Basic Queueing Theory

M/M/-/- Type Queues

Lecture 3

- Kendall’s Notation for Queues

A/B/C/D/E

- Shorthand notation where A, B, C, D, E describe the queue
- Applicable to a large number of simple queueing scenarios

Kendall's Notation for Queues
A/B/C/D/E

A	Inter-arrival time distribution	M: exponential D: deterministic E_k : Erlangian (order k) G : general
B	Service time distribution	
C	Number of servers	
D	Maximum number of jobs that can be there in the system (waiting and in service) *Default ∞ for infinite number of waiting positions*	
E	Queueing Discipline (FCFS, LCFS, SIRO etc.) *Default is FCFS*	

Examples

- M/M/1 or M/M/1/∞
  - Single server queue with Poisson arrivals, exponentially distributed service times and infinite number of waiting positions
- $M/E_2/2/K$
  - Poisson Arrivals, Erlangian of order-2 Service time distribution, two servers, maximum number K in system (waiting and in service)
- G/M/2
  - Generalized Arrivals, Exponentially service time distribution, 2 servers, infinite number of waiting positions

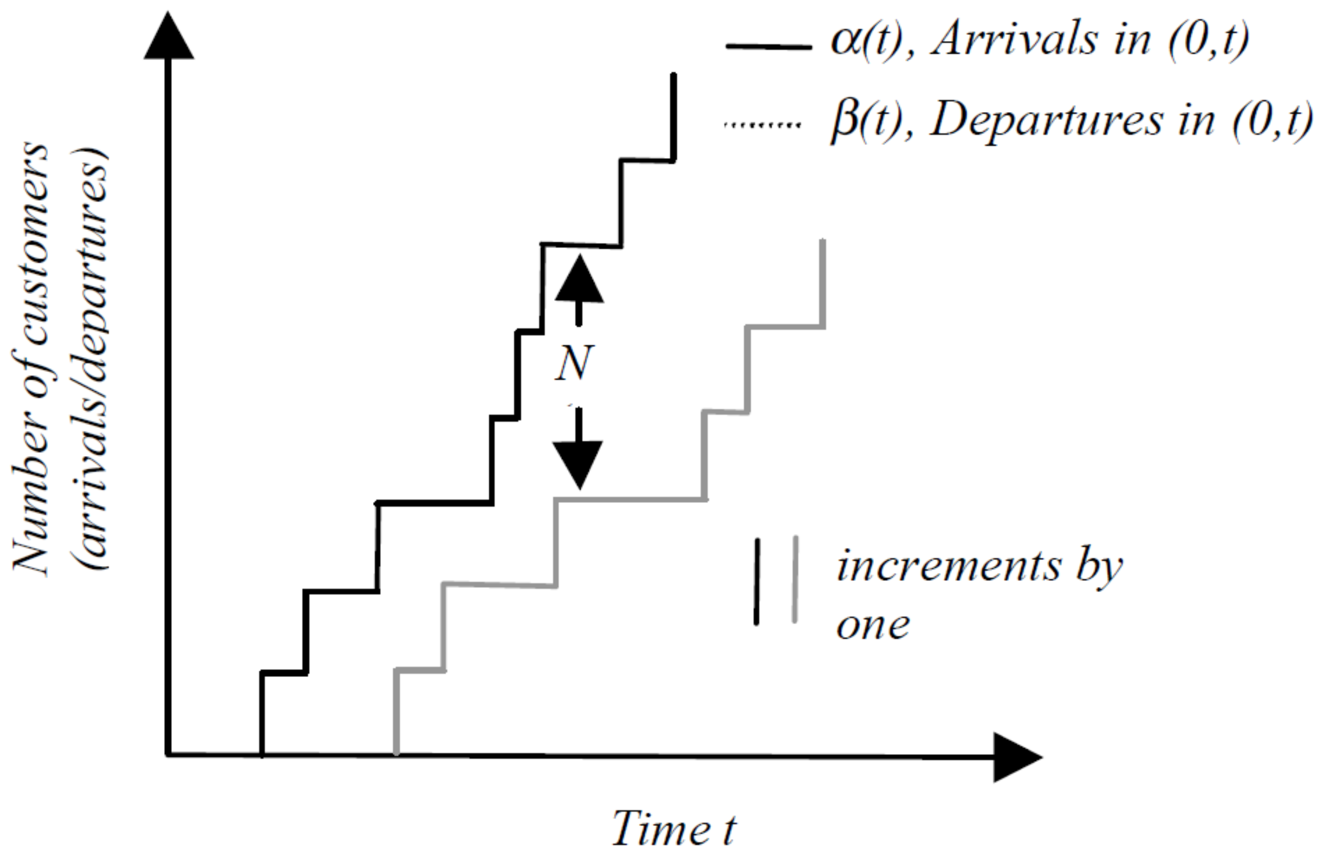
- Little's Result

$$N = \lambda W \quad (2.9)$$

$$N_q = \lambda W_q \quad (2.10)$$

Result holds in general for virtually all types of queueing situations where
λ = Mean arrival rate of jobs that actually enter the system

Jobs blocked and refused entry into the system will not be counted in λ



Graphical Illustration/Verification of Little's Result

- Consider the time interval (*0,t*) where *t* is large, i.e. $t \to \infty$
