho}$ is an integer, $j = 1, 2, \ldots, r$. For some integer $m \geq 1$ let us divide class j into $2m \cdot n_j$ artificial classes, each with arrival rate $\lambda_j / (2m \cdot n_j)$ and the same service time distribution as class j jobs. Then the workload of each artificial class is $\hat{\rho}/2m$. With the new classification we have totally $2m \sum_{j=1}^r n_j$ job classes, each with an equitable work load rate of $\hat{\rho}/2m$.

Using the previous result, the optimal grouping of the classes is $\{1, 2, \ldots, m\sum_{j=1}^{r} n_j\}$ and $\{m\sum_{j=1}^{r} n_j + 1, 2m\sum_{j=1}^{r} n_j\}$. Let

$$k = \min\{l : \; 2m\sum_{j=1}^{l} n_j \geq m\sum_{j=1}^{r} n_j, \; l = 1, \ldots, r\}. \tag{3.26}$$

Then k corresponds to the class whose jobs are split probabilistically between the two stations, where the rate of going to station i, $i = 1, 2$, are

$$\begin{aligned} \lambda_{k1} =& \theta_{k1}\lambda_k = \frac{(m\sum_{j=1}^{r} n_j - 2m\sum_{j=1}^{k-1} n_j)\hat{\rho}}{2mE[S_k]}, \\ \lambda_{k2} =& \lambda_k - \lambda_{k1}. \end{aligned} \tag{3.27}$$

Then, the original classes $\{1, 2, \ldots, k\}$ with allocation rate $\{\lambda_1, \ldots, \lambda_{k-1}, \lambda_{k1}\}$ are routed to station 1, and the original classes $\{k, \ldots, r\}$ with allocation rate $\{\lambda_{k2}, \lambda_{k+1}, \ldots, \lambda_r\}$ are routed to station 2. Because $n_j = \rho_j/\hat{\rho}$, we can rewrite (4.30) as

$$k = \min\{l : \; 2\sum_{j=1}^{l} \rho_j \geq \rho, \; l = 1, \ldots, r\}, \tag{3.28}$$

$$\lambda_{k1} = \frac{\rho/2 - \sum_{j=1}^{k-1} \rho_j}{E[S_k]} \qquad \text{and} \qquad \lambda_{k2} = \lambda_k - \lambda_{k1}. \tag{3.29}$$

Observe that this optimal grouping is independent of m. Because the work load rate of each artificial class $\hat{\rho}/2m \to 0$ as $m \to \infty$, the class grouping problem becomes the same as the probabilistic routing problem. Hence the independence of the solution of the grouping problem to the value of m implies that this solution is strongly asymptotically optimal to the probabilistic routing problem as well.

Finally, using pairwise application of this result to s stations, one concludes:

Theorem 3.2: The following allocation is strongly asymptotically optimal: Allocate classes $\{i_l, \ldots, i_{l+1}\}$ with probabilities $(\lambda_{l,i_l}/\lambda_l, 1, \ldots, 1, \lambda_{l,i_{l+1}})$ to station l, where $i_1 = 1$,

$$i_{l+1} = \min\{k : \sum_{j=1}^{k} \rho_j \geq l\rho, \; k = 1, \ldots, r\} \tag{3.30}$$

$$\lambda_{l,i_{l+1}} = \frac{l\rho - \sum_{j=1}^{i_{l+1}} \rho_j}{E[S_{i_{l+1}}]} \qquad \text{and} \qquad \lambda_{l+1,i_{l+1}} = \lambda_{i_{l+1}} - \lambda_{l,i_{l+1}}. \tag{3.31}$$

In words, Theorem 3.2 says that the strongly asymptotically routing has the property that the classes with similar processing times are routed to the same station. Hence our solution supports the traditional wisdom of forming stations to process similar items and to reduce wide fluctuations in processing requirement.

4. CONCLUSION AND SUGGESTION FOR FUTURE RESEARCH

Assuming that the level of operational control available at the factories are limited, we were able to show that the use of focused factories is advisable. However, it is important to understand the level of operational control that is needed to make a more generalized factory desirable. One may formulate queueing systems with different service policies at the station level and carry out a similar analysis to see at what level of operational control, the desirability for focused factory will disappear.

REFERENCES

[1] Buzacott, J. A. and Shanthikumar, J. G. (1992). Design of manufacturing systems using queueing models, *Queueing Systems: Theory and Applications,* **12**, 135-214.

[2] Buzacott, J. A. and Shanthikumar, J. G. (1993). *Stochastic Models of Manufacturing Systems*, Prentice Hall Inc., Englewood Cliffs, N.J.

[3] Shanthikumar, J. G. and Xu, S. H. (1997). Asymptotically optimal routing and service rate allocation in a multi-server queueing system, *Operations Research,* **45**, 464-469.

[4] Shanthikumar, J. G. and Xu, S. H. (1998). Strongly asymptotically optimal design and control of production and service system, Working paper, Walter A. Haas School of Business, University of California, Berkeley, CA.

[5] Souza, G. C., Wagner, H. M. and Whybark, D. C. (1998). Evaluating focused factory benefits with queueing theory, Working paper, The University of North Carolina, Chapel Hill, NC