11 Multi-machine systems

In the previous chapters we discussed single-machine systems. This chapter is devoted to systems consisting of a group of multiple but identical machines. Each job requires service from only one of the machines. There are several ways in which these parallel machine systems can be controlled. The most common situation is the one in which in front of the machines there is one common (central) queue. As soon as one of the machines completes a job it picks a job from the queue according to the FCFS rule.

Another situation is the one in which the machines are, in some sense, at different locations. In that case it might be more attractive to have a (local) queue of jobs for any one of the machines separately. Of course, then one needs a rule to decide to which queue a job has to be sent. The most natural rule would be the *shortest queue* rule.

A far more difficult system is the one in which the machines are identical, but to be able to process a job the machine has to be tooled with the right parts or equipment. So if the jobs belong to a number of classes, which differ with respect to the tools required, each job can only be processed by a subset of the set of all machines. For these systems, the decision which job to do next may seriously affect the performance. Of course, as result of the tooling, the machines are not really identical anymore.

In this chapter we treat the simplest case of truely identical machines with one common queue. In the next section we consider the system with Poisson arrivals and exponential production times, i.e., the $M/M/c$ system. In the sections 11.2 and 11.3, we will present some approximations for $M/G/c$ and $G/G/c$ systems. Finally, in section 11.4 we look at the effect of machine pooling, and in section 11.5 we analyze a system with two non-identical machines.

11.1 The $M/M/c$ queue

In this section we will analyze the model with exponential interarrival times with mean $1/\lambda$, exponential production times with mean $1/\mu$ and c parallel, identical machines. Jobs are served in order of arrival. We suppose that the occupation rate per machine,

$$\rho = \frac{\lambda}{c\mu},$$

is smaller than one. The state of the system is completely characterized by the number of jobs in the system. Let p_n denote the equilibrium probability that there are n jobs in the system. Similar as for the $M/M/1$ system, we can derive the equilibrium equations for the probabilities p_n from the flow diagram shown in figure 1.

Instead of equating the flow into and out of a single state n, we get simpler equations by equating the flow out of and into the set of states $\{0, 1, \ldots, n-1\}$. This amounts to equating the flow between the two neighboring states $n-1$ and n yielding

$$\lambda p_{n-1} = \min(n, c)\mu p_n, \qquad n = 1, 2, \ldots$$

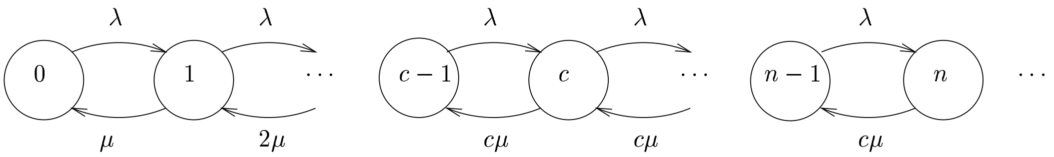

Figure 1: Flow diagram for the $M/M/c$ model

Iterating gives

$$p_n = \frac{(c\rho)^n}{n!} p_0, \qquad n = 0, \ldots, c$$

and

$$p_{c+n} = \rho^n p_c = \rho^n \frac{(c\rho)^c}{c!} p_0, \qquad n = 0, 1, 2, \ldots$$

The probability p_0 follows from normalization, yielding

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!} \cdot \frac{1}{1-\rho} \right)^{-1} .$$

An important quantity is the probability that a job has to wait. Denote this probability by Π_W. By PASTA it follows that

$$\begin{aligned} \Pi_W &= p_c + p_{c+1} + p_{c+2} + \cdots \\ &= \frac{p_c}{1-\rho} \\ &= \frac{(c\rho)^c}{c!} \left((1-\rho) \sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!} \right)^{-1} . \end{aligned} \tag{1}$$

