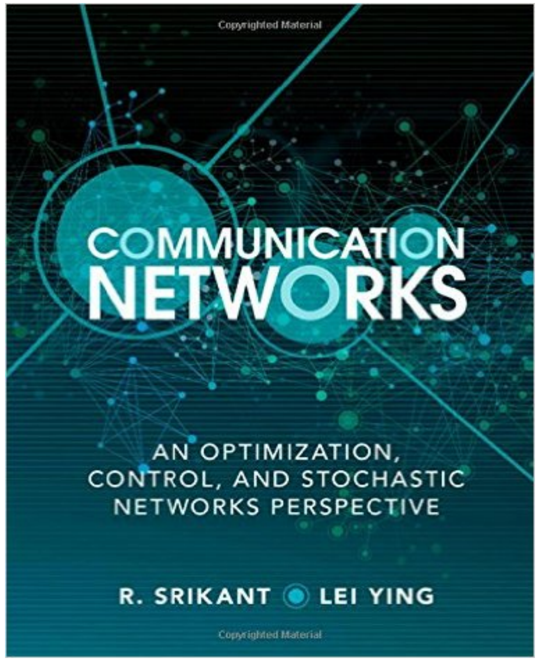
Statistical Multiplexing and Queues

CMPS 4750/6750: Computer Networks

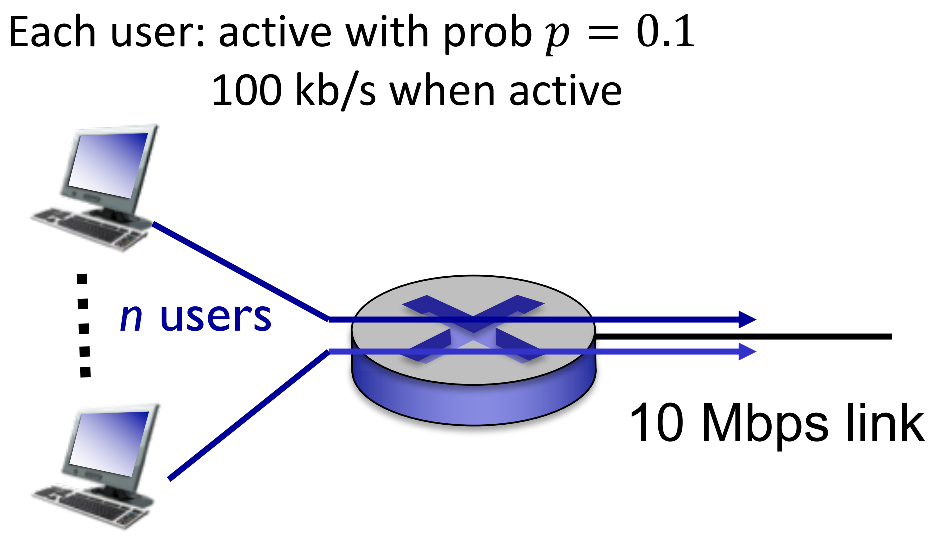
Outline

- The Chernoff bound (3.1)
- Statistical multiplexing (3.2)
- Discrete-time Markov chains (3.3)
- Geo/Geo/1 queue (3.4)
- Little's law (3.4)



Statistical multiplexing

- Example:
  - 10 Mb/s link
  - each user:
    - active with a probability 0.1
    - 100 kb/s when "active"

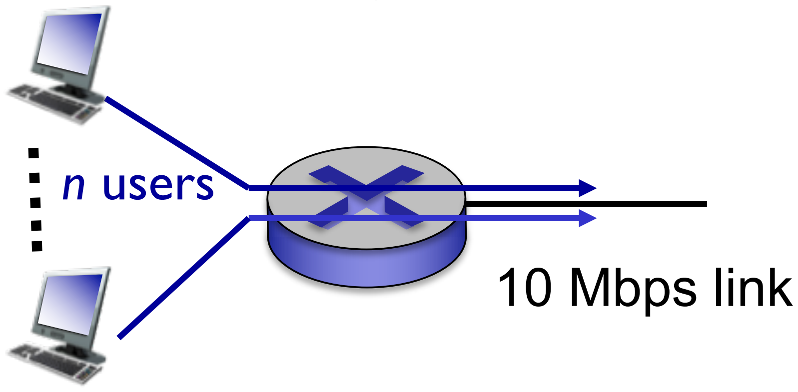


- *How many users can be supported?*
  - assume that there is no output queue
  - 1. allocation according to peak rate (e.g., circuit switching): 10Mbps/100kpbs = 100
  - 2. statistical multiplexing: allow $n \geq 100$ users to share the link
    - What is the overflow probability? Pr(at least 101 users become active simultaneously)

Statistical multiplexing

- Allow $n > 100$ users to share the link
  - For each user i, let $X_i = 1$ if user i is active, $X_i = 0$ otherwise
  - Assume X_i's are *i.i.d.*, $X_i \sim \text{Bernoulli}(0.1)$
  - Overflow probability:

$$\Pr\left(\sum_{i=1}^{n} X_i \geq 101\right) = \sum_{k=101}^{n} \binom{n}{k} 0.1^k (1-0.1)^{n-k}$$



