❑ Simple queuing theory – an introduction

- Queueing system is characterised by
  1. Inter-arrival probability density function
  2. Service time probability density function
  3. Number of servers
  4. The queueing discipline
  5. Amount of buffer space in the queue
- A/B/m queue: A is the inter-arrival probability density function
  B is the service time probanility density
  m is the number of servers
- A & B can be M: Exponential (Markovian system)
  D: All inputs have same known values for inter-arrival & service times (Deterministic)
  G: General, i.e. arbitrary probability distribution
- We will concentrate on the most widely used one – M/M/1 queue
- Assumptions : Infinite number of entities (frames, packets, messages
  Probability of entity/ message arrival within a time interval is dependent only on this time interval
  Input entities, i.e. packets, messages, etc., are independent of each other
- Assuming exponential probability density for inter-arrival time is valid under above assumptions

❑ Simple queuing theory – an introduction

- Poisson's distribution: (Poisson's law)
  For mean arrival rate of λ (1/λ is the mean inter-arrival time) probability $P_n(t)$ of exactly n messages arriving in an interval t is given by

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

- Exponential inter-arrival time
  Probability $a(t)\Delta t$ that inter-arrival time is between t & t+Δt is actually (probability of no arrivals for a time t) x (probability of exactly one arrival in time interval Δt)
  Thus $a(t)\Delta t$ is given by

$$a(t)\Delta t = P_0(t)P_1(\Delta t) = \left(e^{-\lambda t}\right)\left(\lambda\Delta t e^{-\lambda\Delta t}\right)$$

As $\Delta t \rightarrow 0$ we get

$$\Delta t \rightarrow 0, e^{-\lambda\Delta t} \rightarrow 1, \text{so} \lim_{\Delta t \rightarrow 0} a(t)\Delta t = \lambda e^{-\lambda t}dt$$

- Exponential service time
  Following similar reasoning it can be asserted that for mean service rate of μ (mean time to finish servicing a single message is1/μ), probability to finish servicing a message within a time interval Δt is μΔt and service time probability density is also exponential, given by

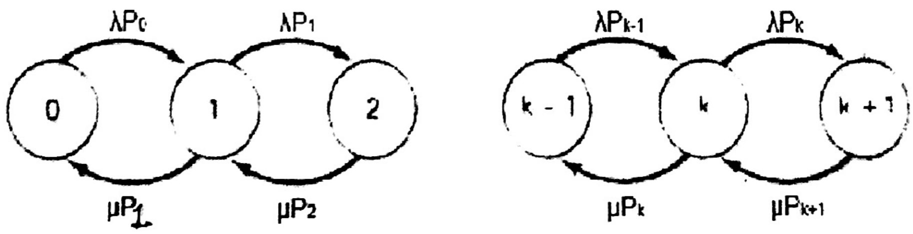
$$\mu e^{-\mu t}dt$$

❑ Simple queuing theory – an introduction

- State of a M/M/1queue
  - Completely defined by state no. k when there are k messages in the system, i.e., k-1 in the queue & 1 in the server. Remaining service time of the one in server is not relevant as exponential density function has no memory
  - P_k : Equilibrium probability that there are exactly messages in the system (k-1 & k)
    At equilibrium this is invariant with time
  - Queue state transition: Arrival of new message moves system in state 'i' to state 'i+1'
    Departure of a processed message causes a system in state 'i' to move to state 'i-1', $i-1 \geq 0$
- Markov's Birth & Death model
  - Key assumptions: 1.Mean no. transitions from state k to k+1 must be the same as transitions from state k+1 to state k (for some k) to maintain equilibrium
    2.Average arrival rate is λ packets/ unit time
    3.Average service rate is μ packets/ per unit time
    4.At equilibrium λ must have the same value as μ
    P_k is the probabilty of a system at equilibrium being in state k when there is an arrival or a departure

❑ Simple queuing theory – an introduction

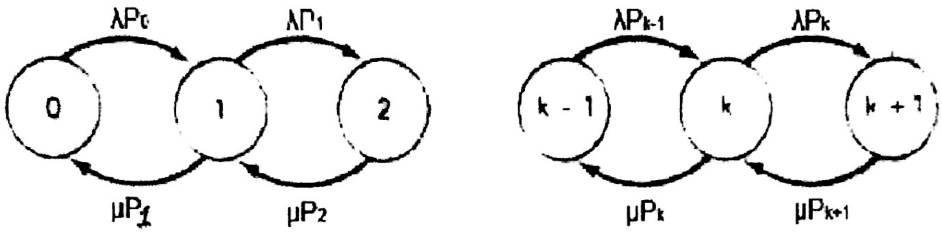
- The Markov Chain



- Out of λ packets arriving/ unit time, λP_k of them will do so when system is in state k & each will trigger a state transition from k to k+1
- Of μ serviced packets leaving/ unit time μP_k of them will find system in state k & each will cause a transition from state k to k-1 (k= 0,1,..., k)
- At equilibrium transitions from state k to k+1 must be equal to k+1 → k state transitions over a period of time

