UNIVERSITY OF WOLLONGONG AUSTRALIA

# Week 10: Performance Architecture

**CSIT985**

**Strategic Network Design**

# Overview

① Definition
② Background
③ Performance Areas
④ Developing Goals for Performance
⑤ Performance Mechanisms
⑥ Quality of Service
⑦ Resource control
⑧ Service Level Agreements
⑨ Architectural Considerations

# Definitions

- Performance is the set of levels for capacity, delay, reliability, maintainability and availability (defined in RMA)
- May consider
  - All traffic flows in the network combined
  - One or more sets of traffic flows, based on users, applications and devices

# Definitions

- A performance architecture is
  - The set of performance mechanisms to
    - ✓ Configure
    - ✓ Operate
    - ✓ Manage
    - ✓ Provision
    - ✓ Account for
  - resources in the network that support traffic flows

# Background

To be able to say we have a network that performs well we need to have measurable goals:

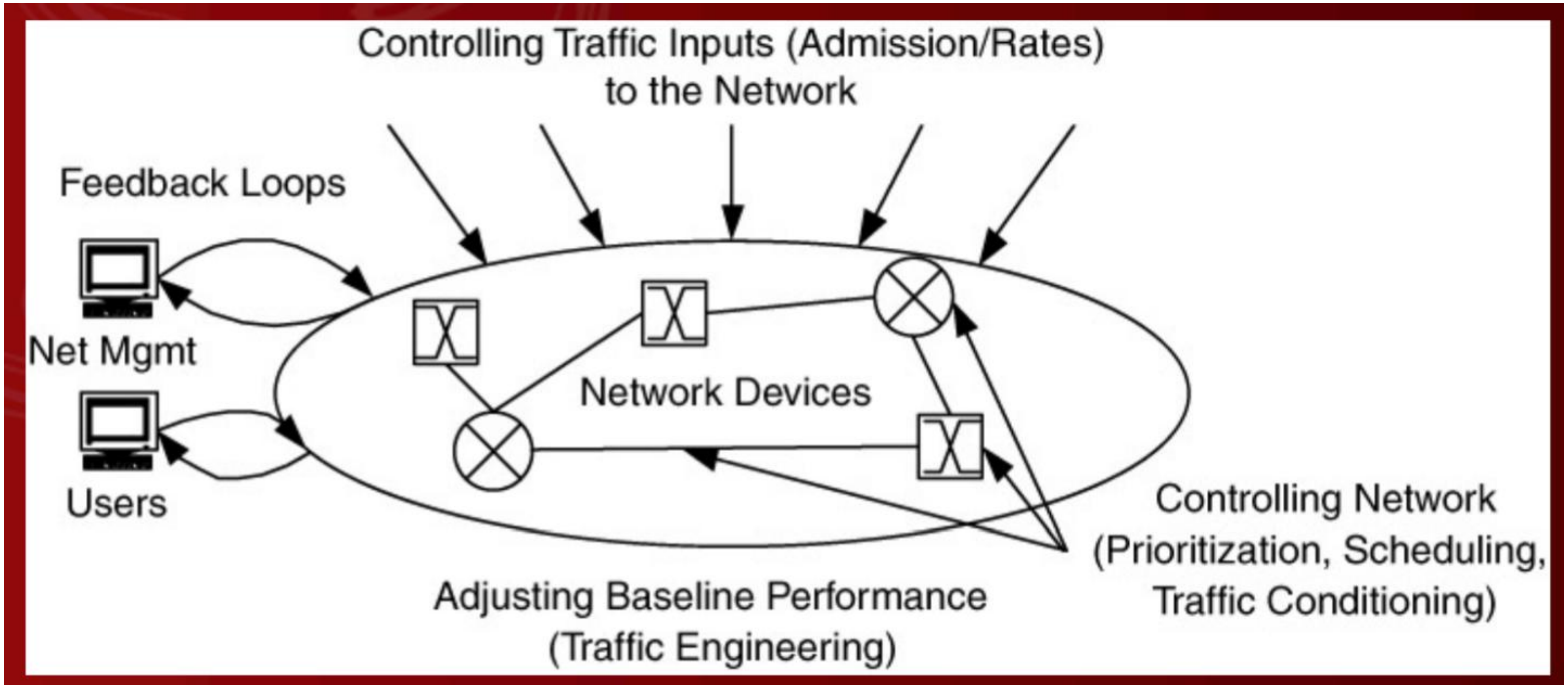
- Improve overall network performance
  - Response times, throughput improved for all users
- Support a particular group of users or application
- Control resource allocation for accounting billing and/or management