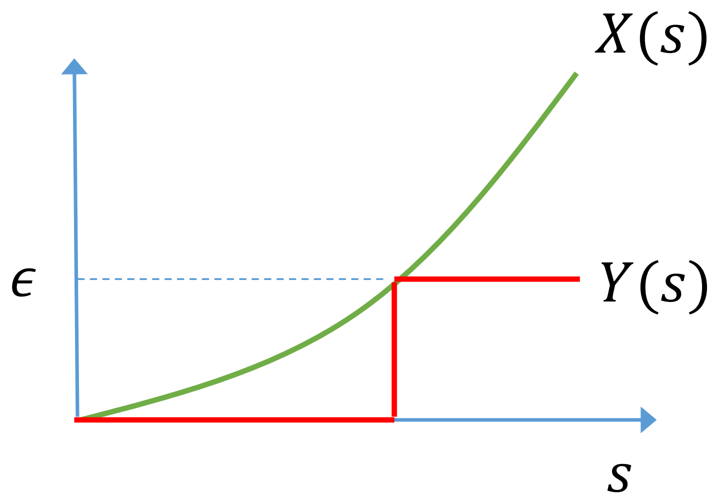
Markov's inequality

Lemma 3.1.1 (Markov's inequality) For a positive $\mathrm{r.v.}\, X$, the following inequality holds for any $\epsilon > 0$:

$$\Pr(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}$$

Proof Define a r.v. Y such that $Y = \epsilon$ if $X \geq \epsilon$ and $Y = 0$ otherwise. So

$$E(X) \geq E(Y)$$
$$= \epsilon \Pr(Y = \epsilon)$$
$$= \epsilon \Pr(X \geq \epsilon)$$



The Chernoff bound

Theorem 3.1.2 (the Chernoff bound) Consider a sequence of independently and identically distributed (*i.i.d.*) random variables $\{X_i\}$. For any constant x, the following inequality holds:

$$\Pr\left(\sum_{i=1}^{n} X_i \geq nx\right) \leq e^{-n \sup\limits_{\theta \geq 0}\{\theta x - \log M(\theta)\}}$$

where $M(\theta) = E(e^{\theta X_1})$ is the moment generation function of X_1

If $X_i \sim$ Bernoulli(p), and $p \leq x \leq 1$, then

$$\Pr\left(\sum_{i=1}^{n} X_i \geq nx\right) \leq e^{-nD(x\|p)}$$

where $D(x \parallel p) = x \log\frac{x}{p} + (1-x)\log\frac{1-x}{1-p}$ (Kullback-Leibler divergence between Bernoulli r.v.s)

Proving the Chernoff bound

$$\Pr\left(\sum_{i=1}^{n} X_i \geq nx\right) \leq \Pr\left(e^{\theta \sum_{i=1}^{n} X_i} \geq e^{\theta nx}\right) \qquad \forall\theta \geq 0$$

Markov inequality $\quad \leq \dfrac{E\left[e^{\theta \sum_{i=1}^{n} X_i}\right]}{e^{\theta nx}}$

$$= \frac{E\left[\prod_{i=1}^{n} e^{\theta X_i}\right]}{e^{\theta nx}}$$

Independent dist. $\quad = \dfrac{\prod_{i=1}^{n} E\left(e^{\theta X_i}\right)}{e^{\theta nx}}$

Identical dist. $\quad = \dfrac{[M(\theta)]^n}{e^{\theta nx}}$

$$= e^{-n(\theta x - \log M(\theta))}$$

$$\forall\theta \geq 0, \Pr\left(\sum_{i=1}^{n} X_i \geq nx\right) \leq e^{-n(\theta x - \log M(\theta))}$$

$$\Rightarrow \Pr\left(\sum_{i=1}^{n} X_i \geq nx\right) \leq \inf_{\theta \geq 0} e^{-n(\theta x - \log M(\theta))}$$

$$= e^{-n \sup_{\theta \geq 0}\{\theta x - \log M(\theta)\}}$$

Proving the Chernoff bound (Bernoulli case)

If $X_i \sim$ Bernoulli(p) $\forall i$, and $p \leq x \leq 1$, then

$$\sup_{\theta \geq 0}\{\theta x - \log M(\theta)\} = x \log \frac{x}{p} + (1-x)\log\frac{1-x}{1-p}$$

Proof Since $X_1 \sim$ Bernoulli(p), $M(\theta) = E\left(e^{\theta X_1}\right) = pe^{\theta} + (1-p)$

Let $f(\theta) = \theta x - \log M(\theta) = \theta x - \log\left(pe^{\theta} + (1-p)\right)$

$$f'(\theta) = x - \frac{pe^{\theta}}{pe^{\theta}+(1-p)}, \quad f'(\theta) = 0 \Rightarrow e^{\theta} = \frac{x}{1-x}\frac{1-p}{p} \quad (\geq 1 \text{ since } x \geq p)$$

$$\Rightarrow \sup_{\theta \geq 0} f(\theta) = x\left(\log\frac{x}{1-x} + \log\frac{1-p}{p}\right) - \log\left(\frac{x}{1-x}(1-p) + 1 - p\right)$$

