Chapter 3: Delay Models in Data Networks

Our goals:

- understand principles of queueing systems:
  - Poisson Process
  - Markov process
  - Little's theorem
  - *Network of queues*
- Determine Performance measures
  - Response time
  - Throughput
  - ?

Introduction- Motivation

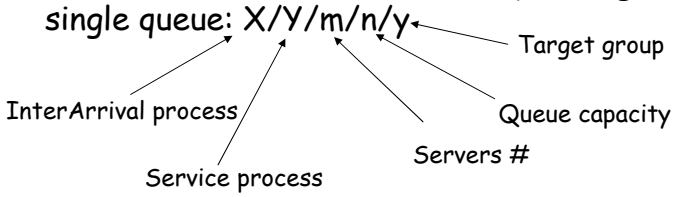
- How to analyze changes in network workloads?
  - Should I add new terminals? How much?
- What percentage of calls will be blocked?
  - Adding more lines would solve the problem?
- Analysis of system (network) load and performance characteristics
  - response time
  - throughput
- Performance tradeoffs are often not intuitive
- Queuing theory, although mathematically complex, often makes analysis very straightforward

Queueing Theory

- Operations Research
- The study of waiting
- Back to early twentieth century
  - Danish mathematician A. K. Erlang (telephone networks), why?
  - Russian mathematician A. A. Markov
- Applied in a broad variety of applications

Queueing Jargons

- Queueing system
  - Customers
  - Queue(s) (waiting room)
  - Server(s)
- Kendall's notation
  - Standard notation to describe queueing containing single queue: X/Y/m/n/y



Common distributions

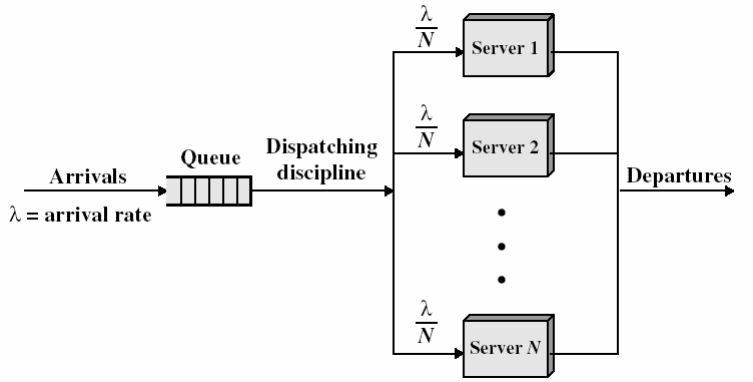
-
  - G = general distribution if interarrival times or service times
  - GI = general distribution of interarrival time with the restriction that they are independent
  - M = negative exponential distribution (Poisson arrivals)
  - D = deterministic arrivals or fixed length service

M/M/1? M/D/1?

General Characteristics of Network Queuing Models

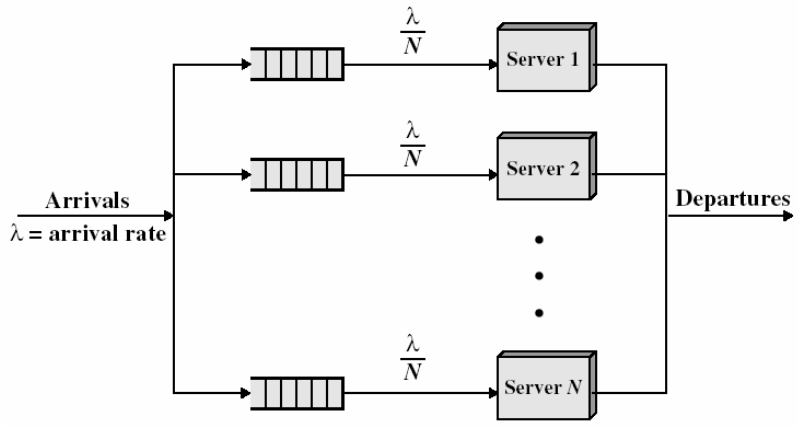
- Item population
  - generally assumed to be infinite therefore, arrival rate is persistent
- Queue size
  - infinite, therefore no loss
  - finite, more practical, but often immaterial
- Dispatching discipline
  - FIFO, typical
  - LIFO
  - Relative/Preferential, based on QoS
  - Processor sharing (PS) discipline
    - Useful for modeling multiprogramming

Multiserver Queue

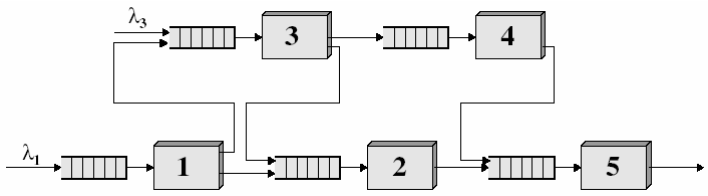


Comments:
1. Assuming N identical servers, and ρ is the utilization of each server.
2. Then, Nρ is the utilization of the entire system, and the maximum utilization is N x 100%.
3. Therefore, the maximum supportable arrival rate that the system can handle is:
   $\lambda_{max} = N / T_s$

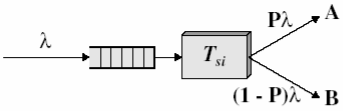
Multiple Single-Server Queues



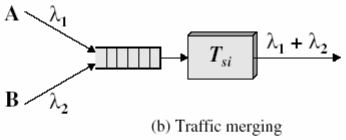
Network of Queues



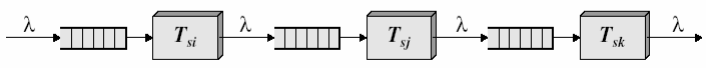
Elements of Queuing Networks



(a) Traffic partitioning

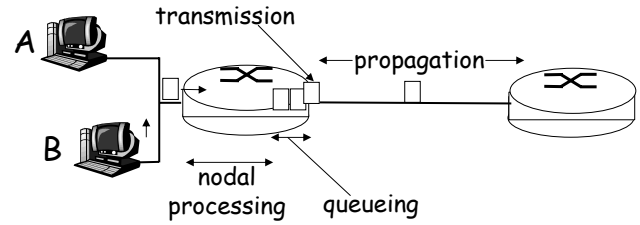


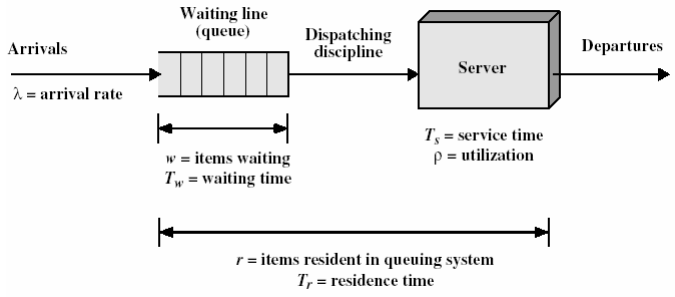
(b) Traffic merging



(c) Simple tandem queue

Delay Components

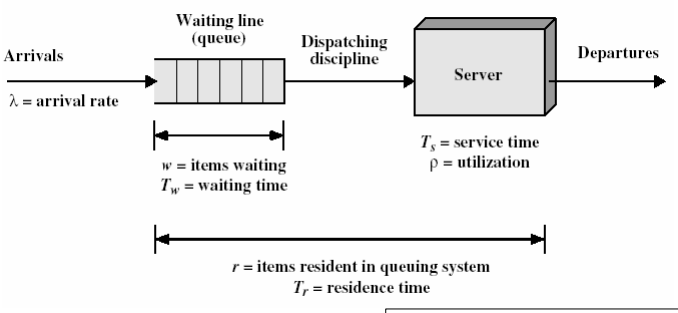




Delay Components (Cont.)

- Packet delay the sum of delays on each link on the path traversed by the packet.
- Each link delay in turns consists of
  - Processing delay: between the time the packet is correctly received at the head node of the link and the time the packet is assigned to an outgoing link queue; is independent of traffic carried.
  - Queueing delay: between the time the packet is assigned to a queue for transmission and the time it starts being transmitted.
  - Transmission delay: between the times that the first and last bits of the packet are transmitted.
  - Propagation delay: between the time the last bit is transmitted at the head node of the link and the time the last bit is received at the tail node; depends on the physical characteristics of the link.

