COMP312-09A Communications and Systems Software

Lecture 8 – Switching and Queues

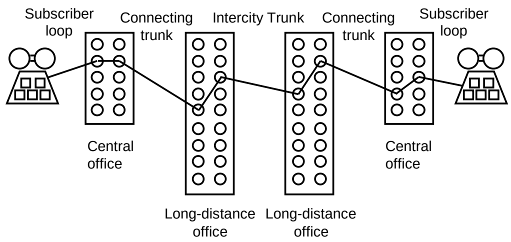
Overview

- Last lectures
  - TCP uses packet loss as an indication of congestion
  - Congestion occurs when a node is unable to forward packets at the rate they arrive at.
- This lecture
  - Switching
  - Queues

First, some assumptions

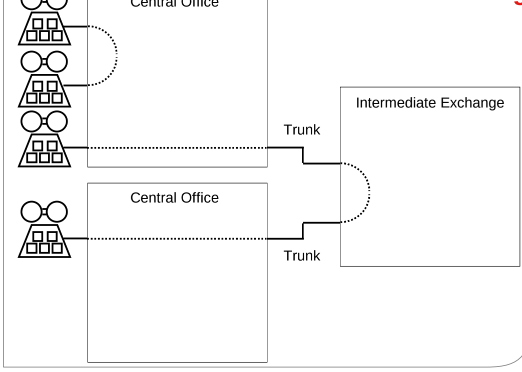
- We're talking about packet switched networks
  - All we've talked about in 202, and so far in 312.
  - Inherent to computer communications.
- Alternative network style: POTS
  - Plain old telephone system, before it was digitised.
  - Circuit switching concept
    - Dedicated, unshared resource.
    - Constant characteristics (bandwidth, data rate, delay)
    - Copper wire, frequency channel

Circuit switching



Galloway Street

Circuit Switching



Circuit Switching

- Three phases of a circuit
  - Setup
  - Information transfer
  - Tear down
- Circuits may be long term
  - “permanent”
  - “leased”
- … or short term.
  - “switched”
- The same infrastructure is used for both

Circuit Switching vs. Packet Switching

- Circuit occupies reserved space.
  - Predictable delay and loss characteristics as no competing traffic
  - Customer pays for reservation … “expensive”
  - Per-circuit (connection) state held throughout network
    - Switches are intelligent
- Packet switching uses whatever capacity is available
  - Often best-effort (Internet Service Model)
  - Unpredictable delay and loss characteristics
  - Customer pays for what they use … “cheap”
  - No connection state held
    - Switches are dumb

Packet switched networks

- Each packet has enough information to allow it to be routed to the destination. Examples:
  - IPv4 header (Internet routing)
    - TCP, UDP headers: routing message to correct socket inside end host
  - Ethernet header (Ethernet forwarding)
- Packet switching allows high speed circuits to be shared efficiently.
  - Link runs at full speed.
  - Matches most observed user behaviour
    - Data is transmitted in bursts, followed by silent periods

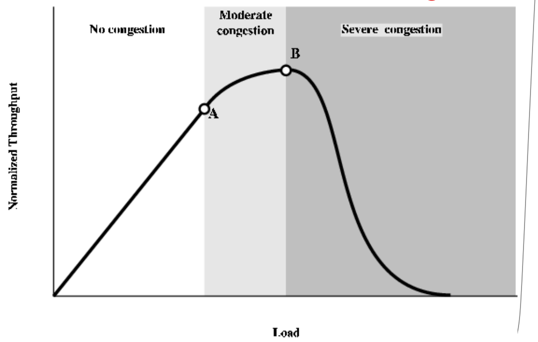
Network Delay

- Network delay is made up of constant and variable factors
- Constant
  - Serialisation delay. Time taken to put packet one bit at a time onto the network.
  - Propagation delay. Time taken for signal to make its way to other end of medium. Bound by speed of light.
- Variable
  - Queuing delay. Time spent waiting for service
  - Routing decision. Time spent determining the output interface. In practice this delay is done in an application-specific integrated circuit (ASIC) and is negligible

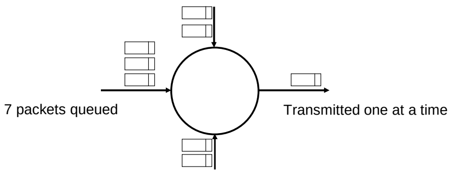
Why do we care about queues?

