Queueing Theory

Frank Y. S. Lin
Information Management Dept.
National Taiwan University
yslin@im.ntu.edu.tw

References

- Leonard Kleinrock, "*Queueing Systems Volume I: Theory*", New York: Wiley, 1975-1976.
- D. Gross and C. M. Harris, "*Fundamentals of Queueing Theory*", New York: Wiley, 1998.

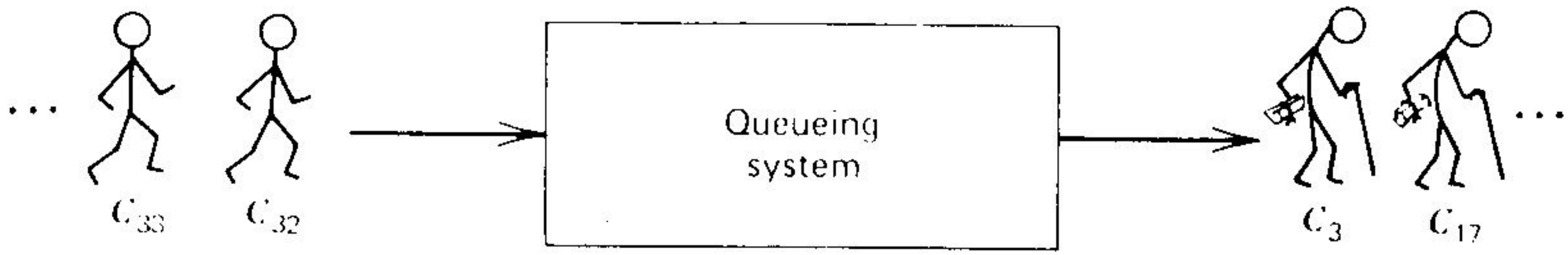
Agenda

- Introduction
- Stochastic Process
- General Concepts
- $M/M/1$ Model
- $M/M/1/K$ Model
- Discouraged Arrivals
- $M/M/\infty$ and $M/M/m$ Models
- $M/M/m/m$ Model

Introduction

Queueing System

- A queueing system can be described as customers ***arriving*** for service, ***waiting*** for service if it is not immediate, and if having waited for ***service***, leaving the system after being served.



Why Queueing Theory

- Performance Measurement
  - Average **waiting time** of customer / distribution of waiting time.
  - Average **number of customers** in the system / distribution of queue length / current work backlog.
  - Measurement of the **idle time** of server / length of an idle period.
  - Measurement of the **busy time** of server / length of a busy period.
  - System **utilization**.

Why Queueing Theory (cont'd)

- Delay Analysis

  Network Delay =

  Queueing Delay

  \+ Propagation Delay (depends on the distance)

  \+ Node Delay
  - Processing Delay (independent of packet length, e.g. header CRC check)
  - Adapter Delay (constant)

Characteristics of Queueing Process

- Arrival Pattern of Customers
  - Probability distribution
  - Patient / impatient (balked) arrival
  - Stationary / nonstationary
- Service Patterns
  - Probability distribution
  - State dependent / independent service
  - Stationary / nonstationary

Characteristics of Queueing Process (cont'd)

- Queueing Disciplines
  - First come, first served (FCFS)
  - Last come, first served (LCFS)
  - Random selection for service (RSS)
  - Priority queue
  - Preemptive / nonpreemptive
- System Capacity
  - Finite / infinite waiting room.

Characteristics of Queueing Process (cont'd)

- Number of Service Channels
  - Single channel / multiple channels
  - Single queue / multiple queues
- Stages of Service
  - Single stage (e.g. hair-styling salon)
  - Multiple stages (e.g. manufacturing process)
  - Process recycling or feedback

Notation

- A queueing process is described by ***A/B/X/Y/Z***

Characteristic	Symbol	Explanation
Interarrival-time distribution (A) Service-time distribution (B)	M	Exponential
	D	Deterministic
	E_k	Erlang type k ($k = 1, 2, \ldots$)
	H_k	Mixture of k exponentials
	PH	Phase type
	G	General
Number of parallel servers (X)	$1, 2, \ldots, \infty$	
Restriction on system capacity (Y)	$1, 2, \ldots, \infty$	
Queue discipline (Z)	FCFS	First come, first served
	LCFS	Last come, first served
	RSS	Random selection for service
	PR	Priority
	GD	General discipline

Notation (cont'd)

- For example, $M/D/2/\infty$/FCFS indicates a queueing process with exponential inter-arrival time, deterministic service times, two parallel servers, infinite capacity, and first-come, first-served queueing discipline.
- Y and Z can be omitted if $Y = \infty$ and $Z =$ FCFS.

