Queueing Theory

Introduction

- In this chapter we will study a class of models in which customers arrive in some random manner at a service facility.

- Upon arrival they are made to wait in queue until it is their turn to be served. Once served they are generally assumed to leave the system.

- Interested parameters
  - the average number of customers in the system (or in the queue)
  - the average time a customer spends in the system (or spends waiting in the queue), etc.

What is Queuing Theory?

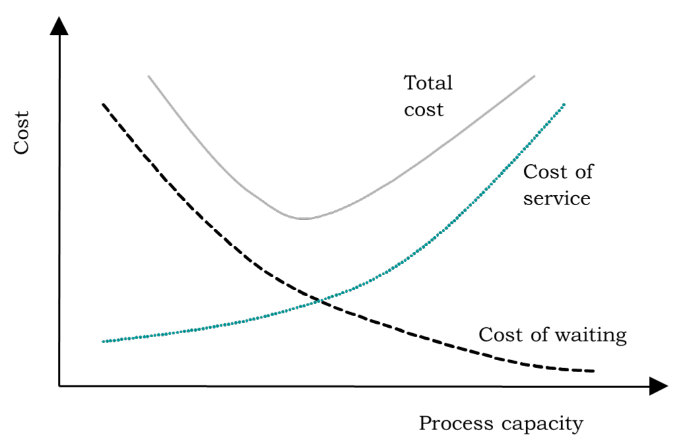
- Mathematical analysis of queues and waiting times in stochastic systems.
  - Used extensively to analyze production and service processes exhibiting random variability in market demand (arrival times) and service times.

- Queues arise when the short term demand for service exceeds the capacity
  - Most often caused by random variation in service times and the times between customer arrivals.
  - If long term demand for service > capacity, the queue will explode!

Why is Queuing Analysis Important?

- Capacity problems are very common in industry and one of the main drivers of process redesign
  - Need to balance the cost of increased capacity against the gains of increased productivity and service

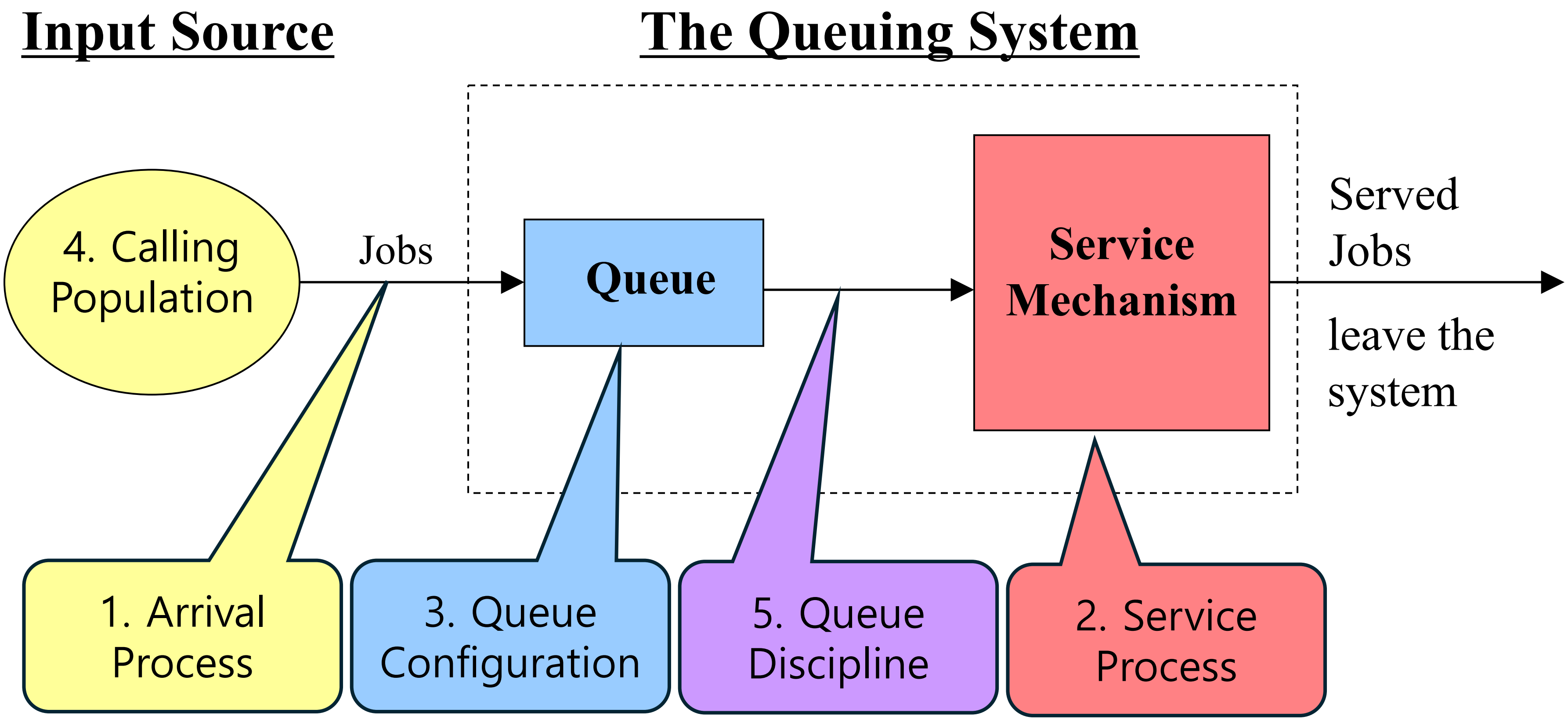
- Queuing and waiting time analysis is particularly important in service systems
  - Large costs of waiting and of lost sales due to waiting



Examples of Real World Queuing Systems?

- Commercial Queuing Systems
  - Commercial organizations serving external customers
  - Ex. Dentist, bank, ATM, gas stations, plumber, garage …
- Transportation service systems
  - Vehicles are customers or servers
  - Ex. Vehicles waiting at toll stations and traffic lights, trucks or ships waiting to be loaded, taxi cabs, fire engines, elevators, buses …
- Business-internal service systems
  - Customers receiving service are internal to the organization providing the service
  - Ex. Inspection stations, conveyor belts, computer support …
- Social service systems
  - Ex. Judicial process, the ER at a hospital, waiting lists for organ transplants or student dorm rooms …

Components of a Basic Queuing Process



Kendall's Notation

- D. G. Kendall proposed describing queueing models using three factors written A/S/c/K/N/D, where
  - A denotes the time between arrivals to the queue
  - S the service time distribution
  - c the number of service channels open at the node.
  - K is the capacity of the queue
  - N is the size of the population of jobs to be served
  - D is the queueing discipline

Kendall's Notation

- Arrival Process
  - M = Memoryless = Poisson
  - E = Erlang
  - H = Hyper-exponential
  - G = General ⇒ Results valid for all distributions

- Service Time Distribution
  - Distribution: M, E, H, or G

- Service Disciplines
  - First-Come-First-Served (FCFS)
  - Last-Come-First-Served (LCFS)
  - Last-Come-First-Served with Preempt and Resume (LCFS-PR)
  - Round-Robin (RR) with a fixed quantum.
  - Small Quantum ⇒ Processor Sharing (PS)
  - Infinite Server: (IS) = fixed delay
  - Shortest Processing Time first (SPT)
  - Shortest Remaining Processing Time first (SRPT)
  - Shortest Expected Processing Time first (SEPT)
  - Shortest Expected Remaining Processing Time first (SERPT).
  - Biggest-In-First-Served (BIFS)
  - Loudest-Voice-First-Served (LVFS)

