*IIE Transactions* (1999) **31**, 195–206

# Optimal allocation of resources in a job shop environment

SRIDHAR SESHADRI and MICHAEL PINEDO

*Leonard N. Stern School of Business, New York University, New York, NY 10012, USA*
*E-mail: sseshadr@stern.nyu.edu*

Received January 1997 and accepted April 1998

In this paper we study the allocation of production services (e.g., maintenance) to machining centers in a job shop when there is a limited amount of resources for such services. Different classes of jobs go through the shop. A job class is characterized by its route, its processing requirements, and its priority. The problem we address is how to optimally allocate production services (the resources) to machining centers so as to minimize the total Work-In-Process in the entire system. In order to analyze this problem we model the job shop as an open queueing network. Assuming certain relationships between performances of machining centers (i.e., speeds) and the amounts of resources allocated, we present methods for allocating the resources to the individual machining centers optimally.

## 1. Introduction

In this paper we study the allocation of production services to machining centers in a job shop environment. Production services, such as material handling, provisions of tooling and fixtures, and maintenance and repair, have a significant impact on throughput. One of the basic assumptions is that there is a limit on the resources that provide these services. The shop floor approach for allocating these services is normally based on heuristics that use bottleneck analysis and the values of the jobs processed at the machining centers. It is clear that when the manager actually allocates services to equipment, the issues mentioned above require an integrated analysis.

To put the problem in perspective, consider the concept of Total Productive Maintenance (TPM). The goal of TPM is to achieve ideal conditions for a plant. Nakajima [1], for example, has suggested that the following operating conditions should be achieved: (1) a machine should be available for more than 90% of the time; (2) its performance efficiency should be greater than 95%; and (3) its yield rate should be greater than 99%. To achieve these operating conditions he focused on six sources of productivity loss that may be present in a manufacturing environment: (1) speed reductions; (2) idling and minor stoppages; (3) set-up and adjustment times; (4) equipment failures; (5) process defects; and (6) yield loss. These losses can be broadly categorized into speed, downtime, and quality related losses. The modeler is faced with the problem of quantifying the losses and analyzing the costs of corrective actions.

In this paper the cost of corrective actions is regarded as the cost of providing production services. It may also be that, in the long term, the losses due to downtime and lower speed can cause additional Work-In-Process (WIP) inventory that has to be maintained. Based on these assumptions we build a model that is amenable to analysis. The model is based on the fact that a machining center performs better when more resources are allocated to it. The relationship between the resource allocated and the speed of the machining center is assumed to be continuous. A machining center's performance is assumed to be a continuous function that is increasingly concave in the amount of resource allocated.

In this paper we show that well known techniques for analyzing the performance measures of queueing networks, and recently developed methods for characterizing these solutions using stochastic convexity concepts [2] can be used to determine analytically the optimal allocation in a job shop. We develop an open queueing network model for a job shop, and based on this model address three issues:

(1) How does the mix of products affect the allocation of services?
(2) How does the arrival pattern of the jobs affect the allocation?
(3) Can an algorithm be designed for solving this problem?

In this paper, we consider for modeling purposes an aggregate allocation of resources. The results can be applied to all services that have an effect on processing rates at the machining centers.

The notion of allocating services on an aggregate basis, which is a key idea in this paper, is not new. Joshi and Gupta [3], and Seshadri [4] have studied this in the context

of allocating preventive maintenance set-ups within a given planning horizon. More recently, Ashayeri *et al.* [5] have used a mathematical program to decide whether or not to schedule maintenance within a given planning horizon. The tradeoffs in their model are between the cost of preventive maintenance versus the costs of holding inventory, costs of incurring backlog, and setup costs. Dekker [6] has used a unified approach to model the costs associated with preventive maintenance. He also shows how these cost functions can be used to plan maintenance in coordination with production. These studies do not consider the stochastic nature of job arrival times and processing times and they restrict their attention to the modeling of the breakdown and repair processes. Kayton *et al.* [7] have described a study carried out at a semiconductor fabrication facility (fab) to determine the impact of maintenance on the different pieces of equipment in the fab. They use a simulation model to demonstrate that downtime at non-bottleneck equipment can affect the throughput rate and thus the average WIP in the facility. In relation to their work, we use an analytical model to develop guidelines that will help a manager decide when utilization alone is not a sufficient factor in deciding where to concentrate maintenance efforts, see Section 5.

Allocation problems in general have been studied extensively, see Ibaraki and Katoh [8]. Examples of work in the context of systems' reliability provide stronger conclusions than our objective criterion of minimizing the value of WIP, [9,10]. Liyanage and Shanthikumar [9] have considered the stochastic allocation problem of determining how many units to allocate to each of $M$ facilities. The response of facility $m$ is random and given by $X_m(m)$. The system's response is a joint function, $h(X_1(r_1), X_2(r_2), \ldots, X_M(r_M))$. The objective is to maximize the expected utility $E(g(h(X_1(r_1), X_2(r_2), \ldots, X_M(r_M))))$. The budget constraint is that the sum of the allocations should not exceed a value $R$. They identified conditions on $X_m(r_m)$ under which, with suitable restrictions on $g$, the allocation can be determined in an easy manner. Shaked and Shanthikumar [10] studied an allocation problem in which they determined how to stochastically maximize the lifetime of parallel and series systems. For examples of similar work see Li [11] and Singh and Singh [12].

The modeling framework adopted in this paper is that of an Open Queuing Network (OQN). It is useful to review the work done on resource allocation within the OQN modeling framework. Kleinrock [13] has reviewed the early work done in the area of capacity assignment in queueing networks. Kelly [14] has discussed several variants of server allocation problems in an OQN. Wein [15] also considers server allocation problems in open queueing networks. However, all server allocation problems considered in the literature assume that the relationship between station performance and resource allocated is linear. Another area of research that is somewhat related to this problem is the allocation of machines in a queueing network. Each machining center in the network consists of a number of machines in parallel and there is a fixed number of machines to be allocated to the various centers. A significant amount of research has been done on the optimal allocation of machines to machining centers in such a network. A number of algorithms have been proposed for this discrete optimization problem. For excellent surveys of this area of research, see Buzacott and Shanthikumar [16] and Bitran and Morabito [17]. Buzacott and Shanthikumar provide a comprehensive review on resource allocation in manufacturing systems.

In the literature dealing with Flexible Manufacturing Systems, Vinod and John [18] have considered a closed queueing network model of a two-stage repair set-up. They have suggested integer programming to solve the problem. They have modeled the dynamics of the actual allocation rather than study the conjecture that an increase in allocation improves performance. Similar comments apply to the recent work of Widmer and Solot [19], and Miriyala and Viswanadham [20], who also model maintenance of a Flexible Manufacturing System. Thus it is seen that the notion of aggregate planning of production services in a multi-product job shop environment is a research area that has not received much attention so far.

This paper is organized as follows. In Section 2 we describe the basic model. In Section 3 we discuss some results concerning open queueing networks. The approach for analyzing the problem is discussed in Sections 4 and 5 under two different assumptions: (i) using machine processing rates that are only a function of the amount of resource allocated; and (ii) using rates that are a function of both the queue length and the amount of resource allocated to the machine. Procedures for solving the problem are also described. In the final section we present a qualitative appraisal of the results.

