Lecture 2
Introduction to Quality of Service (QoS)

Quality of Service (1)

- Two main objectives
  - Provide performance assurance in the managed network
  - Support service differentiation among users/flows
- Formal definition of QoS
  - The capability to provide resource assurance and service differentiation in a network is often referred to as quality of service
- How to provide QoS?
  - Proper definition of the adopted QoS index
  - Resource allocation mechanism (i.e., user/resource scheduling)

Quality of Service (2)

IPTD: IP packet transfer delay
IPDV: IP packet delay variance
IPLR: IP packet loss ratio
IPER: IP packet error ratio

- Service categories (or QoS classes) defined by ITU-T

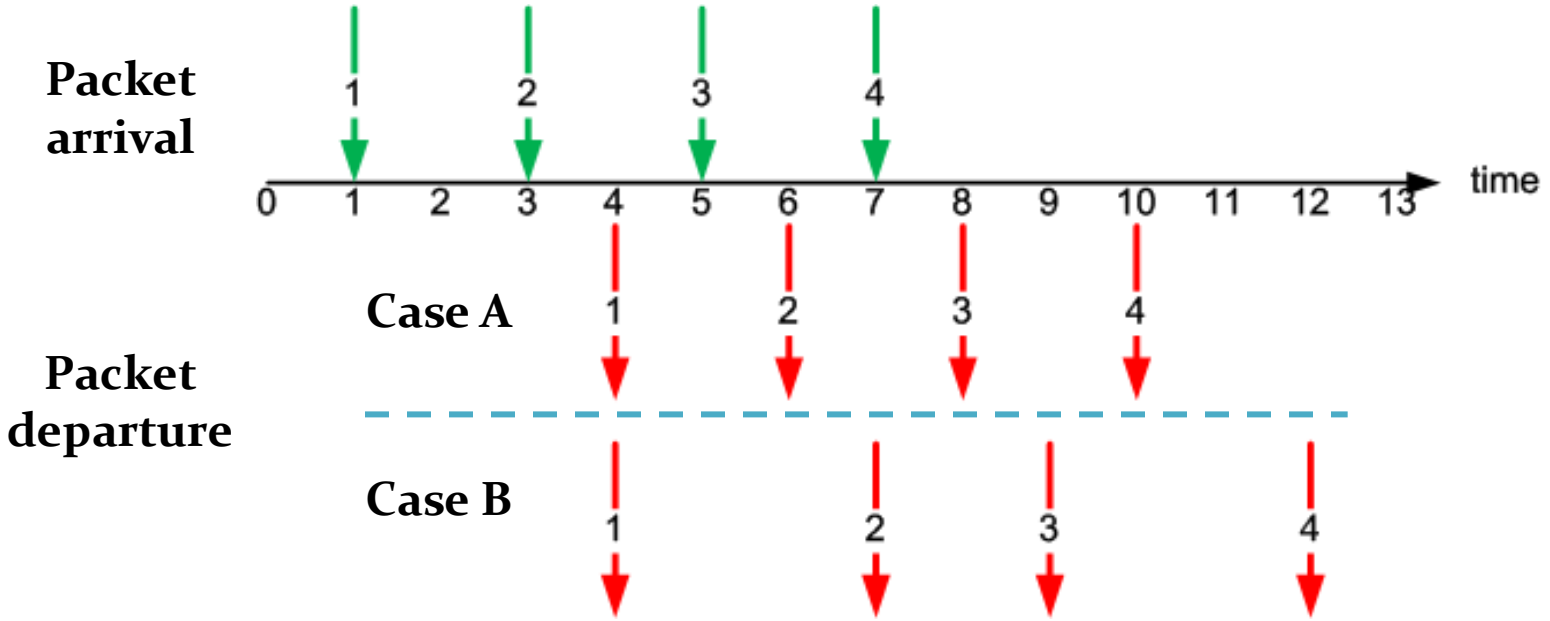
QoS class	Characteristics
0	Real-time, jitter sensitive, highly interactive
1	Real-time, jitter sensitive, interactive
2	Transaction data, highly interactive
3	Transaction data, interactive
4	Low loss only (short transactions, bulk data, video streaming)
5	Traditional applications of default IP neteworks

