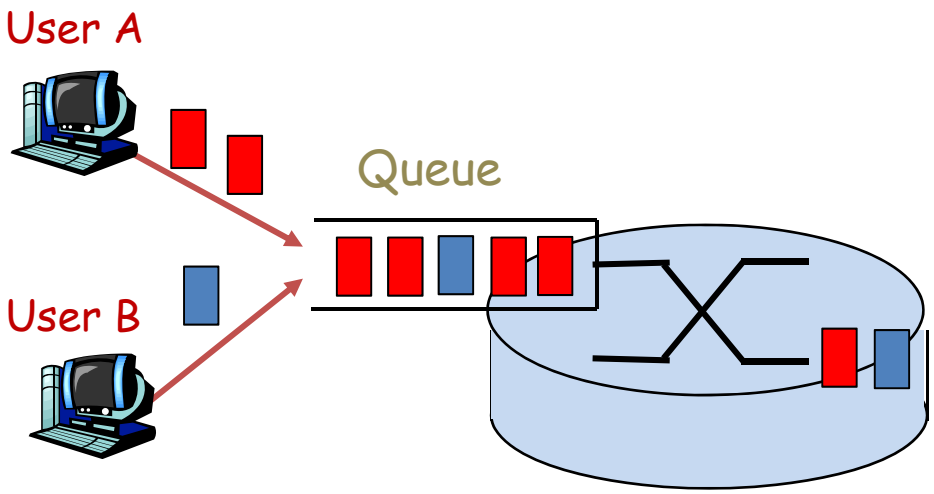
T2 Network Analysis and Queueing Theory



Two approaches to network design

1- "Build first, worry later" approach

- More ad hoc, less systematic

2- "Analyze first build later" is used extensively for telephone networks

- More systematic, optimal, etc.

- A model is a mathematical abstraction: keep only the details that are relevant
- Mathematical modeling is one step that can provide useful approximations

- Mathematical model can be used for:

  1- Evaluate the system performance

  - Average queue length
  - Average waiting time
  - Loss probability due to buffer overflow

  2- Improve the system performance

  - Determine the service rate with tolerable waiting time (upgrade the system with more capacity incurs investment cost. But long waiting time would be annoying to users)

  - Provide guaranteed packet loss probability with large enough buffer (How large is enough?)

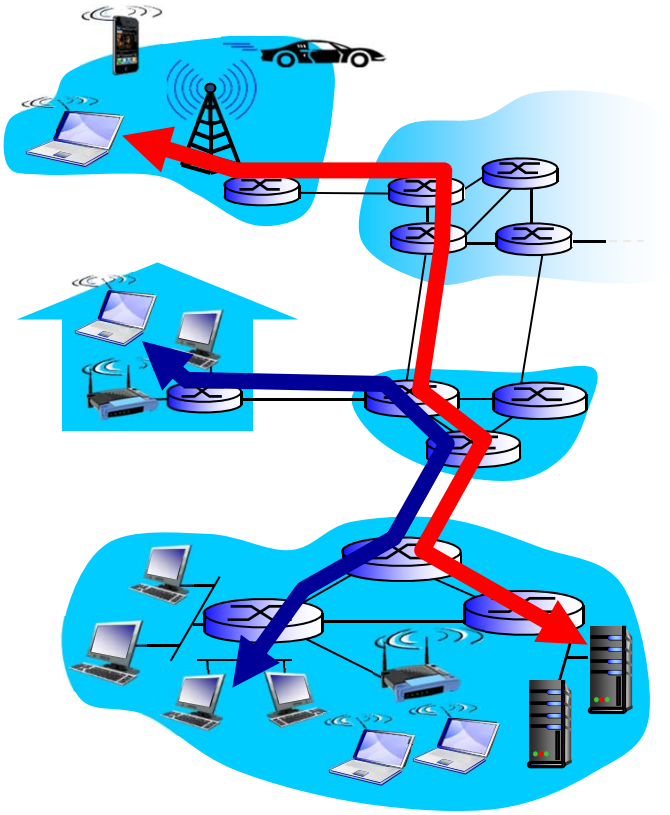
Roadmap

- Randomness in computer networks
- Single server queue

Randomness at the source (1/2)