From the equilibrium probabilities we directly obtain for the mean queue length,

$$\begin{aligned} E(L^q) &= \sum_{n=0}^{\infty} n p_{c+n} \\ &= \frac{p_c}{1-\rho} \sum_{n=0}^{\infty} n(1-\rho)\rho^n \\ &= \Pi_W \cdot \frac{\rho}{1-\rho} , \end{aligned} \tag{2}$$

and then from Little's law,

$$E(W) = \Pi_W \cdot \frac{1}{1-\rho} \cdot \frac{1}{c\mu}. \tag{3}$$

These formulas for $E(L^q)$ and $E(W)$ can also be found by using the mean value technique. If not all machines are busy on arrival the waiting time is zero. If all machines are busy and there are zero or more jobs waiting, then a new arriving job first has to wait until the first departure and then continues to wait for as many departures as there were jobs waiting upon arrival. An interdeparture time is the minimum of c exponential (residual) production times with mean $1/\mu$, and thus it is exponential with mean $1/c\mu$. So we obtain

$$E(W) = \Pi_W \frac{1}{c\mu} + E(L^q)\frac{1}{c\mu}.$$

Together with Little's law we retrieve the formulas (2)–(3). Table 1 lists the waiting probability Π_W and the mean waiting time $E(W)$ in an $M/M/c$ with mean production time 1 for $\rho = 0.9$.

c	Π_W	$E(W)$
1	0.90	9.00
2	0.85	4.26
5	0.76	1.53
10	0.67	0.67
20	0.55	0.28

Table 1: Performance characteristics for the $M/M/c$ with $\mu = 1$ and $\rho = 0.9$

We see that the waiting probability slowly decreases as c increases. The mean waiting time however decreases fast (a little faster than $1/c$). One can also look somewhat differently at the performance of the system. We do not look at the occupation rate of a machine, but at the average number of idle machines. Let us call this the surplus capacity. Table 2 shows for fixed surplus capacity (instead of for fixed occupation rate as in the previous table) and c varying from 1 to 20 the mean waiting time and the mean number of customers in the system.

c	ρ	$E(W)$	$E(L)$
1	0.90	9.00	9
2	0.95	9.26	19
5	0.98	9.50	51
10	0.99	9.64	105
20	0.995	9.74	214

Table 2: Performance characteristics for fixed surplus capacity of 0.1 server

Although the mean number of jobs in the system sharply increases, the mean waiting time remains nearly constant.

The derivation of the distribution of the waiting time is very similar to the one for the $M/M/1$. By conditioning on the state seen on arrival we obtain

$$P(W > t) = \sum_{n=0}^{\infty} P(\sum_{k=1}^{n+1} D_k > t)p_{c+n},$$

where D_k is the kth interdeparture time. Clearly, the random variables D_k are independent and exponentially distributed with mean $1/c\mu$. Hence, we find

$$\begin{aligned} P(W > t) &= \sum_{n=0}^{\infty}\sum_{k=0}^{n} \frac{(c\mu t)^k}{k!} e^{-c\mu t} p_c \rho^n \\ &= \sum_{k=0}^{\infty}\sum_{n=k}^{\infty} \frac{(c\mu t)^k}{k!} e^{-c\mu t} p_c \rho^n \\ &= \frac{p_c}{1-\rho} \sum_{k=0}^{\infty} \frac{(c\mu\rho t)^k}{k!} e^{-c\mu t} \\ &= \Pi_W e^{-c\mu(1-\rho)t}, \qquad t \geq 0. \end{aligned}$$

This yields for the conditional waiting time,

$$P(W > t|W > 0) = \frac{P(W > t)}{P(W > 0)} = e^{-c\mu(1-\rho)t}, \qquad t \geq 0.$$

Hence, the conditional waiting time $W|W > 0$ is exponentially distributed with parameter $c\mu(1-\rho)$.

We finally mention that it can be shown that the departure process of an $M/M/c$ is again Poisson with rate λ.

11.2 The $M/G/c$ queue

For the general $M/G/c$ queue with $c > 1$, no exact results for performance measures like the mean waiting time are available. Only for special types of production time distributions these exact results exist. Fortunately, however, good approximations do exist. These approximations are based on the following two facts:

1. The waiting probability Π_W in the $M/G/c$ system only slightly differs from the waiting probability Π_W in the $M/M/c$ system with the same occupation rate. The latter is known as we have seen in the previous subsection.

2. It is possible to obtain reasonably good approximations for the conditional mean waiting time $E(W\,|\,W > 0)$. These approximations are based on approximations for the time until the next job completes.

