Queueing Theory Notes

Software Systems **Allen B. Downey**
Spring 2008

The M/M/1 model can be extended to deal with a general distribution of service times, in which case it is called M/G/1 (see Chapter 9 of Taylor and Karlin, equations 9.35 and 9.36).

The mean and standard deviation of service times are called ν and τ. So the variance of service times is τ^2 and the service rate is

$$\mu = 1/\nu$$

The load is

$$\rho = \lambda/\mu = \lambda\nu$$

By an analysis similar to what we did in class, we can derive

$$L = \rho + \frac{\lambda^2\tau^2 + \rho^2}{2(1-\rho)}$$

and

$$W = \nu + \frac{\lambda(\tau^2 + \nu^2)}{2(1-\rho)}$$

So for fixed (λ, ν), W and L increase linearly with τ^2! (Or quadratically with standard derviation and coefficient of variation).

If the distribution of service times is exponential, so $\tau = \nu$. With a little algebra, we can derive

$$W = \frac{1}{\nu - \lambda}$$

which is what we derived earlier for the M/M/1 system.

If service times are constant, $\tau = 0$ and we get

$$W = \frac{1 - \rho/2}{\nu - \lambda}$$

So with constant service times we expect average wait time to be less than with exponential service times. For low loads, there is not much difference, but as the load increases, the reduction approaches a factor of 2.