A Review of the Open Queueing Network Models of Manufacturing Systems

by
Gabriel R. Bitran
Sriram Dasu

WP #3229-90-MSA December 1990

A Review of Open Queueing Network Models of Manufacturing Systems

Gabriel R. Bitran

Sloan School of Management, M.I.T., Cambridge MA 02139

Sriram Dasu

Anderson Graduate School of Management, U.C.L.A. , Los Angeles, CA 90024

This research was partially supported by the "Leaders for Manufacturing Program", and by the UCLA Senate Committee on Grants #99

A Review of Open Queueing Network Models of Manufacturing Systems

Abstract: In this paper we review open queueing network models of manufacturing systems. The paper consists of two parts. In the first part we discuss design and planning problems arising in manufacturing. In doing so we focus on those problems that are best addressed by queueing network models.

In the second part of the paper we describe the developments in queueing network methodology. We are primarily concerned with features such as general service times, deterministic product routings, and machine failures - features that are prevalent in manufacturing settings. Since these features have eluded exact analysis, approximation procedures have been proposed. In the second part of this paper we review the developments in approximation procedures and highlight the assumptions that underlie these approaches.

A significant development in the study of queueing network models is the discovery (empirical) that under conditions that are not very restrictive in practice: (i) equilibrium expected queue lengths behave as if they are convex functions of the processing rate of the server, and (ii) altering the processing rate at one station has minimal effect on the equilibrium expected queue lengths at other stations in the network. As a result researchers have been able to approximate some of the optimal design problems by convex programs. In the second part of this paper we describe these developments.

Inspite of the advances made in the analysis of open queueing networks, several of the problems described in the first part of the paper cannot be analyzed without further progress in

methodology. One of the objectives of this paper is to expose the gap between the problems arising in manufacturing and the analytical tools that are currently available. We hope that by first describing the problems and then discussing the methodological developments the gap becomes apparent to the reader.

I. INTRODUCTION

Job shops are complex manufacturing systems which process a wide variety of products in low volumes. Two dominant characteristics of job shops are complex flows through the shop, and long queues of jobs in front of machines. It is not uncommon for a job to spend more than 90% of the time in the facility waiting for machines to become available. The challenge of managing day to day operations in this environment has stimulated an enormous interest in sequencing and scheduling problems. In comparison less attention has been paid to tactical and strategic problems such as choice of equipment, capacity planning, allocation of products to different plants, and determination of lead times. To properly address these issues we need models that provide good estimates of the medium to long term performance of manufacturing systems. Over the last two decades there has been a renewed interest in open queueing network models, and major advances have been made in their (approximate) analysis. These studies were in part motivated by applications in manufacturing settings. In this paper we review open queueing network models of job shops, with primary emphasis on models that facilitate the design of job-shop-like manufacturing systems.

To outline the scope of this survey we start by pointing out what is being excluded. First, we do not extensively survey models proposed to address operational problems arising in job shops.

Almost all the models developed for operational problems assume that the processing requirements of jobs are deterministic, and rely on combinatorial methods to provide solutions. The literature dealing with this subject is huge and several excellent surveys have been written on this topic (Graves 81, Panwalker and Iskander 77). Nevertheless for the sake of completeness we briefly discuss some of the operational problems. Second, we focus essentially on open queueing network models of job shops and do not explore models that are specialized to pure flow shops (tandem queues) (eg. Reich 57, Friedman 65, Whitt 85b), queueing models of flexible manufacturing systems (eg. Buzacott and Yao 86a and 86b), and queueing network models of systems with finite waiting rooms (eg. Gershwin and Schick 83, Brandwajn and Jow 88, Altiok and Perros 89).

Open queueing network models have been applied in many domains including computer science, communication engineering, and manufacturing and service operations, resulting in a large body of literature. Thus to provide an exhaustive review of the developments in the study of open queueing networks is a task of a magnitude well beyond the scope of this paper. An attempt to provide such an overview will also be repetitious of excellent surveys done by Lemoine (77), Disney and Konig (85), and Buzacott and Yao (86a). To minimize the overlap with these surveys we concentrate on recent developments in open queueing network models that incorporate aspects such as general service times, and deterministic product routings, features that are common to manufacturing systems. With these features in place exact analysis has not been possible. However researchers have been very successful in developing good approximation procedures.

I.1 CONTENTS OF THE REVIEW

We start the review by defining key manufacturing terms in section II. In section III we

develop a list of strategic and tactical planning problems that may be addressed by queueing network models.

In sections IV and V we focus on developments in open queueing network models. In section IV we describe methods for evaluating the performance of open network of queues. We are primarily concerned with features such as general service times and deterministic routing. For such networks exact analysis has not been possible and approximation procedures have been developed. Much of section IV is devoted to the parametric decomposition approach which has been very effective in estimating the first moment of the queue length in general networks.

Section V is devoted to recent developments in optimization models. This section in turn consists of two parts. In the first part we describe procedures developed for optimal design of queueing networks. In recent years based on developments in Brownian control theory, progress has been made in deriving near optimal rules for controlling the flow through the network. In the second part of section V we briefly touch upon these advances. Finally in section VI we conclude by pointing out avenues for further research.

II TERMINOLOGY

To avoid ambiguity and for the benefit of readers not familiar with manufacturing systems we specify below the usage of key terms. Our usage of these terms conforms closely to that in the manufacturing and operations management literature.

Operation: An operation is an elemental task which requires resources such as machines, tools, and

labor. An example of a task would be to drill a half inch diameter hole using a drilling machine, a drill bit, an attachment to hold the part that is being drilled, and a drill operator. There may be several drilling machines and operators (not necessarily identical) that can perform this task.

Item: An item is a distinct physical part produced by the facility. Associated with each item are a set of operations, and a precedence relationship that may constrain the sequence in which the operations can be executed.

Product: A distinct commodity produced by the facility. Often a product is produced by assembling together several items. In this survey we exclude facilities that assemble items that are also manufactured in the same facility. Hence the terms items and products will be used interchangeably.

Product families: Set of products that have similar manufacturing requirements. For the purpose of this survey we assume that all items belonging to the same family follow similar routes through the manufacturing facility.

Machine center or stage: A set of machines that are capable of performing similar operations, but are not necessarily identical. An example of a machine center could be a set of drilling machines with varying horse-powers. We distinguish between two types of machines that we call discrete and batch machines. A discrete machine processes one item at a time, whereas a batch machine can operate on several items simultaneously. The processing time of a batch machine does not depend on the number of items being processed, but there is a limit on the number of items that can be processed in a batch.

Job Shops: Based on the number of products that are manufactured in a facility, manufacturing systems are classified into assembly lines or continuous flow systems, and intermittent systems consisting of batch and job shops. Readers are referred to Buffa and Sarin(87), Schmenner(89) or any introductory text on operations management for a detailed description of these three systems. Job shops produce a large number of products and the average demand per period for each item is small relative to the capacities of the machines required to produce the item. Unlike flow shops, in job shops the sequence of visitations to machine centers varies from job to job, and the flow of items through the facility is not unidirectional. As a result of the diversity in the products produced in a job shop the processing requirement at each machine center varies from one job to the next. Also, each job may require different amount of processing at different machine centers. Waiting becomes inevitable due not only to imbalances in the work load at different machine centers but also due to factors such as variability in the arrival of the work orders to the manufacturing facility, machine failures and malfunctioning, absenteeism, and unavailability of proper raw materials and tools. Another significant factor that contributes to queues in job shops is the uncertainty in processing requirements. Typical sources of uncertainty in the processing requirements are production processes that produce defective items that have to be reworked. An example of a manufacturing system that is subject to most of these forms of variability and uncertainty is a semi-conductor wafer fabrication facility (Chen et al 88). It is therefore not surprising that these facilities have motivated considerable interest in open queueing network models(Bitran and Tirupati 88,89a, Chen et al 88, Harrison and Wein 90a).

Job shops have been further classified into open and closed job shops. In open job shops each job is unique and is produced to a custom order. Closed job shops produce only a specific set of products, typically described in a company catalog.

Batch Shops: Almost all the characteristics of job shops are shared by batch shops. Batch shops can be viewed as special cases of closed job shops. The primary difference between batch and job shops is the number of products produced by the facility. Batch shops produce fewer products. Those readers not familiar with manufacturing systems, can for the purpose of this survey, ignore the difference between these two systems.

III. MANAGERIAL PROBLEMS

In what follows we describe a set of the problems associated with job shops, that are best addressed by queueing network models. We partition the problems into three categories : strategic, tactical and operational problems (Anthony 65, Hax and Candea 84).

III.1 STRATEGIC PROBLEMS

One of the objectives of strategic planning is to identify long term goals of the company. To achieve these goals the strategic plan determines the resources that are to be utilized, and the policies that govern the use of these resources (Anthony 65). An important component of the strategic plan is the design of the manufacturing system. System design involves (i) choice of technologies, (ii) acquisition of capital equipment, (iii) the allocation of the products to different plants,(iv) choice of location, and (v) design of the distribution systems. For the purpose on hand we restrict the scope of the design problem to the first three factors.

The design of a plant is a major constraint under which the operating manager makes decisions. For instance, if several machine centers operate at or near their capacities and there is

considerable diversity in the processing requirements of the products assigned to that plant, then queues at machine centers are likely to be very long and, more importantly, the variance of the time a job spends in the system is also likely to be very high. Among the primary tasks of the operating manager are to assign and predict completion times for jobs, control the flow of jobs through the facility, so that due dates are not violated, control the work-in-process inventory levels, respond to unanticipated loss of capacity due to factors such as uncertain yields and machine failures, and strive to improve the processes and products. Since most scheduling, sequencing and routing problems are known to be extremely difficult to solve (Garey, Johnson and Sethi 76, Rinnooy Kan 76, Lenstra, Rinnooy Kan and Brucker 77) the system should be designed so that simple real time control rules are adequate to obtain good performance.

System design involves a trade-off between (i) the fixed cost of facilities and equipment, (ii) variable cost of operating the facility, (iii) processing capabilities of the machines (iv) throughput, (v) lead times, (vi) work-in-process levels and (vii) the complexity of managing the facility.

There are many factors that determine the performance characteristics of a facility including: (i) technology - type of equipment, flexibility of the machines, process controllability, reliability, etc; and (ii) product characteristics - level of standardization among products, tightness of specifications (effects yields), processing flexibility (routing, and machine requirements), variability in demand, etc. Clearly there are a large number of variables and objectives that have to be taken into consideration while designing or re-designing a facility. The tasks are further compounded by the risks and uncertainty induced by the continuous evolution of technologies and products. To provide a meaningful discussion of the design problem we restrict the scope of the problem by fixing some variables and allowing others to be determined by the designer. We only consider those design

objectives that are related to the formation of queues in the shop. Listed below are three important classes of design problems. In all three classes we assume that the demand patterns for the products are known. The principal decision to be made in each of the problems is the selection of physical assets.

In the first class of problems(SP1), the performance characteristics of the system are externally determined, and the designer's task is to determine the lowest cost of the facility. In the second class of problems (SP2) there is a capital budget that constraints the amount that can be spent on acquiring new machines. This constraint is particularly significant when an existing facility is being upgraded. Under such circumstances the models should identify the usage of the available capital that provides the best system performance.