QoS class	Characteristics	IPTD	IPDV	IPLR	IPER
0	Real-time, jitter sensitive, highly interactive	100 ms	50 ms	10^{-3}	10^{-4}
1	Real-time, jitter sensitive, interactive	400 ms	50 ms	10^{-3}	10^{-4}
2	Transaction data, highly interactive	100 ms	U	10^{-3}	10^{-4}
3	Transaction data, interactive	400 ms	U	10^{-3}	10^{-4}
4	Low loss only (short transactions, bulk data, video streaming)	1 s	U	10^{-3}	10^{-4}
5	Traditional applications of default IP networks	U	U	U	U

The delay 5G remote surgery is at most 2 ms

Quality of Service (3)

- Delay variance/Delay jitter: expressed as an average of the deviation from the network mean latency



Case A:

$$\mu = \frac{3+3+3+3}{4} = 3$$

$$\sigma = \sqrt{\frac{4 \times 0^2}{4}} = 0$$

Case B:

$$\mu = \frac{3+4+4+5}{4} = 4$$

$$\sigma = \sqrt{\frac{1^2+0^2+0^2+1^2}{4}} = \sqrt{\frac{1}{2}}$$

5G-ACIA Defined 5QI

5QI value	Resource type	Default priority level	Packet delay budget	Packet error ratio	Default maximum data burst volume	Default averaging window	Example services
2	GBR	40	150 ms	10^{-3}	N/A	2000 ms	Conversational video (live streaming)
6	Non-GBR	60	300 ms	10^{-6}	N/A	N/A	Video (buffered streaming) TCP-based
...	...	...	...	...	...	...	...
82	Delay-critical GBR	19	10 ms	10^{-4}	255 bytes	2000 ms	Discrete automation
83	Delay-critical GBR	22	10 ms	10^{-4}	1354 bytes	2000 ms	Delay-critical GBR

What's Resource Allocation?

- The process that network elements perform to meet the demands of users/applications
- What are the mentioned resources?
  - Bandwidth of the links
  - Buffers at the routers and switches
  - Why?
    - Packets at a router/switch contend for the use of links
    - These contending packets are placed in queues to wait for been transmitted over the links

Resource Allocation v.s. Congestion Control (1)

- What would happen when too many packets are contending for the same link
  - The queue overflows, and thus packets get dropped
  - Packet drop triggers retransmissions
  - Also packet drop is a sign of network congestion
- Network should provide a congestion control mechanism to deal with such a situation

Resource Allocation v.s. Congestion Control (2)

- Allocating resources with any precision is difficult, since resources are distributed throughout the network
- Strategy of traditional Internet
  - Always let the sources send as much data as they want
  - Then recover from the congestion when it occurs
  - Easier approach but it can be disruptive because many packets many be discarded by the network before congestions can be controlled

Resource Allocation v.s. Congestion Control (3)

- Any better cooperation?
  - In network elements
    - Various queuing disciplines can be used to control the order in which packets get transmitted and which packets get dropped
  - At the hosts' end
    - The congestion control mechanism paces how fast sources are allowed to send packets

Classification of Resource Allocation (1)

- Router-centric v.s. host-centric
  - Router-centric: router is responsible for deciding when packets are sent and selecting which packets are to be dropped, and informing the hosts that how many packets they are allowed to send as well
  - Host-centric: end host observes the network conditions and adjusts its behavior accordingly
  - These two are not mutually exclusive
- Reservation-based v.s. feedback-based
  - Reservation-based: end host asks the network for a certain amount of capacity to be allocated for a flow
  - Feedback-based: end host begins sending data without reserving any capacity and then adjusts its sending rate according to the received feedback. Feedback could be “explicit” or “implicit”

Classification of Resource Allocation (2)

- Window-based v.s. rate-based
  - Window-based: the well-known one is TCP, in which the receiver advertises a window to the sender. This window limits how much data the sender can transmit
  - Rate-based: how many bits per second a sender can send is controlled so that the receiver or network can absorb. Rate-based control makes sense for multimedia applications

