The
G/M/1, G/G/1, G/G/m and M/G/m/m
Queues

The G/M/1 Queue

- The G/M/1 queue is the dual of the M/G/1 queue where the arrival process is a general one but the service times are exponentially distributed.

- Service time distribution is exponential with parameter $1/m$

- General Arrival Process with mean arrival rate l.

  Inter-arrival time is random with pdf $a(t)$, cdf $A(t)$ and L.T. of the pdf as $L_A(s)$

- Total Traffic $r = l/m$

Stability consideration require that $r < 1$ for the queue to be at equilibrium

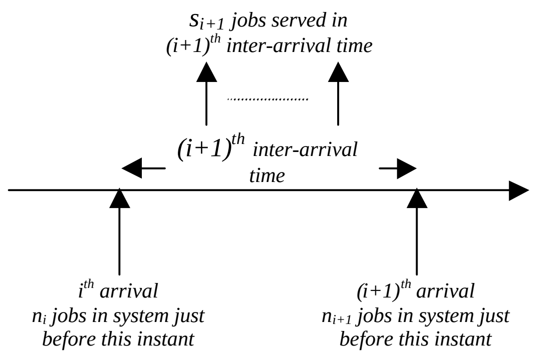
• For analyzing the G/M/1 queue using the Imbedded Markov Chain approach, the *imbedded points are chosen to be the arrival instants of jobs to the system*

• System State = Number in the system *immediately before an arrival instant*

n_i = Number in the system just before the i^{th} arrival

s_{i+1} = Number of jobs served between the i^{th} and the $(i+1)^{th}$ arrivals

The sequence $\{n_i\}$, $i=1,2,\ldots\ldots$ at the imbedded Markov points (i.e. just before arrival instants) forms a Markov Chain.



$$n_{i+1} = n_i + 1 - s_{i+1} \qquad n_i=0,1,\ldots.,\infty \qquad s_{i+1} \le n_i+1 \qquad (1)$$

Consider the chain at equlibrium (i.e. for $i \to \infty$) $\qquad n_{i+1} = n_i + 1 - s_{i+1} \qquad n_i = 0,1,\ldots,\infty \qquad s_{i+1} \le n_i + 1$

One-Step Transition Probabilities

$$\begin{cases} p_{jk} = P\{n_{i+1} = k \mid n_i = j\} \\ p_{jk} = 0 \quad for \quad k > j+1 \end{cases} \qquad (2)$$

Balance Equations

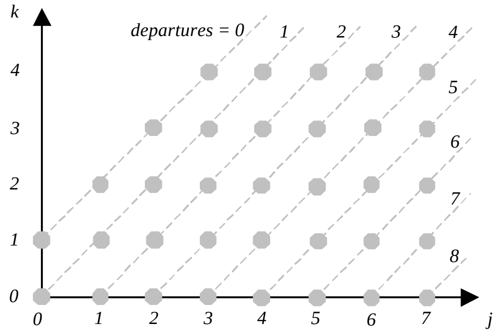
$$p_k = \sum_{j=0}^{\infty} p_j \, p_{jk} \qquad k = 0,1,\ldots,\infty \qquad (3)$$

Normalization Condition

$$\sum_{k=0}^{\infty} p_k = 1$$

 5

Allowable state transitions ($j \to k$) in the G/M/1 queue between successive imbeded points



The points shown are the ones for which the ($j \to k$) transitions can occur. For all other points, the corresponding transitions cannot occur.

The diagonal lines correspond to a constant number of departures (0,1,2,……) between the corresponding ($j \to k$) transitions.

