

Ch.30 Queueing Theory

- Stochastic process : random variable with time
- Discrete/continuous-time : when to change
Discrete/continuous-state : which values to take
 - Changes in the stock price : discrete-time discrete-state
 - Wind speed : continuous-time continuous-state
- # customers, waiting time in a queueing system
 - $n(t)$ is a discrete-state stochastic process = stochastic chain
 - $w(t)$ is a continuous-state stochastic process

State Transition

- Transition probability matrix

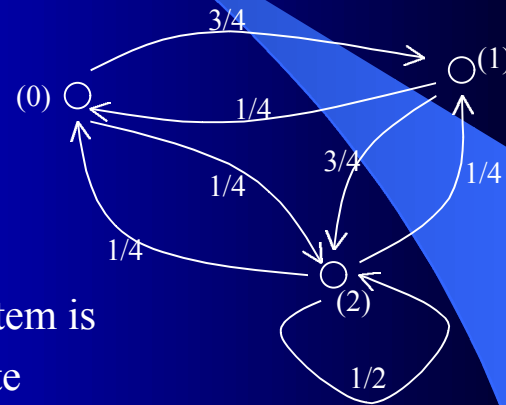
$$P = \begin{bmatrix} 0 & 3/4 & 1/4 \\ 1/4 & 0 & 3/4 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$$

- $\pi = [\pi_0, \pi_1, \pi_2]$

where π_i : probability that the system is
in state E_i at steady-state

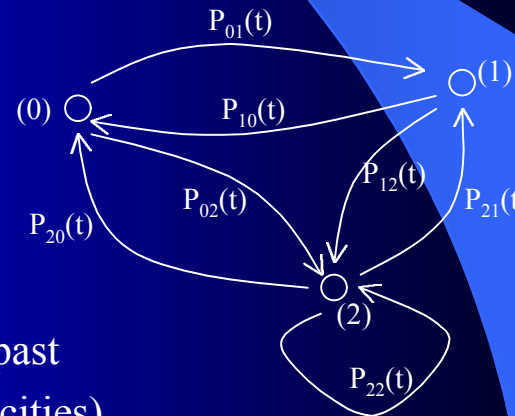
- At steady-state, $\pi = \pi P$

$$\Rightarrow \pi = [0.20, 0.28, 0.52]$$



Markov Process

- The previous one was
“Homogeneous Markov process” (discrete-state discrete-time)
 - Future states of a process are independent of the past
- Non-homogeneous Markov process:
when the transition probabilities are
functions of time : $P_{ij}(t)$
- Non-Markov Process:
when future states are dependent of the past
(think of a traveling salesman visiting 3 cities)



Markov Process (cont'd)

- Transient solution of the state transition

- Let $\pi_i^{(n)}$ be the probability that the system is in state E_i after n steps

- $\pi^{(n)} = [\pi_0^{(n)}, \pi_1^{(n)}, \pi_2^{(n)}]$

- $= \pi^{(n-1)} P$

- $= \pi^{(0)} P^n$

- If $\pi^{(0)} = [1, 0, 0]$, then

n	0	1	2	∞
$\pi_0^{(n)}$	1	0	0.250		0.20
$\pi_1^{(n)}$	0	0.75	0.062		0.28
$\pi_2^{(n)}$	0	0.25	0.688		0.52

- If $\pi^{(0)} = [0, 1, 0]$, then

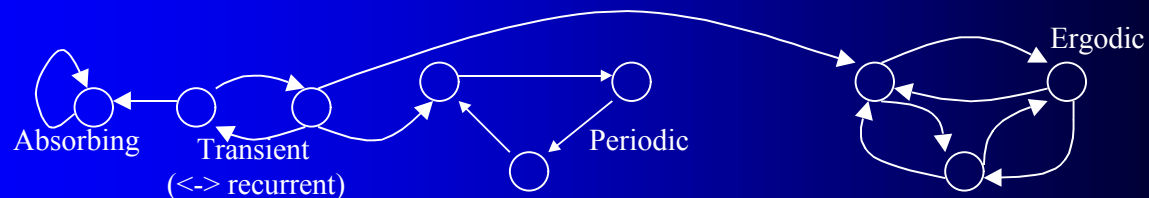
- If $\pi^{(0)} = [0, 0, 1]$, then

Markov Process (cont'd)

- Why should the equilibrium distribution be independent of the starting state ? Why should the distribution converge ?

