14 Open jobs shop systems

The topic of this chapter is the analysis of job shop systems. A job shop consists of groups of similar machines, linked by a material handling system (for transportation of jobs from one machine group to another). A distinguishing feature of a job shop is that it is capable of processing many different types of jobs, each with its own routing and processing characteristics.

One of advantages of jobs shops is *flexibility* in product mix and product volume. A disadvantage, however, is that the production of many different products typically leads to high variations in processing times and job routing, and thus to long (and unpredictable) production lead times and high levels of work in process (WIP). Hence, the dominant concern in managing job shops is almost always trying to deal with the variety of jobs. Typical issues in the design and control of job shops are, e.g., the required capacity, identification of bottlenecks, and setting delivery dates for incoming orders.

In the following section we start with a simple queueing network model, with only one job type.

14.1 Exponential open queueing network model

We consider a production system consisting of M work stations, numbered $1, 2, \ldots, M$; see figure 1. Work station m has c_m parallel identical machines. The production system is processing one type of jobs, arriving according to a Poisson stream with rate λ. The probability that an arriving job joins work station m is denoted by γ_m (thus $\sum_{m=1}^{M} \gamma_m = 1$).

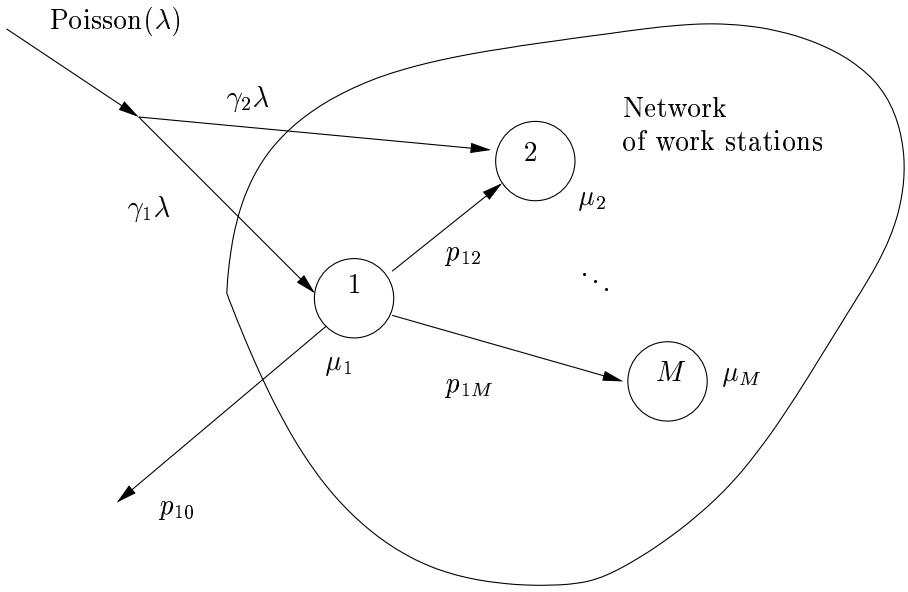

Figure 1: Exponential open queueing network model with M work stations, with c_m machines in station m, $m = 1, 2, \ldots, M$

The processing times at work station m are exponentially distributed with mean $1/\mu_m$, and the processing order is FCFS. The routing of jobs through the system is Markovian: after visiting work station m, a job moves to station n with probability p_{mn} and leaves the system (because all tasks have been completed) with probability p_{m0} (so $\sum_{n=0}^{M} p_{mn} = 1$). Let P be the matrix of routing probabilities p_{mn}, $m, n = 1, \ldots, M$. We assume that P^n tends to 0 as n tends to infinity; this means that each job will eventually leave the network again.

This model is known as an *open Jackson network*; see, e.g., [3, 4]. It is called open, because there is a free inflow of jobs from outside the system.