❑ Simple queuing theory – an introduction

- The Markov Chain



- At equilibrium

$$\lambda P_0 = \mu P_1 \text{ or } P_1 = \left(\frac{\lambda}{\mu}\right) P_0$$

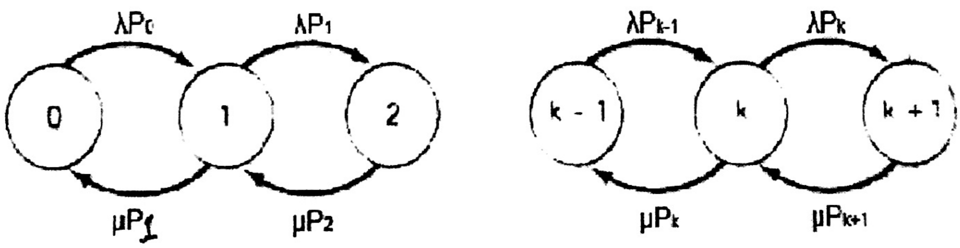
$$\lambda P_1 = \mu P_2 \text{ or } P_2 = \left(\frac{\lambda}{\mu}\right) P_1 = (\lambda/\mu)^2 P_0$$

$$\lambda P_2 = \mu P_3 \text{ or } P_3 = \left(\frac{\lambda}{\mu}\right) P_2 = (\lambda/\mu)^3 P_0$$

$$\lambda P_{k-1} = \mu P_k \text{ or } P_k = \left(\frac{\lambda}{\mu}\right) P_{k-1} = (\lambda/\mu)^k P_0$$

❑ Simple queuing theory – an introduction

- The Markov Chain



- At equilibrium

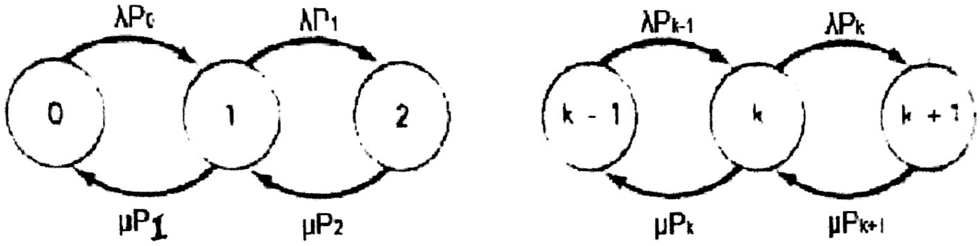
$$\left(\frac{\lambda}{\mu}\right) \text{ is } \frac{\text{average arrival rate}}{\text{average service rate}} = \rho \text{ (say) } \leq 1$$

Also sum of probabilities

$$\sum_{k=0} P_K = 1 \text{ or } \sum_{k=0} (\lambda/\mu)^k P_0 = \sum_{k=0} \rho^k P_0 = 1$$

❑ Simple queuing theory – an introduction

- The Markov Chain



We know (G.P. series) $\sum_{k=0} \rho^k - \dfrac{1}{1-\rho}$

So $\dfrac{P_0}{1-\rho} = 1$ or $P_0 = (1-\rho)$

Rewriting $\rho = (1 - P_0)$ which is the probability that the system is not idle