## 2. The model

The conceptual model of the problem is that the average level of services provided affects the performance of equipment (the service rate), and in turn the service rate affects the work flow. Taking this into account the decision-maker has to use these relationships to allocate and prioritize the allocation of services or resources. The modeling framework adopted is that of an open queueing network with unlimited buffer capacities at each node. In this framework the major components are: (1) the number of product classes; (2) the arrival rate of each job class; (3) the route of each job class and the associated service requirements; (4) the number of machining centers (which we will call machines); (5) the queue priorities; and (6) the relationship between the allocation of the production services and the service rates of machines. In such a framework it is customary to make assumptions that

lead to product form solutions [21–23]. The need for a closed form expression for the expected value of the Work-In-Progress (WIP) at a machine necessitates the assumption that each queue is quasi-reversible [14]. The assumptions are listed below.

Jobs of class $i$, $i = 1, 2, \ldots, C$, arrive according to a Poisson process with a rate $\lambda_i$, have a fixed (deterministic) route structure $s_i(j)$, $j = 1, 2, \ldots, n_i$, where at stage $j$ of the route machine $s_i(j)$ is used. Jobs of class $i$ have a value $w_{ij}$ at stage $j$ of their route.

The service requirement of a class $i$ job at machine $m$, $m = 1, \ldots, M$, has a mean $a_{mj}$. If quasi-reversibility is assumed, then the means of the processing times of all jobs on machine $m$ have to be equal. If the priority rule at a machine is according to a symmetric queue, then the service time distributions can be arbitrary and class dependent. In either case, the service requirements are assumed to be independent of each other. Of a total resource $R$ an amount $r_m$ is allocated to machine $m$. If the queue length is $l$, then machine $m$ provides service at a rate $\phi_m(l, r_m)$. In other words, $\phi_m(l, r_m)$ is the amount of work done in one unit of time. The exact manner in which the allocation $r_m$ affects $\phi_m(l, r_m)$ is described in the next section. The queue priority, which is independent of the classes, is assumed to be such that a product form solution is obtained (see the assumptions at the end of this section).

The simplifying assumptions are quite broad in their scope as far as the processing requirements of jobs are concerned. First, as long as the queue priority leads to a symmetric queue, the service requirement could be a random variable with an arbitrary distribution. Second, if we only assume quasi-reversibility then all the $a_{ij}$ values are equal to a constant and we may assume any of a range of class independent priority rules (including first come first served) and obtain product form expressions for the expected WIP at each machine, as is discussed in (Kelly [14], Section 3.2). In what follows the expected value of the WIP at machine $m$ will be denoted by $E(V_m)$. A limiting assumption is the one made regarding the arrival pattern of jobs being Poisson. The assumptions regarding the effect of service allocation on processing rates are presented in detail in the next section. The assumptions made are partly justified by the fact that they permit an analysis using closed form expressions for an otherwise intractable problem, and partly by the intuitively appealing form of the solution obtained in Sections 4 and 5.

Given the model, the objective is to allocate service to the machines so that the total average value of the WIP in the system is minimized. This objective is subject to the constraints that

(1) the required throughput is achieved;
(2) the resource constraint is satisfied.

It should also be noted that the assumption of a deterministic route structure permits the assumption that the value of work-in-progress increases over the route of a class of jobs, by assuming a different value of $w_{ij}$ at each stage of processing.

One question concerns the types of service requirements we can incorporate into the model and still obtain a product form solution. For example, assume that the service requirements are exponentially distributed. Then one issue that has to be investigated is whether interruptions of service due to machine failures can lead to a significant distortion or deviation from a product form solution. A second issue that needs to be investigated is the effect of the resource allocation on the service times. In Appendix A we show the conditions under which processing times of machines that are subject to breakdown and repair may be considered similar to an exponential distribution. This condition is used below in assumption A2. The effect of resource allocation on service requirements is also derived in Appendix A and will be used in Section 4.

Three different ways of modeling the impact of resource allocation and service time distributions are discussed below. Consider the following three assumptions.

*Assumption 1* (A1): one way of modeling the problem is to assume that the service times are still exponentially distributed. This assumption can be justified if the squared coefficient of variation $c_v^2$ is close to unity. As shown in Appendix A, the $c_v^2$ will be close to unity if either the service rate is smaller than the repair or failure rates, or the failure rate is small relative to the repair rate (this would partly justify the exponential assumption). The assumption that the failure rate is small with regard to the repair rate is justifiable if the machine availability is relatively high. The assumption of exponential service times is harder to justify in the context of job shops. The priority rule can be first come first served. If this is the case, we assume that the processing rate at station $m$ does not depend on the queue length $l$, i.e., $\phi_m(l, r_m)$ does not depend on $l$. We then use the notation $\phi_m(r_m)$ to analyze this case. We assume that the processing rate $\phi_m(r_m)$ is increasing concave in the allocation $r_m$.

*Assumption 2* (A2): if the queue is a symmetric queue, then a product form solution can also be obtained. Some allowable priorities are: (1) last come first served; and (2) server sharing. The service time distribution may be arbitrary but the service times have to be i.i.d. at each machine and independent of everything else. In addition, we assume that the processing rate does not depend on the queue length and denote the rate as $\phi_m(r_m)$. The processing rate is assumed to be increasing concave in the allocation $r_m$.

*Assumption 3* (A3): an alternative approach is to assume that each queue is quasireversible and that the service rate at machine $m$ is queue length dependent. Denote the

service rate to be $g_m(l)$ when the queue length is $l$ and the machine is working at 100% efficiency. Then define the efficiency factor of machine $m$ when $r_m$ of the resource is allocated to it as $\eta_m(r_m)$. We assume that these two factors combine in such a way that

$$\phi_m(l, r_m) = g_m(l)\eta_m(r_m). \tag{1}$$

The efficiency factor $\eta_m(r_m)$ is assumed to be increasing concave in $r_m$. In this approach the average service rate includes both downtime and speed losses.

It is interesting to note that all three assumptions lead to a similar solution and therefore warrant testing using simulation. In the next section the OQN model is described taking all the assumptions into account. In Section 4 either A1 or A2 is assumed to hold. In Section 5 A3 is assumed to hold and $\eta_m(r_m)$ is assumed to be increasing concave in $r_m$.

## 3. Preliminary results

In this section an expression for the expected value of the WIP at machine $m$, $E(V_m)$, is derived. The equilibrium distribution of the queue at machine $m$ is given by [14]:

$$P(N_m = n) = p_m(n) = \frac{b_m a_m^n}{\prod_1^n \phi_m(l, r_m)}, \tag{2}$$

$$a_m = \sum_{i=1}^{C} \lambda_i \sum_{j=1}^{n_i} I[s_i(j) = m] a_{ij}, \tag{3}$$

$$b_m = \left( \sum_{n=0}^{\infty} \frac{a_m^n}{\prod_1^n \phi_m(l, r_m)} \right)^{-1}, \tag{4}$$

$$E(V_m) = \frac{\left(\sum_{n=1}^{\infty} n p_m(n)\right) \sum_{i=1}^{C} \lambda_i \left( \sum_{j=1}^{n_i} w_{ij} I[s_i(j) = m] a_{ij} \right)}{a_m}, \tag{5}$$

where $I[.]$ is the indicator function. Let $E(N_m)$ denote the expected number of jobs at machine $m$. Define $f_{im}$ to be the contribution of a class $i$ job to the value of the WIP at machine $m$. Then

$$f_{im} = \left( \sum_{j=1}^{n_i} w_{ij} I[S_i(j) = m] a_{ij} \right) \lambda_i / a_m.$$

Define $U_m$ to be the expected value of the WIP per job in the queue at machine $m$. The value of $U_m$ can be expressed in terms of the $f_{im}$ as follows,

$$U_m = \sum_{i=1}^{C} f_{im}.$$

With these definitions and the use of Equation (5), the expected value of the jobs in WIP at machine $m$ simplifies to

$$E(V_m) = U_m E(N_m).$$

The expected value of jobs in WIP in the shop is denoted as $E(V)$. The optimization problem is:

$$\min \sum_{m=1}^{M} E(V_m),$$

subject to

$$r_1 + r_2 + \cdots + r_M \le R, \tag{6}$$

$$r_1, r_2, \ldots, r_M \ge 0. \tag{7}$$

Two cases are analyzed separately in the following two sections. Case I is based either on Assumption A1 or on Assumption A2. Case II is based on Assumption A3.