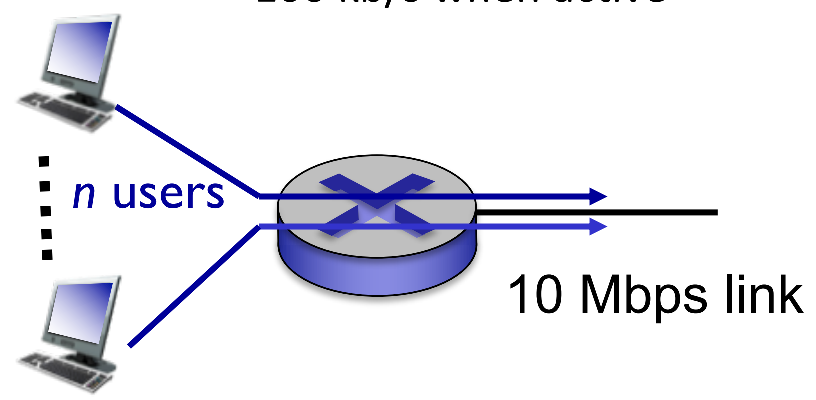
$$= x \log\frac{x}{p} + (1-x)\log\frac{1-x}{1-p}$$

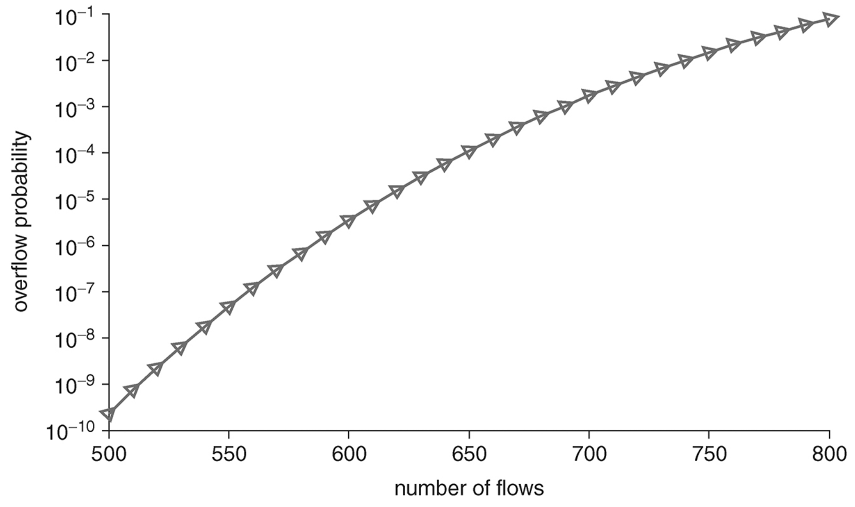
Statistical multiplexing

- Allow $n > 100$ users to share the link
  - For each user i, let $X_i = 1$ if user i is active, $X_i = 0$ otherwise
  - Assume X_i's are *i.i.d.*, $X_i \sim$ Bernoulli(0.1)
  - Overflow probability
    - $\Pr(\sum_{i=1}^{n} X_i \geq 101) = \sum_{k=101}^{n} \binom{n}{k} 0.1^k (1-0.1)^{n-k}$
    - Using the Chernoff bound:

$$\Pr\left(\sum_{i=1}^{n} X_i \geq 101\right) = \Pr\left(\sum_{i=1}^{n} X_i \geq n\frac{101}{n}\right)$$

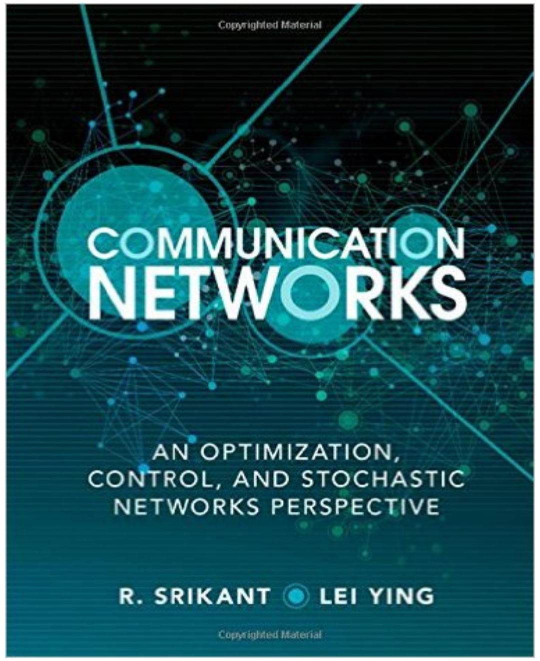
$$\leq e^{-nD\left(\frac{101}{n} \| 0.1\right)}$$





Outline

- The Chernoff bound (3.1)
- Statistical multiplexing (3.2)
- Discrete-time Markov chains (3.3)
- Geo/Geo/1 queue (3.4)
- Little's law (3.4)



Discrete-time stochastic processes

- Let $\{X_k, k \in \mathbb{N}\}$ be discrete-time stochastic process with a countable state space
  - For each $k \in \mathbb{N}$, X_k is a random variable
  - X_k is considered as the state of the process in time-slot k
  - X_k takes on values in a countable set S
  - Any realization of $\{X_k\}$ is called a sample path
- E.g., Let $\{X_k, k \in \mathbb{N}\}$ be an *i.i.d.* Bernoulli process with parameter p
  - $X_k \sim \text{Bernoulli}(p)$, *i.i.d.* over k