# Performance Areas

Performance achievement consists of one or more of the following:

- Controlling traffic inputs
  - ✓Admission and rate control
- Adjusting the baseline performance
  - ✓Traffic or capacity planning
- Controlling all or part of the network for delivery of specific services
  - ✓Prioritizing, scheduling and conditioning traffic flow

# Performance Areas



# Developing Goals for Performance

- The process of developing performance goals begins during the requirements phase
  - The requirements and flow specifications and maps provide inputs
- Although performance is always desirable
  - Need to ensure that the performance mechanisms are both necessary & sufficient to meet the goals set

# Developing Goals for Performance

- **Single-tier** performance where capacity, delay and RMA are optimised for all traffic flows
- **Multi-tier** performance where capacity, delay and RMA are optimised for one or more groups of flows

# Developing Goals for Performance

- A defined level of performance in a data communication system
  - ✓ e.g. to ensure that real-time voice and video are delivered without annoying blips,
  - ✓ a guarantee of bandwidth is required.
- When data is broken into packets that travel through the routers, LANs and WAN with all other data, QoS mechanisms are one way to give real time data priority over non-real time data.
- The only other option is to overbuild the network so there is always sufficient bandwidth.

# Developing Goals for Performance

- In developing performance goals the following questions should be answered
  - ✓Are performance mechanisms necessary for this network?
  - ✓What are we trying to solve, add, or differentiate by adding performance mechanisms to this network?
  - ✓Are performance mechanisms sufficient for this network?

# Developing Goals for Performance

- We should avoid implementing performance mechanisms just because they are interesting or new
  - ✓There should be a clear need and want by customers
  - ✓Performance mechanisms need to be constantly maintained

# Developing Goals for Performance

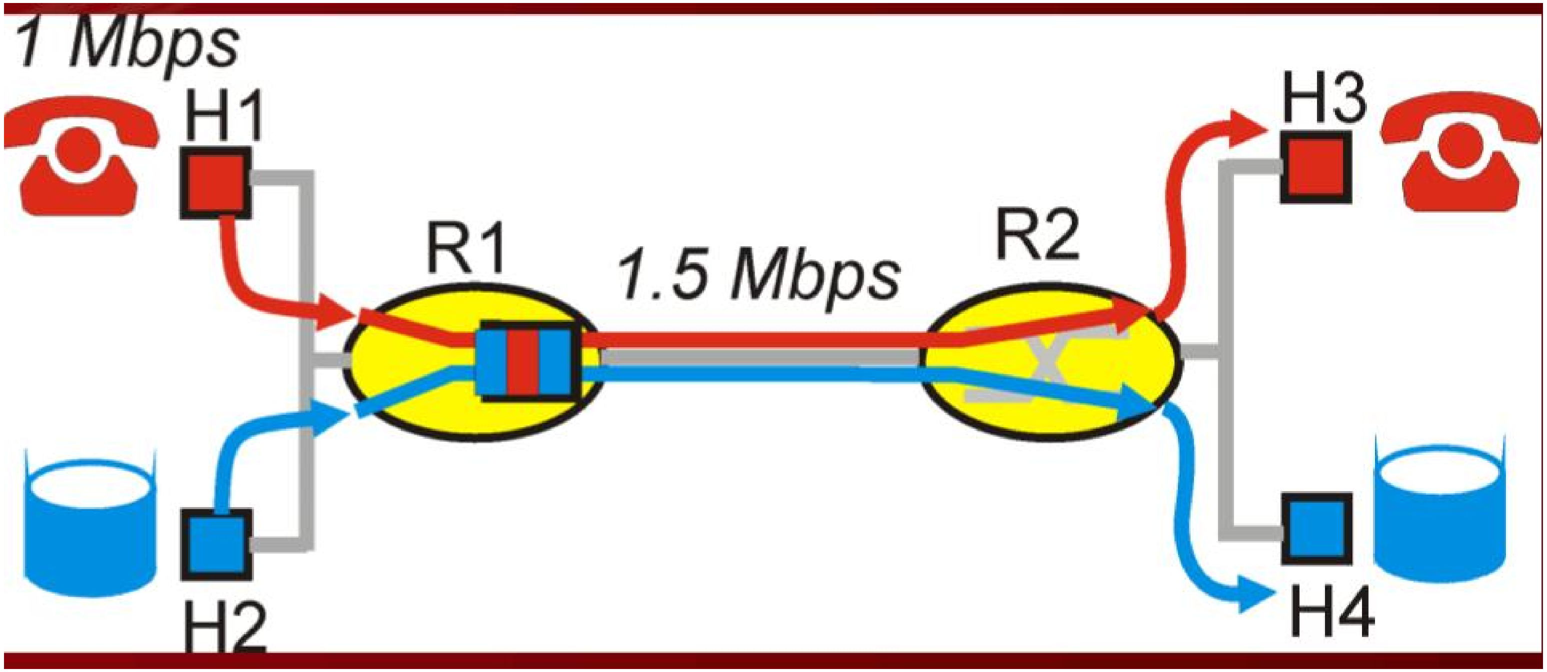
- Start simple and work towards a more complex architecture
  - ✓Implement in selected areas of the network
  - ✓Use only a few mechanisms
  - ✓Select mechanisms that are easy to implement, operate and maintain