⇒ An irreducible, ergodic Markov chain has a limiting distribution

- * Markov chain = discrete-state Markov process
- * MC is irreducible when every state is reachable from every other state
- * MC is ergodic if all states are ergodic
- * A state is ergodic when it is recurrent non-null & aperiodic



Markov Process (cont'd)

- So, what is Markov ?
 - Future state is independent of the past
 - => Future transition time is indep. on how long it has been in the current state
 - => “How long will it stay in the current state” is memoryless
 - => “Time between transitions (τ)” is memoryless
 - => “New arrival or new departure” is memoryless

Markov Process (cont'd)

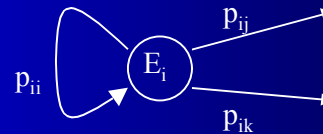
- Discrete-time Markov chain:

- Markov property can be obtained by “one-step transition”

- $\Pr\{\text{system remains } E_i \text{ during the next } m \text{ steps \& leave}\}$

$$= (1-p_{ii}) p_{ii}^m$$

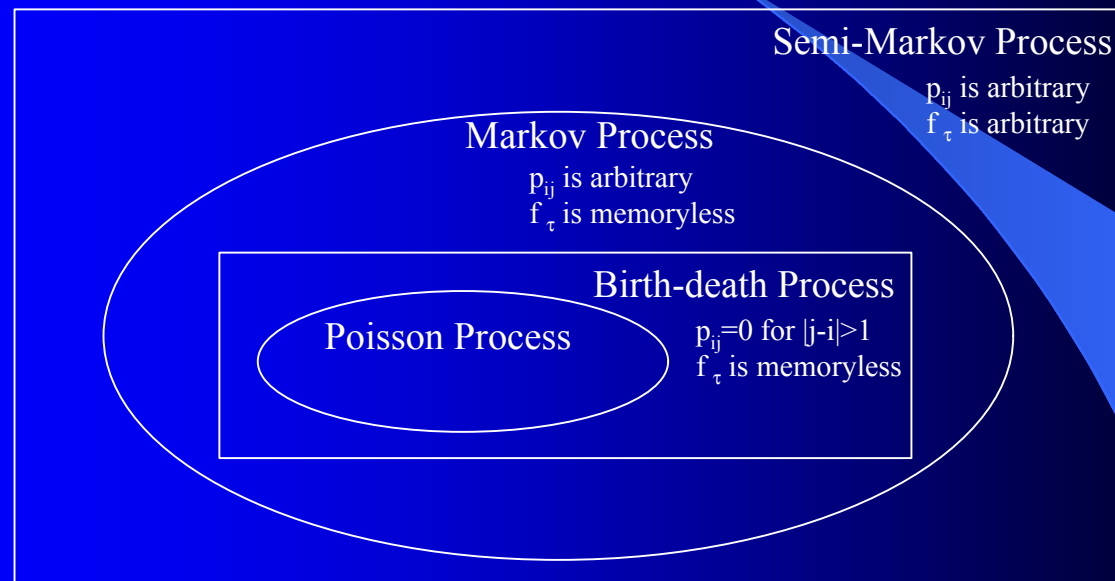
=> Geometric distribution



- How to solve it - “Global Balance Equation” : $\pi = \pi P$

- Continuous-time Markov chain: Exponential distribution

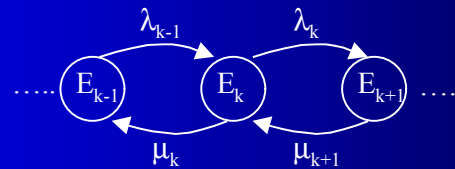
Types of Stochastic Processes



- * Memoryless means geometric or exponential distribution
- * τ is time between transitions
- * Semi-Markov : f_{τ} is not memoryless but at the moment of transitions, the process behaves just like a Markov chain

Birth-Death Process

- Continuous-time (homogeneous) Markov chain where the transitions are restricted to neighboring states only