- *Area(t)* = area between $\alpha(t)$ and $\beta(t)$ at time *t* =

$$\int_0^t [\alpha(t) - \beta(t)]dt$$

- Average Time *W* spent in system = $$\lim_{t\to\infty} \frac{Area(t)}{\alpha(t)}$$
- Average Number *N* in system
- Since, $$= \lim_{t\to\infty} \frac{Area(t)}{t} = \lim_{t\to\infty} \frac{\alpha(t)}{t} \frac{Area(t)}{\alpha(t)}$$

$$\lambda = \lim_{t\to\infty} \frac{\alpha(t)}{t}$$

Therefore, $N = \lambda W$

The PASTA Property

- “Poisson Arrivals See Time Averages”

$p_k(t)$ =P{system is in state k at time t }

$q_k(t)$ =P{an arrival at time t finds the system in state k}

$N(t)$ be the actual number in the system at time t

$A(t,\ t+\Delta t)$ be the event of an arrival in the time interval $(t,\ t+\Delta t)$

Then

$$q_k(t) = \lim_{\Delta t \to 0} P\{N(t) = k \mid A(t, t + \Delta t\}$$

$$= \lim_{\Delta t \to 0} \frac{P\{A(t, t + \Delta t \mid N(t) = k \mid\} P\{N(t) = k\}}{P\{A(t, t + \Delta t\}} = p_k(t)$$

because $\mathrm{P}\{A(t,\ t+\Delta t) | N(t) = k\} = \mathrm{P}\{A(t,\ t+\Delta t)\}$

Equilibrium Solutions for M/M/-/- Queues

- Method 1: Obtain the differential-difference equations as in Section 1.2 or Section 2.2. Solve these under equilibrium conditions along with the normalization condition.

- Method 2: Directly write the flow balance equations for proper choice of closed boundaries as illustrated in Section 2.2 and solve these along with the normalization condition.

- *Method 3*: Identify the parameters of the birth-death Markov chain for the queue and directly use equations (2.7) and (2.8) as given in Section 2.2.