Stochastic Process

Stochastic Process

- Stochastic process: any collection of random variables $X(t)$, $t \in T$, on a common probability space where t is a subset of time.
  - Continuous / discrete time stochastic process
  - Example: $X(t)$ denotes the temperature in the class on t = 7:00, 8:00, 9:00, 10:00, … (discrete time)
- We can regard a stochastic process as a family of random variables which are “indexed” by time.
- For a random process $X(t)$, the PDF is denoted by $F_X(x;t) = P[X(t) <= x]$

Some Classifications of Stochastic Process

- **Stationary Processes**: independent of time

  $$F_X(x; t+\tau) = F_X(x; t)$$

- **Independent Processes**: independent variables

  $$F_X(x; t) = F_{X1,\ldots,Xn}(x_1,\ldots, x_n; t_1,\ldots,t_n)$$
  $$= F_{X1}(x_1; t_1) \ldots F_{Xn}(x_n; t_n)$$

- **Markov Processes**: the probability of the next state depends only upon the current state and not upon any previous states.

  $$P[X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n, \ldots, X(t_1) = x_1]$$
  $$= P[X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n]$$

Some Classifications of Stochastic Process (cont'd)

- **Birth-death Processes**: state transitions take place between neighboring states only.
- **Random Walks**: the next position the process occupies is equal to the previous position plus a random variable whose value is drawn independently from an arbitrary distribution.

General Concepts

Continuous-time Memoryless Property

If X ~Exp(λ), for any $a,b > 0$,

$P[X > a + b \mid X > a] = P[X > b]$

Proof:

$$P[X > a + b \mid X > a]$$

$$= \frac{P[(X > a+b) \cap (X > a)]}{P(X > a)} \qquad (X > a+b) \subset (X > a)$$

$$= \frac{P(X > a+b)}{P(X > a)} = \frac{1 - F_x(a+b)}{1 - F_x(a)} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = e^{-\lambda b} = P(X > b)$$

Global Balance Equation

- Define P_i = P[system is in state i]
  P_{ij} = P[get into state j right after leaving state i]

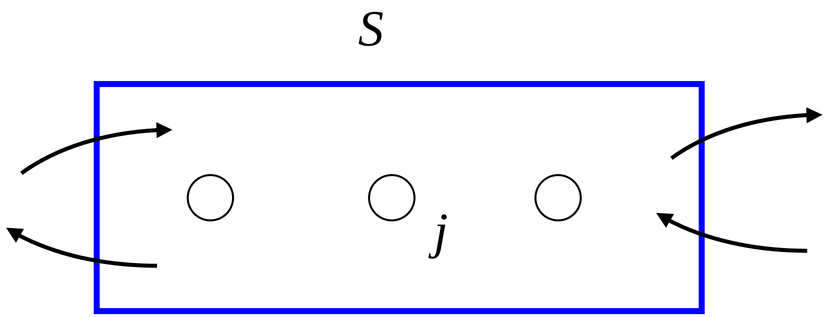
$$\underbrace{P_j \cdot \sum_{\substack{i=0 \\ (i \neq j)}}^{\infty} P_{ij}}_{\text{rate out of state } j} = \underbrace{\sum_{\substack{i=0 \\ (i \neq j)}}^{\infty} P_i \cdot P_{ij}}_{\text{rate into state } j}$$

rate out of state j = rate into state j

General Balance Equation

- Define S = a subset of the state space

$$\sum_{\substack{j=0 \\ (j \in s)}}^{\infty} P_j \cdot \sum_{\substack{i=0 \\ (i \notin s)}}^{\infty} P_{ij} = \sum_{\substack{i=0 \\ (i \notin s)}}^{\infty} P_i \cdot \sum_{\substack{j=0 \\ (j \in s)}}^{\infty} P_{ij}$$



rate in = rate out

General Equilibrium Solution

- Notation:
  - P_k = the probability that the system contains k customers (in state k)

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

  - λ_k= the arrival rate of customers when the system is in state k.
  - μ_k= the service rate when the system is in state k.

