REVUE FRANÇAISE D'AUTOMATIQUE, D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

LIE-FERN HSU
CHARLES S. TAPIERO
CINHO LIN

Network of queues modeling in flexible manufacturing systems : a survey

Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle, tome 27, n° 2 (1993), p. 201-248.

<http://www.numdam.org/item?id=RO_1993__27_2_201_0>

Recherche opérationnelle/Operations Research
(vol. 27, n° 2, 1993, p. 201 à 248)

NETWORK OF QUEUES MODELING IN FLEXIBLE MANUFACTURING SYSTEMS: A SURVEY (*)

by Lie-Fern Hsu ([1]), Charles S. Tapiero ([2]) and Cinho Lin ([3])

Abstract. – *Queueing theory in the past and still to-day, been used intensively for the design and the analysis of Flexible Manufacturing Systems (FMS's). The queueing approach, combines on the one hand important theoretical developments in queueing theory and network of queues and extensive applications to the performance analysis of information systems. The purpose of this paper is to survey the application of queues to the modelling of FMS's and to the management of such manufacturing systems, based on the stochastic environment presumed by networks of queues.*

Keywords: *Queues; Networks; Flexible Manufacturing Systems.*

Résumé. – *Les files d'attentes ont été par le passé et sont jusqu'à aujourd'hui utilisées intensément pour la conception et l'analyse d'ateliers flexibles. Les approches utilisées combinent d'une part un nombre important d'études faites à partir des processus stochastiques et les réseaux de files d'attente et des applications aux systèmes informatiques. Le but de cet article est de résumer l'application des files d'attentes à la modélisation des ateliers flexibles et aux problèmes de gestion industrielle qu'ils engendrent.*

Mots clés : *Files d'attentes; Réseaux; Ateliers flexibles.*

1. INTRODUCTION

Queueing models have been extensively used for the analysis and the design of flexible manufacturing systems (FMS), recognizing on the one hand the simultaneity of job flows in a manufacturing job shop and on the other the stochastic characteristic of job flows and process times. As a result, impetus is now given to research and mostly applications of queueing theory in manufacturing (following an earlier use of queueing models in operations management, Morse, 1963; Prabhu, 1965; Schrage, 1968, 1970, 1974; Schrage and Miller, 166; Jackson, 1957, 1963; Gordon

(*) Received july 1991.

([1]) Department of Management Baruch College, The City University of New York.

([2]) Department of Production, Logistics, and Information Systems and Decisions, ESSEC, Cergy Pontoise, France.

([3]) Department of Industrial Management Science National Cheng Kung University, Taiwan, Republic of China.

Recherche opérationnelle/Operations Research, 0399-0559/93/02 201 48/$ 6.80

and Newell, 1967; Conway *et al.*, 1967; Cooper, 1972; Newell, 1971, 1980; Kleinrock, 1976; and the extensive bibliographical list in Crabill, Gross and Magazine, 1977). Additional references for such applications abound and are partially listed in many references at the end of this paper. Authors such as Sobel, Buzacott, Shanthikumar, Solberg, Stecke, Dubois, Yao, Stidham, Yechiali, Kekre, Hsu, Tapiero and others (*see* section 2) have made additional contributions to the analysis and design of manufacturing systems. A survey is given by Buzacott and Yao (1986) for example on the modeling of FMS using queueing.

Along with simulation, networks of queues provide a modeling framework which is practically helpful to the design and analysis of manufacturing systems. Particularly, preliminary models based on tractable queueing networks can provide a departure point for the study of far more complex models through simulation, providing thereby validation for these complex models.

Application of networls of queues (assuming some strong assumptions regarding the manufacturing process) can provide steady state performance measures regarding machines (or machine group) utilization, expected production rates, mean queue lengths, bottlenecks detection, etc. For example, it has been used to improve the grouping of operations in a job shop (combined with group technology principles) and thus a rationality for the layout of FMS facilities (Co, Wu and Reisman, 1988).

Even though FMSs are systems which are extremely well controlled and therefore cannot be viewed as stochastic, the usefulness of network of queues orignates in the fact that over widely varying demand/load ranges and multiplicity of part types processed, the manufacturing system *behaves on the aggregate as if it were subject to a stochastic behavior which is not always captured by traditional models.*

Today it is agreed that disturbances such as breakdowns, fluctuating demand patterns, quality control, maintenance, etc. should be managed and reduced to render the production process a continuous flawless flow process, integrated and coordinated into a viable whole. To study these effects and the issues of designing, for example, capacity plans, basic sheduling rules, buffer capacity, loading/unloading machinery, etc., and managing them, queueing models may be extremely useful.

They may be needed to provide insights regarding the long run response of a manufacturing system to a given set of operating conditions and a preliminary appreciation of the managerial procedures used in the control of operations.

On the down side, these models which are difficult to analyze, particularly when reasonable assumptions are made regarding the manufacturing process (such as non-Poisson jobs arrivals, limited buffer capacities, application of scheduling techniques seeking to improve system's performance, batch arrival, breakdowns and unreliabilities as well as the relative importance of set-up time and costs in manufacturing systems which are apparently not as important in computer systems). Further, job shops managed for the most part in real time and which are subject to continual changes in work orders and plans cannot be analyzed by queueing models without a critical view of such results. For these resons, an overall appreciation of what queueing theory and its approximations can do for manufacturing systems, is both useful and needed for FMS analysts (although a wide variety of approaches are developed which seem to relieve some of the difficulties encountered when using queueing models. For example, refer to Ho and Cao, 1983; and Cassandra, 1984 and Ho (1985) for an extensive literature survey on Perturbation Analysis).

The purpose of this paper is to provide an outline of queueing-networks modeling in manufacturing and provide a classification which can be used to guide the FMS researcher and planner to what is known and not known in queueing analysis for FMS management. We begin first by a simple descriptive view of queue-like manufacturing cells which are then generalized to a network of queues. Then we review a large number of published papers based on queueing networks.

2. MODELING A FLEXIBLE MANUFACTURING SYSTEM

2.1. Modeling a manufacturing cell (station)

The fundamental elements of a queue-like manufacturing cell can be described by the external input process (or the arrival pattern), the buffer area, loading (*i. e.* the rule by which jobs leave the waiting line and enter for service) as well as the service pattern which expresses how jobs are processed or manufactured (in terms of time, service capacity and the number of servers). The essential ingredients of a manufacturing cell are represented in Figure 1 which is self explanatory.

To use queueing theory, however, it is necessary that we characterize models by basic constituent elements about which specific assumptions are made. The specificity of these assumptions may render the model tractable, while limiting their generality. These elements are defined below.

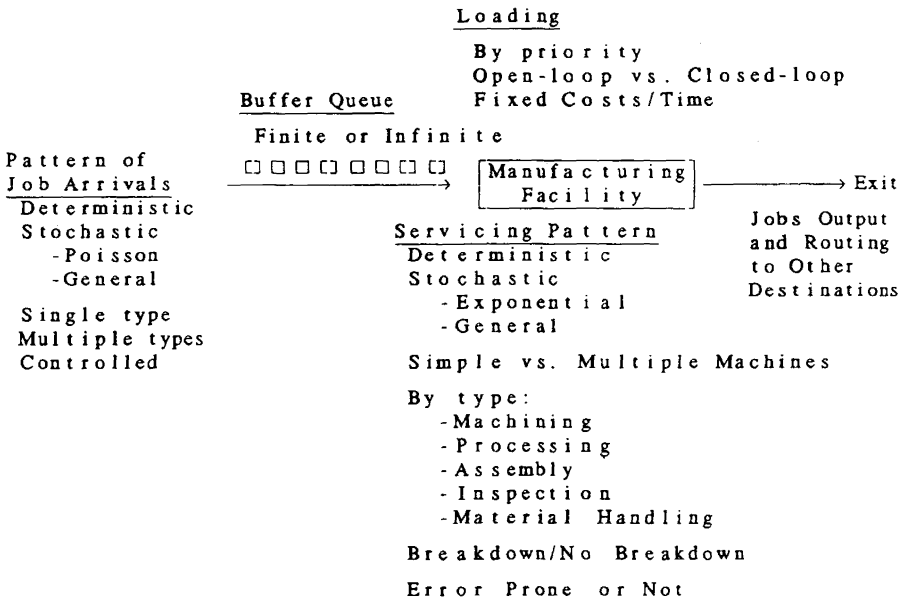

Figure 1. – A queue-like manufacturing cell.

2.1.1. *Input Process or Arrival Pattern (I)*

The arrival pattern or input to a queueing system is often measured by the mean inter arrival time. In general, we use M to symbolize Poisson input, and G to symbolize a general input. For example, the following notations are used:

D, a deterministic inter arrival time;

E_k, an inter-arrival time having the Erlang distribution function with phase k;

H_k, an inter-arrival time having the hyper-exponential distribution function with parameter k, *i. e.*, a mixture of k exponential functions, etc.

Although arrival processes are often assumed to be Poisson, or at least described by renewal processes, job inter-arrival times are dependent (due to application of scheduling rules in job shops). Further, arrivals may occur in batches (lots) of varying sizes. This situation is termed bulk or batch arrivals (e. g., *see* Chaudhry and Templeton, 1983; Chiamsiri and Leonard, 1981; Delbrouck, 1970). Under a bulk-arrival assumption, the number of customers in a batch may be either deterministic or probabilistic. We use superscripts B_D and B_P to represent deterministic and probabilistic bulk arrivals respectively, *i. e.*, M^{B_D} and M^{B_P} will represent a Poisson input with

deterministic and probabilistic bulk arrivals respectively. If we write G^{B_D} and G^{B_P}, then this represents a general input with deterministic and probabilistic bulk arrivals respectively. Such notation, as will become evident, is important to characterize precisely the mathematical nature of the queueing model. No M^{B_P}, G^{B_D}, G^{B_P} FMS models are discussed in this survey however.

There are situations where jobs are impatient. For example, if the waiting time to service is too long, then a customer may balk (Naor, 1969; Yechiali, 1971; Kleinrock, 1967). Similarly, jobs in process for too extended periods of time, may be retrieved. This situation is called reneging. There may also be parallel waiting lines (with various priorities, waiting space and rules) with waiting jobs jockeying for position (*i. e.*, switching) from one line to another. Queues with impatient customers are represented by subscripts B, R, and J. In other words, M_B, M_R, and M_J denote the Poisson input process with balking, reneging, and jokeying respectively. G_B denotes the general input process with balking, etc. Again, no queue-like FMS models belong to the situations of "R" and "J".

2.1.2. *Service Process or Manufacturing Time Requirements (S)*

Service patterns or manufacturing time and requirements are also defined by the time required to process a job. Processing, however, is conditional on the manufacturing system being not empty. Service can be deterministic or probabilistic. We use M to represent exponential service times, L to represent those services with rational Laplace transforms, and G to represent a general service distribution. Just as is the case for input processes, G may also include D, E_k, H_k, etc.

Services may be also described as single or batch service. For example, batch services can represent jobs which are processed in lot sizes. Again, as for input processes, we use the superscript B to represent batch service. Thus, M^B means an exponential bulk service, G^B means a general bulk service, etc.

A service rate may be state dependent, *i. e.*, depending on the number of jobs waiting to be processed. For example, if a queue length is increasing, an operator may work faster or, conversely, he may get flustered and become less efficient. Queues with "impatient" jobs may be viewed as state-dependent arrivals, since the arrival pattern depends on the number of customers in the system. Therefore, we may use subscript D to represent state-dependent service. For example, M_D and G_D denote a state-dependent service with an exponential, and general probability distribution respectively.

Recently, attention has been given to the study of a new class of queueing models, where arriving jobs may require several "servers" at the same time

(e. g., *see* Brill and Green, 1984; Green, 1980; Smith and Whitt, 1981; Whitt, 1985; Courcoubetis and Varaiya, 1984). Servers may be shared by several queues as is the case in cyclic service queues (e. g., Cooper, 1977; Takagi, 1986; Kaufman, 1981; Klimov, 1974; Kuehn, 1979b) or may take vacations (e. g., Doshi, 1986; Keilson and Servi, 1987, 1988; and many recent papers published in Operations Research and Computer Systems Performance Evaluation Review). These models are increasingly important for the study of manufacturing systems. However, no vacation model is surveyed in this paper.

2.1.3. *Number of Service Channels (m)*

The number of service channels refers to the number of parallel servers. There may be a single server (*S*), multiple servers (*M*), random servers and servers with vacations as pointed out earlier. For multiple servers (operators), it is generally assumed that they operate independently of each other. Studies relating to random servers (Brill and Green, 1984; Smith and Whitt, 1981), vacations (Doshi, 1986; Keilson and Servi, 1987) of increasing usefulness in manufacturing, are emerging now in great numbers.

2.1.4. *Waiting of Buffer Capacity (K)*

In some manufacturing systems as with Kanban and Just-in-Time systems, there is a physical limitation to the amount of waiting space. Beyond certain spaces, jobs may no longer join the queue unless a space becomes available by the departure of a customer. These are referred to as queues with finite waiting capacity, and can be interpreted as queues with forced balking (unless arrivals are properly controlled and coordinated as conventionally attempted through MRP II and JIT systems). In general, we use *I*, *F* to indicate whether the waiting capacity is infinite or finite. Such models are used to describe manufacturing systems of the blocking type (*see* Akyldiz, 1988; Altiok and Perros, 1984; Foster and Perros, 1980; Kaufman, 1981; Konheim and Reiser, 1976, 1978; and more particularly Perros, 1984; and Whitt, 1985; for example).

2.1.5. *Queue Dicipline or Scheduling (Q)*

The queue discipline refers to the order in which jobs in the system are processed. The most common and simplest discipline is first-come-first-served (*FCFS*). Other queue disciplines often encountered in manufacturing include last-come-first-served (*LCFS*), random order (*RO*), priority (three types of priority are used–strict priority (*SP*) by which jobs are served according to a fixed scheme of assigned priority; alternating priority (*AP*)

by which jobs are served according to their job class in a cyclic order: head of line (*HOL*) by which each part type may be assigned different priorities at different stations), processor sharing (*PS*), no queue (*NQ*) or infinite server (*I*), SPT (shortest processing time first), *LPT* (largest processing time first), *EDD* (earliest due date first), etc. Early and extensive study of such problems in a manufacturing context have been made by Schrage (1968, 1970, 1974), Gittins and Nash (1974), Harrison (1975), Jaiswal (1968), Lam (1976), Stecke (1984, 1985), Stecke and Solberg (1981), Stidham (1978, 1985), Whinston (1977 *a*, 1977 *b*, 1977 *c*), Conway *et al.* (1967) and others (*see* Courcoubetis and Varaiya, 1984; Meilijson and Yechiali, 1977; Nash and Weber, 1982; Pennotti and Schwartz, 1975; Rubin, 1975). There are other and more general disciplines (*GD*). For example, for priority disciplines, there are two general situations, preemptive and nonpreemptive. In the preemptive case, a job with the highest priority is allowed to enter service immediately, even if a job with lower priority is already in service when the higher priority job enters the system. In other words, the lower priority job in service is preempted, his service stopped, to be resumed again after the higher priority job is served. For the preemptive case, there are two possible variations: a job preempted can be either continued from the point of preemption or start anew. In the nonpreemptive case, the highest priority job goes to the head of the queue but cannot get into service until the job presently in service is completed, even though this job has a lower priority. We use *–P* and *–NP*, in connection with the priority queue disciplines, representing the preemptive and nonpreemptive cases respectively. The number of priority classes is necessarily greater than or equal to two. These five elements (*i. e.*, input process, service process, number of service channels, waiting or buffer capacity and queue discipline) are essential to describe a single queue (station or cell). We will use the vector $C=[I, S, m, K, Q]$ to describe it.

