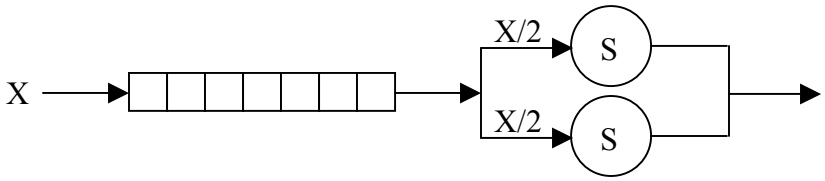
Dr. Jozo Dujmović

A Multiple Server Model

In the case of exponential interarrival and service times a single server model M/M/1 has the response time R=S/(1-SX). If the arrival/departure rate X is constant, then it is limited by X<1/S, and the only way to increase the performance is to have a faster server (i.e. to reduce S). Of course, the server speed is always limited, either by the financial limitations, or by the current technology limits. The only way to go beyond these limits is to use multiple servers. Following is an example of a model with two servers:



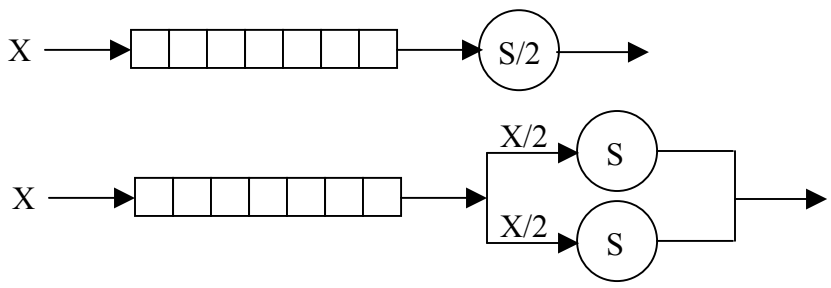
We assume that the servers are equivalent and that the flow of service requests is perfectly balanced: each server handles 50% of the input flow. If the interarrival times and the service times are exponentially distributed, then this model is called M/M/2 and the server utilization is

$$U_2 = SX/2$$
$$0 \le U \le 1, \quad X < X_{\max} = 2/S$$

The mean response time and the mean number of jobs in this model are:

$$R_2 = \frac{S}{1-U_2^2}$$
$$J = XR_2 = \frac{SX}{1-U_2^2} = \frac{2U_2}{1-U_2^2}$$

Let us now compare the following two models:



It might look that these models are equivalent because the single server is now two times faster. But they are not equivalent. Here are the basic performance indicators:

$$U_1 = SX/2, \quad R_1 = \frac{S/2}{1-U_1} = \frac{S/2}{1-SX/2} = \frac{S}{2-SX}$$

$$U_2 = SX/2, \quad R_2 = \frac{S}{1-U_2^2} = \frac{4S}{4-(SX)^2} = \frac{4S}{(2-SX)(2+SX)}$$

$$J_1 = \frac{U_1}{1-U_1}, \quad J_2 = \frac{2U_2}{1-U_2^2}$$

Therefore, $U_1 = U_2$ and we have

$$R_2 = R_1 \frac{2}{1+U_2} = R_1 \frac{4}{2+SX} > R_1, \quad 0 \le X \le 2/S$$

$$J_2 = J_1 \frac{2}{1+U_2} > J_1$$

The similarities between these two systems are:

- All servers have the same utilization $U_1 = U_2 = SX/2$
- Both systems have the same maximum throughput $X < X_{\max} = 2/S$
- For high traffic ($X \to X_{\max}, U_1 \to 1, U_2 \to 1$) we have $R_1 \cong R_2, \; J_1 \cong J_2$

The differences between these two systems are:

- In the case of low traffic ($X \to 0$) the response times are $R_1 = S/2, \quad R_2 = S$
- The response times and the number of jobs are different: $R_1 < R_2, \quad J_1 < J_2$

The fundamental difference between these systems is clearly visible in the case of zero wait time: the response time of the single server system is $R_1 = S/2$, but in the case of two servers, only one server is active yielding a double response time $R_2 = S = 2R_1$. Therefore, the two-server system with servers that have the service time S is slower than the single-server system with server that has the service time S/2.

The reason for using multiple servers is not only performance, but also the reliability of systems. In the case of multiple servers, if one server is down, the system may continue working with reduced performance, while in the case of single server, the system is no longer available.

In the case of k equivalent parallel servers (k>0) and exponential interarrival and service times (M/M/k model) the performance indicators are:

$$U_1 = SX$$

$$U_k = SX/k = U_1/k$$

$$R_k = S\left[1 + \frac{U_1^k}{k!k(1-U_k)^2\left(1 + U_1 + \frac{U_1^2}{2!} + \cdots + \frac{U_1^{k-1}}{(k-1)!} + \frac{U_1^k}{k!(1-U_k)}\right)}\right] = S + W_k$$

$$W_k = S\frac{U_1^k}{k!k(1-U_k)^2\left(1 + U_1 + \frac{U_1^2}{2!} + \cdots + \frac{U_1^{k-1}}{(k-1)!} + \frac{U_1^k}{k!(1-U_k)}\right)}$$

$$J_k = XR_k$$

Here W_k denotes the time customers spend in queue waiting for service. Obviously, if $U_k \to 1$ then $R_k \to \infty$ and $W_k \to \infty$. Special cases of these formulas are:

$$R_1 = S\left[1 + \frac{U_1}{1-U_1}\right] = \frac{S}{1-U_1}$$

$$R_2 = S\left[1 + \frac{U_2^2}{1-U_2^2}\right] = \frac{S}{1-U_2^2}$$

$$R_3 = S\left[1 + \frac{U_1^3}{18(1-U_3)^2\left(1 + U_1 + \frac{U_1^2}{2} + \frac{U_1^3}{6(1-U_3)}\right)}\right] = S\left[1 + \frac{3U_3^3}{2 + 2U_3 - U_3^2 - 3U_3^3}\right]$$

$$R_4 = S\left[1 + \frac{U_1^4}{96(1-U_4)^2\left(1 + U_1 + \frac{U_1^2}{2} + \frac{U_1^3}{6} + \frac{U_1^4}{24(1-U_4)}\right)}\right]$$