General Equilibrium Solution (cont'd)

- Consider state $\leq k$:

$$rate\ in = rate\ out$$

$$P_k \cdot \lambda_k = P_{k+1} \cdot \mu_{k+1}$$

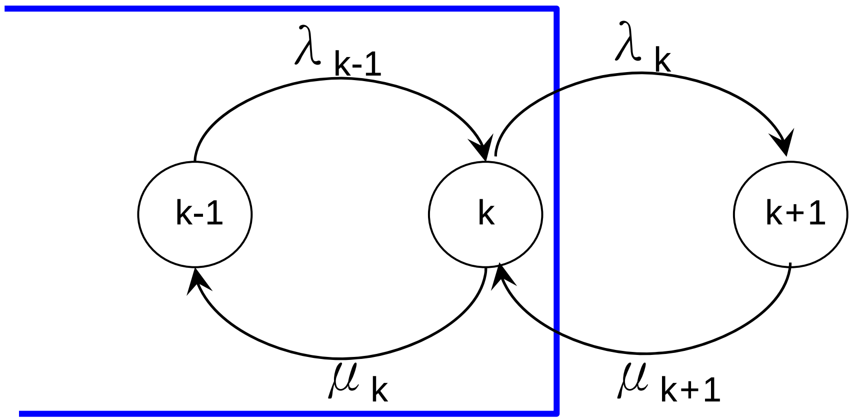
$$\Rightarrow P_{k+1} = \frac{\lambda_k}{\mu_{k+1}} P_k$$

$$\Rightarrow P_k = \frac{\lambda_{k-1}}{\mu_k} P_{k-1}$$

$$\vdots$$

$$\Rightarrow P_k = \frac{\lambda_{k-1} \cdot \lambda_{k-2} \cdots \lambda_0}{\mu_k \cdot \mu_{k-1} \cdots \mu_1} P_0 = \boxed{\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \cdot P_0}$$

#



General Equilibrium Solution (cont'd)

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

$$\Rightarrow \sum_{k=0}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \cdot P_0 = 1$$

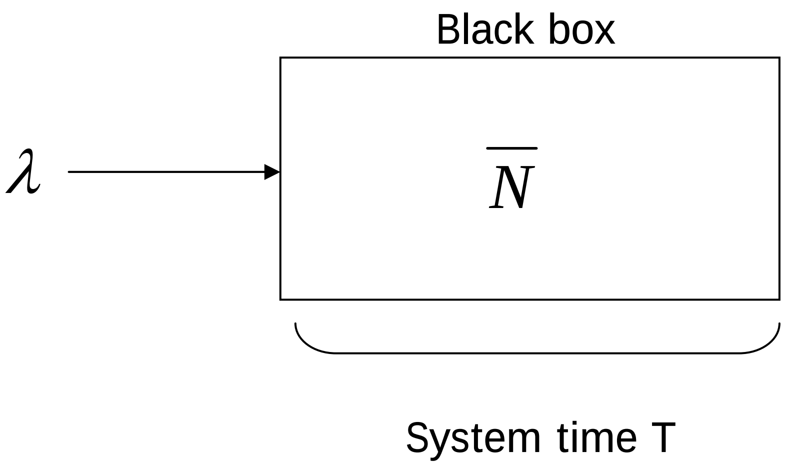
$$\therefore P_0 = \frac{1}{1 + \sum_{k=0}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} \qquad \lambda = \sum_{k=0}^{\infty} P_k \lambda_k, \quad T = \frac{\overline{N}}{\lambda}$$

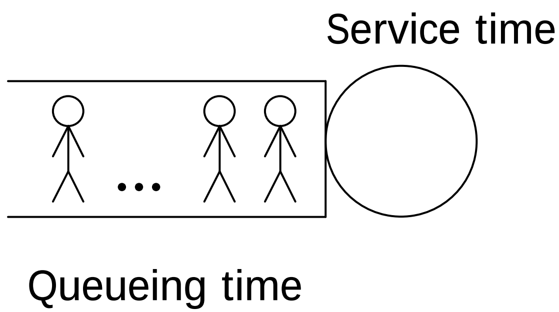
waiting time $w = T - \frac{1}{\mu}$

Little's Result

- $\overline{N}$ = average number of customers in the system
- T = system time (service time + queueing time)
- λ = arrival rate

➔ $\overline{N} = \lambda T$





M/*M*/1 Model

Single Server, Single Queue

(The Classical Queueing System)