For example, a very good approximation for the mean waiting time can be obtained by using Little's formula

$$E(L^q) = \lambda E(W), \tag{4}$$

together with the approximate arrival relation

$$E(W) \approx \Pi_W \cdot \frac{E(R)}{c} + E(L^q) \cdot \frac{E(B)}{c}, \tag{5}$$

where Π_W is the waiting probability in the $M/M/c$ queue with arrival rate λ and mean processing time $1/\mu = E(B)$. In the above arrival relation we assumed, as an approximation, that with c machines the time to clear the queue is c times smaller than with one machine. Combination of (4) and (5) gives the following approximation for the mean waiting time

$$E(W) \approx \frac{\Pi_W \cdot E(R)/c}{1-\rho} = \frac{\Pi_W}{1-\rho} \cdot \frac{1+c_B^2}{2} \cdot \frac{E(B)}{c}, \tag{6}$$

with $\rho = \lambda E(B)/c < 1$. Table 11.2 shows the quality of the approximation for the $M/E_k/c$ queue, i.e., the system with Erlang-k distributed processing times. In all numerical examples we have set $E(B) = 1$.

			$E(W)$	
ρ	k	c	exact	approx
0.2	3	2	0.03	0.028
0.5			0.23	0.22
0.9			2.86	2.84
0.2	3	4	0.0023	0.002
0.5			0.062	0.058
0.9			1.33	1.31
0.5	5	4	0.057	0.052
0.9			1.2	1.18
0.95			2.69	2.67

Table 3: Comparison of the approximation of the mean waiting time with exact results in the $M/E_k/c$ queue.

11.3 The $G/G/c$ queue

For the $G/G/c$ queue one can, for example, use the following approximation for the mean waiting time

$$E(W_{G/G/c}) \approx \frac{E(W_{M/M/c})}{E(W_{M/M/1})} \cdot E(W_{G/G/1}).$$

Here, $E(W_{M/M/c})$ is the mean waiting time in the $M/M/c$ queue with the same arrival rate and the same mean processing time. Furthermore, $E(W_{G/G/1})$ (and similarly $E(W_{M/M/1})$) is the mean waiting time in the system where the c machines are replaced by 1 supermachine, working c times as fast as the original machines. If we substitute the expressions for $E(W_{M/M/c})$ and $E(W_{M/M/1})$ and expression (2) in section 10.2 for $E(W_{G/G/1})$ we get

$$E(W_{G/G/c}) \approx \frac{\Pi_W}{1-\rho} \cdot \frac{c_A^2 + c_B^2}{2} \cdot \frac{E(B)}{c},$$

where $\rho = \lambda E(B)/c$ and Π_W the probability of waiting in the corresponding $M/M/c$ queue. For the special case of Poisson arrivals the above approximation is the same as the one presented in the previous section for the $M/G/c$ system.

In approximating the departure process of the $G/G/c$ system we again act as if the interdepartures times are independent. The mean interdeparture time is equal to the mean interarrival time (by conservation of flow) and the squared coefficient of variation c_D^2 is approximated by

$$c_D^2 \approx 1 + (1-\rho^2)(c_A^2 - 1) + \frac{\rho^2(c_B^2 - 1)}{\sqrt{c}}.$$

For $c = 1$ this approximation is the same as the one proposed for the $G/G/1$ system, and for the $M/M/c$ it yields $c_D^2 = 1$ (which agrees with the property that the output of the $M/M/c$ is again Poisson).

11.4 Pooling

Now we consider a production system with two parallel machines processing two job types, type 1 and 2. The jobs arrive according to Poisson streams with rate λ_1 and λ_2, respectively. The processing times are exponentially distributed with mean $1/\mu_1$ for type 2 jobs, and with mean $1/\mu_2$ for type 2 jobs. Hence, the mean overall processing time is given by

$$\frac{\lambda_1}{\lambda}\frac{1}{\mu_1} + \frac{\lambda_2}{\lambda}\frac{1}{\mu_2} = \frac{\rho_1 + \rho_2}{\lambda},$$

where $\lambda = \lambda_1 + \lambda_2$. The question is whether it is better to dedicate each machine to one job type or to pool the machines and use them for both types. In the first case we have two $M/M/1$ models. This option is only sensible if the amount of work offered per time unit to each of the machines is less than one, so we require that

$$\rho_1 = \frac{\lambda_1}{\mu_1} < 1, \qquad \rho_2 = \frac{\lambda_2}{\mu_2} < 1.$$

For the dedicated system the mean overall throughput time is given by

$$E(S) = \frac{1}{\lambda}\left(\frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2}\right). \tag{7}$$

Note that it may happen that one machine is idle, while there are jobs waiting at the other machine. So the production capacity is not optimally used.