# Developing Goals for Performance

- A well designed performance architecture will address the following
  - ✓Improving overall network performance
  - ✓Improving the performance for select users, applications and/or devices
  - ✓Changing the network from a cost centre to profitability
  - ✓Merging multiple traffic types over a common network infrastructure
  - ✓Differentiating customers for multiple levels of service

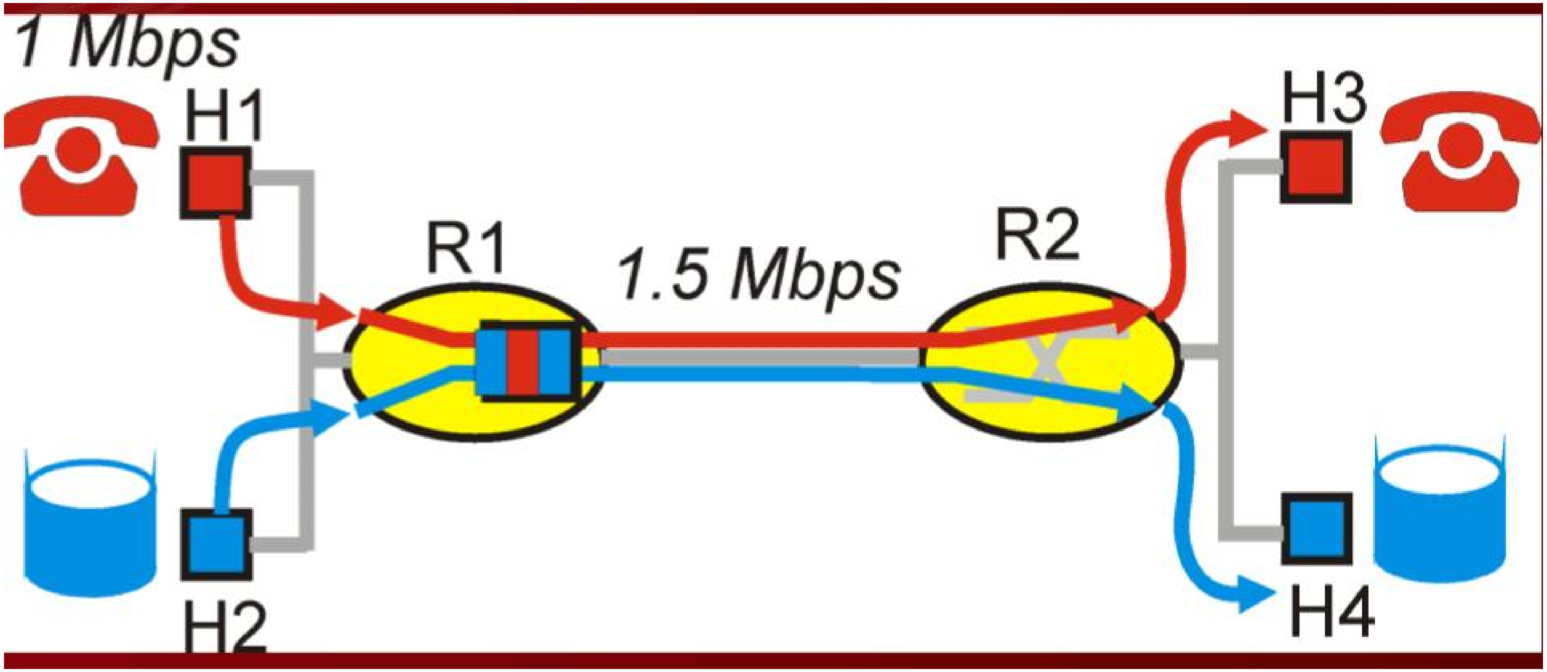
# Developing Goals for Performance

- Example: 1Mbps IP phone, FTP share 1.5 Mbps link.
  - bursts of FTP can congest router, cause audio loss
  - want to give priority to audio over FTP