$M/M/1$ Queue

- Single server, single queue, infinite population:

$$\begin{cases} \lambda_k = \lambda \\ \mu_k = \mu \end{cases}$$

- Interarrival time distribution:

$$p_\lambda(t) = \lambda e^{-\lambda t}$$

- Service time distribution

$$p_\mu(t < t_0) = \int_0^{t_0} \mu e^{-\mu t} dt = 1 - e^{-\mu t_0}$$

- Stability condition $\lambda < \mu$

$M/M/1$ Queue (cont'd)

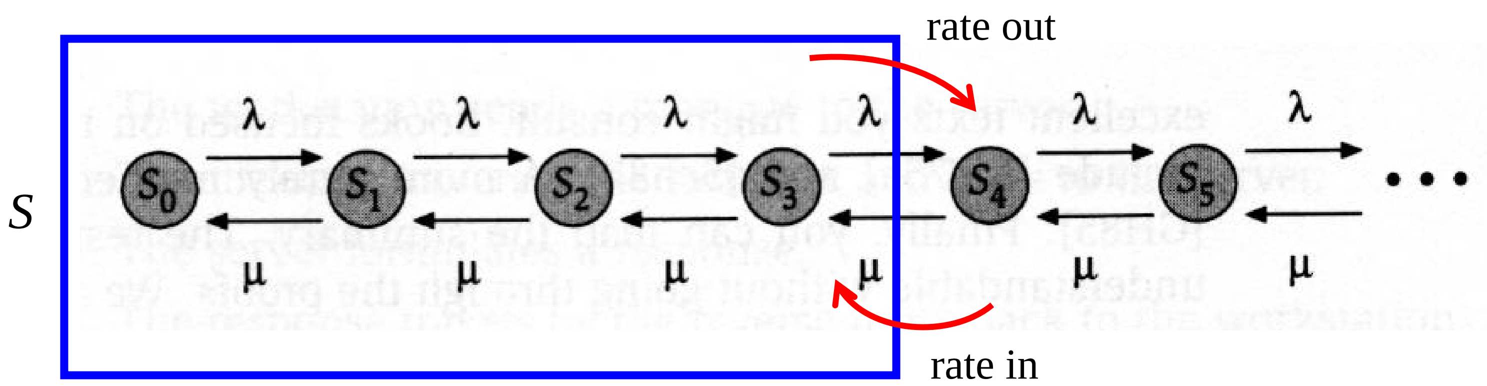
- System utilization

$$\rho = \frac{\lambda}{\mu} = \text{P[system is busy]}, \quad 1\text{-}\rho = \text{P[system is idel]}$$

- Define state S_n = n customers in the system

  (n-1 in the queue and 1 in service)

  S_0 = empty system



$M/M/1$ Queue (cont'd)

- Define p_n = P[n customers in the system]

$$\lambda \times p_n = \mu \times p_{n+1} \quad \text{(rate in = rate out)}$$

$$p_{n+1} = \frac{\lambda}{\mu} \times p_n = \rho \times p_n$$

$$\Rightarrow \quad p_{n+1} = \rho^{n+1} \times p_0$$

Since $\sum_{i=0}^{\infty} p_i = 1 \Rightarrow \sum_{i=0}^{\infty} p_0 \rho^n = 1 \Rightarrow p_0 \sum_{i=0}^{\infty} \rho^n = 1$

$$\Rightarrow \boxed{p_0 = 1 - \rho, \quad p_n = \rho^n \times (1 - \rho)}$$

#

$M/M/1$ Queue (cont'd)

- Average number of customers in the system

$$\overline{N} = \sum k \cdot (1-\rho)\rho^k \;\; = (1-\rho)\sum k \cdot \rho^k$$
$$= (1-\rho)\cdot\rho\cdot\sum d\rho^k / d\rho$$
$$= (1-\rho)\cdot\rho\cdot\frac{d}{d\rho}\sum \rho^k$$
$$= (1-\rho)\cdot\rho\cdot\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$
$$= \frac{\rho}{1-\rho}$$