Let us now consider situation where the two machines are pooled, i.e. they are both used for processing of type 1 and type 2 jobs. Clearly, in this case it will never happen that one of the machines is idle while there are jobs waiting. The pooled system can be modelled as a system with two parallel identical machines where one stream of jobs arrives with rate λ (we merge the type 1 and type 2 streams). When a machine starts processing a job, it is with probability p_1 a type 1 job with mean processing time $1/\mu_1$ and with probability p_2 a type 2 job with mean processing time $1/\mu_2$, where

$$p_1 = \frac{\lambda_1}{\lambda}, \qquad p_2 = \frac{\lambda_2}{\lambda}.$$

So the processing times are hyperexponentially distributed, with mean $(\rho_1 + \rho_2)/\lambda$. It is easily verified that the squared coefficient of variation of the processing time of an arbitrary job is equal to

$$1 + \frac{2\rho_1\rho_2}{(\rho_1+\rho_2)^2}\left(\frac{\mu_1}{\mu_2} + \frac{\mu_2}{\mu_1} - 2\right),$$

so it approximately increases linearly in μ_1/μ_2, provided μ_1/μ_2 is (much) greater than 1. The mean throughput time in the pooled system can be approximated by (see (6))

$$E(S) \approx \frac{\Pi_W}{1-\rho}\left(1 + \frac{\rho_1\rho_2}{(\rho_1+\rho_2)^2}\left(\frac{\mu_1}{\mu_2} + \frac{\mu_2}{\mu_1} - 2\right)\right)\frac{\rho_1+\rho_2}{2\lambda} + \frac{\rho_1+\rho_2}{\lambda},$$

where $\rho = (\rho_1 + \rho_2)/2$ and Π_W is the probability of waiting in the $M/M/2$ system with arrival rate λ and service rate $\lambda/(\rho_1 + \rho_2)$.

ρ_1	ρ_2	μ_1/μ_2	λ_1	λ_2	μ_1	μ_2	$E(S)$	
							Pooled	Dedicated
0.8	0.8	1	0.80	0.80	1.00	1.00	2.8	5.0
		2	1.07	0.53	1.33	0.67	3.0	
		5	1.33	0.27	1.67	0.33	4.2	
		10	1.45	0.15	1.82	0.18	6.4	
		15	1.50	0.10	1.88	0.13	8.6	
0.9	0.7	1	0.90	0.70	1.00	1.00	2.8	7.1
		2	1.15	0.45	1.28	0.64	3.0	
		5	1.38	0.22	1.54	0.31	4.2	
		10	1.48	0.12	1.65	0.16	6.3	
		15	1.52	0.08	1.69	0.11	8.5	

Table 4: Comparison of the mean throughput time in the dedicated and the pooled system; in each example the mean overall processing time is set to 1.

In table 11.4 we list the mean throughput time for the dedicated and the pooled system for various values of the utilization rates ρ_1 and ρ_2 and the ratio μ_1/μ_2. Note that, by setting the mean overall processing time to 1, we have $\lambda = \rho_1 + \rho_2$. Thus it follows from (7) that the mean throughput time in the dedicated system only depends on ρ_1 and ρ_2 (and not on μ_1/μ_2). The results in table 11.4 show that in most cases the pooled system leads to smaller throughput times than in the dedicated system, *except for large values of* μ_1/μ_2.

This may be expected, because pooling makes better use of the capacity. However, when μ_1/μ_2 is large, there will be many small jobs and a few very big ones. So if the machines are pooled, it can occur that both machines are occupied by very big jobs, while there are a lot of small ones waiting. In this situation it may be better to dedicate the machines to one job type.

