

ANALYSIS OF FLEXIBLE MANUFACTURING SYSTEMS WITH PRIORITY SCHEDULING: PMVA

S. SHALEV-OREN and A. SEIDMANN

Department of Industrial Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel
and

P.J. SCHWEITZER

*The Graduate School of Management, The University of Rochester, Rochester,
New York 14627, USA*

Abstract

A new methodology for performance analysis of flexible manufacturing systems (FMSs) with priority scheduling is presented. The analytic model developed extends the mean value analysis of closed networks of queues with multiple product types, various non-preemptive priority service disciplines, and with parallel machine stations. Performance measures derived include the expected throughput per product and per station, utilization of machines and transporters, queuing times and queue length measures for various configurations. Extensive numerical calculations have shown that the algorithm used for solving the problem converges rapidly and retains numerical stability for large models. The paper also illustrates the application of the model to a system with a mixture of FCFS and HOL disciplines which gives insights into various priority assignment policies in FMSs. Special attention was given to the problem of scheduling the robot carriers (transporters).

Keywords and phrases

FMS, scheduling, mean value analysis, HOL, performance analysis, queueing networks

1. Introduction

Several studies have presented mathematical models for analyzing product flow in flexible manufacturing systems (FMSs). Most of these studies have used the closed network of queues approach to model the system and to obtain steady-state performance measures. Using closed network of queues, one assumes that a fixed number of items circulate throughout the system in accordance with specified routing requirements. Finished items are immediately replaced by raw items at the load/unload station. Using the closed network of queues approach, it is also possible to model

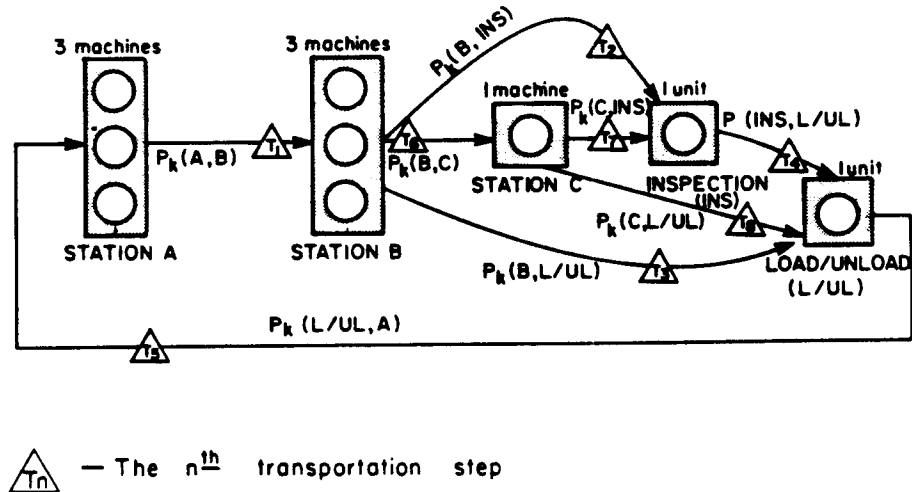


Fig. 1. Logic flow of parts in an FMS.

of one or more identical machines (which corresponds in queuing terminology to parallel servers). The mean service time for each part type at each work station is known. Each station has its own queue of pallets waiting to be processed and the queuing time depends on the priority discipline assigned to that station. Enough space is available for each queue so that blocking is not a problem.

The priority scheme at each station can be either AS (ample-server), FCFS (first come first served), or HOL (head of the line) non-preemptive policy [14]. In HOL, several product types can be assigned the same priority, and the priority level assigned to a specific product type can change from one station to another. Increasing the priority of a given part type decreases its queuing time and its total manufacturing lead times, and increases its throughput (at the expense of other types). This is used to control the number of pallets needed, and to adjust the product mix, machine utilization, and the total flow time per product type.

The model provides steady-state performance measures for both work stations and product types, and predicts major system characteristics such as total throughput and bottlenecks. The analytical model is based on a set of simultaneous equations similar to mean value analysis [20]. In this case, however, additional approximations were developed to handle parallel servers and HOL priority. Unlike product form models (i.e. [22]), it allows multiple product types with distinct processing times. It differs from [4,8,13] by not requiring an open class of customers and by permitting *non-preemptive parallel* stations with several identical machines.