$$\boxed{\therefore \overline{N} = \frac{\rho}{1-\rho}}$$

#

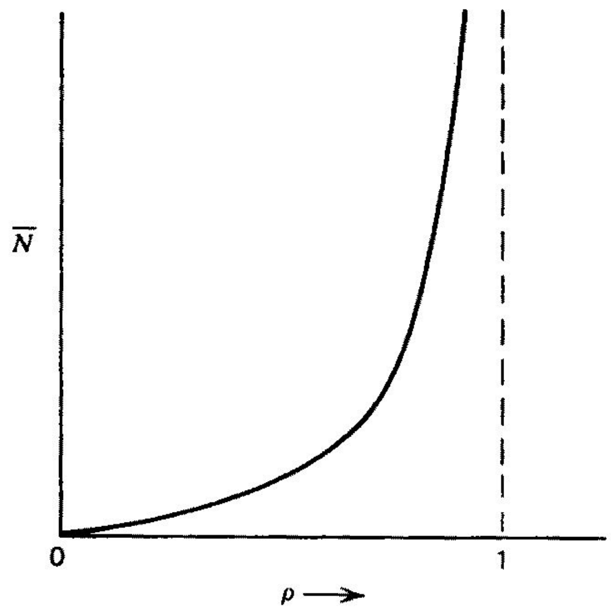


Figure 3.3 The average number in the system M/M/1.

$M/M/1$ Queue (cont'd)

- Average system time

$$T = \frac{\overline{N}}{\lambda} \quad \text{(Little's Result)}$$

$$= \frac{\frac{\rho}{1-\rho}}{\lambda} = \frac{1/\mu}{1-\rho} = \frac{1}{\mu - \lambda}$$ #

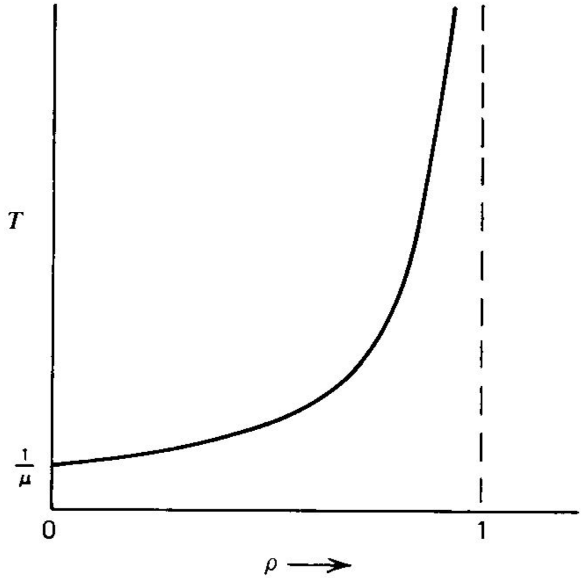


Figure 3.4 Average delay as a function of ρ for M/M/1.

- P[$\geq k$ customers in the system]

$$= \sum_{i=k}^{\infty} (1-\rho)\rho^i = (1-\rho)\frac{\rho^k}{1-\rho} = \rho^k$$

$M/M/1/K$ Model

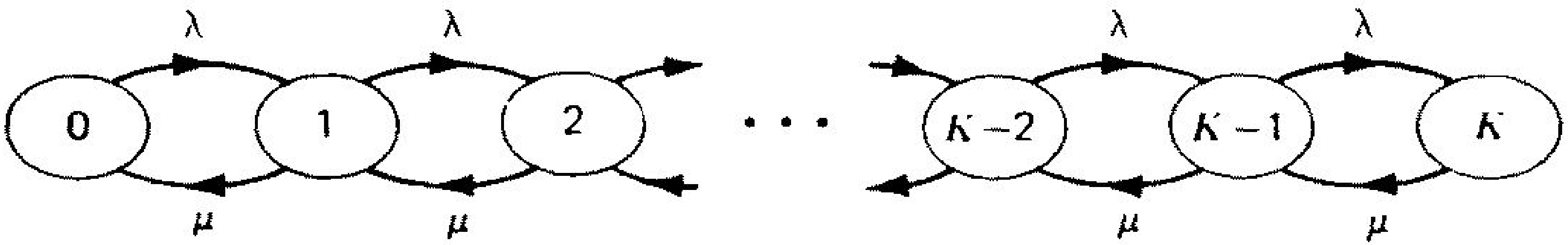
Single Server, Finite Storage

$M/M/1/K$ Model

- The system can hold at most a total of K customers (including the customer in service)

$$\lambda_k = \begin{cases} \lambda & \text{if } k < K \\ 0 & \text{if } k \geq K \end{cases}$$

$$\mu_k = \mu$$



M/*M*/1/*K* Model (cont'd)

$$\begin{cases} P_k = P_0 \prod_{i=0}^{k-1} \dfrac{\lambda}{\mu} = P_0 \left(\dfrac{\lambda}{\mu}\right)^k & k \le K \\ P_k = 0 & k > K \end{cases}$$