## 4. Analysis of case I

Recall that under A1 and A2, the service rate, $\phi_m(r_m)$, is increasing concave in $r_m$. In this case the expression for

$$E(V) = \sum_{m=1}^{M} E(V_m),$$

is equivalent to

$$E(V) = \sum_{m=1}^{M} U_m \left( \frac{a_m}{\phi_m(r_m) - a_m} \right). \tag{8}$$

So

$$\frac{\partial E(V)}{\partial r_m} = - U_m \left( \frac{a_m \phi_m'(r_m)}{(\phi_m(r_m) - a_m)^2} \right), \tag{9}$$

$$\frac{\partial^2 E(V)}{\partial r_m^2} = - U_m \left( \frac{a_m \phi''(r_m)}{(\phi_m(r_m) - a_m)^2} \right) + 2U_m \left( \frac{a_m \phi'(r_m)^2}{(\phi_m(r_m) - a_m)^3} \right). \tag{10}$$

As $\phi_m(r_m)$ is assumed to be increasing concave in $r_m$, the second derivative given in Equation (10) is non-negative and may be assumed to be strictly positive. It can be verified that if the failure rate $\gamma_m(r_m)$ is decreasing convex in $r_m$, it again follows, using Equations (A1) and (A2) in Appendix A, that the second derivative of $E(V)$, Equation (10), is positive.

The objective function is a separable decreasing convex function. There are several efficient procedures available for solving the problem for either case, when $r_m$ is restricted to be integer or allowed to take on continuous values as is discussed by Fox [24], Ibaraki and Katoh

[8, pp. 15–20, 52–77], and Hochbaum and Shanthikumar [25]. Even when there are more constraints, efficient procedures are available for solving large separable convex optimization problems, as is discussed by Hochbaum and Shanthikumar [25], Nemhauser and Wolsey [26], and Hochbaum and Seshadri [27]. Only the nature of the solution is characterized below. It is clear that the unconstrained optimum is achieved when

$$U_m\left(\frac{a_m\phi'_m(r_m)}{(\phi_m(r_m)-a_m)^2}\right) = U_m\left(\frac{a_m}{(\phi_m(r_m)-a_m)}\right)\left(\frac{1}{(\phi_m(r_m)-a_m)}\right)\phi'_m(r_m) = K,$$

where $K$ is a constant that depends on the structure of the network. This expression is a product of four terms, namely

(1) the expected value of the WIP for each job in the queue ($U_m$);
(2) the expected number of jobs in the queue ($E(N_m)$);
(3) the expected time a job spends at the machine (waiting and being processed);
(4) the marginal increase obtained in processing rate with an allocation increase.

In the optimal partition of resources over the network the value of this product has to be the same for each machine. We can interpret this result as follows: the higher the average value of jobs flowing through a machine and the greater the average number of jobs or flow time at the machine, the larger the allocation of resources will be.

In the constrained case a similar result holds. However, some machines may not get any allocation at all. (This supports the traditional view that the bottleneck analysis of machines should be combined with the value of jobs flowing through them for purpose of allocating priorities). As an example, when the resource allocations can assume only integral values then the marginal (greedy) allocation algorithm is optimal, as is discussed in Fox [24] and Appendix A of Buzacott and Shanthikumar [16]. (As a historical note, Fox credits Gross [28] for this result). As the name suggests, the algorithm allocates one unit of a resource at a time. The unit is allocated to that workcenter that stands to gain the most from the allocation, see Section 6 for an example. For our model, the marginal allocation algorithm remains optimal even with given values for upper and lower bounds ($ub(m)$ and $lb(m)$), that is the individual allocations have to be integers and also satisfy constraints of the form $lb(m) \le r_m \le ub(m)$.

One can show for the unconstrained case, see Equation (A4) in Appendix B, that by using the $GI/G/1$ approximation for the delay in queue, at high levels of server utilization the optimal allocation must indeed be proportional to

$$U_m E(N_m)\frac{1}{(\phi(r_m)-\lambda)}\phi'(r_m).$$

The similarity of the expressions for allocating resources under the different assumptions forms the basis for a greedy algorithm proposed in Section 6. The form of the expression above is no coincidence as demonstrated by Shanthikumar and Xu [29] who considered a single stage queueing system with $c$ heterogeneous servers. Customers arrive according to a renewal process. The service times at server $i$ are i.i.d. random variables and each server has a separate queue. There is a cost $h_i$ associated with having customers wait in queue $i$. Customers are routed to queue $i$ with a fixed probability $\theta_i$. They used bounds for the delay in a GI/GI/1 queue and derived the allocation of work to the servers. Therefore the allocation that they derived is similar in form to the formulae derived in this paper. Moreover Shanthikumar and Xu showed that their allocation is strongly asymptotically optimal, that is, under heavy traffic conditions (namely, utilization approaches unity) the ratio of the expected cost under their allocation to the optimal expected cost converges to unity. The method of analysis that uses bounds and approximations for the delay in the queue can be extended to networks of queues, using the Generalized Jackson Networks modeling framework, as is discussed in Buzacott and Shanthikumar [16] and Section 3.2.1 in Bitran and Morabito [17]. Buzacott and Shanthikumar in their review paper discussed the use of this approach in the solution of a number of allocation problems in manufacturing. Note too, that when using the approximation, $E(N_m)$ depends on the variability of inter-arrival as well as service times. The next section also shows the effect of variability.

## 5. Analysis of case II

In this section we use for the service rate the expression in Equation (1). Assume that A3 holds. Assume that the efficiency factor of machine $m$, $\eta_m(r_m)$, is increasing concave in the allocation $r_m$ and assume that the service rate $g_m(l)$ is increasing and concave in the queue length $l$ at machine $m$. Substitute $\eta_m(r_m)$ by $\tau_m(r_m)^{-1}$, consider $\tau_m(r_m)$ as an arrival rate at the machine, and rewrite Equation (2) as:

$$p_m(n) = b_m\tau_m(r_m)^n \Big/ \prod_{l=1}^{n}\left(\frac{g_m(l)}{a_m}\right).$$

It follows from Theorem 5.5 of Shaked and Shanthikumar [2], that the expected number in the queue is stochastically increasing and convex in $\tau_m(r_m)$. (The power of this Theorem is seen from the fact that we do not have to resort to a single calculation of the derivative until the next step!). Then writing $E(V_m)$ as $V_m(\tau_m)$ we obtain

$$\frac{\partial^2 V_m(\tau_m)}{\partial r_m^2} = \left( \left( -\frac{\partial V_m(\tau_m)}{\partial \tau_m} \frac{\partial^2 \eta_m}{\partial r_m^2} \right) \Big/ \eta_m^2 \right) + \frac{\partial^2 V_m(\tau_m)}{\partial \tau_m^2} \left( \left( \frac{\partial \eta_m}{\partial r_m} \right) \Big/ \eta_m \right)^2 + \left( 2 \frac{\partial V_m(\tau_m)}{\partial \tau_m} \left( \frac{\partial \eta_m}{\partial r_m} \right)^2 \right) \Big/ \eta_m^3.$$

Thus we see that, under the assumption that $g_m(l)$ is increasing concave in $l$, and the assumption that $\eta_m(r_m)$ is increasing concave in $r_m$, the objective function is a decreasing (separable) convex function. As noted earlier, there are efficient procedures available for computing a solution to such a problem.

