Equilibrium Analysis
of
the M/G/1 Queue

1. Mean Analysis using Residual Life Arguments
   (Section 3.1)

2. Analysis using an Imbedded Markov Chain Approach
   (Section 3.2)

3. Method of Supplementary Variables (done later!)
   (*http://home.iitk.ac.in/~skb/qbook/MG1_SupVar.PDF*)

Method of Stages or other exact/approximate analytical methods may also be used

Why is the M/M/1 queue so easy to analyze while the analysis of the M/G/1 queue is substantially more difficult ?

• State description for M/M/1 is simple as one needs just one number (i.e. the number in the system) to denote the system state.

• This is possible because the exponential service time distribution is memoryless and service already provided to the customer currently in service need not be considered in the state description.

• This is not true for the M/G/1 queue. Its general state description would require specification of both the number currently in the system and the amount of service already provided to the customer currently being served.

M/G/1/∞ Queue: Single server, Infinite number of waiting positions

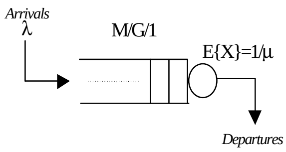
Service discipline assumed to be FCFS unless otherwise specified.
Mean results same regardless of the service discipline

Arrival Process: Poisson with average arrival rate l
Inter-arrival times exponentially distributed with mean $1/l$

Service Times: Generally distributed with pdf $b(t)$, cdf $B(t)$ and L.T.[$b(t)$]=$L_B(s)$

Residual Life Approach for Analyzing the M/G/1 Queue

(Section 3.1)



Note that this approach can only give the mean results for the performance parameters - state distributions cannot be found

We will tacitly assume a FCFS queue. However, since only the mean results are being obtained, these will be the same for queues with other service disciplines, such as LCFS, SIRO etc..

Consider a particular arrival of interest entering the M/G/1 queue

Let r = (random) residual service time of the customer (if any) currently in service

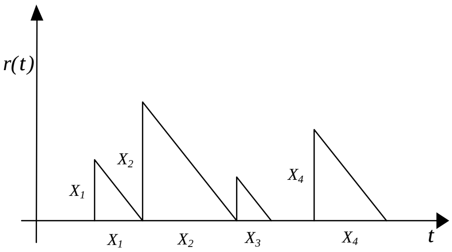
$R=E\{r\}$ Mean Residual Service Time

Then
$$W_q = N_q E\{X\} + R = \lambda W_q E\{X\} + R$$

$$W_q = \frac{R}{(1-\rho)}$$

where
$$\rho = \lambda E\{X\} = \lambda \overline{X} = \frac{\lambda}{\mu}$$

We still need to find R to find W_q. However, once W_q is known, the results N_q, N and W may be found directly from that.



Residual Service Time $r(t)$ as a Function of t

R may be found as the time average of $r(t)$ using a graphical approach, as shown

$R = \lim_{t\to\infty} R_t$ where

$$R_t = \frac{1}{t}\int_0^t r(t)dt \cong \frac{1}{t}\sum_{i=1}^{M(t)} \frac{1}{2} X_i^2$$

$$= \frac{1}{2}\frac{M(t)}{t}\frac{1}{M(t)}\sum_{i=1}^{M(t)} \frac{1}{2} X_i^2$$

For $t \to \infty$,

$$\frac{M(t)}{t} \to \lambda$$

$$\frac{1}{M(t)}\sum_{i=1}^{M(t)} \frac{1}{2} X_i^2 \to \overline{X^2}$$

$$R = \frac{1}{2}\lambda \overline{X^2} \qquad (3.1)$$

$$W_q = \frac{\lambda \overline{X^2}}{2(1-\rho)} \qquad (3.2)$$

Pollaczek-Khinchine or P-K Formula

$$W_q = \frac{\lambda \overline{X^2}}{2(1-\rho)}$$

$$W = W_q + E\{X\} = \frac{\lambda \overline{X^2}}{2(1-\rho)} + \overline{X}$$

$$N_q = \lambda W_q = \frac{\lambda^2 \overline{X^2}}{2(1-\rho)}$$

$$N = \lambda W = \frac{\lambda^2 \overline{X^2}}{2(1-\rho)} + \rho$$

For the M/M/1 queue

$E\{X\}=1/\mu$,

$E\{X^2\}=2/\mu^2$

Substituting these lead to the same results as obtained directly for the M/M/1 queue earlier

