Analytical queuing models

Notation

c_a = coefficient of variation of arrival times

c_e = coefficient of variation of process time

m = number of parallel servers at a station

r_a = arrival rate (items per unit time) = $1/t_a$

r_e = processing rate (items per unit time) = m/t_e

t_a = average time between arrival

t_e = mean processing time

u = utilisation of station = $r_a/r_e = (r_a t_e)/m$

W = expected waiting time in the system (queue time + processing time)

W_q = expected waiting time in the queue

WIP = average work in progress (number of items) in the queue

WIP_q = expected work in progress (number of times) in the queue

Variability

If there were no variability, queues would not have to occur since the capacity of a process could be relatively easily adjusted to match demand

If arrival rate ≤ processing rate && no variation then **WIP_q = 0** and **u = 1**

Utilization = processing rate / (arrival rate · m), m = number of servers

Incorporating variability

Assumption of no variation in arrival or processing times is not realistic. The average or mean arrival and process times can be calculated but only if the variation around these is taken into account – done by using a probability distribution

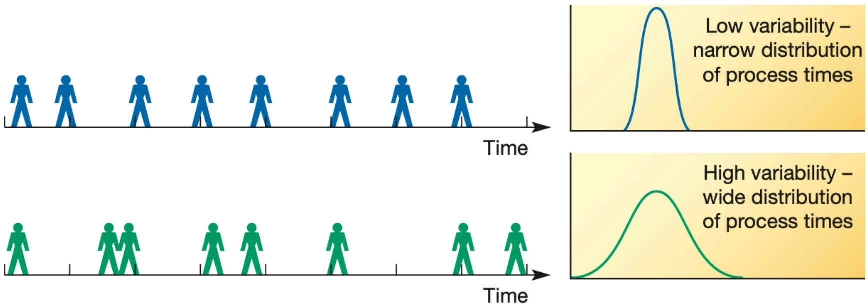

Figure 11.22 Low and high arrival variation

The usual measure for indicating the spread of a distribution is its standard deviation σ. To normalize standard deviation, it is divided by the mean of its distribution

$$c_a = \text{coefficient of variation of arrival times} = \sigma_a / t_a$$

$$c_e = \text{coefficient of variation of processing times} = \sigma_e / t_e$$

Incorporating Little's law

Little's law: $\text{Throughput time} = \text{Work in progress} \times \text{Cycle time}$

Work in progress = Throughput time / Cycle time

WIP = T/C

$$\text{Work in progress in the queue} = \text{the arrival rate at the queue (equivalent to 1/cycle time)} \times \text{waiting time in the queue (equivalent to throughput time)}$$

$$WIP_q = r_a \times W_q$$

$$\text{Waiting time in the whole system} = \text{the waiting time in the queue} + \text{the average process time at the station}$$

$$W = W_q + t_e$$

Types of queueing system

Queuing systems are characterized by four parameters: A/B/m/b

A = distribution of arrival times (interarrival times, the elapsed times between arrivals)

B = distribution of process time

m = number of servers at each station

b = maximum number of items or people allowed in the system

A or B are usually describe as the:

a. The exponential or Markovian distribution denoted by M
b. The general normal distribution denoted by G

Kendall's notation = M/G/1/5 queuing system indicates a system with exponentially distributed arrivals, process times described as a general distribution such as normal distribution, with one server and a maximum number of items allowed of 5.

The most common situations are:

1. M/M/m = the exponential arrival and processing times with m servers and no maximum limit to the queue
2. G/G/m = general arrival and processing distributions with m servers and no limit to the queue

M/M/1 queuing systems

The formula for M/M/1 systems are: $\text{WIP} = \dfrac{u}{1 - u}$

$$\text{WIP} = \text{Cycle time} \times \text{Throughput time}$$

Since $\text{Throughput time} = \text{WIP/Cycle time}$ then

$$\text{Throughput time} = \frac{u}{1 - u} \times \frac{1}{r_a} = \frac{t_e}{1 - u}$$

Since queue = total throughput time – average processing time, then:

$$\begin{aligned} W_q &= W - t_e \\ &= \frac{t_e}{1 - u} - t_e \\ &= \frac{t_e - t_e(1 - u)}{1 - u} = \frac{t_e - t_e - ut_e}{1 - u} \\ &= \frac{u}{(1 - u)} t_e \end{aligned}$$

And Little's law gives $\text{WIP}_q = r_a \times W_q = \dfrac{u}{(1 - u)} t_e r_a$

$$u = \frac{r_a}{r_e} = r_a t_e \qquad \text{WIP}_q = \frac{u}{(1 - u)} \times t_e \times \frac{u}{t_e}$$

Since $r_a = \dfrac{u}{t_e}$ then $= \dfrac{u^2}{(1 - u)}$

M/M/m queuing systems

$$W_q = \frac{u^{\sqrt{2(m+1)}-1}}{m(1 - u)} t_e$$

G/G/1 systems

$$W_q = \left(\frac{c_a{}^2 + c_e{}^2}{2}\right)\left(\frac{u}{(1-u)}\right) t_e$$

The formula is known as the VUT formule because it describes the waiting time as a function of V = variability in the queuing system, U = utilization of the queuing system (demand vs capacity), and T = processing times at the station

G/G/m systems

$$W_q = \left(\frac{c_a{}^2 + c_e{}^2}{2}\right)\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right) t_e$$