# Developing Goals for Performance

- BASIC FACT OF LIFE: the network can not support traffic demands beyond link capacity



# Performance Mechanisms

- Quality of Service (QoS) Mechanisms
- Resource control
  - ✓Prioritization
  - ✓Traffic management
  - ✓Scheduling
  - ✓Queuing
- Service Level Agreements (SLAs)
- Policies

# Quality of Service

- **Quality of service (QoS)** is the description or measurement of the overall performance of a service, such as a telephony or computer network, or a cloud computing service, particularly the performance seen by the users of the network.
- To quantitatively measure quality of service, several related aspects of the network service are often considered, such as:
  - ✓ packet loss
  - ✓ bit rate
  - ✓ throughput
  - ✓ transmission delay
  - ✓ availability
  - ✓ jitter

# Quality of Service

- In the field of computer networking, quality of service refers to traffic prioritization and resource reservation control mechanisms rather than the achieved service quality.
- Quality of service is the ability to provide different priorities to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow.

# Quality of Service

## Why is QoS important?

- Without QoS, network data can become disorganized, clogging networks to the point where performance degrades or, in certain cases, the network shuts down completely.
- QoS is important because enterprises need to provide stable services for employees and customers to use. QoS determines quality of experience (QoE). If the services an organization provides are not reliable, customer and employee relationships may be put at risk.
- Additionally, data integrity and security are more likely to be compromised in a company with a poor QoS. In general, employees and customers depend on communication services to do their jobs. When QoS suffers, quality of work and QoE suffer as well.

# Quality of Service

## How does QoS work?

- When organizations use their networks to send information back and forth between endpoints on the network, the information or data is formatted into packets. Packets are the way computers organize information to be transferred over a network.
- QoS takes on the responsibility of prioritizing packets to get the most out of the finite amount of bandwidth on their network. In other words, the network can only transport so much information in a certain amount of time. QoS prioritizes packets in a way that ensures that bandwidth is used to provide the best internet service possible in that fixed amount of time.

# Quality of Service

## How does QoS work? One example

- For example, packets pertaining to a video call would be prioritized over packets pertaining to an email download. This is because a video call is a more synchronous form of communication than an email -- video needs to happen in real time, whereas emailing is not necessarily time sensitive.
- If a packet is dropped or delayed during a video chat, the end user may experience jitter or latency in the chat. If packets are dropped or delayed in the emailing process, they can still be sent after, and the end user will not experience any lapse in service.

# Quality of Service

**How does QoS work?**

- QoS looks at packet headers to prioritize packets. Packet headers are bits of information that tell QoS and other network components what the packet contains, where it is going (the IP address of its destination) and what it will be used for.
- QoS can also read the packet header, determine a packet is related to video streaming and prioritize it over packets that are less time sensitive. QoS can alter a portion of the packet header to specify priority.

# Quality of Service

## Benefits of QoS?

① Ensures the availability of an organization's network and the applications that run on that network.

② Provides the safe, efficient transfer of data over that network.

③ Allows organizations to use their existing bandwidth more efficiently, instead of upgrading network infrastructure to expand bandwidth.

④ There several specific benefits:

- Mission-critical applications have access to the resources they require.
- Administrators can manage traffic better.
- User experience is improved.

# Quality of Service

## How to implement QoS?

① **Planning.** The organization should gain an understanding of each department's service needs and requirements, choose a fitting model and cultivate buy-in from stakeholders.

② **Design.** The organization should then take note of all significant software and hardware changes and apply the chosen QoS model to the specifics of its network architecture.

③ **Testing.** The organization should test QoS settings and policies in a safe, controlled testing environment where bugs can be worked out.

④ **Deployment.** Policies should be rolled out in phases. An organization may choose to roll out policies by network segment or by separate QoS function.

⑤ **Monitoring and analysis.** Policies should be adjusted to improve performance according to performance data.

# Quality of Service

## Three models to implement QoS

① **Best Effort.** A QoS model where all the packets receive the same priority and there is no guaranteed delivery of packets. Best Effort is applied when networks have not configured QoS policies or when the infrastructure does not support QoS.