Due to diseconomies of scope, the management of a manufacturing facility becomes increasingly difficult as the number of products produced in the facility increases. Hence it is desirable to partition a large facility into smaller more focused sub-units. However, partitioning a facility can result in duplication of equipment. As a result there is a trade-off between increased equipment cost and reduced managerial complexity. In the third class of problems (SP3) we discuss models that can guide managers in partitioning facilities.

SP1: Targeted System Performance:

In this class of problems, the designer has to determine the processing capabilities of each machine center in order to achieve a desired system performance. Performance measures of interest are mean and variance of work-in-process levels at each machine center, mean and variance

of the sojourn time for each product family, and the probability of the sojourn time exceeding a particular value. We assume that the cost of the machines and their capabilities are known to the designer and (s)he may have to choose among alternate technologies. The design problems belonging to this class can be formulated as optimization problems, where the objective is to minimize the cost of the plant, and the constraints are the desired system performance. Given below are two problems belonging to this class:

SP1.1 Targeted Work-in-Process Levels

Objective: Minimize total cost of equipment

Decision Variables: Capacity of Each Machine Center, Technology

Constraints: Upper bound on the mean number of jobs in the system

SP1.2 Targeted Lead Times

Objective and decision variables are the same as in SP1.1,

Constraints: Upper bounds on the means and variances of the sojourn times for each product family.

Problem SP1.1 addresses the relationship between working capital requirements and the cost of equipment. The constraint of SP1.1 corresponds to the total work-in-process investment. Since system design is based on multiple criteria it is useful to develop curves that reflect the trade-off between work-in-process inventory costs and the cost of machinery. This can be done by solving SP1.1 parametrically by varying the upper bound on the permissible average inventory level.

The length of the lead times, and the corresponding reliability and consistency are important

elements of a firm's competitive strategy. This is the motivation for problem SP1.2

Clearly the performance of the system depends not only on the equipment selected, but also on the control rules employed to manage the day-to-day operations. However, since the system must be designed for the long term requirements of the firm, the design decisions are based on imprecise long term forecasts. The forecasts lack details in terms of the timing of the demand, and estimate demands for product families rather than individual products. Consequently at this level of decision making it may be adequate to assume that the shop is operated using very simple rules such as processing the jobs in the sequence in which they arrive at the machine center. This is a concern that the modeler must resolve in all strategic and tactical problems.

SP2: Optimal System Performance:

In this class of problems the designer operates under a budget constraint that limits the amount that can be spent in acquiring new machines. The objective of the design activity is to determine the capacities of each machine center so that some performance measure(s) is optimized. Examples of problems belonging to this class are:

SP2.1 Optimal Work-in-process levels

Objective: Minimize the weighted sum of the average number of jobs at each machine center

Decision Variables: Capacities of Machine centers, Technologies

Constraints: Upper bound on the investment

SP2.1 is analogous to SP1.1, its objective function corresponds to the average investment in work-in-process. Due to Little's law the problem of minimizing the weighted average sojourn times will be similar to SP2.1

As stated earlier, variability in work-in-process levels can be used as a measure of the complexity of managing a facility. Managerial intervention becomes necessary when the work loads are either very high or very low. For instance if the work loads are excessive then extra capacity has to be generated either by sub-contracting some jobs or by working extra hours. Also, expediting is more likely to occur when the work loads are high. Hence it is desirable to design the system so that work-load variations are minimized.

SP2.2 Optimal work load variation:

Objective: Minimize the weighted sum of variances of work loads at each machine center

The constraints and decision variables are the same as in SP2.1

For any choice of physical assets, like other performance measures, the variability in work load will depend considerably on the control rules employed to regulate the work flow through the facility (Graves 86, Matsuo and Gong 90, Denardo and Tang 89). Work load variability is also effected by the diversity in the products produced in the facility. Hence, it may be possible to reduce variability by partitioning the facility into smaller more homogenous units. We will return to this issue while discussing the third class of design problems. Although work load variability is

effected by factors other than capacity levels and technologies, it is an important system characteristic that should be considered while selecting the equipment. It is useful to develop models to measure the effect of technological factors such as machine failure rates and process capability (yield rates, yield variability, and flexibility) on lead time variability. Choice of technology can be an important factor in problems SP2.2 and SP1.2.

SP3: Topology of the Facility (Partitioning of the Facility):

The primary task under consideration in this class is to partition a facility into smaller sub-units. Although we list this class separately, it can be taken into account while solving problems SP1 and SP2. In that case the designer, in addition to selecting the equipment, is permitted to partition the factory into independent sub-networks. We have elected to describe this class separately because inspite of its importance it has received very little attention. We describe these problems assuming that a facility is already in place and we are undertaking a redesign problem. Note that problems SP1 and SP2 are also equally applicable to existing facilities.

In manufacturing systems that produce a small number of products in high volumes it may be possible to set up a flow line for each product by dedicating a set of machines to that product. In such situations the interaction between different products is minimal. For instance the lead time of a product does not depend on the other products that are produced in that facility. However if the demands are not sufficiently high or if the cost of the machines is very high then several items will have to share the same equipment. This is often the case in the semi-conductor industry where machines needed for several processes such as metalization, and photolithography can cost over half a million dollars.

Even if the demand for each item is not sufficiently high to justify a dedicated line, it may be possible to group together a set of items whose requirements are similar and set up flow lines or cells for that group (Burbidge 89, Ahmadi and Matsuo 90). In this case, a single facility is partitioned into several flow lines or group technology cells.

In general, in job shops it will not be economical to partition the products into groups such that flow lines can be set-up for each group. However there are several advantages to partitioning the manufacturing facility into smaller more focused factories. It is well known that beyond some size manufacturing systems become very difficult to manage (Hayes and Wheelwright '84 chapter 3). Hayes and Wheelwright refer to the complexity, and chaos arising from increasing the number of products, processes, and specialists in a given plant as diseconomies of bureaucratization and confusion. Although this phenomena has been recognized, we are unaware of any study that proposes measures for managerial complexity. The primary benefits that are cited for developing focussed factories include simplified product flows, improved control, faster response to changes in demand, and greater predictability of completion times (Skinner 74). Simpler and more manageable facilities also enable product and process improvement.

Facilities that are easy to manage exhibit some important qualitative characteristics such as : job completion times can be predicted accurately (tight confidence intervals), sources of defects and failures can be identified easily and quickly, and the lead times for jobs are relatively small and the facility can readily respond to changes in demands. This suggests that a partial set of surrogate measures for the complexity of a system are (i) variances of the work loads at machine centers, (ii) variances of the time jobs spend in the system, (iii) number of products produced in the facility, (iv) number of different routes through the facility (for instance in a pure flow shop the number of

routes through the shop is only one), and (v) total number of transactions processed on an average day. Transactions are elemental activities such as movement of jobs from one machine center to the next, machine failures, and tool changes. The last two measure attempt to capture the complexity of the flow patterns in the shop. Several other measures can be proposed for the complexity of the flow in the system including the number of intersections between different routes in the shop.

In order to partition a plant into several focussed facilities, we must determine the products that are to be produced in each facility, and allocate equipment to each of the facilities. Ideally facilities should not share equipment and each product should be assigned only to one facility. Thus, the trade-off is between cost of extra capital equipment against reduced managerial complexity. A specific instance of a problem belonging to this class would be to identify the least cost partitioning of the factory subject to an upper bound on the number of products in each plant.

SP3.1 Targeted number in each plant

Objective: Minimize the cost of new machinery

Decision Variables: Products, types of equipment, capacity of each machine centers belonging to each plant

Constraints: Upper bound on number of products in each plant, and upper bounds on the mean lead time for each product family

SP3.2 Optimal route complexity

Objective: Minimize the maximum number of routes in each plant

Design Variables: same as SP3.1

Constraints: Upper bound on the capital cost of redesign, upper bound on the mean lead time for each product family

SP3.1 and SP3.2 are two simple illustrations of the type of problems that have to be solved to facilitate partitioning of a factory. Problem SP2.2 can also be included in this class. To develop formal optimization models that aid mangers in setting-up plants-with-in-plants we have to gain a better understanding of managerial complexity and develop good quantitative measures for complexity. We hope that this paper will draw attention to this problem area.

So far we have focussed on system design, and in particular we were concerned with factors that have to be considered while acquiring physical assets. Once the physical facilities have been decided on, the remaining tasks include determining the number of hours the facility is to operate, determining the appropriate lead times for each product, allocating the resources to various products, and controlling the flow of the jobs through the facility. These problems in turn can be partitioned into two categories tactical and operational control problems. While decisions regarding hard assets are made primarily at the strategic planning level, at the tactical level capacity is adjusted by changing man-power levels and the number of hours a facility operates. At the operational level there is very little opportunity to alter capacity, and the primary concern is with managing the flow (traffic) of jobs through the facility.

III.2 TACTICAL PROBLEMS

The planning horizon for the tactical plan is shorter than the horizon for the strategic plan. The actual horizon length will depend on the specific problem being addressed. Since the horizon

is shorter, the level of detail and the accuracy of the information available for tactical plans is higher. The level of uncertainty in the information is lower. Typically, managers responsible for the tactical plan are likely to report to those responsible for the strategic plan. As a result, the scope of the tactical plan may be restricted to one plant, and the capital that can be expended in acquiring machinery is significantly less than that considered by the strategic planners. Tactical plans are constrained to a large extent by the available physical assets, and the set of products that are produced in each plant can not be changed significantly. To simplify the discussion, and without significantly altering the contents of the discussion, we assume in this paper that the tactical plan can neither acquire new machines nor change the set of products that are produced in the facility.

There are three important decisions at the tactical level that are influenced by the formation of queues in the job shop. They are : (i) operating capacity; (ii) planned lead times for each product family; and (iii) lot sizes.

III.2.1 Capacity Planning

As stated above at the tactical level managers effect capacity primarily through the number of hours the facility is operated and the man-power levels. Problems analogous to SP1 and SP2 also arise at the tactical level, and will be referred to as TP1 and TP2, respectively. However, the scope of the decision variables is significantly less. For example TP1.1 described below is similar to SP1.2. Observe that the scope of the decision variables is restricted. Consequently, unlike SP1.2, TP1.1 may be infeasible.

TP1.1 Targeted Lead Time Performance

Objective: Minimize total man-power costs+ Work-in-process inventory costs

Decision Variables: Number of hours each machine center is operated

Constraints: Existing equipment and technology, upper bounds on the means and variances of the sojourn times for each item.

As stated earlier the trade-off between capacity and work-in-process levels is useful to model and this is done in problem TP1.2.

TP1.2 Optimal Work-in-Process Levels

Objective: Minimize the weighted average of the number of jobs in the system

Decision Variables: Number of hours each machine center is operated

Constraints: Upper bounds on the total number of hours each machine center is operated, and the number of man hours utilized.

III.2.2 Planned Lead Times

In job shops, particularly in closed job shops, it is not uncommon to assign a lead time for each product. The lead time is the difference between the time an order is received by the shop and the delivery date. These lead times are determined periodically (perhaps once every quarter) taking into account external factors such as the market conditions, actions of competitors, total demand faced by the facility etc. The lead times may be announced in the sales catalogues and serve as guidelines for the sales force.