11.5 The $M/M/c$ queue with non-preemptive priorities

In this section we consider an $M/M/c$ system processing different types of jobs. To keep it simple we suppose that there are two types only, type 1 and 2 say, but the analysis can easily be extended the situation with more types of jobs. Type 1 and type 2 jobs arrive according to independent Poisson processes with rate λ_1, and λ_2 respectively. The processing times of all jobs are exponentially distributed with the same mean $1/\mu$. We assume that

$$\rho = \rho_1 + \rho_2 < 1,$$

where $\rho_i = \lambda_i/(c\mu)$, i.e. the occupation rate per machine due to type i jobs. Type 1 jobs are treated with non-preemptive priority over type 2 jobs; i.e., type 1 jobs have priority, but they may not interrupt the processing of type 2 jobs.

For the mean waiting time $E(W_1)$ of type 1 jobs we have

$$E(W_1) = \Pi_W \frac{1}{c\mu} + E(L_1^q)\frac{1}{c\mu},$$

where Π_W is the probability of waiting in the $M/M/c$ with no priorities (note that, for the probability that all machines are busy, it is not relevant in which order jobs are processed, since the processing times do not depend on the job type). Together with Little's law,

$$E(L_1^q) = \lambda_1 E(W_1),$$

we get

$$E(W_1) = \frac{\Pi_W}{1-\rho_1} \cdot \frac{1}{c\mu},$$

and

$$E(L_1^q) = \frac{\Pi_W \rho_1}{1-\rho_1}. \tag{8}$$

Since the processing times of all jobs are exponentially distributed with the same mean, it follows that the total number of waiting jobs does not depend on the order in which the jobs are served. So this number is the same as in the system where all jobs are served in order of arrival. Hence,

$$E(L_1^q) + E(L_2^q) = \frac{\Pi_W \rho}{1-\rho},$$

and thus, by inserting (8), we find

$$E(L_2^q) = \frac{\Pi_W \rho_1}{(1-\rho)(1-\rho_1)}.$$

By using Little's law this yields

$$E(W_2) = \frac{E(L_2^q)}{\lambda_2} = \frac{\Pi_W}{(1-\rho)(1-\rho_1)} \cdot \frac{1}{c\mu}.$$

Note that

$$\frac{E(W_1)}{E(W_2)} = 1 - \rho,$$

which does not depend on how the total flow of jobs is split into a low and high priority group. The conditional wating times $(W_i|W_i > 0)$ can be approximated by exponential distributions, with their means matching

$$E(W_i|W_i > 0) = \frac{E(W_i)}{\Pi_W} =: \frac{1}{\tau_i}, \qquad i = 1, 2.$$

Thus the (unconditional) waiting time distribution may be approximated by

$$P(W_i > t) = \Pi_W P(W_i > t|W_i > 0) = \Pi_W e^{-\tau_i t}, \qquad t \geq 0. \tag{9}$$

We finally note the mean waiting times in case of $r(\geq 2)$ job types are given by

$$E(W_i) = \frac{\Pi_W}{(1 - \sum_{j=1}^{i} \rho_j)(1 - \sum_{j=1}^{i-1} \rho_j)} \cdot \frac{1}{c\mu}, \qquad i = 1, 2, \ldots, r.$$

Remark 11.1 For the highest priority jobs, result (9) is exact.

11.6 The $M/G/c$ queue with non-preemptive priorities

The (approximative) analysis of the $M/G/c$ queue with non-preemptive priorities is very similar to the analysis of the system with exponential processing times. Let us consider the situation with two types, both arriving according to Poisson streams with rate λ_1, and λ_2 respectively. The processing times of all jobs are generally distributed with the same mean $E(B)$. Denote the mean residual processing time by $E(R)$, and assume that

$$\rho = \rho_1 + \rho_2 < 1,$$

where $\rho_i = \lambda_i E(B)/c$. Type 1 jobs have non-preemptive priority over type 2 jobs. For the mean waiting time $E(W_1)$ of type 1 jobs we have, as an approximation,