The Fundamental Task of Queuing Analysis



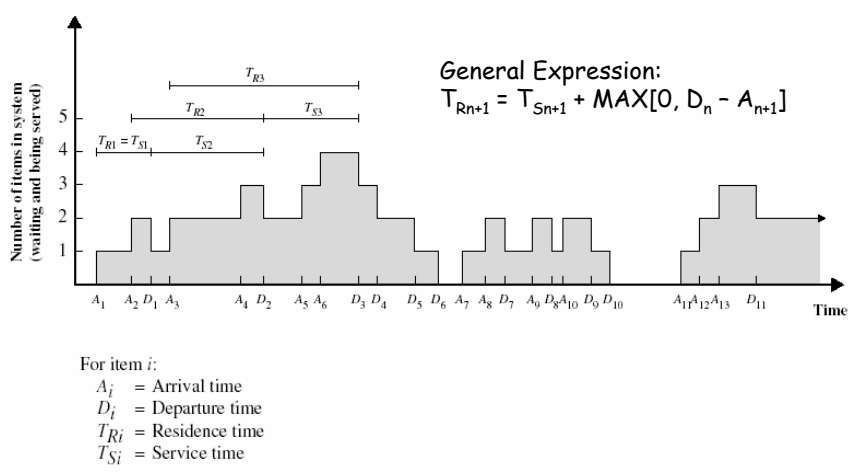
Given:
- Arrival rate, λ
- Service time, T_s
- Number of servers, m

Determine:
- Items waiting, w
- Waiting time, T_w
- Items queued, N_Q
- *Total number, N*
- Residence time, T

Output variables

- Utilisation rate ρ (server utilization, percentage of the time that a server is busy, where m =the number of parallel servers)
- Probability of n customers in the system P_n
- Average number of customers in the system N (service and queue)
- Average number of customers in the queue N_Q
- Average time spent by a customer in the system w (service and queue)
- Average time spent by a customer in the queue w_q

Queuing Process - Example



General Expression:
$T_{Rn+1} = T_{Sn+1} + MAX[0, D_n - A_{n+1}]$

For item i:
A_i = Arrival time
D_i = Departure time
T_{Ri} = Residence time
T_{Si} = Service time

Transient versus steady-state behaviour

- transient behaviour (from t=0)

  performance indicators such as average waiting time, average number of customers in queue, etc. are dependent of the time, e.g. $w_q(t)$, $N_Q(t)$

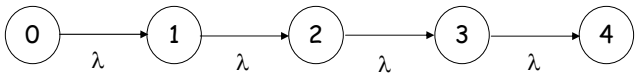
- steady-state (stationary) behaviour ($t \rightarrow \infty$)

  performance indicators such as average waiting time are not dependent of the time anymore; the probability that the system is in a certain state is completely independent of time, e.g. w_q, N_Q

The Poisson distribution

- Three basic assumptions are used to derive a message arrival rate based on the Poisson distribution:
  - Within a very short interval (Δt), the probability of only one message arriving in that interval is high.
    - Prob. Of (exactly one arrival) = $\lambda\Delta t$, λ arrival rate
    - Prob. Of (no arrival) = 1- $\lambda\Delta t$
    - Prob. Of (more than one arrival) = 0

Poisson Process



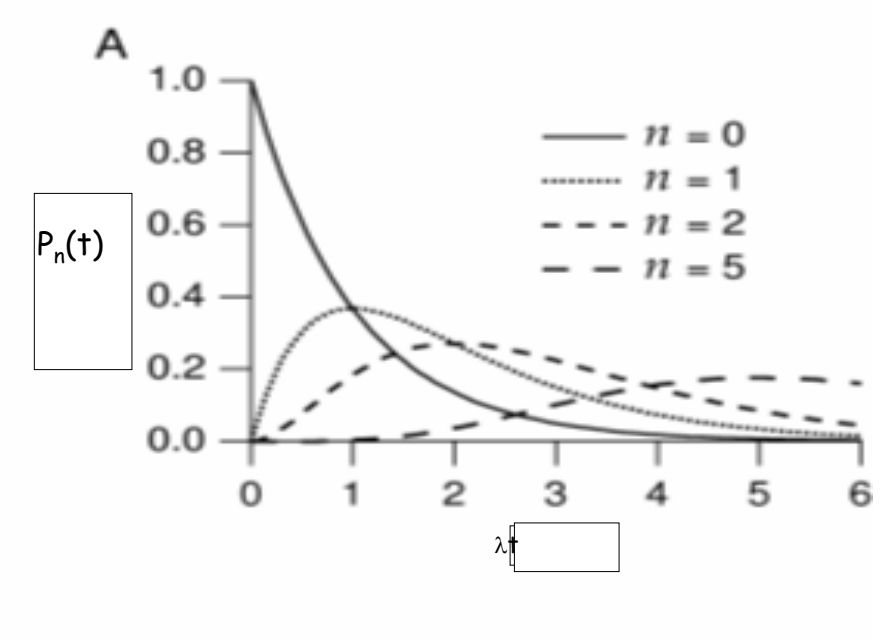
- $P_n(t)$=Prob(n arrivals at t)
  - $p_{n,n}(t)$ is the probability of staying in state n
- $P_0(t+\Delta t) = P_0(t)p_{0,0}(\Delta t)$
  - $P_0(t+\Delta t) = P_0(t)(1-\lambda\Delta t)$ …… (i)
- $P_n(t) = P_n(t)p_{n,n}(\Delta t) + P_{n-1}(t)p_{n-1,n}(\Delta t)$
  - $P_n(t) = P_n(t)(1-\lambda\Delta t) + P_{n-1}(t)(\lambda\Delta t)$ ……..(ii)
- Find a solution for this system of difference of equations

Poisson Process

- We can show that

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

- Then, arrivals are memoryless, i.e. arrivals of messages are independent of each other. This is likely to be the case when the messages are generated from a large number of independent sources.
- The characteristics of the message arrival distribution do not vary depending on the observation period (t).



Exponential Random Variable

- The CDF of an exponential random variable is:

$$F_X(x) = \int_0^x f_X(\bar{x})d\bar{x} = \int_0^x \lambda e^{-\lambda\bar{x}}d\bar{x}$$

- So

$$= \left[-e^{-\lambda\bar{x}}\right]_0^x = 1 - e^{-\lambda x}$$

$$P(X > x) = 1 - F_X(x) = e^{-\lambda x}$$

Memoryless Property of the Exponential

- An exponential random variable ***X*** has the property that "the future is independent of the part", i.e. the fact that it hasn't happened yet, tells us nothing about how much longer it will take.
- In math terms

$$P(X > s + t \mid X > t) = P(X > s) \quad \text{for } s, t > 0$$

Memoryless Property of the Exponential

- Proof:

$$P(X > s+t \mid X > t) = \frac{P(X > s+t, X > t)}{P(X > t)}$$

$$= \frac{P(X > s+t)}{P(X > t)}$$

$$= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}}$$

Example

- Suppose a train arrives at a station according to a Poisson process with average inter-arrival time of 20 minutes
- When a customer arrives at the station the average amount of time until the next arrival is 20 minutes
  - Regardless of when the previous train arrived
- The average amount of time since the last departure is 20 minutes!
- Paradox:
  - If an average of 20 minutes passed since the last train arrived and an average of 20 minutes until the next train, then an average of 40 minutes will elapse between trains
    - But we assumed an average inter-arrival time of 20 minutes!
  - What happened?

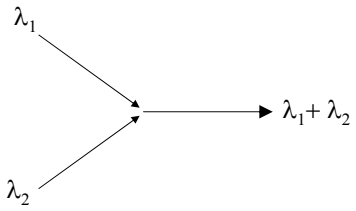
The Poisson distribution

- In the case where in a service facility the next customer is served as soon as the one in service leaves the system, it is apparent that the time between service completions must be equal to the service time.
- Therefore if the time between completions is *exponentially distributed*, then the service time itself is *exponentially distributed* in time. Hence the service time distribution in this case is a *Poisson process*.

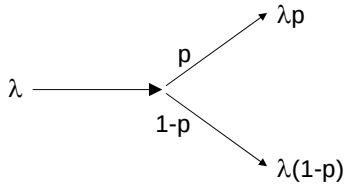
The Poisson distribution

- Other useful properties of the Poisson process:
  - a distribution resulting from the sum of Poisson distributions retains the Poisson distribution.
  - if a Poisson stream is split into multiple substreams, each with a probability P_i of a job going to the ith substream, each substream is also Poisson with a mean rate of P_i.
  - if the arrivals to a multiserver facility are Poisson with each server having exponential service times, the departures also constitute a Poisson stream.

Merging & Splitting Poisson Processes



- $A_1,\ldots, A_k$ independent Poisson processes with rates $\lambda_1,\ldots, \lambda_k$
- Merged in a single process $A= A_1+\ldots+ A_k$
- A is Poisson process with rate $\lambda= \lambda_1+\ldots+ \lambda_k$



- A: Poisson processes with rate λ
- Split into processes A_1 and A_2 independently, with probabilities p and 1-p respectively
- A_1 is Poisson with rate $\lambda_1= \lambda p$
  A_2 is Poisson with rate $\lambda_2= \lambda(1-p)$

Markov Process

- The Poisson process is a special case of a more general process known as the *Markov* process.
- A *Markov* process is a stochastic (random) process that exhibits a particular characteristic, namely that the distribution at any time in the future depends only on the current state of the process and not on how that state was reached (the memoryless property).

