Sādhanā, Vol. 15, Parts 4 & 5, December 1990, pp. 301–317. © Printed in India.

Stochastic analysis of versatile workcentres

R RAM and N VISWANADHAM

Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560012, India

Abstract. We analyse the system consisting of a highly capable workcentre, which processes a variety of part types, using queueing models. The various part types produced by the system have distinct arrival and processing durations that are stochastic in nature. When an arriving workpiece finds the machine busy, it waits in a pre-process storage buffer (queue); this buffer may be common for all the part types, or may be dedicated for that part type. Further, this buffer may be capable of holding only a finite number of workpieces, or may be of infinite capacity. When the machine changes over from producing one type of part to another, a setup operation of stochastic duration is necessary to adjust the machine and load the necessary tools for production of the next part type. This model is representative of a typical machining centre in an Automated Manufacturing System. We focus on GI/G/1 models and multiqueue polling models, and their variants. The important performance measures of the system obtained by queueing analysis are the part-type-wise values of the mean lead time, mean inventory level, and the mean machine utilisation.

Keywords. Stochastic analysis; versatile workcentres; automated manufacturing system; lead time; inventory level; machine utilisation.

1. Introduction

In this paper, we present stochastic models for flexible machining centres which are the basic processing nodes of an Automated Manufacturing System (AMS). The major goal of automated manufacturing, viz., high quality, low volume production of a variety of part types concurrently, with low lead times, is sought to be achieved, based on the capability of the individual workcentres to process different part types. Hence, an understanding of the operation of these workcentres in the face of demands to produce multiple varieties of parts is very important in a performance evaluation study of the automated manufacturing system.

The most important performance measures obtained from a study of this basic system are the *mean lead time* (the time duration from the entry of raw workpiece to the system to the completion of the processing operation on the workcentre), the *mean inventory level* (the mean number of workpieces waiting ahead of a newly arriving workpiece), and the *mean machine utilisation* for each of the part types produced. These performance measures are of great importance in manufacturing, especially so

in automated manufacturing where we seek to minimise the lead times and inventories (ideally, lead time must be just the processing time, and inventory level must be just the workpieces under processing), these performance measures are particularly important. In the terminology of the queueing model approach of our paper, these performance parameters correspond to the mean response time, the mean queue length and mean server utilisation respectively; the different part types are represented by multiple job classes.

1.1 *The model*

We solve the following system: there is a versatile machining centre capable of processing a variety of parts, of which there are N different types. Workpieces of part type i arrive to the system according to a general, independent and identically distributed interarrival process of mean *rate* λ_i. If the machine is unable to serve this arriving workpiece immediately, the workpiece joins a queue to wait for the machine; the queue may be dedicated to that particular part type, or may be common to all part types. Further, the queue may be of finite capacity (capable of accommodating a maximum of a finite number of workpieces, say K_i for part type i). For processing workpieces of part type i, the machine has to be setup for this part type; this requires a setup operation, which has a general independent and identically distributed (i.i.d.) duration represented by the random variable R_i. If this setup is not disturbed, subsequent waiting workpieces of part type i can be processed without incurring additional setup. The machining of a type i workpiece takes a duration which is general i.i.d. and is represented by B_i. Processed parts are assumed to leave the system immediately. Let T_i denote the lead time (system response time) and L_i the mean inventory level (queue length) for part type i; we shall omit the subscript i when we deal with a single part type. The various aspects of the system considered in this paper may be modelled to varying degrees of detail; the issues considered, and their simplest and highest degrees of detail are summarised in table 1.

To complete the description of the model, we need to specify how the machine chooses a particular part type (*scheduling policy*), and, once a particular part type is selected, how the waiting workpieces of that particular part type are processed (*service policy*). We shall refer to the combination of scheduling and service policies as *operation policy*. It must be noted that for certain operation policies, like the first-come-first-served and Bernoulli scheduling policies, the scheduling and service policies overlap; in such cases, for simplicity, we shall continue to refer to them as scheduling policies.

The scheduling policy determines which of the N different part types is taken up

Table 1. Aspects of the systems considered.

Issue	Simplest assumption	Most detailed assumption
Number of queues	Common queue for all parts	Separate queue for each part
Queue capacity	Infinite capacity	Finite capacity (for each part type)
Setup time	Treated as a part of processing time. (Not considered explicitly)	Distinct setup time for each part type. (Explicitly considered)

for processing next when the processing of a particular part type is completed. Examples of scheduling policy that can be clearly demarcated from service policies are:

Cyclic scheduling: We assume that there is a separate queue (which may be of finite capacity or infinite capacity) for each part type. The different part types are taken up in the cyclic order $1, 2, 3, \ldots, N-1, N, 1, 2$, etc. The processing of each part type is preceded by a corresponding setup operation for that part type. This type of scheduling has been widely discussed in the literature on the analysis of computer systems and computer communication networks and is referred to as *cyclic service or polling*. The two extensive survey articles by Takagi (1988, 1990) provide a comprehensive discussion on the modelling, analysis and applications of these and related systems.

Probabilistic scheduling: We assume separate queues for different part types. Whenever the machine finishes processing a part type, the part type i is chosen next with a probability p_i, $(0 < p_i < 1$ for $1 \leqslant i \leqslant N$, and $\Sigma_{i=1}^{N} p_i = 1)$. Each service is preceded by an appropriate setup. These models are referred to as *random polling* models; an analysis of the discrete time version of this system is developed by Kleinrock & Levy (1988).

Markovian scheduling: This is an extension of probabilistic scheduling. Each part type has a separate queue. When the machine completes processing for (say) type i parts, we next take up type j parts for machining with probability $p_{i,j}$, where $0 \leqslant p_{i,j} \leqslant 1$, for $1 \leqslant i, j \leqslant N$, and for each part i, $\Sigma_{j=1}^{N} p_{i,j} = 1$. We assume that the $N \times N$ matrix $[p_{i,j}]$ forms the transition matrix of a irreducible, discrete time Markov chain.