3. System equations and their solution

3.1. INTRODUCTION

This section describes the input parameters to the model, the performance variables employed in the model, the equations satisfied by these variables, and the algorithm for solving these equations. Preliminary observations on the model accuracy and numerical illustrations are provided in sects. 4 and 5, respectively.

3.2. INPUT PARAMETERS

The model is a closed network of queues, with R classes of customers (part types) and $K(r) \geq 1$ customers of class r , $1 \leq r \leq R$. There are M manufacturing stations with $J(m) \geq 1$ parallel identical servers at station m , $1 \leq m \leq M$, all fed by a common queue. A station can be any resource such as machines, inspection units or drying ovens.

$SD(m)$ denotes the service discipline at station m , $1 \leq m \leq M$, with $SD = AS$ or $FCFS$ or HOL . For each station m having HOL discipline, $PR(r, m)$ denotes the priority assigned to customer class r , $1 \leq r \leq R$. Priority 1 is highest, priority 2 the next, etc. We allow several classes of customers to have the same priority at station m ; service is $FCFS$ within each priority. A given customer class may have distinct priorities at different stations. The unit value of class r is given by $UV(r)$.

The service (manufacturing) time of a type r customer at a server in station m is taken as exponentially distributed with mean service time $S(r, m)$, $1 \leq r \leq R$, $1 \leq m \leq M$. The justification for the common use of this assumption is discussed by Solberg [22] and Suri [25].

The routing matrix is given for each part type r , $1 \leq r \leq R$, and from this we calculate the equilibrium distribution $\{P(r, m), 1 \leq m \leq M\}$ of part r 's location at the instants when its transitions occur. The $P(r, m)$ are non-negative and may be scaled arbitrarily. They are proportional to the mean number of visits to station m by part type r [19].

3.3. PERFORMANCE MEASURES

The performance measures computed in the model describe a variety of part types, manufacturing stations and total system variables. The basic variables computed first are:

$G(r, m)$ = throughput of class r parts at station m
(mean number of arrivals or departures per unit time),

$W(r, m)$ = mean time spent by a class r part on queue when it visits station m .

$$T(m) = \sum_{r=1}^R \frac{T(r, m) G(r, m)}{G(m)}$$

= mean sojourn time at station m .

$$N(m) = \sum_{r=1}^R N(r, m)$$

= mean number of parts at station m , in queue and in service.

$$NS(m) = J(m) RO(m) \quad (\leq J(m))$$

= mean number of parts in service at station m .

Systems measures: per class

$$GS(r) = G(r, L/UL)$$

= class r FMS throughput.

$$TFT(r) = K(r)/GS(r)$$

= mean lead time for class r (= mean flow time for class r parts).

$$OV(r) = GS(r) UV(r)$$

= output value per unit time for class r .

$$QS(r) = \sum_{m=1}^M Q(r, m)$$

= mean number of class r parts on queue somewhere.

System measures: aggregate

$$AG = \sum_{r=1}^R GS(r)$$

= system throughput.

FCFS discipline

$$W(r, m) = W_0(r, m) + (1/J(m)) \left(\sum_{t=1}^R G(t, m) W(t, m) S(t, m) - G(r, m) W(r, m) S(r, m)/K(r) \right)$$

$$1 \leq r \leq R, \quad 1 \leq m \leq M, \quad SD(m) = FCFS. \quad (3)$$

Equation (3) expresses the mean queuing delay as the time $W_0(r, m)$ needed to clear one customer from service (if all servers are busy at time of arrival), plus the expected time to clear the queue seen by the last class r arrival to station m . The latter is approximated by the mean workload on queue divided by $J(m)$. This assumes $J(m)$ servers can clear the queue $J(m)$ times as quickly as one server. Similar arguments were used recently by Suri and Hildebrandt [26]. This assumption will be good for stations with heavy utilizations, hence long queues.

The term multiplied by $1/J(m)$ is the standard MVA queue workload estimate including the $1/K(r)$ correction, since the mean queue length observed by an arriving customer of class r to a station is roughly equal to the time average queue length at the station with the arriving customer removed from the system. This correction is exact for $K(r) = 1$ and it is asymptotically correct for the other limiting case of very large $K(r)$ [16].

In computing $W_0(r, m)$, we distinguish between two cases: $J(m) = 1$ and $J(m) > 1$. In the first case, we use the MVA estimate for the residual service time implicit in [20]:

$$W_0(r, m) = \sum_{t=1}^R [G(t, m) S(t, m)^2] - G(r, m) S(r, m)^2 / K(r)$$

$$1 \leq r \leq R, \quad 1 \leq m \leq M, \quad J(m) = 1, \quad SD(m) = FCFS. \quad (4)$$

The $1/K(r)$ expression is the correction term.