From Equation (A11) in Appendix C, it follows that the allocation decision now depends on both the value as well as the variability of the jobs queued at a machine. Seen in this light, the result shows that simple bottleneck analysis would not suffice to plan the allocation of services. This is particularly true when the service rates are increasing concave functions of the line length, or when the variability of the jobs awaiting processing at a machine is not too small. To bring this result in line with those given in Section 4 as well as the results given in the Appendices B and C, define the "average" processing rate, $\overline{\phi}_m(r_m)$, in the queue length dependent case to be such that

$$E(N_m) = \frac{a}{\overline{\phi}_m(r_m) - a}.$$

Then, by differentiating this expression with respect to $r_m$, the optimal allocation in the unconstrained case should be proportional to

$$U_m E(N_m) \left( \frac{1}{\overline{\phi}(r_m) - a} \right) \frac{\partial \overline{\phi}(r_m)}{\partial r_m},$$

as found in Section 4. Of course, we were able to use the first order condition for obtaining the minimum because we had earlier proved the convexity of $E[N_m]$ with respect to $r_m$.

In this analysis, we have assumed that the service rate is increasing concave in the queue length. In Appendix C, we obtain conditions under which this assumption may be dropped. The analysis in Appendix C is considerably more complicated, because we have to resort to algebraic manipulations to derive the results. A probabilistic explanation of the results given in this appendix appears to be a challenging problem.

## 6. The allocation algorithm

In order to illustrate the algorithm, consider a system with two stations in series. All jobs are from the same class. The jobs waiting or under service at the first station have a value $w_1$, and the jobs at the second station a value $w_2$. Of course, $w_1 < w_2$. The speeds of the machines at the two stations are, respectively, $\mu_1(r_1)$, and $\mu_2(r_2)$. The processing time distribution at the first station is exponential while the processing time distribution at the second station is arbitrary. The improvement functions $\phi$'s are increasing concave. They are basically a power function with a power that is less than 1, namely,

$$\mu_m(r_m) = \mu_m(0) + c_m r_m^{d_m} \quad 0 < d_m \leq 1, \ \ m = 1, 2.$$

The total budget is $R$, i.e., $r_1 + r_2 \leq R$. In the experiments we have basically four sets of parameters, namely,

(1) the basic speeds of the two stations when zero resources are allocated, $\mu_m(0) = 1$, for $m = 1, 2$;
(2) the improvement functions corresponding to the two machines

$$\mu_m(r_m) = \mu_m(0) + c_m r_m^{d_m}, \quad 0 < d_m \leq 1, \ \ m = 1, 2;$$

(3) the squared coefficient of variation of the processing time distribution at the second station ($c_v^2$); and
(4) the weights at the two inventory points, $w_1, w_2$.

The results of these experiments are depicted in Fig. 1. (In Fig. 1, scv denotes $c_v^2$.) From the figure it is clear that the ratio of allocations, $r_2/r_1$ increases concavely in the ratio $w_2/w_1$. (The ratio $r_2/r_1$ is also increasing in $c_v^2$, and it is also slightly concave in this coefficient). The monotonicity and concave behavior can be explained intuitively as follows. The monotonicity is obvious. The greater the weight associated with inventories at the second station, the higher have to be the resources allocated to the second machine. However, we see that when $w_2/w_1$ increases, the amount of resource that is shifted from the first station to the second station becomes less and less. The reason for this is that any additional reduction in resource at station 1 has a more significant impact than any of the past reductions.

This particular example is analytically tractable. Larger networks that give rise to a product form solutions (satisfying the assumptions stated in Section 2) can be analyzed in a similar fashion. If we used the generalized Jackson network model, the resource allocation problem can be solved using non-linear programming. Interested readers are refered to Bitran and Morabito [17] for examples.

However, networks with arbitrary processing times and arbitrary servicing disciplines cannot be as easily analyzed, since they do not give rise to product form solutions and may not be amenable to analysis even using approximation techniques. For example, even if we assume that the solutions of their steady state probability can be approximated through one of the well-known approximation techniques [30]; we will still be faced with the problem of determining how the squared coefficients of the inter-arrival times would change with changing allocation of resources. Therefore, with a larger network,