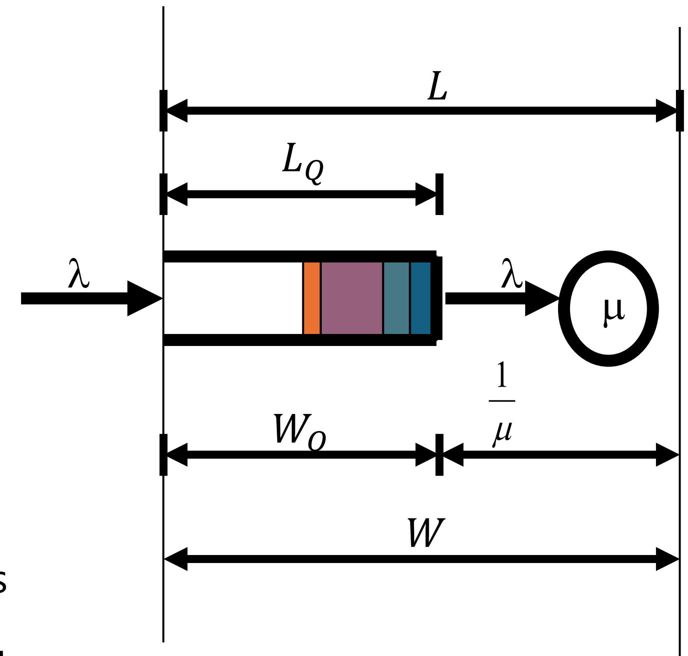
Example: M/M/3/20/1500/FCFS

- Time between successive arrivals is exponentially distributed.
- Service times are exponentially distributed.
- Three servers
- 20 Buffers = 3 service + 17 waiting
  - After 20, all arriving jobs are lost
- Total of 1500 jobs that can be serviced.
- Service discipline is first-come-first-served.
- Defaults:
  - Infinite buffer capacity
  - Infinite population size
  - FCFS service discipline.
- $M/M/1 = M/M/1/\infty/\infty/FCFS$
- $M/G/1/10 = M/G/1/10/\infty/FCFS$

Preliminaries: Cost Equations

- Key variables
  - $N(t)$=Number of customers in the system at time t
  - $P_n(t)$=Probability that at time t, there are n customers in the system.
  - λ_n=Average arrival intensity at n customers in the system
  - μ_n=Average service intensity for the system when there are n customers in it
  - ρ=Utilization factor for the service facility

- Some fundamental quantities of interest for queueing models are
  - L: the average number of customers in the system
  - L_Q: the average number of customers waiting in queue
  - W: the average amount of time a customer spends in the system
  - W_Q: the average amount of time a customer spends waiting in queue



M/M/1 Queue

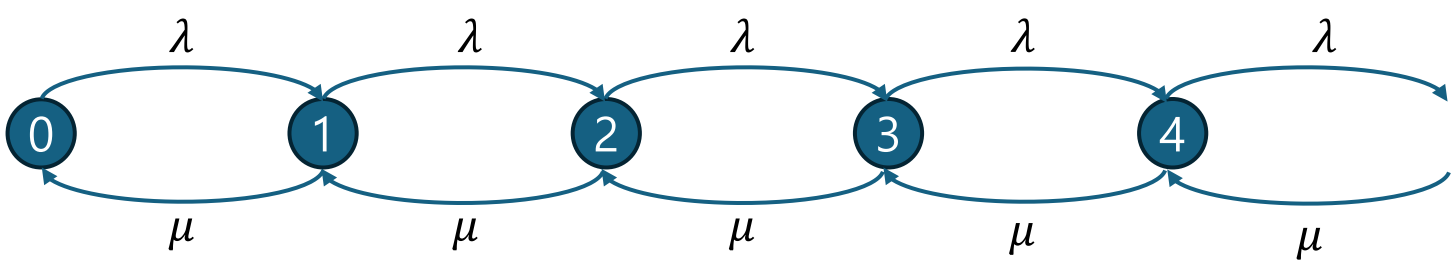
- Arrival process: Poisson process with λ
  - $P\{N(t) = 1\} = \lambda\Delta t + o(\Delta t)$
- Service time distribution: exponential with μ
  - The probability of a service completion in $(t, t + \Delta t) = \mu\Delta t + o(\Delta t)$
- The number of servers: one
- The capacity of queues: infinity
- Let us define p_n as the probability that there are n customers in the queue, including the one in service

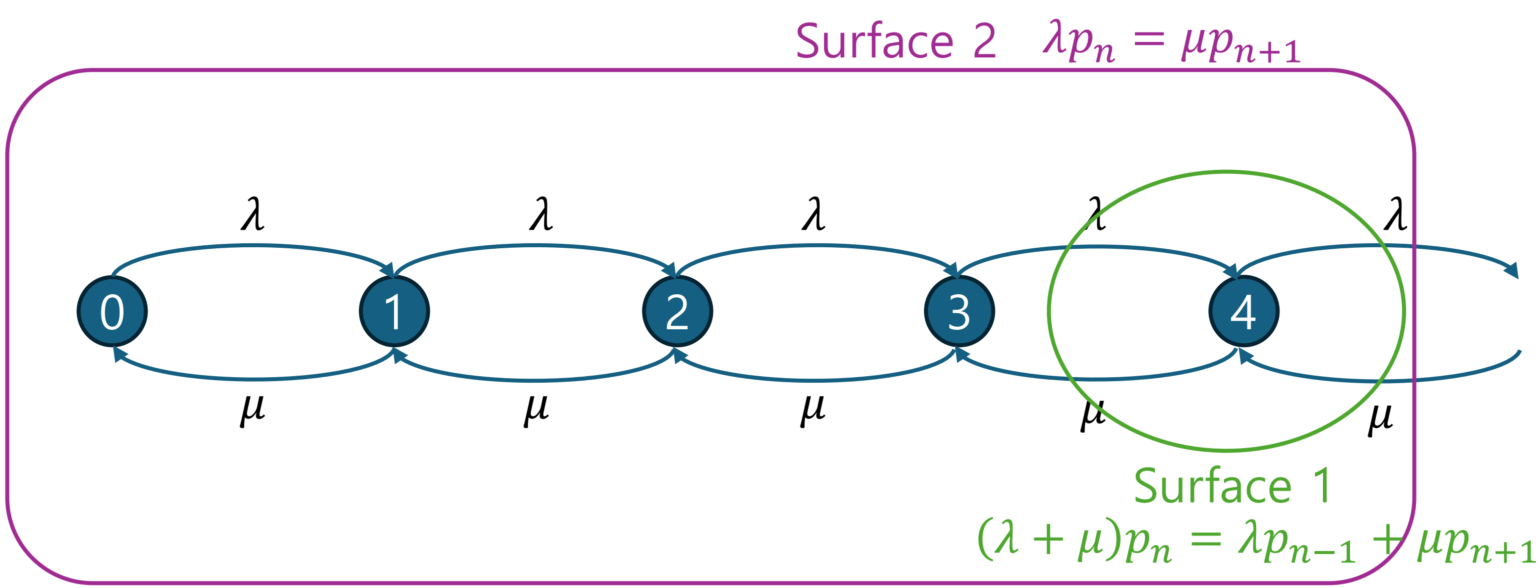
M/M/1 Queue: Transient and Stationary behaviors

