Division of Research
School of Business Administration

March 1988
Revised, March 1989
Revised, July 1989

ON THE OPTIMAL ALLOCATION OF SERVERS AND WORKLOADS IN CLOSED QUEUEING NETWORKS

Working Paper #565-C

Yves Dallery
Universite' Pierre et Marie Curie
Laboratoire MASI
and
Kathryn E. Stecke
The University of Michigan

Operations Research, forthcoming, 1990.

ABSTRACT

In this paper, properties are derived that are useful for characterizing optimal allocations of servers and workloads in single-class, multiserver closed queueing networks (CQNs). The problem is as follows: Suppose a particular workload is allocated to a set of servers within a subnetwork of a CQN. This set of servers is to be partitioned into several multiserver stations. The number of stations, the number of servers, and the workload allocation to each station define a configuration of this subnetwork. Thus, the problem is determining the best configuration of each subnetwork to maximize the throughput in the original CQN.

Decomposition is used to address this problem. Results are obtained for subnetworks in isolation. These results are used to solve the optimal-configuration problem. Applications of the results to design and planning problems of flexible manufacturing are also described.

1. INTRODUCTION

Queueing networks are useful tools for modeling and performance evaluation of integrated systems, such as computer systems (Sauer and Chandy [1981] and Lazowska et al. [1984]), communication networks (Reiser [1979] and Mitrani [1987]), and flexible manufacturing systems (Solberg [1977], Dubois [1983], Stecke and Solberg [1985], Buzacott and Yao [1986a, 1986b] and Dallery [1986]). Under certain assumptions, queueing networks have product form solutions (Baskett et al. [1975]). Efficient algorithms have been developed to calculate the performance parameters of such systems (Buzen [1973], Solberg [1977], Reiser and Lavenberg [1980] and Dubois [1983]). The robustness of queueing network models has been studied experimentally and shown by operational analysis (Denning and Buzen [1978], Suri [1983], Dallery and David [1984]).

There is a growing interest in using queueing network models to address the optimization problems of complex systems such as: routing optimization (Kobayashi and Gerla [1983] and Frein et al. [1988]), server allocation (Vinod and Solberg [1985], Dallery and Frein [1986] and Shanthikumar and Yao [1987, 1988]), workload allocation (Stecke and Morin [1985]), and both workload and server allocation (Stecke and Solberg [1985]).

In this paper, some properties are derived pertaining to optimal allocations of servers and workloads in single-class, multiserver closed queueing network models (CQN). In particular, consider a set of subnetworks of the original CQN. The problem is to determine the best configuration of each subnetwork that yields the highest throughput for the overall original closed queueing network, where the number of stations, the number of servers, and the workload allocated to each station defines a possible configuration of each subnetwork.

In general, a particular workload is allocated to a set of servers that comprise a subnetwork of the original CQN. This workload is fixed and given, and may be a result of this set of servers being of a particular server type. This set of servers can be (and may have to be) partitioned into several multiserver stations and the total workload shared among these stations.

The paper is organized as follows. In §2, the parameters of the closed queueing network models under consideration and some results that are used throughout the paper are presented. In §3, the optimal-configuration problem is stated in detail. §4 provides results pertaining to subnetworks in isolation. In §5, these results are used to solve the optimal-configuration problem for the overall original closed queueing network. Finally, §6 describes the applicability of these results in the design and planning problems of flexible manufacturing systems.

2. THE CLOSED QUEUEING NETWORK MODEL

Consider a single-class closed queueing network consisting of M stations, i = 1,...,M, and N customers. Each station i is composed of a queue and a service facility with a number of identical servers, s_i. The network is characterized by the average visit ratio at station i, v_i, the mean service time at station i, t_i, and the relative service rate of station i when n customers are present, $r_i(n)$, given that $r_i(1) = 1$. The visit ratios are a solution of the following set of equations, where p_{ij} is the average proportion of customers joining station j after completion of their service at station i:

$$v_i = \sum_{j=1}^{M} p_{ji} v_j, \quad \text{for } i = 1,...,M. \tag{1}$$

The relative workload of station i is $w_i = v_i t_i$. For a station with s_i servers, the relative service rate is given by $r_i(n) = \min\{s_i, n\}$, $1 \le n \le N$.

Under certain assumptions, this network has a product form solution. Then, the proportion of time that the system is in state $\vec{n} = (n_1,...,n_i,...,n_M)$, where n_i is the number of customers at station i, is given by:

$$p(\vec{n}) = \frac{1}{G(N)} \prod_{i=1}^{M} \prod_{n=0}^{n_i} f_i(n), \tag{2}$$

$$\text{where } f_i(n) = \begin{cases} 1, & n = 0; \\ \dfrac{w_i}{r_i(n)}, & n \ge 1, \end{cases}$$

and G(N) is a normalizing constant such that $\sum_{|\vec{n}|=N} p(\vec{n}) = 1$.

This product form holds under a multiplicity of stochastic assumptions regarding to service distributions, scheduling disciplines, and customer routing (Gordon and Newell [1967] and Baskett et al. [1975]) and also under the homogeneity assumptions of operational analysis (Denning and Buzen [1978] and Dallery and David [1984, 1986]). A queueing network having a product form solution is said to be separable. In the following, the queueing networks considered here are separable.

One of the most important system performance measures is the throughput of the queueing network, TH(N) = G(N-1)/G(N). If a particular station, say 1, is chosen such that its visit ratio is $v_1 = 1$, then TH(N) can be interpreted as the average number of customers leaving station 1 per unit of time. For manufacturing applications, the throughput corresponds to the production rate.