 6

Let a_n = P{n departures in an inter-arrival time interval | server is always busy during this inter-arrival time}

$$a_j = \int_{x=0}^{\infty} \frac{(mx)^j}{j!} e^{-mx} a(x)dx \qquad j = 0,1,......., \infty \qquad (5)$$

Justification: *Since the service times are exponentially distributed, the number of departures in an inter-arrival time instant where the server is always busy will have the Poisson distribution.*

and $$\sum_{j=0}^{\infty} a_j z^j = \int_0^{\infty} e^{-mx(1-z)} a(x)dx = L_A(m - mz)$$

The a_j's may be found as the coefficient of z^j in the series expansion of $L_A(m\text{-}mz)$

The Balance Equations of (3) then become

$$p_j = a_0 p_{j-1} + \sum_{k=0}^{\infty} a_{k+1} p_{j+k} \qquad j = 1,......., \infty \qquad (4)$$

Solutions to these, satisfying the normalization condition are -

$$p_j = (1 - s)s^j \qquad j = 0,1,......., \infty \qquad (6)$$

where s is a unique root of $s = L_A(m - ms)$

It may be shown that if $r < 1$, then there will always be a unique real solution for $s = L_A(m - ms)$ which will be in the range $0 < s < 1$.

The solution may be verified by direct substitution, as given below -

$$(1-s)s^{j} = a_0(1-s)s^{j-1} + \sum_{k=0}^{\infty} a_{k+1}(1-s)s^{j+k}$$

$$s^{j} = a_0 s^{j-1} + \sum_{k=1}^{\infty} a_k s^{j+k-1}$$

$$s = a_0 + \sum_{k=1}^{\infty} a_k s^{k}$$

$$= \sum_{k=0}^{\infty} a_k s^{k} = L_A(m - ms)$$

• The state distribution $\{p_j\}$ $j=0,1,\ldots.$ is found under equilibrium conditions at the time instants just before job arrivals to the system.

• It is also valid for the departure instants (just after a job leaves the system) as Kleinrock's principle is applicable to this system (i.e. the state changes are at most *+1* or ***-****1*.

• It in **not valid** for arbitrary time instants (or ergodic, time-average results) since PASTA will not be applicable to the system (i.e. the arrival process is not Poisson).

For a FCFS M/G/1 queue at equilibrium, the following queueing delay results may be obtained.

These may be derived using the fact that the service time distribution is exponential and hence memory less

$$W_q = \sum_{n=1}^{\infty} \frac{n}{m}(1-s)s^n = \frac{s}{m(1-s)}$$

$$f_{Wq}(t) = (1-s)d(t) + ms(1-s)e^{-m(1-s)t} \quad \text{for } t \geq 0$$

Other parameters such as W and N_q and the distribution $f_W(t)$ may also be obtained

The multi-server *G/M/m* queue may also be analyzed using a similar approach

The G/G/1 Queue

We cannot analyse this queue exactly but there are useful bounds that have been developed for the waiting time in queue W_q. This can then be used to find bounds on W, N and N_q in the usual fashion, i.e.Little's Result and $W = W_q + \overline{X}$

l = Average arrival rate of jobs (general arrival process)

Let T be the (random) inter-arrival time with (general service time distribution)

$$\begin{cases} E\{T\} = 1/l \\ s_T^2 = E\{T^2\} - [E\{T\}]^2 \end{cases}$$

Let X be the (random) service time with

$$\begin{cases} E\{X\} = \overline{X} \\ s_X^2 = E\{X^2\} - [E\{X\}]^2 \end{cases}$$

$r = l\overline{X}$ = Traffic Offered

$r < 1$ for queue to be stable

If the mean and variance (or second moment) of the inter-arrival times and the service times are known, then the following bounds have been shown to hold for W_q, the waiting time in queue, of any G/G/1 queue.

$$\frac{l\,s_X^2 - \overline{X}(2-r)}{2(1-r)} \leq W_q \leq \frac{l\,(s_X^2 + s_T^2)}{2(1-r)} \qquad (1)$$

Lower Bound *Upper Bound*

The upper bound of (1) is quite useful. The lower bound is actually not very useful as it often gives a negative result which is a trivial conclusion.

 13

Another interesting (and very useful) bound for the G/G/1 queue has been given for the special case where the inter-arrival time *T* satisfies the following property for all values of *t*.

$$E\{T - t\,|\,T > t\} \leq \frac{1}{l} \qquad \text{for all } t \text{ ³ } 0 \qquad (2)$$

 14

• Note that E{T}=$1/l$. The condition of (2) is not very hard to satisfy. If the inter-arrival time is known to be more than t, then (2) requires that the expected length of the remaining inter-arrival time should be less than or equal to the *unconditioned* expected inter-arrival time $1/l$.