- Transient behavior
  - $$p_n(t+\Delta t) = p_n(t)[(1-\lambda\Delta t)(1-\mu\Delta t) + \mu\Delta t\cdot\lambda\Delta t + o(\Delta t)]$$
    $$+p_{n-1}(t)[\lambda\Delta t(1-\mu\Delta t) + o(\Delta t)]$$
    $$+p_{n+1}(t)[\mu\Delta t(1-\lambda\Delta t) + o(\Delta t)]$$
    $$\rightarrow p_n(t+\Delta t) = p_n(t)[1-(\lambda+\mu)\Delta t] + p_{n-1}(t)\lambda\Delta t + p_{n+1}(t)\mu\Delta t$$
  - From a Tailor series such that $p_n(t+\Delta t) = p_n(t) + \frac{dp_n(t)}{dt}\Delta t$
    $$\frac{dp_n(t)}{dt} = -(\lambda+\mu)p_n(t) + p_{n-1}(t)\lambda + p_{n+1}(t)\mu$$

- Stationary behavior (Non-time varying probability) $\frac{dp_n(t)}{dt} = 0$
  - $(\lambda+\mu)p_n = \lambda p_{n-1} + \mu p_{n+1}$ for $n \geq 1$

M/M/1 Queue: State Diagram



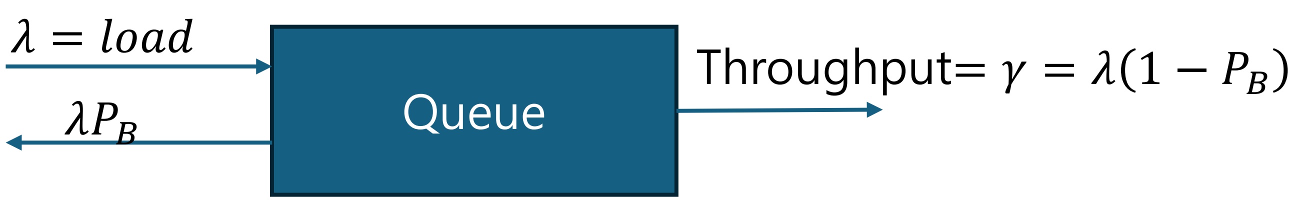


M/M/1 Queue: State Probability

- $\lambda p_n = \mu p_{n+1}$ (balance equation)
- $p_n = \rho^n p_0$, $\rho = \frac{\lambda}{\mu}$
- $\Sigma_n p_n = 1 \rightarrow p_n = (1-\rho)\rho^n$, $\rho = \frac{\lambda}{\mu} < 1$
  - M/M/1 state probability distribution is geometric distribution
- $\rho < 1$ is the necessary condition. If violated, equilibrium is never reached.
- The M/M/1 state probability distribution is geometric
  - p_0 is the probability that the queue is empty.
  - ρ is called utilization, the probability that the queue is non-empty.

M/M/1/N Queue

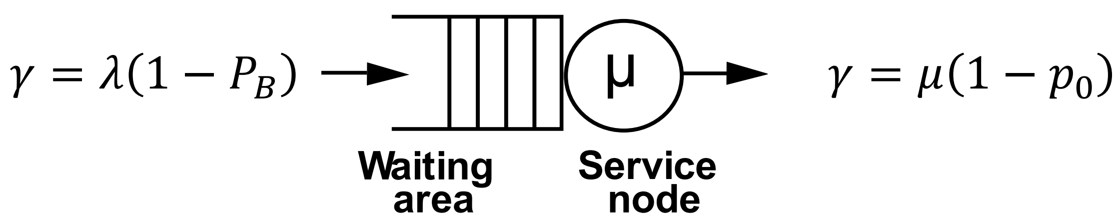
- Accommodating at most N packets
- The balance equation is unchanged except for the two boundary points $n = 0$, and $n = N$
- $\Sigma_n p_n = 1 \rightarrow p_0 = \frac{1-\rho}{1-\rho^{N+1}}, \; p_n = \rho^n p_0 = \rho^n \left(\frac{1-\rho}{1-\rho^{N+1}}\right), n = 1,2 \ldots, N$
- $p_N = \rho^N \left(\frac{1-\rho}{1-\rho^{N+1}}\right)$
    - The probability that the queue is full=the probability of blocking P_B
- With the blocking probability P_B, the net arrival rate is $\lambda(1 - P_B)$
    - Throughput= $\gamma = \lambda(1 - P_B)$



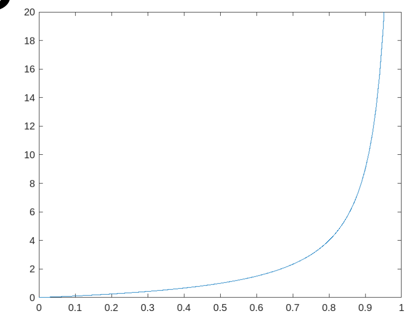
M/M/1/N Queue: Another Interpretation

- For single-server queue, the type of queue, the average rate of service would be μ in customers/sec served on the average, if the queue were always nonempty
- Since the queue is sometimes empty, with probability p_0, the actual rate of service, or throughput γ, is less than μ.
- More precisely, $\gamma = \mu(1 - p_0)$, since $(1 - p_0)$ is the probability that the queue is nonempty
  - ◆There is no blocking and the throughput $\gamma = \lambda$, the average arrival rate. $\lambda = \mu(1 - p_0)$
- $\gamma = \lambda(1 - P_B) = \mu(1 - p_0)$

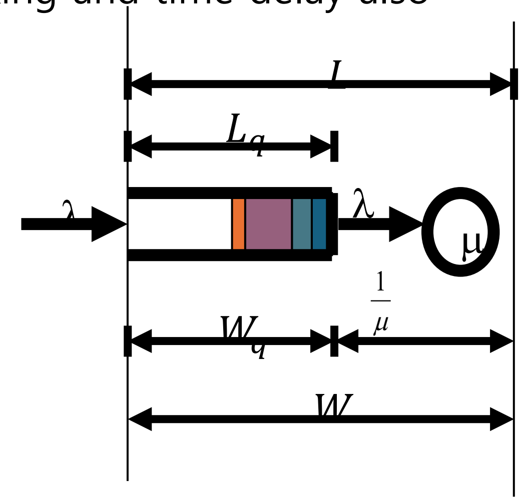
  → $P_B = p_N = \rho^N \left(\frac{1-\rho}{1-\rho^{N+1}}\right) \xrightarrow{\rho^{N+1} \ll 1} (1-\rho)\rho^n$
- The region $\rho > 1$ is said to be the congested region



M/M/1 Queue: Properties



- Statistics of interest: $E(L), E(W_q), E(W)$ etc
- Average of the queue size
  - $E(L) = \Sigma_n n p_n = \frac{\rho}{1-\rho}$
  - As the load increases the throughput goes up but blocking and time delay also increase.