P{server is busy}= $r = lE\{X\}$
P{server not busy}= $(1-r)$
(See page 59 for justification)

$$R = (1-r)E\{r \mid \text{system found empty on arrival}\} + rE\{r \mid \text{system found not empty on arrival}\}$$

$$E\{r \mid \text{system found not empty on arrival}\} = \frac{R}{r} = \frac{\overline{X^2}}{2\overline{X}} = \frac{1}{2}\left(\overline{X} + \frac{s_X^2}{\overline{X}}\right)$$

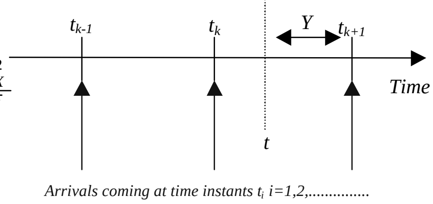
Note that the *counter-intuitive* nature of the above result, i.e. that it is

$\frac{1}{2}\left(\overline{X} + \frac{s_X^2}{\overline{X}}\right)$ rather than $\frac{1}{2}\overline{X}$ illustrates the *Paradox of Residual Life*

Arrival to a non-empty queue samples an ongoing service time but would tend to select longer service times more than shorter ones.

Some Residual Life Results

System examined at arbitrary time instant *t*



$$E\{t_i - t_{i-1}\} = \overline{X}$$
$$E\{(t_i - t_{i-1})^2\} = \overline{X^2}$$
$$\Rightarrow \overline{Y} = \frac{1}{2}\overline{X} + \frac{1}{2}\frac{s_X^2}{\overline{X}}$$

where X is a *lifetime* and Y is then referred to as the *residual lifetime*

$f_{\hat{X}}(x) = \frac{x f_X(x)}{\overline{X}}$ where $\hat{X}$ is the pdf of the selected lifetime

For the distribution of Y, we have the following results

$$f_Y(y)dy = P\{y \le Y \le y+dy\} = \frac{1}{\overline{X}}[1 - F_X(y)]dy \tag{3.7}$$

$$L_Y(s) = L.T.\left(\frac{1}{\overline{X}} - \frac{1}{\overline{X}}\int_0^y f_X(y)dy\right) = \frac{1 - L_X(s)}{s\overline{X}} \tag{3.8}$$

The Imbedded Markov Chain Approach (M/G/1 Queue)

(Section 3.2)

• Choose *imbedded time instants* t_i $i=1, 2, 3\ldots..\infty$ as the instants just after the departure of jobs from the system (after completing service)

• At these time instants, we can describe the system state by the number in the system, i.e.

n_i = *Number left behind in the queue by the* i^{th} *departure*

• We can easily see (shown subsequently) that the sequence n_i forms a Markov Chain, which can be solved to obtain the equilibrium state distribution at these specially chosen time instants ("the departure instants")

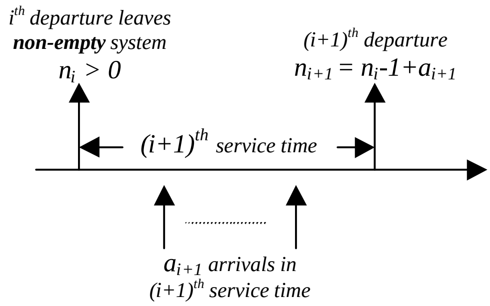
Useful Results Applicable to the M/G/1 Queue

Kleinrock's Result: For systems where the system state can change at most by +1 or -1, the system distribution as seen by an arriving customer will be the same as that seen by a departing customer

State Distribution at the Arrival Instants will be the same as the State Distribution at the Departure Instants

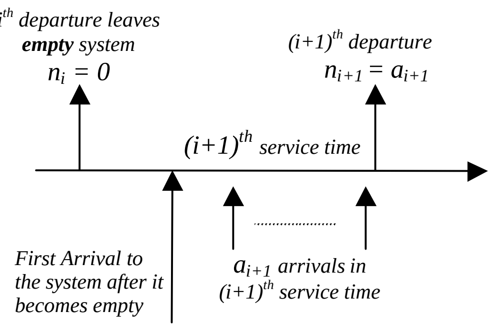
PASTA: Poisson Arrival See Time Averages

State Distributions and Moments seen by an arriving customer will be the same as those observed at an arbitrarily chosen time instant under equilibrium conditions