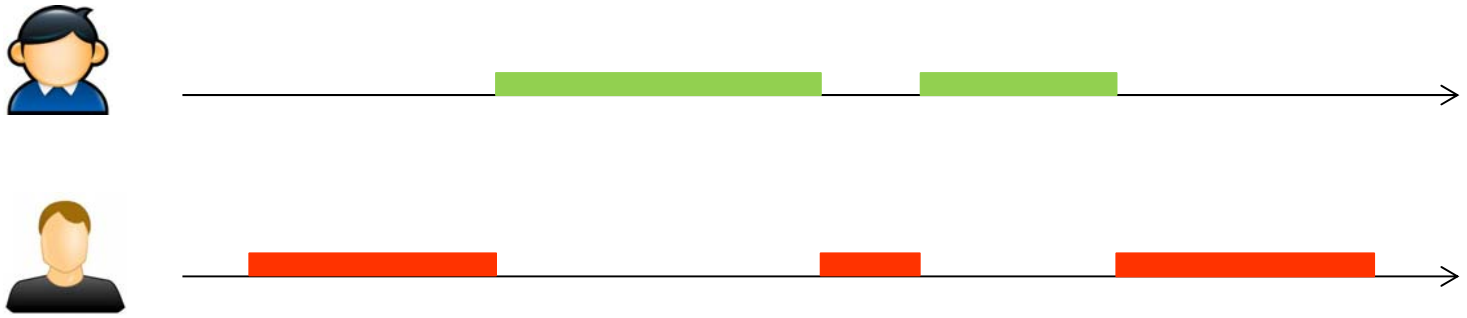
- Communication is full of randomness during the data generation at the source and data transmission at the network

- Random data generation at the source
  - data: users randomly subscribe to download at any time (e.g., web browsing)
  - voice: dynamic changing of data flow based on the interactivity during the conversation
  - video: variable video rate

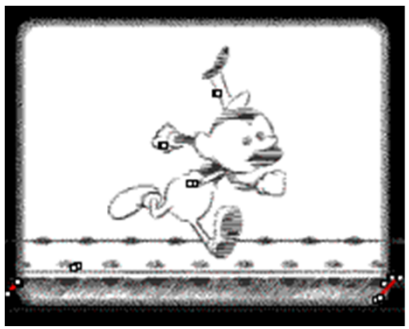


Randomness at the source (2/2)

- Voice communication: interactive communication



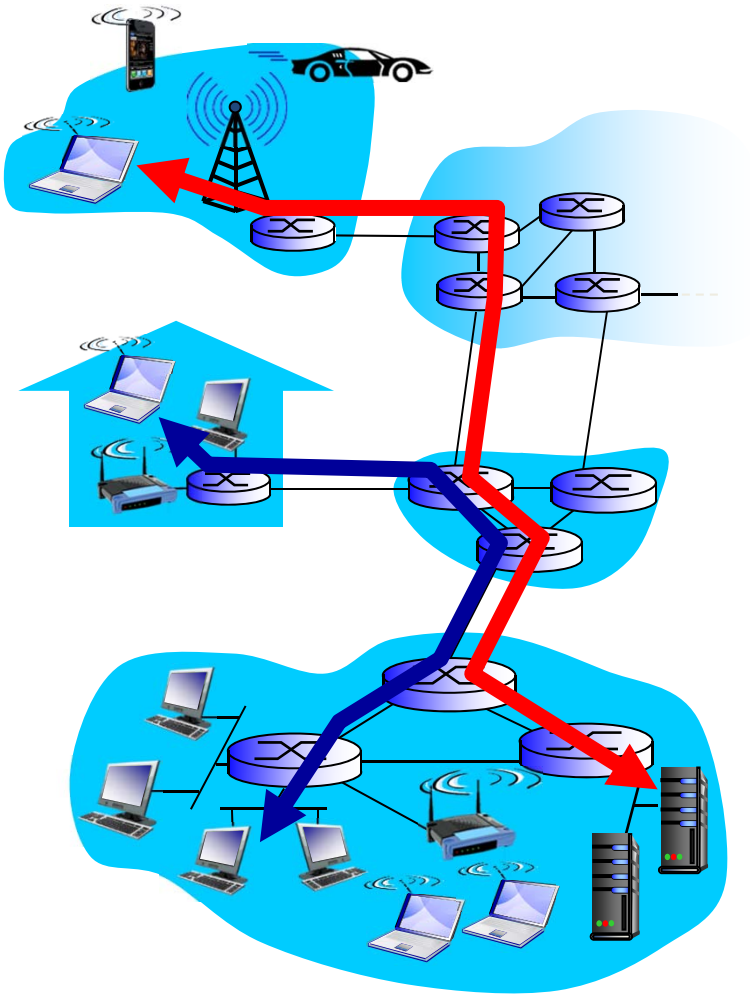
- Video: variable length of video frames