Consider the aggregation property introduced by Chandy et al. [1975], which enables us to replace any subnetwork SN by an equivalent station e. The function $f_e(n)$ (see equation (2)), which characterizes this equivalent station, is defined by: $f_e(n) = \frac{1}{TH(SN,n)}$, where TH(SN,n) is the throughput of the subnetwork in isolation, i.e., considered as a closed queueing network with n customers. It was shown that if a network is separable, then the throughput of the network in which one or more of its subnetworks are replaced by an equivalent station is the same as the throughput of the original network (Balsamo and Iazeolla [1982]).

In the following sections, we compare the throughputs of any two closed queueing networks, $Q^{(1)}$ and $Q^{(2)}$, consisting of the same number of stations, the same number of customers, and such that $f_i^{(1)}(n) \geq f_i^{(2)}(n)$, for all i and n. It may seem obvious that the throughput of network $Q^{(2)}$ ($TH(Q^{(2)},N)$) would then be greater than (or equal to) the throughput of network $Q^{(1)}$ ($TH(Q^{(1)},N)$). However, this is not always true, even if both networks are separable.

This is demonstrated by the following counterexample. Let $Q^{(1)}$ be a separable queueing network consisting of 2 stations and 3 customers. The parameters of station 1

are: $v_1 = 1, t_1 = 5, r_1(1) = r_1(2) = 1$, and $r_1(3) = .5$; the parameters of station 2 are: $v_2 = 1, t_2 = 3, r_2(1) = r_2(2) = r_2(3) = 1$. The throughput of this network is: $TH(Q^{(1)},3) = 0.123$. Consider another network, $Q^{(2)}$, which is identical to $Q^{(1)}$ except that the service time at station 2 is $t_2 = 2$ instead of 3. Then the server at station 2 is faster in $Q^{(2)}$ and so is the throughput of station 2 in isolation. One would expect the overall throughput to be increased. However, the throughput of $Q^{(2)}$ is only: $TH(Q^{(2)},3) = 0.119$. Hence, increasing the speed of the server does not always result in an increase in the overall throughput. Intuitively, this can be explained as follows. Since server 2 is faster in $Q^{(2)}$, customers spend less time at station 2. As a result, the marginal probability of having the three customers at station 1 is larger in $Q^{(2)}$ than in $Q^{(1)}$. Now, when the three customers are at station 1, since the service rate of server 1 is reduced by a factor of two. Then the throughput of $Q^{(2)}$ may indeed be less than that of $Q^{(1)}$.

A similar observation is made in Suri [1985], where monotonicity is defined. A network is N-monotonic if the overall throughput is a nondecreasing function of the population, i.e., $TH(n) \geq TH(n-1)$, for $n \leq N$. A sufficient condition for a network to have the monotonicity property is that $r_i(n) \geq r_i(n-1)$, or equivalently, $f_i(n) \leq f_i(n-1)$, for all i and $n \leq N$ (Suri [1985]).

The following Theorem will be used throughout the paper in order compare the throughputs of two queueing networks. It can be proved by using the results of Shanthikumar and Yao [1986]. An alternative proof based on the product form solution is given in Dallery and Stecke [1988].

Theorem 1: For the separable networks $Q^{(j)}$, j = 1 and 2, consisting of a set of stations I and a population N, assume that:

$$f_i^{(j)}(n) \leq f_i^{(j)}(n-1), \text{ for all } i \in I \text{ and } n \leq N. \qquad (C1)$$

(a) If for all $i \in I$ and $n \leq N$, $f_i^{(1)}(n) \geq f_i^{(2)}(n)$,

then $TH(Q^{(1)},N) \leq TH(Q^{(2)},N)$.

(b) Moreover, if there exists i and n such that $f_i^{(1)}(n) > f_i^{(2)}(n)$,

then $TH(Q^{(1)},N) < TH(Q^{(2)},N)$. ■

Part (a) of Theorem 1 states that if the throughput of a station i considered in isolation is larger in $Q^{(2)}$ than its throughput in network $Q^{(1)}$ for any population n, then the overall throughput of $Q^{(2)}$ is larger than the throughput of $Q^{(1)}$. Part b states that the above property becomes a strict inequality if there is at least one station with a particular population for which the inequality between the throughputs of station i in isolation is a strict inequality.

Most queueing network models of real systems satisfy the condition (C1) of Theorem 1. For example, (C1) holds for load-independent single-server stations, multiple-server stations, and delay stations. Furthermore, if it holds for all stations of a subnetwork, then it holds for the equivalent station of this subnetwork. In the following, all networks that we consider satisfy condition (C1) of Theorem 1.

3. THE OPTIMAL-CONFIGURATION PROBLEM

In §3.1, the optimal-configuration problem is defined. The decomposition approach to solve the problem is described next in §3.2.

3.1 Problem Statement

A closed queueing network is partitioned into a set of K subnetworks. Each subnetwork SN_k, for k = 1,...,K, consists of a set of m_k stations. Each station i of SN_k is defined by its number of servers, $s_{k,i}$, and its relative workload $w_{k,i}$, for i = 1,...,m_k. Let S_k be the total number of servers in subnetwork SN_k and W_k be the total relative workload of subnetwork SN_k. These parameters are defined as:

$$S_k = \sum_{i=1}^{m_k} s_{k,i} \qquad (3)$$

$$W_k = \sum_{i=1}^{m_k} w_{k,i}. \tag{4}$$

The server-allocation vector and the workload-allocation vector, respectively, are $\vec{s}_k = (s_{k,1},...,s_{k,m_k})$ and $\vec{w}_k = (w_{k,1},...,w_{k,m_k})$.