Departure Leaves System Non-empty

$$n_{i+1} = n_i - 1 + a_{i+1} \qquad n_i = 1,2,\ldots\ldots\ldots\ldots$$



Departure Leaves System Empty

$$n_{i+1} = a_{i+1} \qquad n_i = 1,2,\ldots\ldots\ldots\ldots$$

$$n_{i+1} = a_{i+1} \qquad for \quad n_i = 0 \qquad (3.11)$$
$$= n_i - 1 + a_{i+1} \qquad for \quad n_i = 1, 2, 3 \ldots\ldots\ldots$$

or

$$n_{i+1} = n_i - U(n_i) + a_{i+1} \qquad for \quad n_i = 0, 1, 2, 3, \ldots\ldots\ldots \qquad (3.12)$$

 15

Taking expectations of LHS and RHS of (3.11) or (3.12) $\Rightarrow$ $E\{U(n_i)\}=E\{a_{i+1}\}$

Since
$$\begin{cases} E\{U(n_i)\} = 1 - p_0 \\ E\{a_{i+1}\} = \int_0^{\infty} (l\,t) b(t) dt = l\overline{X} = r \end{cases}$$

Therefore $p_0 = 1 - r$ P{System Empty}

 16

P(z) Generating Function for the Number in the System

$$P_i(z) = E\{z^{n_i}\} = \sum_{k=0}^{\infty} z^k P\{n_i = k\}$$

$$P_{i+1}(z) = E\{z^{n_{i+1}}\} = \sum_{k=0}^{\infty} z^k P\{n_{i+1} = k\}$$

$$P_{i+1}(z) = E\{z^{n_i - U(n_i)}\}E\{z^{a_{i+1}}\}$$

Solve for Transient Solution

For Equilibrium State Distribution

1. Drop subscript " *i* " since equilibrium conditions are considered
2. Use $A(z)=L_B(l - lz)$ (3.13)
3. Use the following results -

$$A'(z) = -lL'_B(l - lz) \quad A'(1) = -lL'_B(0) = l\overline{X} = r$$

$$A''(z) = l^2 L''_B(l - lz) \quad A''(1) = l^2 L''_B(0) = l^2\,\overline{X^2}$$

$$P(z) = A(z)E\left\{z^{n-U(n)}\right\} = A(z)\sum_{k=0}^{\infty} z^{k-U(k)} P\{n = k\}$$

$$= A(z)\left[z^0 p_0 + \sum_{k=1}^{\infty} z^{k-1} p_k\right] = A(z)\left[p_0 + \frac{1}{z}\sum_{k=0}^{\infty} z^k p_k - \frac{1}{z} p_0\right]$$

$$= A(z)\left[\frac{1}{z}P(z) - \frac{1}{z} p_0 (1 - z)\right]$$

$$P(z) = \frac{(1-r)(1-z)A(z)}{A(z) - z} = \frac{(1-r)(1-z)L_B(l - lz)}{L_B(l - lz) - z}$$

P-K Transform Equation (3.14)

Under equilibrium conditions, *P(z)* was derived at the customer departure instants. However -

- It will hold at the customer arrival instants (Kleinrock's Result)
- It will also hold for the time averages or at an arbitrary time instant under equilibrium conditions

Expressing $P(z) = \sum_{i=0}^{\infty} a_i z^i$ (*Taylor Series Expansion*)

We can obtain a_i = P{*i* customers in the system} under equilibrium conditions

Moments of the system parameters (e.g. number in the system, may be computed directly from *P(z)*

For this, use *A(1)=1* $A'(1) = l\overline{X} = r$ $A''(1) = l^2\overline{X^2}$

$$P(1) = \lim_{z\to 1} P(z) = \lim_{z\to 1} \frac{(1-r)[(1-z)A'(z) - A(z)]}{A'(z) - 1} = -\frac{(1-r)}{r-1} = 1$$

This result, i.e. *P(1)* must be unity could have been used to obtain p_0 directly, instead of obtaining it as done earlier

Similarly $N = P'(1) = r + \dfrac{l^2\overline{X^2}}{2(1-r)}$ *Mean number in system*

Knowing *N*, the other parameters N_q,*W*, and W_q may be calculated as before

$$W = \overline{X} + \frac{l\,\overline{X^2}}{2(1-r)}$$

$$N = r + \frac{l^2\,\overline{X^2}}{2(1-r)} \quad \Rightarrow \quad \left\{ \begin{array}{l} W = \overline{X} + \dfrac{l\,\overline{X^2}}{2(1-r)} \\[2ex] W_q = W - \overline{X} = \dfrac{l\,\overline{X^2}}{2(1-r)} \\[2ex] N_q = \dfrac{\overline{X^2}}{2(1-r)} \end{array} \right.$$