  - Video = moving picture (30 pics/s)
  - picture (frame) size depends on motion in pictures

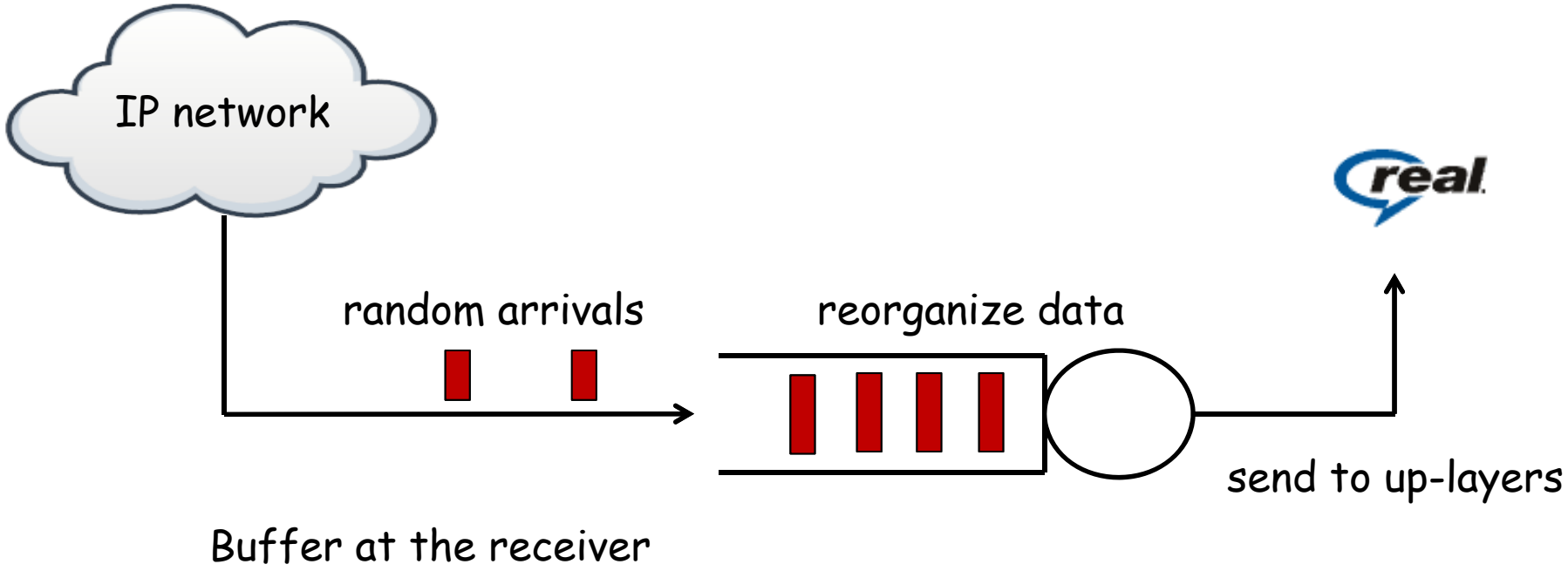
Randomness at the network

- Data transmission encounters random delays and data rate
  - randomly changing physical channels (noise of wireless communication)
  - congestion in the backbone network (statistical multiplexing of data packets)
  - contention of transmission among users in access network