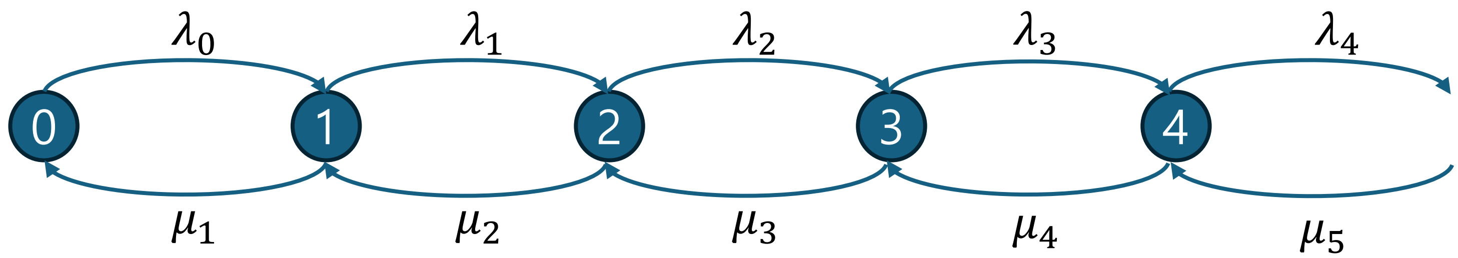
- Little's formula: $\bar{L} = \lambda \bar{W}$
  - We will not drive it here.
  - $\bar{L}$: the average number of customers
  - $\bar{W}$: the average waiting time
  - λ: the arrival rate
- The average wait time and the average delay
  - $E(L_q) = E(W_q) + \frac{1}{\mu}$
- The average number of customers $E(L_q)$ waiting in the queue
  - $E(L_q) = \lambda E(W_q) = \lambda E(W) - \frac{\lambda}{\mu} = \lambda E(W) - \rho$

State Dependent Queues

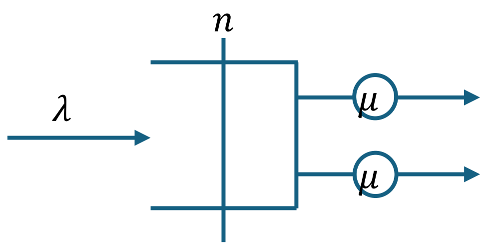
- Arrival and departure rates are dependent on the state of the system (Birth-death process) (e.g) M/M/m
- The balance equation governing the operation of the state-dependent queueing system at equilibrium

$$(\lambda_n + \mu_n)p_n = \lambda_{n-1}p_{n-1} + \mu_{n+1}p_{n+1}, for\ n \geq 1$$
$$\rightarrow \lambda_n p_n = \mu_{n+1}p_{n+1}, for\ n \geq 1$$

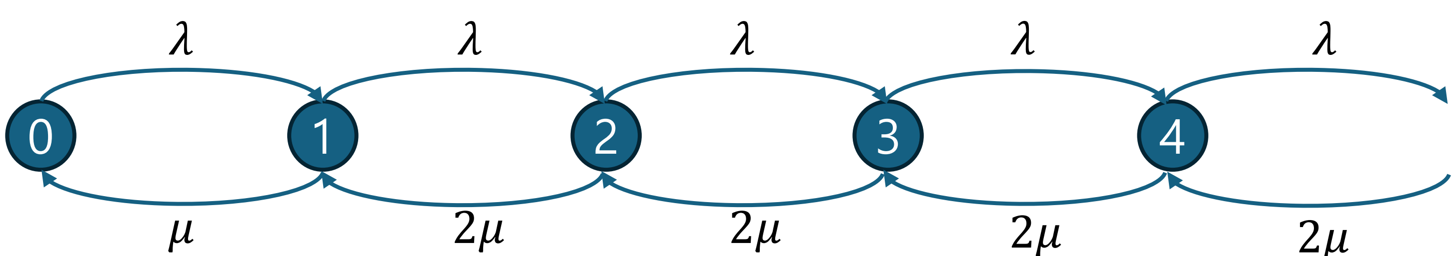
- $p_n/p_0 = \prod_{i=0}^{n-1}\lambda_i \,/\, \prod_{i=1}^{n}\mu_i$
- $\Sigma_n p_n = 1$



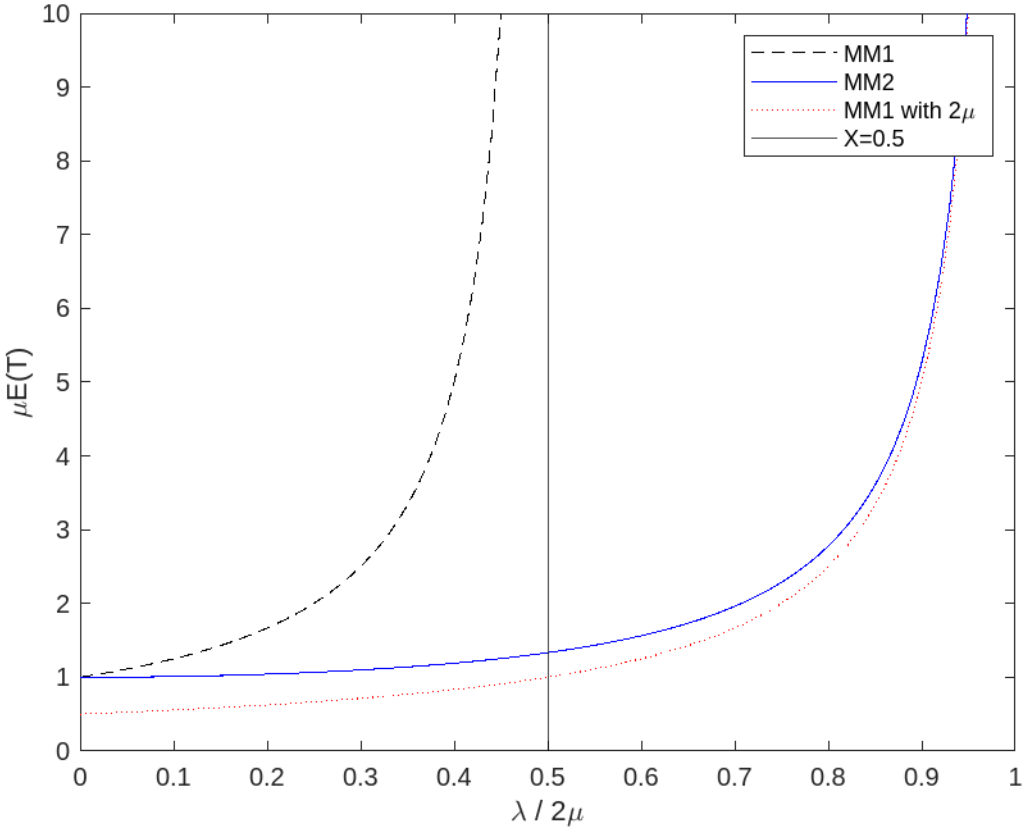
M/M/2



- Two outgoing trunks connecting a packet switch to a neighboring packet switch
  - $\lambda_n = \lambda$
  - $\mu_n = \mu$, $n = 1$, and $\mu_n = 2\mu$, $n > 1$
- $\frac{p_n}{p_0} = \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} = \left(\frac{\lambda}{2\mu}\right)^{n-1}\left(\frac{\lambda}{\mu}\right) = 2\rho^n, n \geq 1, \rho = \frac{\lambda}{2\mu}$
  - $\Sigma_n p_n = 1 \rightarrow p_0 = \frac{1-\rho}{1+\rho}$,
  - $p_n = p_0(2\rho^n) = 2\left(\frac{1-\rho}{1+\rho}\right)\rho^n$
- $E(L) = \Sigma_n n p_n = \frac{2\rho}{1-\rho^2}, \rho = \frac{\lambda}{2\mu}$,
  - $E(W) = \frac{E(L)}{\lambda} = \frac{2}{1-\rho^2}\frac{1}{\mu}$
- The average throughput $\gamma = \mu p_1 + 2\mu(1 - p_0 - p_1)$