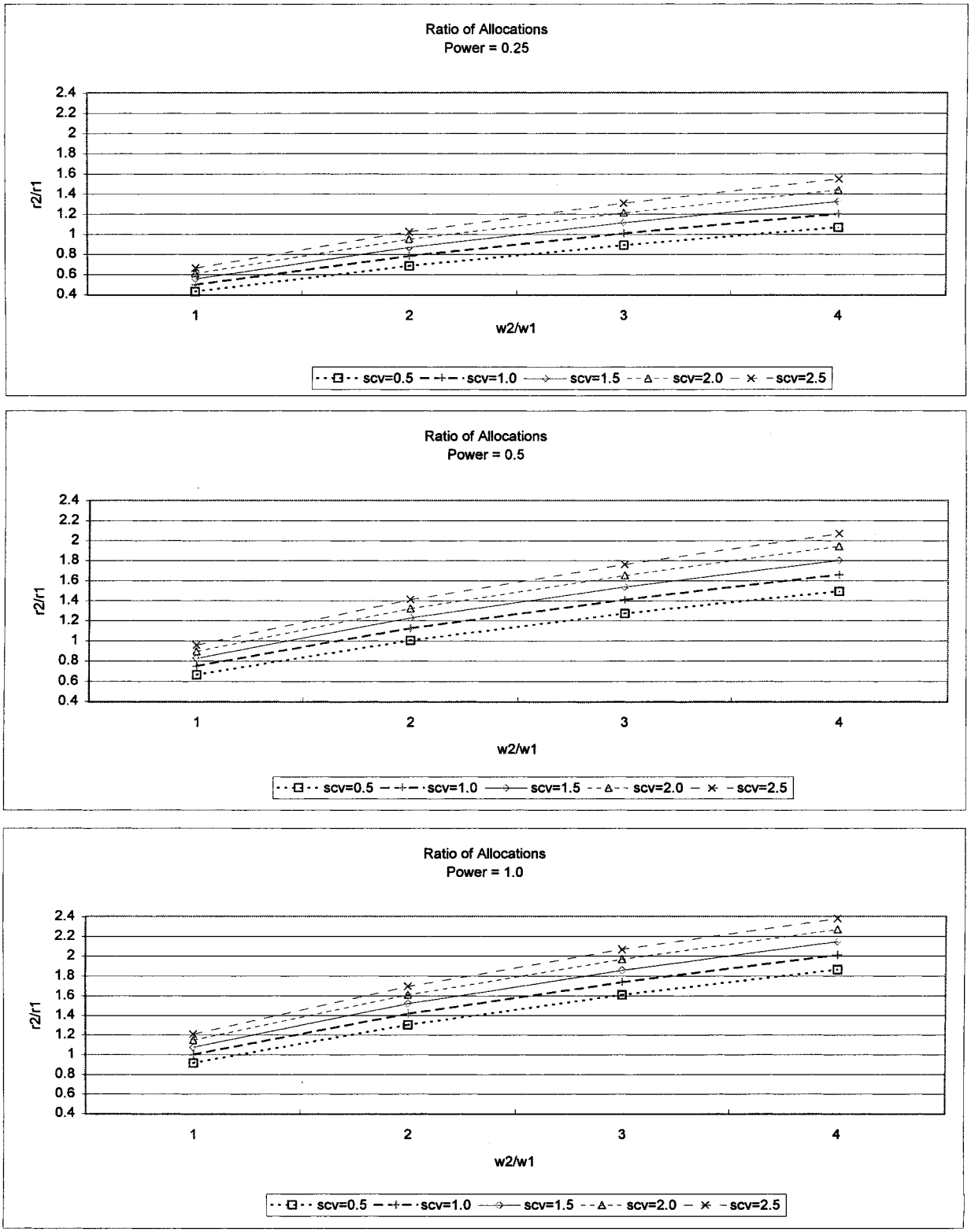


**Fig. 1.** Allocation of resources in a system with two stations in series.

or networks in which there are immediate feedback flows, set up times, lot sizing, machine failures and random yields, we recommend an approach that combines simulation with optimization. A simulation-cum-optimization algorithm for production control is given in Glassey *et al.* [31]. For allocating resources in a more complex setting, it

may be more natural to extend the allocation formula developed in the previous sections in the following straightforward manner:

*Step 0*. Start with zero allocations. Choose a step size, $\Delta R$, for incrementally allocating the budget $R$.
*Step 1*. Simulate the network with the given allocation, $r_1, \ldots, r_m$.
*Step 2*. Determine the machine that has the largest value of:

$$U_m E(N_m)\phi'_m(r_m)/(\phi_m(r_m) - \lambda).$$

Label this machine as $i^*$.
*Step 3*. Allocate $r_{i^*} = r_{i^*} + \Delta R$.
*Step 4*. If the budget is exhausted STOP. Otherwise go to Step 1.

Two examples of applying this method are given in Tables 1 and 2. In the first example, called Example 1, there are three machines, M1, M2, and M3, and three products P1, P2, and P3. The arrival rates, routing information, as well as the mean service times of the three products are shown in Table 1. The inter-arrival time and the service time distributions are gamma, with $c_v^2$ values of 1 and 0.3. We have to allocate two units of resources. Each unit resource will produce an acceleration of 5% in the speed of the equipment. All machines have at the outset a speed of one, i.e., they can perform one unit of work per minute. The objective is to minimize the expected number of jobs in the system. We show the improvement in terms of the total mean flowtime in the system as experienced by each of the three products. In Table 2, we show the effects of an allocation according to three different rules: (1) proportional to the utilization; (2) proportional to the expected number of jobs in the queue at the three machines; and (3) according to the heuristic with a step size, $\Delta R$, equal to 1. The heuristic outperforms the other allocation schemes by 7% or better. In Example 1, because there is no means by which the values of

$$U_m E(N_m)\phi'_m(r_m)/(\phi_m(r_m) - \lambda),$$

can be made equal for the three machines (or over even two of them) using the available two units of resources, the optimal solution is indeed the same as the heuristic solution, that is to allocate all the available resources to M3.

In Example 2, we provide an instance in which the heuristic does not allocate all resources to a single machine. The reduction in mean flow time over the "naive" method of allocating in proportion to utilization is over 25%.

**Table 1.** Routing information

| *Product* | *Route* | *Mean service time (min)* | *scv of service time* | *Arrival rate per day*[1] | *scv of inter-arrival time* |
|---|---|---|---|---|---|
| Example 1 | | | | | |
| P1 | M1 | 1 | 0.3 | 60 | 1 |
| | M2 | 2 | 0.3 | | |
| | M3 | 3 | 0.3 | | |
| P2 | M2 | 1 | 0.3 | 70 | 1 |
| | M1 | 2 | 0.3 | | |
| | M3 | 3 | 0.3 | | |
| P3 | M3 | 1 | 0.3 | 60 | 1 |
| | M2 | 2 | 0.3 | | |
| | M1 | 3 | 0.3 | | |
| Example 2 | | | | | |
| P1 | M1 | 1.7 | 0.3 | 75 | 0.3 |
| | M2 | 2 | 0.3 | | |
| | M3 | 3 | 1 | | |
| P2 | M2 | 1 | 0.3 | 55 | 0.3 |
| | M1 | 2 | 0.3 | | |
| | M3 | 3 | 1 | | |
| P3 | M3 | 1 | 1 | 75 | 0.3 |
| | M2 | 2 | 0.3 | | |
| | M1 | 3 | 0.3 | | |

[1] A day consists of 480 minutes.

## 7. Conclusions

It is evident from the results in Sections 5 and 6 that the optimal allocation and partition of resources over the network has to be such that for each machine the value of

$$U_m E(N_m)\frac{1}{(\phi(r_m) - \lambda)}\phi'(r_m),$$

is equal to the same constant $K$.

In words, the four terms of the formula can be described as follows. The first term represents the expected value of the WIP at machine $m$. The second term represents the expected number of jobs at machine $m$. The third term represents a measure of the flow time of a job at machine $m$ (waiting and being processed). This measure is also proportional to the length of the busy period of the station. The last term represents the marginal increase obtained in the processing rate of machine $m$ with an increase in allocation.

This condition is somewhat intuitive. The marginal increase in processing rate obtained with any additional resource allocation has a certain benefit. The WIP at that machining center will be less. The amount of WIP reduction will be proportional to the total amount of WIP that is residing at that queue. The amount to be allocated is also proportional to the expected length of time that a job remains at a station, which makes sense since one can view this also as proportional to the total expected payout because of that customer.

The fact that the allocation is proportional to the mean busy period needs some explanation. Let $x$ jobs arrive together at a server every $(x + x/3)$ minutes. Let the processing time of a job be 1 minute. The utilization of the server is 0.75, independent of the value of $x$. Then the busy period in this system has a duration of $x$, and the

**Table 2.** Simulation results

| *Allocation method* | *Resource allocated* | | | *Theoretical utilization (%)* | | | *Total mean flowtime*[1] *(min)* | | |
|---|---|---|---|---|---|---|---|---|---|
| Example 1 | | | | | | | | | |
| | M1 | M2 | M3 | M1 | M2 | M3 | P1 | P2 | P3 |
| Base case | 0 | 0 | 0 | 79.2 | 64.6 | 93.8 | 40.7 ± 0.36 | 40.8 ± 0.33 | 40.7 ± 0.36 |
| First 5% | | | | | | | | | |
| Proportional to utilization | 0.333 | 0.272 | 0.395 | 77.8 | 63.7 | 91.9 | 32.1 ± 0.23 | 32.3 ± 0.24 | 32.0 ± 0.24 |
| Proportional to mean number in queue $E(N)$ | 0.17 | 0.06 | 0.77 | 78.5 | 64.4 | 90.1 | 30.5 ± 0.22 | 30.6 ± 0.20 | 30.3 ± 0.23 |
| Heuristic | 0 | 0 | 1 | 79.2 | 64.6 | 89.1 | 27.0 ± 0.18 | 27.3 ± 0.17 | 27.0 ± 0.18 |
| Second 5% | | | | | | | | | |
| Proportional to utilization | 0.333 | 0.272 | 0.395 | 76.6 | 62.8 | 90.1 | 27.7 ± 0.22 | 27.7 ± 0.19 | 27.6 ± 0.21 |
| Proportional to mean number in queue $E(N)$ | 0.224 | 0.089 | 0.687 | 77.6 | 64.1 | 87 | 23.6 ± 0.18 | 23.6 ± 0.17 | 23.4 ± 0.17 |
| Heuristic | 0 | 0 | 1 | 79.2 | 64.6 | 84.4 | 21.9 ± 0.15 | 22.1 ± 0.14 | 21.9 ± 0.15 |
| Example 2 | | | | | | | | | |
| Base case | 0 | 0 | 0 | 96.4 | 74 | 96.9 | 103.1 ± 0.6 | 102.6 ± 0.69 | 103.0 ± 0.61 |
| Proportional to utilization | 0.36 | 0.28 | 0.36 | 92.9 | 71.9 | 93.4 | 49.4 ± 0.34 | 49.0 ± 0.37 | 49.2 ± 0.34 |
| Proportional to mean number in queue $E(N)$ | 0.284 | 0.024 | 0.691 | 93.6 | 73.8 | 90.2 | 40.4 ± 0.22 | 40.0 ± 0.25 | 40.1 ± 0.22 |
| Heuristic | 0.5 | 0 | 0.5 | 92.98 | 73.96 | 90.58 | 32.3 ± 0.20 | 31.5 ± 0.23 | 32.5 ± 0.20 |