Managers can benefit from models that translate the lead time prescriptions into capacity requirements. Problem TP1.1 is an example of an important decision aid in determining the appropriate capacities. A closely related problem is TP2.1

TP2.1 Optimal lead time:

Objective: Minimize the lead time for a specific product family

Decision Variables: Number of hours each machine center is operated

Constraints: Upper bound on the lead times for each product

III.2.3 Lot Sizes

In batch shops that produce to stock, a work order is generated whenever the finished goods inventory falls below a critical level. The finished goods inventory serves as a filter that smooths out some of the variability present in the external environment. By aggregating demand and producing in larger lots the time lost in setting up the machines is also minimized. Thus in addition to reducing the variability observed by the manufacturing systems, finished goods inventory can also increase the throughput of the system. In this environment the number of units produced per order (the lot size) is an important managerial decision, that effects both the lead times and the throughput. Observe that as the lead times increase the finished goods inventory levels will also increase. Models that quantify the relationship between lot sizes, lead times and throughput are very valuable (Karmarkar 87, and Karmarkar, Kekre and Kekre 85). An example of a problem belonging to this class is:

TP3.1 Optimal lot size:

Objective: Minimize the work-in-process inventories

Decision Variables: Lot size for each product

Constraints: Upper bound on the lead times for each product

III.3 OPERATIONAL DECISIONS

After determining the system design and making aggregate allocations of the resources, it is necessary to manage the day-to-day operations of the facility. The operational decisions, belonging to the lowest level of the hierarchy, must take into account the decisions made at the higher levels, and the current status of the facility. At this level there is very little opportunity to either increase the available productive resources, or change the mix of jobs to be processed during a time period. Typical decisions made at this level are:

(i) the sequence of machine visitations for each job , that is the route the job will follow through the shop.

(ii) the sequence in which the jobs are processed at each machine center,

(iii) tracking of jobs through the facility, expediting, and releasing jobs into the facility.

Thus at the operational level detailed sequencing and scheduling decisions are to be made so that orders are completed on time utilizing the resources available. The class of detailed scheduling and sequencing problems that arise in general job shops are extremely difficult to solve exactly (Rinnooy Kan 76). Hence several approximation procedures have been proposed. Readers are referred to Graves 81, Conway Maxwell and Miller 67, and references there in for details of

these scheduling rules. The bulk of the scheduling literature assumes that the production environment is deterministic and static over a specified finite horizon. These environments have been called static job shops (Graves 81). In the context of dynamic job shops, the environment we are concerned with, the focus of researchers has been on dispatch rules. Whenever a machine becomes available, the dispatch rule determines the job that is to be processed next. Examples of dispatch rules are:

(i) shortest processing time,

(ii) least work remaining,

(iii) first to arrive to the shop,

(iv) earliest due date, and

(v) first to arrive at the machine center.

For further details of the dispatching rules the reader is referred to Conway Maxwell and Miller (67), Panwalker and Iskander (77), and the references therein.

IV OPEN QUEUEING NETWORK MODELS OF JOB SHOPS - PERFORMANCE EVALUATION

IV.1 EXACT ANALYSIS

The study of networks of queues was initially motivated by applications in the telephone industry (Erlang 17). However the pursuit of these problems received a significant boost from two seminal papers by J.R. Jackson (57, 63). Jackson's work interestingly, was motivated by job shops. Since the publications of those papers significant theoretical insights have been gained into the properties of queueing networks. In particular, in the last two decades spurred by applications in

computer science, telecommunications, and flexible manufacturing systems, there has been a flurry of activity in this area. Unfortunately exact analysis for open queueing networks with finite number of servers has been possible only for networks that have the following characteristics:

(1) Exponential service time distributions.

(2) Service requirement at each station are independent of the product family. If the service times are allowed to depend on the product family then exact analysis is possible with a preemptive resume, last come first served discipline.

(3) Priority discipline at each queue is independent of the product family.

(4) Arrival process to the network is a Poisson process.

These Markovian systems are also known as reversible networks (Kelly 75, 79). An important property of these networks is the product form of the equilibrium distribution of the number in the system. Consider a network consisting of M stations (machine centers), and let $(C_i)_{i=1,M}$ denote the state of the queue at machine center i. C_i is a vector with elements $(C_i(1),C_i(2), C_i(3),.. C_i(L_i))$, where L_i is the queue length at station i, and $C_i(k)$ specifies the product family of the job in the kth position in the queue.

Let P(C) be the equilibrium probability of observing the network in state C; $C = (C_1,C_2 ,C_3, ..C_M)$. For reversible networks if the equilibrium distribution exists then it is of the following form:

$$P(C) = K\, G(C)\, A_1(C_1)\, A_2(C_2)\, A_3(C_3)... A_M(C_M)$$

where G(.) is a function that depends on the state vector, $A_i(.)$ is a function that depends on the nature of machine center i, $A_i(.)$ is proportional to the equilibrium probability distribution at machine center i with Poisson arrivals, and K is a normalizing constant. For further details regarding these networks the readers are referred to Lemoine (77), Disney and Konig (85), and Buzacott and Yao (86a).

IV.2 APPROXIMATION PROCEDURES

For general job shops the assumptions underlying reversible networks are very restrictive. For instance work by Bitran and Tirupati(88) suggests that the exponential distributions overstate the variability in the service times found in many manufacturing operations, and distributions with squared coefficient of variation (scv) less than one are more appropriate. Since exact analysis has not been possible when the assumption of exponential service times is relaxed, the focus has been on approximation procedures. The approximation schemes can be classified into four categories: decomposition methods, diffusion approximations, mean value analysis, and operational analysis. The procedure that has been employed with considerable success to analyze models of manufacturing systems is the decomposition approach (eg Shanthikumar and Buzacott 81, Bitran and Tirupati 88, Segal and Whitt 88). Only recently diffusion models have been utilized to study scheduling and operational control problems arising in manufacturing. Operational analysis (eg Denning and Buzen 78) has been applied primarily to computer system models, and mean value analysis (eg. Reiser and Lavenberg 80, Sevick and Mitrani 81, and Seidmann et al 87) is concerned with closed queueing networks. Hence we will restrict our attention to the decomposition methods, and applications of diffusion models to manufacturing problems(Harrison and Wein 90a,b, and Wein 90b).

IV.2.1 Decomposition Methods

The decomposition methods are in part motivated by the properties of Jackson networks, and can be viewed as attempts to extend the product form solution to more general networks. In Jackson networks if the arrival rate is a constant that does not depend on the number in the system,

then the equilibrium distribution of the number in the network can be obtained by analyzing each machine center as a M/M/c queue (Jackson 63). Under the decomposition approach this result is mimicked, and each node in the queueing network is approximated by a G/G/c system. There are three basic steps in the decomposition methods :

(1) Characterization of the Arrival Process: At each station the arrival process resulting from the superposition of different streams arriving to that station is (approximately) determined.

(2) Analysis of the queue : Based on the characteristics of the arrival process determined in step 1, the queueing effects at the station are (approximately) computed.

(3) Determination of the departure process : The characteristics of the departure process of each product from the station are (approximately) determined. The departure streams in turn become arrivals at some other stations.

Several variants of the decomposition method can be developed by varying the implementation of the three steps. One of the most often used procedures is the parametric decomposition approach.

The Parametric Decomposition Approach (PDA)

Under the parametric decomposition approach (PDA), in addition to assuming that each node can be treated as being stochastically independent (the decomposition assumption), the arrival process to, the departure process from, and the flow between each node are approximated by renewal processes. Further, it is assumed that two parameters - mean and variance - of the inter-arrival, and service time distributions are adequate to estimate the performance measures at each node. Hence to compute the performance measures we need to (i) approximate all the flows in the

network, and (ii) compute the performance measures based on the first two moments of the interarrival and service times. Accordingly, the description of the PDA will be in two parts : A) Flow analysis, and B) Estimation of performance measures.

The decomposition approach was first proposed by Reiser and Kobayashi (74), and has subsequently been modified and developed by Sevick et al(77), Chandy and Sauer (70), Kuehn(79), Shanthikumar and Buzacott(81), Buzacott and Shanthikumar(85), Whitt(83a), and Bitran and Tirupati(88, 89a, 90). For ease of exposition we first describe the main steps in the parametric approach assuming that there is only one product family and the routing structure is Markovian. Also, unless we state otherwise, a FCFS queue discipline is assumed. Using this as a basis we discuss the assumptions underlying this approach in greater detail, and then briefly describe how it has been augmented to incorporate features such as multiple products, and deterministic routings.

A. Flow Analysis

Let r_{ij} be the probability of a job going to station j, upon completion of service at station i, and let R be an M x M matrix with elements $\{r_{ij}\}$. Because each job eventually leaves the network the matrix $[I - R]^{-1}$ is well defined, where I is the identity matrix. For external arrivals to the network we let i =0, and for departures from the network we let j = 0.

For the renewal process approximating the flow from station i to station j, let l_{ij}, and c_{ij} be the mean, and the squared coefficient of variation (scv) of the renewal interval length. Denote the flow rate from node i to j by a_{ij} ; $a_{ij} = 1/l_{ij}$. In PDA the superposition of the flows arriving at a node are further approximated by a renewal process. We let $a_{.i}$, $a_{i.}$ denote the total flow rate into

and out of node i respectively. Similarly define $c_{.i}$, and $c_{i.}$. The flow rates a_{ij} are determined by the following traffic equations:

$$a_{.i} = a_{0i} + \sum_{j=1}^{m} a_{.j} r_{ji} \qquad (1)$$

$$a_{ji} = a_{.j} r_{ji} \qquad (2)$$

While determining the flow rates is straightforward, approximations are needed for the scvs. In particular we need procedures for approximating by a renewal process each of the following: (i) superposition of renewal processes, (ii) departure processes from queues, and (iii) flow along each arc out of a node (splitting the departure stream).

(i) Approximations for Superposition of Renewal Processes:

In PDA only the mean and the variance of the approximating renewal interval need to be determined. The mean is straight forward to compute - the arrival rate of the approximating process must equal the arrival rate of the superposition process. Whitt(82) considers two basic procedures for determining the variance of the approximating process. He calls them micro and macro approaches.

Assume that the superposition process has been on since $t = -\infty$, and an arrival occurs at time 0. Let S_n be the time of the n^{th} arrival after time 0, and $V(S_n)$ the variance of the random variable S_n. Under the macro approach the variance of the approximating renewal interval is set

at Lim $_{n\to\infty}$ $V(S_n)$ /n . The macro approach is also called the asymptotic method. Henceforth we refer to the limiting variance and scv as the asymptotic variance and asymptotic scv.

Under the micro approach the variance of the approximating renewal interval is set at $V(S_1)$. The time interval starting from 0 until the first arrival after 0 is referred to as the stationary interval of the superposition process. Henceforth we refer to $V(S_1)$ as the stationary interval variance and the corresponding scv as the stationary interval scv.