Discrete-time Markov chains

- Let $\{X_k, k \in \mathbb{N}\}$ be a discrete-time stochastic process with a countable state space. $\{X_k\}$ is called a Discrete-Time Markov Chain (DTMC) if

$$\Pr\left(X_{k+1} = j \mid X_k = i, X_{k-1} = i_{k-1,\ldots,} X_0 = i_0\right) = \Pr(X_{k+1} = j \mid X_k = i) \quad \text{(Markovian Property)}$$

$$= \mathrm{P}_{ij} \quad \text{(“time homogeneous”)}$$

  – P_{ij}: the probability of moving to state j on the next transition, given that the current state is i

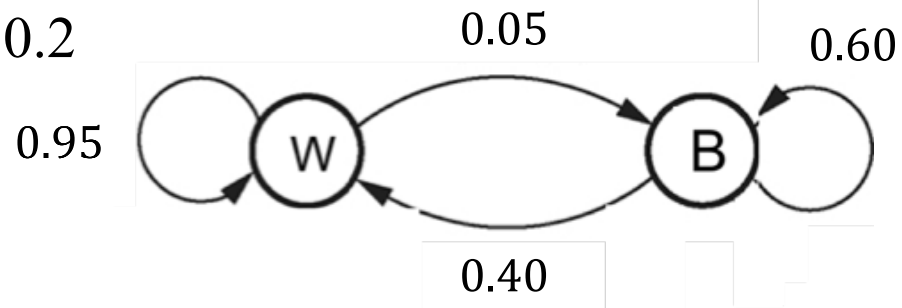
Transition probability matrix

- Transition probability matrix of a DTMC
  - a matrix $\mathbf{P}$ whose (i, j)-th element is P_{ij}
  - $\sum_j P_{ij} = 1, \forall i$ (each row of $\mathbf{P}$ summing to 1)
  - Ex: for an *i.i.d.* Bernoulli process with parameter p, $\mathbf{P} = \begin{pmatrix} p & 1-p \\ p & 1-p \end{pmatrix}$

Discrete-time Markov chains

Repair facility problem: a machine is either working or is in the repair center, with the transition probability matrix:

$$\mathbf{P} = \begin{array}{cc} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} 0.95 & 0.05 \\ 0.40 & 0.60 \end{pmatrix} \end{array}$$



Assume $\Pr(X_0 = \text{“Working”}) = 0.8$, $\Pr(X_0 = \text{“Broken”}) = 0.2$

What is $\Pr(X_1 = \text{“Working”})$?

$$\Pr(X_1 = \text{“W”}) = \Pr(X_0 = \text{“W”} \cap X_1 = \text{“W”}) + \Pr(X_0 = \text{“B”} \cap X_1 = \text{“W”})$$

$$= \Pr(X_0 = \text{“W”}) \Pr(X_1 = \text{“W”} | X_0 = \text{“W”}) + \Pr(X_0 = \text{“B”}) \Pr(X_1 = \text{“W”} | X_0 = \text{“B”})$$

$$= \Pr(X_0 = \text{“W”}) P_{WW} + \Pr(X_0 = \text{“B”}) P_{BW}$$

$$= 0.8 \times 0.95 + 0.2 \times 0.4 = 0.84$$

Discrete-time Markov chains

In general, we have

- $\Pr(X_k = j) = \sum_i \Pr(X_{k-1} = i) P_{ij}$
- Let $p_j[k] = \Pr(X_k = j)$, $p[k] = (p_1[k], p_2[k], \ldots)$. Then

$$p[k] = p[k-1]\mathbf{P}$$

- A DTMC is completely captured by $p[0]$ and $\mathbf{P}$

n-step Transition Probabilities

Let $\mathbf{P}^n = \mathbf{P} \cdot \mathbf{P} \cdots \mathbf{P}$, multiplied n times. Let $P_{ij}^{(n)}$ denote $(\mathbf{P}^n)_{ij}$

Theorem $\Pr(X_n = j \mid X_0 = i) = P_{ij}^{(n)}$

Proof (by induction): $n = 1$, we have $\Pr(X_n = j \mid X_0 = i) = P_{ij} = P_{ij}^{(1)}$

Assume the result holds for any n, we have

$$\Pr(X_{n+1} = j \mid X_0 = i) = \sum_k \Pr(X_{n+1} = j,\ X_n = k \mid X_0 = i)$$

$$= \sum_k \Pr(X_{n+1} = j \mid X_n = k,\ X_0 = i) \Pr(X_n = k \mid X_0 = i)$$

$$= \sum_k P_{kj} P_{ik}^{(n)} = \sum_k P_{ik}^{(n)} P_{kj} = P_{ij}^{(n+1)}$$

Limiting distributions

- **Repair facility problem**: a machine is either working or is in the repair center, with the transition probability matrix:

$$\mathbf{P} = \begin{array}{c c} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} \end{array}$$

$0 < a < 1, 0 < \mathrm{b} < 1$

- *Q:* What fraction of time does the machine spend in the repair shop?

$$\mathbf{P}^n = \begin{pmatrix} \frac{b+a(1-a-b)^n}{a+b} & \frac{a-a(1-a-b)^n}{a+b} \\ \frac{b-b(1-a-b)^n}{a+b} & \frac{a+b(1-a-b)^n}{a+b} \end{pmatrix}$$

$$\lim_{n\to\infty} \mathbf{P}^n = \begin{pmatrix} \frac{b}{a+b} & \frac{a}{a+b} \\ \frac{b}{a+b} & \frac{a}{a+b} \end{pmatrix}$$