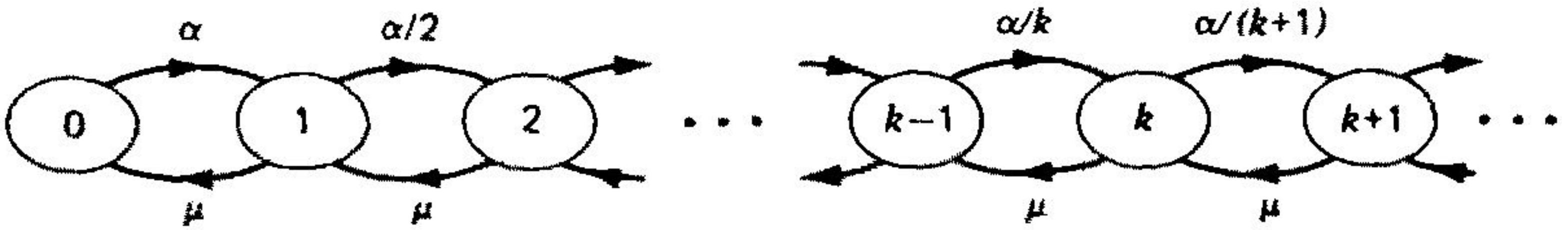
$$\Rightarrow P_0 = \begin{cases} \left[1 + \sum_{k=1}^{K} (\lambda/\mu)^k\right]^{-1} = \dfrac{1-\lambda/\mu}{1-(\lambda/\mu)^{K+1}} & 0 \le k \le K \\ 0 & \text{otherwise} \end{cases}$$

Discouraged Arrivals

Discouraged Arrivals

- Arrivals tend to get discouraged when more and more people are present in the system.

$$\begin{cases} \lambda_k = \dfrac{\alpha}{k+1} \\ \mu_k = \mu \end{cases}$$



Discouraged Arrivals (cont'd)

$$P_k = P_0 \cdot \prod_{i=0}^{k-1} \frac{\alpha/(i+1)}{\mu} = (\alpha/\mu)^k \cdot \frac{1}{k!} \cdot P_0$$

$$P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} (\alpha/\mu)^k \cdot \frac{1}{k!}} = e^{-\frac{\alpha}{\mu}}$$

$$\Rightarrow P_k = \frac{(\alpha/\mu)^k}{k!} \cdot e^{-\frac{\alpha}{\mu}} \qquad \therefore \overline{N} = \frac{\alpha}{\mu}$$

Discouraged Arrivals (cont'd)

$$\bar{\lambda} = \sum_{k=0}^{\infty} \lambda_k P_k = \sum_{k=0}^{\infty} \frac{\alpha}{k+1} \cdot \frac{(\alpha/\mu)}{k!} \cdot e^{-(\alpha/\mu)}$$

$$= \mu\left[1 - e^{-(\alpha/\mu)}\right] \quad (\because \lambda = \mu\rho, \rho = 1 - P_0)$$

$$T = \frac{\bar{N}}{\lambda} = \frac{\alpha/\mu}{\mu(1 - e^{-\alpha/\mu})}$$

$M/M/\infty$ and $M/M/m$

$M/M/\infty$ - Infinite Servers, Single Queue

(Responsive Servers)

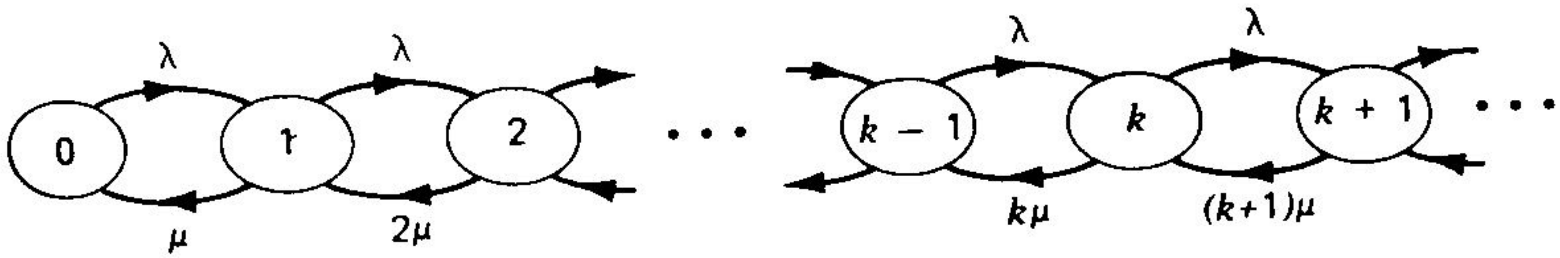
$M/M/m$ - Multiple Servers, Single Queue