- $P_k(t) = P[\text{system is in } E_k]$
- $$\frac{dP_k(t)}{dt} = -(\lambda_k + \mu_k) P_k(t) + \lambda_{k-1} P_{k-1}(t) + \mu_{k+1} P_{k+1}(t) \quad (k \geq 1)$$
- $$\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \quad (k=0)$$
- $$\Rightarrow P_k(t) = e^{-(\lambda+\mu)t} [\rho^{(k-1)/2} I_{k-1}(at) + \rho^{(k+1)/2} I_{k+1}(at) + (1-\rho) \rho^k \sum_{j=k+2}^{\infty} \rho^{-j/2} I_j(at)]$$

if $\lambda_k = \lambda$, $\mu_k = \mu$ for all k , and $\rho = \lambda/\mu$, $a = 2\mu\rho^{1/2}$, $I_k(x)$ = Bessel function
- \Rightarrow complete solution (including transient) is too complex
- We are only interested in steady-state probabilities (not functions of time)
- $$P_k = P_0 \times (\lambda_0 \lambda_1 \dots \lambda_{k-1}) / (\mu_1 \mu_2 \dots \mu_k)$$

Poisson Process

- Birth-death process with $\lambda_k = \lambda$, $\mu_k = 0$ for all k

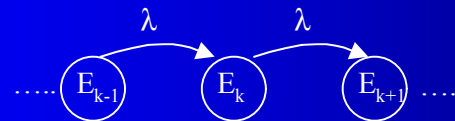
$$- \frac{dP_k(t)}{dt} = -\lambda P_k(t) + \lambda P_{k-1}(t) \quad (k \geq 1)$$

$$= -\lambda P_0(t) \quad (k=0)$$

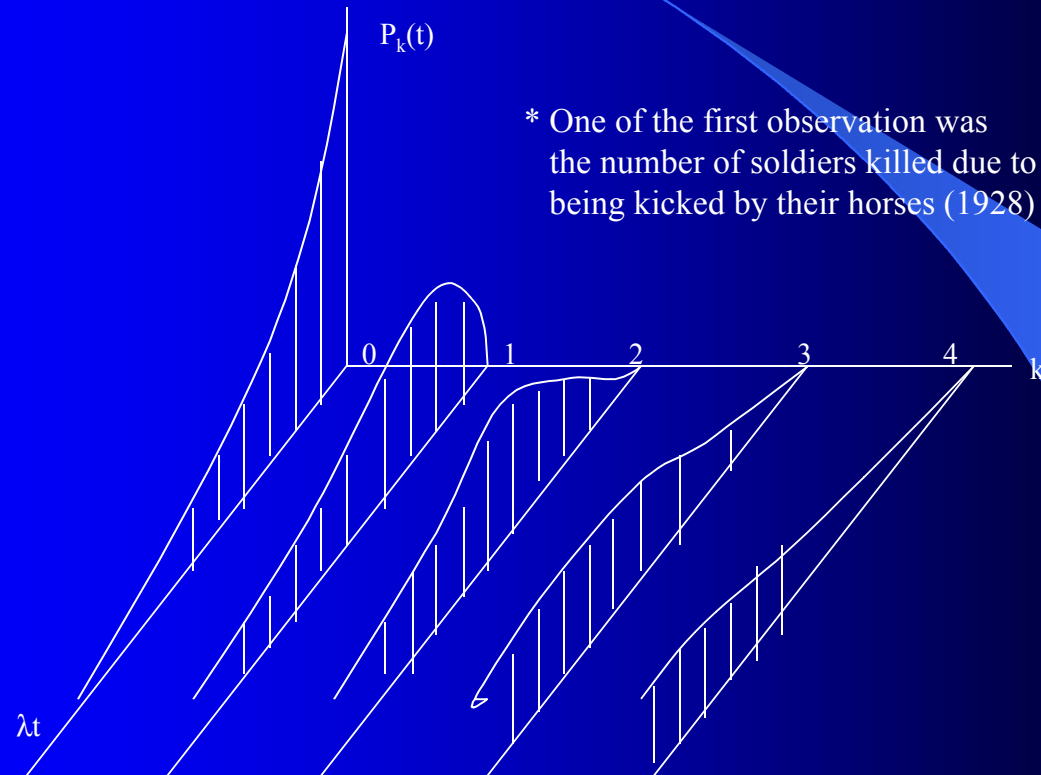
$$\Rightarrow P_k(t) = (\lambda t)^k e^{-\lambda t} / k! : \text{“Poisson” (transient solution !!!)}$$

(what will be the steady-state solution ?)