The first problem is to determine the *capacity of the production network*. This is the maximum number of jobs per time unit that the system is capable to process (or the maximum inflow the system is capable to deal with). Let v_m denote the average number of visits of a job to work station m. Then we have

$$v_m = \gamma_m + \sum_{n=1}^{M} v_n p_{nm}, \qquad m = 1, 2, \ldots, M. \tag{1}$$

This system of equations has a unique solution for $v_1, \ldots, v_M$. So each job has on average v_m/μ_m units work for station m, and thus station m can process at most $c_m\mu_m/v_m$ jobs per time unit. The capacity of the production network is determined by the *bottleneck station*, i.e., the one with the smallest processing capacity. Hence the network capacity is given by

$$\min_{1 \le m \le M} \frac{c_m \mu_m}{v_m}.$$

If the arrival rate λ is equal to or greater than the capacity, then the number of jobs in the system will grow to infinity. From now on we assume that λ is smaller than the capacity, so the network is stable.

Since interarrival times and processing times are assumed to be exponential and the routing is Markovian, this network can be described by a Markov process with states $(k_1, k_2, \ldots, k_M)$ where k_m denotes the number of jobs in work station m. The equilibrium probabilities $p(k_1, k_2, \ldots, k_M)$ exist, since the network is stable. In the following section we first consider the special case in which all stations have exactly one machine. We will derive an explicit form for these probabilities, and based on this result, we can easily obtain mean performance characteristics such as mean number of jobs at the stations and mean production lead times.

Remark 14.1 Note that the stability condition formulated in this section does not depend on exponential processing times and Poisson inflow; it is still valid for general processing and interarrival times. Nor does it depend on Markovian routing; what matters is that p_{mn} denotes the (long-run) fraction of departures from station m that is directed to station n.

14.2 Exponential single-server network

In this section we consider the case where $c_m = 1$ for all m. We first introduce some notation. The vector $(k_1, k_2, \ldots, k_M)$ is denoted by $\underline{k}$, and $\underline{e}_m$ indicates the unity vector

$(0,\ldots,0,1,0,\ldots,0)$ with the one at position m. The function $\epsilon(k)$ is 1 if $k>0$ and 0 otherwise. Then the balance equation in state $\underline{k}$ (flow out is equal to flow in) reads as follows.

$$\begin{aligned} p(\underline{k})\left(\lambda+\sum_{m=1}^{M}\mu_m\epsilon(k_m)\right) &= \sum_{m=1}^{M} p(\underline{k}+\underline{e}_m)\mu_m p_{m0} \\ &\quad +\sum_{n=1}^{M}\sum_{m=1}^{M} p(\underline{k}+\underline{e}_n-\underline{e}_m)\mu_n p_{nm}\epsilon(k_m) \\ &\quad +\sum_{m=1}^{M} p(\underline{k}-\underline{e}_m)\lambda\gamma_m\epsilon(k_m). \end{aligned} \tag{2}$$

The first term at the right-hand side corresponds to a departure from the network, the second one to an internal movement, and the third one to an arrival from outside. As solution we are going to try the form

$$p(\underline{k}) = C x_1^{k_1} x_2^{k_2}\cdots x_M^{k_M}.$$

Substitution of this form into the balance equation (2) and dividing by common powers yields (after rearranging terms)

$$\sum_{m=1}^{M}\left(\mu_m-\sum_{n=1}^{M}\frac{x_n}{x_m}\mu_n p_{nm}-\frac{1}{x_m}\lambda\gamma_m\right)\epsilon(k_m) = \sum_{m=1}^{M} x_m\mu_m p_{m0}-\lambda. \tag{3}$$

The left-hand side is a sum of functions $\epsilon(k_m)$ and the right-hand side is a constant. Thus we can only have equality for all $\underline{k}$ if the coefficients of all $\epsilon(k_m)$ vanish, so the x_m's should satisfy

$$x_m\mu_m = \sum_{n=1}^{M} x_n\mu_n p_{nm}+\lambda\gamma_m, \qquad m=1,2,\ldots,M.$$

If we set $\lambda_m = x_m\mu_m$, then we get

$$\lambda_m = \sum_{n=1}^{M}\lambda_n p_{nm}+\lambda\gamma_m, \qquad m=1,2,\ldots,M. \tag{4}$$