The asymptotic scv can be computed readily from the scvs of the interarrival times of each of the process being merged. For instance the asymptotic scv of the arrivals to station i is given by:

$$\frac{1}{a_{.i}} \sum_{j=0}^{M} a_{ji} c_{ji} \qquad (3)$$

Although the asymptotic scv is easy to compute, the stationary interval scv is cumbersome to determine. As a result Whitt(82) proposes approximation formulae for the stationary interval scv. These formulae are based on the characteristics of hyper-exponential, Erlang, and shifted exponential distributions.

When the two approaches, micro and macro, were used to estimate performance measures of queueing systems, Whitt(82) and Albin(81, 84) found that neither method dominated. Based on their experiments they discovered that a convex combination of the scvs provided by the micro and macro approaches yielded the best results. This approach has been called the hybrid approach (Albin 84). If let c_a, c_s, and c_h denote the asymptotic, stationary, and hybrid scv, respectively, then $c_h = Wc_a + (1-W)c_s$, where $0 \le W \le 1$, and W is a function of the utilization of the server and the

number of arrival streams being merged. As the number of arrival processes being merged goes to infinity, the stationary interval is asymptotically correct. On the other hand as the utilization goes to 1, the asymptotic limit is asymptotically correct. The weighting factor W is so chosen that as the number of process being merged goes to infinity, W goes to zero, and as the utilization goes to 1, W goes to 1.

In the queueing network analyzer proposed by Whitt (83a) the following approximation for the scv of the arrivals at station i is used:

$$c_{.i} = W\sum_{j=0}^{M}\left[\frac{a_{ji}}{a_{.i}}\right] c_{ji} + 1 - W \qquad (4)$$

$$\text{where } W^{-1} = [1 + 4(1 - \rho_i)^2(V - 1)] \qquad (5)$$

$$V^{-1} = \sum_{j=1}^{M} \frac{a_{ji}^2}{a_{.i}^2} \qquad (6)$$

ρ_i : utilization of station i

For further details the reader is referred to Whitt (79, 82, 83a) and Albin (81, 84).

(ii) Approximations for the departure process:

The departure process from a queue is in general not a renewal process(eg Berman and

Westcott 83), however in PDA it is approximated by a renewal process. The mean of the approximating renewal interval is easy to determine. Two alternatives have been considered for the variance - the stationary departure interval variance and the asymptotic limit. Whitt (84) shows that for GI/G/c queues with utilization less than 1, the asymptotic variance of the departure process is the same as the variance of the interarrival times. Hence, once again the asymptotic limit is easy to determine. However the computational tests indicated that the stationary interval provides a better approximation, and that was adopted by Whitt(83). Unfortunately, determining the stationary interval distribution of the departure stream is not easy, and instead of computing the exact stationary departure interval scv, approximations are employed. Combining the formulae for the stationary interval due to Marshall(68) with Kraemer-Langenbach-Belz(76) approximation for the expected waiting time , Whitt (83) obtains the following approximation formulae for the scv of the inter-departure times:

$$c_{i.} = \rho_i^2 cs_i + (1 - \rho_i^2) c_{.i} \qquad (7)$$

where cs_i : the *scv* of the service time at station i

(iii) Approximations for flow along each arc (splitting):

If the routing is Markovian, and the departures from the station are approximated by a renewal process, the flow along each arc will be a renewal process (eg Disney and Konig 85 Theorem 3.1). Under these assumptions, the interdeparture time along each arc out of the station will be the random sum of interdeparture times from the station. The number of interdeparture times (from the station) that have to be convoluted is of course geometrically distributed. Hence the scvs for the flows along each arc can readily be expressed in terms of the scvs of the departure

process from the source station and the routing probabilities:

$$c_{ij} = c_i r_{ij} + 1 - r_{ij} \qquad i,j = 1 \text{ to } M \tag{8}$$

Putting together equations 4,7,and 8, we get the following system of equations for the scvs.

$$a_i c_i - \sum_{j=1}^{M} [a_j(1 - \rho_j^2) r_{ji}^2 c_j] - a_{0i} c_{0i} + \sum_{j=1}^{M} [a_j r_{ji}(\rho_j^2 r_{ji}(cs_j + 1 - r_{ji}))] \qquad i = 1 \text{ to } M \tag{9}$$

Observe that once the a_is are determined from equations 1, and 2 the system of equations (9) is linear in c_i. Since the performance of the queue is estimated on the basis of the first two moments of the inter-arrival and service times, the necessary flow parameters have all been determined.

B. Estimation of Performance Measures

The performance measures at each station are estimated using approximation formulae that are based on the first two moments of the inter-arrival and service times. A wide variety of approximations have been proposed for the analysis of GI/G/c queues (eg. Lindley 52, Page 72, Cosmetatos 75, Kraemer and Langenbach-Belz 76, Marchal 76, Boxma et al 79, and Whitt 85a). We will not review this literature but only identify approximations that have been employed in the parametric decomposition framework.

For the single server queues an approximation formula for the average queue length due

to Kraemer Langenbach-Belz(76) has been extensively used. For estimating the average queue length at a single server station, Shanthikumar and Buzacott(80) tested several formulae, and for different ranges of the scv of the inter-arrival and service times they recommend a different formula. Their recommendations and the corresponding formulae are given in appendix 1.

Whitt(85) proposes approximations for a variety of performance measures in GI/G/c queues. These formulae are based on the behavior of M/M/c , D/M/c and M/D/c queues, heavy traffic approximations for GI/G/c queues, and computational experiments. Although the performance measures considered by Whitt(85a) include the second moment of the queue length, the probability of delay, and the waiting time and queue length distributions, in the context of queueing networks only the estimates for the average queue length have been extensively tested. We present Whitt's formulae for the average queue length in appendix 1.

These observations complete the discussion of the basic elements of the PDA. With in this framework several features relevant to manufacturing system such as deterministic routing for multiproduct networks (Bitran and Tirupati 88), batch machines (Bitran and Tirupati 89c), overtime (Bitran and Tirupati 90), inspection and testing (Segal and Whitt 88), and machine break-down (Segal and Whitt 88, Bugalak and Sanders 89) have been incorporated.

For networks with multiple products and deterministic routing, Bitran and Tirupati(88) modified the procedure for splitting the output from single server stations. For ease of exposition let us assume that upon service completion each product family flows along a different arc. As in the case of Markovian routing, here too the output process from the station and the flow along each arc are approximated by renewal processes. However the procedure for determining the scv of the

flow along each arc is modified. Let us assume that we are interested in the flow of product family k. Henceforth we refer to this product family as the marked family, and all other families arriving at the station are referred to as the aggregate family. Bitran and Tirupati first approximate the arrival process of the aggregate family by a Poisson process and determine the distribution of the number of aggregate arrivals between two marked arrivals. Observe that under the assumption that the aggregate family arrival process is a Poisson process, the number of aggregate arrivals between consecutive marked arrivals are independent and identically distributed. Therefore if the output from the station is assumed to be a renewal process, then the output of the marked product family will also be a renewal process. The interdeparture time of the marked product family is the random sum of the interdeparture times from the station, where the distribution of the number of interdeparture times to be summed is given by the number of aggregate arrivals between marked arrivals. Under these assumptions the scv of the flow along arc ik is given by :

$$c_{iK} = f_K c_{i.} + (1 - f_K)[f_K + (1 - f_K)ca_K]$$

$$\text{where } f_K = \frac{ap_K}{a_{.i}} \qquad (10)$$

ap_K = arrival rate of product K

ca_K = *scv* of the interarrival times of product K

When equations 8 and 10 were tested, equation 10 enabled significantly better estimates of the performance measures. There is also an interesting qualitative difference between equation 8 and 10. In equation 8 observe that as r_{ij} tends to zero the scv of the flow along arc ij approaches 1. On the other hand in equation 10 as the fraction of product k (f_k) arriving at node i goes to zero the scv of the flow along arc k (c_{ik})approaches the scv of the arrival process of product k (ca_k). In fact Bitran and Tirupati show that equation 10 is asymptotically exact. Whitt(88) has further

generalized this result to show that as the proportion of product family k goes to zero, not only does the scv of the departure process for product k approach its arrival scv, but the interdeparture times of product family k become independent and identically distributed, with the same distribution as the interarrival times of product k. Whitt(88) conjectures that this is the case regardless of the number of servers, and priority discipline.

Bitran and Tirupati(88) also develop approximations for the departure process of each family under the assumption that the interarrival times at station i of the aggregate family have an Erlang distribution. In this case the number of aggregate arrivals between marked arrivals are no longer independent. However they assume these random numbers to be independent and identically distributed, and provide two procedures for computing the distribution of the number of aggregate products interfering between consecutive marked arrivals. For further details the readers is referred to Bitran and Tirupati(88).

For networks with batch machines employing ideas similar to those described above, Bitran and Tirupati(89c), develop an approximation for the number of jobs of product family k in each batch. This in turns enables them to approximate the flow process along each arc out of the batch station. Observe that stations down-stream from the batch station will observe bulk arrivals, and the distribution of the number of jobs arriving together will depend on the composition of the batch.

Features such as overtime and machine break-downs have been incorporated primarily by modifying the service times. Appendix 2 illustrates one procedure that incorporates break-downs by modifying the service time distribution. For further details the readers are referred to Whitt and Segal(88), Bitran and Tirupati(90), Bugalak and Sanders(90).

It is clear from the discussion above that the parametric decomposition approach involves several layers of approximations. To start with, each flow process is approximated by a renewal process, next the parameters needed for approximating the flows (such as the stationary interval scv of the superposition process) are computed approximately, and finally the performance measures are estimated using approximation formulae. Inspite of these simplifications the procedure provides remarkably accurate estimates of the first moment of the queue length (eg Whitt 83b, Shanthikumar and Buzacott 81 , Bitran and Tirupati 88, 89c, 90) in very general queueing networks. The parametric approach is particularly appealing because its data and computational requirements are minimal. It only requires the first two moments of the inter-arrival and service times, and the routing matrix. Computations essentially involve solving two systems of linear equations each with M constraints.

Given the success of the PDA in estimating the first moment of the queue length, the idea of decomposing the network into a system of G/G/c queues is very attractive. In order to refine and enhance the decomposition procedure so as to obtain other performance measures such as higher moments of the queue lengths, and waiting times, further study of the three basic decomposition steps is needed. Recall that the 3 basic steps are: (1) Characterization of the Arrival Process, (2) Analysis of the queue, and (3) Determination of the departure process.

In this context the developments in the study of phase type distributions(Ph) and queues with phase type arrival and/or service processes are interesting. Phase type distributions permit detailed analysis of complex point processes, and queues with non-renewal arrivals. For instance, for the superposition of phase renewal processes several parameters such as the stationary interval

moments, and the lag correlations between interarrival times can be readily computed (eg Rudemo 73, Neuts 79, and Bitran and Dasu 90a). In general these parameters are not easy to compute. It is useful to note that phase type distributions are a dense sub-set of all distributions on non-negative real numbers. Exponential, Erlang and Hyper-exponential distributions, whose properties have been exploited in the development of PDA, are special cases of Phase type distributions.

Employing essentially matrix geometric procedures (Neuts 81) detailed analysis of queues with superposition arrivals can be carried out, if the inter-arrival times for each of the streams being merged has a phase type distribution. (Ramaswami 80, Bitran and Dasu 90b). The analysis of these queues determines performance measures as observed by each customer class, and characteristics of the departure process such as the stationary interval distribution and the lag correlations. Matrix geometric techniques can be employed to analyze queues with correlated arrivals, provided the arrival instances can be depicted as transition times in a finite irreducible continuous time Markov process.