2.2. Modeling an integrated system

2.2.1. *Number of Nodes or Manufacturing Cells/Group (N)*

The number of nodes in a queueing network represents the number of manufacturing cells in an *FMS*. Unlike parallel servers cases, who provide the same service in all stations, each node (cell) may provide a different service. A set of such services reflects basically the operations needed to make parts, or assemble parts into finished products. For example, a "jeans" manufacturing system may consist of three cells, namely: cutting, sewing

and pressing. The number of nodes in the network can be either 1 (single) or m (multiple).

2.2.2. *Type of the Network: Open and Closed (T)*

Networks with multiple nodes can be either closed (C) (*i. e.* the number of jobs in the network is fixed wothout any external arrivals or any departure, or more precisely, when a job exits the manufacturing system it is replaced immediately), open (O) (*i. e.* with external arrivals and departures), or mixed (M) (*i. e.* open for some classes of jobs and closed for other classes).

2.2.3. *Sequence of Operations: Transfer Lines, Assembly, etc. (O)*

For multiple-node queueing networks, we can have:

a) Sequential (denoted by S) also called tandem queues, *i. e.*, each node has of most one predecessor and one successor.

b) Sequential with feedback (denoted by S_f).

c) Assembly (denoted by As), in which case each node has any number of predecessors, but at most one successor. Only open queueing networks can have this type of structure, however.

d) Arborescent (denoted by Ar), in which case each node has a single predecessor but any number of successors. Again, only open queueing networks can have this type of structure.

e) Acyclic (denoted by Ac), *i. e.* each node can have any number of predecessors and successors, but a customer cannot visit the same node more than once.

f) Cyclic or general (denoted by G), *i. e.*, feedback flows are permitted. Both open and closed queueing networks can have this type of structure. In all cases, a routing matrix is used to describe the sequence of operations.

2.2.4. *Class of Jobs (C)*

Jobs in a *FMS* network can be of the same class ($C=1$) or belong to different classes ($C=m>1$). Jobs of different classes may follow different sequences of operations (*i. e.* routes). These classes are used to distinguish between "deterministic" and "Random" *FMS*'s. These four elements (*i. e.*, number of nodes, type of the network, sequence of operations, and class of job) will be described by the vector $S=[N, T, O, C]$.

2.2.5. *Description of the System Unreliability (U)*

There are two types of unreliability: equipment failure and defectuous production. Equipment failure include failure time (F), repair time (R),

number of repair stations (N), number of servers (men) at repair station (S). For defectuous production, servers perform their function but their performance is faulty. Policies used to deal with such problems include inspection (I_p), inspection rate (I_r), reworking time (R_t) and feedback for re-processing (F_s).

We use the vector $U=[F, R, N, S, I_p, I_r, R_t, F_s]$ to describe system unreliabilities where

$I_p=CSP$ means that the model uses continuous sampling plans.

$I_p=R$ means that the model uses a random policy.

$F_s=1$ means that the defective products can be repaired.

$F_s=$O means that the defective products cannot be repaired.

2.2.6. *Description of Material Handling Systems (M)*

A material handling system (*MHS*) is an important part of an *FMS*. It can be described by the vector $M=[E, P, T]$, where

$E=R$, C, or A (we use R, C, A to represent a robot, a conveyor, or an *AGV*-automated guided vehicle).

$P=S$ or M (S stands for single type pallet and M for multiple types).

T represents the transportation time.

2.2.7. *System Controls (C)*

System controls include the routing policy (R) and methods for releasing jobs from the entry queue to the *FMS* (Re). We use the vector $K=[R, Re]$ to represent it, where

$R=R, S, D, F_r, F_e$, or PSQ

(with R=random routing, S=symmetric routing, D=Dynamic routing, F_r=fixed routing, F_e=fixed load, and PSO=probability shortest queue routing).

$R_e=IM, FCFS, B, UB, SPT\text{-}WG, PT\text{-}WFBS, SPT\text{-}WPR$

(IM=releasing job to idle machine, B=balance releasing, UB=unbalance releasing, WG=release jobs if and only if the machine is idle, WPR=release jobs to shop at prescribed review time, $WFBS$=release jobs to the shop as soon as the jobs are accumulated to a fixed level at the dispatch area).

We will use this notation to describe the more important queue-like *FMS* models in the third part of the paper.

3. MANAGERIAL ISSUES

FMS survey papers can be classified into two different types:

1) Comprehensive surveys which focus on descriptions of hardware systems (machines). Often cited references include Hutchinson (1979), American Machinist (1981), Gunn (1982), and Dupont-Gatelmand (1982).

2) Model surveys which focus on *FMS* models for the purposes of design, economic justification, and operational problems.

In the latter type, there are four important papers:

(i) Suri (1985) classified the models into two sub areas: Generative (or Prescriptive), and Evaluative (or Descriptive) Models.

(ii) Van Looveren *et al.* (1986) divided *FMS* planning problems into three levels–strategic, tactical, and operational.

(iii) Kalkunte *et al.* (1986) divided *FMS* models into 4 levels: strategic analysis and economic justification, facilities design, intermediate range planning, and dynamic operations planning.

(iv) Buzacott and Yao (1986 *a*, 1986 *b*) reviewed *FMS* models before 1986 and classified them by research problems types.

Our classification however, will be based on the managerial issues encountered in managing *FMSs*. The models we review are classified into eight problem types:

1) Model types.

2) Optimal configuration.

3) Loading models.

4) Part routing models.

5) Part selection models.

6) Releasing and scheluding models.

7) Unreliable system models.

8) Inventory models.

3.1. Models types

We classify the models into two categories:

1) Exact models (based on analytical results).

2) Approximate models (based on approximation algorithms).

For exact models, we have two classes:

1) Infinte Buffer Models (*IBM*).

2) Finite Buffer Models (*FBM*). Most formulas for these models performance characteristics are derived by reversibility theory.

Equilibrium state probabilities can be calculated by a product form formula or by closed a analytic formula. These models basically rely on the assumption of Poisson arrival and departure, which may be overly restrictive in many practical situations. Hence numerous approximations have been developed to overcome these restrictions. Five approaches are used (Bitran and Tirupati, 1988): 1) Decomposition Methods (*DM*), Approximate Mean Value Analysis (*AMVA*), 3) Operational Analysis (*OA*), 4) Exponentialization Approximations (*EA*), and 5) Diffusion Approximations (*DA*).

3.1.1. *Infinite Buffer Model (IBM)*

Jackson (1957 and 1963) pioneered the study of job shop by queueing networks. Solberg (*CAN-Q*, 1977) modeled and analyzed an *FMS* based on a closed queueing network (*CQN*) however which were developed by Gordon and Newell (1967).

Three methods–Convolution Algorithm (*CA*) (Buzen, 1972), Mean Value Analysis (*MVA*) (Reiser Lavenberg, 1980) and the *Z*-Transform Algorithm Maione *et al.* (1986) have been used to compute performance measures such as throughput, waiting length, etc. Solberg (1977) used the Convolution Algorithm to compute performance measures using a software named *CAN-Q*. Subsequently, Co and Wysk (1986) discussed the robustness of *CAN-Q* to predict the performance of *FMSs* and found that it is an expedient tool to evaluate the performance of *FMS* where no exact details are required.

Moore (1972) derived performance formulae by generating functions. Maione *et al.* (1986) extended it to a new model with multiple part types, and solved it by the *Z*-transform method. Their contribution consists in the use of convolutions of *Z*-transforms to construct a decomposition algorithm. This algorithm includes two steps: 1) decompose the *FMS* into one or more completely balanced subsystems and a completely unbalanced subsystem, 2) convolute the *Z*-transform functions of all subsystems. The algorithm leads to a powerful tool for the analysis of complex *FMS* systems. However, the coefficients of the *Z*-transform function are difficult to compute.

Recently, Solot and Bastos (1988), used the *BCMP* algorithm (Baskett, Chandy, Muntz and Palacios, 1975), proposed an exact multiple part types (several pallet types) model called *MULTIO*. This model can be easily modified to find an optimal pallets mix through which a maximum throughput can be obtained. Although all experiments in their paper appear to be more accurate than *CAN-Q* in the multiple part types model, it may not be a good

algorithm when the number of stations and (or) part types increases (because it requires more computer time and memory).

3.1.2. *Finite Buffer Model (FBM)*

According to Hatvany (1983), the storage capacity at each station (local buffer in *FMS*) is very small. In most analytical models, the buffer size is assumed to be infinite, and thus it is not necessary to consider the routing policy. Namely, using reversibility theory (Kelly, 1979), it is only necessary to prove that if a system has a reversible process then it has a steady state product form solution (Melamed, 1983). Yao (1983) and Yao and Buzacott (1985 *a* and 1987) showed that "a system has a reversible process if the routing rule belongs to the following three types: 1) Symmetric routing: all stations are equally visited, 2) Jobs are routed through a central-server, 3) Dynamic routing: the routing rate is dependent on the number of jobs at the dispatching and receiving station.

For a zero local buffer model, Yao and Buzacott (1987) showed that we still have a product form solution even with general service times. Dallery (1986) used the "global reversible routing" to model an *FMS* with a limited global capacity storage (*i. e.*, the total number of jobs at a set of stations has a fixed value), and derived a product form solution. He also showed that this model is especially fitted for an *FMS* with cells of a finite storage capacity.

Dallery and Yao (1986) considered a set of flexible manufacturing cells (*FMC*) linked together by an *MHS* and had a limited storage capacity at each cell. A *CQN* model is developed for the system where each cell is modeled as a sub-network. They also proved that the equilibrium probability distribution at the stations has a product form solution and that the *FMS* can be analyzed cell by cell by aggregating cells into a single station, and then at the system level, by the algorithms proposed by Yao and Buzacott (1985 *a* and 1985 *c*). For general service time distributions, it can be solved also by incorporating the algorithm proposed by Yao and Buzacott (1985 *b*) into the model (although the authors do not mention this property). For general service time distributions, there is no product form solution however.

There is also a Semi-Markov approach to evaluate the performance measures of an *FMS* with finite local buffer. Alam *et al.* (1985) modeled such an *FMS* with multiple servers, *K AGVs* with constant transportation times, and multiple job types. Moreover, service times (dependent on job type as well as station) are assumed to possess a rational Laplace transform. Using a Semi-Markov queueing network system, in which several states were lumped together to reduce the scale of the original state space, they

obtained an exact solution for performance measures. They also presented an approximate procedure for large *FMSs*.

Van Dijk (1989) discussed nonreversible networks with blocking, and found that two special types of *FMS* with this property have the product form solution. However, Van Dijk proved that nonreversible networks can have a product-form solution only when parts are routed for processing at each station along a cyclic order.

Another approach is suggested by Dubois (1983) who analyzed first the asymptotic behavior of the throughput and then used *IBM's* algorithm to evaluate the throughput in case of finite local buffers.

3.1.3. *Approximate Decomposition Method (ADM)*

The underlying idea of *ADM* is to decompose the network into single stations or subsets of stations and analyze each separately. To use this method, two conditions must be satisfied:

1) The nodes are treated as being independent.

2) Reasonably accurate results for means and variances can be obtained.

This method was first developed by Reiser and Kobayashi (1974). There are many papers applying this method to derive the formulae for system performance of queueing networks. In general, this method involves three steps:

1) analyze the interaction between stations;

2) decompose the network into subsystems (a station or a set of stations) and analyze them;

3) "recompose" the results obtained in step 2 to obtain performance formulae.

Here, we only review papers relating to *FMSs* only. Shanthikumar and Buzacott (1981) proposed an algorithm for an *OQN* with single product and single-sever stations whose service times are general, local buffers are unlimited, and arrival times are Poisson. They decomposed the network into a set of *GI/G/1* or *GI/M/1* queues and then approximated the arrival process at each station by a renewal process characterized by a mean and a squared coefficient of variation (*SCV*). This algorithm can be extended to *OQN* with *GI/G/C* ($c>1$) by using the formulae proposed by Whitt (1983 *a*, 1983 *b*). The algorithm can be extended also to models with multiple part types by aggregating the multiple parts into a single part (in general, we use the weight average algorithm). Shanthikumar and Buzacott, focusing on single part types have compared their approximations using simulation results.

Marchal (1985) proposed two approximations for the mean delay formulae in a queue. Then, using *ADM*, the mean delay and departure coefficient of variation of each station (cell) can be calculated and the mean time it takes a job to pass through an *FMS* is obtained.

Bitran and Tirupati (1988) modified the above algorithm and proposed three approximations to capture the interference effect of multiple part types by using the Poisson and Erlang approximations for the aggregate product. They tested these approximations in a semiconductor manufacturing factory with deterministic routing.

Unlike previous algorithms which assume infinite buffer capacity, a decomposition algorithm for finite local buffer was developed by Yao and Buzacott (1985 *a*). They assumptions are: 1) an *OQN* model of *FMS* with a centralized *MHS*; 2) limited local buffers; 3) multiple servers at each station; and 4) general service times. The main points of the algorithm are:

1) Find a relationship between the blocking probability of all stations and inter-arrival time distributions at each station.

2) Decompose the system into a set of *GI/G/C/S* queues and evaluate the performance for each station by the algorithm proposed by Yao and Buzacott (1985 *b*).

Although they did not consider the case of multiple part types, it is possible to approximate the multiple part types case by aggregating all part types into one part type.

For multiple part types and multiple servers, Shanthikumar and Buzacott (1984) suggest that we only evaluate the mean and variance of a job's "flow time" for an *OQN* with general service times and shortest process time scheduling rule. The algorithm decomposes the *FMS* into *M/G/1* queues (Shanthikumar, 1982) and calculates the mean and variance of flow time for each station. The mean and variance of flow time of the *FMS* are obtained by a convolution algorithm. This model assumes that local buffer has an infinite capacity however.

Finally, Kamath, *et al.* (1988) evaluate the system performance in a closed-loop flexible assembly system by an approximate formulae proposed by Shanthikumar and Buzacott (1980 and 1981), Whitt (1983 *a*, 1984 and 1985), Shanthikumar and Gocmen (1983 *a*), Kamath, *et al.* (1988), Kapadia and Hsi (1978).