The service policies can be clearly distingushed from the scheduling policy for the cases mentioned above. The service policy determines how much service is carried out for each part type, once the machine is setup for it. The service policies widely considered in the literature are:

Exhaustive: Once the workcentre is setup for a part, say type i, machining of type i parts is continued till no further workpieces of type i are waiting. In particular, those workpieces that may arrive during the current setup are also machined during this setup.

Gated: The machine processes all and only those workpieces that are waiting when the setup is completed. In particular, those workpieces of this part type that may arrive during the current setup have to wait for the next setup to be processed.

Limited: The amount of machining carried out on a part type is limited by specifying an upper limit on either the number of workpieces that may be processed during a setup or the maximum time for which waiting workpieces of the part type may be taken up for processing. Accordingly, we have either

—*K-limited service*, where a maximum number, say L_i of a part type i workpieces may be processed during the current setup, or

— *T-limited service*, where processing of type i parts is allowed for a maximum duration L_{T_i}.

Further, parts may be taken up in either exhaustive or gated mode for processing

under each of the above service disciplines, giving us a total of four variants of limited service. The K-limited, exhaustive discipline is the one commonly treated in the literature.

Given one of the above mentioned scheduling policies, if all the N part types adapt the same service discipline, e.g., all the part types use the gated service policy, then we have a system with *homogenous* service policy. On the other hand, if different part types have distinct service policies, e.g., part 1 has exhaustive service, part 2 has gated service, part 3 has K-limited-gated service and so on, we have a system with *mixed* service.

The following are examples of operation policies where the scheduling and service policies overlap.

First come first served (*FCFS*): This assumption is commonly made in the literature whenever there are random arrivals: all the arriving workpieces wait in a single, infinite capacity queue, from which they are taken up in an FCFS fashion for processing. The GI/G/1 queueing models are based on this assumption. Such models have been extensively applied in the modelling of manufacturing systems as a network of queues, each queue representing a workcentre.

Bernoulli scheduling: This is an example of a non-exhaustive service discipline applicable to multiqueue systems (i.e., a separate queue for each part type). At each instant of completion of machining of part type i, if the queue of part i is not empty, the machine makes a random decision: with probability p_i, $(0 \leqslant p_i \leqslant 1)$, it processes

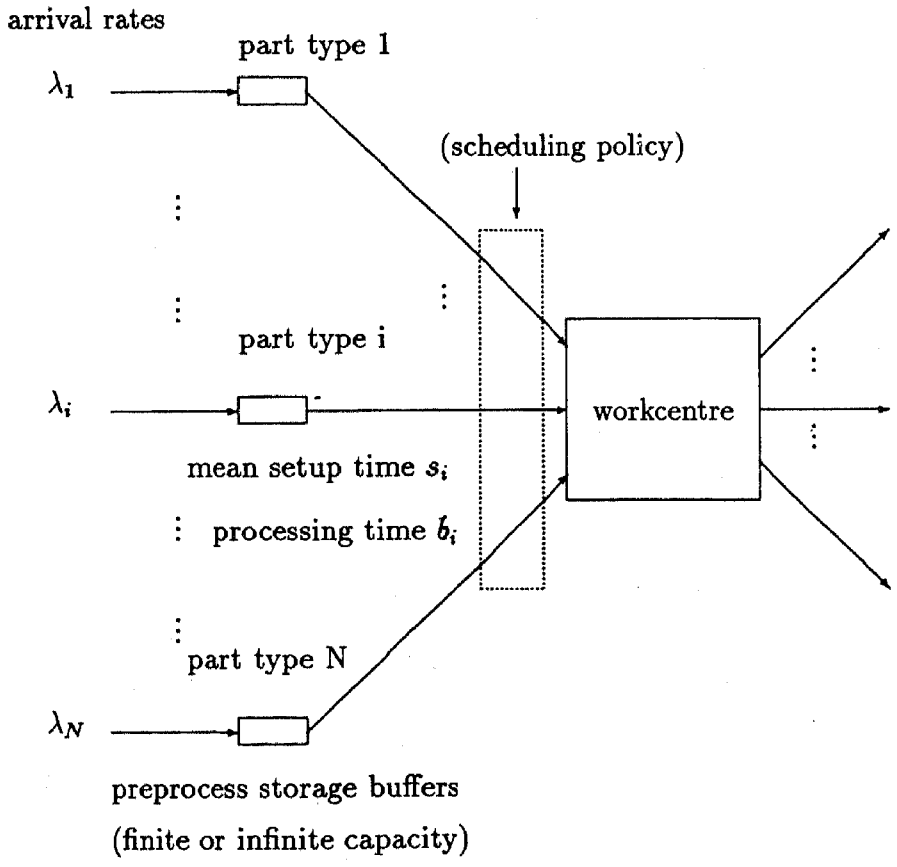

Figure 1. The basic model of a versatile workcentre.

the next waiting workpiece of the same part type, and with probability $1 - p_i$ it changes over to the next part type. If the queue of part i was empty at the completion of a type i workpiece machining, it changes over to the next part type with probability 1. When the machine switches from one part type to the next, a setup is carried out for the new part type before actual machining. Recently, Tedijanto (1990) presented an analysis of this service discipline; the symmetric version of this case has been exactly solved for the mean waiting time of each part type.

The basic model of the workcentre treated in this paper is depicted in figure 1.

The organisation of the paper is as follows: in the rest of this section, we present a brief survey of the literature on work related to this paper. In the next section, we present single queue, single server systems which are simplifications of the $GI_1, GI_2, \ldots, GI_N/G_1, G_2 \ldots, G_N/1/FCFS$ systems. Section 3 focusses on queueing analysis of multi-queue models, particularly cyclic server models. In § 4, we present numerical examples from the manufacturing context based on the models presented in this paper. Section 5 concludes the paper.

