The analysis and the results of the previous two models (the $M/M/1/K$ queue and the $M/M/1$ queue) can be extended to models with more than one server.

We will study the following models:

- The $M/M/s/K$ queue;
- The $M/M/s$ queue;
- The $M/M/\infty$ queue.

The $M/M/s/K$ queue

- Customers arrive according to a Poisson process with rate λ.
- The service times of customers are exponentially distributed with parameter μ.
- There are s servers, serving customers in order of arrival.
- Customers who see at arrival K ($K \geq s$) other customers in the system are lost.

The process $\{X(t), t \geq 0\}$, the number of customers in the system at time t, is again a continuous-time Markov chain with state space $\{0, 1, \ldots, K\}$.

The ‘cut equations’ are given by

$$
\begin{aligned}
\lambda p_{i-1} &= i\mu p_i, \qquad i = 1, \ldots, s, \\
\lambda p_{i-1} &= s\mu p_i, \qquad i = s+1, \ldots, K.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
p_i &= \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} p_0, \qquad i = 0, \ldots, s, \\
p_{s+k} &= \left(\frac{\lambda}{s\mu}\right)^k p_s = \left(\frac{\lambda}{s\mu}\right)^k \left(\frac{\lambda}{\mu}\right)^s \frac{1}{s!} p_0, \qquad k = 0, \ldots, K-s.
\end{aligned}
$$

Finally, from the normalization equation $\sum_{i=0}^{K} p_i = 1$ one can determine the unknown p_0.

Again, from the limiting distribution several long-run performance measures can be calculated.

The $M/M/s$ queue

- Customers arrive according to a Poisson process with rate λ.
- The service times of customers are exponentially distributed with parameter μ.
- There are s servers, serving customers in order of arrival.

Stability condition:

$$\lambda < s \cdot \mu \qquad \text{or alternatively written,} \qquad \rho = \frac{\lambda}{s \cdot \mu} < 1.$$

The process $\{X(t), t \geq 0\}$, the number of customers in the system at time t, is again a continuous-time Markov chain with infinite state space.

The 'cut equations' are given by

$$\begin{aligned}\lambda p_{i-1} &= i\mu p_i, \qquad i = 1, \ldots, s,\\ \lambda p_{i-1} &= s\mu p_i, \qquad i = s+1, \ldots.\end{aligned}$$

Hence,

$$\begin{aligned}p_i &= \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} p_0, \qquad i = 0, \ldots, s,\\ p_{s+k} &= \left(\frac{\lambda}{s\mu}\right)^k p_s = \left(\frac{\lambda}{s\mu}\right)^k \left(\frac{\lambda}{\mu}\right)^s \frac{1}{s!} p_0, \qquad k = 0, \ldots.\end{aligned}$$

Finally, from the normalization equation $\sum_{i=0}^{\infty} p_i = 1$ one can determine the unknown p_0.

Again, from the limiting distribution several long-run performance measures can be calculated.

Performance measures in the $M/M/s$ queue:

$$\begin{aligned}\Pi_W &= \text{probability that a customer has to wait,} \\ &= \sum_{k=0}^{\infty} p_{s+k} = \sum_{k=0}^{\infty} \left(\frac{\lambda}{s\mu}\right)^k p_s = \frac{p_s}{1-\rho}\end{aligned}$$

$$\begin{aligned}B &= \text{expected number of busy servers,} \\ &= \sum_{i=1}^{\infty} \min(i,s)\, p_i = \sum_{i=1}^{\infty} \left(\frac{\lambda}{\mu}\right) p_{i-1} = \frac{\lambda}{\mu}\end{aligned}$$

$$\begin{aligned}L_q &= \text{expected number of waiting customers,} \\ &= \sum_{k=0}^{\infty} k p_{s+k} = \sum_{k=0}^{\infty} k \left(\frac{\lambda}{s\mu}\right)^k p_s = p_s \frac{\rho}{(1-\rho)^2}\end{aligned}$$

$$W_q = L_q/\lambda, \quad L = L_q + B, \quad W = L/\lambda = W_q + 1/\mu$$

The $M/M/\infty$ model

- Customers arrive according to a Poisson process with rate λ.
- The service times of customers are exponentially distributed with parameter μ.
- There is an infinite number of servers, serving the customers.

  (Hence, all customers go immediately into service upon arrival)

The process $\{X(t), t \geq 0\}$, the number of customers in the system at time t, is again a continuous-time Markov chain with infinite state space.