The set of parameters m_k, $\vec{s}_k$, $\vec{w}_k$ define the *configuration* of subnetwork SN_k. A configuration of a subnetwork is said to be *balanced* with respect to the *server-allocation* if the number of servers at each station is the same, i.e., $s_{k,1} = ... = s_{k,m_k}$. A configuration is said to be *balanced* with respect to the *workload-allocation* if the average workload allocated to each server is the same, i.e., $\frac{w_{k,1}}{s_{k,1}} = ... = \frac{w_{k,m_k}}{s_{k,m_k}} = \frac{W_k}{S_k}$. The configuration of the entire closed queueing network is defined as the union of the configurations of each subnetwork.

With this framework, the optimal-configuration problem can be stated as follows: Given the number of subnetworks, and the total number of servers and the total workload of each subnetwork SN_k (K, S_k, and W_k, respectively), what is the best configuration of each subnetwork that yields the highest throughput for the whole closed queueing network?

3.2 Decomposition

Because of the large number of parameters involved in the optimal-configuration problem, finding the best global configuration or even comparing two configurations of the network is very difficult. This problem can be simplified by decomposing it into a set of optimal-configuration problems, each pertaining to a particular subnetwork in isolation.

Consider the problem of comparing two configurations of a queueing network, $C^{(1)}$ and $C^{(2)}$. Each configuration $C^{(i)}$, i = 1 and 2, is the union of the K configurations of subnetworks SN_k, say $C_k^{(i)}$. Our decomposition first examines each subnetwork SN_k in isolation and compares the two configurations $C_k^{(1)}$ and $C_k^{(2)}$. Then we determine how properties that pertain to the subnetwork in isolation can be transferred to the entire queueing network.

4. ANALYSIS OF EACH SUBNETWORK IN ISOLATION

In this section, we first present some results pertaining to comparisons between different configurations of a particular subnetwork, SN_k, in isolation. Isolation implies that this subnetwork is a closed queueing network having the same parameters as the initial network. This problem has been studied by Stecke and Solberg [1985], where the analysis is derived for a particular population n, that is, for a particular number of customers in the subnetwork. As our final goal is to provide results for the entire network, we need to compare several configurations not only for a particular population n, but for populations ranging from 1 to N.

For this purpose, we introduce the useful notion of a *dominant configuration.* Consider two configurations, $C_k^{(1)}$ and $C_k^{(2)}$. Let $TH(C_k^{(i)},n)$ be the throughput of configuration i, i = 1 and 2, of subnetwork SN_k in isolation with population n.

Definition 1. Configuration $C_k^{(2)}$ of subnetwork SN_k is said to dominate configuration $C_k^{(1)}$ if: $TH(C_k^{(2)},n) \geq TH(C_k^{(1)},n)$ for all $n \leq N$. Configuration $C_k^{(2)}$ is then called the N-dominant configuration. ■

Given any two configurations, it is always possible to compare their throughputs for a particular population n. However, if N > 1, it may be that neither is N-dominant, i.e., for some populations, the throughput of one configuration is higher, while for other populations, it can be lower. As we shall see in §5, results pertaining to a subnetwork in isolation can only be useful for the whole network if they are valid in terms of N-dominance.

We now consider comparisons between different configurations of a subnetwork. Several results pertaining to a subnetwork in isolation can be found in Stecke and Solberg [1985]. Issues of balancing or unbalancing both workloads and server allocations are discussed. Most of these results are valid only for a particular population and cannot be stated in terms of dominant configurations. There are however some results that are useful in the context of this paper. A first result (Theorem 1 in Stecke and Solberg [1985]) states

that the throughput of a configuration having a single multiple-server station is higher than the throughput of a configuration with m_k single-server stations. We now generalize this in order to be able to compare many different configurations.

Theorem 2: Consider the following two configurations:

$$C_k^{(1)} = (m_k, \vec{s}_k^{\,(1)}, \vec{w}_k^{\,(1)}), \quad \text{with } m_k \geq 2 \text{ and}$$

$$C_k^{(2)} = (1,(S_k), (W_k)), \quad \text{where}$$

$$S_k = |\vec{s}_k^{\,(1)}| = \sum_{i=1}^{m_k} s_{k,i}^{(1)} \text{ and } W_k = |\vec{w}_k^{\,(1)}| = \sum_{i=1}^{m_k} w_{k,i}^{(1)}.$$

Then: $TH(C_k^{(2)},n) \geq TH(C_k^{(1)},n)$, for any $n = 1,...,N$.

Moreover, if $n > \min_i\{s_{k,i}\}$, then $TH(C_k^{(2)},n) > TH(C_k^{(1)},n)$.

Proof: The proof is given in the Appendix. ■

Theorem 2 states that the throughput of a configuration with only one multiple-server station is greater than the throughput of any configuration with any number of multiple-server stations greater than one, *for any population n.* Throughput is strictly greater when the number of customers is greater than the number of servers at one of the stations. Therefore, the configuration with a single multiple-server station dominates any other configuration.

Now partition the set of stations of SN_k into G_k groups. Let $C_g = (m_g, \vec{s}_g, \vec{w}_g)$ be the configuration of group g. The configuration of SN_k, $C_k = (m_k, \vec{s}_k, \vec{w}_k)$, is then obtained as the union of these configurations, i.e.,

$$C_k = \bigcup_{g=1}^{G_k} C_g,$$

with $m_k = \sum_{g=1}^{G_k} m_g$, $\vec{s}_k = (\vec{s}_g,\ g = 1,...,G_k)$, and $\vec{w}_k = (\vec{w}_g,\ g = 1,...,G_k)$.