[1] The simulations were carried out using the HOM: Process Analysis module (Moses *et al*. [32]). The run length of each simulation was 576 000 minutes. The mean flowtime is based on 71 000 to 84 000 observations (depending upon the product) for example 1, and 66 000 to 90 000 observations for example 2. The three sigma confidence intervals are based on batching 500 observations and setting the standard error equal to the standard deviation across these batched means. The simulations were carried out on a Pentium© 200 MHz Personal Computer.

average waiting time in the system is given by $(x+1)/2$. It then follows that the improvement in waiting time for a given increase in the processing speed of the server is directly proportional to the busy period. The analogy of this example can be carried over to a situation in which arrivals are random and the arrival times of jobs are unaffected by the change in processing speed.

The following question can now be raised. Do the results in Section 5 hold if we do not make any assumption with regard to the service rate $\phi_m(l)$ at machine $m$, i.e., it is not necessarily increasingly concave with the queue length $l$? It turns out that the results obtained in Section 5 often hold under this weaker condition, but not always. This is further discussed in Appendix C.

## Acknowledgements

We gratefully acknowledge the comments of three anonymous referees, whose careful scrutiny helped improve the presentation of this paper considerably.

## References

[1] Nakajima, S. (1988) *Introduction to TPM: Total Productive Maintenance*, Productivity Press, Cambridge, MA.

[2] Shaked, M. and Shanthikumar, J.G. (1988) Stochastic convexity and its applications. *Advances in Applied Probability*, **20**, 427–446.

[3] Joshi, S. and Gupta, R. (1986) Scheduling of routine maintenance using production schedules and equipment failure history. *Computers and Industrial Engineering*, **10**, 11–20.

[4] Seshadri, S. (1988) Determination of aggregate preventive maintenance programs using production schedules. *Computers and Industrial Engineering*, **14**, 193–200.

[5] Ashayeri, J., Teelen, A. and Selen, W. (1996) A production and maintenance planning model for the process industry. *International Journal of Production Research*, **34**, 3311–3326.

[6] Dekker, R. (1995) Integrating optimization, priority setting, planning and combining of maintenance activities. *European Journal of Operational Research*, **82**, 225–240.

[7] Kayton, D., Teyner, T., Schwartz, C. and Uzsoy, R. (1997) Focusing maintenance improvement efforts in a wafer fabrication facility operating under the theory of constraints. *Production and Inventory Management Journal*, **38**, Fourth Quarter, 51–57.

[8] Ibaraki, T. and Katoh, N. (1988) *Resource Allocation Problems: Algorithmic Approaches*, MIT Press, Cambridge, MA.

[9] Liyanage, L. and Shanthikumar, J.G. (1992) Allocation through stochastic schur convexity and stochastic transposition increasingness, in *Stochastic Inequalities*, IMS Lecture Notes/Monograph Series, **22**, Shaked, M. and Tong, Y.L., (eds), Institute of Mathematical Statistics, Hayward, CA. pp. 253–273.

[10] Shaked, M. and Shanthikumar, J.G. (1992) Optimal allocation of resources to nodes of parallel and series systems. *Advances in Applied Probability*, **24**, 894–914.

[11] Li, H. (1994) On a class of stochastic arrangement inequalities arising in optimal allocation of resources. *Probability in the Engineering and Informational Sciences*, **8**, 113–124.

[12] Singh, H. and Singh, R.S. (1997) Optimal allocation of resources to nodes of series systems with respect to failure-rate ordering. *Naval Research Logistics*, **44**, 147–152.

[13] Kleinrock, L. (1976) *Queueing Systems*, Vol. 2: *Computer Applications*, John Wiley, New York, NY.
[14] Kelly, F.P. (1979) *Reversibility and Queueing Networks*, John Wiley, New York, NY.
[15] Wein, L.M. (1989) Capacity allocation in generalized Jackson networks. *Operations Research Letters*, **8**, 143–146.
[16] Buzacott, J.A. and Shanthikumar, J.G. (1992) Design of manufacturing systems using queueing models. *Queueing Systems*, **12**, 135–213.
[17] Bitran, G.R. and Morabito, R. (1996) Open queueing networks: optimization and performance evaluation models for discrete manufacturing systems. *Production and Operations Management*, **5**, 163–193.
[18] Vinod, B. and John, T.C. (1986) On optimal capacities for repair facilities in flexible manufacturing systems, in *Flexible Manufacturing Systems: Methods and Studies*, Kusiak, A. (ed), Elsevier, Amsterdam, pp. 341–349.
[19] Widmer, M. and Solot, P. (1990) Do not forget the breakdowns and the maintenance operations in FMS design problems. *International Journal of Production Research*, **28**, 421–430.
[20] Miriyala, K. and Viswanadham, N. (1989) Reliability Analysis of FMS. *International Journal of Flexible Manufacturing Systems*, **2**, 145–162.
[21] Shanthikumar, J.G. and Buzacott, J.A. (1981) Open queuing models of dynamic job shops. *International Journal of Production Research*, **19**, 255–266.
[22] Solberg, J.J. (1977) A mathematical model of computerized manufacturing systems, in *Proceedings of the Fourth International Conference on Production Research*, Tokyo, Japan. Taylor and Francis, London, pp. 1265–1275.
[23] Yao, D.D. and Buzacott, J.A. (1985) Modeling the performance of flexible manufacturing systems. *International Journal of Production Research*, **5**, 945–959.
[24] Fox, B.L. (1966) Discrete optimization via marginal analysis. *Management Science*, **13**, 210–216.
[25] Hochbaum, D.S. and Shanthikumar, J.G. (1990) Convex separable optimization is not much harder than linear optimization. *Journal of the Association for Computing Machinery*, **37**, 843–862.
[26] Nemhauser G.L. and Wolsey W.A. (1988) *Integer and Combinatorial Optimization*, Wiley, New York.
[27] Hochbaum, D.S. and Seshadri, S. (1993) The empirical performance of a polynomial algorithm for constrained nonlinear optimization. *Annals of Operational Research*, **43**, 229–248.
[28] Gross, O. (1956) A class of discrete-type minimization problems. Report RM-1644-PR, Rand Corporation, Santa Monica, CA.
[29] Shanthikumar, J.G. and Xu, S.H. (1997) Asymptotically optimal routing and service rate allocation in a multiserver queueing system. *Operations Research*, **45**, 464–469.
[30] Buzacott, J.A. and Shanthikumar, J.G. (1993) *Stochastic Models of Manufacturing Systems*, Prentice Hall, Englewood Cliffs, NJ.
[31] Glassey, C.R., Seshadri, S. and Shanthikumar, J.G. (1996) Linear control rules for production control of semiconductor fabs. *IEEE Transactions on Semiconductor Manufacturing*, **9**, 536–549.
[32] Moses, M., Seshadri, S. and Yakirevich, M. (1998) *HOM: Operations Management Software for Windows*, Irwin McGraw-Hill, Burr Ridge, IL.
[33] Feller, W. (1950) *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd edn., Wiley, New York.