Clearly λ_m is the *total arrival rate* (of internal and external arrivals) to work station m; the above set of equations is very similar to (1) and it has a unique solution, namely $\lambda_m = v_m\lambda$. So x_m is given by

$$x_m = \rho_m = \frac{\lambda_m}{\mu_m}, \qquad m=1,\ldots,M,$$

where ρ_m is the occupation rate of work station m. Finally, since the left-hand side of (3) vanishes, the right-hand side should also vanish. This follows by observing that, when the system is stable, the total inflow is equal to the total outflow, so

$$\lambda = \sum_{m=1}^{M}\rho_m\mu_m p_{m0}.$$

Thus we find that

$$p(\underline{k}) = C\rho_1^{k_1}\rho_2^{k_2}\cdots\rho_M^{k_M},$$

where C follows from normalization. This yields

$$C^{-1} = \sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\cdots\sum_{k_M=0}^{\infty}\rho_1^{k_1}\rho_2^{k_2}\cdots\rho_M^{k_M} = \frac{1}{1-\rho_1}\cdot\frac{1}{1-\rho_2}\cdots\frac{1}{1-\rho_M}.$$

Summarizing, the conclusion is that

$$p(\underline{k}) = p_1(k_1)p_2(k_2)\cdots p_M(k_M), \tag{5}$$

where for $m = 1, 2, \ldots, M$,

$$p_m(k_m) = (1-\rho_m)\rho_m^{k_m}, \qquad k_m = 0, 1, 2, \ldots \tag{6}$$

Solution (5) is a *production form solution*; there is a lot of literature on queueing networks with product form solutions, see, e.g., [1, 2, 5, 6, 7]. The *marginal distribution* $p_m(\cdot)$ of the number of jobs at work station m is exactly the same as the queue length distribution of the $M/M/1$ system with arrival rate λ_m and service rate μ_m. This is a surprise, since in general the inflow at station m is *not Poisson* (see remark 14.2). But, clearly, to find the marginal distribution at station m we may act as if the inflow is Poisson! Another important observation is that the queue length distributions of the work stations are *independent*, since the simultaneous queue length distribution (5) is the product of the marginal distributions.

Remark 14.2 Let us consider the feedback queue in figure 2.

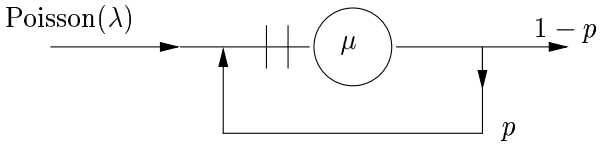

Figure 2: Simple feedback queue with Poisson arrivals, exponential processing times and feedback probability p

It is readily verified that

$$p_k = (1-\rho)\rho^k, \qquad k = 0, 1, 2, \ldots,$$

where

$$\rho = \frac{\lambda}{\mu(1-p)}.$$

The external arrivals are Poisson, but the *total inflow* (of external and feedback arrivals) is not Poisson; take $\mu = 1/\epsilon$, $p = 1-\epsilon$ (so $\mu(1-p) = 1$) and $\lambda \ll 1$. Then the arrival pattern at the workstation looks as in figure 3.

Figure 3: Pattern of clustered arrivals at the feedback queue

From the product form solution (5) we can immediately obtain mean performance characteristics. Let L_m denote the number of jobs at station m and S_m the production lead time at station m. Then we have

$$E(L_m) = \frac{\rho_m}{1-\rho_m}, \qquad E(S_m) = \frac{E(L_m)}{\lambda_m} = \frac{1/\mu_m}{1-\rho_m}, \qquad m = 1, 2, \ldots, M.$$

For the total number of jobs in the system, L, and the total production lead time S it follows that

$$E(L) = \sum_{m=1}^{M} E(L_m) = \sum_{m=1}^{M} \frac{\rho_m}{1-\rho_m}, \qquad E(S) = \frac{E(L)}{\lambda}.$$

Remark 14.3 The product form result is also valid for non-Markovian routing, such as a *fixed* route for jobs through the network. However, if jobs have to visit a work station more than once (possibly for different operations), we have to require that the processing times are exponentially distributed with the *same mean* for each visit.