In the following, we have used this approach

M/M/1 (or M/M/1/∞) Queue

$$\lambda_k = \lambda \qquad \forall k$$

$$\mu_k = 0 \qquad k = 0$$

$$= \mu \qquad k = 1, 2, 3, \ldots\ldots$$

For $\rho < 1$

$$p_k = p_0 \left(\frac{\lambda}{\mu}\right)^k = p_0 \rho^k$$

$$p_0 = (1 - \rho)$$

$$N = \sum_{i=0}^{\infty} i p_i = \sum_{i=0}^{\infty} i \rho^i (1 - \rho) = \frac{\rho}{1 - \rho}$$

$$W = \frac{N}{\lambda} = \frac{1}{\mu(1 - \rho)}$$

Using Little's Result

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

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

Using Little's Result

M/M/1/∞ Queue with Discouraged Arrivals

$$\lambda_k = \frac{\lambda}{k+1} \qquad \forall k$$

$$\mu_k = 0 \qquad k = 0$$
$$= \mu \qquad k = 1, 2, 3, \ldots\ldots$$

For $\rho = \lambda/\mu < \infty$

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu(i+1)} = p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \qquad (2.14)$$

$$p_0 = \exp\left(-\frac{\lambda}{\mu}\right) \qquad (2.15)$$

$$N = \sum_{k=0}^{\infty} k p_k = \frac{\lambda}{\mu}$$

$$W = \frac{N}{\lambda_{eff}} = \frac{\lambda}{\mu^2 \left[1 - \exp\left(-\frac{\lambda}{\mu}\right)\right]}$$

Little's Result

$$\lambda_{eff} = \sum_{k=0}^{\infty} \lambda_k p_k = \mu \left[1 - \exp\left(-\frac{\lambda}{\mu}\right)\right]$$

Effective Arrival Rate

M/M/1/∞ Queue with Discouraged Arrivals

- In this case, *PASTA is not applicable* as the overall arrival process is not Poisson

π_r = P{arriving customer sees r in system (before joining the system)}

ΔE be the event of an arrival in $(t, t+\Delta t)$

E_i is the event of the system being in state i

$$P\{E_i\} = p_i = e^{-\lambda/\mu}\left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}$$

$$P\{\Delta E \mid E_i\} = \frac{\lambda \Delta t}{i+1}$$

$$\pi_r = P\{E_r \mid \Delta E\} = \frac{P\{E_r\}P\{\Delta E \mid E_r\}}{P\{\Delta E\}} = \frac{P\{E_r\}P\{\Delta E \mid E_r\}}{\sum_{i=0}^{\infty} P\{E_i\}P\{\Delta E \mid E_i\}}$$

$$\pi_r = \left(\frac{\lambda}{\mu}\right)^{r+1} \frac{1}{(r+1)!}\left(\frac{e^{-\lambda/\mu}}{1-e^{-\lambda/\mu}}\right) \qquad W = \sum_{k=0}^{\infty} \frac{k+1}{\mu}\pi_k = \frac{\lambda}{\mu^2(1-e^{-\lambda/\mu})}$$

- M/M/m/∞ Queue (*m* servers, infinite number of waiting positions)

$$\lambda_k = \lambda \qquad \forall k \qquad \mu_k = k\mu \qquad 0 \le k \le (m-1)$$
$$= m\mu \qquad k \ge m$$

For $\rho = \lambda/\mu < m$

$$p_k = p_0 \frac{\rho^k}{k!} \qquad for \quad k \le m$$
$$= p_0 \frac{\rho^k}{m!\,m^{k-m}} \qquad for \quad k > m \tag{2.16}$$

$$p_0 = \left(\sum_{k=0}^{m-1} \frac{\rho^k}{k!} + \frac{m\rho^m}{m!(m-\rho)} \right)^{-1} \tag{2.17}$$

Erlang's C-Formula

$$P\{queueing\} = \sum_{k=m}^{\infty} p_k = C(m,\rho) = p_0 \frac{m\rho^m}{m!(m-\rho)} \tag{2.18}$$

- M/M/m/m Queue (m server loss system, no waiting)

$$\lambda_k = \lambda \qquad k < m$$
$$= 0 \qquad otherwise \quad (Blocking\ \ or\ Loss\ Condition)$$
$$\mu_k = k\mu \qquad 0 \le k \le m$$
$$= 0 \qquad otherwise$$