$$E(W_1) = \Pi_W \frac{E(R)}{c} + E(L_1^q) \frac{E(B)}{c},$$

where Π_W is the probability of waiting in the corresponding $M/M/c$ with no priorities (i.e., the $M/M/c$ with arrival rate λ and service rate $\mu = 1/E(B)$). Then, together with Little's law, we get

$$E(W_1) = \frac{\Pi_W}{1 - \rho_1} \cdot \frac{E(R)}{c},$$

In exactly the same way as in the previous section we obtain the following approximation for the mean waiting time of the low priority jobs,

$$E(W_2) = \frac{\Pi_W}{(1-\rho)(1-\rho_1)} \cdot \frac{E(R)}{c}.$$

And also, the (conditional) waiting time distributions of the low and high priority jobs may be approximated by exponential distributions.

11.7 Non-identical machines

So far we considered systems with identical parallel machines. In this section we will study a system with two different machines, a fast and a slow one. Jobs arrive according to a Poisson stream with rate λ. The processing times are exponentially distributed with mean $1/\mu_1$ on machine 1 and $1/\mu_2$ on machine 2 ($\mu_1 > \mu_2$). Jobs are processed in order of arrival. A job arriving when both machines are idle is assigned to the fast machine. We assume that

$$\rho = \frac{\lambda}{\mu_1 + \mu_2} < 1.$$

As state description we use the number of jobs in the system, and if the number of jobs is equal to 1, we distinguish between state $(1, f)$ in which the fast server is working and state $(1, s)$ in which the slow server is working. Then it is readily verified that balance equations are solved by

$$\begin{aligned} p_0 &= \frac{1-\rho}{1-\rho+C}\,, \\ p_1 &= p_{1,f} + p_{1,s} = C p_0\,, \\ p_n &= \rho^{n-1} p_1\,, \quad n > 1, \end{aligned}$$

where we used the notation $\mu = \mu_1 + \mu_2$, $\rho = \lambda/\mu$ and

$$C = \frac{\lambda\mu(\lambda + \mu_2)}{\mu_1\mu_2(2\lambda + \mu)}.$$

For the mean number of jobs in the system we find

$$E(L) = \sum_{n=1}^{\infty} n p_n = \frac{C}{(1-\rho)(1-\rho+C)},$$

from which mean throughput time can be obtained by applying Little's law. In table 11.7 we list the mean number in the system for various values of ρ and μ_2/μ_1. The results show that some difference in machine speed may reduce the mean throughput time, but when the difference becomes too large, the mean performance will detoriate. Also, a lightly loaded system seems more sensitive to differences in machine speed than a heavily loaded system.

ρ	μ_2/μ_1	$E(L)$
0.5	1.0	1.33
	0.5	1.30
	0.1	1.56
0.8	1.0	4.44
	0.5	4.43
	0.1	4.72
0.9	1.0	9.47
	0.5	9.47
	0.1	9.75

Table 5: Mean number in the system.

It is interesting to investigate whether it is better not to use that slower machine at all, assuming of course that $\lambda < \mu_1$. Let $E(L^f)$ denote the mean number of jobs in the system that only uses the fast machine, so

$$E(L^f) = \rho_1/(1-\rho_1),$$

where $\rho_1 = \lambda/\mu_1 < 1$. In case $\mu_1 = 5$ and $\mu_2 = 1$, we get for $\lambda = 3$,

$$E(L^f) = \frac{3}{2} > \frac{24}{17} = E(L),$$

but when $\lambda = 2$,

$$E(L^f) = \frac{2}{3} < \frac{81}{104} = E(L).$$

In [1] it is shown that one should not remove the slow machine if $r > 0.5$ where $r = \mu_2/\mu_1$. But when $0 \leq r < 0.5$ the slow machine should be removed (and the resulting system is stable) whenever $\rho \leq \rho_c$, where

$$\rho_c = \frac{2 + r^2 - \sqrt{(2+r^2)^2 + 4(1+r^2)(2r-1)(1+r)}}{2(1+r^2)}.$$

For example, for $r = 0.4$ the critical value ρ_c is equal to 0.14, and for $r = 0.1$ it is 0.65.

References

[1] M. Rubinovitch, *The slow server problem*, J. Appl. Prob., 22 (1985), pp. 205–213.