



Fig. 2.17. Normalized response times (R/S) for multiple servers $m = 1, 4, 16, 64$. The upper curve corresponds to the single server case

$$R = \frac{C(m, \rho)S}{m(1 - \rho)} + S, \quad (2.64)$$

where the first term is the expected waiting time W . The corresponding queue length (number of customers in the queueing center, including the ones in service) is given by Little's law $Q = \lambda R$:

$$Q = \frac{\rho C(m, \rho)}{m(1 - \rho)} + m\rho. \quad (2.65)$$

$C(m, \rho)$ is the probability that all the servers are busy and therefore customers arriving with traffic intensity (2.47) will have to wait for service. This probability is defined by the rather complicated function:

$$C(m, \rho) = \frac{\frac{(m\rho)^m}{m!}}{(1 - \rho) \sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!}}, \quad (2.66)$$

which is known as *Erlang's C function*. Notice that if $m = 1$ then $C(1, \rho) = \rho$ (the probability that the single server is busy) so (2.64) reduces to the residence time (2.35) for a uniserver queue, and similarly the queue length (2.65) reduces to (2.36).

Excerpted from "Analyzing Computer System Performance with Perl::PDQ", by Neil J. Gunther, Springer-Verlag 2005. ISBN 3540208658