A probability distribution $\pi = (\pi_1, \pi_2, \ldots)$ is called a limiting distribution of the DTMS if

$$\pi_j = \lim_{n\to\infty} P_{ij}^{(n)} \quad \text{and} \quad \sum_j \pi_j = 1$$

Stationary distributions

- A probability distribution $\pi = (\pi_1, \pi_2, \dots)$ is said to be stationary for the DTMS if

$$\pi \cdot \mathbf{P} = \pi$$

  - $\pi \cdot \mathbf{P} = \pi \iff \sum_i \pi_i P_{ij} = \pi_j \ \forall j$
  - If $p[0] = \pi$, then $p[k] = \pi$ for all k

- **Theorem** If a DTMS has a limiting distribution π, then π is also a stationary distribution and there is no other stationary distribution

- *Q1*: under what conditions, does the limiting distribution exist?

- *Q2*: how to find a stationary distribution?

Irreducible Markov chains

- Ex: A Markov chain with two states a and b and the transition probability matrix given by:
$$\mathbf{P} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$
  - If the chain started in one state, it remained in the same state forever
  - $\lim_{n\to\infty} \mathbf{P}^n = \mathbf{P}$
  - $\pi \cdot \mathbf{P} = \pi$ for any distribution π (not unique)

- State j is said to be reachable from state i if there exists $n \geq 1$ so that $P_{ij}^{(n)} > 0$
- A Markov chain is said to be irreducible if any state i is reachable from any other state j

Aperiodic Markov chains

- Ex: A Markov chain with two states a and b and the transition probability matrix given by: $\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
  - $\pi \cdot \mathbf{P} = \pi \Rightarrow \pi = (0.5, 0.5)$
  - $\lim_{n\to\infty} P_{jj}^{(n)}$ does not exist for any j (a state is only visited every other time step.)
- Period of state i: $d_i = \gcd\{n > 0: P_{ii}^{(n)} > 0\}$
  - State i is said to be aperiodic if d_i=1
- A Markov chain is said to be aperiodic if all states are aperiodic
- **Theorem** Every state in an irreducible Markov chain has the same period.

Big Theorem

Consider a DTMC that is irreducible and aperiodic

- If the chain has a finite state-space, it always has a limiting distribution.
- There must be a positive vector π such that $\pi = \pi\mathbf{P}$ (an invariant measure)
  - If $\sum_i \pi_i = 1$, then π it is the unique stationary distribution and $\lim_{n\to\infty} P_{ij}^{(n)} = \pi_j$
  - If $\sum_i \pi_i = \infty$, a stationary distribution does not exist and $\lim_{n\to\infty} P_{ij}^{(n)} = 0$

How to find stationary distributions?

- Using the definition:

$$\pi_j = \sum_i \pi_i P_{ij} \ \forall j$$

$$\Leftrightarrow \ \pi_j = \sum_{i \neq j} \pi_i P_{ij} + \pi_j P_{jj} \quad \forall j$$

$$\Leftrightarrow \ \pi_j \left(1 - P_{jj}\right) = \sum_{i \neq j} \pi_i P_{ij} \quad \forall j$$

$$\Leftrightarrow \ \pi_j \sum_{i \neq j} P_{ji} = \sum_{i \neq j} \pi_i P_{ij} \quad \forall j$$

(global balance equations)

- Ex: given the transition matrix P of a DTMC, find its stationary distribution.

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ \frac{3}{4} & 0 & \frac{1}{4} \\ 1 & 0 & 0 \end{pmatrix}$$

$$\pi = \left(\frac{4}{9}, \ \frac{4}{9}, \ \frac{1}{9}\right)$$

How to find stationary distributions?

- Using the definition:

$$\pi_j = \sum_i \pi_i \, P_{ij} \;\; \forall j$$

$$\Leftrightarrow \;\; \pi_j = \sum_{i \neq j} \pi_i \, P_{ij} + \pi_j P_{jj} \quad \forall j$$

$$\Leftrightarrow \;\; \pi_j \left(1 - P_{jj}\right) = \sum_{i \neq j} \pi_i \, P_{ij} \quad \forall j$$

$$\Leftrightarrow \;\; \pi_j \sum_{i \neq j} P_{ji} = \sum_{i \neq j} \pi_i \, P_{ij} \quad \forall j$$

(global balance equations)

- Using the local balance equations:

$$\pi_j P_{ji} = \pi_i P_{ij} \quad \forall i, j$$

$$\Rightarrow \;\; \sum_i \pi_j P_{ji} = \sum_i \pi_i \, P_{ij} \;\; \forall j$$

$$\Rightarrow \;\; \pi_j \sum_{i \neq j} P_{ji} = \sum_{i \neq j} \pi_i \, P_{ij} \;\; \forall j$$