14.3 Exponential multi-server network

We now extend the results of the previous section for single-server stations to multi- and infinite-server stations. Recall that the queue length probabilities $p(k)$ for an $M/M/c$ system with arrival rate λ and service rate μ are given by

$$p(k) = \begin{cases} \dfrac{1}{k!}(c\rho)^k p(0), & k = 0, 1, \ldots, c-1; \\ \dfrac{1}{c!c^{k-c}}(c\rho)^k p(0), & k = c, c+1, \ldots, \end{cases}$$

where $\rho = \lambda/(c\mu) < 1$ and

$$p(0) = \left(\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!(1-\rho)} \right)^{-1}.$$

In case $c = \infty$ (so there is always a server available) the queue length probabilities are Poisson distributed with mean λ/μ (let c tend to infinity in the expressions above), so

$$p(k) = e^{-\lambda/\mu} \frac{1}{k!} \left(\frac{\lambda}{\mu} \right)^k, \qquad k = 0, 1, 2, \ldots \tag{7}$$

By direct substitution into the balance equations it may be verified that the simultaneous queue length probabilities $p(\underline{k})$ again have a product form solution, i.e.,

$$p(\underline{k}) = p_1(k_1) p_2(k_2) \cdots p_M(k_M), \tag{8}$$

where the marginal queue length probabilities $p_m(k_m)$ at work station m are distributed as the queue length probabilities in the $M/M/c$ with c_m servers, arrival rate λ_m and service rate μ_m. As before the arrival rates λ_m satisfy the set of equations (4).

For the mean number of jobs and the mean production lead time in station m we find, if $c_m < \infty$,

$$E(L_m) = \frac{\Pi_W \rho_m}{1-\rho_m} + \frac{\lambda_m}{\mu_m},$$

and (by Little's law),

$$E(S_m) = \frac{\Pi_W}{1-\rho_m} \cdot \frac{1}{c_m \mu_m} + \frac{1}{\mu_m},$$

where $\rho_m = \lambda_m/(c_m\mu_m)$ and Π_W the probability of waiting in the $M/M/c$ queue with c_m servers, arrival rate λ_m and service rate μ_m. If $c_m = \infty$, these expressions simplify to

$$E(L_m) = \frac{\lambda_m}{\mu_m}, \qquad E(S_m) = \frac{1}{\mu_m}.$$

The mean *total* production lead time can be determined by application of Little's law, or alternatively as

$$E(S) = \sum_{m=1}^{M} v_m E(S_m).$$

Remark 14.4 Expression (7) and product form result (8) remain valid if the processing times in an infinite server station are *generally distributed* (for example, constant processing times). Or, in other words, these results are insensitive to the distribution of the processing times in an $M/M/\infty$ queue

Example 14.5 Let us assume that we have to allocate operators to the M workstations. There are N operators available and they can operate any of the machines. In each work station there are c_m machines, and each operator may be assigned to exactly one machine. Further,

$$M \leq N \leq \sum_{m=1}^{M} c_m.$$

Clearly, when N is strictly less than the right-hand side, we cannot allocate an operator to every machine. The problem is to allocate the operators to the machines such that the mean total number of jobs in the system (or equivalently, the mean total production lead time) is minimized. Let $f_m(c)$ denote the mean number of jobs in an $M/M/c$ queue with arrival rate λ_m and service rate μ_m. Then we have to solve the following optimization problem:

$$\begin{aligned}
&\min \sum_{m=1}^{M} f_m(a_m) \\
&\text{subject to} \\
&\sum_{m=1}^{M} a_m = N, \\
&1 \leq a_m \leq c_m, \quad m = 1, 2, \ldots, M.
\end{aligned}$$

It can be shown that the optimal solution can be found by a *greedy algorithm*: start with assigning to each station exactly one operator and then subsequently add an operator to the station where the maximal reduction in the mean number of jobs is achieved.