Theorem 3: Consider the following two configurations:

$$C_k^{(1)} = (m_k, \vec{s}_k^{(1)}, \vec{w}_k^{(1)}) = \bigcup_{g=1}^{G_k} (m_g, \vec{s}_g, \vec{w}_g)$$

and $\quad C_k^{(2)} = (G_k, \vec{s}_k^{(2)}, \vec{w}_k^{(2)})$,

with $\vec{s}_k^{(2)} = (|\vec{s}_g|,\ g = 1,...,\ G_k)$ and $\vec{w}_k^{(2)} = (|\vec{w}_g|,\ g = 1,...,\ G_k)$.

Then $\quad TH(C_k^{(2)},n) \geq TH(C_k^{(1)},n), \quad$ for any n.

Proof: Because of the aggregation property recalled in §2, we can replace, for $C_k^{(1)}$, all stations of each group by an equivalent station. The resulting subnetwork has the same number of stations as $C_k^{(2)}$. From Theorem 2, any equivalent station has a smaller throughput than the corresponding station in $C_k^{(2)}$, when considered in isolation. From Theorem 1, the overall throughput of SN_k is higher for configuration $C_k^{(2)}$ than for $C_k^{(1)}$. ■

Theorem 3 states that starting with a particular configuration and grouping several stations into a single station (having a number of servers and workload equal to the total number of servers and the total workload, respectively) increases throughput for the overall subnetwork. Since this is true for any number of customers, the resulting configuration dominates the initial configuration.

Now, consider the problem of comparing different configurations having the same number of stations, m_k, and the same server-allocation vector, $\vec{s}_k$, such that each station contains the same number of servers, i.e., $s_{k,i} = \frac{S_k}{m_k}$, for all i, but having different workload-allocation vectors $\vec{w}_k$. Let $C_k^{(1)}$ be any configuration and $C_k^{(2)}$ be the balanced

workload configuration, i.e., the configuration such that $w_{k,i} = \frac{W_k}{m_k}$, for all i. The following theorem was proved in Stecke and Solberg [1985] and Yao and Kim [1986].

Theorem 4: If the server-allocation is balanced, then the configuration with a balanced workload-allocation is N-dominant, i.e.,

$$TH(C_k^{(2)},n) \geq TH(C_k^{(1)},n), \qquad \text{for any } n = 1,...,N. \qquad \blacksquare$$

Theorem 5: Consider a particular (balanced or unbalanced) configuration of server-allocation $\vec{s}_k$. Then for any unbalanced configuration of workload-allocation, $C_k^{(1)}$, there exists a number of customers, N_0, such that:

$$TH(C_k^{(2)},n) > TH(C_k^{(1)},n), \qquad \text{for any } n \geq N_0,$$

where $C_k^{(2)}$ is the balanced workload-allocation configuration.

Proof: Let W_k be the total workload. Let $\vec{w}_k$ be the total workload-allocation vector of $C_k^{(1)}$, with $|\vec{w}_k| = W_k$, and let $Y_b = \max_i \left\{\frac{w_{k,i}}{s_{k,i}}\right\}$ be the maximum average workload per server of $C_k^{(1)}$. Let $Y = \frac{W_k}{S_k}$ be the average workload per server in the balanced configuration, $C_k^{(2)}$. Since $C_k^{(1)}$ is unbalanced, $Y_b > Y$. The following *upper bound* on throughputs of closed queueing networks is provided by the asymptotic bound analysis (ABA) (Denning and Buzen [1978]):

$$TH(C_k^{(1)},n) \leq \frac{1}{Y_b}.$$

The following *lower bound* is provided by the SSD (single-server disaggregation) method of Dallery and Suri [1986]:

$$TH(C_k^{(2)},n) \geq \frac{n}{W_k + (n-1)Y}.$$

It is easy to check that if N_0 is the first integer greater than the quantity $\frac{W_k - Y}{Y_b - Y}$, then for any $n \geq N_0$, $TH(C_k^{(2)},n) > TH(C_k^{(1)},n)$. ■

The following result provides a proof of Conjecture 6 of Stecke and Solberg [1985].

Corollary 6: For a totally saturated system ($n = \infty$), the optimal workload allocation is balanced, even for unbalanced server-allocation configurations.

Proof: The result follows directly from Theorem 5. ■

We now provide a conjecture pertaining to the situation where the number of servers at any station of SN_k must not exceed a given value $\bar{s}_k$.

Conjecture 7: If $s_{k,i} \leq \bar{s}_k$ and if the total number of servers s_k is such that there exists an $a_k \in \aleph$ such that $S_k = a_k \bar{s}_k$, then for any population n, the optimal configuration of the subnetwork is balanced with respect to both the server-allocation and the workload-allocation and is defined by: $m_k = a_k = \frac{S_k}{\bar{s}_k}$; $s_{k,i} = \bar{s}_k$; $w_{k,i} = \frac{W_k}{m_k}$. ■

This result is a conjecture because it is partly based on Conjecture 8 of Stecke and Solberg [1985]. This conjecture can be intuitively described with an example. Consider a subnetwork SN_k for which $S_k = 15$, $W_k = 30$, and $\bar{s}_k = 5$. The optimal configuration specified by Conjecture 7 is such that: $m_k = 3$, $\vec{s}_k = (5, 5, 5)$, $\vec{w}_k = (10, 10, 10)$. Consider any other configuration, for instance, a configuration such that: $m_k = 4$ and $\vec{s}_k = (4, 4, 4, 3)$. For any population n, the maximum throughput of the configuration (obtained with the best workload-allocation) is less than the maximum throughput of a configuration such that $m_k = 4$ and $\vec{s}_k = (5, 5, 4, 1)$ (from Conjecture 8 of Stecke and Solberg [1985]). This in turn is less than the maximum throughput of a configuration such that $m_k = 3$ and $\bar{s}_k = (5, 5, 5)$ (from Theorem 3), which is equal to the throughput of the configuration defined by Conjecture 7 (from Theorem 4).