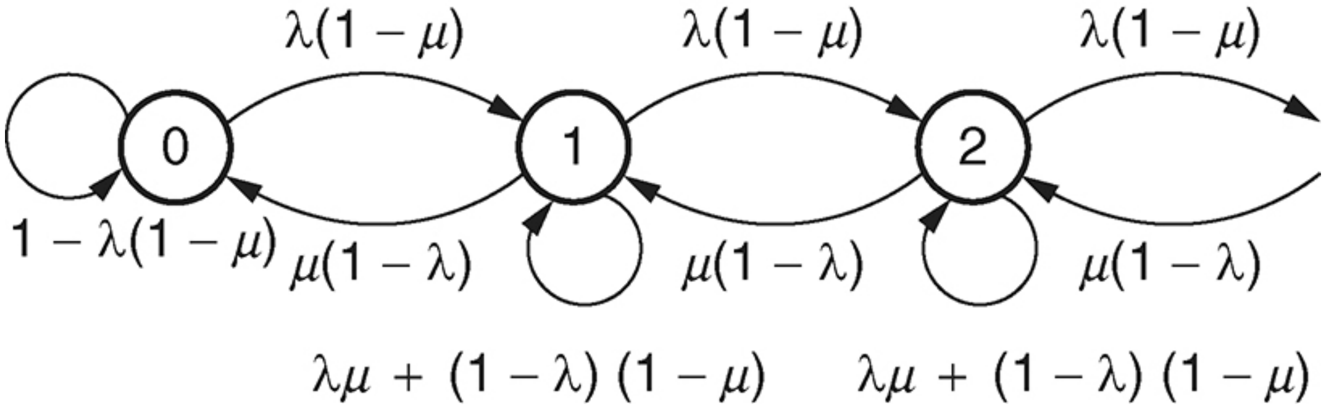
Geo/Geo/1 queue

buffer with infinite size

- A single server queue with infinite buffer size
- $a(k)$ - number of packets arrive in time-slot k
  - $a(k)$~Bernoulli(λ), *i.i.d.* over k $\Rightarrow$ inter-arrival time ~ Geometric (λ)
- $s(k)$ - number of packets served in time-slot k
  - $s(k)$~Bernoulli(μ), *i.i.d.* over k $\Rightarrow$ service time ~ Geometric (μ)
  - $s(k)$ and $a(k)$ are independent processes
- $q(k)$ - number of packets in the queue at the beginning of time-slot k (before packet arrivals occur)
- Queueing dynamics: $q(k+1) = [q(k) + a(k) - s(k)]^+$ $\qquad (x)^+ = \max(x, 0)$
  - Arrival occurs before any departure in each time-slot
  - $q(k)$ includes the packet that is being processed

Geo/Geo/1 queue

$q(k)$ is an infinite state Markov chain

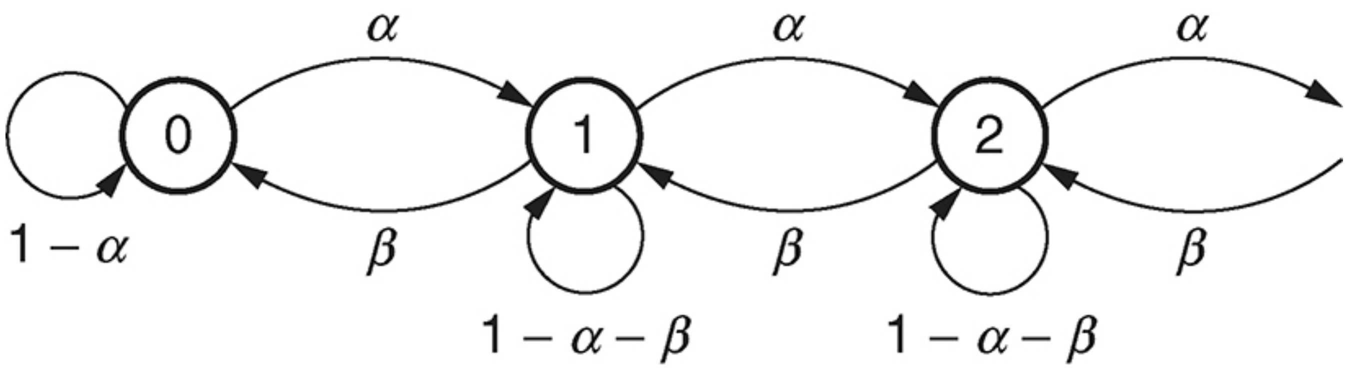


$P_{i,i+1} = \lambda(1-\mu)$, $P_{i,i} = \lambda\mu + (1-\lambda)(1-\mu)$ for $i > 0$, $P_{0,0} = 1 - \lambda(1-\mu)$

Let $\alpha = \lambda(1-\mu)$ = Pr(1 arrival, no departure)
$\beta = \mu(1-\lambda)$ = Pr(no arrival, 1 departure)

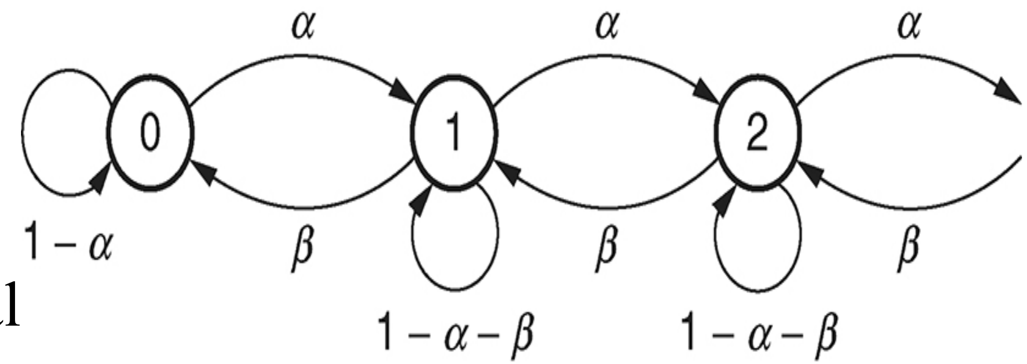
We will assume $0 < \lambda, \mu < 1$ which implies $0 < \alpha, \beta < 1$