Evaluation Criteria (1)

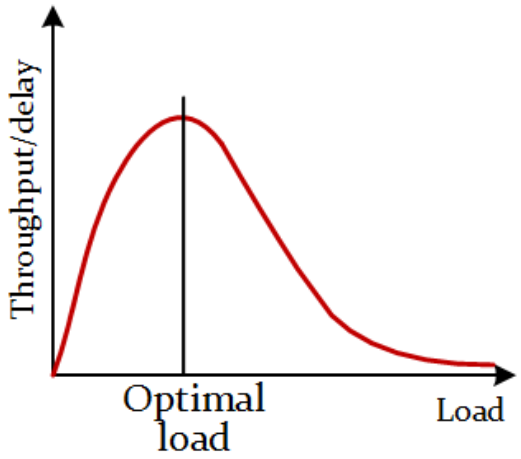
- What do we expect from the resource allocation mechanism?
  - How a network effectively and fairly allocates its resources?
- How to evaluate the effectiveness of the adopted resource allocation?
  - Through throughput and delay performance
  - However, throughput and delay conflict with each other
  - We want send as many packets into the network as possible to drive the utilization of all links up to 100%
  - While this would result in increasing the number of queued packets in the network; and thus the queue length and queueing delay increase as well

Evaluation Criteria (2)

- One common metric to evaluate the effectiveness of a resource allocation scheme is the ratio of throughput to delay, i.e.,

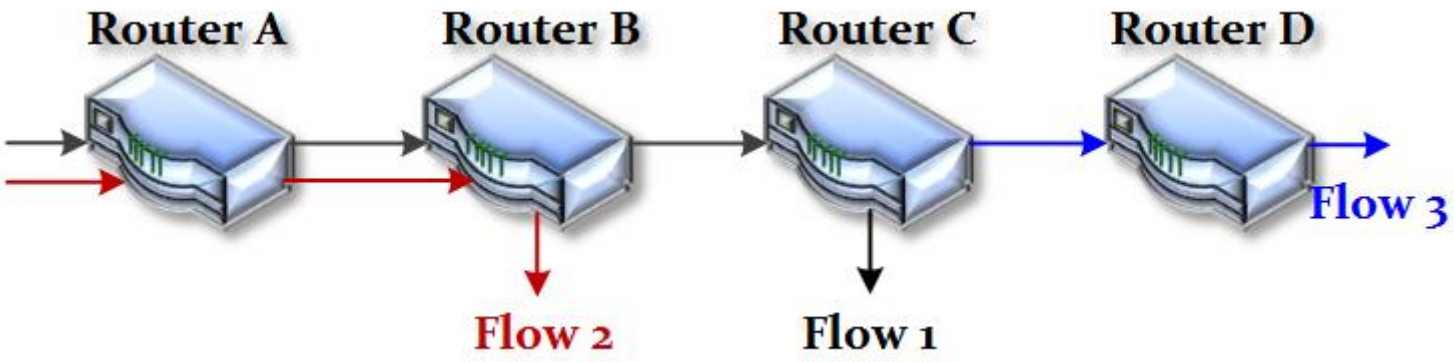
$$\text{Power} = \frac{\text{Throughput (bits)}}{\text{Delay (s)}}$$

- The objective is to maximize this ratio
- A representative power curve is shown below
- The power curve looks very much like the system throughput curve in a timesharing computer system



Evaluation Criteria (3)

- How to evaluate the fairness of a resource allocation?
  - Fairness means that each flow receives an equal share of the bandwidth
  - However, equal shares may not equate to fair shares
    - Locally fair share v.s. globally fair share of the three flows in the illustrative example below



Evaluation Criteria (4)

- Jain's fairness index
  - Given a set of flow throughputs (x_1, x_2, ..., x_n), the fairness index to these flows is

$$f(x_1, x_2, .., x_n) = \frac{(\sum_{i=1}^{n} x_i)^2}{n(\sum_{i=1}^{n} x_i^2)}$$