3.1.4. *Approximate Mean Value Analysis (AMVA)*

This is a heuristic approach based on *MVA*. Schweitzer (1979) first introduced it and considered a case with a single server and a single job

type. Neuse and Chandy (1981) extended it to the multiple servers model. Krzesinske and Grcyling (1984) independently developed an improved algorithm for the same problem. Recently, a more powerful algorithm has been developed by Akyldiz and Bolch (1988). Approximate calculations of performance measures for this algorithm have less than four percent error on the average in all the examples investigated. This algorithm is also easy to implement and requires a small amount of computer run-time. The papers discussed above have focused their attention on computer systems only.

Cavaille and Dubois (1982) have also applied the *AMVA* approach to modeling an *FMS*. Other than exponential service time, they also consider deterministic service times. Their algorithm is not adequate for service time with a *SCV* greater than one however. An unreliable machine model was also discussed in their paper.

Based on Mean Value Analysis (*MVA*), Suri and Hildebrant (1984) built a computer program called *MVAQ*. By using *MVAQ* the part production rates, machine utilization, and average work-in-process inventories of an *FMS* with single part or multiple part types can be easily obtained.

There are at least two computer packages available which use a heuristic approach: *MVA-MVHEUR* (Schweitzer, 1979) and *PMVA* (Shalev-Oven, *et al.* 1985). The *MVHEUR* is based on an algorithm proposed by Schweitzer (1979). The PMVA is an extension of *MVHEUR* and considers an *FMS* with multiple servers, *FCFS*, infinite server or *HOL* (head of line, which means each part type may be assigned different priorities at different stations) and several transportation mechanisms operating with partitioned service responsibilities by which the transportation time may be dependent on actual distance.

Menga *et al.* (1984) first used the *AMVA* to analyze an "intelligent dispatching problem". Conterno *et al.* (1986) studied a routing policy by which incoming parts are sent to the work center with the "least unfinished work". They derived the "routing coefficients" and "adaptive dispatching speed-up coefficients" first, and then evaluated the system throughput by *AMVA*.

3.1.5. *Operational Analysis Approximation (OAA)*

OAA was first proposed by Buzen (1976) and extended by Denning and Buzen (1978). Four assumptions of *QAA* (Denning and Buzen, 1978) are one step behavior, flow balance, routing homogeneity, and homogeneous service time (*HST*). Their approach consists in measuring quantities directly through a data set observed on-line or through simulation and used to predict

the performance of the network based on the theory of closed exponential network models.

Dallery and David (1983 and 1986) use this approach to develop an iterative algorithm for generating a data set with which we can characterize the processing time, called apparent processing time, and then derive the performance of the *FMS* with multiple part types and multiple servers.

Fanti (1988) assessed the *OAA's* robustness in evaluating the system performance of *FMSs* and derived an algorithm to obtain the infimum and maximum throughput and utilization.

3.1.6. *Exponentialization Approximation (EA)*

This approach was first used to model a computer system (Shum and Buzen, 1977; Marie, 1979). Yao and Buzacott (1986 *a*) adopted it to analyze a *CQN FMS* model. It may be extended to networks with limited local buffer space and multiple server stations however. The approach consists in transforming the general network into an exponential network. Based on selected problems, Yao and Buzacott (1986 *a*) conclude that this algorithm is more accurate than the approximate *MVA* algorithm developed by Cavaille and Dubois (1982).

3.1.7. *Diffusion Approximations (DA)*

Diffusion approximations are based on the assumptions that the number of events in a given time interval is approximately normally distributed and that queues are almost always non empty. Application of the central limit theorem is then used to capture the variations in queue lengths and to approximate the discrete-value queueing process $N(t)$ by a continuous-path Markov process $X(t)$ (*i. e.* diffusion process). A diffusion equation is then used to describe a continuous process $X(t)$ with a reflecting boundary at $x=0$. For these reasons, DA is useful in heavy traffic conditions. By modeling the process behavior at the reflection boundary, Gelenbe (1975, 1979) developed an approximation which is less dependent on heavy traffic assumptions. Reiser and Kobayashi (1974) showed that in most cases the accuracy of the diffusion approximation is quite adequate and is much higher than results which assume the service times to be exponentially distributed. Many papers (Cox and Miller, 1965; Gaver, 1968; Newell, 1980; Gaver and Shedler, 1973 *a*, 1973 *b*; Kobayashi, 1974; Reiser and Kobayashi, 1974; Gelenbe, 1975; Gelenbe and Pujolle, 1976; Halachmi and Franta, 1977; Newell, 1980; Pujolle and Ai, 1986) have used DA to evaluate the system performance of queueing networks.

3.1.8. *Summary and Comments*

Essential contributions and models' assumptions, methodologies, as well as results derived are shown in Table 1.

Overall, we can conclude that queueing models may be a good way to evaluate the *FMS's* performance when they are at the design phase. For example, Chen *et al.* (1988) used a simple queueing network, to predicted certain key system performance measures in a semiconductor manufacturing factory and found that the predicted values are within 10% of those actually observed in the factory. Shortcomings of this model include the following:

1) Queueing models evaluate the system performance under a given configuration of the *FMS*, thus alternative configurations are difficult to reprogram and calculate.

2) Queueing models assume a steady state operation.

3) They describe the system in an aggregate way which omits the detailed operations such as transportation time, set up time, etc. which are very important in manufacturing.

Compared to stimulation models, queueing network models require relatively little input data and do not use much computer time. Therefore, they can be viewed as suitable when we need to evaluate preliminary *FMS* designs.

3.2. Optimum configuration model

The design of an *FMS* configuration involves determination of the number of machines, pallets, buffers, process rate of tools or machines, etc., under the requirement of production rates, budget limitations, etc. These problems are generally formulated as mathematical programming problems coupled with a queueing network which is used to compute a system's predicted performance. They can be classified into the following categories: Server (machine) models, buffer models, coupling model of server and buffer, coupling model of server, pallets, and work load, tool management models and layout models.

3.2.1. *Server (Machine) Model*

Vinod and Solberg (1985) first coupled the *CQN* with mathematical programming to find the optimal number of servers at each station. Dallery and Frein (1986) modified it and develop a more powerful procedure to find the optimal solutions. All of these papers modeled the *FMS* by a closed,

TABLE 1
Summary of Major Models of System Performance

Author	Assumption	Metho-dology	Main result
Solberg (1977) (CAN-Q)	*C*=[*M, M, m, I, FCFS*] *S*=[*m, C, G, M*] *M*=[–, *S*, –]* *K*=[*R*, –]	*IBM*	Throughput. Marginal distribution. Mean number of part
Shanthikumar and Buzacott (1981)	*C*=[*G, G, m, I, FCFS*] *S*=[*m, O, G, M*] *M*=[–, *S*, –] *K*=[*R*, –]	*ADM*	Same as Solberg (1977)
Cavaille and Dubois (1982)	*C*=[*Ek, D*, or *M, m, I, FCFS*] *S*=[*m, O, G, M*] *M*=[–.*M*.–] *U*=[*G, G, S, S*, –, –, –, –] *K*=[*R*, –]	*AMVA*	Same as Solberg (1977)
Dubois (1983)	*C*=[*M, M, m, F, FCFS*] *S*=[*m, O, G, S*] *M*=[–, *S*, –] *K*=[*R*, –]	*FBM*	Same as Solberg (1977)
Dallery and David (1983)	*C*=[*G, G, m, I, FCFS*] *S*=[*m, C, G, M*] *M*=[*R, M*, –] *K*=[*R*, –]	*OAA*	Same as Solberg (1977)
Shanthikumar and Gocmen (1983)	*C*=[*G, G, m, I, FCFS*] *S*=[*m, C, G, S*] *M*=[–, *S*, –] *K*=[*R*, –]	*ADM*	Same as Solberg (1977)
Yao (1983) Yao and Buzacott (1985 *a*, 1987)	*C*=[*Mb, M, m, F, FCFS*] *S*=[*m, C* or ,*O G, S*] *M*=[–, *S*, –] *K*=[*R* or *S* or *D*, –]	*FMB*	Same as Solberg (1977)
Shanthikumar and Buzacott (1984)	*C*=[*M, G, m, I, SPT*] *S*=[*m, O, G, M*] *M*=[–, *S*, –] *K*=[*R*, –]	*ADM*	Flow time for each job type
Suri Hildebrant (1984)	*C*=[*M, M, m, I, FCFS*] *S*=[*m, C, G, S*] *M*=[*C* or *Ca, S*] *K*=[*R*, –]	*IBM* *MVA*	Throughput. Machine utilization. Average W-I-P sizes.
Conterno *et al.* (1986)	*C*=[*M, M, m, I*, –] *S*=[*m, C, G, M*] *M*=[–, *S*, –] *K*=[*, –] *: least unfinished work	*AMVA*	Same as Solberg (1977)
Alam *et al.* (1985)	*C*=[*Mb, M, m, F, FCFB*] *S*=[*m, O, G, M*] *M*=[–, *S, D*] *K*=[*R*, –]	*FMB*	Same as Solberg (1977)
Shalev-Oven *et al.* (1985) (PMVA)	*C*=[*M, M, m, I, FCFS* or *I* or *HOL*] *S*=[*m, C, G, M*] *M*=[*R, M, M*] *K*=[*R*, –]	*AMVA*	Same as Solberg (1977)

TABLE 1 (*continued*)

Author	Assumption	Metho-dology	Main result
Marchal (1985)	*O*=[*G, G, m, I, FCFS*] *S*=[*m, O, G, S*] *K*=[*R*, –]	*ADM*	Mean flow time
Yao and Buzacott (1985)	*C*=[*M, M, m, F, FCFS*] *S*=[*m, O, G, S*] *M*=[*K, S, G*] *R*=[*R, FCFS*]	*ADM*	Same as Solberg (1977)
Co *et al.* (1986)	*C*=[*M, M, m, F, FCFS*] *S*=[*m, C, G, S* or *M*] *M*=[*C* or *K, S, D*] *K*=[*R*, –]	*IBM*	Robustness of *CAN-Q*
Dallery and Yao (1986)	*C*=[*Mb, M, m, F, FCFS*] *S*=[*M, O, G, S*] *M*=[*C* or *K, S, D*] *K*=[*R*, –]	*FBM*	The performance of FMC and each cell.
Yao and Buzacott (1986 *a*)	*C*=[*M, G,* or *M, m, F, FCFS*] *S*=[*m, C, G, S*] *M*=[*C* or *A, S*, –] *K*=[–, *D*, or *Fr* or *Fe*]	*EA*	Same as Solberg (1977)
Dallery (1986)	*C*=[*M, M, m, F, FCFS*] *S*=[*m*, –, *G, S*] *K*=[*R*, –]	*FBM*	Same as Solberg (1977)
Fanti *et al.* (1988)	*C*=[*G, G, m, I*, –] *S*=[*m, O, G, M*] *K*=[*R*, –]	*OAA*	Same as Solberg (1977)
Kamath *et al.* (1988)	*C*=[*M, M, M, I, FCFS*] *S*=[*m, C, As, S*] *M*=[–, *S*, –] *K*=[*R*, –]	*ADM*	Same as Solberg (1977) for a closed-loop flexible assembly systems.
Solot Bastos (1988)	*C*=[*M, M, m, I, FCFS*] *S*=[*m, C, G, M*] *M*=[–, *M*, –] *K*=[*R*, –]	*ECM*	Throughput. Margiral distribution. Server utilization. Mean waiting number of each part type. Mean spent time of each part type pallet.
Van Dijk (1989)	*C*=[–, *M* or *G, m, F, FCFS*] *S*=[*m*, –, *S, S*] *K*=[*R*, –]	*FBM*	Steady state distribution.

* –: indicates that feature is not considered.

single-job type queueing network with M stations and N customers. There are two differences between these two algorithms however:

1) The *V-S* (Vinod and Solberg) algorithm starts with an arbitrary upper bound while Dallery and Frein use a lower bound determined by asymptotic analysis.

2) The *V-S* algorithm uses a so-called "bisection search" which is similar to the bisection line search method, while Dallery and Frein use a "gradient analysis", similar to the gradient method in nonlinear programming.

It appears however that the algorithm proposed by Dallery and Frein (1986) is more efficient than the *V-S* algorithm.

An alternative approach was suggested by Shanthikumar and Yao (1988). Using a *CQN FMS*, they derived an algorithm to allocate *C* servers among *M* work centers when the *CQN* throughout is maximized. Finally, they proved an arrangement increasing property expressed in terms of server's assignment objective function (e. g., the more servers in the system, the higher the throughput will be) and derived an upper bound formula for the throughput. Thus, they solve the optimal servers assignment mathematical programming problem. They also proposed a greedy heuristic algorithm by which they obtain an optimal solution for a system with two working centers (although it cannot guarantee that the optimal solution in general cases is obtained). This approach is similar to Fox's marginal allocation scheme however (Fox 1966). The server allocation model with a maximum profit objective has been solved by Shanthikumar and Yao, 1987.

Stecke and Solberg (1986) also discussed the problem above but solved for the optimal servers allocation in a system with three-machines and two groups (stations). For cases with more than three machines, they used *CAN-Q* to calculate the throughput for all allocations in a two-stations system and showed that unbalanced configurations of allocated servers are superior to balanced ones when groups of pooled machines sizes are unequal.

3.2.2. *Buffer Model*

In practice, buffer capacity (size) at each station in an *FMS* is limited, affecting thereby the throughput. Buzacott and Shanthikumar (1980) first discussed the relationships between the local buffers space and throughput. They consider a case with two balanced machines and a single-class job shop and proved that the objective function is concave with respect to the buffer size (Yao and Shanthikumar, 1986).

So (1989) proposed a model to determine the buffer capacities required to achieve a specified performance in an *FMS* with multiple products. So's approach is based on a measure of performance often used in pull production systems; the average proportion of demands backlogged. A *GI/G/C* queue together with a decomposition method are introduced to approximate the mean and variance of the total process times (including the waiting time). Assuming a fixed blacklog. So (1989) estimated the aggregate buffer size

required for a total system demand, and then decomposed these buffers to each station.

3.2.3. *Coupling Model of Server and Buffer*

The combined server and capacity allocation problem was dealt with by Shanthikumar and Yao (1987b). Their approach is similar to that of Yao and Shantikumar (1986) and garantees an optimal solution in a single-station system only.

3.2.4. *Coupling Model of Server, Pallets and Work load*

Throughout the papers discussed above, work load allocation and the number of pallets were given. In practice, we can allocate the total work load among the machines and find the optimal number of pallets that maximize some measures of system performance. Using a simple *FMS* with only one machine type, Lee *et al.* (1989) formulated a mathematical programming model to determine the optimal number of machines (servers), pallets, and optimum work load allocation subject to requirements on the system's throughput and work load bound at each station. They derived a fathoming rule to eliminate dominated alternatives (in which a *CQN* is used to calculate the throughput). The initial feasible solution is obtained by the approach used by Dallery and Frein (1986). Although this model can yield optimal allocations of work load at each station, it assumes that each machine can perform one function only.