For the case $J(m) > 1$, we use

$$W_0(r, m) = B(r, m) DT(r, m)$$

$$1 \leq r \leq R, \quad 1 \leq m \leq M, \quad SD(m) = FCFS, \quad J(m) > 1, \quad (5)$$

$$V(m) = \sum_{t=1}^R [K(t) VT(t, m)] - 1, \tag{7}$$

where

$$VT(t, m) = \begin{cases} 1 & \text{if class } t \text{ visits station } m \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

If P_n is the probability that there are n parts in the station, then the probability that all $J(m)$ servers are busy is therefore

$$B(r, m) = \sum_{n=J(m)}^{V(m)} P_n$$

$$= \frac{U(r, m)^{J(m)}}{J(m)!} \frac{(1 - X(r, m))^{V(m) - J(m) + 1}}{(1 - X(r, m))} + \sum_{t=0}^{J(m)-1} \frac{U(r, m)^t}{t!}$$

$$= \begin{cases} \frac{U(r, m)^{J(m)} (1 - X(r, m))^{V(m) - J(m) + 1}}{J(m)! [1 - X(r, m)] \left\{ \sum_{t=0}^{J(m)-1} \frac{U(r, m)^t}{t!} + \frac{U(r, m)^{J(m)} (1 - X(r, m))^{V(m) - J(m) + 1}}{J(m)! (1 - X(r, m))} \right\}} & \text{if } V(m) \geq J(m) \\ 0 & \text{if } V(m) < J(m) \end{cases} \tag{9}$$

where

$$U(r, m) = \sum_{t=1}^R G(t, m) S(t, m) - G(r, m) S(r, m) / K(r) \tag{10}$$

server stations where every $K(r) \rightarrow \infty$ and when all classes share the same priority, eq. (12) corresponds with the FCFS, single-server MVA equation (3).

The values for $W(r, m)$ were not computed directly from eq. (12) because iterations sometimes lead to utilizations above unity. Since the term $t = r$ on the right-hand side involves $W(r, m)$, eq. (12) represents a *quadratic equation* for $W(r, m)$ and it turned out to be convenient to solve this quadratic equation, obtaining an explicit expression for $W(r, m)$. To do so, we first rewrite eq. (1) as

$$G(r, m) = \frac{K(r)}{W(r, m) + \theta(r, m)}, \quad (13)$$

where

$$\theta(r, m) = S(r, m) + \left\{ \sum_{\substack{t=1 \\ t \neq m}}^M \frac{P(r, t)}{P(r, m)} [S(r, t) + W(r, t)] \right\}$$

= mean time between entry of a class r part into service at station m until its next arrival to the queue at station m . (14)

Then rewrite eq. (12) as

$$W(r, m) = \frac{W_0(r, m)}{D(r, m) [D(r, m) - E(r, m) - F(r, m)]}, \quad (15)$$

where

$$D(r, m) = 1 - \frac{1}{J(m)} \sum_{t=1}^R G(t, m) S(t, m)$$

$$PR(t, m) < PR(r, m) \quad (16)$$

$$E(r, m) = \frac{1}{J(m)} \sum_{\substack{t=1 \\ t \neq r}}^R G(t, m) S(t, m)$$

$$PR(t, m) = PR(r, m) \quad (17)$$

by

$$W(r, t) + S(r, t) + W(r, M + 1) + S(r, M + 1), \quad \text{for } 1 \leq t \leq M.$$

The number of transporters is $J(M + 1)$.

Since the transporters serve all the M manufacturing stations, we augment eq. (1) with the definition:

$$G(r, M + 1) = \sum_{m=1}^M G(r, m).$$

We augment eq. (2) by an equation for $W(r, M + 1)$ which has exactly the same form as the previous equations for $W(r, m)$ in the relevant service disciplines. This model assumes that the $J(M + 1)$ transporters do not interfere with the motions of each other.