\Rightarrow Distribution of number of customers in the system is Poisson



Poisson Process (cont'd)



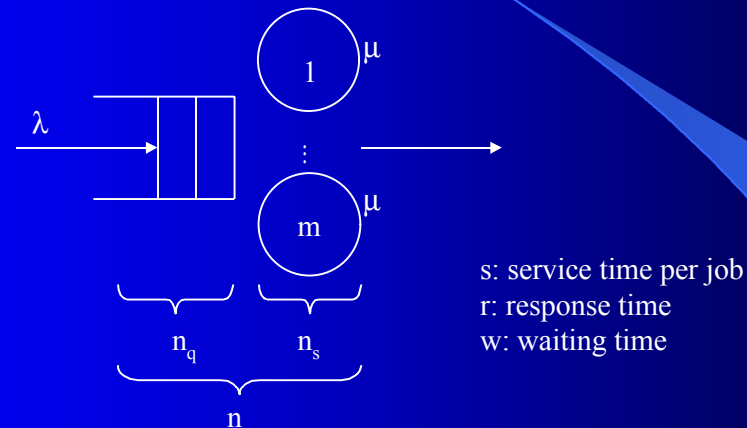
Poisson Process (cont'd)

- $A(t), a(t)$: cdf & pdf for the time between adjacent arrivals
 - $A(t) = 1 - \Pr\{\text{no arrivals during } t\} = 1 - P_0(t) = 1 - e^{-\lambda t}$
 - $a(t) = \lambda e^{-\lambda t}$: “exponential distribution”
- Constant Birth Rate
 - $\Pr\{\text{next arrival during } \Delta t\} = 1 - e^{-\lambda \Delta t} = \lambda \Delta t + o(\Delta t)$
 - $\Pr\{\text{no arrival during } \Delta t\} = 1 - \lambda \Delta t + o(\Delta t)$
 - $\Pr\{2 \text{ or more arrivals during } \Delta t\} = o(\Delta t)$
- Characteristics
 - merging k Poisson processes results in a Poisson stream ($\lambda = \sum \lambda_i$)
 - splitting a Poisson process results in Poisson streams ($\lambda_i = p_i \lambda$)
 - departure process of a system (μ) with Poisson arrival ($\lambda < \mu$) is Poisson (λ)

Queueing Model

- Component of Queueing Model
 - Arrival process
 - Exponential distribution (M for memoryless)
 - Erlang (E_k) and hyperexponential (H_k) distributions
 - Service time distribution
 - Number of servers
 - System capacity
 - Population size
 - Service discipline
- Notation: A/S/m/B/K/SD
 - Example: M/M/3/20/1500/FCFS
 - “M/M/3/ ∞ / ∞ /FCFS” is written as “M/M/3”
 - Bulk arrivals, bulk service : group of jobs - $M^{[x]}$

Rules for all queues



- Stability condition : $\lambda < m \mu$ (for infinite buffer and infinite population system)
- Number in system vs number in queue : $n = n_q + n_s$
 - $E(n) = E(n_q) + E(n_s)$, $\text{Var}(n) = \text{Var}(n_q) + \text{Var}(n_s)$, $\text{Cov}(n_q, n_s) = 0$
- Number vs Time : Little's Law (1961) if no creation/no lost of jobs
 - mean # in system = arrival rate x mean response time
- Time in system vs in queue : $r = w + s$
 - $E(n) = E(w) + E(s)$, $\text{Var}(r) = \text{Var}(w) + \text{Var}(s)$, $\text{Cov}(w, s) = 0$

Ch.31 Analysis of a Single Queue

- Interests in Queueing Systems
 - Traffic Intensity (λ/μ)
 - Server Utilization
 - Probability that N customers are in the system at time t
- Relationships (Little's Law)
 - $L = \lambda W$ (L: avg # in the system)
 - $L_q = \lambda W_q$ (L_q : avg # in queue)
 - $W = W_q + 1/\mu$ (W: avg waiting time in sys.)
(W_q : avg waiting time in queue)

Birth-Death Queueing System

- Homogeneous continuous-time Markov chain (discrete-state Markov)

- $P_k(t) = P[\text{system is in } E_k]$

- $\frac{dP_k(t)}{dt} = -(\lambda_k + \mu_k) P_k(t) + \lambda_{k-1} P_{k-1}(t) + \mu_{k+1} P_{k+1}(t) \quad (k \geq 1)$