For $\rho = \dfrac{\lambda}{\mu} < \infty$

$$\begin{cases} p_k = p_0 \dfrac{\rho^k}{k!} & for \quad k \le m \qquad (2.19) \\ \quad = 0 & otherwise \\ p_0 = \dfrac{1}{\sum_{k=0}^{m} \dfrac{\rho^k}{k!}} & \qquad (2.20) \end{cases}$$

- M/M/m/m Queue (m server loss system, no waiting)

Simple model for a telephone exchange where a line is given only if one is available; otherwise the call is lost

Blocking Probability B(m,ρ)
= P{an arrival finds all servers busy and leaves without service}

$$B(m,\rho) = p_0 \frac{\rho^m}{m!} \qquad \text{Erlang's B-Formula} \tag{2.21}$$

$$B(0,\rho) = 1 \qquad B(m,\rho) = \frac{\dfrac{\rho B(m-1,\rho)}{m}}{1 + \dfrac{\rho B(m-1,\rho)}{m}} \tag{2.22}$$

- M/M/1/K Queue (single server queue with K-1 waiting positions)

$$\lambda_k = \lambda \qquad k < K$$
$$= 0 \qquad otherwise \quad (Blocking\ or\ Loss\ Condition)$$
$$\mu_k = \mu \qquad k \le K$$
$$= 0 \qquad otherwise$$

For

$$\rho = \frac{\lambda}{\mu} < \infty$$

$$p_k = p_0 \rho^k \qquad for \quad k \le K \qquad (2.23)$$
$$= 0 \qquad otherwise$$

$$p_0 = \frac{(1-\rho)}{(1-\rho^{K+1})} \qquad (2.24)$$

- M/M/1/-/K Queue (single server, infinite number of waiting positions, finite customer population *K*)

$$\lambda_k = \lambda(K-k) \qquad k < K$$
$$= 0 \qquad otherwise \quad (Blocking \ \ or\ Loss\ Condition)$$
$$\mu_k = \mu \qquad k \le K$$
$$= 0 \qquad otherwise$$

For

$$p_k = p_0 \rho^k \frac{K!}{(K-k)!} \qquad k=1,\ldots..,K \qquad (2.25)$$

$$\rho = \frac{\lambda}{\mu} < \infty \qquad p_0 = \frac{1}{\sum_{k=0}^{K} \rho^k \frac{K!}{(K-k)!}} \qquad (2.26)$$

Delay Analysis for a FCFS M/M/1/∞ Queue (Section 2.6.1)

Q: Queueing Delay (not counting service time for an arrival

pdf $f_Q(t)$, cdf $F_Q(t)$, $L_Q(s) = LT(f_Q(t))$

W=Q+T W: Total Delay (waiting time and service time) for an arrival

pdf $f_W(t)$, cdf $F_W(t)$, $L_W(s) = LT(f_W(t))$

T: Service Time

$$f_T(t) = \mu e^{-\mu t} \quad F_T(t) = e^{-\mu t} \quad L_T(s) = \frac{\mu}{(s+\mu)}$$

Since $Q \perp T$

$$L_W(s) = \frac{\mu}{(s+\mu)} L_Q(s) \quad f_W(t) = f_Q(t) * [\mu e^{-\mu t}] \qquad (2.30)$$

Knowing the distribution of either *W* or *Q*, the distribution of the other may be found

- For a particular arrival of interest -

$F_Q(t)$ = P{queueing delay $\le t$}
= P{queueing time=0} + [$\Sigma_{n\ge 1}$ P{queueing time$\le t$ | arrival found n jobs in system}]p_n