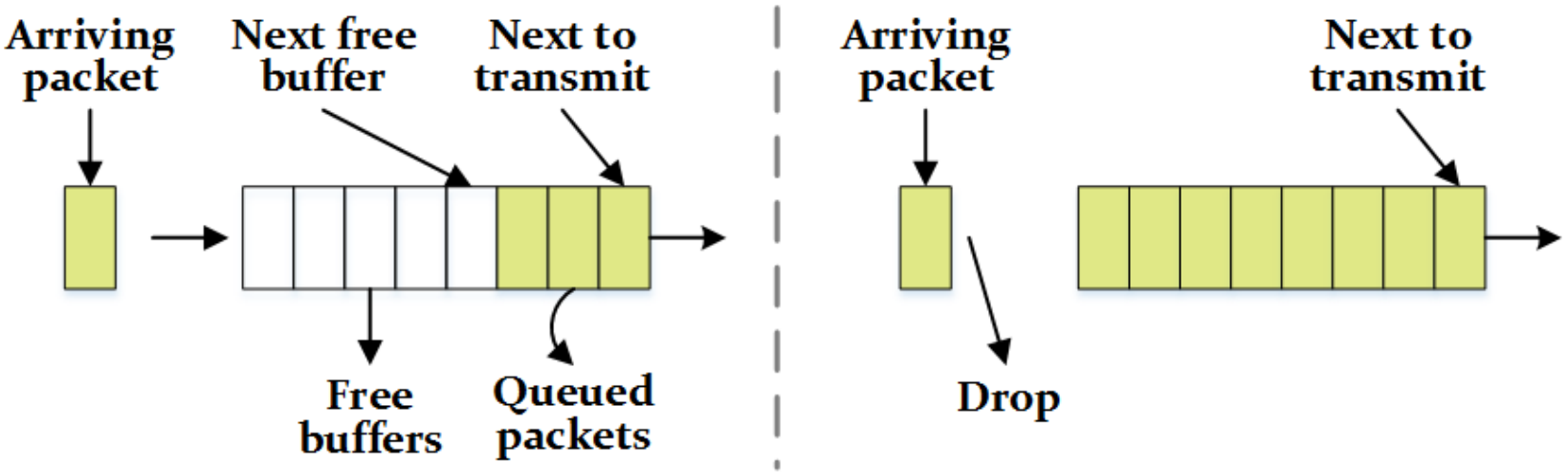
  - The fairness index always results in a number between 0 and 1, with 1 representing the greatest fairness
  - Consider *n* flows, each one receives a throughput of 1 unit of data per second, therefore the fairness index is $\frac{n^2}{n \times n} = 1$
  - Suppose one flow receives a throughput of 1+Δ. Now the fairness index is $\frac{((n-1)+1+\Delta)^2}{n(n-1+(1+\Delta)^2)} = \frac{n^2+2n\Delta+\Delta^2}{n^2+2n\Delta+n\Delta^2} < 1$

Queueing Disciplines

- What's queueing discipline?
  - Method implemented in routers to govern how packets are buffered while waiting to be transmitted
- Importance of queueing discipline
  - Queueing algorithm can be thought of as allocating both bandwidth (which packets get transmitted) and buffer space (which packet get discarded)
  - Queueing algorithm directly affects the latency experienced by a packet by determining how long a packet waits to be transmitted
- Two common queueing algorithms are first-in first-out (FIFO) and fair queueing (FQ)

FIFO

- The first packet that arrives at a router is the first packet to be transmitted
- Packets that arrives at the tail end of the FIFO are dropped (tail drop)
- Regardless of which flow the packet belongs to or how important the packet is
- Note that scheduling discipline differs from drop policy