14.4 Incorporation of material handling

Infinite server stations are very useful to describe transportation delays in production systems. At least, when there are always sufficiently many transporters for transporting a job from one station to another (so no waiting for transportion occurs). This is, for example, the situation when jobs are transported on a conveyor system.

Suppose the mean transportation time from station m to n is T_{mn} time units. Then, to incorporate transportation in our queueing network model, we add between any two stations m and n, an infinite server station with mean processing time T_{mn}. From the results of the previous section it follows that the number of jobs in transit from station m to n is Poisson distributed with mean $\rho_{mn} = \lambda_m p_{mn} T_{mn}$. Thus

$$P(k \text{ jobs in transit from station } m \text{ to } n) = e^{-\rho_{mn}} \frac{\rho_{mn}^k}{k!}.$$

Since the sum of Poisson random variables is again Poisson, we also find

$$P(k \text{ jobs in transit in the system}) = e^{-\rho_T} \frac{\rho_T^k}{k!},$$

with

$$\rho_T = \sum_{m=1}^{M} \sum_{n=1}^{M} \rho_{mn}.$$

Note that ρ_T is the mean number of jobs in transit in the production system, and by Little's law, the mean total time spent in transit of a job is ρ_T/λ.

14.5 General multi-server network

In this section we consider the situation where the interarrival times and processing times have general (instead of exponential) distributions. The lesson we have learned from the exponential job shops is that each work station can be analyzed *in isolation* with an appropriate arrival process, and these results can be combined to yield the overall performance. In the general setting we are going to adopt this lesson to derive approximate results.

We model each work station m as a $G/G/c_m$ system with interarrival times with mean $1/\lambda_m$ and an appropriate coefficient of variation c_{A_m}. Let $p_m(\cdot)$ denote the (approximate) queue length distribution of this $G/G/c_m$ system, with mean $E(L_m)$. For the overall performance we now obtain

$$\begin{aligned} p(\underline{k}) &\approx p_1(k_1)p_2(k_2)\cdots p_M(k_M), \\ E(L) &\approx E(L_1) + E(L_2) + \cdots + E(L_M), \\ E(S) &\approx E(L)/\lambda. \end{aligned}$$

The remaining problem is to find good estimates for the coefficients of variation c_{A_m}. Here we restrict ourselves to the observation that in *large randomly routed jobs shops* the arrival process in each work station can be safely approximated by a Poisson process (so we may model each station m as an $M/G/c_m$ system).

14.6 General multi-class multi-server network

So far we have looked at job shops processing one job type. Now we are going to treat the situation in which the job shop processes R job types, numbered $1, 2, \ldots, R$. Typically R may be very large. We suppose that each job type requires exactly n_r operations. These operations have to be performed in a fixed (predetermined) sequence. The work station for the first operation is C_{1r} with processing time B_{1r}, the second one is C_{2r} with processing time B_{2r} and so on, up to the n_rth operation. The arrival rate of type r jobs is Λ_r; the total arrival rate is

$$\lambda = \sum_{r=1}^{R} \Lambda_r.$$

For the total flow into (or out of) station m we have

$$\lambda_m = \sum_{r=1}^{R} \Lambda_r \sum_{i=1}^{n_r} I(C_{ir} = m),$$

where $I(C_{ir} = m)$ indicates whether the ith station in the production sequence of a type r job is station m or not, so

$$I(C_{ir} = m) = \begin{cases} 1 & \text{if } C_{ir} = m; \\ 0 & \text{otherwise.} \end{cases}$$