Bitran and Dasu(90a) use the term generalized phase process (GPh) for point processes generated by transitions on a subset of arcs in finite irreducible continuous time Markov chains. Special case of GPhs are phase renewal processes, alternating phase renewal processes, and superposition of phase renewal processes ($\mathfrak{s}Ph_i$). Superposition of GPhs ($\mathfrak{s}GPh_i$) is also a GPh. GPhs in turn are special cases of the N -process (N-p) identified by Neuts (79).

Since queues with correlated arrivals (GPh or N-p type) can be analyzed, it is not necessary to approximate either the departure or the superposition arrival process by a renewal process. Nevertheless approximations will be needed if either (i) the actual process is not a GPh (N-p), or

(ii) the size of the exact representation is too large and is computationally prohibitive - the size of a GPh is the number of states of the underlying Markov process. The departure process from a GPh/G/1 queue is an example of the former case, and the departure process from a GPh/Ph/1 is an example of the latter.

Bitran and Dasu(90) study the problem of approximating ΣPh_i by a GPh that is smaller in size. (As the number of process superposed increases the size of the exact representation grows very rapidly.) They propose an approximation scheme that takes into account only the asymptotic and stationary interval moments of the ΣPh_i process. To evaluate the procedure they compare the performance measures of ΣPh_i /$E_2(M)/1$ with that of the corresponding GPh/$E_2(M)/1$ queue. Their limited computational tests are very encouraging in that the error in the first three moments of the queue length are less than 5%, and the error in the scv of the stationary departure interval is less than 10% .

Although matrix geometric techniques and phase type distributions permit analysis of fairly complex queues and point processes, a drawback of the approach of Bitran and Dasu(90a, b) is that it is computationally very expensive as compared to PDA. Therefore, it may not be suitable for analyzing large networks. Nevertheless, the developments in this domain are useful for developing and evaluating approximations needed to enhance the decomposition methods.

Recently, Bertsimas and Nakazato(90) have derived the exact characteristics of the departure process from GI/G/1 queues. Using Hilbert factorization they have determined the transforms of the stationary interdeparture interval and the lag-correlations. These developments should prove useful for further refining approximations for the departure processes.

V OPEN QUEUEING NETWORK MODELS OF JOB SHOPS - OPTIMIZATION MODELS

So far the discussion has been about how to evaluate the performance of a given queueing system. Since we are concerned with models that aid the (re)design of manufacturing systems we next look at procedures that determine the configuration of a queueing network that achieves a particular performance objective. Clearly, if the system design problem is one of selecting from a limited number of alternative system configurations (designs), then performance evaluation models will be adequate. Otherwise, more sophisticated optimization routines become necessary.

As stated several times in earlier sections, the actual performance of a system is inextricably linked to the control rules employed to operate the system. It is therefore tempting to require algorithms that determine optimal system configuration to simultaneously determine the optimal control policies. Such a monolithic approach has so far eluded analysis. Considerations such as the detailed data requirements of operational control rules also diminishes the practical appeal of the monolithic approach.

The optimization procedures developed for open queueing networks can be partitioned into two categories : those that design the network assuming simple operational rules such as FCFS, and those that determine the optimal operational rules for an existing system. Given the bias of this paper our focus will be on the first set of models. In recent years based on the developments in the theory of dynamic control for Brownian networks, researchers have proposed control rules for job shops. The second part of this section contains a brief review of this literature.

V.1 OPTIMAL DESIGN OF NETWORKS

Optimal design of computer systems and communication networks based on queueing network models have been studied for several years (Kleinrock 76 chapter 5, Chandy et al 77, Tantawi and Towsley 85). Emergence of flexible manufacturing systems has also motivated developments in design of queueing networks (Buzacott and Yao 86b). Much of this work is concerned with closed queueing networks. It is only recently that open queueing network models have been proposed for designing job shops (Bitran and Tirupati 89a,b).

In section 2 we identified three classes of problems related to the design of manufacturing systems - SP1, SP2, SP3. We now discuss developments in queueing networks that address similar problems. The earliest result in design of open queueing networks is due to Kleinrock(64). This problem is similar to problem SP2.1. Hence we begin by describing procedures developed for SP2.1, which can be restated as follows:

Q2.

$$\min \sum_{i=1}^{M} L_i(m_1,n_1;m_2,n_2;\ldots m_M,n_M)$$

$$\sum_{i=1}^{M} g_i(m_i,n_i) = D$$

$$[m_i,n_i] \in X_i \quad i = 1 \text{ to } M$$

where :
n_i : number of machines at station i
m_i : processing rate of each machine at station i
$g_i(m_i,n_i)$ - cost of equipping station i with capabilities (m_i,n_i)
$L_i(m_i,\ldots)$: Expected numbers of jobs at station i as a function of the processing capabilities

We refer to this problem as Q2, to emphasis that it is based on a queueing network model of a manufacturing facility. (Problem Q1 which corresponds to the first class of strategic problems is formulated later in the paper.) The layout of the network, arrivals rates for different product families, and the routing matrix are predetermined, and are not decision variables in Q2. The decision variables are the processing capabilities of each station which are determined by the number of servers (n_i)s, and the processing rate of each server at a station (m_i)s. Observe that m_i should be a vector with n_i elements, however we assume that all servers are identical. Hence m_i is a scalar.

In section 2 we identified two basic procedures by which capacity is altered in manufacturing systems - by purchasing machines, or by increasing the number of hours the facility is operated each day. In queueing models capacity is primarily altered by either increasing the number of servers or changing the processing rate. Adding servers at a station is clearly equivalent to acquiring machines. However the correspondence between increasing the processing rate and adding extra shifts is loose. If the entire facility is operated for an extra shift, then it has the effect of increasing the number of jobs processed each day at a station. But if only some machine centers are operated for extra time then it is not equivalent to increasing the processing rate of the corresponding station in the queueing network model of the facility. However, Bitran and Tirupati(90) show that the effect of operating a machine center extra hours can be closely approximated by altering the processing rate at that station. Queueing models in which capacity is altered (continuously) by changing the processing rate have also been justified as being approximations for the discrete process of (a) determining the capability of the machine to be acquired or (b) adding more machines. In some situations such as metal cutting operations, the processing rate can be altered

almost continuously by changing the cutting speeds.

Several special cases of Q2 have been analyzed. We develop next a notation to identify these models. The problems will be denoted by Q2.TN.D.

- TN is the type of queueing network. For our purpose J (G) denotes a Jackson (General) network, and S (M) will be used in conjunction with J (G) to denote that all stations have single (multiple) servers. For example JS denotes a Jackson network with only single server stations, and GM denotes a general open queueing network with multiple servers at one or more stations.
- D denotes the decision variable. There are two possible values for D, R and N. R (N) denotes problems where the service rates (number of servers) are the decision variables.

Kleinrock(64) considered the problem of determining the processing rate (m_i) at each node in an open Jackson network so as to minimize the expected number of jobs at each station. He assumed that all stations in the network are single server stations and that the cost of a machine is proportional to the processing rate. We denote this problem as Q2.JS.R.

Q2.JS.R

$$\min \sum_{i=1}^{M} L_i(m_i)$$

$$L_i(m_i) = \frac{a_{.i}}{(m_i - a_{.i})} \quad i = 1 \text{ to } M.$$

$$\sum d_i m_i = D$$

$$m_i \geq 0$$

$\sum d_i m_i = D$ is the Budget constraint, where d_i is the unit cost of capacity at station i. The optimal rates for this problem are:

$$m_i^* = a_i + \frac{\sqrt{d_i a_i}}{\sum_{j=1}^{M} \sqrt{d_j a_i}} \frac{(D - \sum_{j=1}^{M} d_j a_j)}{d_i} \qquad i = 1,M \qquad (11)$$

If the unit cost of capacity is the same at each station the solution above is equivalent to first allocating just enough capacity to each station to satisfy the arrival rate, and then allocating the excess capacity among stations in proportion to the square roots of their arrival rates. Observe that the optimal allocations do not result in a system where all the server utilizations are equal.

Problem Q2.JS.R has an elegant solution because : (C1) the average number in the system is a convex function of the processing rate, (C2) capacity addition at one station has no effect on the expected queue length at other stations, (C3) processing rates (decision variables) are continuous variables, (C4) Cost of the machines are convex (linear) functions of the rates, and (C5) the sojourn time at each station can be expressed in closed form. Conditions C1 - C4 permit marginal analysis, and C5 enables a closed form solution.

Q2.JM.R is the equivalent of Q2.JS.R obtained by permitting multiple server stations in the network. Note that the decision variables in Q2.JM.R are the processing rates and not the number of servers at each station. Harel and Zipkin (87) have established that in an M/M/m queue the

expected sojourn time is a convex function of the service rate. As a result Q2.JM.R enjoys properties C1 through C4; hence it is a convex program that can be solved via marginal analysis (eg. Avriel 76, and Bertsekas 82).

Q2.JM.N is a variant of Q2.JM.R in which the decision variables are the number of servers (n_i) at each station. Hence the budget constraint has to be modified to $\sum d_i n_i = D$. In this case condition C2 continues to be valid, and condition C1 is valid because the average sojourn time in an M/M/c queue is a convex function of c (Dyer and Proll 77). However, since the decision variables are no longer continuous, marginal analysis will not yield the optimal solution. Q2.JM.N has the structure of the Knapsack problem and can be solved by any algorithm proposed for such problems (Nemhauser and Wolsey 88)

In general networks conditions C1 and C2 do not apply and we are left with difficult non-linear optimization problems. As a result approximations have been proposed. In all the approximation procedures that are described in this section (for problems Q1 defined below, and Q2), the non-convex program is approximated by a convex program. In order to do this, the equilibrium queue length distributions are first approximated by a product form distribution. As a result of this approximation condition C2 holds. Next the approximation formulae for the expected queue lengths are shown to be convex functions of the processing rate (the decision variable). Consequently all 5 conditions are met by the approximating convex program.

Harrison and Williams (87) have shown that in networks in which the utilization of each station is between 0.9 and 1.0, the stationary distribution of the number in queue can be

approximated by a product form distribution. Under this approximation scheme the expected number of customers at station i is given by :

$$L_i = \frac{\sigma_i^2}{2(m_i - a_i)} \qquad (12)$$

$$\text{where, } \sigma_i^2 = a_{0i}c_{0i} + a_i cs_i + \sum_{j=1}^{M} a_j f_{ji}(cs_j f_{ji} + 1 - r_{ji}) \qquad (13)$$

Observe that the formula for the average number at the station is convex in m_i, and capacity addition at one station has no effect on the average queue length at other stations. Wein(90) employs these heavy traffic approximations to develop a solution for Q2.GS.R. Under this approximation scheme all 5 conditions are valid and the following solution is readily obtained :

$$m_i^* = a_i + \frac{\sqrt{d_i\sigma_i^2}}{\sum_{j=1}^{M}\sqrt{d_j\sigma_j^2}} \frac{(D - \sum_{j=1}^{M} d_j a_j)}{d_i} \quad \text{for } i = 1,M \qquad (14)$$

The allocations made under this procedure (eq 14) reduces to that due to Kleinrock (eq 11) if the network is a Jackson network. Observe that equation 14 compensates for the variability (σ_i) at a station.