Buffers

- Buffers are used to absorb the randomness of network
  - Sender, routers, receivers all have buffers



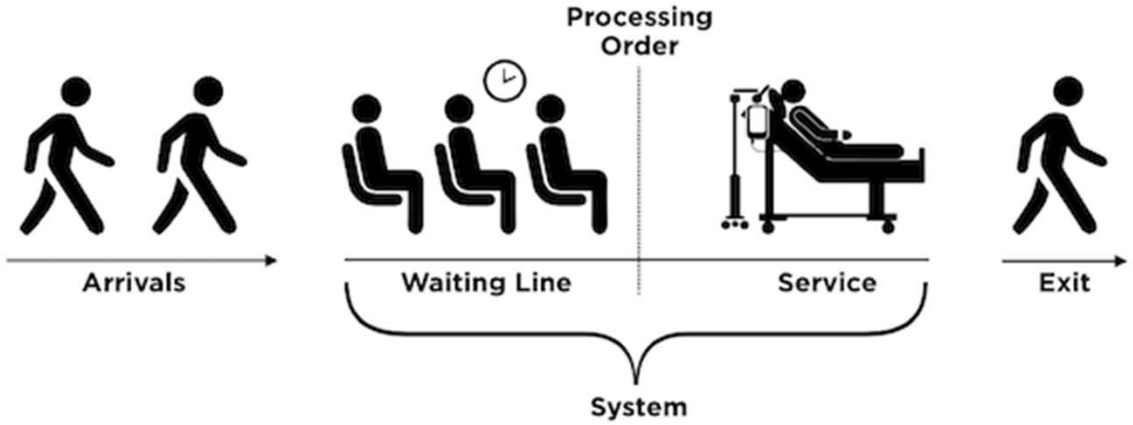
Roadmap

- Randomness in computer networks
- Single server queue

Note: the following part was originally prepared by R. Srikant (UIUC) and Lei Ying (ASU)

Queueing System

- Queueing System:
  - Incoming traffic: packet arrivals
  - Outgoing traffic: packet departures
  - Buffer: storing packets waiting for service
  - Server: process each packet before it departs
  - Service discipline: First-In First-Out, Processor sharing, etc.



Single Server Queue

- One server, operating in discrete-time
  - Time is slotted (equal duration) and indexed as 0, 1, 2,…
- Packet arrivals
  - The time between the arrival of two packets is geometrically distributed with parameter $\lambda \in (0,1)$
  - In other words, if a packet arrives at time t=10, then the next packet will arrive at time 10+X, where X is geometrically distributed with mean $1/\lambda$
- Service time
  - Each packet takes a geometrically distributed amount of time to process, with the parameter of geometric distribution denoted by $\mu \in (0,1)$, which is called service time
- The inter-arrival times and service times of packets are independent

Queueing Assumptions

- Packets arriving at the server are enqueued in the buffer

- The first packet in the buffer is processed by the server. When a packet finishes service, it departs the system. The next packet is then processed by the server

- In each time slot, we assume that any arriving packets arrive first, and then any departing packets depart. In particular, a packet that arrives in the system is available for service in the same time slot. We will use this assumption to develop a mathematical model for the queue later