1.2 *Related literature*

The issue of scheduling multiple part varieties on a single machine under *known* demands, processing times and setup times (i.e., *the deterministic scheduling problem for a single machine*) has been well researched (Dobson *et al* 1987 and references therein). Several heuristic algorithms have been proposed for the control of this system under time varying stochastic demands, but assuming fixed production rates, e.g., Leachman & Gascon (1988). Recent investigations on the scheduling of manufacturing systems have been carried out using a hierarchical approach; in particular, the work of Gershwin (1989), Perkins & Kumar (1989), and Kumar & Seidman (1990) are particularly relevant. The hierarchical approach is based on a *time scaling* in the activities occurring in a manufacturing system: a part processing may take about an hour; a setup operation may take a duration an order of magnitude longer, say 10 hours; a machine failure may occur once in 200 hours. In these hierarchical production scheduling systems, the processing operations carried out by a workcentre are at the lowest level of the time scaling (fastest activities), and the setup activities are at the immediately higher level. The queueing models in this paper are utilised to analyse the system consisting of a versatile workcentre, along with its preprocess buffers, at these two lower time scales. However, in automated manufacturing an important aspect is the reduction of setup times through the use of mechanisms like automatic tool changers, and hence setup and operation times are typically of comparable magnitude. We have basically drawn upon the relevant results in the vast literature on single server queues the GI/G/1 queue, and its variants, and from the literature on *polling models* (Takagi 1988, 1990) developed particularly in the context of computer communication. Relevant literature on these models are surveyed and referenced at the appropriate point in the next two sections.

2. Single queue models

We start with queueing models where the different part types share a single common queue, which may be of infinite or finite capacity (see figure 2).

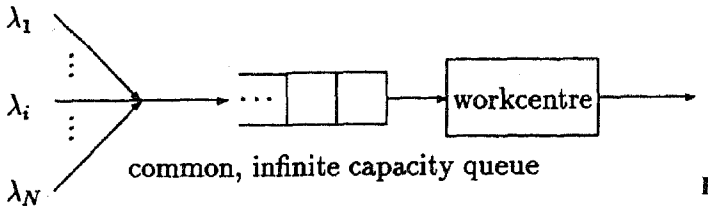

Figure 2. The common queue model.

2.1 *M/G/1 queueing model*

The simplest model of the system is the well known M/G/1 queue, widely described in the queueing literature. This entails that we make the following simplifying assumptions on the model:

(i) All arriving workpieces share a common, infinite capacity queue in an FCFS fashion.

(ii) The processing time is interpreted as the sum of the setup time and the actual machining time (i.e., the setup time is not considered explicitly). If B_i' represents the modified processing time of a workpiece of part type i, then

$$B_i' = B_i + p_{\text{setup},i} R_i,$$

where $p_{\text{setup},i}$ is the probability that a setup operation has to be carried out afresh. Since we have assumed exponential arrivals, $p_{\text{setup},i}$ is the same as the probability that the previous part type processed was *not* of type i, so $p_{\text{setup},i} = 1 - (\lambda_i / \Sigma_{j=1}^{N} \lambda_j)$. In terms of Laplace transforms, we have $\hat{B}_i'(s) = \hat{B}_i(s) \times [p_{\text{setup},i} \hat{R}_i(s) + 1 - p_{\text{setup},i}]$.

(iii) The total arrival rate to the system is $\lambda = \Sigma_{i=1}^{N} \lambda_i$, and the service time of an arbitrary workpiece has the transform

$$\hat{B}'(s) = \Sigma_{i=1}^{N} (\lambda_i / \lambda) \hat{B}_I'(s)$$

We are in effect analysing the $M_1, \ldots, M_N/G_1, \ldots, G_N/1$ queue by an M/G/1 model. The analysis of the system now proceeds along the standard method for the M/G/1 queue, which may be found in any standard text on queueing theory or stochastic modelling. A simple variant of this system, where each part type has distinct exponential arrival and service rates, i.e., the $M_1, \ldots, M_N/M_1, \ldots, M_N/1$ system, is discussed in an early paper by Ancker & Gafarian (1961).

2.2 *GI/G/1 queueing model*

2.2a *Related work*: Several investigations have been carried out on the approximate analysis of the GI/G/1 queueing system, which constitutes a general model for the single machine multiple part type processing system; see Shanthikumar & Buzacott (1980) for a survey of important results, and recommended methods applicable for specific parameter ranges. This is also useful in approximate analysis of non-product-form open general queueing networks of automated manufacturing systems, by adapting the product-form networks idea of decomposing the network into independent nodes corresponding to machining centres. Approximate analysis of general open queueing networks, and further refinements, have been proposed based on this independence concept by several authors – Kuehn (1979), Marie (1979), Whitt (1983), Shanthikumar & Buzacott (1981, 1985) and Bitran & Tirupati (1988, 1989). In

particular, the investigations of Shanthikumar and Buzacott, and Bitran and Tirupati are directed towards general open queueing network models of manufacturing systems, the former dealing with single product networks, and the latter with multi-product networks. The core of all these investigations is an appropriate analysis of individual workcentres processing multiple types of parts, and different processing times for different parts, by a GI/G/1 model under a multi-product assumption. In fact, several of the approximate models for this system have been developed especially for application to the analysis of networks of GI/G/1 systems with arbitrary routing. Shanthikumar & Gocmen (1983) have applied the principle of decomposition of a queueing network into individual independent nodes in developing a heuristic analysis of a closed network of GI/G/1 queues. In this subsection, we summarise the recent approaches to solving this model in the manufacturing context.