Discrete-Time Markov Chain

- Discrete-time stochastic process $\{X_n: n = 0,1,2,\ldots\}$
- Takes values in {0,1,2,…}
- Memoryless property:

$$P\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} = P\{X_{n+1} = j \mid X_n = i\}$$

$$P_{ij} = P\{X_{n+1} = j \mid X_n = i\}$$

- Transition probabilities P_{ij}

$$P_{ij} \geq 0, \quad \sum_{j=0}^{\infty} P_{ij} = 1$$

- Transition probability matrix $P=[P_{ij}]$

State Probabilities – Stationary Distribution

- State probabilities (time-dependent)

$$\pi_j^n = P\{X_n = j\}, \qquad \pi^n = (\pi_0^n, \pi_1^n, \ldots)$$

- In matrix form:

$$P\{X_n = j\} = \sum_{i=0}^{\infty} P\{X_{n-1} = i\} P\{X_n = j \mid X_{n-1} = i\} \Rightarrow \pi_j^n = \sum_{i=0}^{\infty} \pi_i^{n-1} P_{ij}$$

- If time-dependent distribution converges to a limit

$$\pi^n = \pi^{n-1} P = \pi^{n-2} P^2 = \ldots = \pi^0 P^n$$

- π is called the *stationary distribution*

$$\pi = \lim_{n \to \infty} \pi^n$$

- Existence depends on the structure of Markov chain

$$\pi = \pi P$$

Ergodic Markov Chains

- Markov chain with a stationary distribution

$$\pi_j > 0, \quad j = 0,1,2,...$$

- States are positive recurrent: The process returns to state j "infinitely often"
- A positive recurrent and aperiodic Markov chain is called ergodic
- Ergodic chains have a unique stationary distribution

$$\pi_j = \lim_{n\to\infty} P_{ij}^n$$

- Ergodicity ⇒ Time Averages = Stochastic Averages

Calculation of Stationary Distribution

A. Finite number of states

- Solve explicitly the system of equations

$$\pi_j = \sum_{i=0}^{m} \pi_i P_{ij}, \quad j = 0,1,...,m$$

$$\sum_{i=0}^{m} \pi_i = 1$$

- Numerically from P^n which converges to a matrix with rows equal to π
- Suitable for a small number of states

B. Infinite number of states

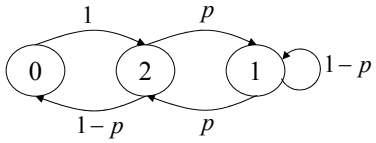
- Cannot apply previous methods to problem of infinite dimension
- Guess a solution to recurrence:

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}, \quad j = 0,1,...,$$

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

Example: Finite Markov Chain

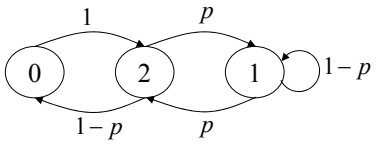
Absent-minded professor uses two umbrellas when commuting between home and office. If it rains and an umbrella is available at her location, she takes it. If it does not rain, she always forgets to take an umbrella. Let p be the probability of rain each time she commutes. What is the probability that she gets wet on any given day?



- Markov chain formulation
- i is the number of umbrellas available at her current location
- Transition matrix

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1-p & p \\ 1-p & p & 0 \end{bmatrix}$$

Example: Finite Markov Chain



$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1-p & p \\ 1-p & p & 0 \end{bmatrix}$$

$$\begin{cases} \pi = \pi P \\ \sum_i \pi_i = 1 \end{cases} \Leftrightarrow \begin{cases} \pi_0 = (1-p)\pi_2 \\ \pi_1 = (1-p)\pi_1 + p\pi_2 \\ \pi_2 = \pi_0 + p\pi_1 \\ \pi_0 + \pi_1 + \pi_2 = 1 \end{cases} \Leftrightarrow \pi_0 = \frac{1-p}{3-p}, \pi_1 = \frac{1}{3-p}, \pi_2 = \frac{1}{3-p}$$

$$P\{\text{gets wet}\} = \pi_0 p = p\frac{1-p}{3-p}$$

Example: Finite Markov Chain

- Taking $p = 0.1$:

$$\pi = \left(\frac{1-p}{3-p}, \frac{1}{3-p}, \frac{1}{3-p}\right) = (0.310, 0.345, 0.345)$$

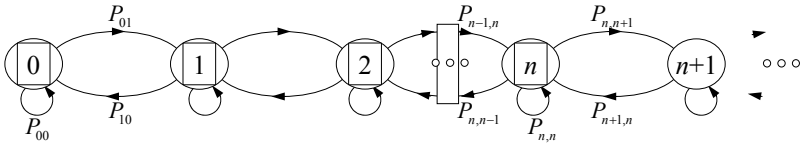
$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0.9 & 0.1 \\ 0.9 & 0.1 & 0 \end{bmatrix}$$

- Numerically determine limit of P^n

$$\lim_{n\to\infty} P^n = \begin{bmatrix} 0.310 & 0.345 & 0.345 \\ 0.310 & 0.345 & 0.345 \\ 0.310 & 0.345 & 0.345 \end{bmatrix} \quad (n \approx 150)$$

- Effectiveness depends on structure of P

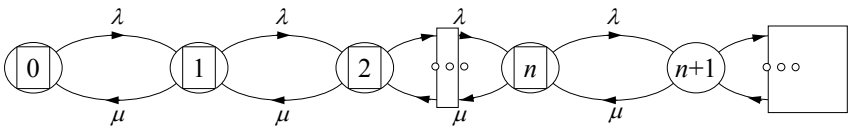
M/M/1 QUEUE ANALYSIS



$$P_n(t+\Delta t) = P_n(t)\,p_{n,n}(\Delta t) + P_{n-1}(t)\,p_{n-1,n}(\Delta t) + P_{n+1}(t)\,p_{n+1,n}(\Delta t)$$

$$P_0(t)(t+\Delta t) = P_0(t)\,p_{0,0}(\Delta t) + P_1(t)\,p_{1,0}(\Delta t)$$

Now, consider the same assumptions used in previous slides when we proved that the arriva
Process is indeed a Poisson Process



Birth-Death Process

Then,

$$P_n(t+\Delta t) = P_n(t)(1-\lambda\Delta t)(1-\mu\Delta t) + P_{n-1}(t)(\lambda\Delta t) + P_{n+1}(t)(\mu\Delta t) \quad \ldots\ldots(1)$$

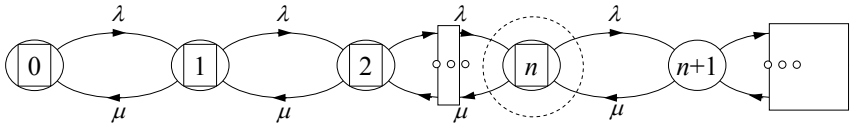
$$P_0(t)(t+\Delta t) = P_0(t)(1-\lambda\Delta t) + P_1(t)(\mu\Delta t) \quad \ldots\ldots(2)$$

Simplify these expressions and let $\Delta t \rightarrow 0$

$$\frac{dP_n(t)}{dt} = -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t),$$

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t) \quad \ldots\ldots(3)$$

Equilibrium or steady-state analysis



• Analogy with Electric Circuits

•***Probability flux*** is the product of the probability of the state at which the transition originates and the transition rate associated with the transition

•Then, input flow = output flow

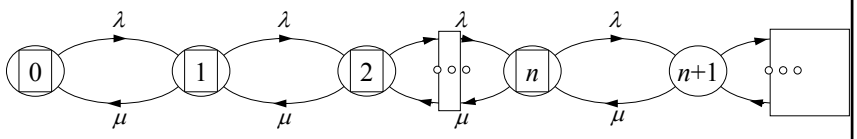
$$\frac{dP_n(t)}{dt} = -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t),$$

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t)$$

•At Equilibrium: $dP_n(t)/dt = 0$

The M/M/1 Queue

- Arrival process: Poisson with rate λ
- Service times: iid, exponential with parameter μ
- Service times and interarrival times: independent
- Single server
- Infinite waiting room
- $N(t)$: Number of customers in system at time t (state)



Exponential Random Variables

- X: exponential RV with parameter λ
- Y: exponential RV with parameter μ
- X, Y: independent

Then:

1. $\min\{X, Y\}$: exponential RV with parameter $\nu = \lambda+\mu$
2. $P\{X<Y\} = \lambda/(\lambda+\mu)$

[Exercise 3.12]

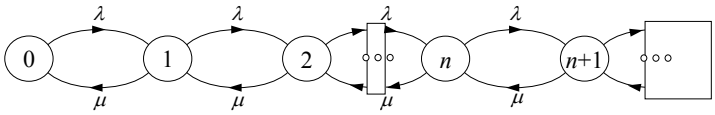
M/M/1 Queue: Markov Chain Formulation

- Jumps of $\{N(t): t \geq 0\}$ triggered by arrivals and departures
- $\{N(t): t \geq 0\}$ can jump only between neighboring states