3.2.5. *Tool Management Model*

Using the fact that changing the tools processing rate will change the relative queue lengths in a system, Schweitzer and Seidmann (1989) proposed a processing rate optimization model for an *FMS* with multiple job visits to work centers. They formulated a mathematical programming model which is solved by *MVA* and an algorithm for problems with convex resource-allocations (Bitran and Hax, 1981).

For spare tools allocation problems, Vinod and Sabbagh (1986) presented a closed queueing network optimization model. Their man objective is to study the tradeoffs between tool spares and the cost of repairs subject to a tool constraint availability. This research does not directly deal with the performance problem of *FMS*.

3.2.6. *Layout Model*

A throughput-maximizing algorithm for facility planning and layout of *FMS* was proposed by Co, *et al.* (1989). They extended the package

of *CRAFT* (computerized relative allocation of facility) by embedding an approximate *MVA*, called *FMS-Q* into the *CRAFT*. Therefore, the system throughput can be obtained by *FMS-Q* for specific facility configurations. It can be used to generate and evaluate alternative *FMS* configurations at any stage within a planning horizon.

3.2.7. *Summary and Comments*

The major contributors and their assumptions, methodologies, as well as results of the optimal configuration models are shown in Table 2.

TABLE 2

Summary of Major Models of Optimum Configuration

Author	Assumption	Methodology	Main result
Vinod and Solberg (1985)	*C*=[*M*, *M*, *m*, *I*, *FCFS*] *S*=[*m*, *D*, *G*, *S*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *MPM* 2. *IBM* 3. Arbitrary choice for initial solution 4. Bisection search	Optimal number of server at each station.
Dallery and Frein (1986)	*C*=[*M*, *M*, *m*, *I*, *FCFS*] *S*=[*m*, *C*, *G*, *S*] *M*=[–, *S*, –] *S*=[*R*, –]	1. *MPM* 2. *IBM* 3. Calculate the low bound of initial solution; 4. Gradient analysis	Optimal number of server at each station.
Yao and Shanthikumar (1986)	*C*=[*M*, *M*, *m*, *F*, *FCFS*] *S*=[*m*, *C*, *G*, *S*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *MPM* 2. *IBM* 3. Marginal allocation	Optimal number of local buffer.
Shanthikumar and Yao (1987 *b*, 1988)	*C*=[*M*, *M*, *m*, *I*, *FCFS*] *S*=[*m*, *C*, *G*, *S*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *MPM* 2. *IBM* 3. Marginal allocation with upper bound formula of throughput	Optimal number of server at each station.
Co, Wu, and Reisman (1989)	*C*=[*M*, *M*, *m*, *I*, *FCFS*] *S*=[*m*, *C*, *G*, *M*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *AMVA* 2. Layout Theory	Optimal layout
Lee *et al.* (1989)	*C*=[*M*, *M*, *m*, *I*<*FCFS*] *S*=[*m*, *C*, *G*, *S*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *MPM* 2. *IBM* 3. Fathoming method	1. Optimal allocation of workload at each station. 2. Optimal number of servers and pallets.
Schweitzer and Seidmann (1989)	*C*=[*M*, *M*, *m*, *F*, *FCFS* or *S*=[*m*, *C*, *G*, *M*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *IBM* (*MVA*) 2. *MPM* 3. Convex resource allocation algorithm	Optimal process tool rate.
So (1989)	*C*=[*G*, *G*, *m*, *F*, *FSFS*] *S*=[*m*, *C*, *S*, *S*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *MPM* 2. *ADM* 3. Backlogged rate measurement	Optimal number of local buffer.

These models, (expected tools speed models) did not consider the tool configuraton problem, which is a critical weakness. Industry data indicates that tooling accounts for 25-30% of the fixed and variable cost of production in an automated machining environment (Ayres, 1988). In other words, tool management directly affects production costs.

3.3. Loading models

An *FMS* may consist of many manufacturing cells (stations) which process independently certain type of jobs and have their own production and buffer capacities. An important management issue to consider is thus the suitable allocation of the aggregate production rate (which comes from an upstream production stage and is assumed constant) to stations (machines). This differs from short term scheduling or releasing rules because it basically focuses on a design problem. Load objectives may be numerous, including (Stecke, 1986):

1) Balance the processing times on assigned machines.

2) Minimize the parts movement between machines.

3) Balance the work load per machine for a system of groups of pooled machines with equal size.

4) Unbalance the work load per machine for a system of groups of pooled machines with unequal sizes.

5) Fill the tool magazines as densely as possible.

6) Maximize the number of operation assignments. There are two types of load models–rule model and optimal model.

3.3.1. *Rule Model*

Yao (185, 1987) used "Majorization Order Theory" to compare the loading policies with regard to throughput, number of jobs in the system, and queue length when there are single servers at each station. Stecke and Morin (1985) and Shanthikumar and Stecke (1986) also discussed the foregoing problem by another approach and obtained similar results. Yao and Kim (1987) extended some of these results to cases where each station has the same number of multiple parallel servers. Finally, Stecke and Solberg (1985) discuss unbalancing problems, and mathematical models used to find optimal work loads and machine group sizes in an *FMS*.

3.3.2. *Optimum Models*

Yao and Shanthikumar (1986, 1987) considered the case in which each station in an *FMS* is regarded as an Erlang loss system of the type $M/G/n_i/n_i$

and proposed a mathematical programming model to find the optimal load (input) rate at each station and to obtain the maximum throughput. They used a convex objective function with respect to the load and proposed an iterative algorithm, based on the Frank-Wolfe algorithm (1956) to obtain an optimal solution.

3.3.3. *Summary and Comments*

The major contributions and their assumptions, methodologies, as well as results are shown in Table 3.

TABLE 3

Summary of Major Models of Loading

Author	Assumption	Methodology	Main result
Yao (1985, 1987, 1986 *b*)	*C*=[*G, G, m, I, FCFS*] *S*=[*m, o* or *C, G, S*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *ECM* 2. Majorization order theory	Rule of loading
Yao and Kim (1987)	*C*=[*G, G, m, I, FCFS*] *S*=[*m, o* or *C, G, S*] *M*=[–, *S*, –] *K*=[*R*, –]	Same as Yao (1985, 1987)	Same as Yao (1985, 1987)
Yao and Shanthikumar (1986, 1987)	*C*=[*M, G, m, F, FCFB*] *S*=[*m, o* or *C, G, S*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *MPM* 2. Queueing theory	Optimal load (input) rate.

In practice, loading problems are far more complex however. Specifically, tools assignment to machines may involve humans and not only automatic retooling systems. None of the models reviewed suggested how to deal with these problems when man-machine systems are taken into consideration.

3.4. Part routing models

Part routing schemes are ued to devise process plans for each part. In an *FMS*, each part can be produced using alternative routes to take advantage of system flexibility and machines versatility. Below, we consider routing rules which maintain the product form solutions when local buffers are limited. Subsequently, we discuss optimal part routing models.

3.4.1. *Routing Models with Limited Local Buffers*

Yao and Buzacott (1985 *a*, 1986 *b* and 1987) studied an *FMS* model with limited local buffers and proposed three types of models according to the system operational strategies:

1) Fixed routing models where the proportion of parts delivered to a station from the central buffer is (on average) constant.

2) Fixed loading models where routing is derived from a given operation mix required by a production task.

3) Dynamic routing models where the routing rate from a central buffer to work stations depends on the number of jobs waiting at the central buffer and in work stations. Although there can be numerous rules, the routing rule used consists in routing to the shortest queue with highest probability. Hence it is called the probabilistic shortest queue routing (*PSQ*).

The paper's conclusions are:

1) Dynamic routing scheme has an obvious advantage in increasing throughput but it requires more information and more complex control processses.

2) The *PSQ* routing is not significantly different from deterministic shortest queue routing (*DSQ*; *i. e.*, routing jobs to the shortest queue with probability one) when the local buffers are small compared with the total job population.

3) A good (robust) routing scheme for a real system is the *DSQ*, implying that the material handing system will always send parts to the shortest queue.

3.4.2. *Optimum Models*

Kimenia and Gershwin (1983, 1985) used a *CQN* model to study part routing problems in *FMS*. Their objective function is to minimize the congestion and delay within the system subject to a production rate requirement determined at a high level of the decision hierarchy. They solved it by a "Three Levels *LP* Model" in which a classical *CQN* is embedded to obtain the optimal routing policy.

Avonts and Wassenhove (1988) proposed a coupling model in which the *LP* was combined with the *CQN* to solve the part mix and routing mix problems simultaneously. Menga *et al.* (1984) decomposed the decision/control process of the *FMS* in two hierarchical levels. At the lower level, they developed an algorithm for finding the optimal routing coefficients by using *MVA* formulas.

3.4.3. *Summary and Comments*

The major contributions and assumptions, methodologies, as well as results for routing models reviewed are shown in Table 4.

TABLE 4

Summary of Major Models of Part Routing and Selection

Author	Assumption	Methodology	Main result
Buzacott and Shanthikumar (1980)	*C*=[*M, M, m, I, FCFS*] *S*=[*m, C, G, S*] *M*=[–, *S*, –] *M*=[–, *S*, –] *K*=[*R*, –]	*IBM*	Optimal production capacity.
Kimemia and Gershwin (1983, 1985)	*C*=[*M, M, m, I, FCFS*] *S*=[*m, C, G, S*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *IBM* 2. *MPM*	Optimal routing policy.
Menga *et al.* (1984)	*C*=[*M, M, m, I, FCFS*] *S*=[*m, C, G, S*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *IBM* 2. *MPM*	Optimal routing coefficients.
Yao and Buzacott (1985 *a*, 1986 *b*, 1987)	*C*=[*M, M, m, F, FCFS*] *S*=[*m, C* or *O, G, S*] *M*=[–, *S*, –] *K*=[*R* or *Fe* or *Fr*, –]	1. *IBM* 2. Resersiblity theory.	System performance and it's corresponding routing rule.
Avonts and Wassenhove (1988)	*C*=[*M, M, m, I, FCFS*] *S*=[*m, C, G, S*] *M*=[–, *S*, –] *K*=[*R*, –]	1. *ECM* 2. *MPM*	Optimal routing mix and part selection.

The routing scheme is highly related to a machine's flexibility. If machines in an *FMS* are totally versatile then the number of alternative routes is maximal. Hence it may be possible to find more efficient schedules. However, in practice, this is not the case since machines cannot be always substitutes to one another. Therefore, flexibility in part process plans as well as highly adaptive routing algorithms are very important factors for throughput maximization. As a result, we have to combine the scheduling and routing problems. Moreover, the routing problem should also be solved jointly with part type selection and part mix problems since these three combined problems constitute the batching problem.

3.5. Part selection models

Parts selection is a fundamental decision (Kalkunte *et al.*, 1986) reached to accomplish long range profitability and flexibility goals. Unlike parts mix

problems which seek to define the number of units of each part type which will be processed in the *FMS*, it consists in selecting the part types.

Two papers discuss this problem using a network of queues. A first general model describes the relationships between part type selection and work load balance while a second uses a mathematical programming model through which the optimal part selection and routing mix problems are solved simultaneously.

3.5.1. *General Model*

Buzacott and Shanthikumar (1980) studied the problem with a classical *OQN* and found that work loads can be balanced by selecting a variety of parts or using flexible machines (thereby, maximum production capacity can be obtained). However, the number of parts or (and) part types is (are) in reality restricted. Hence optimal production capacity is not be obtained when we balance the system work load. As a result, assuming some restrictions on parts or (and) on the number of part types, part selection is not an easy problem to deal with.

3.5.2. *Optimum Model*

A complete decision model for part selection was proposed by Avonts and Wassenhove (1988). Assuming that a new *FMS* is introduced and some products have to be shifted from a standard production system to an *FMS*, Avonts and Wassenhove formulate a model to select which part types should be shifted and in what quantities should it be produced on the *FMS* to obtain maximum savings. Two *LP* formulations are used to deal with the short and medium terms routing mix problems. The constraints considered are:

1) Limited capacity of the *FMS* machine tools.

2) Demand requirements are given.

3) Demand requirements can also be met by the standard system.

4) Work load is balanced within the allowed deviation percentage relative to the average work load.

They then developed a solution procedure which consists in the combination of an *LP* model and a queueing network model. The iterative algorithm of the solution procedure is briefly described as follows:

1) Initial step: Solve the *LP* with machines capacity utilization at 100%.

2) Iterative steps:

a) compute the part mix and routing mix for all parts;

b) use *CAN-Q* to determine the utilization of all machines;

c) solve an *LP* under the solution of step (*b*) until a satisfactory solution is found, *i. e.*, until part mix and routing mix stabilize.

3.5.3. *Summary and Comments*

Part selection problems have to be solved simultaneously with some other problems (*see* section 3.4.3.). No paper available in the literature has combined these related problems and therefore are of limited value. A summary of the assumptions, methodologies, as well as results of these two papers are also shown in Table 4.

3.6. Releasing or scheduling model

Releasing or scheduling models deal with dynamic and operation planning problems. The problems dealt with are:

1) At what time should a job (part) be released.

2) What job (part) types should be released.

The models are categorized as follows:

1) General rule models, providing general scheduling rules adapted to specific situations (e. g., balance rule and idle machine rule).

2) Single-level optimal models, which determine optimal schedules by a single-level decision making process.

3) Two-level optimal models which determine optimal schedules by a two-level hierarchical decision making process.

3.6.1. *General Rule Model*

The balanced queue and idle machine releasing rule are first discussed by Buzacott (1976). Under the balanced queue rule, a part is released from the machine with the shortest queue. Buzacott (1976) showed that the job shop capacity depends on jobs selection and release to a machine's queue. It is shown that the optimum release rule maintains a balanced queue when the number of jobs in the shop is kept to a constant level (but the optimum rule will be an idle machine rule when the number of jobs in the shop can vary).

Buzacott and Shanthikumar (1980) also compared the *FCFS* with the idle machine rule in cases in which a maximum number of jobs are allowed (in a two or three machines balanced system) and showed that the production capacity is improved by using the idle machine rule).

However, the models above only deal with small size systems (of two or three machines). The general case remains open. Thus, we cannot conclude

which rule is generally better. Stecke and Morin (1985) used a single server *CQN* model to analyze the optimality of balance for an adequately buffered *FMS* in which each operation is assigned to only one machine, and showed that the balancing releasing rule maximizes the expected production.

Buzacott and Gupta (1986) studied different scheduling policies for an *FMS* with two job types, single server *OQN*, and non-negligible set up times. The assumptions are as follows:

1) Each job visits a machine no more than once.

2) The processing times are generally distributed.

3) Set up times follow a general distribution.

4) The machine has an infinite local buffer.

5) Jobs are served according to one of the following three queue disciplines:

(i) *FCFS*-Jobs are served in order of their arrivals.

(ii) Strict priority (*SP*)-All jobs are ordered according to a fixed scheme of assigning priorities such that jobs belonging to the higher priority class are always served first under a preemptive priority.