4. Solution procedure

Equations (1)–(5) and (21) constitute a set of 2 MR simultaneous non-linear equations for the 2 MR unknowns, the $G(r, m)$'s and the $W(r, m)$'s. They are easily solved by successive substitutions, one iteration of which computes the $\theta(r, m)$'s and the $G(r, m)$'s and then passes sequentially through all M stations, and for each station, passes through all customer classes which can visit this station and computes the $W(r, m)$'s. Convergence is significantly accelerated by scaling the throughput of each center, if necessary, to maintain utilization below 95% (assuming, of course, that no utilization exceeds 95%; if this assumption is incorrect, the algorithm will not converge and the parameter 95% must be increased). Thus the algorithm maintains for all m :

$$\sum_{r=1}^R G(r, m) S(r, m) \leq 0.95 J(m). \quad (24)$$

Applications of eqs. (4) and (5) result in new estimates for $W_0(r, m)$, which are used in the next step of the iteration. Each iteration of the algorithm passes through all M stations, and at each station, computes $W(r, m)$ according to their respective priority discipline; for fixed station m , the algorithm passes through all customer classes r in increasing order of $PR(r, m)$ if $SD(m)$ is HOL, and in increasing order of r otherwise. This accelerates convergence at the HOL stations due to the

the $W_0(r, m)$ estimates since they appear at the numerator of the HOL equations. We also observed that as the population became larger, the estimates became more accurate. The MVA correction term was negligible for large populations, and apparently important for smaller populations.

In addition, several experiments studied the asymptotic behaviour of the *analytic* model at certain extreme cases of the priority disciplines. For example, comparing FCFS with a HOL case where all customers had the same priority, the results agreed within 1% accuracy. In other cases, having few classes with 10 members each and visiting a HOL (or FCFS) station, the classes were partitioned into several subclasses of the same priority. The results for the $W(r, m)$'s of these subclasses agreed with each other and with the original (unpartitioned) class. This showed that the $1/K(r)$ correction appeared to be working properly for small populations.

Finally, we ran HOL and FCFS stations with $J(m)$ parallel servers such that

$$J(m) = \sum_{r=1}^R K(r).$$

Comparing the results with AS discipline revealed deviations of less than 1%.

The analysis of numerical accuracy indicates that the proposed methodology is capable of providing accurate results for a mixture of several operating disciplines with minimal computational effort.

6. Numerical illustration

Consider a system of three manufacturing stations (A, B, C) producing three product types. The system consists of three machines of type A (station 1), two machines of type B (station 2) and one machine of type C (station 3). It uses two transporters for material handling and one load/unload (L/UL) station (station 4). Ten percent of the produced items are randomly sequenced to the inspection station (INS) (station 5) before they reach the load/unload station. Table 1 presents the process requirements (i.e. the routes and service times) for the three product types.

The mean transport time is assumed to be 5 time units. The system operates with 7, 7 and 10 pallets for product types 1, 2 and 3, respectively.

All stations, except the transporters, are FCFS and the HOL priority assignments at the transporters was changed from run to run. The unit values of all parts are equal to one.

Table 2 presents the numerical results of five runs with various service disciplines and priority assignments. It shows that by changing the priorities among the product types, one can obtain significant changes in the throughput and in the output mix while still retaining about the same machine utilization and average total flow

Table 2
The effects of priority changes on the transporters

Run no.	Class r	Priority PR(r , Trans.)	Throughput $G(r) \times 10^4$	Utilization: RO(m)				Lead time ATFT
				A	B	C	Trans.	
1	1	1	373	0.870	0.835	0.811	0.893	229
	2	1	351					
	3	1	324					
2	1	1	440	0.883	0.842	0.633	0.894	224
	2	2	379					
	3	3	253					
3	1	3	270	0.870	0.816	0.876	0.876	236
	2	2	397					
	3	1	351					
4	1	1	497	0.826	0.837	0.846	0.897	229
	2	3	212					
	3	2	338					
5	1	2	349	0.854	0.824	0.879	0.885	233
	2	2	328					
	3	1	351					

simulation model (PSIM); both models refer to run number 1 and run number 5 in table 2. It seems that these PMVA results are slightly conservative in predicting throughputs and station utilization and, as expected, these predictions are more accurate than those for the mean sojourn time and the mean queue length. The simulations were stopped by using a sequential stopping rule that employed batch means analysis to estimate the confidence interval for the estimates of the total times in the system. Once the half-width of the confidence interval was less than 5% of the estimated mean, the simulation was stopped [3].