- Packet switched networks perform well while the arrival rate is less than the network’s capacity.
- Network conditions deteriorate rapidly when it is not
  - Similar to a traffic jam where system collapses and little productive work (travelling) is done.

Congestion



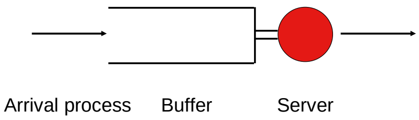
Queues



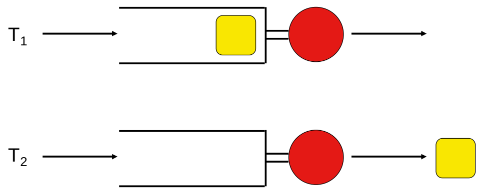
Queues

- A queue consists of
  - Buffer space
  - One or more servers
- Customers arriving go into the buffer.
- One customer at a time receives service from a server
- After a service time the customer leaves the queue and another queue receives service

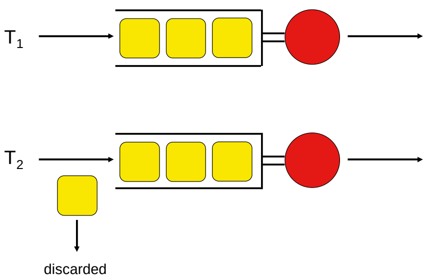
Queue



Queue



Queue



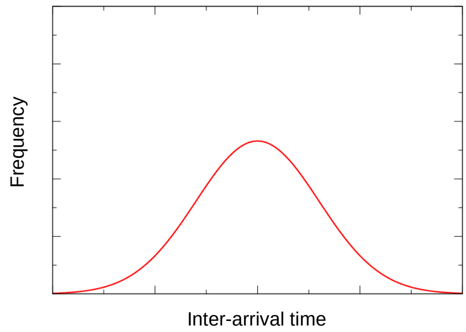
Queues: Characterised by

- Arrival process
  - Characteristics of packet inter-arrival times.
- Distribution of service times
  - How long it takes forwarding decision to be made.
- The number of servers
- The size of the buffer
- The order in which customers receive service
  - Normally FIFO (first in, first out)

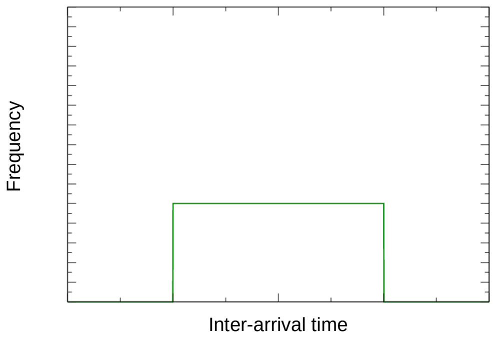
Arrival processes

- A deterministic arrival process has a constant inter-arrival time
- Common random distributions include
  - Normal
  - Uniform
  - Bi-modal
  - Negative Exponential