2.2b *Simplifying assumptions:* The analyses in these investigations are carried out using the knowledge of the first two moments of the interarrival and service processes (the '*parametric analysis*' of general queueing networks). In the context of the versatile machine-multiple part types case, the application of the GI/G/1 queueing model entails the following assumptions:

- the setup time is not explicitly considered; it may be treated as part of the machining time.
- the arrival and service processes, which can be different for different part types, have to be suitably combined. In the approach of the earlier papers, the system performance measures are obtained for a 'typical' part type, which is representative of all the part types; in a recently developed alternative approach (particularly Bitran & Tirupati 1988, 1990), performance measures are obtained for each part type individually by simplifying the system into a two-part type system: one part type is the particular part type of interest, and the second is an aggregate part type representative of all the other part types.
- the arrivals wait in an infinite capacity queue common to all part types, from which workpieces are taken up for machining on a first-come-first-served basis.

2.2c *Approximations based on the aggregation of all part types into a single 'typical' part type:* Let λ_j and c_{aj}^2 denote respectively the mean *rate* and the squared coefficient of variation (s.c.v.) of the interarrival process for part type $j (1 \leqslant j \leqslant N)$, and τ_j and c_{sj}^2 denote the mean processing *time* and the squared coefficient of variation of the service process respectively. We have to approximate the system at two stages to apply the model: (i) the parameters corresponding to the various part types have to be combined into those corresponding to a 'typical' part type; (ii) depending on the composite parameters obtained for the 'typical' part type, a suitable approximation of the GI/G/1 queue is to be applied.

(i) *Arrival process of the typical part type* – The mean total arrival rate of the 'typical' part to the system is given by the sum of the arrival rates of individual parts: $\lambda = \Sigma_{j=1}^{N} \lambda_j$.

The s.c.v. of the arrival process is obtained by the hybrid approximation of Albin used by Whitt (1983) as

$$c_A^2 = w \sum_{j=1}^{N} (\lambda_j/\lambda) c_{aj}^2 + 1 - w,$$

where w is a weight, given by

$$w = [1 + 4(1 - \rho)^2(N - 1)]^{-1}.$$

(Here, ρ denotes the offered load, or the utilisation, of the machine, and is given by $\rho = \lambda\tau$.)

(ii) *The service process of the typical part* – The mean service time of the typical workpiece is given by $\tau = (\Sigma_{j=1}^{N} \lambda_j \tau_j)/\lambda$, and its squared coefficient of variation of the service process is obtained as (Whitt 1983)

$$\tau^2(c_S^2 + 1) = \left[\sum_{j=1}^{N} \lambda_j \tau_j^2 (c_{sj}^2 + 1)\right] \Big/ \left[\sum_{j=1}^{N} \lambda_j\right].$$

Depending on the values obtained for c_S^2 and c_A^2, we can adopt the methods suggested by Shanthikumar & Buzacott (1980) and Whitt (1983) for the approximate analysis of the GI/G/1 system. We omit the mathematical details of these computations. See Whitt (1983) for a further discussion on approximate computation of distribution of the waiting time. In all these cases, the mean system lead time (waiting time in queue plus processing time) for a 'typical' part type is obtained from Little's Law as $E[R] = E[L]/\lambda$.

The work of Shanthikumar & Buzacott (1980, 1981) and Whitt (1983) is based on representing a multiproduct, general open queueing network in terms of a single representative part type typical of all the part types; the service and arrival parameters at each node are modified to aggregate the behaviour of all the part types by this typical part type. The routing information of individual parts are usually deterministic; the aggregation methodology converts this into Markovian routing. (The deterministic routing of different part types is retained in an exact analysis if we model the system by special classes of queueing networks, like open product-form queueing network models (Baskett *et al* 1975) or the closely related quasi-reversible networks (Kelly 1979); in this paper we are dealing with more general queueing models of individual machines (nodes) a network of which does not fall in these exactly solvable categories.)

2.2d *Approximate analyses explicitly considering distinct part types*: Bitran & Tirupati (1988) point out that the randomisation of deterministic routing of the various part types to Markovian routing in the aggregation process can lead to significant errors in the computation of system performance measures. They propose a refinement for computing the squared coefficient of variation of the departure process of a specific part type, say i, from a node (which is a GI/G/1 queue). To compute the s.c.v. of the departure process of part type i from a particular workcentre, we consider the following simplified view of each workcentre: consider the specific part type i, and an aggregate part type, which represents all part types other than i. The s.c.v. $c_D^2(i)$ of the departure process of part i from the workcentre may be computed from

$$c_D^2(i) = [(\lambda_i/\lambda)c_D^2] + c_{n(i)}^2,$$

where the new term $n(i) = z(i) + 1$, where $z(i)$ is the number of workpieces of the aggregate type that arrive during an interarrival time of part type i, $c_{n(i)}^2$ represents the s.c.v. of $n(i)$. The difficult part is estimating the s.c.v. $c_{n(i)}^2$. The authors propose three approximations for this purpose, which may be briefly stated as follows: (i) the arrivals of the aggregate product may be treated as Poisson; (ii) during any interarrival

period of the part type i (aggregate part type), the interarrival process of the aggregate part type (part type i) may be treated as Erlangian; and (iii) the arrival of part type i is treated as a random incidence in the arrival stream of the aggregate product; both the part type i and the aggregate part type have Erlang arrivals. Extensive numerical investigations by Bitran & Tirupati (1988) indicate that the proposed approximations provide significant improvements in accuracy over the aggregation approach previously employed. In a subsequent paper (Bitran & Tirupati 1989), they develop an approximate analysis for a single workcentre multiple-item system, where the arriving workpieces of the different part types have to form a batch of a given (fixed) size before service can be started.