M/M/1 vs M/M/2 vs M/M/1 with 2μ



Other examples

- $M/M/\infty$
- Queue with discouragement
- M/M/N/N

Other examples: M/M/∞

- The number of trunks is equal to the number of calls. No queueing up for service nor a probability of blocking
  - $\lambda_n = \lambda$
  - $\mu_n = n\mu$, $n \geq 1$
- $\frac{p_n}{p_0} = \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} = \frac{\rho^n}{n!}, n \geq 1, \rho = \frac{\lambda}{\mu}$
  - $\Sigma_n p_n = 1 \rightarrow p_0 = e^{-\rho}$,
  - $p_n = p_0 \frac{\rho^n}{n!} = e^{-\rho} \frac{\rho^n}{n!}$ (Poisson!!!)
- $E(L) = \Sigma_n n p_n = \rho = \frac{\lambda}{\mu}$,
  - $E(T) = \frac{E(n)}{\lambda} = \frac{1}{\mu}$
- The average throughput

$$\gamma = \Sigma_{n\geq 0} \mu_n p_n = \mu \Sigma_{n\geq 0} n p_n = \mu E(L) = \lambda$$

Other examples: Queue with discouragement

- A system with customer flow control at the input (e.g.) moviegoers and shoppers at a single line to serve
  - $\lambda_n = \frac{\lambda}{n+1}$
  - $\mu_n = \mu$, $n \geq 1$
- $\frac{p_n}{p_0} = \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} = \frac{\rho^n}{n!}, n \geq 1, \rho = \frac{\lambda}{\mu}$ (the same as M/M/∞)
  - $\Sigma_n p_n = 1 \rightarrow p_0 = e^{-\rho}$,
  - $p_n = p_0 \frac{\rho^n}{n!} = e^{-\rho} \frac{\rho^n}{n!}$
- The average throughput

$$\gamma = \Sigma_{n \geq 0} \lambda_n p_n = \mu(1 - e^{-\rho})$$

- $E(L) = \Sigma_n n p_n = \rho = \frac{\lambda}{\mu}$,
  - $E(W) = \frac{E(n)}{\gamma} = \frac{1}{\mu} \frac{1}{(1-e^{-\rho})}, \rho = \frac{\lambda}{\mu}$

Other examples: M/M/N/N

- The number of trunks is equal to the number of calls. No queueing up for service nor a probability of blocking
  - $\lambda_n = \lambda$
  - $\mu_n = n\mu,\ n \geq 1$
- $\frac{p_n}{p_0} = \frac{\prod_{i=0}^{n-1} \lambda_i}{\prod_{i=1}^{n} \mu_i} = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} = \frac{\rho^n}{n!}, n \geq 1, \rho = \frac{\lambda}{\mu}$
  - $\sum_{n=0}^{N} p_n = 1 \rightarrow p_0 = \frac{1}{\sum_{n=0}^{N} \frac{\rho^n}{n!}},$
  - $p_n = \frac{\rho^n}{n!} \frac{1}{\sum_{n=0}^{N} \frac{\rho^n}{n!}}$
  - $p_B = \frac{\rho^N}{N!} \frac{1}{\sum_{n=0}^{N} \frac{\rho^n}{n!}}$ (Erlang-B distribution)
- $E(L) = \rho(1 - P_B), \rho = \frac{\lambda}{\mu}$
- The average throughput

$$\gamma = \lambda(1 - P_B) = \sum_{n=0}^{N} \mu_n p_n = \mu E(L)$$

- $E(W) = \frac{E(L)}{\gamma} = \frac{1}{\mu}$

M/G/1 Queue: Mean Value Analysis

- The arrival process is Poisson, but a general service-time distribution
- The Pollaczek-Khinchine formulas
  - $E(L) = \left(\frac{\rho}{1-\rho}\right)\underbrace{\left[1 - \frac{\rho}{2}(1 - \mu^2\sigma^2)\right]}_{correction\ factor}$
  - $E(W) = \frac{E(L)}{\lambda} = \left(\frac{\frac{1}{\mu}}{1-\rho}\right)\left[1 - \frac{\rho}{2}(1 - \mu^2\sigma^2)\right]$
  - $\rho = \frac{\lambda}{\mu} = \lambda E(\tau)$
    - λ: the average Poisson arrival rate
    - $E(\tau) = \frac{1}{\mu}$: the average service time
  - σ^2: the variance of the service-time distribution
    - For $\sigma^2 > \frac{1}{\mu^2}$, $E(L)$ and $E(W)$ increase as σ^2 increases
    - For $\sigma^2 < \frac{1}{\mu^2}$, $E(L)$ and $E(W)$ decrease relative to M/M/1 as σ^2 decreases
  - $E(W_q) = E(W) - \frac{1}{\mu} = \frac{\lambda E(\tau^2)}{2(1-\rho)}$
    - $E(\tau^2) = \sigma^2 + \frac{1}{\mu^2}$: the second moment of the service-time distribution
- M/M/1: $\sigma^2 = \frac{1}{\mu^2}$
- M/D/1: : $\sigma^2 = 0$, $E(L) = \left(\frac{\rho}{1-\rho}\right)\left(1 - \frac{\rho}{2}\right)$, $E(W) = \left(\frac{\frac{1}{\mu}}{1-\rho}\right)\left(1 - \frac{\rho}{2}\right)$

Example 8.2 M/M/1

- Suppose that customers arrive at a Poisson rate of one per every 12 minutes, and that service time is exponential at a rate of one service per 8 minutes. What are $E(L)$ and $E(W)$?
- [Answer] $\lambda = \frac{1}{12}$, $\mu = \frac{1}{8}$, $\rho = \frac{\frac{1}{12}}{\frac{1}{8}} = \frac{8}{12} = \frac{2}{3}$
  - $E(L) = \Sigma_n n p_n = \frac{\rho}{1-\rho} = 2$
  - $E(W) = \frac{E(n)}{\lambda} = \frac{1/\mu}{1-\rho} = 24$

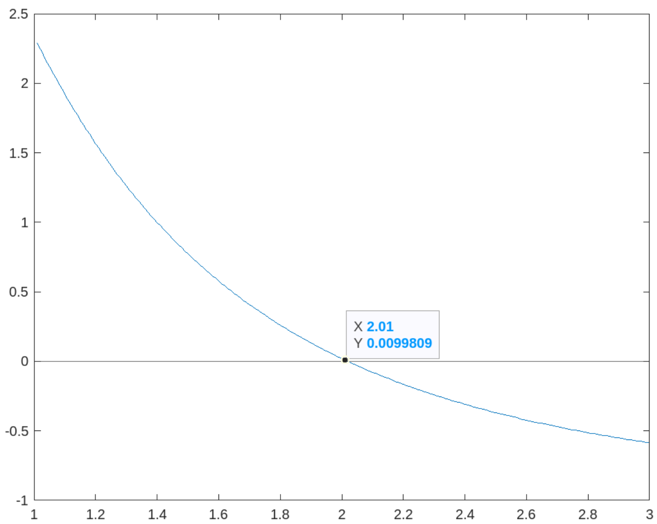
Example 8.3 M/M/N/N

- Suppose that it costs $c\mu$ dollars per hour to provide service at rate μ. Suppose also that we incur a gross profit of A dollars for each customer served. If the system has a capacity N, what service rate maximizes our total profit? Assume that the arrival rate is λ.
- [Answer]
  - $P_n = \rho^n \left(\frac{1-\rho}{1-\rho^{N+1}}\right), \rho = \frac{\lambda}{\mu}$
  - Throughput is $\gamma = \lambda(1 - P_N)$
  - Profit per hour = $\lambda(1 - \mathrm{P_N})\mathrm{A} - \mathrm{c}\mu$

$$= \lambda\left(1 - \rho^{N+1}\left(\frac{1-\rho}{1-\rho^{N+1}}\right)\right)\mathrm{A} - \mathrm{c}\mu$$