(iii) Alternating priority (*AP*)-Jobs are served according to their job class in a cyclic order. This means that if a machine is currently processing jobs of class i, it will start to serve the next class (say j) only when there are no jobs of class i waiting to be served.

With the above assumptions they derived appoximate formulae for the flow time distribution, compared the scheduling disciplines, and showed that:

1) the Ap rule yields smaller mean flow times than the *FCFS* and *SP* rules do:

2) for under-utilized systems, scheduling rules do not affect flow time substantially; and

3) set up times have a significant effect on system performance.

3.6.2. *Single Level Optimum Models*

Hildebrant (1980) developed a scheduling approach in an *FMS* in which machines' failure is explicity considered. The model is formulated as a nonlinear programming problem (*NLP*) in which a queueing network is used for the measurement of performances. Since machines' failure affect a system's configuration, the time horizon for the problem is partitioned into a series of disjoints segments. We can then schedule parts by individual time intervals and predict future machine failure using an appropriate probability model. Thus, certain parts may be deferred until some later time when

conditions are perceived as more favorable. This formulation however, predetermines all parts mix and parts routing in terms of failure types. The problem's objective is to minimize the completion time under the following restrictions:

1) a production target (demand);

2) a maximum number of fixtures can reside within the system during any failure condition; and

3) a maximum number of fictures available for parts.

In this model the completion time is a function of machine failure types, production target (requirement of demend) and part routing. It is solved by a variation on the successive linearization method, which was developed by Suri (1978), together with a *MVA*, evaluating throughput at each iteration.

For an *FMS* with limited buffers at work stations and a centralized material handling station, Seidmann and Tennenbaum (1986) proposed two objective functions for releasing policies. One minimizes the weighted starvation penalty and the other maximizes the weighted throughput. The releasing control problem consists then in determining which work station should be fed whenever it becomes available. Unfortunately, product form solutions do not hold here since the model has a state dependent arrival process. However, a tractable analytical formulation for various distribution functions of *MHS* can be solved by the Schweitzer's Value Iteration Scheme (Schweitzer, 1971). Nevertheless, this model only focuses on "one" *MHS*, *i. e.*, an FMS with a robot or a cart in its *MHS* which is not always the case in reality.

3.6.3. *Two-Level Optimum Models*

Shanthikumar (1984) used a single server *OQN* to model a single machine dynamic job shop. There are two decision levels:

1) The first decides the time at which the batch of jobs in the dispatch area are to be released to the shop.

2) The second decides the scheduling discipline to be used within each batch.

Shanthikumar (1984) investigated four different release policies for first level problems:

1) *FCFS*: release without any delay at the dispatch area (this can be then modeled as an *M/G/1/ FCFS* queue).

2) *SPT-WG*: release jobs if and only if the machine is idle, *i. e.* , scheduling within generations under the shortest process time rule (Nair and Neuts, 1969 and 1971).

3) *SPT-WPR*: release jobs to shop at prescribed review time, *i. e.*, scheduling within period review under the shortest process time rule.

4) *SPT-FBS*: release jobs to the shop as soon as a fixed level of jobs are accumulated at the dispatch area, *i. e.*, scheduling within fixed batch size.

Moreover, all jobs within a batch are scheduled according to a *SPT* service rule. Then, if mean waiting times are used, then *SPT-WG* is better than *SPT-WFBS*, *SPT-WPR* and *FCFS*. Furthermore, *SPT-WPR* and *SPT-WFBS* have a higher mean waiting time than *FCFS*. However, since the *SPT-WG* policy requires a continuous update on the shop state, we cannot conclude it is the best. The costs associated to these rules should be compared as well. Shanthikumar (1984), thus derived a combined cost function expressed in terms of the dispatch, monitoring and job waiting cost of each rule.

Menga *et al.* (1984) also proposed a two-level hierarchical approach to solve the scheduling problem in an *FMS*. At the higher level, a lot r is characterized by a ready time, a due date and a given size. Assuming a given planning horizon, lots are scheduled to maximize a production performance index and to satisfy release and due date constraints. Time-phased order for materials and releasing (scheduling) for lots are determined at each period. The release rule at the lowest level consists in selecting optimal routing coefficients which are derived by "balancing the utilization idle machine rule". By using the *MVA*, an iterative algorithm was developed to compute the routing coefficients.

3.6.4. *Summary and Comments*

Major contributions and their assumptions, methodologies, as well as results for releasing and scheduling models are shown in Table 5.

In general rule models, only the single server case is discussed (these rules need not be true for multiple servers problems where their number in each station need not be equal. The second level problem in Shanthikumar's (1984) is not expicitly formulated. He basically adopted SPT service discipline at the second level since his analysis centers around the mean waiting time of jobs under different releasing policies. In Menga *et al.* (1984), the authors introduced the lot size which is assumed constant and determined by customers order size.

3.7. Unreliable system Models

There are two types of unreliabilities in a manufacturing system–equipment failure and product defect. Three kinds of models–general breakdown

TABLE 5
Summary of Major Models of Scheduling, Releasing

Author	Assumption	Metho-dology	Main result
Buzacott (1976)	*C*=[*M, M,* 1, *I,* –] *S*=[3, *o, G, S*] *K*=[*R, IM* or *FCFS*]	*IBM*	Better policy is balanced queue rule when the number of part is fixed, otherwise it is idle machine rule.
Buzacott and Shanthikumar (1980)	*C*=[*M, M,* 1, *I,* –] *S*=[3, *o, G, S*] *K*=[*R, IM* or *FCFS*]	Same as Buzacott (1976)	Idle machine rule is better than *FCFS* under a constant part number in the system.
Hildebrant (1980)	*C*=[*M, M, m, I, FCFS*] *S*=[*m, C, G, M*] *U*=[*M, M,* 1, 1, –, –, –, –] *M*=[*Ca, M,* –] *K*=[*R,* –]	1. *IBM* (*MVA*) 2. *MPM*	Optimal parts scheduling.
Menga *et al.* (1984)	*C*=[*M, M, m, I, FCFS*] *S*=[*m, C, G, M*] *M*=[–, *M,* –] *K*=[*R, IM*]	1. *IBM* (*MVA*) 2. *MPM*	1. Scheduling for each lot. 2. Optimal routing. Coefficients.
Shanthikumar (1984)	*C*=[*M, G, m, I, SPT*] *S*=[*m, O, G, S*] *M*=[–, *S,* –] *K*=[*R, FCFS* or *SPT-WFBS, SPT-WPR*]	1. *IBM* 2. Traditional optimization model	1. Releasing time of the batch of jobs. 2. Scheduling rule within each job.
Stecke and Morin (1985)	*C*=[*M, M, m, I*<*FCFS*] *S*=[*m, C, G, S*] *M*=[–, *S,* –] *K*=[*R, B* or *UB*]	*IBM*	Balanced queue rule maximizes the expected production.
Buzacott and Gupta (1986)	*C*=[*G, G, m, I, FCFS* or *SP-P* or *AP-P*] *S*=[1, *o, G, M*] *M*=[–, *S,* –] *K*=[*R,* –]	*ADM* for flow time distribution	Comparsions of *FCFS*, *SP*, and AP scheduling rule.
Seidmann and Tennenbaum (1986)	With a state dependent arrival *C*=[*M, M*<*m, F, FCFS*] *S*=[*m, C, G, S*] *M*=[*R* and *Ca, S, M*] *K*=[*PSQ,* –]	1. *MPM* 2. *IBM* 3. Semi-Markovian decision process.	Releasing or loading rule for *MHS*.

models, optimal breakdown models, and quality control models, have been studied.

3.7.1. *General breakdown Models*

Buzacott and Shanthikumar (1980) demonstrated the effect of breakdown on production capacity, but they didn't present any optimal rule for managing the FMS in such circumstances.

A general model for *FMS* with machines breakdown was first proposed by Vinod and Solberg (1984). In their model, each work station was considered as a delay station which was visited by the preemptive failing customers. Assuming ample supply of repairmen, they analyzed it by a *CQN* and proposed two approximate methods–*MCA* (multiple class approximate method) which is a modification of *MVA*, and *APE* (approximate product form) which is based on the concept of product form solution, to calculate the throughput, sojourn time and utilization. Using simulations, they found:

1) *MCA* always underestimates the actual throughput.

2) *MCA* is consistent and accurate. The relative throughput error decreases with increasing N while the *APF* does not have this property.

3) *MCA* and *APF* estimate the mean queue length fairly accurately. The error, though small, frequently increases with increasing N, but the rate of increase decreases.

Vinod and Altiok (1986) modeled the foregoing problem by an exact equivalent network in which the service time was represented with a two-stage Coxian distribution. They approximate it by *CAN-Q* and validate it for a wide range of model parameters. The results indicate that approximate formulae are robust.

3.7.2. *Optimum breakdown models*

Vinod and John (1986) studied a failure prone *FMS*, with a repair facility. The *FMS* in their model has M stations that perform distinct operations and each station has one or more identical machine(s). Each station represents a distinct input resource for the repair system. Hence, there are $M+2$ stations in the system and its stationary probability distribution can be obtained from the *QN* formulae with multiple job types. A mathematical model was built for determining the optimal capacities of repair facilities. They proved the monotonicity property in steady state. Namely, if the number of repair channels at either one or both repair facilities is increased by one, the (steady state) number of operating machines at each station increases stochastically. Using this property, it can then be solved by traditional discrete optimization algorithms (Lawler and Bell, 1966).

Vinod and Sabbagh (1986) proposed a tool availability model which is used to determine the optimal level of spares for each tool type and optimal capacity level of repair facility. The model is a nonlinear integer program which is solved by a modified version of lexicographic partial enumeration algorithm (Sabbagh, 1983).

3.7.2. *Quality control models*

The great majority of *FMSs* discussed have presumed a zero-defects technology. In practice, this is rarely the case. Tapiero and Hsu (1988) considered an unreliable *FMS*, and proposed an algorithm to compute the average outgoing quality (*AOQ*) of the system based on an approximation of Whitt (Whitt, 1983a and 1983b) for a *GI/G/S* network of queues. They derived a relationship between the inspection effort and the output quality under Bernoulli (Duncan, 1974) inspection plan. Unfortunately, under a *CSP*-1 (Duncan, 1974) inspection plan, they could not obtain the above results. They also do not consider optimal inspection plans for *FMSs* however.

Hsu and Tapiero (1989) modify the forgoing model (Tapiero and Hsu, 1988) and explicitly discuss the following three issues in the *FMS*:

1) Introduce a general procedure for measuring and managing the in-process quality control of an *FMS* described by an *OQN*.

2) Provide some managerial insights regarding the role, position and distribution of the quality control effort in the *FMS*; and

3) Formulate the intricate relationships between the *FMS's* operating characteristics and its control.

They used a nonlinear programming problem to find an optimal quality control plan. The objective function includes the following elements:

1) Inspection cost of each part type at each process step.

2) Waiting cost per unit time.

3) Post-sales failure cost for each part type.

4) A benefit (revenue) for each part type; and

5) A scrapping cost for each part type.

The constraints are the inspection rate for each part type. By the *ADM*, they calculated the throughput and *AOQ*. Subsequently, the value of an objective function is obtained uner a fix inspection rate for each part type. They did not propose a complete algorithm to solve this nonlinear programming problem however but provided some sensitivity analysis regarding profits, inspection rate and the *AOQ*.

3.7.3. *Summary and Comments*

The major contributions and their assumptions, methodologies, as well as results for unreliable systems are shown in Table 6.

Two assumptions in Hsu and Tapiero (1989) seem to be unreasonable. These are:

TABLE 6

Summary of major models of unreliable system and inventory

Author	Assumption	Metho-dology	Main result
Vinod and John (1980)	C=[M, M, m, I, *FCFS*] S=[m, C, G, m] U=[M<M, 2, 2, 2, 2, 2, 2] M=[–, S, –] K=[R, –] two stage repair facility.	1. *IBM* 2. *MPM*	Optimal repair capacity.
Vinod and Solberg (1984)	C=[M, M, m, I, *FCFS*] S=[m, C, G, m] U=[M, M, 1, –, –, –, –, –] M=[–, S, –] K=[R, –] ample repair capacity.	Two approximated method-*MCA* and *APF* which are modified form *MVA* and solution respectively.	1. Throughtput. 2. Sojourn time. 3. Utilization product of unreliable *FMSs*.
Vinod and Altiok (1986)	C=[M, M, m, I<*FCFS*] S=[m, C, G, S] U=[M, M, 2, –, –, –, –, –] M=[–, S, S] K=[R, –] Ample repair capacity. Service completion time=two-stage Coxian distribution.	Approximate it by *CAN-Q*	Throughput
Buzacott and Shanthikumar (1980)	C=[M, M, 1, I, –] S=[3, o, G, S] K=[R, *IM* or *FCFS*]	Same as Buzacott (1976)	Idle machine rule is better than *FCFS* under a constant part number in the system.
Yao (1986)	C=[M^{BD}, M, m, F, *FCFS*] S=[m, C, G, S] M=[–, S, –] K=[R, –] *: Optimal delivery point.	1. *IBM* 2. Classical optimal theory	1. Optimal delivery point. 2. Optimal batch size.
Vinod and Sabbagh	C=[M, M, m, I<*FCFS*] S=[m, C, G, M] U=[M<M<1, –, –, –, –, –] M=[–, S, –] K=[R, –] Ample repair capacity.	1. *IBM* with multiple class 2. *MPM*	Optimal allocation of spare tools.
Tapiero and Hsu (1988)	C=[G, G, m, I, *FCFS*] S=[m, O, G, M] U=[–, –, –, –, R or *CSP*, D, –, O] K=[R, –]	*ACM*	*AOQ*
Hsu and Tapiero (1989)	C=[G, G, m, I, *FCFS*] S=[m, O, G, m] U=[–, –, –, –, R, D, , –, O] K=[R, –]	1. *ADM* 2. *MPM*	Relationships between profit, inspection plan and *AOQ*.

1) Infinite local buffer (almost all real *FMSs* have finite local capacity (Buzacott and Yao, 1986 *a* and 1986 *b*); and

2) All defective parts are scrapped (this is impossible when the part value is high and defective parts can be repaired in a reasonable cost).

3.8. Inventory models

Yao (1986) considered the following inventory problems in a generic *FMS*,

1) When to deliver a batch from the central warehouse to the buffer storage of the *FMS*, and

2) How large the batch size should be.

Clearly, this is traditional inventory problem which can be solved by an (s, Q) inventory model (Silver 1981). It can then be formulated using a total expected cost function which includes the order, holding and production losses. These correspond to the lost sales case in (s, Q) models. By using queueing network performance measures and classical optimization, Yao (1986) proved the convexity of the cost function and solved it using a "one-dimensional search", *i. e.*, for a collection of s values, compute Q from first order conditions and its corresponding total expected cost, and then find the optimal solution based on the local optimal solutions. The routing and part mix plicies were not considered in this model however.