In a related paper, Whitt (1988) has developed the theory for the output process for a particular part type from a node, when that part type is in light traffic. The basic idea of the light traffic approximation for the departure process of a single part type from a node may be stated thus: "If the arrival rate of one class (part type) to some queue (machine) is a small proportion of the total arrival rate there, then the departure process for that class from that queue tends to be nearly the same as the arrival process for that class to that queue". This principle can be used in an approximate analysis of a general multiclass queueing network. A simplified characterisation of the departure process of a GI/G/1 system with multiple arrival streams (part types) by a renewal process is developed by Albin (1986).

2.3 *Finite capacity queues*

Thus far we have not put any restriction on the capacity of the preprocess buffer queue. In real life manufacturing, the number of fixtures available is limited, and so is the space available to hold waiting workpieces in front of a machine. Hence, it is more realistic to assume a finite capacity for the preprocess queue. In the following, we summarise the results for finite buffer systems.

2.3a M/G/1/N *queue model*: This is finite capacity analog of the M/G/1 model. We may think of the well known M/M/1/N system as an elementary version of this model, where all the part types are aggregated (logically) into one representative part type, which has exponential arrival and service in a single server system with a total buffer capacity of N, including the server. Basharin (1965) has analysed the finite queue analog of the case where each part type has distinct exponential arrival and service in a finite capacity queue, single server system, i.e., the $M_1,\ldots, M_N/M_1,\ldots,M_N/1/N$. The general M/G/1/N system is analysed by Lavenberg (1975). An arrival that finds all the N buffer spaces full is assumed to be lost. Lavenberg presents an expression for the Laplace–Stieltjes Transform (LST) of the distribution of the queueing time in the system, in terms of the steady state probability of the imbedded Markov chain at the departure epochs, and the LST of the service time distribution. The relevant mathematical results are highly detailed, hence we omit them.

2.3b *The GI/G/1/N model*: This is the finite queue analog of the previous model, representing an important real life issue, viz., the availability of only a finite number of buffer spaces to hold waiting part types in front of a workcentre. An analysis of this system will also be fundamental to an analysis of a network of finite capacity queues, i.e., general open queueing networks with blocking.

3. Multiqueue models

We next turn our attention to queueing models, where each part type forms a separate queue, which are attended to by the versatile workcentre represented by a single server. Typically, the different part types are served in cyclic order, and each queue is assumed to have infinite capacity. Few results are available for the finite-capacity queues case.

3.1 *Cyclic server (polling) models with infinite buffer queues*

This model exhibits a separate queue for each part type. Workpieces of type i arrive with an exponential interarrival time or rate λ_i. The service time of a workpiece of a part type is given by an independent random variable of general distribution, denoted by B_i. The different part types are taken up for processing in the cyclic order $1, 2, \ldots, N-1, N, 1, 2$, etc. The processing of type i workpieces is preceded by a setup operation for part type i, given by a generally distributed duration R_i, which depends on part i, but is independent of other system parameters. When the machine has served waiting workpieces of type i (according to the given service policy), it changes over to the next part type $(i+1)$ modulo N by initiating a setup for this part type. Such a cyclic server or polling model has been solved exactly for the mean waiting time in queue of each part type, for the following cases.

(i) The service policy within a queue is the same for all the N part types, and can be exhaustive or gated. Within a queue, workpieces are processed in FCFS order
(ii) The special case where K-limited-exhaustive discipline is applied to the system for all the N part types, under the assumption that all the N parts have identical parameters, i.e., same arrival, setup and processing times (i.e., the *symmetric* case).

3.1a *Exhaustive service*: The system may be solved for the mean waiting times as follows: let b_i and $b_i^{(2)}$ denote the mean and second moment of the processing time for part i, B_i. Let $\rho_i = \lambda_i \times b_i$ be the utilisation of the machine by part type i, and let $\rho = \Sigma_{i=1}^{N} \rho_i$ be the total utilisation of the machine (the utilisation values exclude the time spent on setups). The mean of the total time spent on setup in each cycle is $R = \Sigma_{i=1}^{N} r_i$, and the variance of this time is $\Delta^2 = \Sigma_{i=1}^{N} \delta_i^2$.

The system is stable, i.e., the queue lengths for each part type do not build up to infinity, if $\rho < 1$.

The waiting time in queue of a type i workpiece is given by (Takagi 1988)

$$E[W_i] = E[I_i^2]/\{2E[I_i]\} + \lambda_i b_i^{(2)}/\{2(1-\rho_i)\},$$

where I_i denotes the intervisit time for the queue corresponding to part i. The intervisit time is defined as the duration starting the instant the machine leaves queue i and ending the instant the machine finishes setup for part i in the next cycle. The mean $E[I_i]$ and the variance $Var[I_i]$ of the intervisit time are given by

$$E[I_i] = (1-\rho_i)R/\{(1-\rho)\}$$

and

$$Var[I_i] = \delta_i^2 + \{(1-\rho_i)/\rho_i\} \sum_{j=1, j\neq i}^{N} r_{ij}$$

where $\{r_{ij}\}$, $1 \leqslant i, j \leqslant N$ are the set of covariances for the station times of the queues

for part types i and j. The station time for queue i is defined as the time interval between successive instants the machine starts setup for parts i and $i+1$. The values of the covariances r_{ij} are obtained by solving the set of the following $O(N^2)$ linear equations (Takagi 1988).

$$r_{ij} = \frac{\rho_i}{1-\rho_i}\left(\sum_{m=i+1}^{N} r_{jm} + \sum_{m=1}^{j-1} r_{jm} + \sum_{m=j}^{i-1} r_{mj}\right), \quad \text{for } j < i;$$

$$r_{ij} = \frac{\rho_i}{1-\rho_i}\left(\sum_{m=i+1}^{j-1} r_{jm} + \sum_{m=j}^{N} r_{mj} + \sum_{m=1}^{i-1} r_{mj}\right), \quad \text{for } j > i;$$

and

$$r_{ii} = \frac{\delta_i^2}{(1-\rho_i)^2} + \frac{\lambda_i b_i^{(2)} E[I_i]}{(1-\rho_i)^3} + \frac{\rho_i}{1-\rho_i}\left(\sum_{j=1, j\neq i}^{N} r_{ij}\right).$$