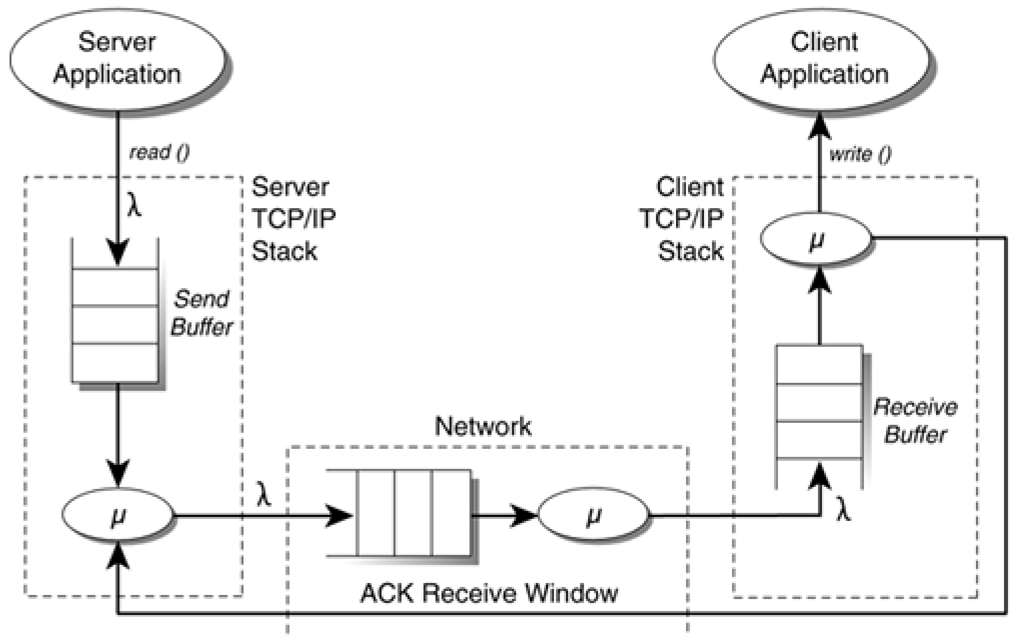
$$= \frac{\lambda \mathrm{A}(1-\rho^N)}{1-\rho^{N+1}} - \mathrm{c}\mu$$

  - If $\mathrm{N} = 2, \lambda = 1, A = 10, c = 1$, then
    - Profit per hour = $\frac{10(\mu^3-\mu)}{\mu^3-1} - \mu$
    - $\frac{d}{du}[profit\ per\ hour] = \frac{10(2\mu^3-3\mu^2+1)}{(\mu^3-1)^2} - 1$



Network of Queues

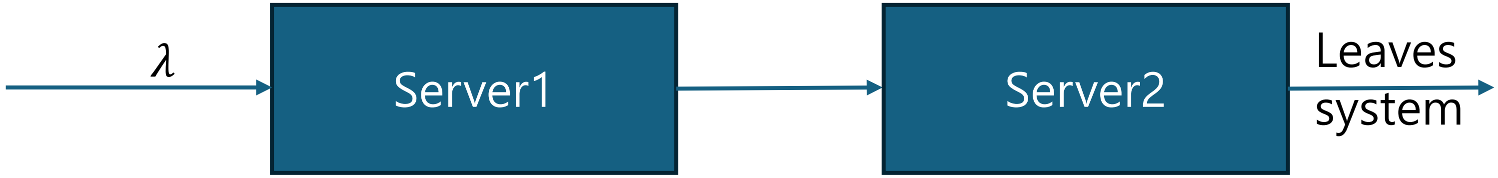
- Network of Queue: model in which jobs departing from one queue arrive at another queue (or possibly the same queue)
- Open systems
  - Customers arrive from outside the system are served and then depart.
  - Example: Packet switched data network.
- Closed systems
  - Fixed number of customers (K) are trapped in the system and circulate among the queues.
  - Example: CPU job scheduling problem
- Mixed systems



Server Application
read ()
λ
Server TCP/IP Stack
Send Buffer
μ
λ
Network
μ
ACK Receive Window
Client Application
write ()
Client TCP/IP Stack
μ
Receive Buffer
λ

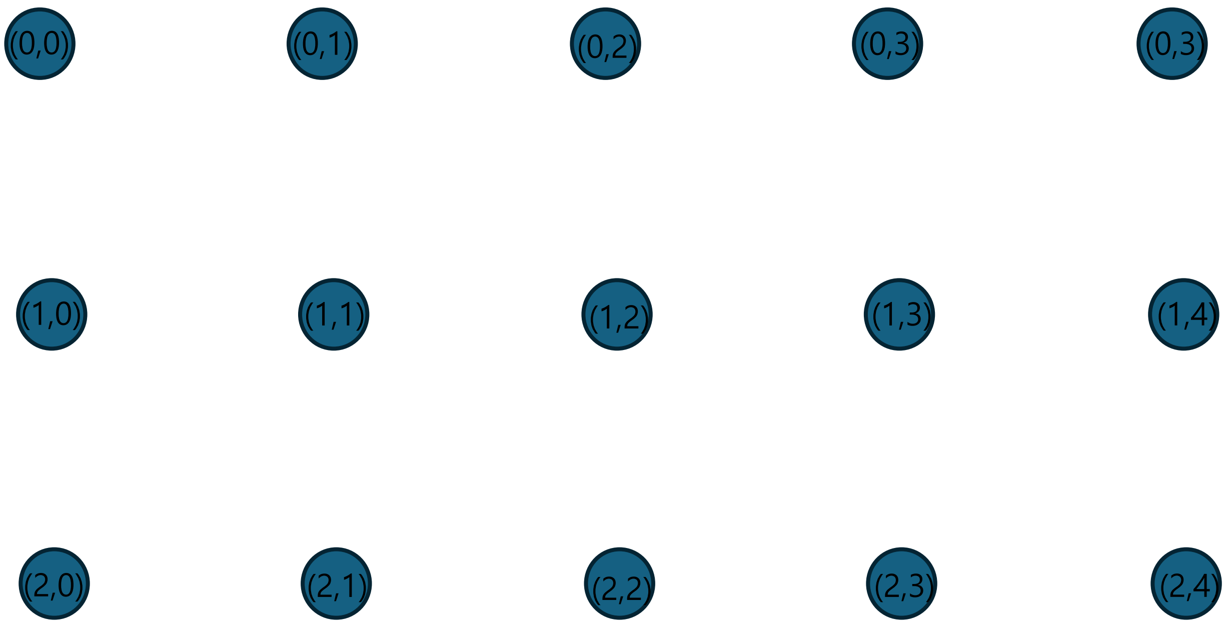
Example of Open Systems: Tandom system

- Consider a two-server system in which customers arrive at a Poisson rate λ at server 1. After being served by server 1 then join the queue in front of server 2
- There is infinite waiting space at both servers
- Each server serves on customer at a time which server i taking an exponential time with rate μ_i for a server $i = 1, 2$.
- A tandom (or sequential) system