② **Integrated Services (IntServ).** A QoS model that reserves bandwidth along a specific path on the network. Applications ask the network for resource reservation, and network devices monitor the flow of packets to make sure network resources can accept the packets.

③ **Differentiated Services (DiffServ).** A QoS model where network elements, such as routers and switches, are configured to service multiple classes of traffic with different priorities. Network traffic must be divided into classes based on a company's configuration.

# Quality of Service

- Determining, setting and acting on priority levels for traffic flows
- Includes
  - IP QoS
  - ~~Type of Service (ToS)~~
  - ~~ATM Class of Service (CoS)~~
  - ~~Frame Relay Committed Information Rate (CIR)~~

# Quality of Service

- **Focus on IP QoS**
  - DiffServ (Differentiated Services)
    - Aggregation of traffic flows on a per-hop basis
    - Good scalability
    - Best-effort, Assured forwarding, Expedited forwarding
  - IntServ (Integrated Services)
    - End-to-end support for individual flows

# Quality of Service

- A bit of history…
- ATM (Asynchronous Transfer Mode) was one of the first packet technologies to have “built-in” modes of service.
- As everything was built around IP, a lot of work has been done to “retro-fit” QoS mechanisms into IP like:
  - IntServ
    - RSVP (Resource Reservation Protocol)
  - DiffServ

# Quality of Service

- MPLS is now the QoS mechanism of choice for telecomm providers and is making its presence felt in corporate environments
  - ✓MPLS – Multiprotocol label Switching
  - ✓MPLS is an ethernet based technology
  - ✓MPLS is a routing technique in telecommunications networks that directs data from one node to the next based on labels rather than network addresses

# Quality of Service – IntServ

- **IntServ** is an architecture that specifies the elements to guarantee QoS on networks.
- IntServ can for example be used to allow video and sound to reach the receiver without interruption.
- IntServ specifies a fine-grained QoS system, which is often contrasted with DiffServ's coarse-grained control system.
- Under IntServ, every router in the system implements IntServ, and every application that requires some kind of QoS guarantee has to make an individual reservation.
- Flow specs describe what the reservation is for, while RSVP is the underlying mechanism to signal it across the network.

# Quality of Service – IntServ

- Sometimes traffic flows need to be treated individually
  - IntServ defines value and mechanisms for allocating resources across end-to-end paths.
  - Works in tandem with RSVP – Resource ReSerVation Protocol (RFC 2205)
  - Source routers request capacity via a RESV request. Once other routers have agreed to this request they must carry the traffic.

# Quality of Service – IntServ

- What is RSVP--Resource ReSerVation Protocol?
  - ✓ RSVP is a transport layer protocol that is used to reserve resources in a computer network to get different QoS while accessing Internet applications. It operates over IP and initiates resource reservations from the receiver's end.
- Features of RSVP?
  - ✓ RSVP is a receiver oriented signalling protocol. The receiver initiates and maintains resource reservation.
  - ✓ It is used both for unicasting (sending data from one source to one destination) and multicasting (sending data simultaneously to a group of destination computers).
  - ✓ RSVP supports dynamic automatic adaptation to changes in network.
  - ✓ It provides a number of reservation styles. It also provides support for addition of future styles.

# ReSerVation Protocol

- A communications protocol that signals a router to reserve bandwidth for realtime transmission.
- RSVP is designed to clear a path for high priority traffic
- Works at Layer Transport Layer – is alike to ICMP and IGMP

# DiffServ

- Differentiated services or DiffServ is a computer networking architecture that specifies a mechanism for classifying and managing network traffic and providing QoS on modern IP networks.
- DiffServ can, for example, be used to provide low-latency to critical network traffic such as voice or streaming media while providing best-effort service to non-critical services such as web traffic or file transfers.
- DiffServ uses a 6-bit differentiated services code point (DSCP) in the 6-bit differentiated services field (DS field) in the IP header for packet classification purposes. The DS field, together with the ECN field, replaces the outdated IPv4 TOS field.