## Appendix A

In this appendix we analyze stations that are subject to breakdowns. The service time requirements are assumed to be exponentially distributed. We derive conditions under which the time that a job occupies a machine can still be regarded as close to the exponential.

Assume that the service time of a job on a machine is an exponentially distributed random variable with parameter $\mu$. Assume that the up and down times of the machine are also exponentially distributed random variables with parameters $\gamma$ and $\delta$ respectively, [30]. Assume that the processing of an interrupted job is resumed after a breakdown. Let the random variable $X$ denote the service requirement of a job and let the random variable $S$ denote the total time spent on the machine by a job. Let the random variable $D$ denote the number of failures (down times) during a service, and assume that each failure requires a random repair time $R_j$, $j = 1, 2, \ldots, D$. So

$$S = X + \sum_{j=1}^{D} R_j.$$

The Laplace transform of $S$ is

$$F_S(s) = \frac{\mu}{\mu + s + \gamma - (\gamma\delta/(\delta + s))},$$

and the mean and variance of $S$ are

$$E(S) = \frac{1}{\mu} + \frac{\gamma}{\mu\delta},$$

and

$$\mathrm{Var}(S) = \left(\frac{1}{\mu} + \frac{\gamma}{\mu\delta}\right)^2 + 2\frac{\gamma}{\mu\delta^2}.$$

The squared coefficient of variation $c_v^2$ of $S$ is close to unity if

$$2\frac{\gamma}{\mu\delta^2} \Big/ \left(\frac{1}{\mu} + \frac{\gamma}{\mu\delta}\right)^2 = 2\mu\frac{\gamma}{(\delta + \gamma)^2} \approx 0.$$

This condition is used in assumption A1, see Section 2.

If we assume that the down times are not greatly affected by the service provided (as in a well-maintained system), then the first and second derivatives of the average time spent by a job on the machine, with respect to the maintenance service (i.e., the allocation $r_m$) provided, are given by

$$E'(S) = \frac{\gamma'}{\mu\delta}, \tag{A1}$$

$$E''(S) = \frac{\gamma''}{\mu\delta}. \tag{A2}$$

These two equations are used in Section 4.

## Appendix B

Bounds for the expected number in the queue in a GI/G/1 queue are usually given in terms of the squared coefficient of variation of the inter-arrival time and the service time

distributions. Denote these to be $c_a^2$ and $c_S^2$ respectively. Denote the quantity of resources allocated to the server as $r$. Let the arrival rate of customers be $\lambda$ and the service rate be $\phi(r)$. Kingman's bound for the expected number of jobs, $N(r)$, in the system is, [30]:

$$N(r) \leq \frac{\lambda(\phi^2(r)c_a^2 + \lambda^2 c_S^2)}{2(\phi^2(r) - \lambda\phi(r))} + \frac{\lambda}{\phi(r)}.$$

We can verify that: (1) if $\phi > \lambda$, then $\phi^2/(\phi^2 - \lambda\phi)$ and $1/(\phi^2 - \lambda\phi)$ are convex functions of $\phi$; (2) that $1/\phi$ is a convex decreasing function of $\phi$; and (3) that $N$ is a decreasing function of $\phi$. Therefore $N$ is a convex decreasing function of $\phi$. Now if $\phi(r)$ is a concave non-decreasing function of $r$, then

$$\frac{\mathrm{d}^2 N(\phi(r))}{\mathrm{d}r^2} \approx \frac{\partial N(\phi(r))}{\partial\phi}\frac{\mathrm{d}^2\phi(r)}{\mathrm{d}r^2} + \frac{\partial^2 N(\phi(r))}{\partial\phi^2}\left(\frac{\mathrm{d}\phi(r)}{\mathrm{d}r}\right)^2 \geq 0.$$

The optimal allocation for minimizing the average number in the queue in the unconstrained case is given by,

$$\frac{\partial N(\phi(r))}{\partial\phi}\frac{\mathrm{d}\phi(r)}{\mathrm{d}r} = \text{constant}.$$

or

$$\left(\frac{\lambda(2\phi(r)c_a^2)}{2(\phi^2(r) - \lambda\phi(r))}\right)\phi'(r) - \left(\frac{\lambda(\phi^2(r)c_a^2 + \lambda^2 c_S^2)}{2(\phi^2(r) - \lambda\phi(r))^2}(2\phi(r) - \lambda)\right)\phi'(r) - \left(\frac{\lambda}{\phi^2(r)}\right)\phi'(r) = \text{constant}. \quad \text{(A3)}$$

For relatively high levels of utilization, the second term in the parentheses on the right hand side of equation (A3) will tend to dominate. This term can be written as,

$$\frac{\lambda(\phi^2(r)c_a^2 + \lambda^2 c_S^2)}{2(\phi^2(r) - \lambda\phi(r))^2}(2\phi(r) - \lambda)\phi'(r) \approx N(r) \times \frac{1}{(\phi(r) - \lambda)} \times \phi'(r), \quad \text{(A4)}$$

thus yielding an expression similar to the ones obtained in Sections 4 and 5.

## Appendix C

In this appendix we consider the same problem as studied in Section 5. The service rate is written as the product of two functions, $\phi_m(l, r_m) = g_m(l)\eta_m(r_m)$. However, now the function $g_m(l)$ is not necessarily increasing concave in the queue length $l$ at machine $m$. However, the efficiency factor $\eta_m(r_m)$ is still increasing concave in the allocated resource $r_m$.

Writing $\eta_m(r_m)$ as $\tau_m(r_m)^{-1}$, we obtain analogous to Equations (2), (3), (4), and (5):

$$\phi_m(l, r_m) = g_m(l)\eta_m(r_m) = g_m(l)/\tau_m(r_m). \quad \text{(A5)}$$

$$p_m(n) = \frac{b_m(r_m)a_m^n\tau_m(r_m)^n}{\prod_1^n g_m(l)}. \quad \text{(A6)}$$

$$b_m(r_m) = \left(\sum_{n=0}^{\infty}\left(\frac{a_m^n\tau_m(r_m)^n}{\prod_1^n g_m(l)}\right)\right)^{-1}. \quad \text{(A7)}$$

$$E(V_m) = U_m\sum_{n=1}^{\infty}\frac{b_m(r_m)a_m^n\tau_m(r_m)^n}{\prod_1^n g_m(l)}n, = U_m b_m(r_m)\sum_{n=1}^{\infty}\tau_m^n(r_m)C_{mn}n, \quad \text{(A8)}$$

where

$$C_{mn} = \frac{a_m^n}{\prod_1^n g_m(l)},$$

is a constant and $b_m(r_m)$ is the normalizing constant. Assuming differentiability of $\eta(.)$, and using Equations (A7) and (A8), we obtain,

$$\frac{\partial E(V_m)}{\partial r_m} = U_m b_m(r_m)\sum_{n=1}^{\infty}\tau_m^n(r_m)C_{mn}n^2\left(\frac{\tau_m'(r_m)}{\tau_m(r_m)}\right) - U_m b_m'(r_m)\sum_{n=1}^{\infty}\tau_m^n(r_m)C_{mn}n. \quad \text{(A9)}$$

$$b_m'(r_m) = -\left(\sum_{n=0}^{\infty}n\frac{a_m^n\tau_m(r_m)^n}{\prod_1^n g_m(l)}\left(\frac{\tau_m'(r_m)}{\tau_m(r_m)}\right)\right)\Bigg/\left(\sum_{n=0}^{\infty}\frac{a_m^n\tau_m(r_m)^n}{\prod_1^n g_m(l)}\right)^2, = -b_m(r_m)E(n)\left(\frac{\tau_m'(r_m)}{\tau_m(r_m)}\right). \quad \text{(A10)}$$