5. OPTIMAL CONFIGURATION OF THE NETWORK

In this section, results pertaining to the comparison of different configurations of the entire closed queueing network defined in §3.1 are presented.

Theorem 8: Consider two configurations, $C^{(1)}$ and $C^{(2)}$, of a network as defined in §3.1.

(a) If for all k = 1,...,K and n = 1,...,N, $TH(C_k^{(2)},n) \geq TH(C_k^{(1)},n)$,

then $TH(C^{(2)},N) \geq TH(C^{(1)},N)$.

(b) Moreover, if there exists k and $n \leq N$ such that $TH(C_k^{(2)},n) > TH(C_k^{(1)},n)$,

then $TH(C^{(2)},N) > TH(C^{(1)},N)$.

Proof: Let $Q^{(i)}$, i = 1 and 2, be a network composed of K stations, where station k, k = 1,...,K, is the equivalent station obtained by the aggregation of the stations of subnetwork SN_k in configuration $C_k^{(i)}$. The function $f_k^{(i)}(n)$, which characterizes this equivalent station, is given by (see §2):

$$f_k^{(i)}(n) = \frac{1}{TH(C_k^{(i)},n)}, \text{ for } n = 1,...,N. \tag{5}$$

Then the throughput of $Q^{(i)}$ is equal to the throughput of the original network with configuration $C^{(i)}$, from the aggregation property (Balsamo and Iazeolla [1982]), i.e.,

$$TH(Q^{(i)},N) = TH(C^{(i)},N). \tag{6}$$

As a result, it is equivalent to compare the throughputs of $Q^{(1)}$ and $Q^{(2)}$. On the other hand, since the original network is separable and consists of a set of multiple-server stations, the functions $f_k^{(i)}(n)$ satisfy condition (C1) of Theorem 1, i.e.,

$$f_k^{(i)}(n) \leq f_k^{(i)}(n-1), \quad \text{for all k and n.} \tag{7}$$

The result follows by first applying the results of Theorem 1 to networks $Q^{(1)}$ and $Q^{(2)}$ and then using equation (6). ■

The following results solve the optimal configuration problem as stated in §3.1.

Proposition 9: The optimal configuration of a subnetwork, SN_k, is such that SN_k consists of a single multiple-server station, if this is feasible. This is independent of the configuration of the other subnetworks.

Corollary 10: The optimal configuration of the network, i.e., that which yields the highest throughput, is such that each subnetwork SN_k consists of a single multiple-server station, if this is feasible: ($m_k = 1, \vec{s}_k = (S_k)$, and $\vec{w}_k = (W_k)$).

Proposition 9 follows from Theorem 2 and Theorem 9 and Corollary 10 follows from Proposition 9. Corollary 10 states that within each subnetwork, grouping all servers into a single station maximizes the throughput. However, because of physical constraints on the system being modeled, it may not be possible to group all servers into one station.

An example of such constraints is the following: for each subnetwork SN_k, the number of servers at any station is limited, i.e., $s_{k,i} \leq \bar{s}_k$, $i = 1,...,m_k$. In such cases, it is not easy to determine the optimal configuration, in general. Even if we could determine the optimal configuration of each subnetwork, SN_k, for each population n (from Conjectures 5 and 8 in Stecke and Solberg [1985]), this would often be useless because the optimal configuration of SN_k might be different for each population n. Therefore, no configuration would be N-dominant. However, the following useful results can be provided:

Proposition 11: Consider a configuration $C^{(1)}$ of the queueing network, and a configuration $C^{(2)}$ obtained from $C^{(1)}$ by grouping at least two stations of a subnetwork into a single station. Then the throughput of the network is higher with configuration $C^{(2)}$ than with configuration $C^{(1)}$.

Proposition 12: If the server-allocation of subnetwork SN_k is balanced, then the optimal workload-allocation of subnetwork SN_k is balanced, independent of the configurations of the other subnetworks.

Corollary 13: If the server-allocations of all subnetworks are balanced, then the optimal configuration of the network is provided by a balanced workload-allocation at each subnetwork.

Proposition 11 follows directly from Theorems 3 and 8. Proposition 12 follows directly from Theorems 4 and 8. Then Corollary 13 follows directly from Proposition 12.

Finally, Conjecture 7 may be used at the entire network level in a similar way as follows.

Conjecture 14. If the number of servers at any station of subnetwork SN_k is constrained by $s_{k,i} \leq \bar{s}_k$, and if there exists an integer a_k such that $S_k = a_k \bar{s}_k$, then the optimal configuration of SN_k is defined by: $m_k = a_k$, $s_{k,i} = \bar{s}_k$, and $w_k = \frac{W_k}{a_k}$. This is independent of the configuration of the other subnetworks. Then both the workload and server-allocations are balanced.

Conjecture 14 directly follows from Conjecture 7 and Theorem 8.

Conjecture 15: If for each subnetwork SN_k, the number of servers at any station must not exceed a maximum number $\bar{s}_k$, and the total number of servers S_k is a multiple of $\bar{s}_k$, then the optimal configuration of the network is such that the configuration of each subnetwork SN_k is defined by:

$$m_k = \frac{S_k}{\bar{s}_k};\ s_{k,i} = \bar{s}_k,\ \text{for any i;}\ \text{and}\ w_{k,i} = \frac{\bar{s}_k}{S_k} W_k,\ \text{for any k.}$$

For each subnetwork, the workload and server-allocations are balanced.

Conjecture 15 follows directly from Conjecture 14.