Erlang-n distribution for sum of n exponential r.v.s

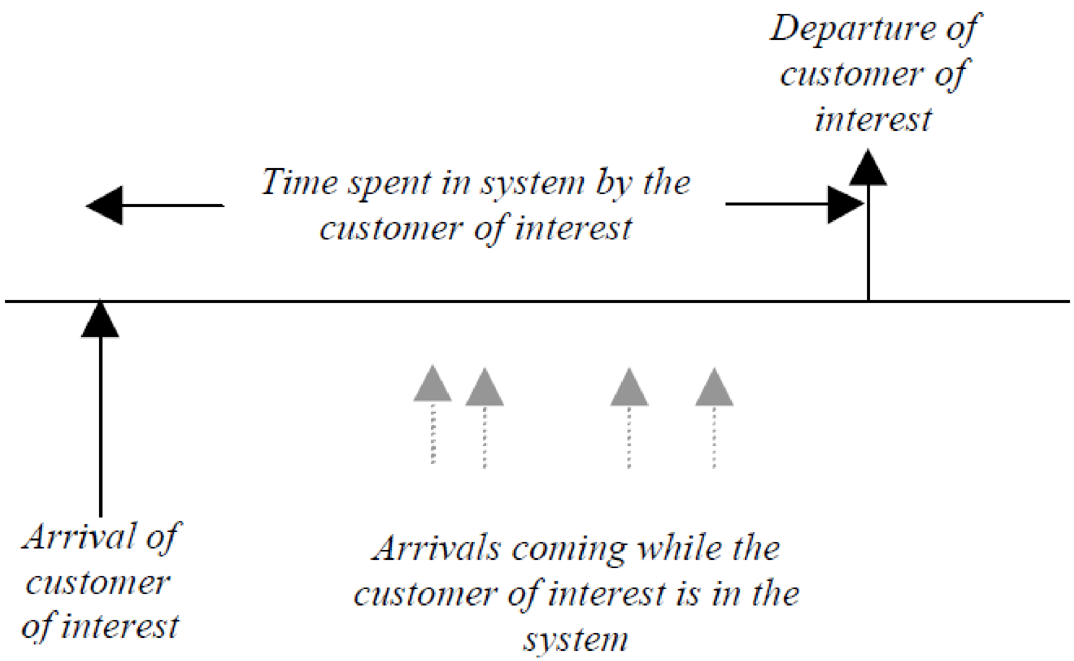
$$F_Q(t) = (1-\rho) + (1-\rho)\sum_{n=1}^{\infty}\rho^n \int_{x=0}^{t} \frac{\mu(\mu x)^{n-1}}{(n-1)!} e^{-\mu x} dx$$

$$= (1-\rho) + (1-\rho)\rho \int_0^t \mu e^{-\mu x} \sum_{n=1}^{\infty} \frac{(\mu x \rho)^{n-1}}{(n-1)!} dx \qquad (2.31)$$

$$= (1-\rho) + (1-\rho)\rho \int_0^t \mu e^{-\mu x(1-\rho)} dx = (1-\rho) + \rho(1-e^{-\mu t(1-\rho)})$$

$$f_Q(t) = \frac{dF_Q(t)}{dt} = \delta(t)(1-\rho) + \lambda(1-\rho)e^{-\mu t(1-\rho)} \qquad (2.32)$$

$$f_W(t) = (1-\rho)\mu e^{-\mu t} + \lambda(1-\rho)\mu \int_0^t e^{-\mu(1-\rho)(t-x)} e^{-\mu x} dx = (\mu-\lambda)e^{-(\mu-\lambda)t}$$