Table 3(b)
 Comparison of PMVA and PSIM predictions
 (iii) STATION UTILIZATION: $RO(m)$

Station (m)	Run no. 1			Run no. 5		
	PMVA	PSIM	$ \bar{\Delta} $ (%)	PMVA	PSIM	$ \Delta $ (%)
A	0.870	0.871	0.1	0.854	0.864	1.2
B	0.835	0.848	1.5	0.824	0.835	1.3
C	0.811	0.830	2.4	0.879	0.919	4.4
INS	0.402	0.394	2.0	0.393	0.387	1.6
L/UL	0.769	0.782	1.6	0.753	0.764	1.4
TRANS.	0.893	0.898	0.6	0.885	0.890	0.6
			$ \bar{\Delta} = 1.36$ (%)	$ \bar{\Delta} = 1.75$ (%)		

(iv) MEAN QUEUE LENGTH: $Q(m)$

Station (m)	Run no. 1			Run no. 5		
	PMVA	PSIM	$ \Delta $ (%)	PMVA	PSIM	$ \Delta $ (%)
A	3.59	3.52	2.0	3.14	3.21	2.2
B	2.96	3.23	8.4	2.73	3.03	9.9
C	2.19	1.88	16.5	3.33	3.04	8.9
INS	0.25	0.24	4.2	0.24	0.22	9.1
L/UL	2.14	2.41	11.2	1.95	2.11	7.6
TRANS.	4.83	4.62	4.5	4.58	4.20	9.0
			$ \bar{\Delta} = 7.8$ (%)	$ \bar{\Delta} = 7.78$ (%)		

where $G^*(t, m)$ is the average arrival rate of class t parts to m . Considering the mean service times for those arrivals, it yields the expected third phase delay:

$$\frac{W(r, m)}{J(m)} = \sum_{\substack{t=1 \\ \text{PR}(t, m) < \text{PR}(r, m)}}^R G^*(t, m) S(t, m). \quad (\text{A.2})$$

In order to eliminate self-queuing errors, the following approximations are used [21]:

$$G^*(t, m) = \begin{cases} G(t, m) & t \neq r \\ G(t, m) \frac{K(t) - 1}{K(t)} & t = r \end{cases} \quad (\text{A.3})$$

$$Q^*(t, m) = G^*(t, m) W(t, m). \quad (\text{A.4})$$

We can now write the expression for the total mean delay of a class r part as a function of the delays in other classes:

$$\begin{aligned} W(r, m) &= W_0(r, m) + (1/J(m)) \sum_{\substack{t=1 \\ \text{PR}(t, m) < \text{PR}(r, m)}}^R Q^*(t, m) S(t, m) \\ &+ \frac{W(r, m)}{J(m)} \sum_{\substack{t=1 \\ \text{PR}(t, m) < \text{PR}(r, m)}}^R G^*(t, m) S(t, m) \\ &= \left[\frac{W_0(r, m) + \frac{1}{J(m)} \sum_{\substack{t=1 \\ \text{PR}(t, m) < \text{PR}(r, m)}}^R G^*(t, m) W(t, m) S(t, m)}{1 - \frac{1}{J(m)} \sum_{\substack{t=1 \\ \text{PR}(t, m) < \text{PR}(r, m)}}^R G^*(t, m) S(t, m)} \right] \end{aligned} \quad (\text{A.5})$$

Inserting (A.3) into (A.5) and solving for $W(r, m)$, we obtain eq. (12).

- [20] P.J. Schweitzer, Approximate analysis of multiclass closed networks of queues, Int. Conf. on Stochastic Control and Optimization, Free University, Amsterdam (1979).
- [21] P.J. Schweitzer, A. Seidmann and S. Shalev-Oren, The correction term in approximate mean value analysis, Letters of Operations Research (1985), to appear.
- [22] J.J. Solberg, A mathematical model of computerized manufacturing systems, *Proc. 4th Int. Conf. on Production Research*, Tokyo, Japan (1977).
- [23] K.E. Stecke and J.J. Solberg, Loading and control policies for a flexible manufacturing system, *Int. J. of Production Research* 19, 5(1981)481.
- [24] R. Suri, New techniques for modeling and control of flexible automated manufacturing systems, IFAC Meeting, Kyoto, Japan (1981).
- [25] R. Suri, Robustness of queuing network formulas, *J. ACM* 30, 3(1983)564.
- [26] R. Suri and R.R. Hildebrant, Modeling flexible manufacturing systems using mean-value analysis, *Journal of Manufacturing Systems* 1, 3(1984).
- [27] H.A. Taha, *Operations Research* (Macmillan, New York, 1976).