Priority Queueing

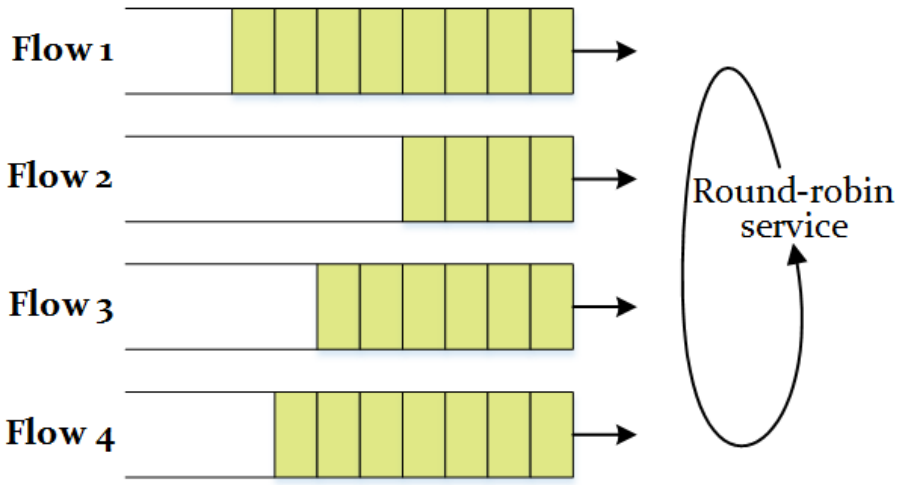
- A variation on basic FIFO is priority queueing
- Idea:
  - Each packet is with a priority
  - Router implements multiple FIFO queues, one for each priority class
  - Router always transmits packets out of the highest-priority queue if the queue is nonempty before moving on to the next priority queue
  - Within each priority, packets are still managed in a FIFO manner
- Problem of priority queueing: high priority queue can starve out all the other queues
  - From your view point, how to deal with this problem?

Fair Queueing (1)

- Main problem with FIFO queuing: it does not discriminate between different traffic sources, or it does not separate packets according to the flow to which they belong
- Fair queuing (FQ) is to address this problem
- Idea of FQ
  - Router maintains a separate queue for each flow currently being handled by the router
  - The router then serves these queues in a sort of round-robin

Fair Queueing (2)

- The main complication with FQ is that the packets being processed at a router are not necessarily the same length
- To truly allocate the bandwidth of the outgoing link in a fair manner, it is necessary to take packet length into consideration



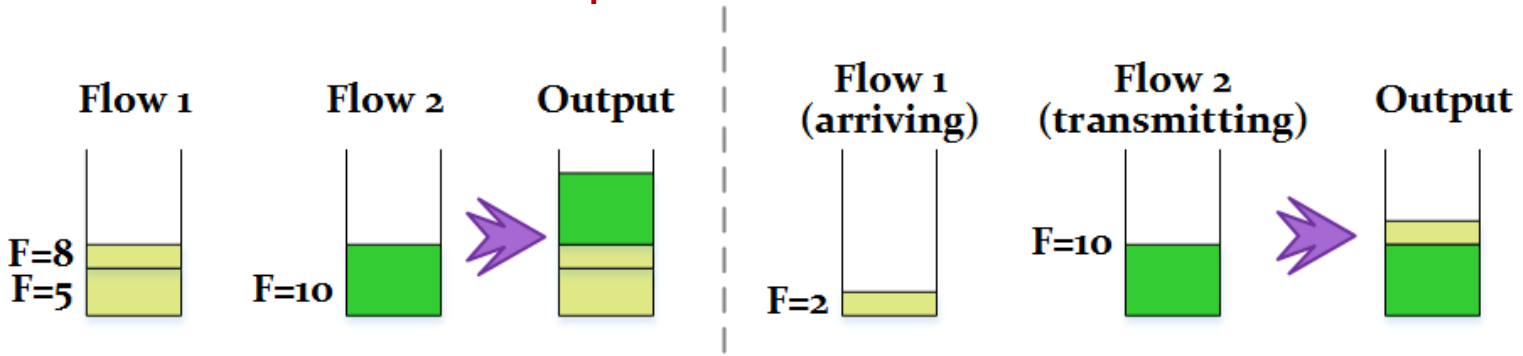
- For example, if a router is managing two flows, one with 1000-byte packets and the other with 500-byte packets (perhaps because of fragmentation upstream from this router), then a simple round-robin servicing of packets from each flow's queue will give the first flow two thirds of the link's bandwidth and the second flow only one-third of its bandwidth

Fair Queueing (3)

