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 $[0, T] : \lambda T$.

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 $M/D/1$.

The $G/M/1$ queue

we will now look at a model in which not the service times but the interarrival times are generally distributed, the $G/M/1$ queue.

Model:

- The arrival process is a process in which the interarrival times $(A_1, A_2, \ldots)$ of customers are identically distributed with an arbitrary distribution function $G(\cdot)$. The mean interarrival time equals $E(A_1) = 1/\lambda$. The function $\tilde{G}(s)$ is defined as

$$\tilde{G}(s) = E(e^{-sA_1}).$$

- Service times are exponentially dsistributed with parameter μ.
- There is a single server and the capacity of the queue is infinite.

Unfortunately, also for this model the process $\{X(t) : t \geq 0\}$, the number of customers in the system at time t, is not a CTMC. Also we can not use the mean-value analysis, as presented before for the $M/G/1$ queue, because the PASTA property does not hold anymore (the arrival process is not a Poisson process here).

We will restrict ourselves to stating results for the limiting distribution of the number of customers at arrival instants $(\pi_j^*, j = 0, 1, 2, \ldots)$ en and at arbitrary instants $(p_j, j = 0, 1, 2, \ldots)$.

Stability condition:

Just as for the $M/G/1$ queue, the stability condition for the $G/M/1$ 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.,

$$\rho := \frac{\lambda}{\mu} < 1.$$

The function $\tilde{G}(s) = E(e^{-sA_1})$ is called the *Laplace-Stieltjes transform* of the random variable A_1 and can be calculated as follows.

- If A_1 is a continuous random variable with probability density function $g(\cdot)$, then
$$\tilde{G}(s) = \int_0^\infty e^{-sx} g(x) dx.$$
- If A_1 is a discrete random variable with probability mass function $p(x_i) = P(A = x_i), i = 1, 2, \ldots$, then
$$\tilde{G}(s) = \sum_{i=1}^\infty e^{-sx_i} p(x_i).$$

Examples:

- If A_1 is exponential with parameter λ, then $\tilde{G}(s) = \lambda/(\lambda + s)$.
- If A_1 is deterministic and equal to $1/\lambda$, then $\tilde{G}(s) = e^{-s/\lambda}$.

Limiting distribution of the number of customers at arrival instants

The limiting distribution of the number of customers at arrival instants is given by

$$\pi_j^* = (1-\alpha)\alpha^j, \quad j \geq 0,$$

where α is the unique solution in the interval (0,1) of the equation

$$u = \tilde{G}(\mu(1-u)).$$

Example:

If the interarrival times are exponentially distributed with parameter λ, then $\alpha = \rho$ (check!) and hence

$$\pi_j^* = (1-\rho)\rho^j, \quad j \geq 0.$$

Limiting distribution of the number of customers at arbitrary instants

The limiting distribution of the number of customers at arbitrary instants is given by

$$p_0 = 1 - \rho, \quad p_j = \rho\pi^*_{j-1} = \rho(1-\alpha)\alpha^{j-1}, \quad j \geq 1.$$

Idea proof:

The long-run rate at which the number of customers in the system jumps from $j-1$ to j equals $\lambda\pi^*_{j-1}$.

The long-run rate at which the number of customers in the system jumps from j to $j-1$ equals μp_j.

Because these two rates have to be equal, we have $p_j = \rho\pi^*_{j-1}$.

Expected number of customers in the system

The expected number of customers in the system is given by

$$L = \sum_{j=1}^{\infty} j p_j = \rho(1-\alpha) \sum_{j=1}^{\infty} j \alpha^{j-1} = \frac{\rho}{1-\alpha}.$$

Expected time in the system

The expected time customers spend in the system is given by

$$W = \frac{L}{\lambda} = \frac{1}{\mu(1-\alpha)}.$$

Alternative derivation

$$W = \sum_{j=0}^{\infty} \pi_j^* \frac{j+1}{\mu} = \frac{1-\alpha}{\mu} \sum_{j=0}^{\infty} (j+1)\alpha^j = \frac{1}{\mu(1-\alpha)}.$$

The $G/G/1$ queue

The last single-station queueing model we discuss will be the $G/G/1$ queue. In this model, both the interarrival times and the service times have a *general* distribution.

For this model, an exact analysis is in general impossible. Therefore, we restrict ourselves to giving *approximations* for the following performance measures:

- W_q, the expected time in the queue;
- W, the expected time in the system;
- L_q, the expected number of customers in the queue;
- L, the expected number of customers in the system.

Model:

- The arrival process is a process for which the interarrival times $(A_1, A_2, \ldots)$ of customers are identically distributed random variables with an *arbitrary* distribution function.

  Mean interarrival time: $E(A_1)$.

  Variance of the interarrival time: $E((A_1 - E(A_1))^2) = \sigma^2_{A_1}$.

  Coefficient of variation of the interarrival time: $c_{A_1} = \frac{\sigma_{A_1}}{E(A_1)}$.
- Service times of customers $(B_1, B_2, \ldots)$ are identically distributed random variables with an *arbitrary* distribution function.

  Mean service time: $E(B_1)$.

  Variance of the service time: $E((B_1 - E(B_1))^2) = \sigma^2_{B_1}$.

  Coefficient of variation of the service time: $c_{B_1} = \frac{\sigma_{B_1}}{E(B_1)}$
- There is a single server and the capacity of the queue is infinite.

Stability condition:

Just as in the $M/M/1$, $M/G/1$ and $G/M/1$ queue, the stability condition for the $G/G/1$ 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.,

$$\rho := \frac{E(B_1)}{E(A_1)} < 1.$$

Approximation W_q:

An often used approximation for the expected time in the queue is given by

$$W_q \approx \frac{\rho}{1-\rho} \cdot \frac{c_{A_1}^2 + c_{B_1}^2}{2} \cdot E(B_1)$$

Special cases:

For the $M/M/1$ queue the approximation is equal to the exact value:

$$W_q = \frac{\rho}{1-\rho} \cdot \frac{1+1}{2} \cdot E(B_1) \qquad (M/M/1)$$

For the $M/G/1$ queue the approximation is equal to the exact value:

$$W_q = \frac{\rho}{1-\rho} \cdot \frac{1+c_{B_1}^2}{2} \cdot E(B_1) \qquad (M/G/1)$$

For the $G/M/1$ queue the approximation is NOT equal to the exact value:

$$W_q \approx \frac{\rho}{1-\rho} \cdot \frac{c_{A_1}^2+1}{2} \cdot E(B_1) \qquad (G/M/1)$$

Approximations for W, L_q and L:

From the approximation for W_q,

$$W_q \approx \frac{\rho}{1-\rho} \cdot \frac{c_{A_1}^2 + c_{B_1}^2}{2} \cdot E(B_1),$$

we immediately obtain approximations for W, L_q and L via the formulas

$$\begin{aligned} W &= W_q + E(B_1), \\ L_q &= \frac{W_q}{E(A_1)}, \qquad \text{(Little)} \\ L &= \frac{W}{E(A_1)}. \qquad \text{(Little)} \end{aligned}$$

Example:

- In a workstation jobs are delivered at a rate of one job every 8 hours.
- The standard deviation of the time between successice delivery times is 4 hours.
- The average production time of a job is 6 hours with a standard deviation of 2 hours.

Question:
What would be the reduction in the expected time in the system if the deliveries could be made more regular, for instance with a standard deviation of only one hour?

Answer:
In this case the expected time in the system is reduced from roughly 9.25 hours to approximately 7 hours.