Assume process at time t is in state i: $N(t) = i \geq 1$

- X_i: time until the next arrival – exponential with parameter λ
- Y_i: time until the next departure – exponential with parameter μ
- $T_i = \min\{X_i, Y_i\}$: time process spends at state i
  - T_i : exponential with parameter $\nu_i = \lambda+\mu$
  - $P_{i,i+1} = P\{X_i < Y_i\} = \lambda/(\lambda+\mu)$, $P_{i,i-1} = P\{Y_i < X_i\} = \mu/(\lambda+\mu)$
  - $P_{01} = 1$, and T_0 is exponential with parameter λ

M/M/1 Queue: Stationary Distribution



- Birth-death process → DBE

$$\mu p_n = \lambda p_{n-1} \Rightarrow$$
$$p_n = \frac{\lambda}{\mu} p_{n-1} = \rho p_{n-1} = \ldots = \rho^n p_0$$

- Normalization constant

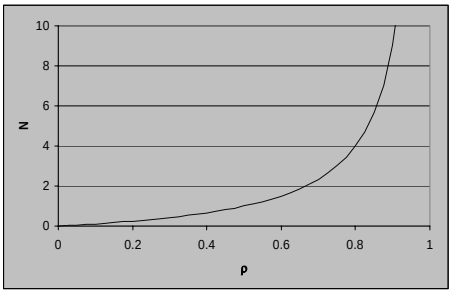
$$\sum_{n=0}^{\infty} p_n = 1 \Leftrightarrow p_0 \left[1 + \sum_{n=1}^{\infty} \rho^n \right] = 1 \Leftrightarrow p_0 = 1 - \rho \ , \ \text{if } \rho < 1$$

- Stationary distribution

$$p_n = \rho^n (1-\rho), \quad n = 0, 1, \ldots$$

The M/M/1 Queue

- ρ=λ/μ: utilization factor
- Long term proportion of time that server is busy
- ρ=1-p_0: holds for any M/G/1 queue
- Stability condition: ρ<1
  - Arrival rate should be less than the service rate



The M/M/1 Queue

- Average number of customers in system

$$N = \sum_{n=0}^{\infty} np_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = (1-\rho)\rho\sum_{n=0}^{\infty} n\rho^{n-1}$$

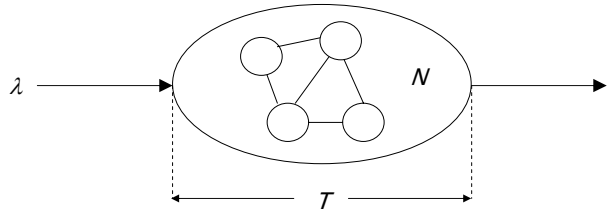
$$\Rightarrow N = \rho(1-\rho)\frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

- Variance of waiting customers in system

$$\sigma_n^2 = \sum_{n=0}^{\infty} (n-\overline{n})^2 p_n$$

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

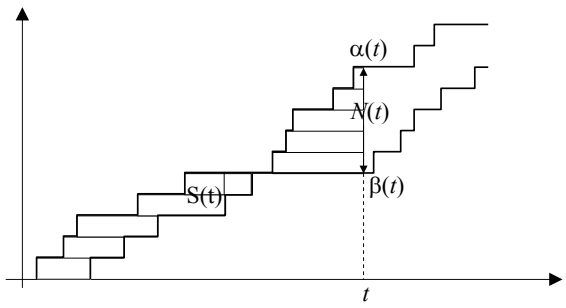
Little's Theorem



- λ: customer arrival rate
- N: average number of customers in system
- T: average delay per customer in system
- Little's Theorem: System in steady-state

$$N = \lambda T$$

Counting Processes of a Queue



- $N(t)$: number of customers in system at time t
- $\alpha(t)$: number of customer arrivals till time t
- $\beta(t)$: number of customer departures till time t
- T_i : time spent in system by the i^{th} customer
- S(t): the cumulative area b/w $\alpha(t)$ and $\beta(t)$

Time Averages

- Time average over interval [0,t]
- Steady state time averages

$$\lambda_t = \frac{\alpha(t)}{t}$$

$$N(t) = \alpha(t) - \beta(t)$$

$$\overline{T_t} = \frac{S(t)}{\alpha(t)}$$

$$\overline{n_t} = \frac{S(t)}{t}$$

subsitute

$$\overline{n_t} = \frac{S(t)}{t} = \frac{\overline{T_t}\alpha(t)}{t} = \lambda_t \overline{T_t}$$

take the limit as $t \to \infty$

$$\overline{n} = \lambda \overline{T}$$

- Little's theorem $\overline{n} = \lambda \overline{T}$
- Applies to any queueing system provided that:
  - Limits $\overline{n}, \lambda, and\ \overline{T}$ exist, and
  - We give a simple graphical proof under a set of more restrictive assumptions

The M/M/1 Queue

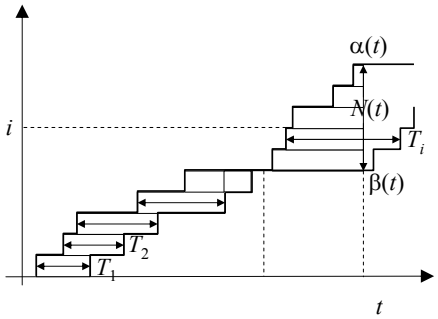
- Little's Theorem: average time in system

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

- Average waiting time and number of customers in the queue – excluding service

$$W = T - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda} \text{ and } N_Q = \lambda W = \frac{\rho^2}{1 - \rho}$$

Proof of Little's Theorem for FCFS



- FCFS system, N(0)=0
  - $\alpha(t)$ and $\beta(t)$: staircase graphs
  - $N(t) = \alpha(t) - \beta(t)$
  - Shaded area between graphs

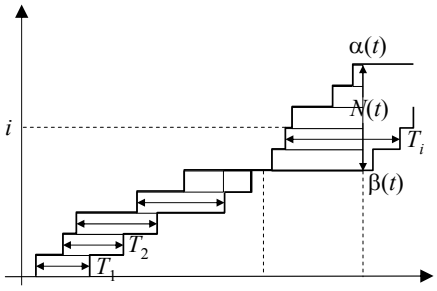
$$S(t) = \int_0^t N(s)ds$$

- Assumption: N(t)=0, infinitely often. For any such t

$$\int_0^t N(s)ds = \sum_{i=1}^{\alpha(t)} T_i \Rightarrow \frac{1}{t}\int_0^t N(s)ds = \frac{\alpha(t)}{t}\frac{\sum_1^{\alpha(t)} T_i}{\alpha(t)} \Rightarrow N_t = \lambda_t T_t$$

  - If limits $N_t \to N$, $T_t \to T$, $\lambda_t \to \lambda$ exist, Little's formula follows
  - We will relax the last assumption

Proof of Little's for FCFS (cont.)



- In general – even if the queue is not empty infinitely often:

$$\sum_{i=1}^{\beta(t)} T_i \le \int_0^t N(s)ds \le \sum_{i=1}^{\alpha(t)} T_i \Rightarrow \frac{\beta(t)}{t}\frac{\sum_1^{\beta(t)} T_i}{\beta(t)} \le \frac{1}{t}\int_0^t N(s)ds \le \frac{\alpha(t)}{t}\frac{\sum_1^{\alpha(t)} T_i}{\alpha(t)}$$

$$\Rightarrow \delta_t T_t \le N_t \le \lambda_t T_t$$

- Result follows assuming the limits $T_t \to T$, $\lambda_t \to \lambda$, and $\delta_t \to \delta$ exist, and $\lambda=\delta$

Probabilistic Form of Little's Theorem

- Have considered a single sample function for a stochastic process
- Now will focus on the probabilities of the various sample functions of a stochastic process
- Probability of n customers in system at time t

$$p_n(t) = P\{N(t) = n\}$$

- Expected number of customers in system at t

$$E[N(t)] = \sum_{n=0}^{\infty} n.P\{N(t) = n\} = \sum_{n=0}^{\infty} np_n(t)$$

Probabilistic Form of Little (cont.)

- $p_n(t)$, $E[N(t)]$ depend on t and initial distribution at t=0
- We will consider systems that converge to steady-state
- there exist p_n independent of initial distribution

$$\lim_{t\to\infty} p_n(t) = p_n, \quad n = 0,1,...$$

- Expected number of customers in steady-state [stochastic aver.]

$$EN = \sum_{n=0}^{\infty} np_n = \lim_{t\to\infty} E[N(t)]$$

- For an ergodic process, the time average of a sample function is equal to the steady-state expectation, with probability 1.

$$N = \lim_{t\to\infty} N_t = \lim_{t\to\infty} E[N(t)] = EN$$

Probabilistic Form of Little (cont.)