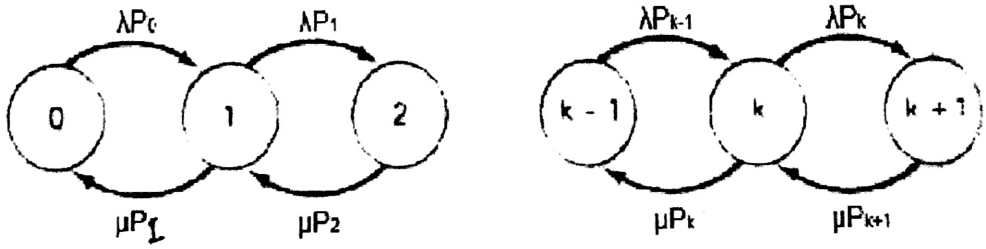
$$P_k = \rho^k P_0 = \rho^k(1-\rho)$$

Mean number of packets in the system is given by

$$N = \sum_{k=0} kP_K = (1-\rho)\sum_{k=0} k\rho^k$$

❑ Simple queuing theory – an introduction

- The Markov Chain



$$N = \sum_{k=0} kP_K = (1-\rho)\sum_{k=0} k\rho^k$$

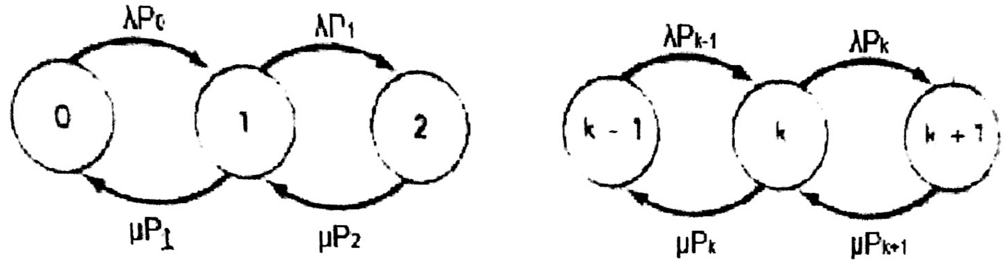
This eqn., can be solved by differentiating $\sum_{k=0} \rho^k = \dfrac{1}{1-\rho}$ with respect to ρ

We get $\sum_{k=0} k\rho^{k-1} = \dfrac{1}{(1-\rho)^2}$, multiplying both sides by ρ we get

$$\sum_{k=0} k\rho^k = \frac{\rho}{(1-\rho)^2}$$

❑ Simple queuing theory – an introduction

- The Markov Chain



Putting $\sum_{k=0} k\rho^k - \frac{\rho}{(1-\rho)^2}$ in $N = (1-\rho)\sum_{k=0} k\rho^k$ we get

$N = \frac{(1-\rho)\rho}{(1-\rho)^2} = \frac{\rho}{(1-\rho)}$ which is intutively ok, as $\rho = \frac{\lambda}{\mu} \to 1$ queue length N grows very rapidly

- Delay/ waiting time (Little's law)

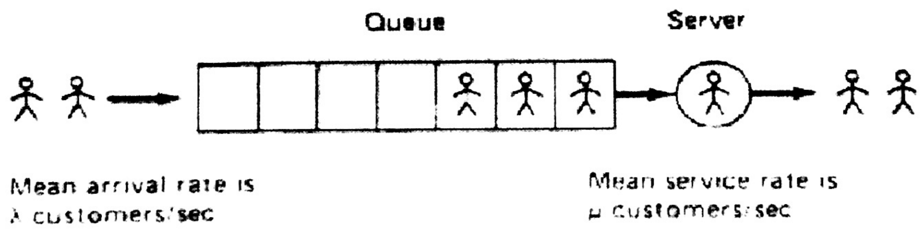
  Let mean waiting time for all packets be T

  A particular packet marked just as it enters the system

  Number of packets arriving during the time the marked packet stays in the system is λT

❑ Simple queuing theory – an introduction

- Little's law/ Theorem



Number of packets in the system at the point of time when the marked packet leaves is λT

So mean number of packets in the system is given by

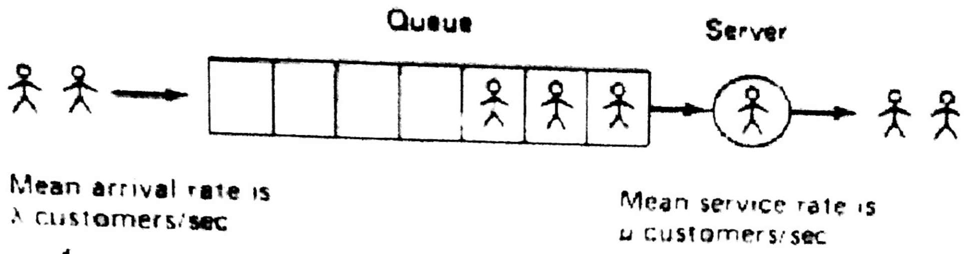
$$N = \lambda T \text{ or } T = {}^{N}/_{\lambda} = \frac{{}^{\rho}/_{\lambda}}{(1-\rho)} = \frac{({}^{\lambda}/_{\mu})/\lambda}{1-({}^{\lambda}/_{\mu})}$$

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

So mean wait time $T = \frac{1}{\mu-\lambda}$

❑ Simple queuing theory – an introduction

▪ Little's law/ Theorem



$$T = \frac{1}{\mu - \lambda}$$

λ: Meaningful to specify in packets/ messages per unit time/ second; transmitting station will never normally transmit fragment of a packet or message

μ: above expression for T assumes that μ is also expressed in packets/ messages, but packet/ message size varies from one system to another; usual way of specifying service rate is bits per unit time/ second

Assumptions: Packet size x exponentially distributed with density function $\mu e^{-\mu x}$ with mean of 1/μ bits/packet

Channel capacity is C bits/ sec, i.e., service rate is C/(1/μ) = Cμ

Thus now wait time becomes $T = \frac{1}{\mu C - \lambda}$