6. APPLICATIONS TO FLEXIBLE MANUFACTURING

In this section, we demonstrate how the previous results can be used in the context of flexible manufacturing systems (FMSs). An FMS is able to process simultaneously several types of parts. During their processing in an FMS, parts are fixtured onto pallets. The total number of pallets available is a constant, N. When all operations on a part are completed, it is unloaded and a new part is input into the system. An FMS is composed of K types of resources (machining centers, inspection stations, loading/unloading (L/U) stations, a material handling system (MHS),...). In the following, we use the term machine type for each resource type, and machine for server. Each machine type k, k = 1,...,K, consists of a set of S_k identical machines. The average workload required by a part on a machine of type k, W_k, can be derived from the manufacturing plans of each part type and the prescribed production ratios (Dallery [1986]).

Such an FMS can be modeled as a closed single-class, multiserver queueing network, as described in §3.1. Each machine type k can be modeled by a subnetwork SN_k. The total number of machines in subnetwork SN_k is S_k and the total workload of subnetwork SN_k is the average workload of a part at machine type k, W_k. The number of customers (parts) is equal to the total number of pallets, N.

The configuration of each subnetwork depends on how the FMS is designed and operated. Let P(k) be the set of parts which, at a given time, are waiting for a machine of type k. The machines of type k may be partitioned into several subsets (called groups) such that any part of P(k) is waiting for any machine of a particular group. Then the configuration of the corresponding subnetwork is such that the number of stations, m_k, is equal to the number of groups. For each group, the number of servers, $s_{k,i}$, is equal to the number of machines in this group. The workload, $w_{k,i}$, is equal to the average workload of a part required in this group.

To illustrate this application, consider the following: If a part that is waiting for a particular type of machine can be processed by any machine of this type, then there is only one group. However, because of physical constraints, such as tooling, this may not be

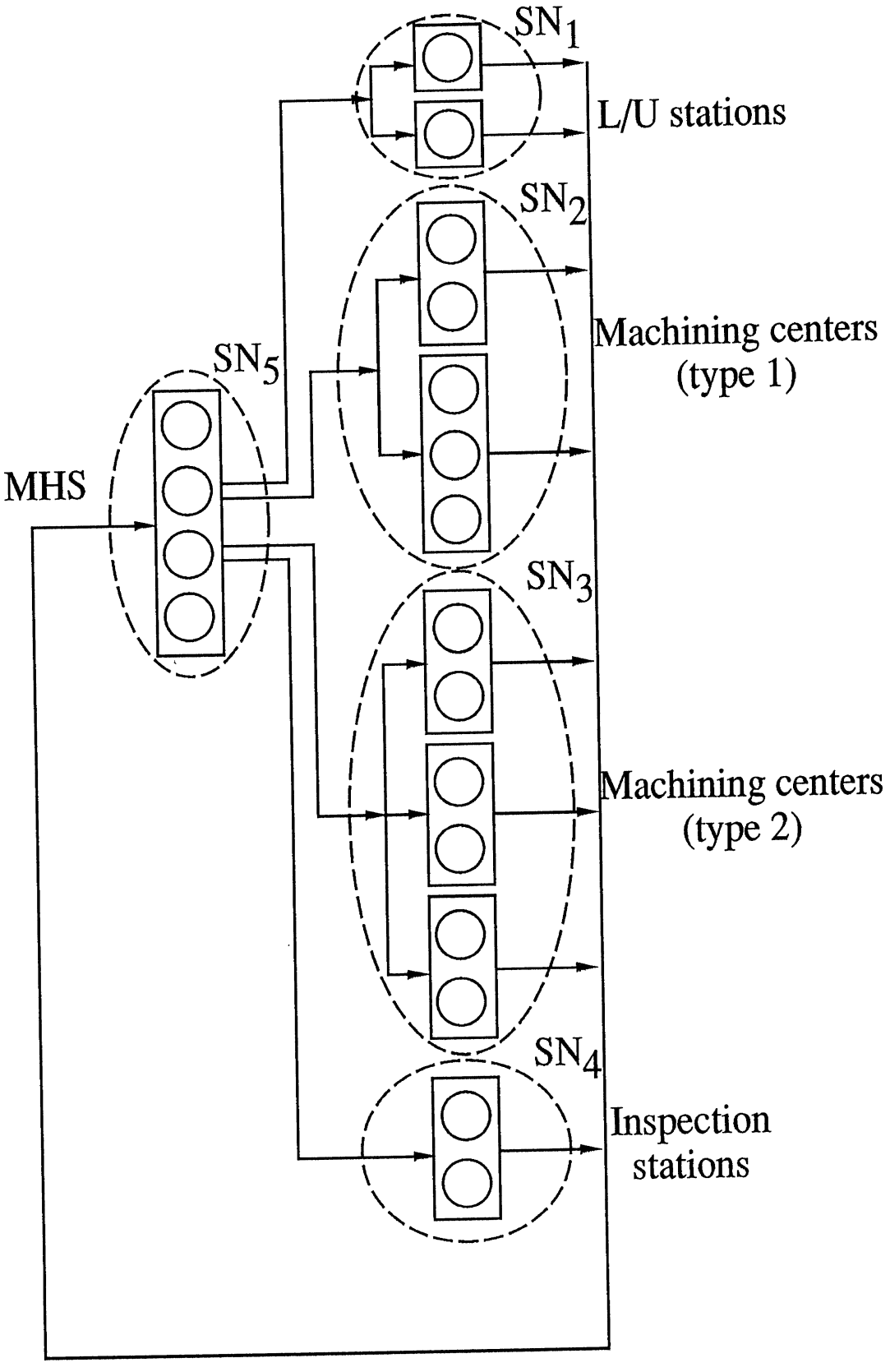
possible. For instance, it is usually impossible to place all cutting tools required for all operations to be performed by a particular machine type in only one limited-capacity tool magazine. As another example, there may not be a single storage area for all machines. Therefore the parts in P(k) would be located at several local storage areas, each feeding a group of machines. In both cases, partitioning of machines into groups must occur. The results presented in the previous sections can then be used to determine the optimal configuration of the FMS.