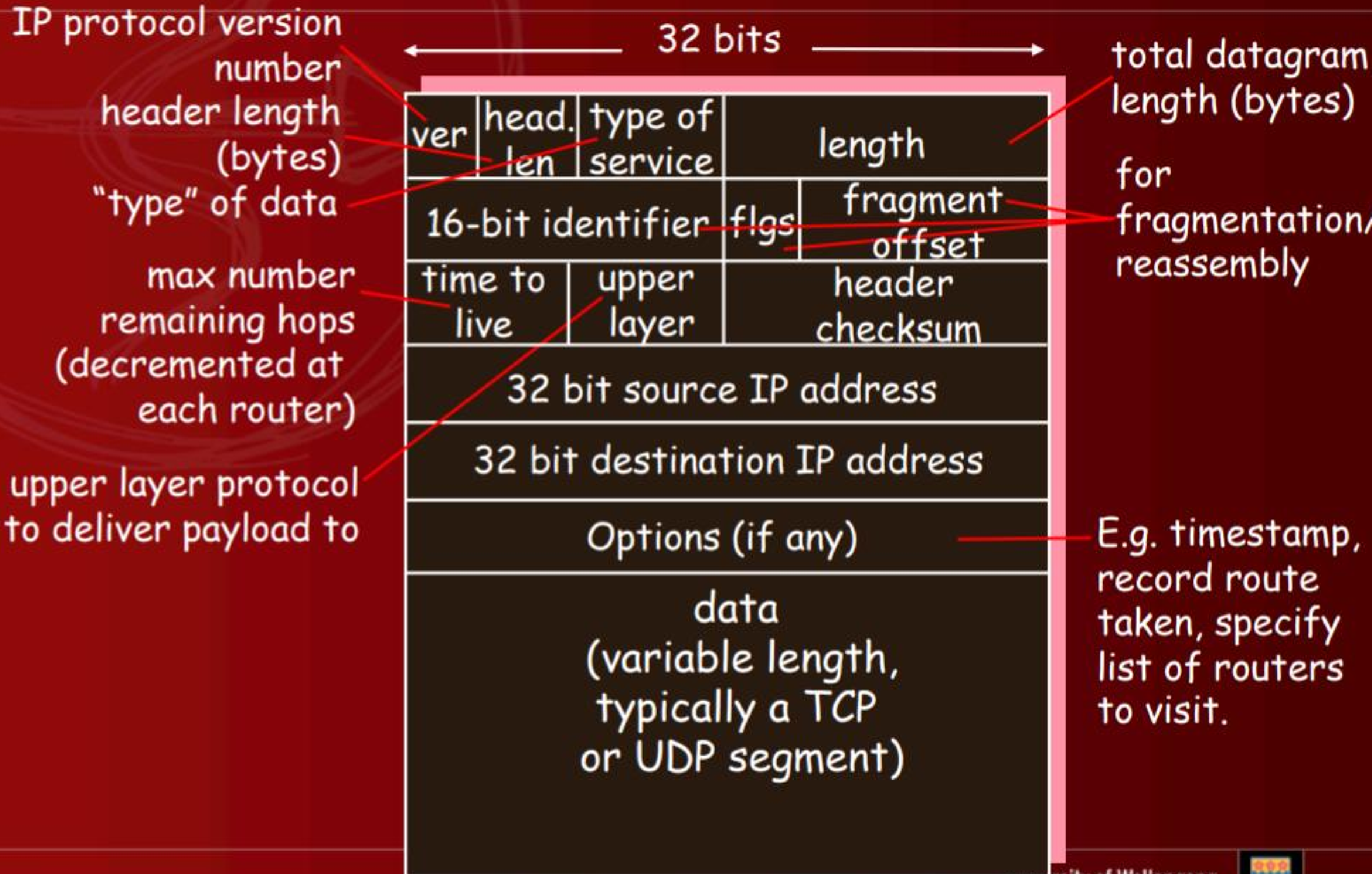
# DiffServ

- While IntServ is a “fine-grained” QoS system, DiffServ's is a “coarse-grained” control system.
  - IntServ – individual flows are treated specially
  - DiffServ – flows are aggregated into common classes and then treated specially

# DiffServ

- Can be invoked when IP packets are marked in the Type of Service (ToS) byte for IPv4 (or traffic class in IPv6).
- DiffServ enables three (or more) classes of flow to be handled
  - Best effort
  - Assured Forwarding (AF)
  - Expedited Forwarding (EF)

# IPv4 datagram format



# DiffServ

- Second generation DiffServ uses the Differentiated Services Code Point (DSCP) in the ToS byte to specify precedence “class”.
- Benefits of implement DiffServ
  - ✓ Reduces the burden on network devices and easily scales as the network grows.
  - ✓ Allows customers to keep any existing Layer 3 ToS prioritization scheme that may be in use.
  - ✓ Allows customers to mix DiffServ-compliant devices with any existing ToS-enabled equipment in use.
  - ✓ Alleviates bottlenecks through efficient management of current corporate network resources.