Geo/Geo/1 queue



- The Markov chain $q(k)$ is
  - irreducible: any state is reachable from any other state
  - aperiodic: $P_{00} > 0$

Geo/Geo/1 queue



To find the stationary distribution, apply the local balance equation:

$$\beta\pi_{i+1} = \alpha\pi_i$$

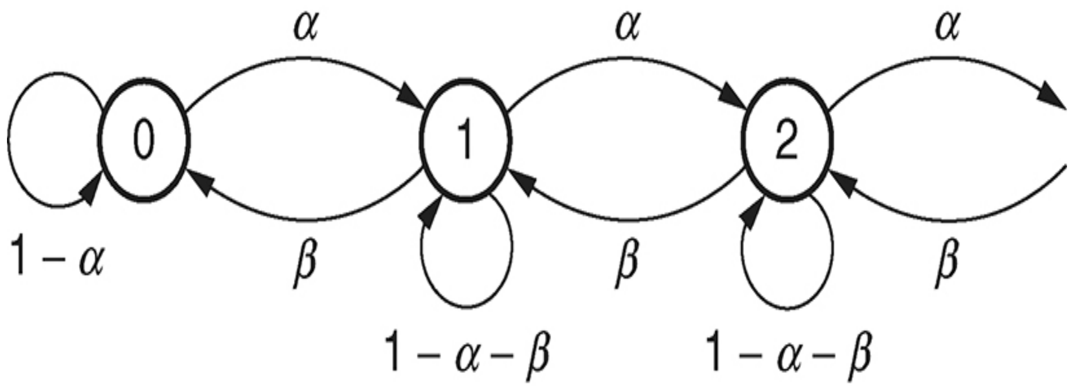
$$\Rightarrow \pi_{i+1} = \rho\pi_i \text{ where } \rho = \frac{\alpha}{\beta} = \frac{\lambda(1-\mu)}{(1-\lambda)\mu}$$

$$\left.\begin{aligned} &\Rightarrow \pi_i = \rho^i\pi_0 \\ &\textstyle\sum_i \pi_i = 1 \end{aligned}\right\} \Rightarrow \textstyle\sum_i \pi_i = \pi_0 \sum_i \rho^i = 1$$

The Markov chain has a stationary distribution iff $\rho < 1$, or equivalently $\lambda < \mu$

- If $\rho < 1$, $\sum_i \rho^i = \frac{1}{1-\rho}$ $\Rightarrow \pi_0 = 1 - \rho,\ \ \pi_i = \rho^i(1-\rho)$
- If $\rho \geq 1, \pi_0 \sum_i \rho^i = 1$ never hold

Geo/Geo/1 queue



Assume $\rho < 1$, then $\pi_i = \rho^i(1-\rho)$

The average queue length is

$$E(q) = \sum_i i\rho^i(1-\rho)$$

$$= (1-\rho)\rho\sum_i i\rho^{i-1}$$

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

$$= \frac{\rho}{1-\rho}$$

What is the average waiting time of a packet?

Little's law

Informally, "the mean queue length is equal to the product of the mean arrival rate and the expected waiting time"

- holds for very general arrival processes and service disciplines
- $A(t)$ – number of packet arrivals up to (and including) time-slot t
- $I_i(t) = 1$ if packet i arrived in a time-slot $< t$ and departs in a time-slot $\geq t$, $I_i(t) = 0$ otherwise
  - $I_i(t) = 1$ if packet i remains in the system at the beginning of time-slot t
  - $q(t) = \sum_{i=1}^{A(t-1)} I_i(t)$
  - the waiting time of packet i, denoted by w_i, is defined to be $w_i = \sum_{t=1}^{\infty} I_i\,(t)$

Little's law

$\lambda(T) = \frac{A(T)}{T}$: average arrival rate by time-slot T

$L(T) = \frac{1}{T}\sum_{t=1}^{T} q(t)$: average queue length by time-slot T

$W(n) = \frac{1}{n}\sum_{k=1}^{n} w_k$: average waiting time of the first n packets

Define the following limits (with probability 1)

$$\lambda = \lim_{T\to\infty} \lambda(T), \quad L = \lim_{T\to\infty} L(T), \; W = \lim_{n\to\infty} W(n)$$

Theorem 3.4.1 (Little's law) Assuming that λ and W exists and are finite, L exists and $L = \lambda W$.

- also applies to the stead-state expectations (from the ergodicity of Markov chains)

Proof of Little's law (sketch)

Theorem 3.4.1 (Little's law) Assuming that λ and W exists and are finite, L exists and $L = \lambda W$.