Figure 1 provides an example FMS configuration. The FMS consists of two L/U stations, five machining centers of a first type, six machine tools of a second type, two inspection stations, and an MHS consisting of four carts. The machining centers of type 1 (subnetwork SN_2) are distributed into two groups, one with two machines and the other group with three machines.

The FMS can be modeled in terms of the queueing network described in §3.1. Then the results provided in §5 can be used to determine the best FMS configuration. Corollary 10 says that the optimal FMS configuration is where all machines of a particular type are gathered into a single group. However, as previously mentioned, because of technological considerations, this is not always possible. In such cases, the following conclusions follow from the Theorems and Conjectures of §5:

1. For any machine type for which there are no physical constraints, group all machines into a single group (Proposition 9).
2. For any machine type for which the group sizes are prescribed and equal for the different groups, then the optimal workload-allocation is balanced within this machine type (Proposition 12).
3. For any machine type for which the maximum group size is limited and such that the total number of machines of this type is a multiple of this maximum size, then the server-allocation is balanced and each group size is equal to the maximum group size. The workload-allocation is balanced (Conjecture 14).

FIGURE 1. An example FMS configuration.



ACKNOWLEDGEMENTS

We are grateful for the detailed and helpful comments of the anonymous referees on an earlier version of this manuscript. Kathy Stecke also acknowledges a Summer Research Grant from the Business School of The University of Michigan.

BIBLIOGRAPHY

Balsamo, S. and G. Iazeolla, "An Extension of Norton's Theorem for Queueing Networks", *IEEE Transactions on Software Engineering*, Vol. SE. 8, No. 4, pp. 298-305 (1982).

Baskett, F., K. Mani Chandy, R. R. Muntz, and F. G. Palacios, "Open, Closed, and Mixed Networks of Queues with Different Classes of Customers", *Journal of the Association of Computing Machinery*, Vol. 22, pp. 248-260 (1975).

Buzacott, John A. and David. D. W. Yao, "Flexible Manufacturing Systems: A Review of Analytical Models", *Management Science*, Vol. 31, pp. 890-905 (1986a).

Buzacott, John A. and David. D. W. Yao, "On Queueing Network Models of Flexible Manufacturing Systems", *Queueing Systems Theory and Applications*, Vol. 1, No. 1 (1986b).

Buzen, Jeffrey P., "Computational Algorithms for Closed Queueing Networks with Exponential Servers", *Communications of the Association of Computing Machinery*, Vol. 16, pp. 527-531 (1973).

Chandy, K. Mani, U. Herzog, and L. Woo, "Parametric Analysis of Queueing Networks", *IBM Journal of Research and Development*, Vol. 19, pp. 43-49 (January 1975).

Dallery, Yves, "On Modeling Flexible Manufacturing Systems Using Closed Queueing Networks", *Large Scale Systems*, Vol. 11, pp. 109-119 (1986).

Dallery, Yves and Rene' David, "Some New Results on Operational Analysis", *Performance '84*, (E. Gelenbe, Editor), North-Holland, Amsterdam, pp. 119-134 (1984).

Dallery, Yves and Rene' David, "Operational Analysis of Multiclass Queueing Networks", *Proceedings of the IEEE Conference on Decision and Control*, Athens, Greece (1986).

Dallery, Yves and Yannick Frein, "An Efficient Method to Determine the Optimal Configuration of a Flexible Manufacturing System", *Annals of Operations Research*, Vol. 15, pp. 207-225 (1988).

Dallery, Yves and Kathryn E. Stecke, "On the Optimal Allocation of Servers and Workloads in Closed Queueing Networks with Applications to Flexible Manufacturing", Working Paper No. 565, SBA, The University of Michigan, Ann Arbor, MI (March 1988).

Dallery, Yves and Rajan Suri, "Approximate Disaggregation and Performance Bounds for Queueing Networks with Multiple-server Stations", *Performance Evaluation Review*, Vol. 14, No. 1, pp. 111-128 (May 1986).

Denning, Peter J. and Jeffrey P. Buzen, "The Operational Analysis of Queueing Network Models", *Computing Surveys*, Vol. 10, No. 3, pp. 225-262 (1978).

Dubois, Didier, "A Mathematical Model of a Flexible Manufacturing System with Limited In-process Inventory", *European Journal of Operational Research*, Vol. 14, No. 1, pp. 66-78 (September 1983).

Frein, Yannick, Yves Dallery, Jean-Jacques Pierrat, and Rene' David, "Optimisation du Routage des Pièces dans un Atelier Flexible par des Méthodes Analytiques", *RAIRO APII,* Vol. 22, pp. 489-508 (1988).

Gordon, W. J. and G. F. Newell, "Closed Queueing Networks with Exponential Servers", *Operations Research,* Vol. 15, pp. 252-267 (1967).

Kobayashi, H. and M. Gerla, "Optimal Routing in Closed Queueing Networks", *ACM Transactions on Computers,* Vol. 1, No. 4 (1983).

Lazowska, Edward D., John Zahorjan, G. Scott Graham and Kenneth C. Sevcik, *Quantitative System Performance: Computer System Analysis Using Queueing Network Models,* Prentice-Hall, Englewood Cliffs, NJ (1984).

Mitrani, Isi, *Modeling of Computer and Communication Systems,* Cambridge University Press, Cambridge, UK (1987).