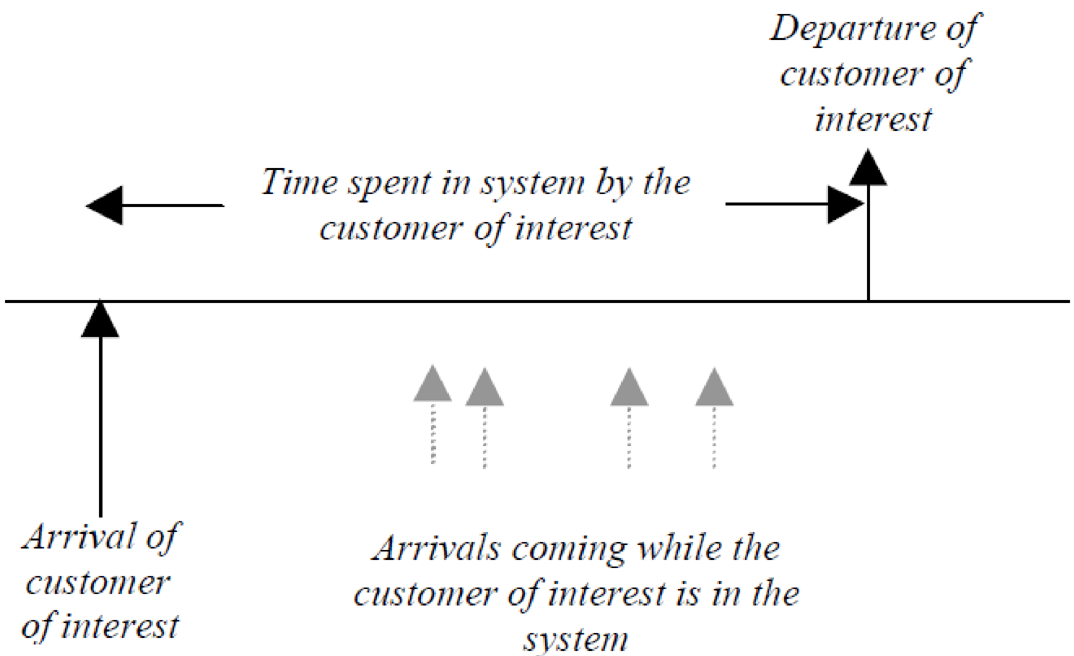
PASTA applicable to this queue
N and N_Q seen by an arrival same as the time-averaged values

Arrival/Departure of Customer/Job of Interest from a FCFS M/M/1Queue

Let

N^* = Number in the system that a job will see left behind when it departs

$p_n^* = P\{N^* = n\}$ for N^*=0, 1,....,∞



For a FCFS queue, number left behind by a job will be equal to the number arriving while it is in the system.

$$G^*(z) = \sum_{n=0}^{\infty} z^n p_n^* = \sum_{n=0}^{\infty} z^n \int_{t=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_W(t)dt \qquad (2.36)$$

$$= \int_0^{\infty} e^{-\lambda t(1-z)} f_W(t)dt = L_W(\lambda - \lambda z)$$

$$E\{N^*\} = \left.\frac{dG^*(z)}{dz}\right|_{z=1} = -\lambda \left.\frac{dL_W(s)}{ds}\right|_{s=0} = \lambda W \qquad (2.37)$$

An important general observation can also be made along the lines of Eq. (2.36).

Consider the number arriving from a Poisson process with rate λ in a random time interval T where $L_T(s)=L_T\{f_T(t)\}$. The generating function $G(z)$ of this will be given by

$$G(z) = L_T(\lambda - \lambda z)$$

and the mean number will be $E\{N\}= \lambda E\{T\}$

This result will be found to be useful in various places in our subsequent analysis.

Delay Analysis for the FCFS M/M/m/∝ Queue (Section 2.6.2)

Using an approach similar to that used for the M/M/1 queue, we obtain the following

$$f_Q(t) = \left\{1 - p_0\left[\frac{m\rho^m}{m!(m-\rho)}\right]\right\}\delta(t) + \left[\frac{\mu p_0 \rho^m e^{-\mu(m-\rho)t}}{(m-1)!}\right]u(t)$$

$$f_W(t) = \left\{1 - p_0\left[\frac{m\rho^m}{m!(m-\rho)}\right]\right\}\mu e^{-\mu t} - \left[\frac{\mu p_0 \rho^m [e^{-\mu(m-\rho)t} - e^{-\mu t}]}{(m-1)!(1-m-\rho)}\right]$$

See Section 2.6.2 for the details and the intermediate steps

Example

- Analyze a queue with a single server where the average arrival rate of customers in (N-i)λ per second from a Poisson process, when the system is in state i. Assume that service time required by a customer is exponentially distributed with mean $1/\mu$ seconds. Assume that N is the highest state of the system.