Delay Distribution in a FCFS M/G/1 Queue

T	Total time spent in system (r.v.) by an arrival
Q	Total waiting time (r.v.) before service begins for an arrival

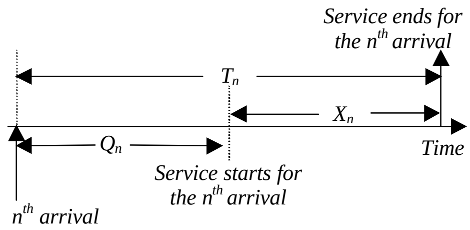
$$T = Q + X \qquad Q \perp X$$

$$L_T(s) = E\{e^{-sT}\} = L_Q(s) L_B(s)$$

$\boxed{L_B(s) \text{ is known if the distribution of the service time } X \text{ is given}}$

Consider a particular job arrival and its departure (say the n^{th} one) in a FCFS M/G/1 queue

The number of customers that the n^{th} user will see left behind in the queue when it departs will be the number of arrivals that occurred while it was in the system.



Therefore $L_T(l - lz) = P(z)$

Substituting $s = (l - lz)$ ⇨ $L_T(s) = \dfrac{s(1-r)L_B(s)}{s - l + lL_B(s)}$ (3.15)

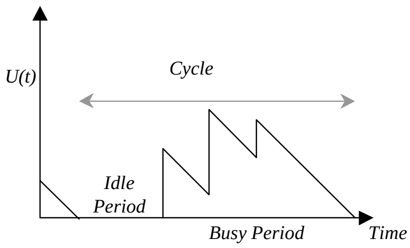
Substituting $T = Q + X$, Q^X and $L_B(s) = E\{e^{-sX}\}$

$$L_Q(s) = \frac{L_T(s)}{L_B(s)} = \frac{s(1-r)}{s - l + lL_B(s)} \qquad (3.16)$$

$L_T(s)$ and *$L_Q(s)$* are the L.T.s of the pdfs of the total delay and the queueing delay as seen by an arrival in a FCFS M/G/1 queue.

An alternate approach for deriving *$L_T(s)$* and *$L_Q(s)$* may be found in Section 3.7

Busy Period Analysis of a M/G/1 Queue (Section 3.4)



U(t)=0 during Idle Period

U(t)>0 during Busy Period

A Cycle consists of the combination of successive Idle and Busy periods

Unfinished Work *U(t)* in a M/G/1 Queue

Idle Period

Exponentially distributed with mean *1/l*

$$f_{IP}(t) = me^{-mt} \qquad t \geq 0$$

$$L_{IP}(s) = \frac{l}{s + l}$$

This will have the same distribution as an *inter-arrival time*

Busy Period

Consider a busy period that starts with the arrival of customer A_1.

Let X_1 be the service time for A_1.

Let there be n^* arrivals (A_2,................, A_{n^*+1}) that arrive during the service time X_1, in the sequence A_2,............, A_{n^*+1}.

Note that the busy period BP will consist of the sum of X_1 and n^* sub-busy periods.

Each of the sub-busy periods are i.i.d. random variables with the same distribution as that of the busy period BP (to be found)

$$BP = X_1 + BP_2 + ...+ BP_{n^*+1} \qquad BP_j \perp BP_k \qquad BP_j \perp X_1 \qquad \forall j,k$$

$$E\{e^{-s(BP)} \mid X_1 = x, n^* = k\} = e^{-sx}[L_{BP}(s)]^k$$

$$E\{e^{-s(BP)} \mid X_1 = x\} = e^{-sx}\sum_{k=0}^{\infty} e^{-lx}\frac{(lx)^k}{k!}[L_{BP}(s)]^k$$

$$= e^{-sx}e^{-lx}e^{\,lxL_{BP}(s)}$$

$$= e^{-x(s+l-lL_{BP}(s))}$$

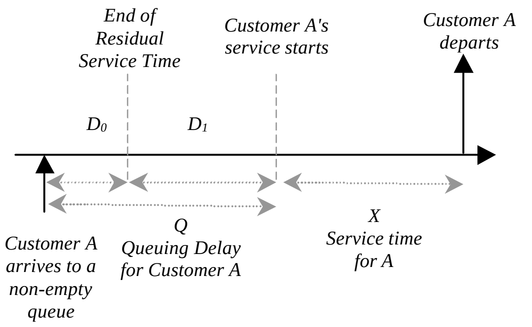
$$L_{BP}(s) = E\{e^{-s(BP)}\} = \int_{x=0}^{\infty} e^{-x[s+l-lL_{BP}(s)}b(x)dx$$

$$L_{BP}(s) = L_B\left(s + l - lL_{BP}(s)\right) \qquad (3.19)$$

Solve (3.19) to obtain $L_{BP}(s)$

The moments of BP may be obtained directly from (3.19) using the moment generating properties of the L.T. $L_{BP}(s)$. See Section 3.4 for the mean and some higher moments of BP.