- What we really want is bit-by-bit round-robin; that is, the router transmits a bit from flow 1, then a bit from flow 2, and so on
- Clearly, it is not feasible to interleave the bits from different packets
- The FQ mechanism therefore simulates this behavior by first determining when a given packet would finish being transmitted if it were being sent using bit-by-bit round-robin, and then using this finishing time to sequence the packets for transmission
- For a flow, let
  - P_i : the length of packet i
  - S_i: time when the router starts to transmit packet i
  - F_i: time when router finishes transmitting packet i
  - $F_i = S_i + P_i$

Fair Queueing (4)

- When do we start transmitting packet *i*?
  - Depends on whether packet *i* arrived before or after the router finishes transmitting packet *i*-1 for the flow
  - Let A_i denote the time that packet *i* arrives at the router
  - Then $S_i = \max(F_{i-1}, A_i)$
  - $F_i = \max(F_{i-1}, A_i) + P_i$
  - Now for every flow, we calculate F_i for each packet that arrives using our formula
  - We then treat all the F_i as timestamps
  - Next packet to be transmitted is always the packet that has the lowest finish timestamp



Fair Queueing (5)

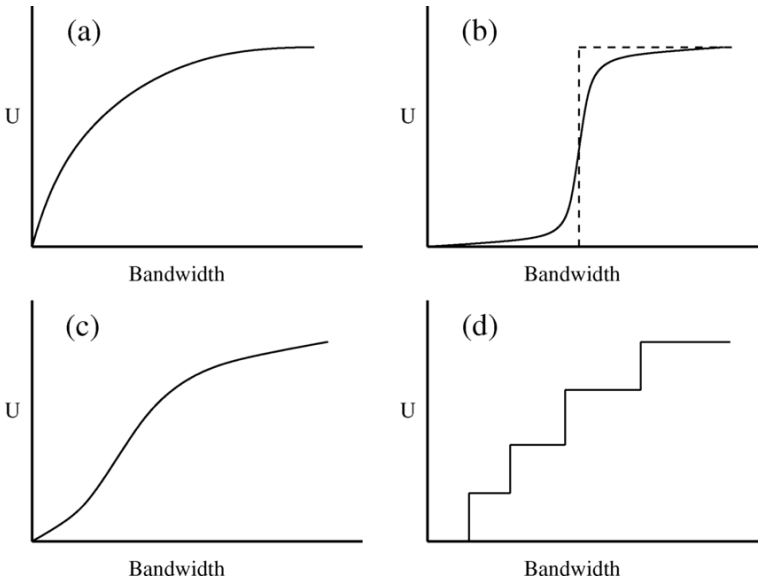
- Two notes of fair queueing
  - The link is never idle as long as there is at least one packet in the queue (we call this *work conserving*)
  - If the link is fully loaded and there are *n* flows sending data, one cannot use more than 1/*n* of the link bandwidth
    - If one tries to send more than that, it itself would incur long delay
- Weighted fair queueing (WFQ) is a variation of FQ
  - Each flow (queue) is assigned a weight
  - This weight logically specifies how many bits to transmit each time the router serves that queue
  - Therefore, this weight controls the percentage of the link's bandwidth that the flow will get

Proportional-Fair Resource Allocation

- Motivation of proportional fairness
  - Maximize the total wireless network throughput while at the same time allowing all users at least a minimal level of service
- Formal definition
  - let T be the observation time duration
  - Let T_i be the time utilized by flow i
  - Let R_i be the data rate of flow i
  - The throughput of flow i is $S_i=(T_iR_i)/T$
  - Proportional fairness is to achieve

$$\max \sum_{\forall i} \mathrm{U}(S_i)$$

  - U(•) is utility function
  - The common utility function is log



Max-Min Fair Resource Allocation

- Idea: maximize the minimum resources the user gets among all
- An illustrative example is shown here
  - The bottleneck link is link A-C
  - Do resource allocation on link A-C first
  - Since the resources allocated to the dark green flow has been determined, the next bottleneck link is link A-B. As a result, the rate assigned to yellow flow is 1-1/3=2/3
  - The rate allocated to blue flow is 2-2/3=4/3