3.1b *Gated service*: We follow the same notation as in the exhaustive service case. The mean waiting times in queue are obtained by solution of a set of $O(N^2)$ linear equations. The mean waiting time for a workpiece of part type i is given by (Takagi 1988)

$$E[W_i] = (1+\rho_i)E[C_i^2]/\{2E[C]\},$$

where C_i is the random variable denoting the cycle time for part type i, defined as the time interval between successive instants of setup initiation for part type i. The expected value of the cycle time is independent of the part type and is given by

$$E[C] = E[C_i] = R/(1-\rho).$$

The condition for system stability is $\rho < 1$.

To obtain $E[C_i^2] = (E[C_i])^2 + Var[C_i]$, we solve for $Var[C_i]$ from:

$$Var[C_i] = \left(\frac{1}{\rho_i}\right)\sum_{j=1, j\neq i}^{N} r_{ij} + \sum_{j=1}^{N} r_{ji}.$$

The values $\{r_{ij}; 1 \leqslant i,j \leqslant N\}$ are again the set of covariances of station times. Here the station time for part type i is defined as the time interval between the successive instants the setups for part types i and $i+1$ are completed. The values of r_{ij} are obtained by solving the following set of linear equations (Takagi 1988)

$$r_{ij} = \rho_i\left(\sum_{m=i}^{N} r_{jm} + \sum_{m=1}^{j-1} r_{jm} + \sum_{m=j}^{i-1} r_{mj}\right), \quad \text{for } j < i;$$

$$r_{ij} = \rho_i\left(\sum_{m=i}^{j-1} r_{jm} + \sum_{m=j}^{N} r_{mj} + \sum_{m=1}^{i-1} r_{mj}\right), \quad \text{for } j > i;$$

and

$$r_{ii} = \delta_{i+1}^2 + \lambda_i \times b_i^{(2)} \times E[C] + \rho_i \times \sum_{j=1, j\neq i}^{N} r_{ij} + \rho_i^2 \times \sum_{j=1}^{N} r_{ji}.$$

3.1c *Limited service*: The limited service discipline has not been solved for the general asymmetric case (the solution is known for the symmetric case, Fuhrmann & Wang 1988). Several approximate approaches have been proposed in the literature; nearly all of them deal with the K-limited discipline with limit 1 for all the queues

(since the maximum number of workpieces of a particular part type processed in a cycle is limited to one, the K-limited-exhaustive and K-limited-gated variants become identical); the work reported in Ibe & Cheng (1989), Boxma & Meister (1987) and Srinivasan (1988) are representative of the approximate approaches to this special case of the problem. Fuhrmann & Wang (1988) present bounds for the computation of the mean waiting times under the more general situation where the limit on the number served is different for different part types; these bounds are extended to provide approximations to compute the mean waiting times. The survey articles by Takagi (1988, 1990) summarise the recent approximate analyses.

3.1d *Bernoulli scheduling*: The Bernoulli scheduling discipline constitutes a generalisation of the exhaustive and the K-limited-exhaustive service discipline with limit 1 for all part types; when p_i is 1 for all the N part types, it reduces to the exhaustive service, and when $p_i = 0$ for all parts it is the K-limited service discipline of limit 1. In the symmetric case, i.e., all the N part types have completely identical parameters, Tedijanto (1990) has shown that the mean waiting time of a workpiece can be given by

$$E[W] = \frac{1}{2(1-\rho-\lambda R(1-p))} \times \left(\lambda \sum_{j=1}^{N} b^{(2)} + \frac{(1-\rho)\Delta^2}{R} + R(1+\lambda b - 2\lambda bp) \right).$$

3.2 *Cyclic server models with limited buffers*

These models represent one important class, the general version of which is as yet unsolved. However, the single-buffer case, where each part type has a single buffer, has been successfully analysed; the recent results are given by Takine *et al* (1988), and Ibe & Cheng (1989). The solution of the symmetric version requires solving $O(2^{N-1})$ linear equations, and the general (asymmetric) version needs the solution of $O(2^{N-1})$ linear equations. Tran-Gia & Raith (1988) present an approximate analysis for more general systems where each part type may have a finite queue of (distinct) finite capacity, under non-exhaustive service assumptions. Seidmann *et al* (1985) have analysed a related model, in which a single server (a manufacturing cell) processes different types of part types, according to a probabilistic schedule and 1-limited service discipline; but they assume that at each service completion, a new workpiece is always available for each part type, i.e., no arrival process is explicitly considered; this assumption greatly simplifies the analysis.

4. An example

In this section, we consider the application of the various models to an example system – a single versatile machining centre processing three different part types. We restrict our attention to a few models due to space limitations. The mean values of the interarrival duration, setup time and processing time per workpiece for these part types are given in table 2.

Since the assumptions underlying the various models are different, the results produced by these will also differ; we shall concentrate on mean lead time (response time), or, equivalently, the mean inventory level (queue length) for each part type, and the mean utilisation.

Table 2. Parameters for the example.

	Mean time for		
Part type	Arrival	Processing	Setup
1	10·0	3·0	15·0
2	16·0	2·0	10·0
3	20·0	5·0	8·0

4.1 *M/G/1 queueing model*

In this model, the analyst assumes that all part types arrive with exponential interarrival times. All the parts wait in a common, infinite capacity queue from which they are taken up for processing in an FCFS fashion. Setup times are ignored. The mean lead time obtained is for a 'representative' part type. Let us consider the mean response time when the service time of each workpiece is k-stage Erlang with the mean value given above. A simple calculation gives the mean queue lengths when the service time has $k = 1, 2, 3$ stages (table 3).