Geometric Distribution: Arrivals

- Let λ be the probability with which a packet arrives in a time slot (Bernoulli random variable).
  - Then, the next packet will arrive after X time slots, where
$$P(X = i) = (1 - \lambda)^{i-1}\lambda, \qquad i \geq 1$$
  - since this is simply the probability that there are no arrivals for the next (i–1) time slots and there is an arrival in the i^{th} time slot
- A random variable X with such a distribution is called a geometric distribution
- Important: based on the discussion above, note that we can think of the packet arrival process in one of two equivalent ways:
  - Inter-arrival distribution is geometric
  - A packet arrives with probability λ in each time slot, i.e., the number of packets arriving in a time slot is a Bernoulli random variable

Mean of the Geometric Distribution

$$E(X) = \sum_{i=1}^{\infty} i * P(X = i) = \sum_{i}^{\infty} i(1-\lambda)^{i-1}\lambda = \frac{1}{\lambda}$$

- Thus, the inter-arrival time of packets has mean $\frac{1}{\lambda}$ time slots
- Since the probability of one arrival is λ, and the probability of no arrivals is $1-\lambda$, the mean number of arrivals in each time slot is λ. This is called the arrival rate (packet arrivals/time slot)

Geometric Distribution: Service Times

- In a similar manner, one can think of the amount of time required to process or serve a packet as being geometrically distributed with mean $1/\mu$, or equivalently, the probability with which a packet departs the system is μ

- Thus, the mean service time of a packet is $1/\mu$ time slots, and the mean service rate of the server is μ packets/time slot

Queueing Model

- $q(k)$: Queue length (number of packets in the queue) at the beginning of time slot k
- $a(k)$: Variable which takes on a value 1 if there was an arrival in time slot k and 0 otherwise, i.e., it is an indicator variable indicating if an arrival occurred
- $d(k)$: Indicator variable indicating if there was a departure in time slot $k$or not. Note that $d(k)$ has to be 0 if there is no packet in the queue at the beginning of the time slot and if there was no arrival, i.e., $d(k)$=0 if $q(k)$+$a(k)$=0. Otherwise, it can be either 1 or 0 depending on whether there was a departure or not.

Queuing dynamics

$$q(k+1) = q(k) + a(k) - d(k),$$

Where

$$a(k) = \begin{cases} 1 \quad with\ probability\ \lambda \\ 0\ with\ probability\ 1-\lambda \end{cases}$$

and if $q(k) + a(k) > 0$,

$$d(k) = \begin{cases} 1\ with\ probability\ \mu \\ 0\ with\ probability\ 1-\mu \end{cases}$$

Probabilistic Description

- Define $\mathrm{P}_{i,i+1} := \mathrm{P}(q(k+1) = i+1 | q(k) = i)$ to be the conditional probability that the queue length increases from i to i+1 in one time slot.
- Thus, it is the probability that there is one arrival and no departures in a time slot, i.e.,