# DiffServ – RFC 2745

- DiffServ operates over a DiffServ domain that is defined by routers designated as “boundary nodes”
  - (RFC 2745 2.1) – next slide for further reading.

# DiffServ - RFC 2745 2.1

**Differentiated Services Domain**

- A DS domain is a contiguous set of DS nodes (i.e. routers) which operate with a common service provisioning policy and set of PHB (per-hop-behavior) groups implemented on each node.
- A DS domain has a well-defined boundary consisting of DS boundary nodes which classify and possibly condition ingress traffic to ensure that packets which transit the domain are appropriately marked to select a PHB from one of the PHB groups supported within the domain.
- Nodes within the DS domain select the forwarding behavior for packets based on their DS codepoint, mapping that value to one of the supported PHBs using either the recommended codepoint->PHB mapping or a locally customized mapping [DSFIELD].

# DiffServ - RFC 2745 2.1

**Differentiated Services Domain**

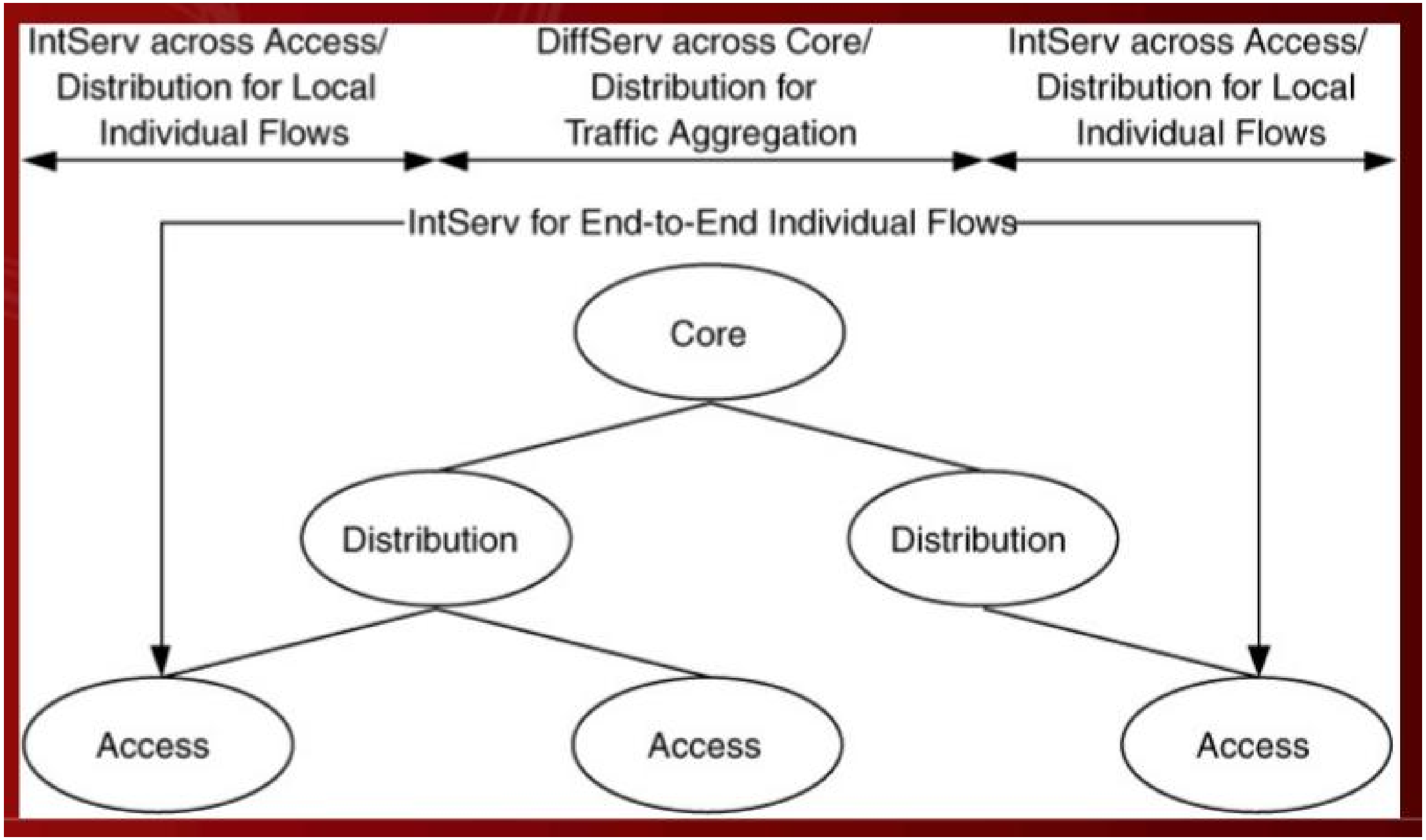
- Inclusion of non-DS-compliant nodes within a DS domain may result in unpredictable performance and may impede the ability to satisfy service level agreements (SLAs).
- A DS domain normally consists of one or more networks under the same administration; for example, an organization's intranet or an ISP. The administration of the domain is responsible for ensuring that adequate resources are provisioned and/or reserved to support the SLAs offered by the domain.