A simple calculation gives the mean utilisations as $\rho_1 = 0{\cdot}3$, $\rho_2 = 0{\cdot}125$, and $\rho_3 = 0{\cdot}25$, and the total utilisation is $\rho = 0{\cdot}675$. An elementary application of Little's Law gives the mean lead times.

4.2 *GI/G/1 queue models*

We relax the exponential arrival assumption of the previous model and allow the arrival processes also to be GI. Other assumptions remain unchanged. The results are obtained for a typical part type which is representative of all the part types. Assume that the arrival and service processes of the three parts are all 1, 2 or 3 stage Erlang. Table 4 summarises the mean queue length values.

4.3 *Cyclic server queueing models*

Let us take a more detailed view of the example. Let parts 1 and 2 be members of a group technology part family, and part 3 be a member of another group technology part family. The workcentre produces them in the order 1, 2, 3 so that part family 1 is produced first, and then part family 2. Assume that arriving workpieces wait in a dedicated, infinite capacity queue. The processing of each part type is preceded by a setup for that part type. In such a case, the cyclic server multiqueue model of § 3 is

Table 3. Mean queue lengths, M/G/1 model.

Service time is		
1-Stage Erlang (exponential)	2-Stage Erlang	3-Stage Erlang
2·2442	1·85192	1·72115

Table 4. Mean queue lengths, GI/G/1 model.

	Service process		
Arrival process	1-Stage Erlang	2-Stage Erlang	3-Stage Erlang
1-Stage Erlang	2·2442	1·8519	1·7642
2-Stage Erlang	2·0543	1·6083	1·4897
3-Stage Erlang	1·9919	1·5263	1·4126

appropriate. Applying the exhaustive service policy would imply that all existing requirements of production for a particular part type are completed, and the machine changes over only when no further demand is present. Under the gated service policy, we process all outstanding demands for a particular part type that were present at the instant of setup completion. With limited service, we process each part until either a time limit expires or a maximum number has been processed, or there are no further waiting workpieces of that part type.

For simplicity, the setups are assumed to take an exponential duration. Let the machining times be k-Erlang for all parts; we consider $k = 1, 2, 3$. In table 5 we present the mean waiting times (this excludes the processing time) for each part type; the mean lead time is the sum of the mean waiting time and the mean processing time.

The machine utilisation values for the different part types are the same as before. For the 1-limited service discipline, the system is unstable, as an elementary check shows.

4.4 *Cyclic server model with single buffers*

Let there be N fixtures, one fixture each dedicated to each of the part types in the previous example; the infinite buffer assumption is no longer appropriate. We use the

Table 5. Mean waiting times, multiquene model.

Exhaustive service	Service time is		
Part type	1-Stage Erlang	2-Stage Erlang	3-Stage Erlang
1	47·66	45·96	45·39
2	50·98	57·68	56·91
3	51·60	49·81	49·21

Gated service	Service time is		
Part type	1-Stage Erlang	2-Stage Erlang	3-Stage Erlang
1	79·24	77·38	76·76
2	69·28	67·61	67·06
3	77·35	75·43	74·79

Table 6. Multiquene, single buffer model.

Performance measure	Part 1	Part 2	Part 3
Throughput	0·0224	0·0205	0·0188
Maching utilisation	0·0672	0·0411	0·0939
Blocking probability (Mean queue length)	0·7662	0·6712	0·6244

cyclic server model where each part type has exactly one buffer (the fixture). When a fixture is occupied by a workpiece, further arrivals of workpieces of the same type are prohibited (loss system); when the fixture is released, a new arrival can take place. Assume, for simplicity, that all setup and processing times are exponential; also, ignore the fixturing time, or treat it as part of the processing time. Due to the single buffer per part assumption, exhaustive, gated and limited service disciplines become identical. We are interested in parameters like throughput of the parts (parts produced per unit time), the utilisation of the machine and the probability of an arrival finding the corresponding buffer full and hence being lost, for each part type. Under the assumption of independent exponential arrivals for each part type, applying the PASTA property (Wolff 1982), the probability that an arrival finds the corresponding buffers are full (and is hence lost), is the same as the steady state probability that the buffers are full. This result will hold even if the setup and processing activities have general i.i.d. durations; they need not be exponential. For our example, owing to the single buffer assumption, the probability that the buffer is occupied is the same as the mean number of workpieces of a particular part type in the system. We present these performance parameters for our example in table 6.

5. Conclusion

We considered a series of stochastic models for the analysis of a highly capable machining centre processing different types of parts. The issues addressed include using a single aggregate representative part type versus explicitly considering multiple part types, inclusion or non-inclusion of setup times, single common queue versus multiple queues (one for each part type), finite capacity queues versus infinite capacity queues, and exponential versus general arrival and service processes. Table 7 compares the degree of detail captured by each of the two major modelling approaches considered in this paper – (i) the GI/G/1 queue and variants (§ 2), (ii) the multiqueue cyclic server queueing model and variants (§ 3).

From a computational point of view, the GI/G/1 models are the simplest, since performance measures are easily determined by a straightforward computation using an explicit formula. The cyclic server systems (polling models with exponential arrivals) can be analysed by solving $O(N^2)$ linear equations. However, important variants like systems with mixed service, and systems with finite buffer capacities are solved only approximately.

The models considered in this paper can be used to analyse the performance of a single machine under varying demand rates, changing setup or processing times, or varying product mixes. In particular, the effect of introduction of new part types, or

Table 7. Comparison of the modelling approaches.