- In principle, we can find the probability distribution of the delay T_i for customer i, and from that the expected value $E[T_i]$, which converges to steady-state

$$ET = \lim_{i\to\infty} E[T_i]$$

- For an ergodic system

$$T = \lim_{i\to\infty} \frac{\sum_1^\infty T_i}{i} = \lim_{i\to\infty} E[T_i] = ET$$

- Probabilistic Form of Little's Formula: $EN = \lambda.ET$
- Arrival rate define as

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

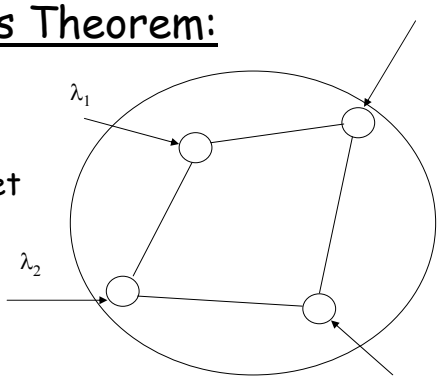
Time vs. Stochastic Averages

- "Time averages = Stochastic averages," for all systems of interest in this course
- It holds if a single sample function of the stochastic process contains all possible realizations of the process at $t\to\infty$
- Can be justified on the basis of general properties of Markov chains

Applications of Little's Theorem:

- Example1:
  - Average delay per packet

$$T = \frac{N}{\sum_{i=1}^{n} \lambda_i}$$



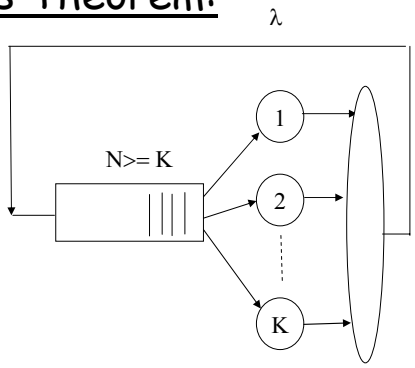
  - Regardless of packet length distribution and method of routing packets

Applications of Little's Theorem:

- Example2:
  - *N=λ T*
  - *Let* $\overline{X}$ *be the average service time*
  - *Then, as all severs are busy*

$$K = \lambda \overline{X}$$

  - $$T = \frac{N\overline{X}}{K}$$



  - *Let* $\overline{K}$ *be the average number of busy servers and* β *is the portion of blocked customers*
  - Then, $\overline{K} = (1-\beta)\lambda\overline{X}$
  - Solve for β $\beta \geq 1 - \frac{K}{\lambda\overline{X}}$

Example3: Estimating throughput in a time-sharing system

- *Maximum attainable throughput happens when we have N waiting*
- *using little's theorem:*

$$\lambda = \frac{N}{T}$$

- *Average delay T=R+D*

$$NP \geq D \geq P$$

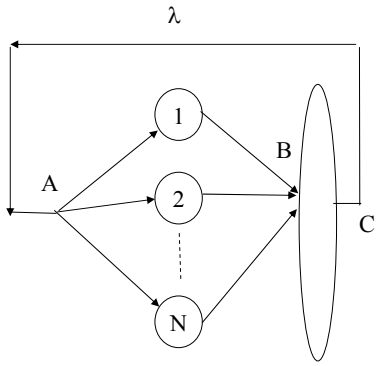
$$NP + R \geq T \geq P + R$$

- *Using these two relations,*

$$\frac{N}{R+NP} \leq \lambda \leq \frac{N}{R+P}$$

- *Considering the processing unit*

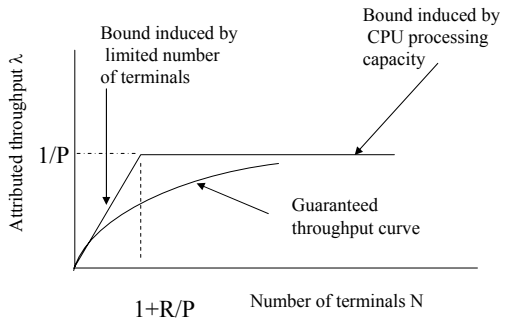
$$\lambda \leq \frac{1}{P} \Rightarrow \frac{N}{R+NP} \leq \lambda \leq \min\left\{\frac{1}{P}, \frac{N}{R+P}\right\}$$



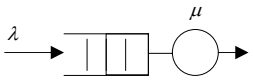
Example3: Cont.

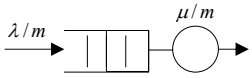
- *Again using little's theorem,*

$$\max\{NP, R+P\} \leq T \leq R+NP$$



M/M/1: Example-I Slowing Down





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

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

$$W = \frac{\rho}{\mu-\lambda} = \frac{\lambda/\mu}{\mu-\lambda}$$

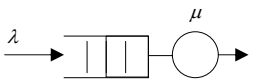
$$N' = \frac{\rho'}{1-\rho'} = \frac{\lambda/\mu}{1-\lambda/\mu} = \frac{\lambda}{\mu-\lambda} = N$$

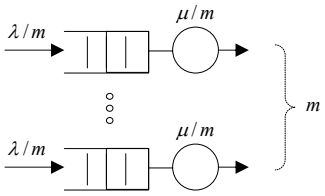
$$T' = \frac{N'}{\lambda/m} = \frac{m}{\mu-\lambda} = mT$$

$$W' = \frac{\rho'}{\mu/m-\lambda/m} = \frac{m(\lambda/\mu)}{\mu-\lambda} = mW$$

- M/M/1 system: slow down the arrival and service rates by the same factor m
- Utilization factors are the same ⇒stationary distributions the same, average number in the system the same
- Delay in the slower system is m times higher
- Average number in queue is the same, but in the 1st system the customers move out faster

Example-II: Statistical MUX-ing vs. TDM





$$T' = \frac{m}{\mu-\lambda} = mT$$

- m identical Poisson streams with rate λ/m; link with capacity 1; packet lengths iid, exponential with mean 1/μ
- Alternative: split the link to m channels with capacity 1/m each, and dedicate one channel to each traffic stream
- Delay in each "queue" becomes m times higher
- Statistical multiplexing vs. TDM or FDM
- When is TDM or FDM preferred over statistical multiplexing?

"PASTA" Theorem

- Markov chain: "stationary" or "in steady-state:"
  - Process started at the stationary distribution, or
  - Process runs for an infinite time $t \to \infty$
- Probability that at any time t, process is in state i is equal to the stationary probability

$$p_i = \lim_{t\to\infty} P\{N(t) = i\} = \lim_{t\to\infty} \frac{T_i(t)}{t}$$

- Question: For an M/M/1 queue: given t is an arrival time, what is the probability that $N(t)=i$?
- Answer: Poisson Arrivals See Time Averages!

PASTA Theorem

- Steady-state probabilities:

$$p_n = \lim_{t\to\infty} P\{N(t) = n\}$$

- Steady-state probabilities upon arrival:

$$a_n = \lim_{t\to\infty} P\{N(t^-) = n \mid \text{arrival at } t\}$$

- Lack of Anticipation Assumption (LAA): Future inter-arrival times and service times of previously arrived customers are independent

- Theorem: In a queueing system satisfying LAA:
  1. If the arrival process is Poisson:

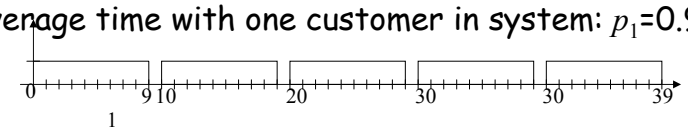
$$a_n = p_n, \quad n = 0, 1, \ldots$$

  2. Poisson is the only process with this property (necessary and sufficient condition)

PASTA Theorem

Doesn't PASTA apply for all arrival processes?

- Deterministic arrivals every 10 sec
- Deterministic service times 9 sec
- Upon arrival: system is always empty a_1=0
- Average time with one customer in system: p_1=0.9



- "Customer" averages need not be time averages
- Randomization does not help, unless Poisson!