Delay Distribution in a LCFS M/G/1 Queue



Customer arrival/departure instants and delays in a LCFS M/G/1 Queue

 29

Queueing Delay $Q = D_0 + D_1$ waiting time in queue before service

D_0 = Residual service time of job during whose service *A* arrives

$D_0 = 0$ if *A* arrives to an empty queue (probability = $1\text{-}r$)

$$f_{D_0}(t) = \frac{1 - B(t)}{\overline{X}} \qquad (3.20)$$

$$L_{D_0}(s) = \frac{1 - L_B(s)}{s\overline{X}} \qquad (3.21)$$

D_1 will consist of *sub busy periods*, one associated with each of the customer arrivals in D_0

Note that D_0 and D_1 are not independent of each other

 30

$$L_Q(s) = E\left\{e^{-sQ}\right\} = (1-r) + rE\left\{e^{-sQ} \middle| \mathit{arrival\ to\ non-empty\ queue}\right\}$$
$$= (1-r) + rE\left\{\exp(-s(D_0 + D_1)) \middle| \mathit{arrival\ to\ non-empty\ queue}\right\}$$

For the case where *the arrival A comes to a non-empty queue*

$$E\left\{e^{-sD_1} \mid D_0 = y, N_0 = n\right\} = [L_{BP}(s)]^n$$
$$E\left\{e^{-sQ} \mid D_0 = y, N_0 = n\right\} = e^{-sy}[L_{BP}(s)]^n$$

which leads to

$$E\left\{e^{-sD_1} \mid D_0 = y\right\} = \sum_{n=0}^{\infty} \frac{(ly)^n}{n!} e^{-ly}[L_{BP}(s)]^n = \exp[-y\{l - lL_{BP}(s)\}]$$
$$E\left\{e^{-sQ} \mid D_0 = y\right\} = \exp[-y\{s + l - lL_{BP}(s)\}]$$

 31

For the case where *the arrival A comes to a non-empty queue*

$$E\left\{e^{-sD_1}\right\} = \int_{y=0}^{\infty} E\left\{e^{-sD_1} \middle| D_0 = y\right\} f_{D_0}(y)dy$$
$$= \int_{y=0}^{\infty} \left[\exp[-y\{l - lL_{BP}(s)\}\right] f_{D_0}(y)dy$$
$$= L_{D_0}\left(l - lL_{BP}(s)\right)$$

Using (3.21), we then get

$$E\left\{e^{-sD_1}\right\} = \frac{1 - L_B(l - lL_{BP}(s))}{\overline{X}(l - lL_{BP}(s))} \qquad (3.22)$$

 32

Similarly

For the case where *the arrival A comes to a non-empty queue*

$$E\{e^{-sQ}\} = \int_{y=0}^{\infty} E\{e^{-sQ} \mid D_0 = y\} f_{D_0}(y)dy$$

$$= \int_{y=0}^{\infty} [\exp[-y\{s + l - lL_{BP}(s)\}] f_{D_0}(y)dy$$

$$= L_{D_0}\left(s + l - lL_{BP}(s)\right)$$

Using (3.21), we then get

$$E\{e^{-sQ}\} = \frac{1 - L_B(s + l - lL_{BP}(s))}{\overline{X}(s + l - lL_{BP}(s))} = \frac{1 - L_{BP}(s)}{\overline{X}(s + l - lL_{BP}(s))} \qquad (3.23)$$

 33

Therefore, considering both the cases where Customer A finds the queue empty and non-empty -

$$L_Q(s) = (1 - r) + r\,\frac{1 - L_{BP}(s)}{(s + l - lL_{BP}(s))\overline{X}} \qquad (3.24)$$

and

$$L_T(s) = L_Q(s) L_B(s) \qquad (3.25)$$

$L_T(s)$ and *$L_Q(s)$* are the L.T.s of the pdfs of the total delay and the queueing delay as seen by an arrival in a LCFS M/G/1 queue.

The results obtained for the M/G/1 queue may be used to obtain the delay distributions for the M/D/1 queue as well. This is given in Section 3.6.

 34