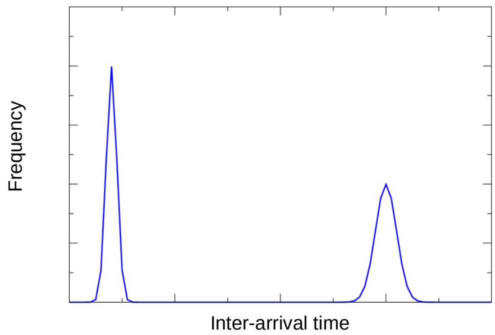
Normal



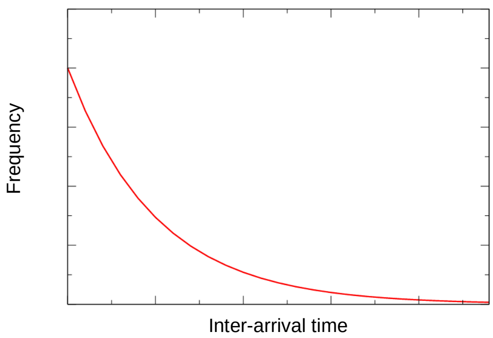
Uniform



Bimodal



Negative Exponential



Arrival distributions

- Normal
  - Packets arrive at a rate with a distribution that can be accurately characterised by a mean and standard deviation
- Uniform
  - Packets arrive at a constant rate
- Bi-modal
  - Packets tend to arrive separated by one of two delays, each of which can be accurately characterised by a mean and standard deviation
- Negative exponential
  - Most packets arrive separated by a short amount of time, though some arrive with larger separation

Poisson process

- A process with uncorrelated negative-exponentially distributed inter-arrival times is known as a Poisson process
- A Poisson process is truly random
  - there is no relationship between any of the arrivals.
- This process is reasonably common in queuing situations
  - Including arrival processes at Internet routers

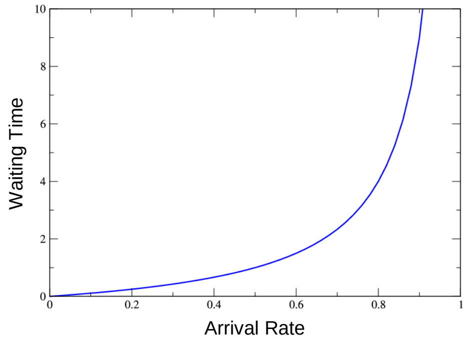
Queuing notation

- Queues are described by three or four parameters in the form A/B/C/D
- First two are arrival and service processes
  - M: Markov means Poisson arrivals or negative-exponential service times
  - G: General. Can be any distribution
  - D: Deterministic
- Third is the number of servers
- Fourth (optional) is size of the buffer. If absent, infinite buffer is assumed

M/M/1 Queue

- Simplest possible queue.
  - Poisson arrivals
  - Exponential service times
  - One server
  - Infinite buffer space
- Interested in the mean waiting time given an offered load

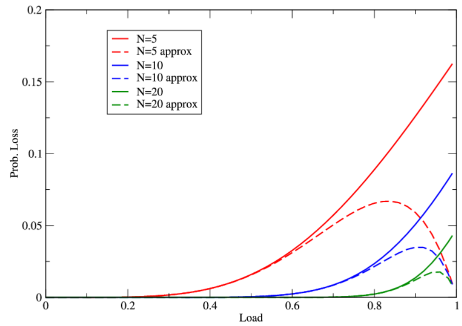
M/M/1 Waiting Time



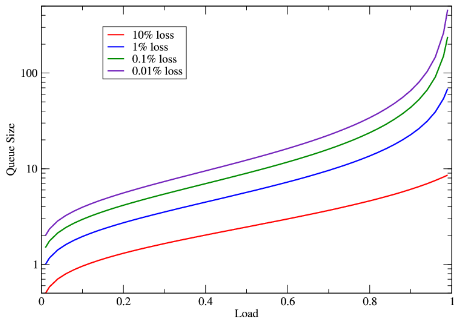
M/M/1/N

- In reality, queues do not have infinite buffer space
- M/M/1/N is the same as M/M/1 but buffer is not infinite
  - The queue size never exceeds N
- Interested to know probability of packets being lost due to queue overflowing
- Interested to also know minimum size of queue to achieve a specified packet loss rate

M/M/1/N Probability of Loss



M/M/1/N Buffer Space



Summary

- Utilisation must be kept much lower than 100% for good performance.
- Loss can be reduced by increasing queue length
  - At the expense of increasing delay variation through network
  - Current thought in research community is small queue sizes are good

Next lectures

- Other queue disciplines
- Virtual circuits
- ECN, RED.

Resources

- http://media.pearsoncmg.com/aw/aw_kurose_network_2/applets/queuing/queuing.html