Let us now turn to the first class of strategic problems identified in section 2. Recall that problem class SP1 deals with designing a least cost facility to achieve specified performance characteristics. Queueing problem Q1, which corresponds to problem SP1.1 is formulated below.

The objective of SP1.1 is to design a least cost facility, given an upper bound on the expected number of jobs in the system.

Q1.

$$\min \sum g_i(m_i,n_i)$$

$$\sum_{i=1}^{M} L_i(m_1,n_1;m_2,n_2;...m_M,n_M) \leq U$$

$$(m_i,n_i) \in X_i \qquad i = 1,M$$

where U: is an upper bound on expected numbers of jobs in the system

Below we formulate a special case of Q1 for Jackson networks with single server stations, with processing rates as the decision variables. We continue to use the notation described above, so we denote this problem as Q1.JS.R :

Q1.JS.R.

$$\min \sum_{i=1}^{M} d_i m_i$$

$$\sum v_i L_i \leq u$$

$$L_i = \frac{a_{.i}}{m_i - a_{.i}} \qquad i = 1,M$$

$$m_i \geq 0 \qquad i = 1,M$$

where v_i : weighting factor

The objective function, and L_i are convex in m_i. Observe that conditions C1 - C4 hold, therefore Q1.JS.R is a convex program and can be solved through marginal analysis. For the same

reasons, Q1.JM.R is also a convex program. In general conditions C1 - C2 do not hold. Therefore approximations are needed for Q1.GS.R, Q1.GM.R, and Q1.GM.N.

Bitran and Tirupati (89b) employ the parametric decomposition scheme to provide an approximate solution for Q2.GS.R. Recall that in PDA each node is approximated by a GI/G/c queue, and the queue length at each station is estimated based on the two moments of the inter-arrival and service times. Bitran and Tirupati first assume that altering the processing rate (m_i) does not effect the scv of the service process. This is equivalent to assuming that the machine operates faster.

For networks which process a large number of products, Bitran and Tirupati provide experimental evidence that the scvs of the departures from a station (c_i)changes very little even if the processing rate at the station is changed. Hence they assume that the scv of the departure process is unaffected by the processing rate. This assumption enables condition C1.

Next, they show that the modified Kraemer and Langenbach-Belz formula (given below) employed for estimating the expected number at each station is convex in the processing rate.

$$L_i(KLB) = \rho_i + \frac{\rho_i^2}{2(1-\rho_i)}(c_{.i} + cs_i)\, g(\rho_i, c_{.i}, cs_i)$$

$$g(\rho_i, c_{.i}, cs_i) = \left\{\exp \frac{-2(1-c_{.i})}{3(c_{.i}+cs_i)} \frac{(1-\rho_i)}{\rho_i}\right\} \text{ if } c_{.i} \le 1 \qquad (15)$$

$$= 1 \text{ otherwise}$$

Thus Q2.GS.R is approximated by the following convex program:

Q2.GS.R*

$$\min \sum_{i=1}^{M} d_i m_i$$

$$\sum_{i=1} v_i L_i(KLB) \leq U$$

$$m_i \geq 0$$

Although Bitran and Tirupati(89b) assume that the scv of the flows between stations are unaffected by changes in processing rates, in order to solve Q2.GS.R*, estimates of $c_{.i}$ are needed. Bitran and Tirupati assume that a facility is being redesigned and employ the existing utilizations to compute the $c_{.i}$s. Bitran and Sarkar(90) refine this procedure by relaxing the assumption of unchanging $c_{.i}$s in the following manner: (i) As in Bitran and Tirupati (89b), they assume a facility is being redesigned, and hence initially compute the $c_{.i}$ values based on the existing utilizations; (ii) They then solve Q2.GS.R* assuming that $c_{.i}$s are unaffected by changes in processing rates; (iii) using the new processing rates obtained in step ii they recompute values of $c_{.i}$s (i.e, they update the scv values to take into account the modified processing rates); and (iv) with the modified scv values they repeat step ii. This process is repeated until the procedure converges. If the scvs of the arrivals at each station are allowed to vary, then Q2.GS.R* is no longer a convex program, hence even if Bitran and Sarkar's procedure converges it need not converge to the optimal solution. In fact the procedure need not converge at all. However Bitran and Sarkar provide conditions on the network data, under which their procedure is guaranteed to converge.

In a paper that closely parallels that of Bitran and Tirupati (89b), Boxma, Rinnooy Kan and

van Vliet (90) provide an approximate solution for Q1.GM.N. Once again Q1.GM.N is approximated by a convex program by assuming that conditions C1 and C2 hold. It is interesting to note that although in a GI/G/c system the expected queue length is not a convex function of the processing rate, the approximation formula due to Kraemer Langenbach-Belz is convex in the processing rates.

The work of Bitran and Tirupati (89a,b), Wein (90a), and Boxma, Rinnooy Kan and van Vliet(90) suggests that fairly general queueing networks behave as if conditions C1 and C2 are valid. This is very encouraging because it enables us to approximate by convex programs, optimization problems where

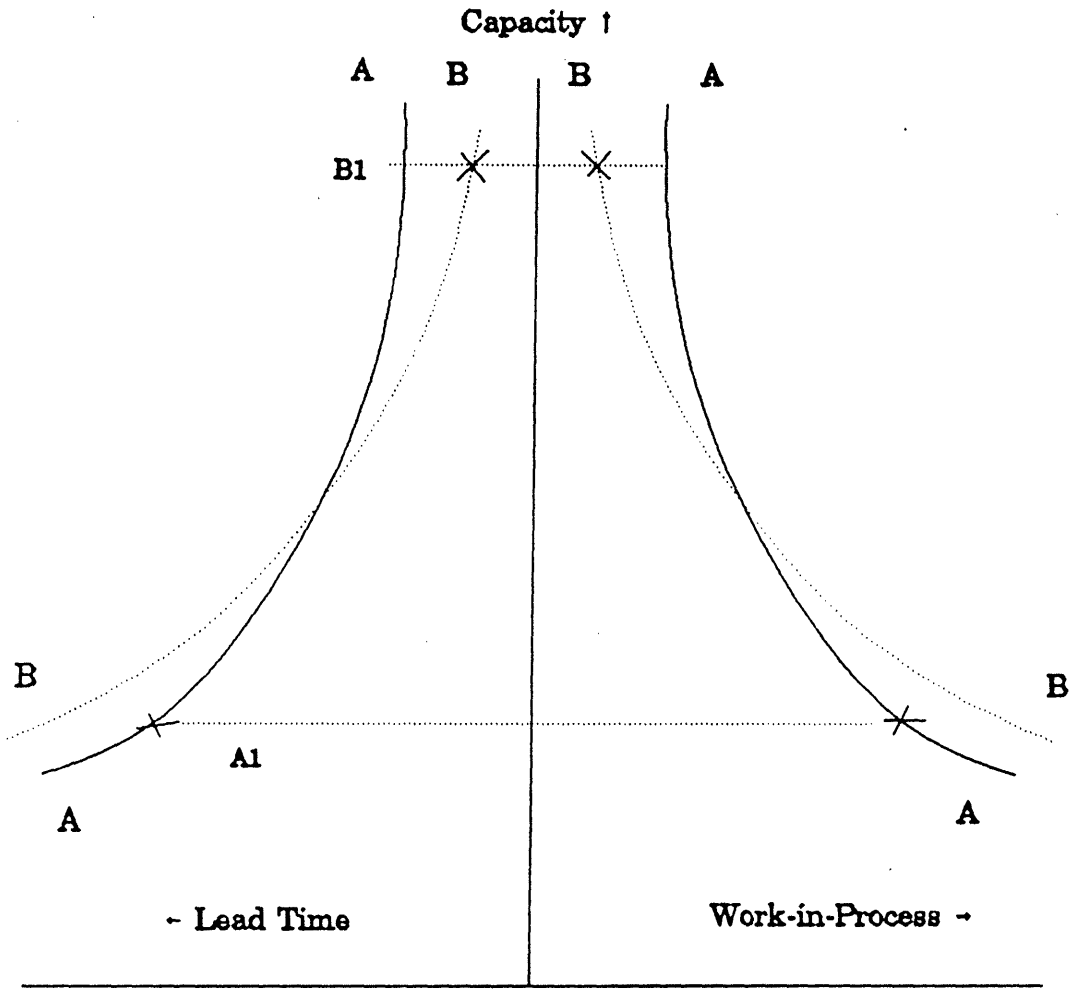
(i) the decision variables are either processing rates or number of servers at each station,

(ii) the objective function is a non-decreasing convex functions of the expected queue lengths, and

(iii) the constraints are upper bounds on a non-decreasing convex function of the queue lengths.

Design of a manufacturing system must take into account several, often conflicting factors. Hence solving problems Q1 or Q2 to optimality is not sufficient for selecting a design. It is useful to develop curves that for a given product mix, throughput and technology, describe the trade-off between the cost of machines (fixed capital costs) and work-in-process costs that are proportional to the number of jobs in the system. These curves can be developed by parametrically solving Q2 for different values of D, the available budget. Bitran and Tirupati (89b) describe a greedy heuristic that can also be employed for developing these curves.

Figure 1 illustrates the usage of these curves. In this figure we have two curves AA and BB each corresponding to a different technology, which we denote as technology TA and TB,



The tradeoff curves illustrate, as an example, the choices presented by two different technologies A and B. The figure indicates that A is better than B for low capacity investment while B is more desirable for shorter lead times.

Figure 1: An Illustration of Tradeoff Curves.

respectively. A firm competing on the basis short lead times may choose technology TB, and operate at point B. Where as a firm competing on the basis of lower costs may choose technology TA and operate at point A.

V.2 OPTIMAL CONTROL OF QUEUEING NETWORKS

We now turn our attention to the second class of optimization problems that determine control rules for a predetermined system. Here the network layout, the product mix, throughputs, and capacities are predetermined. The objective of the optimization problem is to control the flow of jobs or customers through the system. Under the hierarchical framework developed in section III, the class of queueing problems discussed in this sub-section address operational problems.

There is a large body literature concerning the optimal scheduling of multiclass queueing systems, where the scheduling decisions are to dynamically decide which class of customers to serve (eg. Sobel 79, Federgruen and Groenevelt 88). Although the theory for single-station systems is well developed, no papers exist on the optimal scheduling of a multiclass queueing network, which appears to be mathematically intractable by the standard semi-Markov decision process approach. However, Harrison (88) has shown how to approximate a queuing network scheduling problem by a dynamic control problem for Brownian motion. This heavy traffic approximation assumes that each station in the network is busy the great majority of the time, and thus focuses on the bottleneck, or most heavily loaded stations in the network. Fortunately, this is where most of the congestion and queueing occurs, and where scheduling can have its biggest impact.