4. CONCLUSION

This survey has reviewed a large number of papers which have dealt with a network of queues approach to the design and the management of *FMSs*. We categorized this survey into eight management problems and compared various research papers relating to each of the problems. We note throughout our survey that the essential measures of performance used were throughout or flow time. Using these measures, many authors have derived formulae for various optimal decision rules. In some cases, exact results are obtained while in the greater part, they are only approximations.

Optimum configuration models are most useful for determining the initial investment in an FMS and for specific production goals. These models require the solution of complicated nonlinear programming problems however, although their formulation seems to be straithforward. Loading, routing, and part selection models are basically intermediate *FMS* policies. Nonetheless, they ought to be combined to determine an optimal configuration. We find no papers that dealt simultaneously with these management problems. Releasing and scheduling models belong to a class of real time control problems. However, all optimal models studied are not easy to execute in a real time.

Breakdown and quality control models usually assume breakdown rates which are too simple and therefore not realistic. Further, associated problems of maintenance are rarely dealt with conjointly. In quality control models, inspection rates are used to find a "good" quality control policy. We note here that traditional quality control methods are not suitable for *FMSs*, and that flexibility in routing compounds the difficulties to measure a process reliability (but it augments the potential for reliability management).

To use profitably queueing networks in *FMS* systems, Dallery (1986) suggests that we respond to the following questions:

1) Does the *FMS* have universal or dedicated pallets for each part?

2) Is the *FMS* controlled according to fixed queueing disciplines, or in such a way that prescribed production ratio for each part type is achieved? *i. e.*, features such as identical work stations, storage location, and a material handling system can be easily incorporated in a given queueing network model.

Although the queueing network models reviewed here can be used to solve many of the design and operation problems of *FMSs*, there are some limitations.

1) They only model the equilibrium-system-behavior which is appropriate for long-term planning problems or for screening systems design. Operational issues require that a transient-system-behavior which is far more complicated, be used.

2) They neglect the tool management system. Tool management system is a critical issue in the *FMS* (Gray, Seidmann and Stecke, 1993) since it largely affects the productivity of a facility. Kiran and Drason (1988) argued that tooling has become one of the hindrances to efficient *FMS* performance. Using industrial data, it has been pointed out that tooling accounts for 25-30% of the fixed and variable costs of production in an automated machining environment (Ayres, 1988).

3) Queueing networks can be applied primarily at the design stage. They are very useful at the preliminary design stage where we seek information regarding the structural performance of the system under a broad range of parameters variation. However, at a detailed stage, they lose some of their usefulness. To compensate these deficiencies, simulation models must be built to capture the detailed system operations, processing requirements, and resource allocation. Nevertheless, combining queueing networks with simulation can provide an efficient means to validate complex models by

testing special cases whose performance are, in the long run, predictable (Shanthikumar and Sargent, 1983; Buzacott and Yao, 1986).

4) They rarely respond to any unexpected situation in real time when the environment changes.

5) No model here considers the costs of information acquisition in the decision process and its execution.

6) They seldom consider flexibility as a measure of system performance. Further, flexibility pertains not only to a potential for productivity enhancement but also to an ability to respond to and adapt to changes in the market or to other unpredictable factors. As a result, it is necessary to stress flexibility as a design objective, rather than just estimate it once the *FMS* structure has been designed in terms of standard performance measures.

7) The performance measures used throughout this survey do not always measure the productivity of an *FMS*. We found that the throughput and flow time are widely used in performance evaluation. Sometimes the facility utilization has also been used. If we use the throughput or the flow time as a measure together with a utilization and/or budget constraint, it may lead to a sub-optimal solution because it does not capture the marginal effects of constraints (e. g., the throughput may increase substantially due to a minor change in the constraint (Nankeolyar and Christy 1989). They do not capture the essential benefits of flexibility. Some models (Yao, 1986; Dallery and Frein, 1986) use some other cost factors, such as *WIP*, interest, and discount rate. These performance measures, seem to be more reasonable in practice on the one hand but do not value the advantages of flexibility on the other. Nandkeolyar and Christy (1989) suggested a measure based on weighted productivity (weighted output/weighted input). This approach seems to be better then traditional measures if a proper set of weights is selected. There are some shortcomings however:

a) It is difficult to identify all the relevant factors to be included in the measure.

b) The factors weights are difficult to assign.

c) It does not include the value of flexibility.

d) It is difficult to relate to specific production activities of *FMSs*.

8) Heuristic algorithms for solving the mathematical programming problems are not efficient for large *FMS*. Although *FMS* design models can be formulated, their solution is very complicated and time consuming (especially for large *FMSs*). For this reason, only simple numerical examples were

considered in the papers we reviewed. Thus, the development of efficient heuristic algorithms is an important issue to be dealt with in the future.

9) No paper discusses bulk arrival and bulk service. While in manufacturing, lot (bulk) sizes are conventionally used in processing jobs, their study in a network of queues is lagging.

10) The integration of scheduling algorithms in controlling jobs flow in *OQN* and *CQN* manufacturing systems have not been addressed in sufficient depth. No work has studied practical scheduling rules such as earliest due date first (EDD), largest processing time first (*LPT*), etc. which are often used in scheduling job shops.

Finally, network of queues are by no means the only technique to study *FMSs*. Although this is perhaps one of the first techniques applied successfully to design such systems. These last few years, Petri Nets, Max-Plus algebra, fuzzy sets, expert and decision support systems of various sorts and philosophies have been conceived and extensive studies are still being performed to design, assess and manage in real time the extremely complex problems which arise in the management of flexible manufacturing systems.

REFERENCES

I. F. Akyldiz, Exact Product Form Solution for Queueing Networks with Blocking, *I.E.E.E. Trans. on Computers*, 1984, *C-36*, No. 1.

I. F. Akyldiz and G. Bolch, Mean Value Analysis Approximation for Multiple Server Queueing Networks, *Performance Evaluation*, 1988, *8*, pp. 77-91.

M. Alam, D. Gupta, S. I. Ahmad and A. Raouf, Performance Modeling and Evaluation of Flexible Manufacturing Systems Using a Semi-Markov Approach, *Manufacturing Research and Technology - Flexible Manufacturing*, 1985, A. Raouf and S. E. Ahmad Eds., Elsevier, Amsterdam.

N. Alberti, U. La Commare and S. Nofo Ladiga, Cost Efficiency: An Index of Operational Performance of Flexible Automated Production Environments, *Proceedings of the Third O.R.S.A./T.I.M.S. Conference on Flexible Manufacturing Systems*, 1989, pp. 67-72.

T. Altiok, Approximate Analysis of Exponential Tandem Queues With Blocking, *Euro. J. of Operational Research*, 1982, *11*, pp. 390-398.

American Machinist, CAM: An International Comparison, Special report 740, *American Machinist*, August 1981, pp. 207-226.

L. H. Avonts and L. N. Van Wassenhove, The Part Mix Problem in *FMS*: a Coupling Between an *LP* Model and a Closed Queueing Network, *Int. J. Prod. Res.*, 1988, *26*, No. 12, pp. 1891-1902.

R. V. Ayres, Future Tends in Factory Automation, *Manufacturing Review*, 1988, *1*, No. 2, pp. 93-103.

F. Baskett, K. M. Chandy, R. R. Muntz and F. G. Palacios, Open, Closed, and Mixed Networks of Queues with Different Classes of Customers, *J. A.C.M.*, 1975, *22*, No. 2, pp. 248-260.

G. R. BITRAN and A. C. HAX, Disaggregation and Resource Allocation Using Convex Knapsack Problems with Bounded Variable, *Management Science*, 1981, *7*, No. 4, pp. 431-441.

G. R. BITRAN and D. TIRUPATI, Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference, *Management Science*, 1988, *34*, No. 1, pp. 75-100.

P. H. BRILL and L. GREEN, Queues in which Customers Receive Simultaneous Service from a Random Number of Servers: A Systems Point Approach, *Management Science*, 1984, *30*, pp. 51-68.

J. A. BUZACOTT, The Production Capacity of Job Shops with Limited Storage, *Int. J. Prod. Res.*, 1976, *14*, No. 5, pp. 597-605.

J. A. BUZACOTT, Optimal Operating Rules for Automated Manufacturing Systems, *I.E.E.E. Transactions on Automatic Control*, 1982, *AC-27*, No. 1, pp. 80-86.

J. A. BUZACOTT, Modelling Automated Manufacturing Systems, *Proceedings* of Fall Industrial Engineering Conference, 1983, pp. 130-137.

J. A. BUZACOTT and D. GUPTA, Impact of Flexible Machines Automated Manufacturing Systems, *International J. of Flexible Manufacturing Systems*, 1992.

J. A. BUZACOTT and J. G. SHANTHIKUMAR, Models for Understanding Flexible Manufacturing Systems, *A.I.I.E.*, 1980, *12*, No. 4, pp. 339-350.

J. A. BUZACOTT and J. G. SHANTHIKUMAR, On approximate Queueing Models of Dynamic Job Shops, *Management Science*, 1985, *31*, No. 7, pp. 870-887.

J. A. BUZACOTT and D. D. YAO, Flexible Manufacturing Systems: a Review of Analytical Models, *Management Science*, 1986 *a*, *32*, No. 7, pp. 890-905.

J. A. BUZACOTT and J. G. SHANTHIKUMAR, On Queueing Network Models of Flexible Manufacturing Systems, *Queueing Systems*, 1986 *b*, *1*, pp. 5-27.

P. J. BUZEN, Computational Algorithms for Closed Queueing Networks with Exponential Servers, *Comm. A.C.M.*, 1973, *16*, No. 9, pp. 527-531.

P. J. BUZEN, Fundamental Laws of Computer System Performance, *Proceeding* of Int. Symp. on Computer Modeling, Measurement and Evaluation, 1976, pp. 200-210.

C. CASSANDRAS, A Hierarchical Control Scheme for Material Handling Systems, *Proceedings of the First O.R.S.A./T.I.M.S. Special Interest Conference on FMSs*: Operations Research Models and Applications, Ann. Arbor, MI, 1984, pp. 397-402.

J. B. CAVAILLE and C. DUBOIS, Heuristic Methods Based on Mean-Value Analysis for Flexible Manufacturing Systems Performance Evaluation, *Proceedings* of the 21st I.E.E.E. Conference on Decision and Control, 1982, pp. 1061-1065.

K. M. CHANDY and D. NEWSE, Linearizer: A Heuristic Algorithm for Queueing Network Models of Computing Systems, *Comm. A.C.M.*, 1982, *25*, No. 2, pp. 126-134.

M. L. CHAUDHRY and J. G. C. TEMPLETON, *A first Course in Bulk Queues*, Wiley, New York, 1983.

H. CHEN, J. M. HARRISON, A. MANDELBAUM and A. VAN ACKERE, Empirical Evaluation of a Queueing Network Model for Semiconductor Wafer Fabrication, *Operations Research*, 1988, *36*, No. 2, pp. 202-215.

S. CHIAMSIRI and M. S. LEONARD, A Diffusion Approximation for Bulk Queues, *Management Science*, 1981, *27*, pp. 1188-1199.

H. C. CO and R. A. WYSK, The Robustness of *CAN-Q* in Modelling Automated Manufacturing Systems, *Int. J. Prod. Res.*, 1986, *24*, No. 6, pp. 1485-1503.

H. C. CO, A. WU and A. REISMAN, A Throughput-Maximizing Facility Planning and Layout Model, *Int. J. Prod. Res.*, 1989, *27*, No. 1, pp. 1-12.

R. W. CONWAY, W. L. MAXWELL and L. W. MILLER, *Theory of Scheduling*, Addison-Wesley, New York, 1967.

R. CONTERNO, G. MENGA and S. QUAGLINO, Performance Evaluation of *FMS* by Heuristic Queueing Network Analysis, *I.E.E.E. International Conference on Robotics and Automation*, San Francisco, CA., 1986, pp. 959-964.

R. B. COOPER, *Introduction to Queueing Theory*, MacMillan, New York, 1972.

C. COURCOUBETIS and P. P. VARAIYA, Serving Process with Least Thinking Time Maximizes Resource Utilization, *I.E.E.E. Automatic Control*, 1984, *29*, No. 11.

D. R. COX and H. D. MILLER, *The Theory of Stochastic Processes*, 1965, John Wiley and Sons, Inc., New York, Chapters 2 and 5.

T. B. CRABILL, D. GROSS and M. J. MAGAZINE, A Classified Bibliography on Research on Optimal Design and Control of Queues, *Operations Research*, 1977, *25*, pp. 219-232.

Y. DALLERY, On Modeling Flexible Manufacturing Systems Using Closed Queueing, *Large Scale System*, 1986, *11*, No. 2, pp. 109-119.

Y. DALLERY, A Queueing Network Model of Flexible Manufacturing Systems Consisting of cells, *I.E.E.E. International Conference on Robotics and Automation*, 1986, San Francisco, CA., pp. 951-958.

Y. DALLERY, and R. DAVID, A New Approach Based on Operational Analysis for Flexible Manufacturing Systems Performance Evaluation, *Proceedings of 22th I.E.E.E. Conference on Decision and Control*, 1983, pp. 1056-1061.

Y. DALLERY and R. DAVID, Operational Analysis of Multiclass Queueing Networks, *Proceedings* of 25th *I.E.E.E. Conference on Decision and Control*, 1986, pp. 1728-1732.

Y. DALLERY and Y. FREIN, An Efficient Method to Determine the Optimal Configuration of a Flexible Manufacturing System, *Proceedings of the Second O.R.S.A./T.I.M.S. Conference on FMSs*, 1986, pp. 269-282, Elsevier, Amsterdam.

Y. DALLERY, T. J. JAW and S. K. CHEN, Sequencing in Flexible Manufacturing Systems and Other Short Queueing Length Systems, *J. Manufacturing System*, 1988, *7*, No. 1, pp. 1-8.

Y. DALLERY and D. D. YAO, Modeling a System of Flexible Manufacturing Cells, *Modeling and Design of FMSs*, edited by A. Kusiak, 1986, pp. 289-299, Elsevier, Amsterdam.

L. E. N. DELBROUCK, A Feedback Queueing System with Batch Arrivals, Bulk Service and Queue Dependent Service Time, *J. Assoc. Compt. Mach.*, 1970, *17*, pp. 314-323.

P. J. DENNING and J. P. BUZEN, The Operational Analysis of Queueing Networks Models, *Computing Surveys*, 1978, *10*, No. 3, pp. 225-261.

R. L. DISNEY and D. KONIG, Queueing Networks: A Survey of Their Random Processes, *S.I.A.M. Rev.*, 1985, *27*, No. 3, pp. 335-403.

B. T. DOSHI, Continuous Time Control of the Arrival Process in an *M/G/1* Queue, *Stoch. Proc. and Appl.*, 1977, *5*, pp. 265-284.