- $\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t) \quad (k=0)$

$\Rightarrow P_k(t) = e^{-(\lambda+\mu)t} [\rho^{(k-i)/2} I_{k-i}(at) + \rho^{(k-i-1)/2} I_{k+i+1}(at) + (1-\rho)\rho^k \sum_{j=k+i+2}^{\infty} \rho^{-j/2} I_j(at)]$
if $\lambda_k = \lambda$, $\mu_k = \mu$ for all k , and $\rho = \lambda/\mu$, $a = 2\mu\rho^{1/2}$, $I_k(x)$ = Bessel function

Birth-Death Queueing System (cont'd)

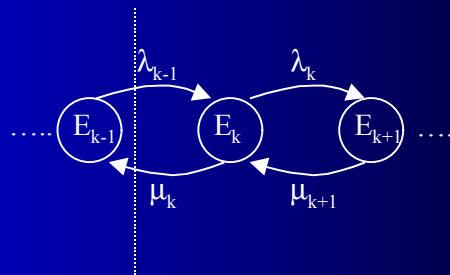
- In equilibrium, “rate in = rate out principle” : Local Balance Equation

- $p_k \triangleq \lim_{t \rightarrow \infty} P_k(t)$: steady-state probability

- conservation of flow : $\lambda_{k-1} p_{k-1} = \mu_k p_k$

$$\Rightarrow p_k = \lambda_{k-1} p_{k-1} / \mu_k$$

$$\Rightarrow p_k = p_0 (\lambda_0 \lambda_1 \dots \lambda_{k-1}) / (\mu_1 \mu_2 \dots \mu_k)$$

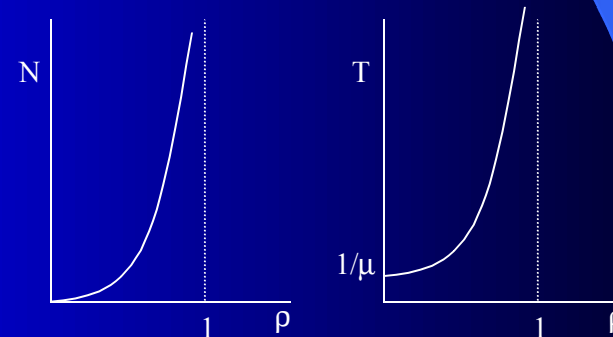
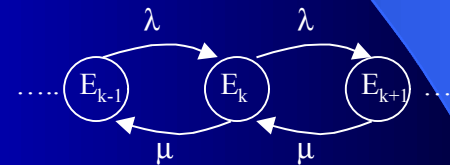


M/M/1: Classical Queueing System

- Birth-death process with $\lambda_k = \lambda$, $\mu_k = \mu$ for all k
 - $p_k = p_0 (\lambda_0 \lambda_1 \dots \lambda_{k-1}) / (\mu_1 \mu_2 \dots \mu_k) = p_0 (\lambda/\mu)^k$
 - $p_0 = 1 / [1 + \sum_{k=1}^{\infty} (\lambda/\mu)^k] = 1 / [1 + (\lambda/\mu) / (1 - \lambda/\mu)] = 1 - \lambda/\mu = 1 - \rho$
 - ⇒ $p_k = (1 - \rho) \rho^k$ where $\rho = \lambda/\mu$

- average number of customers
 $N = \sum_{k=1}^{\infty} k p_k = \rho / (1 - \rho)$

- average time spent in the system
 $T = N / \lambda$ (Little's Law)
 $= \rho / (1 - \rho) \lambda$
 $= 1 / (1 - \rho) \mu$