Proof: $L(T) = \frac{1}{T}\sum_{t=1}^{T} q(t)$

$$= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{A(t-1)} I_i(t)$$

$$= \frac{1}{T}\sum_{i=1}^{A(T-1)}\sum_{t=1}^{T} I_i(t)$$

$$\leq \frac{1}{T}\sum_{i=1}^{A(T)}\sum_{t=1}^{\infty} I_i(t) = \frac{1}{T}\sum_{i=1}^{A(T)} w_i$$

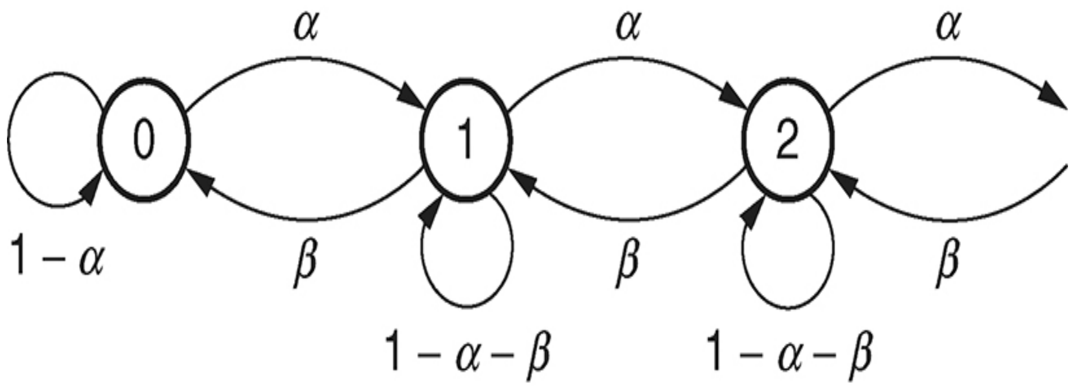
$$\lim_{T\to\infty} L(T) \leq \lim_{T\to\infty} \frac{1}{T}\sum_{i=1}^{A(T)} w_i$$

$$= \lim_{T\to\infty} \frac{A(T)}{T}\frac{\sum_{i=1}^{A(T)} w_i}{A(T)}$$

$$= \lambda\, W$$

It remains to show $L \geq \lambda\, W$ (see the textbook [SY])

Geo/Geo/1 queue



Assume $\rho < 1$, then $\pi_i = \rho^i(1-\rho)$

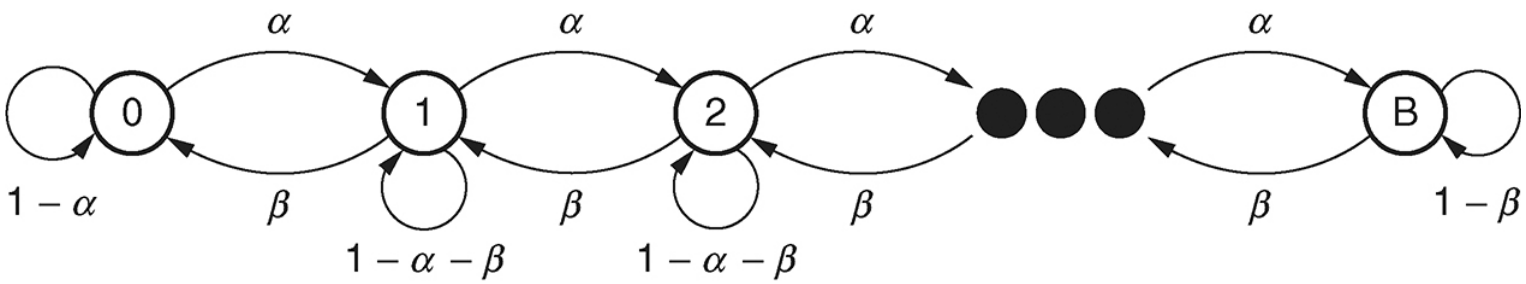
The average queue length is

$$L = E(q) = \sum_i i\rho^i(1-\rho)$$

$$= \frac{\rho}{1-\rho}$$

The mean waiting time of a packet $W = \frac{L}{\lambda} = \frac{\rho}{\lambda(1-\rho)}$

Geo/Geo/1/B queue



- Same setting as Geo/Geo/1 except that the buffer size is $B < \infty$
  - $q(t)$ is a irreducible and aperiodic DTMC with a finite state space

$$\beta\pi_{i+1} = \alpha\pi_i \quad \text{for } 0 \le i \le B-1,$$

$$\Rightarrow \pi_{i+1} = \rho\pi_i \text{ where } \rho = \frac{\alpha}{\beta} = \frac{\lambda(1-\mu)}{(1-\lambda)\mu} \quad \text{for } 0 \le i \le B-1,$$

$$\Rightarrow \pi_i = \rho^i\pi_0 \quad \text{for } 0 \le i \le B,$$

$$\Rightarrow \pi_0 \sum_{i=0}^{B} \rho^i = 1 \;\; \Rightarrow \pi_0 = \frac{1-\rho}{1-\rho^{B+1}} \;\; \Rightarrow \pi_i = \frac{(1-\rho)\rho^i}{1-\rho^{B+1}}, i = 0, 1, \dots, B$$

- What is the fraction of arriving packets that are dropped?
  - $p_d = \Pr(q(t) = B | a(t) = 1) = \Pr(q(t) = B) = \pi_B$