Reiser, Martin, "A Queueing Network Analysis of Computer Communication Networks with Window Flow Control", *IEEE Transactions on Communications,* Com. 27, No. 8, pp. 1199-1209 (1979).

Reiser, Martin and Steven S. Lavenberg, "Mean Value Analysis of Closed Multi-chain Queueing Networks", *Journal of the Association of Computing Machinery,* Vol. 27, No. 2, pp. 313-323 (1980).

Sauer, C. H. and K. Mani Chandy, *Computer Systems Performance Modeling,* Prentice-Hall, Englewood Cliffs, NJ (1981).

Shanthikumar, J. George and David D. W. Yao, "The Effect of Increasing Service Rates in a Closed Queueing Network", *Journal of Applied Probability,* Vol. 23, pp. 474-483 (1986).

Shanthikumar, J. George and David D. W. Yao, "Optimal Server Allocation in a System of Multiserver Stations", *Management Science,* Vol. 33, No. 9, pp. 1173-1180 (September 1987).

Shanthikumar, J. George and David D. W. Yao, "On Server Allocation in Multiple Center Manufacturing Systems", *Operations Research,* Vol. 36, No. 2, pp. 333-342 (1988).

Solberg, James J., "A Mathematical Model of Computerized Manufacturing Systems", *Proceedings of the 4th International Conference on Production Research,* Tokyo, Japan (1977).

Stecke, Kathryn E. and Ilyong Kim, "Performance Evaluation for Systems of Pooled Machines of Unequal Sizes: Unbalancing Versus Balancing", *European Journal of Operational Research,* forthcoming (1989).

Stecke, Kathryn E. and Thomas L. Morin, "The Optimality of Balancing Workloads in Certain Types of Flexible Manufacturing Systems", *European Journal of Operational Research,* Vol. 20, No. 7, pp. 68-82 (1985).

Stecke, Kathryn E. and James J. Solberg, "The Optimality of Unbalancing Both Workloads and Machine Group Sizes in Closed Queueing Networks of Multiserver Queues", *Operations Research,* Vol. 33, No. 4, pp. 882-910 (July-August, 1985).

Suri, Rajan, "Robustness of Queueing Network Formulas", *Journal of the Association of Computing Machinery,* Vol. 30, No. 3, pp. 564-594 (July 1983).

Suri, Rajan, "A Concept of Monotonicity and its Characterization for Closed Queueing Networks", *Operations Research,* Vol. 33, No. 3, pp. 606-624 (1985).

Vinod, B. and James J. Solberg, "The Optimal Design of Flexible Manufacturing Systems", *International Journal of Production Research,* Vol. 23, No. 6, pp. 1141-1151 (1985).

Yao, David D. and S. C. Kim, "Some Order Relations in Closed Networks of Queues with Multi-server Stations", *Naval Research Logistics Quarterly,* Vol. 34, No. 1, pp. 53-66 (1987).

APPENDIX

Proof of Theorem 2: The throughput of configuration $C_k^{(2)}$ is:

$$TH(C_k^{(2)},n) = \begin{cases} \dfrac{n}{W_k}, & n \le S_k; \\ \dfrac{S_k}{W_k}, & n > S_k. \end{cases}$$

Consider configuration $C_k^{(1)}$.

Case 1: $n \le \min_i\{s_{k,i}\}$.

Since the number of customers is never greater than the number of servers at any station, no queueing occurs. Therefore, the overall response time at each station i, $i = 1,...,m_k$, is $R_i = w_{k,i}$. Therefore,

$$TH(C_k^{(1)},n) = \frac{n}{\sum_{i=1}^{m_k} R_i} = \frac{n}{W_k} = TH(C_k^{(2)},n), \quad \text{for any n.}$$

Case 2: $\min_i\{s_{k,i}\} < n \le S_k$.

Without loss of generality, we can assume that $s_{k,1} = \min_i\{s_{k,i}\}$. This subnetwork in isolation has a product form solution. From equation (2), it is easy to check that $p(\vec{n}) > 0$ for any $\vec{n}$ such that $|\vec{n}| = n$. As a result, we have $p(n_1 > s_{k,1}) > 0$, where $p(n_1 > s_{k,1})$ is the proportion of time that station 1 has more than $s_{k,1}$ customers. Then queueing does occur at station 1 and therefore $R_1 > w_{k,1}$. We also have $R_i \ge w_{k,i}$, for any $i \ne 1$. Therefore,

$$TH(C_k^{(1)},n) = \frac{n}{\sum_{i=1}^{m_k} R_i} < \frac{n}{\sum_{i=1}^{m_k} w_{k,i}} = \frac{n}{W_k} = TH(C_k^{(2)},n), \quad \text{for any } \vec{w}_k.$$

Case 3: $n > S_k$.

Without loss of generality, we can assume that $\frac{s_{k,1}}{w_{k,1}} = \min_i \left\{\frac{s_{k,1}}{w_{k,1}}\right\}$. Again from product form, we obtain $p(n_1 < s_{k,1}) > 0$, where $p(n_1 < s_{k,1})$ is the proportion of time that station 1 has less than $s_{k,1}$ customers. Then at least one server of station 1 is idle. Therefore, throughput is strictly less than the maximum throughput that could be obtained if all servers of station 1 were always busy, i.e., $\frac{s_{k,1}}{w_{k,1}}$. Then we obtain:

$$TH(C_k^{(1)},n) < \frac{s_{k,1}}{w_{k,1}} \le \frac{\sum_{i=1}^{m_k} s_{k,1}}{\sum_{i=1}^{m_k} w_{k,1}} = \frac{S_k}{W_k} = TH(C_k^{(2)},n), \quad \text{for any } \vec{w}_k. \qquad \blacksquare$$