Effective scheduling heuristics have been derived for a variety of problems by obtaining an

optimal solution to the Brownian control problem and interpreting this solution in terms of the original queueing system. Harrison and Wein (90a) and Laws and Louth (89) have each found policies for minimizing the long run expected average number of customers in the system for specific open network problems, and have shown that the performance of the proposed policy is very close to a lower bound on the best achievable performance. Harrison and Wein (90b) have developed an effective static scheduling policy for maximizing the average throughput rate of a two-station queueing network (where the customer population size is held fixed), and have derived an analytic comparison between the proposed policy and that of any static policy, such as the shortest expected processing time rule and the shortest expected remaining processing time rule. If ρ_i is the relative traffic intensity at station i and RW_{ik} is the expected remaining amount of work that station i needs to devote to a class k customer before it exits, then the policy ranks all classes by the index $\rho_2 RW_{1k} - \rho_1 RW_{2k}$, and awards higher priority at station 1 (respectively, station 2) to classes with smaller (respectively, larger) values of this index. Thus, each station feeds the other station as much work as it can, and in this way the overall idleness of the two servers is minimized.

Wein (90a,b) also considers the problem of releasing customers into the network (from an infinite buffer and subject to a specified class mix) and scheduling the customers at each station of a two-station network. The objective is to minimize the mean number of customers in the network subject to a lower bound constraint on the mean throughput rate. The proposed scheduling policy, called a workload balancing policy, is a dynamic index policy, where the indices are dynamic reduced costs derived from a linear program. If we let $W_i(t)$ be the total expected amount of work anywhere in the network for station i at time t, then the customer release policy, called a workload regulating policy, injects a customer into the system whenever the workload process $(W_1(t), W_2(t))$ enters a particular region in the nonnegative orthant of R^2. This analysis was generalized to the multistation

setting by Wein (90c), but the derivation of the release region becomes quite tedious.

All of the papers employing the Brownian analysis contain simulation experiments demonstrating the effectiveness (with respect to traditional policies) of the proposed customer release and priority scheduling policies. Thus, this approximation procedure has made progress in this difficult problem area.

VI CONCLUSIONS

In this paper we reviewed manufacturing problems that can be modelled as open queueing networks. Over the last decade three major developments have occurred in this area : (i) approximation techniques that provide good estimates of the performance of open queueing networks with general service times and deterministic routings, (ii) empirical discovery that some of the optimal network design problems can be closely approximated by convex programs, and (iii) identification of near optimal rules for control of flows through networks using Brownian control models.

Although these developments are of significant practical interest, there are still many other features, encountered in manufacturing settings that are not incorporated into the existing queueing network models. These features include:

(i) Arrival of Jobs to the Shop: All the queueing models assume that the arrival process is time homogenous and evaluate the equilibrium performance of the system. However in practice the arrival rates of jobs is likely to vary due to factors such as seasonality in demand, introduction of new products and elimination of old (obsolete) products. Also the correlation between the demands

for different products is ignored. Correlation may be induced, for instance, if customers typically order a set of products from the facility.

(ii) Priority Rules: Performance measures of the queueing network are estimated assuming a FCFS queue discipline, and ignore control rules for regulating the flow into and through the facility.

(iii) Equilibrium Analysis: Even in absence of seasonality in the demands the arrival pattern changes over time, thus the utility of equilibrium analysis has to be evaluated. This assumption requires empirical testing to identify when it is not appropriate.

(iv) Performance measures: The existing models concentrate on the first moment of the queue length. However, the second moment of the time spent in the shop, as an example of other measures, can be of great help in design and management of manufacturing systems that behave like open queueing networks.

Hence, although significant progress has been achieved in solving complex manufacturing problems, more research needs to be done to allow for the incorporation of features of practical importance and to validate many of the models proposed trough theory and empirical studies.

Acknowledgements: The authors thank Dr. D. Sarkar and Professors P.Kouvelis, G.Shirley, D. Tirupati, and L. Wein for their comments and suggestions.

REFERENCES

Ahmadi, R. and H. Matsuo (1990), " The line segmentation problem," Working paper, Anderson Graduate School of Management, UCLA. to appear in Operations Research

Albin, S.L. (1981), "Approximating queues with superposition arrival processes, "PhD dissertation, Dept of IE and OR, Columbia University, New York.

Albin, S.L. (1984), "Approximating a point process by a renewal process. II. Superposition arrival processes to queues," Operations Research, 32, 1133-1162

Anthony, R.N. (1965), Planning and Control Systems: A Framework for Analysis, Graduate School of Business Administration, Harvard University, Boston

Avriel, M. (1976) Nonlinear Programming, Analyses and Methods, Prentice Hall, Englewood Cliffs, N.J.

Berman, M. and M. Westcott(1983), " On queueing systems with renewal departure processes," Advances in Applied Probability, 15, 657-673.

Bertsekas, D.P. (1982), Constrained Optimization and Lagrange Multiplier Methods. Academic Press, New York.

Bertsimas, D. and D. Nakazato (1990), "The departure process from a GI/G/1 queue and its application to the analysis of tandem queues," Working Paper, Sloan School of Management, M.I.T.

Bitran, G.R. and S. Dasu (1990a), "Approximating non-renewal processes by Markov chains," Working Paper, Anderson Graduate School of Management, UCLA.

Bitran, G.R. and S. Dasu (1990b), "Analysis of $Ph_i/Ph/1$ queues," Working Paper, Anderson Graduate School of Management, UCLA.

Bitran, G.R. and D. Sarkar (1990), "Throughput analysis in manufacturing networks," Working paper, Sloan School of Management, M.I.T.

Bitran, G.R. and D. Tirupati (1988), " Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference," Management Science, 35, 851-878

Bitran, G.R. and D. Tirupati (1989a), "Capacity planning with discrete options in manufacturing networks," Annals of Operations Research, 17, 119-136

Bitran, G.R. and D. Tirupati (1989b), " Trade-off curves, targeting and balancing in manufacturing networks," Operations Research,37, 547 -564

Bitran, G.R. and D. Tirupati (1989c), "Approximations for product departures from a single server station with batch processing in multi-product queues," Management Science, 35, 851-878

Bitran, G.R. and D. Tirupati (1990), "Approximations for network of queues with overtime, " to appear in Management Science

Boxma, O.J., J.W. Cohen and N.Huffels (1979), "Approximations of the mean waiting time in an M/G/s queueing system," Operations Research, 27, 1115-1127

Boxma, O.J., A.H.G. Rinnooy Kan, and M. van Vliet (1990), "Machine allocation algorithms for job shop manufacturing," Econometric Institute Report 9014/A, Erasmus University, Rotterdam, The Netherlands, To appear in Journal of Intelligent Manufacturing

Brandwajn, A. and Y.L. Jow (1988) "An approximation method for tandem queues with blocking", Operations Research, 36,

Bugalak, A.A. and J.L. Sanders (1989), "Modelling and design optimization of asynchronous flexible assembly systems with statistical process control and repair," Technical report, University of Wisconsin - Madison

Burbidge, J.L. (1989), Production Flow Analysis for Planning Group Technology, Oxford University Press, New York.

Buzacott, J.A,. and J.G. Shanthikumar (1985), "Approximate queueing models of dynamic job shops," Management Science, 31, 870-887.

Buzacott, J.A. and D.D. Yao (1986a) "On queueing network models of flexible manufacturing systems," Queueing Systems 1, 1, 5-27

Buzacott, J.A. and D.D. Yao (1986b) "Flexible manufacturing systems : A review of analytic models, " Management Science 32, 7, 890-905

Chen, H., J.M. Harrison, A. Mandelbaum, A.A. Ackere and L.M. Wein (1988), "Empirical evaluation of a queueing network model for semiconductor wafer fabrication," 36, 202-215.

Conway, R.W., W.L. Maxwell, and W. Miller, (1967) Theory of Scheduling. Addison-Weseley, Reading, Mass

Cosmetatos, G.P. (1975), "Approximate explicit formulae for the average queueing time in the process (M/D/r) and (D/M/r)," INFOR, 13, 328-331

Denardo, E.V. and C.S. Tang (1989), "Bilinear control of Markov production systems, " Working paper, Anderson School of Management, UCLA, to appear in Operations Research.

Denning, P.J. and J.P. Buzen (1978), "The operational analysis of queueing network models," Computing Surveys, 10, 225-261

Disney, R.L. and D. Konig (1985), "Queueing networks: A survey of their random processes, " SIAM review 27, 335 - 403

Dyer, M.E., and L.G. Proll (1977), "On the validity of marginal analysis for allocating servers in M/M/c queues," Management Science, 23, 1019-1022

Erlang, A.K. (1917), "Solution of some problems in the theory of probabilities of some significance in automatic telephone exchanges," Post Office Electrical Engineer's Journal, 10, 189 -197

Federgruen, A. and H. Groenevelt (1988), "M/G/c queueing systems with multiple customer classes: characterization and control of achievable performance under nonpreemptive priority rules," Management Science, 34, 1121-1138

Friedman , H.D. (1965) "Reduction methods for tandem queueing systems," Operations Research, 13, 121-131

Garey, M.R., D.S. Johnson and R. Sethi (1976), "Complexity of flow shop and job shop scheduling algorithms," Operations Research, 24, 117 - 129

Gershwin, S.B., and I.C. Schick (1983) " Modelling and analysis of three stage transfer lines, with unreliable machines and finite buffers," Operations Research, 31, 354 -380

Graves S.C. (1981), "A review of production scheduling, " Operations Research, 29, 646 -675

Graves, S.C. (1986), " A tactical planning model for a job shop, " Operations Research, 34, 522 - 533

Harel, A. P.H. Zipkin (1987), "Strong Convexity results for queueing systems," Operations Research, 35, 405 - 418.

Harrison, J.M. (1988), "Brownian models of queueing networks with heterogeneous customer populations," in W.Fleming and P.L. Lions (eds.), Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA vol 10, Springer-Verlag, New York, 147 -186.

Harrison, J.M. and R.J. Williams, (1987) "Brownian models of open queueing networks with homogenous customer populations," to appear in Stochastics.

Harrison, J.M. and L.M. Wein (1990a), "Scheduling networks of queues: Heavy traffic analysis of a simple open network," Queueing Systems, 5, 265-280.

Harrison, J.M. and L.M. Wein (1990b), " Scheduling Networks of queues: Heavy traffic analysis of a two-station closed network," to appear in Operations Research.

Hax, A.C. and D. Candea (1984), Production and Inventory Management, Prentice-Hall, New Jersey

Hayes, R.H. and S.C. Wheelwright (1984), Restoring Our Competitive Edge; Competing Through Manufacturing, John Wiley and sons, New York.

Jackson, J.R. (1957), " Networks of waiting lines," Operations Research, 5, 518 -521

Jackson, J.R. (1963), "Jobshop-like queueing systems, " Management Science, 10, 131-142

Karmarkar, U (1987), "Lot sizes, lead times and in-process inventories, " Management Science, 33, 409 - 418

Karmarkar, U., S. Kekre, and S. Kekre (1985), "Lotsizing in multi-item, multi-machine job shops," IIE Transactions, 17, 290-298

Kelly, F.P. (1975), "Networks of queues with customers of different types, " Journal of Applied Probability, 12, 542-554.

Kelly, F.P. (1979), Reversibility and Stochastic Networks, John Wiley, New York

Kleinrock, L. (1964), Communication Nets: Stochastic Message Flow and Delay, Dover Publications, New York.