(The m-Server Case)

$M/M/\infty$ Queue

- There is always a new server available for each arriving customer.

$$\begin{cases} \lambda_k = \lambda \\ \mu_k = k\mu \end{cases}$$



$M/M/\infty$ Queue (cont'd)

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = \frac{(\lambda/\mu)^k}{k!} e^{-\lambda/\mu}$$

$$\Rightarrow \overline{N} = \frac{\lambda}{\mu}$$

$$\Rightarrow T = \frac{1}{\mu}$$ (Little's Result)

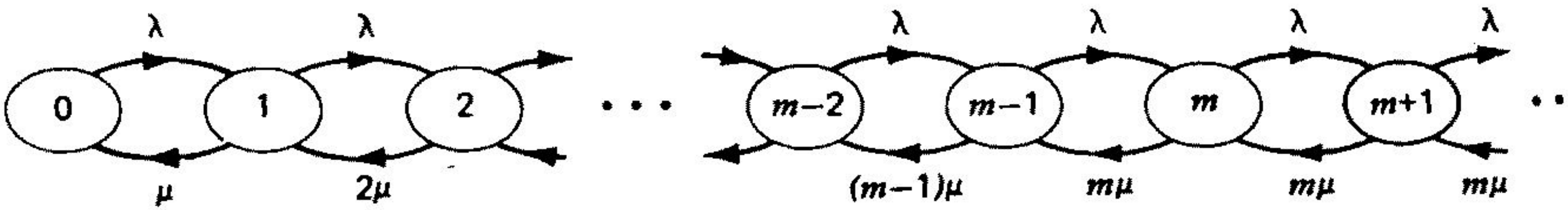
M/*M*/*m* Queue

- The *M*/*M*/*m* queue
  - An *M*/*M*/*m* queue is shorthand for a single queue served by multiple servers.
  - Suppose there are *m* servers waiting for a single line. For each server, the waiting time for a queue is a system with service rate μ and arrival rate λ/m.
  - The *M*/*M*/1 analysis has been done, at risk conclusion:

delay = $\dfrac{1}{\mu - \lambda / n}$

throughput $\rho = \dfrac{\lambda / n}{\mu} = \dfrac{\lambda}{n\mu}$

M/M/m Queue (cont'd)



$$\lambda_k = \lambda$$

$$\mu_k = \begin{cases} k\mu & \text{if } k \le m \\ m\mu & \text{if } k > m \end{cases}$$

For $k \le m$ $\quad p_k = p_0 \dfrac{\lambda}{\mu}\dfrac{\lambda}{2\mu}\cdots\dfrac{\lambda}{k\mu} = p_0 (\dfrac{\lambda}{\mu})^k \dfrac{1}{k!}$

For $k > m$ $\quad p_k = p_0 \dfrac{\lambda}{\mu}\dfrac{\lambda}{2\mu}\cdots\dfrac{\lambda}{n\mu}\cdots\dfrac{\lambda}{n\mu} = p_0 (\dfrac{\lambda}{\mu})^k \dfrac{1}{n!}(\dfrac{1}{n})^{k-n}$

M/M/n Queue (cont'd)

$$\sum_{i=0}^{\infty} p_i = 1$$

$$\therefore \quad p_0 = \frac{1}{\sum_{k=0}^{n-1} p_i \frac{(np)^k}{k!} + \frac{(np)^n}{n!} \frac{1}{(1-\rho)}} \qquad where\ \rho = \frac{\lambda}{n\mu}$$

P[queueing] = $\sum_{k=m}^{\infty} p_k$

Total system time = $\frac{1}{\mu} + \frac{\lambda(/\mu)^n \mu}{(n-1)!(n\mu - \lambda)^2} \times p_0$

Comparisons (cont'd)

- $M/M/1$ v.s $M/M/4$

  If we have 4 $M/M/1$ systems: 4 parallel communication links that can each handle 50 pps (μ), arrival rate λ = 25 pps per queue.

  ➔average delay = 40 ms.