PASTA Theorem: Proof

- Define A(t,t+δ), the event that an arrival occurs in [t, t+ δ)
- Given that a customer arrives at t, probability of finding the system in state n:

$$P\{N(t^-)=n \mid \text{arrival at t}\} = \lim_{\delta\to 0} P\{N(t^-)=n \mid A(t,t+\delta)\}$$

- A(t,t+δ) is independent of the state before time t, N(t$^-$)
  - N(t$^-$) determined by arrival times <t, and corresponding service times
  - A(t,t+δ) independent of arrivals <t [Poisson]
  - A(t,t+δ) independent of service times of customers arrived <t [LAA]

$$a_n(t) = \lim_{\delta\to 0} P\{N(t^-)=n \mid A(t,t+\delta)\} = \lim_{\delta\to 0} \frac{P\{N(t^-)=n, A(t,t+\delta)\}}{P\{A(t,t+\delta)\}}$$

$$= \lim_{\delta\to 0} \frac{P\{N(t^-)=n\}P\{A(t,t+\delta)\}}{P\{A(t,t+\delta)\}} = P\{N(t^-)=n\}$$

$$a_n = \lim_{t\to\infty} a_n(t) = \lim_{t\to\infty} P\{N(t^-)=n\} = p_n$$

PASTA Theorem: Intuitive Proof

- t_a and t_r: randomly selected arrival and observation times, respectively
- The arrival processes prior to t_a and t_r respectively are *stochastically* identical
  - The probability distributions of the time to the first arrival before t_a and t_r are *both* exponentially distributed with parameter λ
  - Extending this to the 2nd, 3rd, etc. arrivals before t_a and t_r establishes the result
- State of the system at a given time t depends *only* on the arrivals (and associated service times) before t
- Since the arrival processes before arrival times and random times are identical, so is the state of the system they see

Arrivals that Do not See Time-Averages

Example 1: Non-Poisson arrivals

- IID inter-arrival times, uniformly distributed between in 2 and 4 sec
- Service times deterministic 1 sec
- Upon arrival: system is always empty
- λ=1/3, T=1 → N=T/λ=1/3 → p_1=1/3

Example 2: LAA violated

- Poisson arrivals
- Service time of customer i: $S_i = \alpha T_{i+1}$, $\alpha < 1$
- Upon arrival: system is always empty
- Average time the system has 1 customer: $p_1 = \alpha$

Distribution after Departure

- Steady-state probabilities after departure:

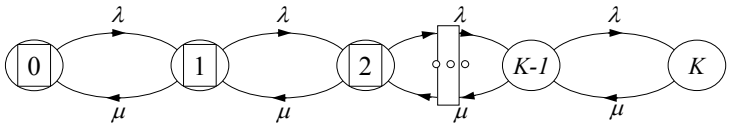
$$d_n = \lim_{t\to\infty} P\{X(t^+) = n \mid \text{departure at } t\}$$

- Under very general assumptions:
  - N(t) changes in unit increments
  - limits a_n and exist d_n
- $a_n = d_n$, $n=0,1,\ldots$
- In steady-state, system appears stochastically identical to an arriving and departing customer
- Poisson arrivals + LAA: an arriving and a departing customer see a system that is stochastically to the one seen by an observer looking at an arbitrary time

M/M/* Queues

- Poisson arrival process
  - Interarrival times: iid, exponential
- Service times: iid, exponential
- Service times and interarrival times: independent
- $N(t)$: Number of customers in system at time t (state)
- $\{N(t): t \geq 0\}$ can be modeled as a continuous-time Markov chain
- Transition rates depend on the characteristics of the system
- PASTA Theorem always holds

M/M/1/K Queue



- M/M/1 with finite waiting room
  - At most K customers in the system
  - Customer that upon arrival finds K customers in system is dropped
- Stationary distribution

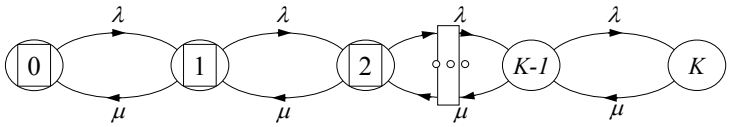
$$p_n = \rho^n p_0, n = 1,2,...,K$$

$$p_0 = \frac{1-\rho}{1-\rho^{K+1}}$$

- Stability condition: always stable – even if $\rho \geq 1$
- Probability of loss

$$P\{\text{loss}\} = P\{N(t) = K\} = \frac{\rho^K(1-\rho)}{1-\rho^{K+1}}$$

M/M/1/K Queue (proof)



- Exactly as in the M/M/1 queue:

$$p_n = \rho^n p_0, \ n = 1,2,...,K$$

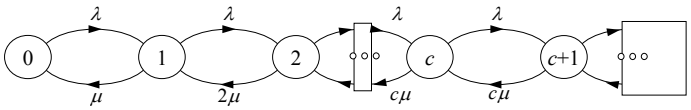
- Normalization constant:

$$\sum_{n=0}^{K} p_n = 1 \Rightarrow p_0 \sum_{n=1}^{K} \rho^n = 1 \Rightarrow p_0 \frac{1-\rho^{K+1}}{1-\rho} = 1$$

$$\Rightarrow p_0 = \frac{1-\rho}{1-\rho^{K+1}}$$

- *Generalize: Truncating a Markov chain*

M/M/c Queue



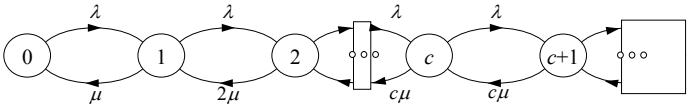
- Poisson arrivals with rate λ
- Exponential service times with parameter μ
- c servers
- Arriving customer finds n customers in system
  - $n < c$: it is routed to any idle server
  - $n \geq c$: it joins the waiting queue – all servers are busy

➧ Birth-death process with state-dependent death rates

$$\mu_n = \begin{cases} n\mu, & 1 \leq n \leq c \\ c\mu, & n \geq c \end{cases}$$

[Time spent at state n before jumping to n -1 is the minimum of $B_n = \min\{n,c\}$ exponentials with parameter μ]

M/M/c Queue



- Detailed balance equations

$$1 \leq n \leq c: \quad p_n = \frac{\lambda}{n\mu} p_{n-1} = \cdots = \frac{\lambda}{n\mu}\frac{\lambda}{(n-1)\mu}\cdots\frac{\lambda}{\mu} p_0 = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0 = \frac{(c\rho)^n}{n!} p_0, \qquad \rho \equiv \frac{\lambda}{c\mu}$$

$$n > c: \quad p_n = \left(\frac{\lambda}{c\mu}\right)^{n-c} p_c = \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \left(\frac{\lambda}{c\mu}\right)^{n-c} p_0 = \frac{c^c}{c!}\left(\frac{\lambda}{c\mu}\right)^n p_0 = \frac{c^c \rho^n}{c!} p_0$$

- Normalizing

$$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow p_0 = \left[1 + \sum_{k=1}^{c-1}\frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!}\sum_{k=c}^{\infty}\rho^{k-c}\right]^{-1} = \left[\sum_{k=0}^{c-1}\frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!}\frac{1}{1-\rho}\right]^{-1}$$

M/M/c Queue

- Probability of queueing – arriving customer finds all servers busy

$$P_Q = P\{\text{queueing}\} = \sum_{n=c}^{\infty} p_n = p_0 \frac{(c\rho)^c}{c!} \sum_{n=c}^{\infty} \rho^{n-c} = \frac{(c\rho)^c}{c!} \frac{1}{1-\rho} p_0$$

- Erlang-C Formula: used in telephony and circuit-switching
  - Call requests arrive with rate λ; holding time of a call exponential with mean 1/μ
  - c available circuits on a transmission line
  - A call that finds all c circuits busy, continuously attempts to find a free circuit – "remains in queue"
- M/M/c/c Queue: c-server loss system
  - A call that finds all c circuits busy is blocked
  - Erlang-B Formula: popular in telephony

M/M/c Queue

- Expected number of customers waiting in queue – not in service

$$N_Q = \sum_{n=c}^{\infty} (n-c) p_n = p_0 \frac{(c\rho)^c}{c!} \sum_{n=c}^{\infty} (n-c)\rho^{n-c} = p_0 \frac{(c\rho)^c}{c!} \frac{\rho}{(1-\rho)^2}$$

$$= P_Q (1-\rho) \frac{\rho}{(1-\rho)^2} = P_Q \frac{\rho}{1-\rho}$$

- Average waiting time per customer (in queue)

$$W = \frac{N_Q}{\lambda} = P_Q \frac{\rho}{\lambda(1-\rho)}$$

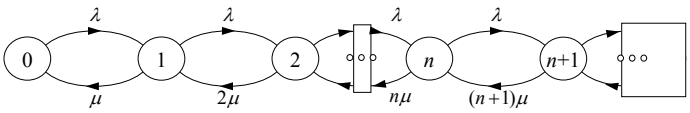
- Average time in system per customer (queued + serviced)

$$T = W + \frac{1}{\mu} = P_Q \frac{\rho}{\lambda(1-\rho)} + \frac{1}{\mu}$$

- Expected number of customers in system

$$N = \lambda T = P_Q \frac{\rho}{(1-\rho)} + c\rho$$

M/M/∞ Queue: Infinite-Server System



- Infinite number of servers – no queueing
- Stationary distribution:

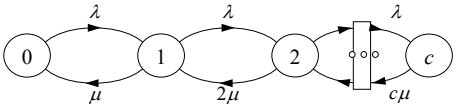
$$p_n = \frac{(\lambda/\mu)^n}{n!} e^{-\lambda/\mu}, \quad n = 0,1,...$$