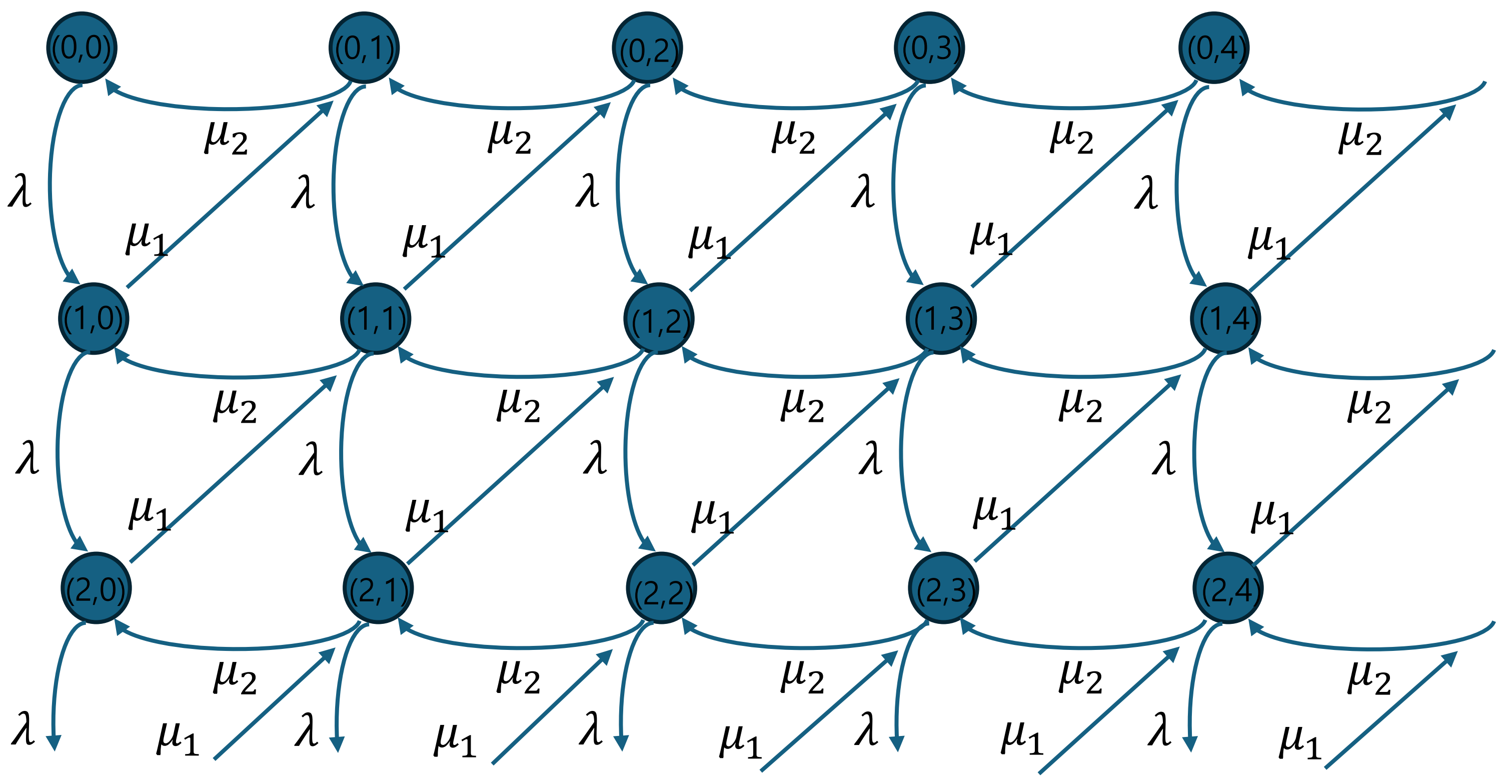
Analysis of a Tandom System

- To analyze this system, keep tack of the number of customers at the servers and define the state by the pair (n, m), where n customers at server 1 and m customers at server 2



Analysis of a Tandom System

- To analyze this system, keep tack of the number of customers at the servers and define the state by the pair (n, m), where n customers at server 1 and m customers at server 2



Balance Equation

State	Rate that the process leaves = rate that it enters
$0,0$	$\lambda P_{0,0} = \mu_2 P_{0,1}$
$n, 0; n > 0$	$(\lambda + \mu_1) P_{n,0} = \mu_2 P_{n,1} + \lambda P_{n-1,0}$
$0, m: m > 0$	$(\lambda + \mu_2) P_{0,m} = \mu_2 P_{0,m+1} + \mu_1 P_{1,m-1}$
$n, m; nm > 0$	$(\lambda + \mu_1 + \mu_2) P_{n,m} = \mu_2 P_{n,m+1} + \mu_1 P_{n+1,m-1} + \lambda P_{n-1,m-1}$
$\Sigma_{n,m} P_{n,m} = 1$	

Probability of State

- The departure process of an M/M/1 queue is a Poisson process with rate λ, it follows that what server 2 faces is also an M/M/1 queue.
  - $P\{n\ at\ server\ 1\} = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right)$
  - $P\{m\ at\ server\ 2\} = \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right)$

- If the numbers of customers at servers 1 and 2 were independent random variables, then it would follow that
  - $P_{n,m} = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right) \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right)$

- Verification is left for assignment

Properties

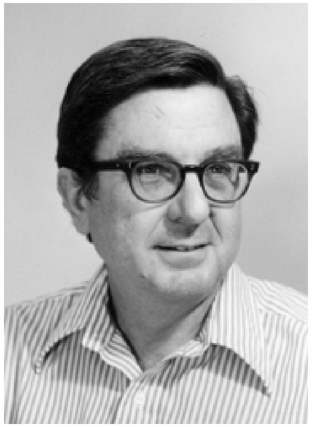
- The average number of customers in the system
  - $E[L] = \sum_{n,m}(n+m)P_{n,m}$
    $$= \sum_n n\left(\frac{\lambda}{\mu_1}\right)^n\left(1-\frac{\lambda}{\mu_1}\right) + \sum_m m\left(\frac{\lambda}{\mu_2}\right)^m\left(1-\frac{\lambda}{\mu_2}\right)$$
    $$= \frac{\lambda}{\mu_1-\lambda} + \frac{\lambda}{\mu_2-\lambda}$$
  - $E[W] = \frac{L}{\lambda} = \frac{1}{\mu_1-\lambda} + \frac{1}{\mu_2-\lambda}$

Generalization of Tandom System: Jackson Network

- Consider system of k servers.
- Customers arrive from outside the system to server i, $i = 1, \dots, k$ in accordance with independent Poisson processes at rate r_i
- Once a customer is served by server i, the then joins the queue in front of server $j, j = 1, \dots, k$ with probability P_{ij}.
- μ_j is the exponential service rate at server j
  - $\lambda_j = r_j + \Sigma_{i=1}^{k} \lambda P_{ij}, i = 1, \dots, k$
  - $P\{n\ at\ server\ j\} = \left(\frac{\lambda_j}{\mu_j}\right)^n \left(1 - \frac{\lambda_j}{\mu_j}\right), n \geq 1, \frac{\lambda_j}{\mu_j} < 1$
  - $P_{n_1,\dots,n_k} = \Pi_{j=1}^{k} \left(\frac{\lambda_j}{\mu_j}\right)^{n_j} \left(1 - \frac{\lambda_j}{\mu_j}\right)$
  - $E[L] = \Sigma_{i=1}^{k} average\ numer\ at\ server\ j = \Sigma_{j=1}^{k} \left(\frac{\lambda_j}{\mu_j - \lambda_j}\right)$
  - $E[W] = \frac{L}{\lambda} = \frac{L}{\Sigma_{j=1}^{k} r_j} = \frac{\Sigma_{j=1}^{k} \left(\frac{\lambda_j}{\mu_j - \lambda_j}\right)}{\Sigma_{j=1}^{k} r_j}$