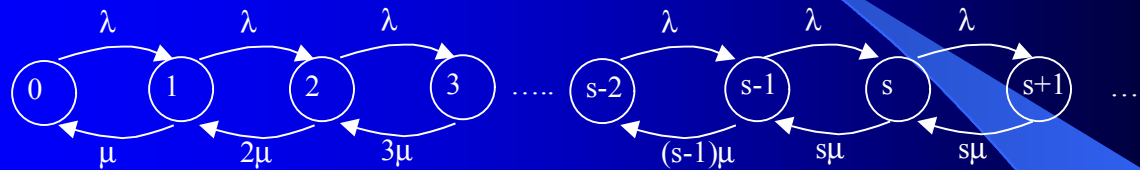


M/M/1: Example

- Engineers use one terminal for serious computation. Arrival pattern is Poisson with a mean of 10 people each day. The distribution of time spent at a terminal is exponential with a mean of 30 minutes. Engineers complain about the terminal service but the manager finds the terminal is in use only 5 hours out of 8-hour working day.
 - $\lambda = 10 \text{ persons/day} \times 1 \text{ day/8hours} \times 1 \text{ hour/60minutes} = 1/48 \text{ persons/min}$
 - $\mu = 1/30 \text{ persons/min}$
 - $\rho = \lambda / \mu = 30/48 = 0.625$: server utilization
 - $L = \rho / (1 - \rho) = 1.667 \text{ persons}$: average number of customers in system
 - $L_q = 1.667 - 0.625 = 1.042$: average number of customers in queue
 - $T = L / \lambda = 80 \text{ minutes}$: average time spent in the system
 - $T_q = 80 - 30 = 50 \text{ minutes}$: average time wasted in queue

* $T_q = L_q / \lambda \leq$ Little's Law can be applied in the subsystem

M/M/s : s-Server Queue



- $\lambda_n = \lambda$ for all n , $\mu_n = n\mu$ for $n=0,1,2,\dots,s$
 $= s\mu$ for $n=s,s+1,\dots$
- State Rate In = Rate Out (Flow conservation law)

0	$\mu P_1 = \lambda P_0$
1	$2\mu P_2 + \lambda P_0 = (\lambda + \mu) P_1$
2	$3\mu P_3 + \lambda P_1 = (\lambda + 2\mu) P_2$
....
s-1	$s\mu P_s + \lambda P_{s-2} = \{\lambda + (s-1)\mu\} P_{s-1}$
s	$s\mu P_{s+1} + \lambda P_{s-1} = (\lambda + s\mu) P_s$
s+1	$s\mu P_{s+2} + \lambda P_s = (\lambda + s\mu) P_{s+1}$
....

M/M/s : s-Server Queue (cont'd)

- Stability condition: $\lambda < s\mu$

$$\begin{aligned}
 - P_0 &= 1 / \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s}^{\infty} \frac{(\lambda/s\mu)^{n-s}}{1 - (\lambda/s\mu)} \right] \\
 &= 1 / \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1 - (\lambda/s\mu)} \right]
 \end{aligned}$$

$$\begin{aligned}
 - P_n &= \frac{(\lambda/\mu)^n}{n!} P_0, & \text{if } 0 \leq n \leq s \\
 &= \frac{(\lambda/\mu)^n}{s! s^{n-s}} P_0, & \text{if } s \leq n
 \end{aligned}$$

- $P[L(\infty) \geq s] = \text{Prob that an arriving customer is forced to join the queue/wait}$
 $= \{(\lambda/\mu)^s P_0\} / \{s! (1 - \lambda/s\mu)\}$
- “M/M/∞” is a special case of M/M/s, where jobs do not wait
 - response time = service time : called “delay center”

M/M/s: Example

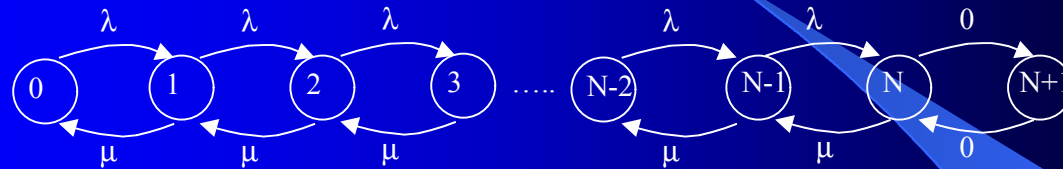
- Two car wash machines with exponentially distributed service time (mean 40 seconds) and cars arrive at rate 2 cars/minute with Poisson distribution.
 - $\lambda = 2$ cars/minute, $\mu = 60/40 = 1.5$ cars/minute, $s = 2$
 - $$P_0 = \left\{ \left[\sum_{n=0}^{\infty} \frac{(4/3)^n}{n!} \left(\frac{4}{3} \right)^2 \frac{1}{2!} \frac{2(3/2)}{2(3/2) - 2} \right] \right\}^{-1}$$

$$= \{1 + 4/3 + (16/9)(1/2)(3)\}^{-1}$$

$$= \{15 / 3\}^{-1} = 1/5 = 0.2$$
 - The probability that all servers are busy is given by (new car must wait)

$$P(L(\infty) \geq 2) = \{(4/3)^2 (1/5)\} / \{2!(1 - 2/3)\} = (8/3) (1/5) = 0.533$$
 - $L_q = \{(2/3)(8/15)\} / (1 - 2/3) = 1.07$ cars
 $L = L_q + \lambda/\mu = 16/15 + 4/3 = 12/5 = 2.4$ cars
 $W = L / \lambda = 2.4 / 2 = 1.2$ minutes
 $W_q = W - 1/\mu = 1.2 - 2/3 = 0.533$ minute