B. T. DOSHI, Vacation Queues, A Survey, *Queueing Systems*, 1986, *1*.

A. J. DUNCAN, *Quality Control and Industrial Statistics*, 4th ed., 1974, Irwin, Illinois.

C. DUPONT-GATELMAND, A Survey of Flexible Manufacturing Systems, *J. Manufacturing Systems*, 1982, *1*, No. 1, pp. 1-16.

M. P. FANTI, B. MAIONE, Q. SEMERARO and B. TURCHIANO, *International Journal of Systems Science*, 1988, *19*, No. 11, pp. 2381-2394.

W. FELLER, *An Introduction to Probability Theory and Its Application*, 1971, *II*, 2nd ed. Wiley, New York.

F. G. FOSTER and H. G. PERROS, On the Blocking Process in Queue Networks, *European J. of Operational Research*, 1980, *5*, pp. 276-283.

B. FOX, Discrete Optimization via Marginal Analysis, *Management Science*, 1966, *13*, No. 2, pp. 210-216.

D. P. GAVER, Diffusion Approximations and Models for Certain Congestion Problems, *J. Appl. Prob.*, 1968, *5*, pp. 607-623.

D. P. GAVER and G. S. SHEDLER, Approximate Models for Processor Utilization in Multi-Programmed Computer Systems, *S.I.A.M. J. Comput.*, 1973 *a*, *2*, pp. 183-192.

D. P. GAVER and G. S. SHEDLER, Processor Utilization in Multiprogramming Systems via Diffusion Approximations, *Operations Research*, 1973 *b*, *21*, pp. 569-576.

E. GELENBE, On Approximate Computer System Models, *J. Assoc. Comput. Mach.*, 1975, *22*, pp. 261-269.

E. GELENBE, Probabilistic Models of Computer Systems, Part. II: Diffusion Approximations, Waiting Times, and Batch Arrivals, *Acta Informatica*, 1979, *12*, pp. 285-303.

R. GELENBE and G. PUJOLLE, The Behavior of a Single Queue in a General Queueing Networks, *Acta Informatica*, 1976, *7*, pp. 123-136.

E. GELENBE and G. PUJOLLE, A Diffusion Model for Multiple Class Queueing Networks, *Measuring Modelling and Evaluating Computer Systems*, 1977, H. BEILNER and E. GELEBE Eds., North-Holland.

J. C. GITTINS and P. NASH, Scheduling, Queues, and Dynamic Allocation Indices, *Proc. E.M.S.*, Prague, 1974, pp. 191-202, Prague, Czech. Academy of Sciences.

W. J. GORDON and G. J. NEWELL, Cosed Queueing Systems with Exponential Servers, *Operations Research*, 1967, *15*, No. 3, pp. 254-265.

L. GREEN, A Queueing System in which Customers Require a Random Number of Servers, *Operations Research*, 1980, *28*, pp. 1335-1346.

D. GROSS and C. M. HARRIS, *Fumdamentals of Queueing Theory*, John Wiley and Sons, New York, 1974.

S. C. GRAVES and J. KEILSON, A Methodology for Studying the Dynamics of Extended Logistic Systems, *Naval Res. Logist Quart.*, 1979, *26*, pp. 169-197.

S. C. GRAVES and J. KEILSON, Systems Balance for Extended Logistic Systems, *Operations Research*, 1983, *31*, No. 2, pp. 234-249.

A. E. GRAY, A. SEIDMANN and K. E. STECKE, A Synthesis of Decision Models for Tool Management in Automated Manufacturing, *Management Science, Forthcoming*, 1993.

N. R. GREENWOOD, *Implementing Flexible Manufacturing Systems*, Wiley, New York, 1988.

T. G. GUNN, The mechanization of design and manufacturing, *Scientific American*, Sept. 1982, pp. 115-130.

B. HALACHMI and W. R. FRANTA, A Diffusion Approximation Solution to the *G/G/K* Queueing System, *Computers and Operations Research*, 1977, *4*, pp. 37-46.

J. HARRISON, Dynamic Scheduling of a Multiclass Queue: Discount Optimality, *Operations Research*, 1975, *23*, pp. 260-269.

J. HATVANY, World Survey on C.A.M., Butterworths, Kent, U.K., 1983.

R. R. HILDENBRANT, Scheduling Flexible Machining Systems Using Mean Value Analysis, *Proceedings* of 19th *I.E.E.E. Conference on Decision and Control*, 1980, pp. 701-706.

Y. C. Ho, A Survey of the Perturbation Analysis of Discrete Event Dynamic Systems, *Annals of Operations Research*, 1985, *3*, pp. 393-402.

Y. C. Ho and X. CAO, Perturbation Analysis and Optimization of Queueing Networks, *Journal of Optimization Theory and Applications*, 1983, *40*, No. 4, pp. 559-582.

G. K. HUTCHISON, Flexible Manufacturing Systems in the United States, Management Research Center, University of Wisconsin, Milwaukee, 1979.

G. K. HUTCHINSON and B. E. WYNNE, A Flexible Manufacturing System, *Industrial Engineering*, Dec. 1973, pp. 10-17.

L. F. Hsu and C. S. Tapiero, Quality Control of an Unreliable Flexible Manufacturing System: with Scrapping and Infinite Buffer Capacity, *Int. J. FMSs*, 1989, *1*, pp. 325-346.

L. F. Hsu and C. S. Tapiero, Inspection of an Unreliable Flexible Manufacturing System: with Repairable Parts and Non-Negligible Inspection Times, *Production Planning and Control*, 1993.

J. R. Jackson, Networks of Waiting Lines, *Operations Research*, 1957, *5*, No. 2, pp. 518-521.

J. R. Jackson, Jobshop-Like Queueing Systems, *Management Science*, 1963, *10*, No. 1, pp. 131-142.

N. Jeiswal, *Priority Queues*, Academic Press, New York, 1968.

M. V. Kalkunte, C. Saring and W. E. Wilhelm, Flexible Manufacturing Systems: a Review of Modeling Approaches for Design, Justification and Operation, *Flexible Manufacturing Systems: Methods and Studies*, A. Kusiak Ed., 1986, pp. 3-25, Elsevier, Amsterdam.

A. S. Kapadia and B. P. Hsi, Steady State Waiting Time in a Multicenter Job Shop, *Naval Res. Logist. Quart.*, 1978, *25*, pp. 149-154.

M. Kamath, R. Suri and J. L. Sanders, Analytical Performance Models for Closed-Loop Flexible Assembly Systems, *Int. J. FMSs*, 1988, *1*, pp. 51-84.

J. S. Kaufman, Blocking in a Shared Resource Environment, *I.E.E.E. Trans. on Communications*, 1981, *Com-29*, pp. 1474-1481.

J. Keilson and L. D. Servi, Dynamics of the *M/G/1* Vacation Model, *Operations Research*, 1987, *35*, pp. 575-582.

J. Keilson, Blocking Probabilities for *M/G/1* Vacation Systems with Occupancy Level Dependent Schedules, *Operations Research*, Forthcoming.

F. P. Kelly, *Reversibility and Stochastic Networks*, Wiley, New York, 1979.

D. G. Kendall, Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Imbedded Markov Chains, *Ann. Math. Statist.*, 1953, *24*, pp. 338-354.

A. S. Kiran and R. J. Krason, Automating Tooling in a Flexible Manufacturing System, *Industrial Ingineering*, April 1988, pp. 52-57.

H. T. Klahorst, Flexible Manufacturing Systems: Combining Elements to Lower Cost, and Flexibility, *A.I.I.E.*, 1981, *13*, No. 11, pp. 112-117.

L. Kleinrock, *Queueing Systems*, *2*, Computer Applications, Wiley, New York.

G. Klimov, Time Sharing system I, *Theor. Probability Appl.*, 1974, *19*, pp. 532-551.

J. Kelemenia and S. B. Gershwin, An Algorithm for the Computer Control of a Flexible Manufacturing System, *A.I.I.E.*, 1983, *15*, No. 4, pp. 353-362.

J. Kelemenia and S. B. Gershwin, Flow Optimization in Flexible Manufacturing Systems, *Int. J. Prod. Res.*, 1985, *23*, No. 1, pp. 81-96.

H. Kobayashi, Application of the Diffusion Approximation to Queueing Networks. Part I: Equilibrium Queue Distributions, *J. Assoc. Comput. Mach.*, 1974, *21*, pp. 316-328.

A. G. Konheim and M. Reiser, A Queueing Model with Finite Waiting Room and Blocking, *J. Assoc. Comput. Mach.*, 1976, *23*, pp. 328-341.

A. G. Konheim and M. Reiser, Finite Capacity Queueing Systems with Applications in Computer Modeling. *S.I.A.M. J. Comput.*, 1978, *7*, pp. 210-229.

A. E. Krzesinske and A. Grcycling, Improved Linearizer Methods for Queueing Networks with Queue Dependent Centers, *A.C.M. SIGMETRICS Conf. Proc.*, Cambridge, MA, 1984, pp. 41-51.

A. Kusiak, *Modeling and Design of FMS*, Elsevier, Amsterdam, 1986.

P. J. KUEHN, Approximate Analysis of General Networks by Decomposition, *I.E.E.E. Trans. on Commun.*, 1979 *a*, *Com-27*, No. 1, pp. 113-126.

P. J. KUEHN, Analysis of Switching System Control Structure by Decomposition, *Proceeding* of 9th International Teletraffic Congress, Spain, 1979 *b*.

S. S. LAM, Store-and- Forward Buffer Requirements in a Packet Switching Network, *I.E.E.E. Trans. Commun*, 1976, *Com-24*, pp. 394-403.

E. L. LAWLER and M. D. BELL, A Method for Solving Discrete Optimization Problems, *Operations Research*, 1966, *14*, No. 3, pp. 1098-1112.

H. F. LEE, M. M. SRINIVASAN and C. A. YANO, An Algorithm for the Minimum Cost Configuration Problem in Flexible Manufacturing Systems, *Int. J. of Flexible manufacturing Systems*, Forthcoming.

B. MAIONE, Q. SEMERARO and B. TURCHIANO, Closed Analytical Formulae for Evaluating Flexible Manufacturing System Performance Measures, *Int. J. Prod. Res.*, 1986, *24*, No. 3, pp. 583-592.

R. MALHAME and K. BOUKAS, Transcient and Steady-States of Statistical Flow Balance Equations in Manufacturing Systems, *Proceedings* of the Third *O.R.S.A./T.I.M.S.* Conference on *FMSs*, 1989, pp. 339-345.

W. G. MARSHAL, Numerical Performance of Approximate Queueing Formulae with Application to Flexible Manufacturing Systems, *Annals of Operations Research*, 1985, No. 3, pp. 141-152.

R. MARIE, An Approximate Analytical Method for General Queueing Networks, *I.E.E.E. Transactions on Software Engineering*, 1979, *SE-5*, No. 3, pp. 530-541.

J. MEILIJSON and U. YECHIALI, On Optimal Right of Way Policies at a Single Server Station when Insertion of Idle Times is Permitted, *Stochastic Processes and Applications*, 1977, *6*, pp. 25-32.

B. MELAMED, On Reversibility of Queuing Networks, *Stoch. Proc, Appl.*, 1982, *3*, pp. 227-236.

G. MENGA, B. BRUNO, R. CONTERNO and M. A. DATO, Modeling *FMS* by Closed Queuing Network Analysis Methods, *I.E.E.E. on Components, Hybrids, and Manufacturing Technology*, 1984, *BHMT-7*, No. 3, pp. 241-248.

F. R. MOORE, Computational Model of a Closed Queueing Network with Exponential Servers, *IBM J. R. and D.*, 1972, *16*, Dec., pp. 567-581.

P. M. MORSE, *Queues, Inventories and Maintenance*, Wiley, New York, 1963.

S. S. NAIR and M. F. NEUTS, A Priority Rule Based on the Ranking of the Service Times for *M/GI* Queues, *Operations Research*, 1969, *17*, No. 2, pp. 466-473.

S. S. NAIR and M. F. NEUTS, An Exact Comparison of Waiting Times Under Three Priority Rules, *Operations Research.*, 1971, *19*, No. 2, pp. 414-423.

V. NANDKEOLYAR and D. P. CHRISTY, Evaluating the Design of Flexible Manufacturing Systems, *Int. J. of Flexible Manufacturing Systems*, 1992.

P. NAOR, On the Regulation of Queue Size by Levying Tolls, *Econometrica*, 1969, *27*, pp. 15-24.

P. NASH and R. R. WEBER, Dominant Strategies in Stochastic Allocation and Scheduling Problems, in *Deterministic and Stochastic Scheduling*, Ed., DORDRECHT The Netherlands, Reidel, 1982, pp. 343-353.

E. D. NEUSE and K. M. CHANDY, SCAT: A Heuristic Algorithm for Queueing Network Models of Computing Systems, *A.C.M. SIGMETRICS* Conf. Proc. 10, 1981, No. 3, pp. 59-79.

G. F. NEWELL, *Application of Queueing Theory*, Chapman and Hall, London, 1971, Chapter 6.

G. F. NEWELL, *Approximate Behavior of Tandem Queues*, Springer-Verlag, Berlin, 1980.

M. PENNOTTI and M. SCHWARTZ, Congestion Control in Store and Forward Tandem Links, *I.E.E.E. Trans. Commun.*, 1975, *Com-23*, pp. 1434-1443.

H. G. PERROS, Queueing Networks with Blocking: A Bibliograhy, *Performance Evaluation Review*, 1984, *12*, pp. 8-12.

N. U. PRABHU, *Queues and Inventories*, Wiley, New York, 1965.

G. PUJOLLE and W. AI, A Solution for Multiserver and Multiclass Open Queueing Networks, *INFOR*, 1986, *24*, No. 3, pp. 221-230.

M. REISER and H. KOBAYASHI, Accuracy of Diffusion Approximation of Some Queueing Systems, *IBM J. of R. and D.*, 1974, *8*, March, pp. 114-124.

M. REISER and S. S. LAVENBERG, Mean Value Analysis of Closed Multichain Queueing Networks, *Comm. A.C.M.*, 1980, *27*, No. 2, pp. 313-322.

I. RUBIN, Path Delays in Communication Networks, *Appl. Math. Optimiz.*, 1975, *1*, 3, pp. 193-221.

M. S. SABBAGH, A General Lexicographic Partial Enumeration Algorithm for the Solution of Integer Nonlinear Programming Problems, *Ph. D. dissertation*, the School of Engineering and Applies Science, George Washington University, 1983.

C. H. SAUER and K. M. CHANDY, Approximate Analysis of Central Server Models, *IBM J. Res. Develop.*, 1975, *19*, pp. 301-313.

L. E. SCHRAGE, A Proof of Optimality on the Shortest Remaining Processing Time Discipline, *Operations Research*, 1968, *16*, pp. 687-690.

L. E. SCHRAGE, An Alternative Proof of a Conservation Law for the Queue *G/G/1*, *Operations Research*, 1970, *18*, pp. 185-187.

L. E. SCHRAGE, Random Results in Scheduling: Implications of Queueing Theory for Scheduling, Lecture note, University of Chicago, 1974.