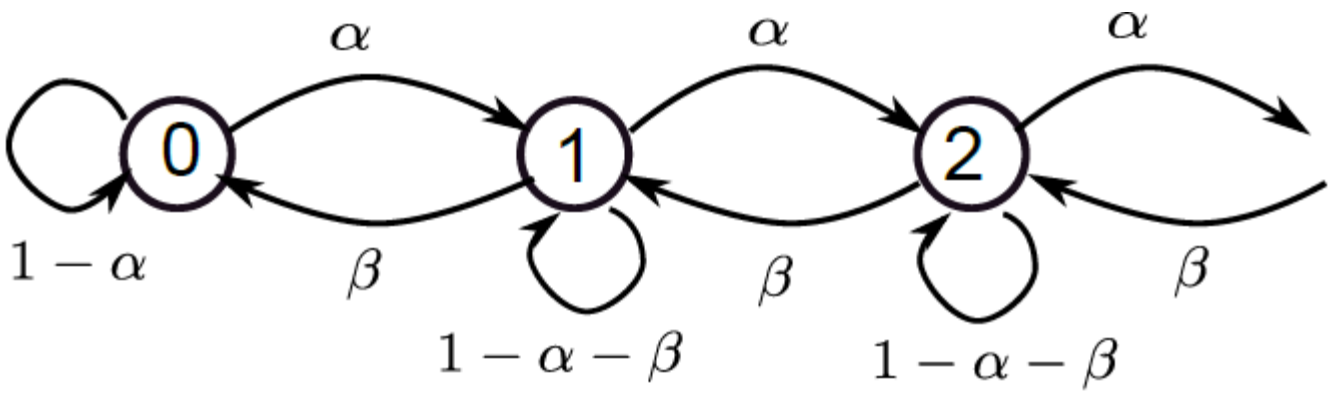
$$\mathrm{P}_{i,i+1} = \lambda(1-\mu)$$

- Similarly, $P_{i+1,i}$, the conditional probability that the queue length decreases by 1 is given by

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

Pictorial Description

- Each circle represents the state of the system, i.e., the number of queue at the beginning of a time slot
- The arrows represent transitions from one state to another (including the possibility of staying in the same state) and the expressions on the arrows represent the probability of the transition occurring.



Performance Analysis

- How do we compute the performance measure of such a system?
- One can try to compute the expected number of packets in the system at any time instant
- Or, one can try to compute the mean waiting time of a packet entering the system
  - The waiting time of a packet is the amount of time that a packet stays in the system
  - If a packet arrives in time slot t and departs in time slot $t+n$, then its waiting time is n
  - We would be interested in the average of the waiting time over all packets

Expected Queue Length

- Let $p_i(k) = P(q(k) = i)$ be the probability that the queue length is equal to i at the beginning of time slot k
- If we can compute $p_i(k)$, then we can calculate the mean number of packets in the queue in time slot k as

$$E\big(q(k)\big) = \sum_{i=0}^{\infty} i p_i(k)$$

- In principle, one can calculate $p_i(k)$ by using the fact that the queue length in the previous slot must have been i–1, i, or i+1
- The queue length at time slot k depends on the queue length at time slot k-1

Markov Chain

- Thus, (for $i > 0$),

$$p_i(k) = p_{i-1}(k-1)P_{i-1,i} + p_i(k-1)P_{i,i} + p_{i+1}(k-1)P_{i+1,i}$$

- In the above expression, the first term on the right-hand side is the probability that the queue was in state (i–1) in the previous time-slot and made a transition from (i–1) to i. The other terms can be interpreted similarly.
- In principle, if $p_i(0)$ was known for all i, then we can calculate $p_i(k)$ for all i and k. But this would be cumbersome. So we will do something much simpler
- The queueing system above is an example of a Markov chain. A Markov chain is a stochastic system where the probabilistic description of the system in time slot k+1 can be written in terms of the probabilistic description in the previous time slot k, and one does not need information about past time slots k–1, k–2,….

Steady-State

- It is much easier to compute $p_i(\infty)$, instead of $p_i(k)$ for any finite time k
- We will use the special notation π_i to denote $p_i(\infty)$
- π_i is called the steady-state or stationary probability of being in state i
- If the system has been in operation for a long time, then π_i is a good approximation to the probability that the queue length at the current time is i

http://setosa.io/markov/random-sequence-markov.html

How to calculate π_i?

- Interpret π_i as the long-term fraction of time that the queue spends in state i. Thus, over a large time interval T, the queue will be in state i for $\pi_i T$ time slots
- Further, over this long time interval, the number of times the queue length jumps from i to $i+1$ is $\pi_i T P_{i,i+1}$
  - Recall $P_{i,i+1}$ is the probability of jumping from i to $i+1$
- Similarly the number of times the queue length jumps from $i+1$ to i is $\pi_{i+1} T P_{i+1}$

How to calculate π_i?