• For the special case when the arrival process is Poisson, the inter-arrival times will be exponentially distributed and will satisfy (2) as an equality.

• Note that many distributions (like say the uniform distribution) will satisfy (2). An exception to this are *hyper-exponential* type distributions

If the arrival process is such that (2) is satisfied, then the following bounds have been shown to hold

$$W_{qU} - \frac{1+r}{2l} \le W_q \le W_{qU} \qquad (3)$$

where W_{qU} is the upper bound of (1), i.e. $W_{qU} = \dfrac{l(s_X^2 + s_T^2)}{2(1-r)}$

To see the tightness of the bounds of (3), consider using it to find the bounds on N_q, the mean number waiting in queue for a G/G/1 system.

We get

$$\lambda W_{qU} - \frac{1+\rho}{2} \leq N_q \leq \lambda W_{qU} \qquad (4)$$

• The difference between the upper and the lower bounds is only $0.5(1+\rho)$.

• Note that ρ will be small anyway as $0<\rho<1$ for a stable queue.

• In any case, the difference between the upper and the lower bounds will be between 0.5 and 1. In percentage terms, as $\rho \rightarrow 1$, i.e the traffic increases, this will get increasingly smaller compared to N_q.

Heavy Traffic Approximation for the G/G/1 Queue

As $\rho \rightarrow 1$, (i.e. when the offered traffic is high), the distribution of the waiting time in a G/G/1 queue will be approximately an exponentially distributed radom variable with mean given by

$$W_q = \frac{\lambda(\sigma_X^2 + \sigma_T^2)}{2(1-\rho)}$$

Note that this results is an interesting one as it not only provides a mean but also a distribution for the waiting time in queue under very general conditions.

The G/G/m Queue

m = Number of Servers
$r = l\overline{X}$ = Offered Traffic

W_{q1} = Average waiting time in queue for the equivalent G/G/1 queue.

Other notations same as for the G/G/1 queue

The Equivalent G/G/1 Queue

Same arrival process of jobs as for the G/G/m queue

For the service times, use $\frac{\overline{X}}{m}$ and $\frac{s_X^2}{m^2}$ as the mean and variance, respectively.

Note that the server here works *m times faster* than a server in the original G/G/m queue

The following bounds then hold for the G/G/m queue

$$W_{q1} - \frac{(m-1)\overline{X^2}}{2m\overline{X}} \le W_q \le l \frac{\left[s_T^2 + \frac{s_X^2}{m} + \frac{(m-1)(\overline{X})^2}{m^2} \right]}{2\left[1 - \frac{r}{m}\right]}$$

- The value of W_{q1} may be computed as a lower bound on W_q
- These bounds are rather loose and may not be very useful in practice

Heavy Traffic Approximation for the G/G/m Queue

For *(r/m)®1* in a G/G/m queue, a *heavy traffic approximation result* holds in a manner similar to that given for the G/G/1 case.

Specifically, for *(r/m)®1* in a G/G/m queue, the waiting time in queue at steady-state tends towards a random variable with an exponential distribution which has a mean given by -

$$W_q \approx l \frac{\left[s_T^2 + \frac{s_X^2}{m}\right]}{2\left[1 - \frac{r}{m}\right]}$$

The M/G/m/m Queue

• Even though the M/G/m queue is hard to analyze, the finite capacity M/G/m/m queue (*m server queue without waiting positions*) is surprisingly easy to analyze. (See additional notes)

• Even more remarkably, its state probability distribution and blocking probability results are identical to those obtained for the corresponding M/M/m/m queue

$$p_k = \frac{\frac{r^k}{k!}}{\sum_{n=0}^{m} \frac{r^n}{n!}} \qquad P_B = B(m, r) = \frac{\frac{r^m}{m!}}{\sum_{j=0}^{m} \frac{r^j}{j!}} \qquad r = l\overline{X}$$

for *k=0,1,……..,m*