Thus using Equations (A9) and (A10), we obtain the expression

$$\frac{\partial E(V_m)}{\partial r_m} = U_m\left(E(N_m^2) - E(N_m)^2\right)\left(\frac{\tau_m'(r_m)}{\tau_m(r_m)}\right). \quad \text{(A11)}$$

A similar exercise results in

$$\frac{\partial^2 E(V_m)}{\partial r_m^2} = U_m\left(E(N_m^2) - E(N_m)^2\right)\left(\left(\frac{\tau_m''(r_m)}{\tau_m(r_m)}\right) - \left(\frac{\tau_m'(r_m)}{\tau_m(r_m)}\right)^2\right) + U_m\left(\frac{\partial E(N_m^2)}{\partial r_m} - \frac{\partial E(N_m)^2}{\partial r_m}\right)\left(\frac{\tau_m'(r_m)}{\tau_m(r_m)}\right). \quad \text{(A12)}$$

$$\frac{\partial E(N_m^2)}{\partial r_m} = (E(N_m^3) - E(N_m^2)E(N_m))\left(\frac{\tau_m'(r_m)}{\tau_m(r_m)}\right). \quad \text{(A13)}$$

$$\frac{\partial E(N_m)^2}{\partial r_m} = (2E(N_m)E(N_m^2) - 2E(N_m)^3)\left(\frac{\tau_m'(r_m)}{\tau_m(r_m)}\right). \quad \text{(A14)}$$

Using the fact that the efficiency is an increasing function of the resources assigned to the machine, we obtain

$$\frac{\tau_m'(r_m)}{\tau_m(r_m)} = \frac{-\eta_m'(r_m)}{\eta_m(r_m)} \leq 0. \quad \text{(A15)}$$

$$\left(\frac{\tau_m''(r_m)}{\tau_m(r_m)}\right) - \left(\frac{\tau_m'(r_m)}{\tau_m(r_m)}\right)^2 = \frac{-\eta_m''(r_m)}{\eta_m(r_m)} + \frac{\eta_m'(r_m)^2}{\eta_m(r_m)^2}. \quad \text{(A16)}$$

It immediately follows from Equations (A11) and (A15) that,

$$\frac{\partial E(V_m)}{\partial r_m} \leq 0. \quad \text{(A17)}$$

As efficiency is assumed to be an increasing concave function, we may ignore

$$\frac{-\eta_m''(r_m)}{\eta_m(r_m)},$$

in (A15). Thus using Equations, (A12), (A13), (A14), (A15) and (A16), in order to test the sign of the second derivative in (A12), it suffices to examine the sign of the expression

$$(E(N_m^3) - E(N_m^2)(3E(N_m) - 1) + E(N_m)^2(2E(N_m) - 1)). \quad \text{(A18)}$$

Using the moment-inequality (Feller [33, 8.10, p. 153]), we have the bound

$$E(N_m^3)E(N_m) = \alpha E(N_m^2)^2 \quad \alpha \geq 1. \quad \text{(A19)}$$

Using (A19), and setting $t = E(N_m^2)/E(N_m)^2$, we can rewrite (A18) as,

$$\alpha E(N_m^2)^2/E(N_m) - E(N_m^2)(3E(N_m)-1) + E(N_m)^2(2E(N_m)-1),$$

$$= \frac{\left(\alpha E(N_m)t^2 - t(3E(N_m) - 1) + (2E(N_m) - 1)\right)}{E(N_m)^2}. \quad \text{(A20)}$$

Note that $t$ is greater than or equal to unity. The expression given in (A20) is non-negative either: (1) if $\alpha = 1$ and $t = 1$ or $t \geq 2$; or (2) $\alpha \geq 9/8$.

Remark: in the former case, i.e., (1), the equation

$$E(N_m)t^2 - 3E(N_m)t + 2E(N_m) = 0,$$

has two roots at $t = 1$ and $t = 2$, and in the latter case, (2), the equation

$$\alpha E(N_m)t^2 - 3E(N_m)t + 2E(N_m) = 0,$$

has no real roots. The condition given in (2) is achieved for constant (line independent) service rates and $\alpha$ can be shown to be 1.5. Thus these conditions are not limiting, and under either of them the objective function remains a separable convex function. In general, it can be shown that the expression in Equation (A8) is equal to

$$E[(N_m - E(N_m))^3 + (N_m - E(N_m))^2],$$

and is positive if the distribution of the number of jobs at the machine is either symmetric or skewed to the right. Note that we have not made any assumptions about the service rates in this analysis, except that they are increasing in $r_m$.

It is also of interest to note that if we substitute the moments of a gamma distribution of parameter $r$ in Equation (A18), we will be left with checking the sign of,

$$r(r+1)(r+2) - r(r+1)(3r-1) + r(r+1)(2r-1)$$
$$= r(r+1)(r+2-(3r-1) + (2r-1) = 2r(r+1) \geq 0.$$

## Biographies

Sridhar Seshadri is currently an Associate Professor in the Department of Statistics and Operations Research and the Operations Management Area, at the Leonard N. Stern School of Business, New York University. He received the degree of Bachelor of Technology from the Indian Institute of Technology, Madras, India, in 1978, the Post Graduate Diploma in Management from the Indian Institute of Management, Ahmedabad, India, in 1980 and a Ph.D. degree in Management Science from the University of California at Berkeley in 1993. His research interests are in the area of stochastic modeling and optimization, with applications to manufacturing, distribution, telecommunications, database design and finance. He has worked as an engineer in medium and heavy engineering firms in India and in the Middle East, and taught for five years in the Operations Management Area, at the Administrative Staff College of India, Hyderabad, India. He is an Associate Editor for *Naval Research Logistics* and *Telecommunication Systems*.

Michael Pinedo is Research Professor in Operations Management at the Sten School of Business at NYU. He received his Ph.D. in Operations Research from the University of California at Berkeley in 1978. His research focuses on the modeling of production and service systems, and, more specifically, on the planning and scheduling of these systems. He is the author of the test *Scheduling: Theory, Algorithms and Systems*, (Prentice-Hall), has recently completed a second book on *Operations Scheduling with Applications in Manufacturing and Services*, (to appear with McGraw-Hill/Irwin) and is co-author of the monograph *Queueing Networks: Customers, Signals and Product Form Solutions* (to appear with Wiley). Over the last decade he has been involved in industrial systems development. He participated in the design, development and implementation of planning and scheduling systems at International Paper, Philips Electronics, Siemens, and Merck. He is Editor of the *Journal of Scheduling* (Wiley).