Issue	GI/G/1 queue	Multiqueue cyclic server
Part types	All part types aggregated to a single representative part type	N distinct parts
Setup time	Ignored	GI (explicit)
No. of queues	One common	One per part type
Queue capacity	Infinite (special cases like M/GI/1/N are solved)	Infinite (approximate solution for finite capacity case)
Arrivals	GI	Exponential (for exact solution)
Processing	GI	GI
Scheduling	FCFS	Cyclic service exactly solved; approximate solutions for others
Service policy	FCFS	Exhaustive, gated, limited

taking some part types out of production, on the lead times of parts already being manufactured can be studied. Besides, these models can form the building blocks of a more elaborate model of an AMS, consisting of several machines and material handling system equipment. Such a model of an AMS, will be typically solved by a decomposition approach (owing to the largeness of the system), by analysing individual equipment (e.g., deriving their flow equivalents) and combining these individual results into those of the entire system. The models proposed in this paper will enable explicit consideration of multiple part types and setup activities. Further, the introduction of the issues of machine failures and repair in the model, in the spirit of queueing systems with server vacations, will enhance the accuracy of the model. Development of solution methodologies for models of the AMS incorporating these features is part of our continuing work in this area.

References

Albin S L 1986 Delays for customers from different arrival streams to a queue. *Manage. Sci.* 32: 329–340

Ancker C Jr, Gafarian A V 1961 Queueing with multiple poisson inputs and exponential service times. *Oper. Res.* 9: 321–327

Basharin G 1965 A single server with a finite queue and items of several types. *Theory Probab. Its Appl. (USSR)* 10: 261–274

Baskett F, Chandy K M, Muntz R R, Palacios F G 1975 Open, closed and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.* 22: 248–260

Bitran G R, Tirupati D 1988 Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference. *Manage. Sci.* 34: 75–100

Bitran G R, Tirupati D 1989 Approximations for product departures from single-server station with batch processing in multi-product queues. *Manage. Sci.* 35: 851–878

Boxma O J, Meister B W 1987 Waiting-time approximations for cyclic-service systems with switchover times. *Performance Eval.* 7: 299–308

Buzacott J A, Shanthikumar J G 1985 On approximate queueing models of dynamic job shops. *Manage. Sci.* 31: 870–887

Dobson G, Karmarkar U S, Rummel J 1987 Batching to minimise flow times on one machine. *Manage. Sci.* 33: 784–799

Fuhrmann S W, Wang Y T 1988 Analysis of cyclic service systems with limited service: Bounds and approximations. *Performance Eval.* 9: 35–54

Gershwin S B 1989 Hierarchical flow control: A framework for scheduling and planning discrete events in manufacturing systems. *Proc. IEEE* 77: 195–209

Ibe O C, Cheng X 1989a Approximate analysis of asymmetric single-service token-passing systems. *IEEE Trans. Commun.* 37: 572–577

Ibe O C, Cheng X 1989b Performance analysis of asymmetric single-buffer polling systems. *Performance Eval.* 10: 1–14

Kelly F P 1979 *Reversibility and stochastic networks* (Chichester: John Wiley and Sons)

Kleinrock L, Levy H 1988 The analysis of random polling systems. *Oper. Res.* 36: 716–732

Kuehn P J 1979 Approximate analysis of general networks by decomposition. *IEEE Trans. Commun.* 27: 113–126

Kumar P R, Seidman T I 1990 Dynamic instabilities and stabilisation methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Autom. Control.* 35: 289–298

Lavenberg S S 1975 The steady-state queueing time distribution for the M/G/1 finite capacity queue. *Manage. Sci.* 21: 501–506

Leachman R C, Gascon A 1988 A heuristic scheduling policy for multi-item single-machine production systems with time-varying, stochastic demands. *Manage. Sci.* 34: 377–390

Marie R A 1979 An approximate analytical method for general queueing networks. *IEEE Trans. Software Eng.* 5: 530–538

Perkins J R, Kumar P R 1989 Stable, distributed, real-time scheduling of flexible manufacturing/assembly/ disassembly systems. *IEEE Trans. Autom. Control* 34: 139–148

Seidmann A, Schweitzer P J, Nof S Y 1985 Performance evaluation of a flexible manufacturing cell with random feedback flow. *Int. J. Product. Res.* 23: 1171–1184

Shanthikumar J G, Buzacott J A 1980 On the approximations to the single server queue. *Int. J. Product. Res.* 18: 761–773

Shanthikumar J G, Buzacott J A 1981 Open queueing network models of dynamic jobshops. *Int. J. Product. Res.* 19: 255–266

Shanthikumar J G, Gocmen M 1983 Heuristic analysis of closed queueing networks. *Int. J. Product. Res.* 21: 675–690

Srinivasan M M 1988 An approximation for mean waiting times in cyclic server systems with nonexhaustive service. *Performance Eval.* 9: 17–33

Takagi H 1988 Queueing analysis of polling models. *ACM Comput. Surv.* 20: 5–28

Takagi H 1990 Analysis of polling systems: An update. *Stochastic analysis of computer and communication systems* (Amsterdam: North Holland)

Takine T, Takahashi Y, Hasegawa T 1988 Exact analysis of asymmetric polling systems with single buffers. *IEEE Trans. Commun.* 36: 1119–1127

Tedijanto 1990 Exact results for the cyclic-service queue with a Bernoulli schedule. *Performance Eval.* 11: 107–115

Tran-Gia P, Raith T 1988 Performance analysis of finite capacity polling systems with nonexhaustive service. *Performance Eval.* 8: 1–16

Whitt W 1983 The queueing network analyser. *Bell Syst. Tech. J.* 62: 2779–2815

Whitt W 1988 A light-traffic approximation for single-class departure process from multi-class queues. *Manage Sci.* 34: 1333–1346

Wolff R W 1982 Poisson arrivals see time averages. *Oper. Res.* 30: 223–231