To predict the performance we are going to model each station m as an $M/G/c_m$ system; it is reasonable to approximate the inflow at station m by a Poisson stream, since the number of job types R is large and each type has its own production sequence. Denote the processing time of an *arbitrary job* in station m by B_m; so we have $B_m = B_{ir}$ with probability $\Lambda_r I(C_{ir} = m)/\lambda_m$. Hence,

$$E(B_m) = \frac{1}{\lambda_m} \sum_{r=1}^{R} \sum_{i=1}^{n_r} \Lambda_r I(C_{ir} = m) E(B_{ir}),$$

and

$$E(B_m^2) = \frac{1}{\lambda_m} \sum_{r=1}^{R} \sum_{i=1}^{n_r} \Lambda_r I(C_{ir} = m) E(B_{ir}^2).$$

The workload per machine in station m is given by

$$\rho_m = \lambda_m E(B_m)/c_m,$$

which is assumed to be lesss than 1. Now we have all ingredients to approximate the mean waiting time in workstation m. As approximation we may use

$$E(W_m) = \frac{\Pi_W}{1-\rho_m} \cdot \frac{E(B_m^2)}{2c_m E(B_m)},$$

where Π_W is the probability of waiting in an $M/M/c_m$ with arrival rate λ_m and service rate $1/E(B_m)$. For the mean production lead time $E(S_{ir})$ of the ith operation for a type r job we get

$$E(S_{ir}) = \sum_{m=1}^{M} E(W_m) I(C_{ir} = m) + E(B_{ir}).$$

The mean *total* production lead time $E(S_r)$ of a type r job follows by adding up the mean production lead times for each operation required; so

$$E(S_r) = \sum_{i=1}^{n_r} E(S_{ir}).$$

Finally, by application of Little's law, we obtain for the mean total number of jobs in the system $E(L)$ (or WIP level),

$$E(L) = \sum_{r=1}^{R} \Lambda_r E(S_r).$$

Example 14.6 Consider a production system consisting of two single machine work stations. The system is processing two job types. The processing characteristics for each job type are presented in table 1.

r	Λ_r (jobs/hour)	C_{ir}	$E(B_{ir})$ (min)	$\sigma(B_{ir})$	$E(B_{ir}^2)$
1	3	1,2,1	10,5,6	2,5,2	104,50,40
2	2	2	20	0	400

Table 1: Processing characteristics for job types 1 and 2

In table 2 we translate the characteristics above to the processing time characteristics of an arbitrary (or aggregate) job processed by machine 1 and 2, respectively.

m	λ_m (jobs/hour)	$E(B_m)$ (min)	$E(B_m^2)$	ρ_m
1	6	8	72	0.80
2	5	11	190	0.92

Table 2: Processing characteristics for an arbitrary job in station 1 and 2

Hence, the mean waiting time at workstation 1 and 2 may be approximated by

$$E(W_1) = \frac{0.8}{0.2} \cdot \frac{72}{2 \cdot 8} = 18 \text{ (min)}, \qquad E(W_2) = \frac{0.92}{0.08} \cdot \frac{190}{2 \cdot 11} = 99.3 \text{ (min)}.$$

For the mean total production lead time of a type 1 job we get

$$E(S_1) = (18 + 10) + (99.3 + 5) + (18 + 6) = 156.3 \text{ (min)}$$

and for a type 2 job,

$$E(S_2) = 99.3 + 20 = 119.3 \text{ (min)}.$$

References

[1] F. Baskett, K.M. Chandy, R.R. Muntz, F.G. Palacios, *Open, closed, and mixed networks of queues with different classes of customers*, J. ACM, 22 (1975), pp. 248–260.

[2] N.M. van Dijk, *Queueing networks and product forms: a systems approach*, Wiley, 1993.

[3] J.R. Jackson, *Networks of waiting lines*, Oper. Res., 5 (1957), pp. 518–521.

[4] J.R. Jackson, *Networks of waiting lines*, Mgmt. Sci., 10 (1963), pp. 131–142.

[5] F.P. Kelly, *Reversibility and stochastic networks*, Wiley, 1979.

[6] S.S. Lavenberg (ed.), *Computer Performance Modeling Handbook*, Academic Press, 1983.

[7] J. walrand, *An introduction to queueing networks*, Prentice-Hall, 1988.