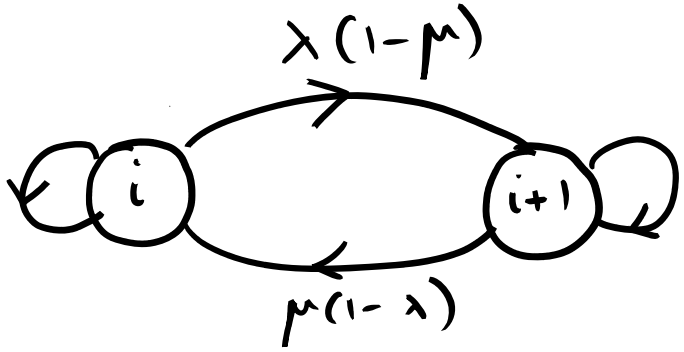
- The number of jumps from i to $i+1$ over any time interval cannot be different from the number of jumps from $i+1$ to i by more than 1 Why?
  - Each time the queue length jumps from i to $i+1$, it has to get back i to jump again from i to $i+1$
- Thus,

$$|\pi_i T P_{i,i+1} - \pi_{i+1} T P_{i+1,i}| \leq 1$$

- Dividing by T and letting $T \to \infty$, we get

$$\pi_i P_{i,i+1} = \pi_{i+1} P_{i+1,i}$$

Solving for π_i



$$\pi_i \lambda(1-\mu) = \pi_{i+1}\mu(1-\lambda)$$

- Equivalently,

$$\pi_{i+1} = \rho\pi_i$$

where $\rho = \frac{\lambda(1-\mu)}{\mu(1-\lambda)}$

Solving for π_i

- Thus, $\pi_i = \rho^i \pi_0$. Next, using the fact that $\sum_{i=1}^{\infty} \pi_i = 1$ (since π_i are probabilities), we get

$$\pi_0 \sum_{i=0}^{\infty} \rho^i = 1$$

- Assume $\rho < 1$. Then, using the fact that $\sum_{i=0}^{\infty} \rho^i = \frac{1}{1-\rho}$, we get

$$\pi_0 = 1 - \rho,$$

and

$$\pi_i = \rho^i(1-\rho), i = 0,1,2, \ldots$$

The Condition $\rho < 1$

- We have obtained the steady-state distribution under the condition $\rho < 1$, which is equivalent to $\lambda < \mu$
  - Interpretation: we need arrival rate to be less than service rate
- When $\rho \geq 1$, which is equivalent to $\lambda \geq \mu$, the sum $\sum_{i=0}^{\infty} \rho^i$ becomes infinity, and we cannot compute the steady-state distribution
  - Interpretation: when the arrival rate is greater or equal to the service rate, the queue will not be stable; the queue lengths will blow up to ∞
  - One can use Markov chain theory to rigorously establish that a steady-state distribution does not exist when $\lambda \geq \mu$

Interpretation of Steady-State

- Once the system reaches steady-state, then the probability distribution over the queue lengths will not change.

- In other words, if $p_i(k) = \pi_i \ \forall i$, then $p_i(k+1) = \pi_i \ \forall i$

- We can verify this by substituting for $p_i(k) = \pi_i \ \forall i$ in the equation on Slide 43.

Expected Queue Length in Steady State

- Let $L = E(queue\ length)$ in steady-state. Then,

$$L = \sum_{i=0}^{\infty} i\pi_i = \sum_{i=0}^{\infty} i\rho^i(1-\rho) = \frac{\rho}{1-\rho}$$

- Note that as $\rho \to 1$, or equivalently $\lambda \to \mu$, the expected steady-state queue length goes to ∞

Expected Waiting Time of a Packet

- Little's law (intuition in the next slide) is a relationship between the expected queue length (L) and expected waiting time (W)

$$L = \lambda W$$

  - Recall that λ is the arrival rate
  - The above relationship holds for very general queueing systems
- Thus, for our specific model

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

More on Little's Law

- Suppose a restaurant earns $1 per hour from each customer
- If there are L customers on average in the restaurant, the restaurant's income is L dollars per hour
- Another way to calculate the income is as follows: suppose each customer stays in the restaurant for W hours on average, then the income per customer is W dollars
- If customers arrive at the rate of λ customers/hour, then the restaurant earns λW dollars per hour
- These two ways of calculating income must yield the same answer, so $L=\lambda W$