Kleinrock, L. (1976), Queueing Systems, Vol II: Computer Applications, John Wiley and sons, New York.

Kraemer, W. and M. Langenbach-Belz(1976), "Approximate formulae for the delay in the queueing system GI/G/1," Congressbook, Eighth Int. Teletraffic congress, Melbourne, 235- 1/8

Laws, C.N. and G.M. Louth, 1989, "Dynamic scheduling of four station network," Probability in the Engr. and Inf. Sciences.

Lemoine, A.J. (1977), "Network of queues - A survey of equilibrium analysis," Management Science, 24, 464 -481

Lenstra, J.K., A.H.G. Rinnooy Kan and P. Brucker (1977), "Complexity of machine scheduling problems," Annals of Discrete Mathematics, 1, 343-362

Lindley, D.V. (1952), "The theory of queues with a single server," Proc. Camb. phil. Soc. math. physc Sci., 48, 277- 295

Marchal, W.G. (1976), "An approximate formula for waiting time in single server queues," A.I.I.E. Transactions, 8, 473 - 486

Marshall, K.T. (1968), "Some inequalities in queueing," Operations Research, 16, 651-665.

Matsuo, H. and L. Gong (1990) "Smoothing production and stabilizing WIP in a production system with yield loss, " Working paper, University of Texas, Austin.

Nemhauser, G.L. and L.A. Wolsey (1988), Integer and Combinatorial Optimization, John Wiley and sons, New York.

Nuets, M.F. (1979), " A versatile Markovian point process," Journal of Applied Probability, 16, 764-779.

Neuts, M.F.(1981), Matrix Geometric Solutions in Stochastic Models, Johns Hopkins Univ. Press, Baltimore.

Page, E. (1972), Queueing Theory in O.R., Operational Research Series, Edited by K.B.Haley

Panwalker, S.S. and W. Iskander (1977), " A survey of scheduling rules," Operations Research, 25, 45 -61

Perros, H.G. and T. Altiok (eds.) (1989), Queueing Networks with Blocking, Elsevier Science Publishers, New York

Ramaswami, V. (1980), " The N/G/1 queue and its detailed analysis," Advances in Applied Probability, 12, 222- 261

Reich, E. (1957) "Waiting times when queues are in tandem, " Annals of Mathematical Statistics, 28, 768 -773

Reiser, M. and H. Kobayashi (1974), " Accuracy of diffusion approximations for some queueing systems," IBM journal of Research and Development, 18, 110 -124

Reiser,M. and S.S. Lavenberg (1980) "Mean value analysis of closed multichain queueing networks," JACM, 27, 313-322

Rinnooy Kan, A.H.G. (1976), Machine Scheduling Problems: Classification, Complexity and Computations. Nijoff, The Hague, Netherlands.

Rudemo, M. (1973), "Point processes generated by transitions of Markov chains," Advances in Applied Probability, 5, 262 -286

Segal, M. and W. Whitt (1988), " A queueing network analyzer for manufacturing, " Proce. of 12th int. Teletraffic Congress. Torine, Italy

Seidman, A., P.J. Schweitzer, and S. Shalev-Oren (1987), "Computerized closed queueing network models of flexible manufacturing systems: A comparative evaluation, " Large scale systems, 12, 91-107

Sevick, K.C. and I. Mitrani (1981), " The distribution of queueing network states at input and output instants," JACM 28, 358-471

Sevick, K.C., A.I. Levy, S.K.Tripathi, and J.L. Zahorjan (1977), "Improving approximations of aggreagated queueing network systems," Proc. Computer Performance, Modeling, Measurement and Evaluation.

Shantikumar, J.G., and J.A. Buzacott (1980), " On the approximations to the single server queue," IJPR, 18, 761-773.

Shanthikumar, J.G. and J.A. Buzacott (1981), "Open queueing network models of dynamic job shops," International Journal of Production Research, 19, 255-266

Skinner, W. (1974), " The focussed factory, " Harvard Business Review, May-June, 113 -121.

Sobel, M.J. (1979), "Optimal operation of queues," in Mathematical models in queueing theory, Lecture Notes in Economical and Mathematical Systems, vol 98, Springer Verlag.

Tantawi, A.N., and D. Towsley (1985), "Optimal static load balancing in distributed computer systems," JACM, 32, 445-465

Wein, L.M. (1990a), " Capacity allocation in Generalized Jackson Networks," Operations Research Letters, 8, 143-146.

Wein, L.M. (1990b), " Optimal control of a two-station brownian network," Mathematics of Operations Research, 15, 215-242

Wein, L.M.(1990d), " Scheduling networks of queues: Heavy traffic analysis of a multistation network with controllable input," to appear in Operations Research.

Wein, L.M. (1990c), " Scheduling networks of queues: heavy traffic analysis of a two-station network with controllable inputs," to appear in Operations Research.

Whitt, W.(1979), "Approximating a point process by a renewal process: A general framework," Bell Laboratories.

Whitt, W. (1982), "Approximating a point process by a renewal process: Two basic methods," Operations Research, 30, 125 -147.

Whitt, W. (1983a), "The queueing network analyzer," Bell Systems Technical Journal, 62, 2779-2843.

Whitt, W. (1983b), "Performance of the queueing network analyzer," Bell Systems Technical Journal, 63, 1911-1979.

Whitt, W. (1984), "Approximations for departure processes and queues in series," Naval Research Logistics Quarterly, 31, 499 -521

Whitt, W. (1985a), " Approximations for the GI/G/m queue," to appear in Advances in Applied Probability.

Whitt, W. (1985b) "Best order for queues in series," Management Science, 31, 745 -487

Whitt, W. (1988), " A light traffic approximation for single-class departures from multi-class queues," Management Science, 34, 1333-1346.

Appendix 1

Notation

c = number of servers,

ρ = station utilization,

ca = scv of interarrival time,

cs = scv of service time,

L_q = mean number of jobs in queue (excluding jobs in service).

L = mean number of jobs in the system (in queue and in service)

L_q^* = mean number of jobs in queue in our M/M/c system with utilization ρ

Whitt (1985)

$$L_q = \frac{\rho^2}{(1-\rho)} \frac{(ca+cs)}{2} g(\rho,ca,cs)$$

$$\text{where } g(\rho,ca,sc) = \exp\left[-\frac{2(1-\rho)}{3\rho} \frac{(1-ca)^2}{(ca+cs)}\right], \quad ca \le 1,$$

$$= \exp\left[-\frac{(1-\rho)(ca-1)}{(1+\rho)(ca+10cs^2)}\right], \quad ca > 1.$$

Shanthikumar and Buzacott (1980)

The table below corresponds to Table 2 of Shanthikumar and Buzacott (1980), and classifies approximations for mean number of jobs in queue in GI/G/1 system

cs ca	[0,0.3]	[0.3,0.7]	[0.7,0.9]	[0.9,1.0]
[0,0.1]	L_q(KL)	L_q(KL)	L_q(KL)	L_q(KL)
[0.1,0.3]	L_q(KL)	L_q(KL)	L_q(MARC)	L_q(MARC)
[0.3,0.7]	L_q(KL)	L_q(MARC)	L_q(MARC)	L_q(MARC)
[0.7,0.9]	L_q(KL)	L_q(MARC)	L_q(PAGE)	L_q(PAGE)
[0.9,1.0]	L_q(PAGE)	L_q(PAGE)	L_q(PAGE)	L_q(PAGE)

where

$$L_q(KL) = \frac{\rho^2}{(1-\rho)} \frac{(ca+cs)}{2} g(\rho,ca,cs),$$

$$g(\rho,ca,cs) = \exp\left[\frac{-2(1-\rho)}{3\rho} \frac{(1-ca)^2}{(ca+cs)}\right], \; ca \le 1,$$

$$= \exp\left[\frac{-(1-\rho)\,(ca-1)}{(ca+4cs)}\right], \; ca > 1,$$

$$L_q(MARC) = \frac{\rho^2(1+cs)}{(1+\rho^2 cs)} \frac{(ca+\rho^2 cs)}{2} (1-\rho),$$

$$L_q(PAGE) = \frac{\rho^2}{(2(1-\rho))} \left[ca(1+cs)+cs(1-ca) \exp \frac{-2(1-\rho)}{\rho}\right].$$

GI/G/c System

Whitt (1985)

$$L = L_q + c\rho$$

$$L_q = \phi(\rho,ca,cs,c) \cdot \frac{ca+cs}{2} \cdot L_q^*,$$

where $\phi(\rho,ca,cs,c)$

$$= \frac{4(ca-cs)}{4ca-3cs}\phi 1(c,\rho) + \frac{cs}{4ca-3cs}\,\theta((ca+cs)/2,c,\rho), \; ca > cs$$

$$= (cs-cs)/(2(ca+cs))\,\phi 3(c,\rho) + (cs+3ca)/(2(ca+cs))\,\theta((ca+cs)/2,c,\rho), \; ca \le cs$$

$$\theta(\alpha,c,\rho) = 1, \; \alpha > 1$$

$$= (\phi 4(c,\rho))^{2(1-\alpha)}, \; 0 \le \alpha \le 1.$$

$$\delta(c,\rho) = \min\{0.24,(1-\rho)(c-1)[(4+5c)^{0.5}-2]/(16c\rho)\}$$
$$\phi 1(c,\rho) = 1+\delta(c,\rho)$$
$$\phi 2(c,\rho) = 1-4\delta(c,\rho)$$
$$\phi 3(c,\rho) = \phi 2(c,\rho)\exp(-2(1-\rho)/(3\rho))$$
$$\phi 4(c,\rho) = \min\{1,(\phi 1(c,\rho)+\phi 3(c,\rho))/2\}.$$

Appendix 2

Computation of mean and scv of modified service time

Let t, t and ct respectively denote process time, its mean and its scv,

g, g and cg respectively denote repair time, its mean and its scv,

s, s and cs respectively denote modified service time, its mean and its scv,

u expected time between successive failures, and

E() the expectation operator

Then the modified service time, s has the following characteristics

$$s \sim t \text{ with probability } (1-p)$$
$$\sim t+g \text{ with probability } p$$

where p = the probability that a job has a breakdown while in process

= t/u

Noting that $E(t) = \bar{t}$, $E(t^2) = (1+ct)t^{-2}$,
$E(g) = \bar{g}$, $E(g^2) = (1+cg)g^{-2}$,
and $E(t+g) = \bar{t}+\bar{g}$, $E((t+g)^2) = t^{-2}(1+ct)+g^{-2}(1+cg)+\overline{tg}$,

We obtain

Note that the procedure above can be modified to include different types of breakdowns with associated repair characteristics.

$$\bar{s} = E(s) = \bar{t}+p\bar{g} = \bar{t}\left(1+\frac{\bar{g}}{\bar{u}}\right),$$

$$E(s^2) = (1-p)E(t^2) + pE((t+g)^2)$$
$$= \overline{t^2}(1+ct) + p[\overline{g^{-2}}(1+cg)+2\overline{tg}],$$

$$\text{and } cs = \frac{1}{\bar{s}^2}\left[\overline{t^2}ct + p\overline{g^2}(1-p+cg)\right].$$