  Whereas for an $M/M/4$ system,

  ➔average delay = 21.7 ms.

Comparisons (cont'd)

- Fast Server v.s A Set of Slow Servers #1

If we have an $M/M/4$ system with service rate μ =50 pps for each server, and another $M/M/1$ system with service rate 4μ = 200 pps. Both arrival rate is λ = 100 pps

➔delay for $M/M/4$ = 21.7 ms

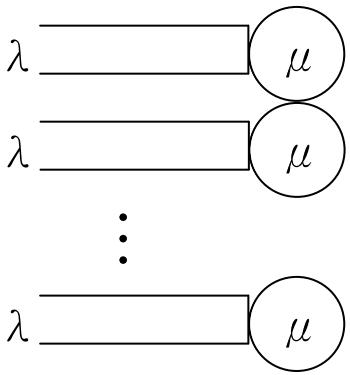
➔delay for $M/M/1$ = 10 ms

Comparisons (cont'd)

- Fast Server v.s A Set of Slow Servers #2

If we have n $M/M/1$ system with service rate μ pps for each server, and another $M/M/1$ system with service rate $n\mu$ pps. Both arrival rate is $n\lambda$ pps

S_1

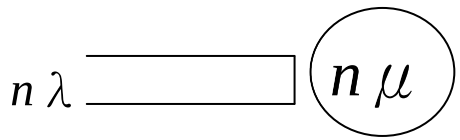


$$T_1 = \frac{1/\mu}{1-\rho}$$

$$T_2 = \frac{1/n\mu}{1-\dfrac{n\lambda}{n\mu}} = \frac{1/n\mu}{1-\rho}$$

$$\therefore \quad T_2 = \frac{T_1}{n}$$

S_2



$M/M/m/m$

Multiple Servers, No Storage

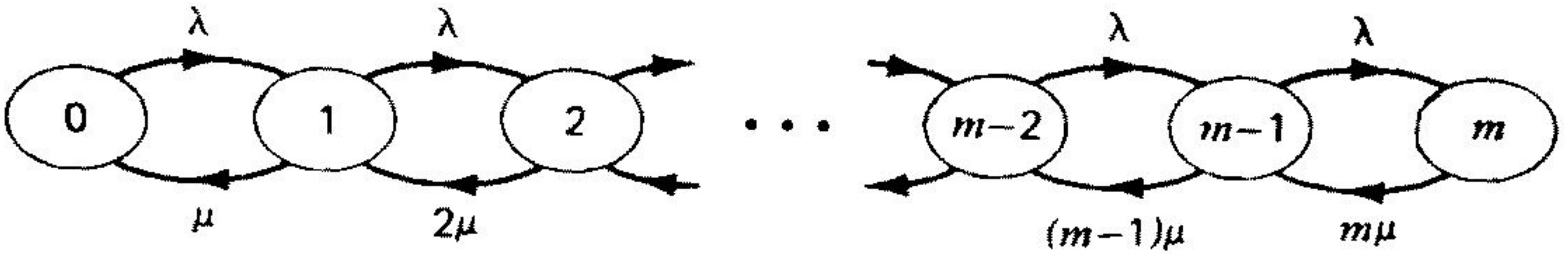
(m-Server Loss Systems)

$M/M/m/m$

- There are available m servers, each newly arriving customers is given a server, if a customers arrives when all servers are occupied, that customer is lost

  e.g. telephony system.

$$\begin{cases} \lambda_k = \begin{cases} \lambda & if\ k < m \\ 0 & if\ k < m \end{cases} \\ \mu_k = k\mu \end{cases}$$



M/M/m/m (cont'd)

$$P_k = \begin{cases} P_0 \cdot (\lambda/\mu)^k \dfrac{1}{k!} & \text{if k} \le \text{m} \\ 0 & \text{if k} > \text{m} \end{cases}$$

$$\Rightarrow P_0 = \left[\sum_{k=0}^{\infty} (\lambda/\mu)^k \frac{1}{k!} \right]^{-1}$$

M/M/m/m (cont'd)

- Let p_m describes the fraction of time that all m servers are busy. The name given to this probability expression is *Erlang's loss formula* and is given by

$$p_m = \frac{(\lambda/\mu)^m / m!}{\sum_{k=0}^{m} (\lambda/\mu)^k / k!}$$

- This equation is also referred to as *Erlang's B formula* and is commonly denoted by $B(m, \lambda/\mu)$
- http://www.erlang.com