M/M/1/N: Finite Storage



- $\sum_{n=0}^N P_n = 1$
 $\Rightarrow P_0 + (\lambda/\mu)^1 P_0 + \dots + (\lambda/\mu)^N P_0 = 1$
 $\Rightarrow P_0 \{1 + (\lambda/\mu)^1 + \dots + (\lambda/\mu)^N\} = 1$
 $\Rightarrow P_0 = 1 / \{\sum_{n=0}^N (\lambda/\mu)^n\}$
 $= 1 / \left\{ \frac{1 - (\lambda/\mu)^{N+1}}{1 - (\lambda/\mu)} \right\}$
 $= (1 - \rho) / (1 - \rho^{N+1})$
- $P_n = \left\{ \frac{1 - \rho}{1 - \rho^{N+1}} \right\} \rho^n$, for $n = 0, 1, 2, \dots, N$
- Effective arrival rate $\lambda_e = \lambda (1 - P_N) \Rightarrow W = L/\lambda_e, W_q = L_q/\lambda_e$

M/M/1/N: Example

- One car wash machine but the parking lot allows two waiting cars

- assume that $\lambda = 2$ cars/minute, $\mu = 3$ cars/minute

- traffic intensity $\rho = \lambda / \mu = 2 / 3$

- $P_N = P_3 = \{(1-2/3) (2/3)^3\} / \{1 - (2/3)^4\} = 8 / 65 = 0.123$

- Average # of cars

$$L = \frac{2/3 \{1 - 4(2/3)^3 + 3(2/3)^4\}}{\{1 - (2/3)^4\} (1 - 2/3)} = \frac{66}{65} = 1.015 \text{ cars}$$

- Effective arrival rate

$$\lambda_e = \lambda (1 - P_n) = 2(1 - 8/65) = 2 \cdot 57 / 65 = 114/65 = 1.754 \text{ (cars/minute)}$$

- W is given by

$$W = L / \lambda_e = 1.015 / 1.754 = 0.579 \text{ (minutes)}$$

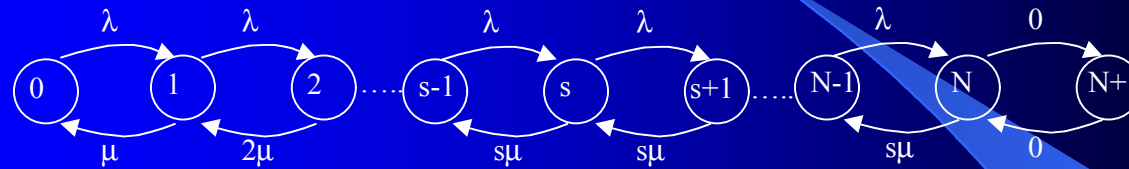
- $P_0 = (1 - \rho) / (1 - \rho^{N+1}) = 0.415$

$$\Rightarrow \text{effective utilization } \rho_e = 1 - P_0 = \lambda_e / \mu = 0.585$$

- $L_q = L - (1 - P_0) = 0.43$ cars

$$\Rightarrow W_q = L_q / \lambda_e = 0.43 / 1.754 = 0.245 \text{ minutes}$$

M/M/s/N



Other models

– M/M/1/∞/M

- finite customer population (M) with single server
- $\lambda_n = \lambda(M-n)$ for $0 \leq n \leq M$, 0 otherwise and $\mu_n = \mu$

– M/M/∞/∞/M

- finite population with a separate server for each customer
- $\lambda_n = \lambda(M-n)$ for $0 \leq n \leq M$, 0 otherwise and $\mu_n = n\mu$

– M/M/s/N/M

- finite population, s -server, finite storage
- $\lambda_n = \lambda(M-n)$ for $0 \leq n \leq M$, 0 otherwise and $\mu_n = n\mu$ for $0 \leq n \leq s$, $s\mu$ otherwise