  Poisson with rate λ/μ
- Average number of customers & average delay:

$$N = \frac{\lambda}{\mu}, \qquad T = \frac{N}{\lambda} = \frac{1}{\mu}$$

- *The results hold for an M/G/∞ queue*

M/M/c/c Queue: c-Server Loss System



- c servers, no waiting room
- An arriving customer that finds all servers busy is blocked
- Stationary distribution:

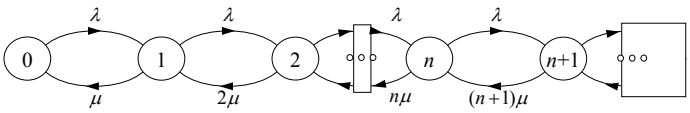
$$p_n = \frac{(\lambda/\mu)^n}{n!}\left[\sum_{k=0}^{c}\frac{(\lambda/\mu)^k}{k!}\right]^{-1}, \quad n = 0,1,...,c$$

- Probability of blocking (using PASTA):

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

- Erlang-B Formula: used in telephony and circuit-switching
- *Results hold for an M/G/c/c queue*

M/M/∞ and M/M/c/c Queues (proof)



- DBE:

$$(n\mu)p_n = \lambda p_{n-1} \Rightarrow p_n = \frac{\lambda}{n\mu}p_{n-1} = \frac{\lambda}{n\mu}\frac{\lambda}{(n-1)\mu}p_{n-2} = \cdots = \frac{\lambda\cdot\lambda\cdots\lambda}{n\mu\cdot(n-1)\mu\cdots\mu}p_0$$

$$\Rightarrow p_n = \frac{(\lambda/\mu)^n}{n!}p_0, \quad n = 0,1,...$$

- Normalizing:

$$p_0 = \left[\sum_{k=0}^{c}\frac{(\lambda/\mu)^k}{k!}\right]^{-1}, \quad \text{for M/M/c/c}$$

$$p_0 = \left[\sum_{k=0}^{\infty}\frac{(\lambda/\mu)^k}{k!}\right]^{-1} = e^{-\lambda/\mu}, \quad \text{for M/M/c/c}$$

Multidimensional Markov Chains

Theorem 8:

- $\{X_1(t)\}$, $\{X_2(t)\}$: independent Markov chains
- $\{X_i(t)\}$: reversible
- $\{X(t)\}$, with $X(t)=(X_1(t), X_2(t))$: vector-valued stochastic process
- ➧ $\{X(t)\}$ is a Markov chain
- ➧ $\{X(t)\}$ is reversible

Multidimensional Chains:

- Queueing system with two classes of customers, each having its own stochastic properties – track the number of customers from each class
- Study the "joint" evolution of two queueing systems – track the number of customers in each system

Example: Two Independent M/M/1 Queues

- Two independent M/M/1 queues. The arrival and service rates at queue i are λ_i and μ_i respectively. Assume $\rho_i = \lambda_i/\mu_i < 1$.
- $\{(N_1(t), N_2(t))\}$ is a Markov chain.
- Probability of n_1 customers at queue 1, and n_2 at queue 2, at steady-state

$$p(n_1,n_2) = (1-\rho_1)\rho_1^{n_1} \cdot (1-\rho_2)\rho_2^{n_2} = p_1(n_1) \cdot p_2(n_2)$$

- "Product-form" distribution
- Generalizes for any number K of independent queues, M/M/1, M/M/c, or M/M/∞. If $p_i(n_i)$ is the stationary distribution of queue i :

$$p(n_1,n_2,\ldots,n_K) = p_1(n_1)p_2(n_2)\ldots p_K(n_K)$$

Example: Two Independent M/M/1 Queues

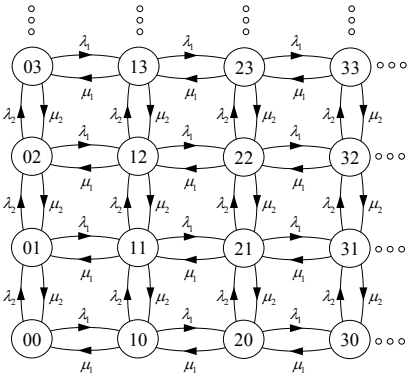
- Stationary distribution:

$$p(n_1,n_2) = \left(1-\frac{\lambda_1}{\mu_1}\right)\left(\frac{\lambda_1}{\mu_1}\right)^{n_1}\left(1-\frac{\lambda_2}{\mu_2}\right)\left(\frac{\lambda_2}{\mu_2}\right)^{n_2}$$

- Detailed Balance Equations:

$$\mu_1 p(n_1+1,n_2) = \lambda_1 p(n_1,n_2)$$
$$\mu_2 p(n_1,n_2+1) = \lambda_2 p(n_1,n_2)$$

- Verify that the Markov chain is reversible – Kolmogorov criterion



Example: Two Queues with Joint Buffer

- The two independent M/M/1 queues of the previous example share a common buffer of size B – arrival that finds B customers *waiting* is blocked
- State space restricted to

$$E = \{(n_1, n_2) : (n_1 - 1)^+ + (n_2 - 1)^+ \le B\}$$

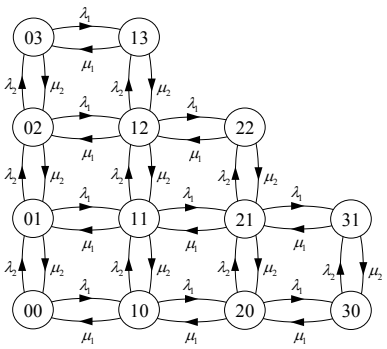
- Distribution of truncated chain:

$$p(n_1, n_2) = p(0,0) \cdot \rho_1^{n_1} \rho_2^{n_2}, \ (n_1, n_2) \in E$$

- Normalizing:

$$p(0,0) = \left[\sum_{(n_1, n_2) \in E} \rho_1^{n_1} \rho_2^{n_2} \right]^{-1}$$

- Theorem specifies joint distribution up to the normalization constant
- Calculation of normalization constant is often tedious



State diagram for B =2

M/G/1 Queues

$$\rho = \left(\frac{\lambda}{\mu} \right) \qquad W_q = \frac{N_q}{\lambda}$$

$$P_0 = \left(1 - \frac{\lambda}{\mu} \right) \qquad W = \frac{N}{\lambda}$$

$$N_q = \left(\frac{\lambda^2 \sigma s^2 + \rho^2}{2(1-\rho)} \right)$$

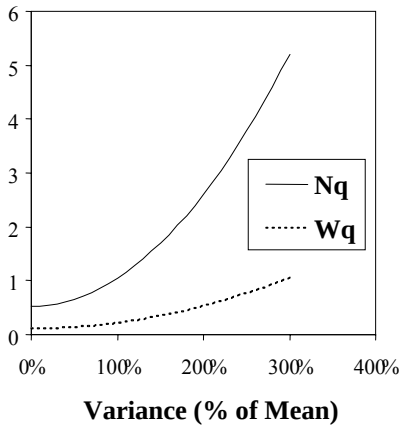
$$N = N_q + N_s = N_q + \rho$$

ρ is the average number of users in the server!

Observations:

- M/G/1 queue with < 1 may have infinite W if the second moments $\overline{X}^2 = \infty$
  - Small fraction of users have incredibly long service time (e.g. WWW traffic)
- P-K formula is valid for any order of servicing customers as long as the order id independent of the required service time

Service Variance and Wait Time



- Note in the previous two examples, that the mean service time was the same. Only the variance changed.
- Wait time is a function of the mean and variance of the service process.
- All other things being equal, the greater the service variance, the larger the wait time.

Ref: www.dal.ca/~jblake

An Example: Secretary Hiring

Suppose you must hire a secretary and you have to select one of two candidates.

Secretary 1 is very consistent, typing any document in exactly 15 minutes.

Secretary 2 is somewhat faster, with an average of 14 minutes per document, but with times varying according to the exponential distribution.

The workload in the office is 3 documents per hour, with interarrival times varying according to the exponential distribution. Which secretary will give you shorter turnaround times on documents?

