

Queuing

Miles Jefferson

January 6, 2011

This is a brief description of how queues can be modeled. Software implementing the formulas described can be found at <http://www.milesj.com/queue.htm>

The binomial distribution

In general, for n number of independent trials each with a probability p of success and $q = (1 - p)$ probability of failure, the probability of r successful events occurring in n is given by the binomial distribution

$$Pr \{r \text{ in } n, p\} = {}^n C_r \cdot p^r \cdot q^{(n-r)} = {}^n C_r \cdot p^r \cdot (1-p)^{(n-r)}$$

where ${}^n C_r$ is the number of combinations,

$${}^n C_r = \frac{n!}{r!(n-r)!}$$

giving

$$Pr \{r \text{ in } n, p\} = \frac{n! \cdot p^r \cdot (1-p)^{(n-r)}}{r!(n-r)!}$$

Calls and traffic

If n calls are made from an infinite source over time T , then the average call arrival rate, v , is

$$v = \frac{n}{T}$$

When there are an infinite number of sources, the rate of calls being offered to the system, v , is not affected by how many calls are currently in the system so the the probability, p , of a 1 call in an interval dt is

$$Pr \{1 \text{ call arrives in } dt\} = p = v \cdot dt = \frac{n}{T_S} dt$$

If each call has an average call duration T_s then the call load, A , is

$$A = p.T_s$$

If these calls are handled by m channels, then the agent occupancy, ρ , is found by dividing the call load by the number of agents

$$\rho = \frac{A}{m}$$

Since the mean call duration T_s tells us how long a channel will remain occupied once a call arrives, the probability that an occupied channel will complete a call in dt is

$$Pr \{1 \text{ call completed in } dt\} = \frac{1}{T_s} .dt$$

The arrival distribution

If the n calls are generated independently, then the probability of r arriving in T will follow a binomial distribution

$$Pr_A \{r \text{ in } T, p\} = \frac{T!.p^r.(1-p)^{(T-r)}}{r!(T-r)!}$$

For large n the Poisson distribution can be used, which is a limiting case to the binomial distribution as $n \rightarrow \infty$ and p remains fixed.

Blocking probability

If all m channels are occupied when a call arrives the arriving call cannot be completed. It can then be handled in one of several ways: either "cleared," in effect forgotten; "returned", where after a delay it reiterates, or the call is put on hold ("held") in a queue of a certain size, Q . Here we will only consider the latter situation. The probability that all m channels are blocked with calls being held in a queue is given by Erlang C formula,

$$P_Q(m, A) = \frac{\left(\frac{A^m}{m!}\right) \left(\frac{m}{m-A}\right)}{\sum_{i=0}^{m-1} \frac{A^i}{i!} + \left(\frac{A^m}{m!}\right) \left(\frac{m}{m-A}\right)}$$

Having calculated P_Q , the average waiting time, T_W can be found as

$$T_w = \frac{P_Q(m, A) .T_s}{m.(1-\rho)}$$

A table of synonyms

The queue modeling approach described above can be used in many situations. The table below shows how terms are synonymous across different scenarios.

scenario	Term			
	n	m	A	ρ
call center	calls	agents	call load	agent occupancy
telephone exchange	calls	lines	offered traffic	utilization
network	requests	servers	traffic intensity	utilization
doctors on-call	assessments	doctors	work load	intensity
ambulances	999s	vehicles	call load	intensity
checkouts	customers	cashiers	customer load	cashier occupancy

Further reading

There are many resources and articles on-line describing the modeling of queuing systems. I would suggest starting at [http://en.wikipedia.org/wiki/Erlang_\(unit\)#Erlang_C_formula](http://en.wikipedia.org/wiki/Erlang_(unit)#Erlang_C_formula). I will update this document with further information at some point.