Example

- Ans:
- Given that

$$\lambda_i = (N-i)\lambda \qquad i = 0,1,2,\ldots\ldots.,N$$

$$\mu_i = \mu \qquad i = 1,2,\ldots\ldots\ldots.,N$$

- the flow balance equations may then be written as

$$N\lambda p_0 = \mu p_1 \qquad p_1 = N\rho p_0 \qquad \rho = \frac{\lambda}{\mu}$$

$$(N-k)\lambda p_k = \mu p_{k+1} \qquad p_{k+1} = (N-k)\rho p_k$$

$$\lambda p_{N-1} = \mu p_N \qquad p_N = \rho p_{N-1}$$

$$\text{or} \quad p_k = \frac{N!}{(N-k)!}\rho^k p_0 \qquad \text{for } k=0,1,2.........N$$

The probability p_0 may be found by applying the normalization condition that $\sum_{k=0}^{\infty} p_k = 1$

$$p_0 \sum_{k=0}^{\infty} \frac{N!}{(N-k)!}\rho^k = 1 \quad \text{giving} \quad p_0 = \frac{1}{\sum_{k=0}^{\infty} \frac{N!}{(N-k)!}\rho^k}$$

Example

- Consider a M/M/2/4 queue at equilibrium. Its state probabilities are observed to be 1/16, 4/16, 6/16, 4/16 and 1/16 respectively for system states 0,1,2,3 and 4. For this queue, determine N and N_q. If the mean arrival rate (from a Poisson process) is observed to be 2 customers per hour, determine the mean delay quantities W and W_q and estimate the mean service time.

- Ans:

Using the system state probabilities, we get

$$N = \frac{4}{16} + \frac{12}{16} + \frac{12}{16} + \frac{4}{16} = 2 \quad \text{and} \quad N_q = \frac{4}{16} + \frac{2}{16} = \frac{3}{8}$$

The effective arrival rate $: \lambda_{eff} = 2\left(\frac{15}{16}\right) = \frac{15}{8}$ Using this,

$$W = \frac{16}{15} \text{ and } W_q = \frac{1}{5}$$

Example

- Analyze a M/M/1/ queue with the following parameters

  $\lambda_k = \alpha^k \lambda$ for k=0,1,2,….. And $0 <= \alpha <= 1$
  $\mu_k = \mu$ for k=1,2,3

  Obtain the steady state distribution of the queue and the conditions under which such a distribution will exist.

- Ans:

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = p_0 \rho^k \alpha^{\left(\sum_{i=1}^{k-1} i\right)} = p_0 \rho^k \alpha^{\frac{k(k-1)}{2}} \quad \text{with } \rho = \frac{\lambda}{\mu},\ k=1,2,\ldots\infty$$

and $$p_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \rho^k \alpha^{\frac{k(k-1)}{2}}}$$

from the normalization conditions

For ensuring the existence of the equilibrium solution, we need

$$\alpha = \sum_{k=0}^{\infty}\left[\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}\right] = \sum_{k=0}^{\infty}\rho^k\alpha^{\frac{k(k-1)}{2}} < \infty$$

$$\beta = \sum_{k=0}^{\infty}\frac{1}{\lambda_k\prod_{i=0}^{k-1}\frac{\lambda_i}{\mu_{i+1}}} = \frac{1}{\lambda}\sum_{k=0}^{\infty}\rho^k\alpha^{\frac{k(k-1)}{2}} = \infty$$