L. E. SCHRAGE and L. W. MILLER, The Queue *M/G/1* with the Shortest Remaining Processing Time Discipline, *Operations Research*, 1966, *14*, pp. 670-683.

P. J. SCHWEITTZER, Iterative Solution of the Functional Equations of Undiscounted Markov Renewal Programming, *J. Mathematical Analysis and Applications*, 1971, *34*, No. 3, pp. 495-501.

P. J. SCHWEITZER, Maximum Throughput in Finited Capacity Open Queueing Networks with Product-Form Solutions, *Management Science*, 1977, *24*, No. 2, pp. 217-223.

P. J. SCHWEITZER, Approximate Analysis of Multiclass Closed Networks of Queue, International Conference of Stochastic Control and Optimization, Free University, Amsterdam, 1979.

P. J. SCHWEITZER and A. SEIDMANN, Production Rate Optimization for *FMSs* with Distinct Multiple Job Visits to Work Centers, *Int. J. of Flexible Manufacturing Systems*, Forthcoming.

A. SEIDMANN, One-Line Scheduling of Flexible Manufacturing Cell with Stochastic Sequence-Dependent Processing Rates, *Int. J. Prod. Res.*, 1987, *25*, No. 6, pp. 907-924.

P. J. SCHWEITZER and S. SHALEV OREN, Computerized Close Queueing Network Models of Flexible Manufacturing Systems: A Comparative Evaluation, *Large Scale Systems*, 1987, *12*, No. 2, pp. 91-107.

P. J. SCHWEITZER and P. F. SCHWEITZER, Part Selection Policy for a Flexible Manufacturing Cell Feeding Several Production Lines, *A.I.I.E.*, 1984, *16*, No. 4, pp. 355-362.

P. J. SCHWEITZER and A. TENENBAUM, Optimal Stochastic Scheduling of Flexible Manufacturing Systems with Finite Buffers, *Proceedings of the Second O.R.S.A./T.I.M.S. Conference on FMSs*, Elsevier, Amsterdam, 1986.

K. C. SEVICK, A. I. LEVY, S. K. TRIPATHI and J. L. ZAHORJAN, Improving Approximations of Aggregated Queueing Network Subsystems, *Comp. Performance Modeling Measurement and Evaluation*, 1977.

S. SHALEV-OREN, A. SEIDMANN and P. J. SCHWEITZER, Analysis of Flexible Manufacturing Systems with Priority Scheduling: PMVA, *Annals of Operations Research*, 1985, *3*, pp. 115-139.

J. G. SHANTHIKUMAR, On Reducing Time Spent in *M/G/I* Systems, *E.U.R.. J. Operations Research*, 1982, *9*, pp. 286-294.

J. G. SHANTHIKUMAR, Comparison of Dispatch Polices for a Single Server Queueing Model with Limited Operational Control, *Int. J. Prod. Res.*, 1984, *22*, No. 3, pp. 389-403.

J. G. SHANTHIKUMAR and M. GOCMEN, Heuristic Analysis of Closed Queueing Networks, *Int. J. Prod. Res.*, 1983, *21*, No. 5, pp. 675-681.

J. G. SHANTHIKUMAR and J. A. BUZACOTT, On the Approximations to the Single Server Queue, *Int. J. Prod. Res.*, 1980, *18*, No. 6, pp. 255-263.

J. G. SHANTHIKUMAR and J. A. BUZACOTT, 1981, Open Queueing Network Models of Dynamic Job Shops, *Int. J. Prod. Res.*, 1981, *19*, No. 3, pp. 255-266.

J. G. SHANTHIKUMAR and J. A. BUZACOTT, The Time Spent in a Dynamic Job Shop, *EURO J. Operations Research*, 1984, *17*, pp. 215-216.

J. G. SHANTHIKUMAR and K. E. STECKE, Reducing Work-in-Process Inventory in Certain Classes of Flexible Manufacturing Systems, *EURO J. Operations Research*, 1986, *26*, pp. 266-271.

J. G. SHANTHIKUMAR and R. G. SARGENT, A Unifying View of Hybrid Simulation Analytic Models and Modelin, *Operations Research*, 1983, *31*, No. 6, pp. 1030-1052.

J. G. SHANTHIKUMAR and D. D. YAO, Stochastic Monotonicity of the Queue Lengths in Closed Queueing Networks, *Operations Research*, 1987 *a*, *35*, No. 4, pp. 583-588.

J. G. SHANTHIKUMAR and D. D. YAO, Optimal Server Allocation in a System of Multi-Server Stations, *Management Science*, 1987 *b*, *33*, No. 9, pp. 1173-1180.

J. G. SHANTHIKUMAR and D. D. YAO, On Server Allocation in Multiple Center Manufacturing Systems, *Operations Reseach*, 1988, *36*, No. 2, pp. 333-342.

A. SHUM and J. P. BUZEN, The E.P.F. Technique: A Method for Obtaining Approximate Solutions to Closed Queueing Networks with General Service Times, *Measuring, Modeling and Evaluating Computer Systems*, BEILNER and E. GELENBE Ed., North-Holland, 1977.

D. R. SMITH and W. WHITT, Resource Sharing for Efficiency in Traffic Systems, *Bell System Tech. J.*, 1981, *60*, pp. 39-55.

M. L. SMITH, R. RAMESH, R. A. DUDER and E. E. BLAIR, Characteristic of U.S. Flexible Manufacturing Systems - a Survey, *Proceedings of the second O.R.S.A./T.I.M.S. Conference on FMSs*, 1986, pp. 477-485, Elsevier, Amsterdam.

K. T. So, Allocating Buffer Storage in a Flexible Manufacturing System, *Inst. J. FMSs*, 1989, *1*, No. 3, pp. 223-237.

J. J. SOLBERG, A Mathematical Model of Computerized Manufacturing Systems, 4th International Conference on Production Research, Tokyo, 1977.

P. SOLOT and J. M. BASTOS, MULTIQ: a Queueing Model for *FMSs* with Several Pallet Types, *Journal of the Operational Research Society*, 1988, *39*, No. 9, pp. 811-821.

M. M. SRINIVASAN, On Extending the Scope of Bounding Techniques for Close Queueing Networks, *Large Scale Systems*, 1987, *12*, No. 2, pp. 125-142.

J. SPRAGINS, Analytical Queueing Models: Guest Editor's Introduction, *I.E.E.E. Transactions on Computers*, 1980, *13*, 4, pp. 9-11.

K. E. STECKE, Formulation and Solution of Nonlinear Integer Production Planning Problems for Flexible Manufacturing System, *Management Science*, 1983, *29*, No. 3, pp. 273-288.

K. E. STECKE, Design, Planning, Scheduling, and Control Problems of Flexible Manufacturing System, *Proceedings* of the First *O.R.S.A./T.I.M.S.* Special Interest Conference

on Flexible Manufacturing Systems: *Operations Research Models and Applications*, Ann Arbor, Michigan, 1984.

K. E. STECKE, Useful Models to Address *FMS* Operating Problems, *Proceedings of the I.F.I.P. Conference*, Advances in Production Management Systems, Budapest, Hungary, 1985.

K. E. STECKE, A Hierarchical Approach to Solving Machine Grouping and Loading Problems of Flexible Manufacturing Systems, *EJOR*, 1986, *26*, pp. 212-243.

K. E. STECKE and T. L. MORIN, The Optimality of Balancing Workloads in Certain Types of Flexible Manufacturing Systems, *EURO J. Operations Research*, 1985, *25*, pp. 68-82.

K. E. STECKE ans J. J. SOLBERG, Loading and Control Policies for a Flexible Manufacturing System, *Int. J. Prod. Res.*, 1981, *19*, No. 5, pp. 481-490.

K. E. STECKE and J. J. SOLBERG, The Optimality of Unbalancing Both Workloads and Machine Group Sizes in Closed Grueueing Networks of Multiserver Queues, *Operations Research*, 1985, *33*, No. 4, pp. 882-910.

S. STIDHAM, Socially and Individually Optimal Control of Arrivals to a *GI/M/1* Queue, *Management Science*, 1978, *24*, pp. 1598-1610.

S. STIDHAM, Optimal Control of Admission to a Queueing System, *I.E.E.E. Trans. On Automatic Control*, 1985, *AC-30*, pp. 705-713.

R. SURI, Resource Management in Large Systems, PHD, *Thesis*, Harvard, Division of Applied Science, 1979.

R. SURI, An Overview of Evaluative Models for Flexible Manufacturing Systems, *Annals of Operations Research*, 1985, *3*, pp. 61-69.

R. SURI and R. R. HILDEBRANT, Modelling Flexible Manufacturing Systems Using Mean-Values Analysis, *J. Manufacturing system*, 1984, *3*, No. 1, pp. 27-38.

TAKAGI, *Polling Systems*, MIT Press, Cambridge, Mass, 1986.

TAKAGI, Blocking when Service is Required from Several Facilities Simultaneously, *A.T.T. Tech. J.*, 1985, *64*, pp. 1807-1856.

J. TALAVAGE, R. G. HANNAN, *FMSs in Practice, Application, Design, and Simulation*, Marcel Dekker Inc., New York, 1988.

C. S. TAPIERO and L. F. HSU, Quality Control of an Unreliable Random *FMS*: with Bernoulli and CSP Sampling, *Int. J. Prod. Res.*, 1988, *26*, No. 6, pp. 1125-1135.

H. THOMAS, Flexible Manufacturing Systems: Combining Elements to Lower Costs and Flexibility, *A.I.I.E.*, 1981, Nov., pp. 113-116.

N. M. VAN DIJK, Comment on Yao and Buzacotts' Modeling a Class of Flexible Manufacturing Systems With Reversible Routing, *Operations Research*, 1989, *37*, No. 5, pp. 845-846.

A. J. VAN LOOVEREN, L. F. GELDERS and L. N. VAN WASSENHOVE, A Review of *FMS* Planning Models, *Modeling and Design of FMSs*, A. KUSIAK Ed., 1986, pp. 3-31, Elsevier, Amsterdam.

B. VINOD and T. ALTIOK, Approximating Unreliable Queueing Networks Under the Assumption of Exponentiality, *J. Opl. Res. Soc.*, 1986, *37*, No. 3, pp. 309-316.

B. VINOD and T. C. JOHN, On Optimal Capacities for Repair Facilities in Flexible Manufacturing Systems, *Flexible Manufacturing Systems: Methods and Studies*, A. KUSIAK Ed., 1986, pp. 61-73, Elsevier, North-Holland.

N. VINOD and M. SABBAGH, Optimal Performance Analysis of Manufacturing Systems Subject to Tool Availability, *E.U.R.O. J. Operations Reseach*, 1986, *24*, pp. 398-409.

B. VINOD and J. J. SOLBERG, Performance Models for Unreliable Flexible Manufacturing Systems, *OMEGA*, 1984, *12*, No. 3, pp. 299-308.

T. R. WILLEMAIN, Approximate Analysis of a Hierarchical Queueing Network, *Operations Research*, 1972, *20*, pp. 522-544.

W. L. WINSTON, Assignment of Customers to Servers in a Heterogeneous Queueing System with Switching, *Operations Research*, 1977 *a*, *25*, pp. 468-483.

W. L. WINSTON, Optimal Dynamic Rules for Assigning Customers to Servers in a Heterogeneous Queueing System, *Naval Research Logistics*, 1977 *b*, *24*, pp. 293-300.

W. L. WINSTON, Optimality of the Shortest Line Discipline, *J. Appl. Prob.*, *14*, pp. 181-189.

W. WHITT, The Queueing Network Analyzer, *Bell System Tech. J.*, 1983 *a*, *62*, No. 9, pp. 2779-2815.

W. WHITT, Performance of Queueing Network Analyzer, *Bell system Tech. J.*, 1983 *b*, *63*, No. 9, pp. 2817-2843.

W. WHITT, Approximations to Departure Processes and Queues in Series, *Naval Res. Logist. Quart.*, 1984, *31*, pp. 499-521.

W. WHITT, The Best Order of Queues in Series, *Management Science*, 1985, *31*, No. 3, pp. 475-487.

D. D. YAO, Queueing Models of Flexible Manufacturing Systems. *Ph. D. dissertation*, Dept., of I.E., University of Toronto, Canada, 1983.

D. D. YAO, Some Properties of the Throughput Function of Closed Networks of Queue, *OR Letters*, 1985, *3*, No. 6, pp. 313-317.

D. D. YAO, An Optimal Storage Model for a Flexibles Manufacturing System, *Flexible Manufacturing Systems: Methods and Studies*, A. KUSIAK Ed., 1986, pp.113-125.

D. D. YAO, Majorization and Arrangement Orderings in Open Queueing Networks, *Annals of Operations Research*, 1987, *9*, pp. 531-543.

D. D. YAO, The Arrangement of Servers in an Ordered-Entry System, *Operations Research*, 1987, *35*, No. 5, pp. 759-763.

D. D. YAO and J. A. BUZACOTT, Modelling the Performance of Flexible Manufacturing Systems, *Int. J. Prod. Res.*, 1985 *a*, *23*, No. 5, pp. 945-959.

D. D. YAO and J. A. BUZACOTT, Queueing Models for a Flexible Machining Station Part I: The Diffusion Approximation Part II: The Method of Coxian Phases, *Eur. J. Operations Research*, 1985 *b*, *19*, pp. 233-252.

D. D. YAO and J. A. BUZACOTT, Modeling a Class of State-Dependent Routing, *Annals of Operations Research*, 1985 *c*, *3*, pp. 153-167.

D. D. YAO and J. A. BUZACOTT, The Exponentialization Approach to Flexible Manufacturing System Models with General Processing Times, *E.U.R.. J. Operations Research*, 1986 *a*, *24*, pp. 410-416.

D. D. YAO, J. A. BUZACOTT, Models of Flexible Manufacturing Systems with Limited Local Buffers, *Int. J. Prod. Res.*, 1986 *b*, *24*, No. 1, pp. 107-118.

D. D. YAO and J. A. BUZACOTT, Modeling a Class of Flexible Manufacturing Systems with Reversible Routing, *Operations Research*, 1987, *35*, No. 1, pp. 87-93.

D. D. YAO and G. SHANTHIKUMAR, Some Resource Allocation Problems in Multi-Cell Systems, Proceedings of the Second O.R.S.A./T.I.M.S. Conference on *FMSs*, 1986, pp. 245-255.

D. D. YAO and G. SHANTHIKUMAR, The Optimal Input Rates to a System of Manufacturing Cells, *INFOR*, 1987, *25*, No. 1, pp. 57-65.

D. D. YAO and S. C. KIM, Some Order Relations in Closed Metworks of Queues with Multiserver Stations, *Naval Res. Logis.*, 1987, *34*, pp. 53-66.

U. YECHIALI, On Optimal Balking Rules and Toll Charges in a *GI/M*/1 Queueing Process, *Operations Research*, 1971, *19*, pp. 349-370.