Ref: faculty.cox.smu.edu/~cchamber/itom6203/course_files/Session5.PPT

Secretary Hiring - Queuing Model

Secretary 1:

λ = 3 doc/hr μ = 4 docs/hr $\rho = \frac{3}{4} = .75$

$N_q = \rho^2 / 2 (1 - \rho) = 2.8125$ $N = N_q + \rho = 3.5625$

$W_q = N_q / \lambda = 0.9375$ hr

$W = W_q + 1/\mu = 1.1875$ hrs = 75 mins

Secretary 2:

λ = 3 doc/hr μ = 4.29 docs/hr ρ = 3/4.29 = .70

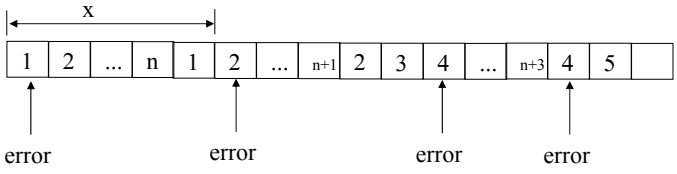
$N_q = \{\lambda^2 \sigma^{2+} \rho^2\} / 2 (1 - \rho) = 1.633$ $N = N_q + \rho = 2.333$

$W_q = N_q / \lambda = 0.7777$ hr

$W = W_q + 1/\mu = 1.01$ hrs = 60.65 mins

Example: Delay analysis of an ARQ system

- Go back n ARQ
  - packet transmission in frames of one time unit
  - Maximum waiting time for acknowledgment is n -1 frames
  - Retransmission is due to receiver rejection
    - Transmit packets in frame i+1, i+2, ..., i+n-1
    - Then, go back to transmit the given packet in frame i+n



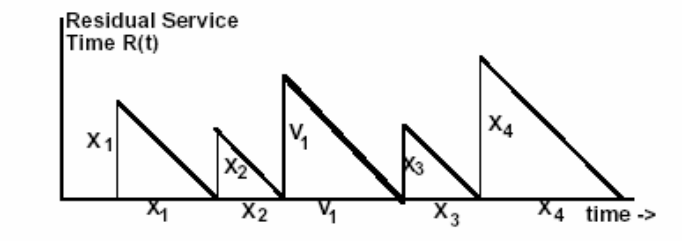
  - Transmitter queue behaves as M/G/1

Delay analysis of an ARQ system (Cont.)

- What is the service time (x) distribution?
  - X= 1 + kn with probability $(1-p)p^k$
  - p *is error probability*
- Using Pollaczek-Khintchine formula, N_q, W_q, W and N can be found

M/G/1 with vacations

- Useful for polling and reservation systems
  - (e.g., token rings) sending various kinds of control and record-keeping packets
- When the queue is empty, the server takes a vacation
- Vacation times are IID and independent of service times and arrival times
  - If system is empty after a vacation, the server takes another vacation
  - The only impact on the analysis is that a packet arriving to an empty
  - system must wait for the end of the vacation



- It can be proved that the queuing waiting time is

$$W_q = \frac{\lambda \overline{X}^2}{2(1-\rho)} + \frac{\overline{V}^2}{2\overline{V}}$$

- Mutual independence of the vacation intervals is not required

Example: Slotted Frequency- and Time-Division Multiplexing

- FDM
  - m Poisson traffic streams each with λ / m arrival rate
  - Each packet is *m* time units
  - Therefore, M/D/1 queueing system
  - Using P-K formula,

$$W_{FDM} = \frac{\lambda m}{2(1-\lambda)}$$

- SFDM
  - Consider slotted scheme where transmissions can start only at times *m,2m,3m, ….*
  - *M/D/1 with vacations represent this queueing system, then*

$$W_{SFDM} = \frac{\lambda m}{2(1-\lambda)} + \frac{m}{2}$$

Example: Slotted Frequency- and Time-Division Multiplexing

- TDM
  - Very similar to SFDM
- Look now at the total delay time
  - $W = W_q + W_s$
- Therefore,

$$T_{FDM} = W_{FDM} + m$$

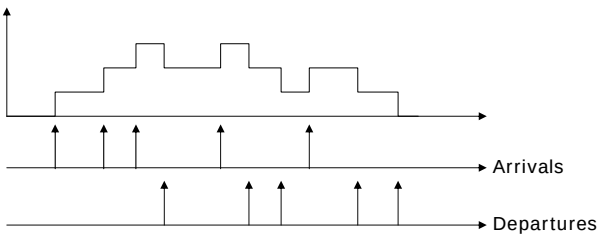
$$T_{SFDM} = W_{SFDM} + m = T_{FDM} + \frac{m}{2}$$

$$T_{TDM} = W_{TDM} + 1 = T_{FDM} - \left(\frac{m}{2} - 1\right)$$

  - $T_{TDM} < T_{FDM}$ Longer waiting time compensated by faster service time

Burke's Theorem

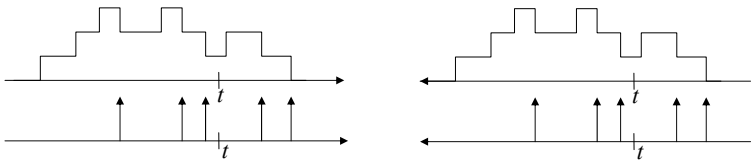
- {X(t)} birth-death process with stationary distribution $\{p_j\}$
- Arrival epochs: points of increase for {X(t)}
  Departure epoch: points of increase for {X(t)}
- {X(t)} completely determines the corresponding arrival and departure processes



Burke's Theorem

- Poisson arrival process: $\lambda_j=\lambda$, for all j
  - Birth-death process called a (λ, μ_j)-process
  - Examples: M/M/1, M/M/c, M/M/∞ queues
- Poisson arrivals → LAA:
  For any time t, future arrivals are independent of {X(s): s≤t}
- (λ, μ_j)-process at steady state is reversible: forward and reversed chains are stochastically identical
- ➧ Arrival processes of the forward and reversed chains are stochastically identical
- Arrival process of the reversed chain is Poisson with rate λ
- The arrival epochs of the reversed chain are the departure epochs of the forward chain
- ➧ Departure process of the forward chain is Poisson with rate λ

Burke's Theorem

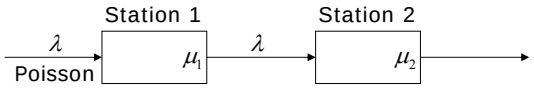


- Reversed chain: arrivals after time t are independent of the chain history up to time t (LAA)
- Forward chain: departures prior to time t and future of the chain {X(s): s≥t} are independent

Burke's Theorem

- Theorem: Consider an M/M/1, M/M/c, or M/M/∞ system with arrival rate λ. Suppose that the system starts at steady-state. Then:
  1. The departure process is Poisson with rate λ
  2. At each time *t*, the number of customers in the system is independent of the departure times prior to *t*
- Fundamental result for study of networks of M/M/* queues, where output process from one queue is the input process of another

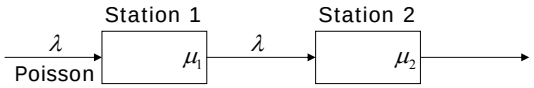
Single-Server Queues in Tandem



- Customers arrive at queue 1 according to Poisson process with rate λ.
- Service times exponential with mean $1/\mu_i$. Assume service times of a customer in the two queues are independent.
- Assume $\rho_i=\lambda/\mu_i<1$
- What is the joint *stationary* distribution of N_1 and N_2 – number of customers in each queue?
- Result: in *steady state* the queues are independent and

$$p(n_1,n_2)=(1-\rho_1)\rho_1^{n_1}\cdot(1-\rho_2)\rho_2^{n_2}=p_1(n_1)\cdot p_2(n_2)$$

Single-Server Queues in Tandem



- Q1 is a M/M/1 queue. At steady state its departure process is Poisson with rate λ. Thus Q2 is also M/M/1.
- Marginal stationary distributions:

$$p_1(n_1) = (1-\rho_1)\rho_1^{n_1}, \quad n_1 = 0,1,... \qquad p_2(n_2) = (1-\rho_2)\rho_2^{n_2}, \quad n_2 = 0,1,...$$

- To complete the proof: establish independence at steady state
- Q1 at steady state: at time t, $N_1(t)$ is independent of departures prior to t, which are arrivals at Q2 up to t. Thus $N_1(t)$ and $N_2(t)$ independent:

$$P\{N_1(t) = n_1, N_2(t) = n_2\} = P\{N_1(t) = n_1\}P\{N_2(t) = n_2\} = p_1(n_1) \cdot P\{N_2(t) = n_2\}$$

- Letting $t \to \infty$, the joint stationary distribution

$$p(n_1, n_2) = p_1(n_1) \cdot p_2(n_2) = (1-\rho_1)\rho_1^{n_1} \cdot (1-\rho_2)\rho_2^{n_2}$$

Queues in Tandem

- Theorem: Network consisting of *K* single-server queues in tandem. Service times at queue *i* exponential with rate μ_i, independent of service times at any queue $j \neq i$. Arrivals at the first queue are Poisson with rate λ. The stationary distribution of the network is:

$$p(n_1,\ldots,n_K) = \prod_{i=1}^{K}(1-\rho_i)\rho_i^{n_i}, \quad n_i = 0,1,\ldots; i = 1,\ldots,K$$

- At *steady state* the queues are independent; the distribution of queue *i* is that of an isolated M/M/1 queue with arrival and service rates λ and μ_i

$$p_i(n_i) = (1-\rho_i)\rho_i^{n_i}, \quad n_i = 0,1,...$$

- Are the queues independent if not in steady state? Are stochastic processes $\{N_1(t)\}$ and $\{N_2(t)\}$ independent?