Jackson Network

- James Jackson (UCLA Math professor) did the basic work on queueing networks
- Jackson Networks – special class of open queueing networks
  - Network of M queues
  - There is only one class of customers in the network
  - A job can leave the network from any node
  - All service times are exponentially distributed with rate μ_j at queue j
  - The service discipline at all nodes is FCFS.
  - All external customer arrival processes are Poisson processes with rate r_j at queue j



Jackson's Theorem

- If in an open network $\lambda_i < \mu_i$ holds for all queues $i = 1, .., M$
  - the arrival rates λ_i can be computed by

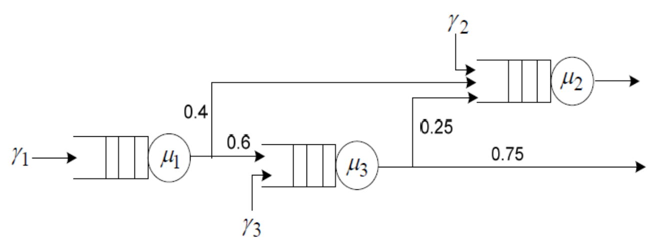
    $$\boldsymbol{\lambda} = \boldsymbol{r}(\boldsymbol{I} - \boldsymbol{R})^{-\mathbf{1}}$$

    where $\boldsymbol{\lambda} = [\lambda_1, \dots, \lambda_M]$
    $\boldsymbol{r} = [r_1, r_2, \dots, r_M]$
    $\boldsymbol{P} = [P_{ij}]$
  - The steady-state probability of the network can be expressed as the product of the state probabilities of the individual queues.

    $$P_{n_1, \dots, n_M} = P_{n_1} P_{n_2} \cdots P_{n_M}$$
  - The nodes of the network can be considered at independent M/M/1 queues with arrival rate λ_i and service rate μ_i

Open Networks: Example

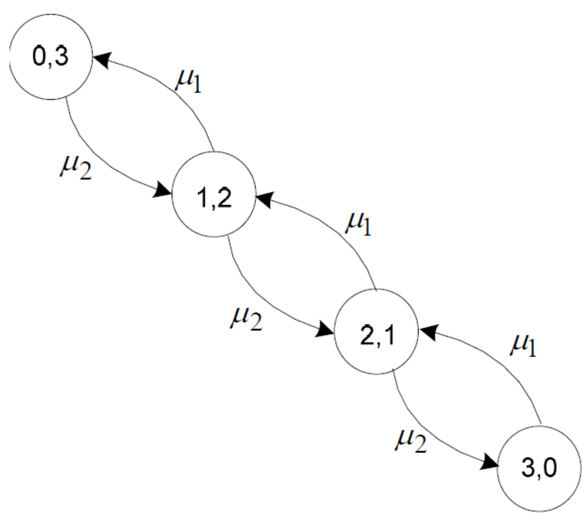
- Three node network shown below
- Poisson external arrivals with $\lambda_1 = 0.5$, $\lambda_2 = 0.25$, $\lambda_3 = 0.25$
- Exponential service at each queue with $\mu_1 = 1$, $\mu_2 = 1$, $\mu_3 = 1$

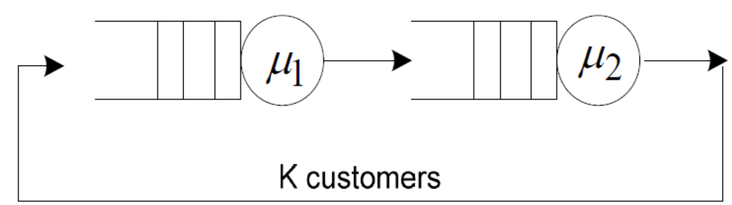


- From the diagram $P_{12} = 0.4, P_{13} = 0.6, P_{32} = 0.25, P_{24} = 1.0, P_{34} = 0.75$
- $R = \begin{bmatrix} 0 & 0.4 & 0.6 \\ 0 & 0 & 0 \\ 0 & 0.25 & 0 \end{bmatrix}$ $\lambda = r(I - R)^1 = [0.5, 0.5875, 055]$

Closed System

- A system in which new customers never enter and existing ones never depart
  - Simplest case k customers circulating among m queues
- Each queue i has exponentially distributed service time μ_i
- State of network defined by $(n_1, n_2, \ldots, n_m)$
- Example m=2, k=3





Closed System

- The limiting probability

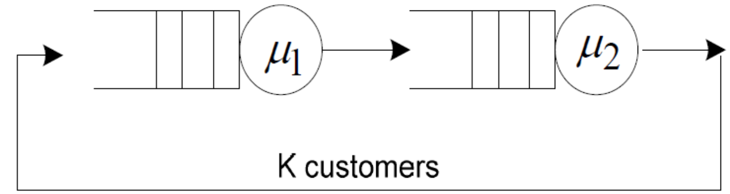
$$P_k(n_1, \dots, n_m) = P\{n_j\ customers\ ate\ server\ j, j = 1, .., m\}$$

- From the balance equation

$$P_k(n_1, \dots, n_m) \begin{cases} M_k \prod_{j=1}^{m} \left(\frac{\lambda_k(j)}{\mu_j}\right)^{n_j} if \sum_{j=1}^{m} n_j = k \\ 0 \end{cases}$$

◆where $M_k = \left[\sum_{\substack{n_1,\dots,n_k: \\ \Sigma n_j = k}} \prod_{j=1}^{m} \left(\frac{\lambda_k(j)}{\mu_j}\right)^{n_j}\right]^{-1}$

Example of Closed Systems



- Example m=2, k=3
- From the diagram $P_{12} = P_{21} = 1$
- State space $S = \{(0,3), (1,2), (2,1), (3,0)\}$
- $M_k = \left[\sum_{\substack{n_1,\dots,n_k:\\ \Sigma n_j = k}} \prod_{j=1}^{m} \left(\frac{\lambda_k(j)}{\mu_j}\right)^{n_j}\right]^{-1} = \left[\left(\frac{\lambda_k(2)}{\mu_2}\right)^3 + \left(\frac{\lambda_k(1)}{\mu_1}\right)^1 \left(\frac{\lambda_k(2)}{\mu_2}\right)^2 + \left(\frac{\lambda_k(1)}{\mu_1}\right)^2 \left(\frac{\lambda_k(2)}{\mu_2}\right)^1 + \left(\frac{\lambda_k(1)}{\mu_1}\right)^3\right]^{-1}$
- Let $\lambda_k(1) = 1 \rightarrow \lambda_k(2) = 1$, then $M_k = 1.875^{-1}$

$$\mathrm{P}(0,3) = M_k \left(\frac{\lambda_k(2)}{\mu_2}\right)^3 = 0.0667$$

$$\mathrm{P}(1,2) = M_k \left(\frac{\lambda_k(1)}{\mu_1}\right)^1 \left(\frac{\lambda_k(2)}{\mu_2}\right)^2 = 0.1333$$

$$\mathrm{P}(2,1) = M_k \left(\frac{\lambda_k(1)}{\mu_1}\right)^2 \left(\frac{\lambda_k(2)}{\mu_2}\right)^1 = 0.2667$$

$$\mathrm{P}(3,0) = M_k \left(\frac{\lambda_k(1)}{\mu_1}\right)^3 = 0.5333$$