The limiting distribution is given by

$$p_i = \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}\, e^{-\lambda/\mu}, \qquad i = 0, \ldots .$$

Remark that this is a Poisson distribution with parameter λ/μ.

The $M/G/1$ queue

In many applications, the assumption of exponentially distributed service times is not realistic (e.g., in production systems). Therefore, we will now look at a model with *generally* distributed service times.

Model:

- Arrival process is a Poisson process with rate λ.
- Service times of customers $(Y_1, Y_2, \ldots)$ are identically distributed with an arbitrary distribution function.

  Mean service time: $E(Y_1) = \tau$.

  Variance of the service time: $E((Y_1 - E(Y_1))^2) = \sigma^2$.

  Second moment of the service time: $E(Y_1^2) = \sigma^2 + \tau^2 = s^2$.
- There is a single server and the capacity of the queue is infinite.

Unfortunately, in this model the process $\{X(t) : t \geq 0\}$, the number of customers in the system at time t, is not a CTMC. Hence, determination of the limiting distribution of the process $\{X(t) : t \geq 0\}$) should be done in a different way.

We will restrict ourselves, however, to a so-called *mean-value analysis*: determination of the expected time in the system, the expected number of customers in the system, ...

Stability condition:

Just as for the $M/M/1$ queue, the stability condition for the $M/G/1$ queue is that the amount of work offered per time unit to the server should be less than the amount of work the server can handle per time unit, i.e.,

$$\rho := \lambda\tau < 1.$$

Occupation rate of the server:

Because the expected amount of work offered to the server per time unit equals $\rho < 1$, the fraction of time the server is busy (= occupation rate of the server) is also equal to ρ. The fraction of time the server is idle is hence equal to $1 - \rho$.

Expected time in the queue, W_q:

The time a customer is waiting in the queue consists of two parts:

- the *remaining* service time of the customer in service;
- the service times of the customers in the queue.

Hence, in order to calculate W_q we first have to obtain the expected remaining service time of the customer in service.

Expected remaining service time of the customer in service

Here is figure of the remaining service time of the customer in service as function of time.

Take a big interval of length T.

Expected number of served customers in $[0, T]$: λT.

Contribution of one customer to the expected area: $E(Y_1^2/2) = s^2/2$.

=> Total expected area in figure: $\lambda T \cdot s^2/2$.

=> Expected remaining service time: $\lambda s^2/2$.

The expected time in queue, W_q, now can be determined using the following *mean-value relations*:

$$
\begin{aligned}
W_q &= \lambda s^2/2 + L_q \tau, \\
L_q &= \lambda W_q.
\end{aligned}
$$

Remark that in the first relation we use the PASTA property and that the second relation is Little's formula applied to the queue.
Hence we have

$$
\begin{aligned}
W_q &= \frac{\lambda s^2}{2(1-\lambda\tau)} = \frac{\lambda s^2}{2(1-\rho)}, \\
L_q &= \lambda W_q = \frac{\lambda^2 s^2}{2(1-\rho)}.
\end{aligned}
$$

Once we know W_q and L_q, then W and L of course follow from

$$W = W_q + \tau \quad \text{and} \quad L = L_q + \rho.$$

Example: $M/M/1$ **queue**

In the case of exponentially distributed service times with parameter μ we have

$$\tau = \frac{1}{\mu}, \quad \sigma^2 = \frac{1}{\mu^2}, \quad s^2 = \frac{2}{\mu^2},$$

and hence the expected remaining service time equals

$$\frac{\lambda s^2}{2} = \frac{\lambda}{\mu^2} = \rho \cdot \frac{1}{\mu}.$$

This also follows from the memoryless property of the exponentisl dsitribution (explain).

For the quantities W_q and L_q we find (as before)

$$W_q = \frac{1}{\mu}\frac{\rho}{1-\rho}, \quad L_q = \frac{\rho^2}{1-\rho}.$$

Example: $M/D/1$ queue

In the case of deterministic service times equal to τ we have

$$\sigma^2 = 0, \quad s^2 = \tau^2,$$

and hence the expected remaining service time equals

$$\frac{\lambda s^2}{2} = \frac{\lambda \tau^2}{2} = \rho \cdot \frac{\tau}{2}.$$

For the quantities W_q and L_q we find

$$W_q = \frac{\tau}{2}\frac{\rho}{1-\rho}, \quad L_q = \frac{\rho^2}{2(1-\rho)}.$$

Remark that in the $M/D/1$ queue, the quantities W_q and L_q are smaller than in the corresponding $M/M/1$ queue. This is due to the smaller variance of the service times in the $M/D/1$.