# DiffServ

- It is expected that service providers will use MPLS within their networks and use DiffServ at the edges of the network for classification and assignment to the right connection

# QoS – before MPLS

- IntServ is best implemented in resources the network administrator has direct control over.



# QoS

- IntServ (RSVP) does not scale well and is not commonly found in corporate networks
- DiffServ works on all routers
- RSVP has been re-purposed by the IETF to work on MPLS (see link on RSVP-TE)

# Quality of Service

- Practical advice on implementation

| Function/ feature | DiffServ | IntServ |
|---|---|---|
| Scalability | Scalable to large enterprises | Limit to small or medium enterprise network |
| Granularity of control | Traffic aggregated into classes | Per-flow or group of flows |
| Scope of control | Per hop | All network devices in end-to-end path or flow |

# MultiProtocol Label Switching

- MPLS is used to ensure that all packets in a particular flow take the same route over a backbone.
- Deployed by many telcos and service providers, MPLS can deliver the QoS required to support realtime voice and video as well as service level agreements (SLAs) that guarantee bandwidth. (i.e. able to deliver hard service metrics like ATM)
- Large enterprises may also use MPLS in their national networks

# MultiProtocol Label Switching

- A MPLS header is attached to a IP packet by a Label Edge Router (LER)
- The entry and exit points of an MPLS network are defined by LERs.
- Routers in between are called Label Switch Routers (LSR) which use the MPLS header to route packets
- VLSI technology (i.e. hardware as opposed to CPU and Router lookup tables) use the MPLS header to make routing decisions.

# Performance Mechanisms

- Quality of Service (QoS) Mechanisms
- Resource control
  ① Prioritization
  ② Traffic management
  ③ Scheduling
  ④ Queuing
- Service Level Agreements (SLAs)
- Policies

# Resource control

- Prioritization
  - A ranking based upon importance and urgency relative to other network entities
  - Users, applications, devices, flows and connections
  - Priorities should be determined during the requirements and flow analysis phases
  - May be based on
    - ✓ Type of protocol (TCP vs UDP)
    - ✓ Port number

# Resource control - conditioning

- Traffic conditioning
- Set of mechanisms that increase or decrease performance to traffic flows
- Taken in order of nominal traffic flow across the network, traffic conditioning consists of:
  - 1. Classification
  - 2. Metering
  - 3. Shaping
  - 4. Dropping

# Resource control

- Traffic management
  - Admission control
    - Ability to refuse access
    - Uses priority level to change access behavior
  - Traffic conditioning
    - Modify performance to traffic flows
    - Classification
    - Metering
      - Temporal characteristics of traffic flow
    - Shaping
      - Delaying non-conforming traffic to match performance characteristics
    - Dropping

- Classification
  - Packets can be marked with Differentiated Services Code Points (DSCPs) for Best Effort, Assured Forwarding or Expedited Forwarding
- Metering
  - Temporal characteristics of traffic flow (traffic rates or burst sizes) are measured
- Metering is performed in routers.
- This information is compared with the traffic profile detailed in the SLA

# Resource control

- Shaping (example)
  - Consider an SLA that specifies a peak rate of 1.5Mb/s
  - A meter is measuring that traffic and calculates a rate of
    - (200 packets/sec) x (1500 byte/pkt)=2.4 Mbits/sec
  - This non-conforming flow is subsequently forwarded to a shaper queue where packets are delayed by an average of 10ms ➔ 100 pkts/sec
  - ➔ (100 pkts/sec) x (1500 byte/pkt) = 1.2 Mbits/sec
  - Shaping continues for either a set period or until flow conforms
- Dropping – Packets are discarded

# Resource control

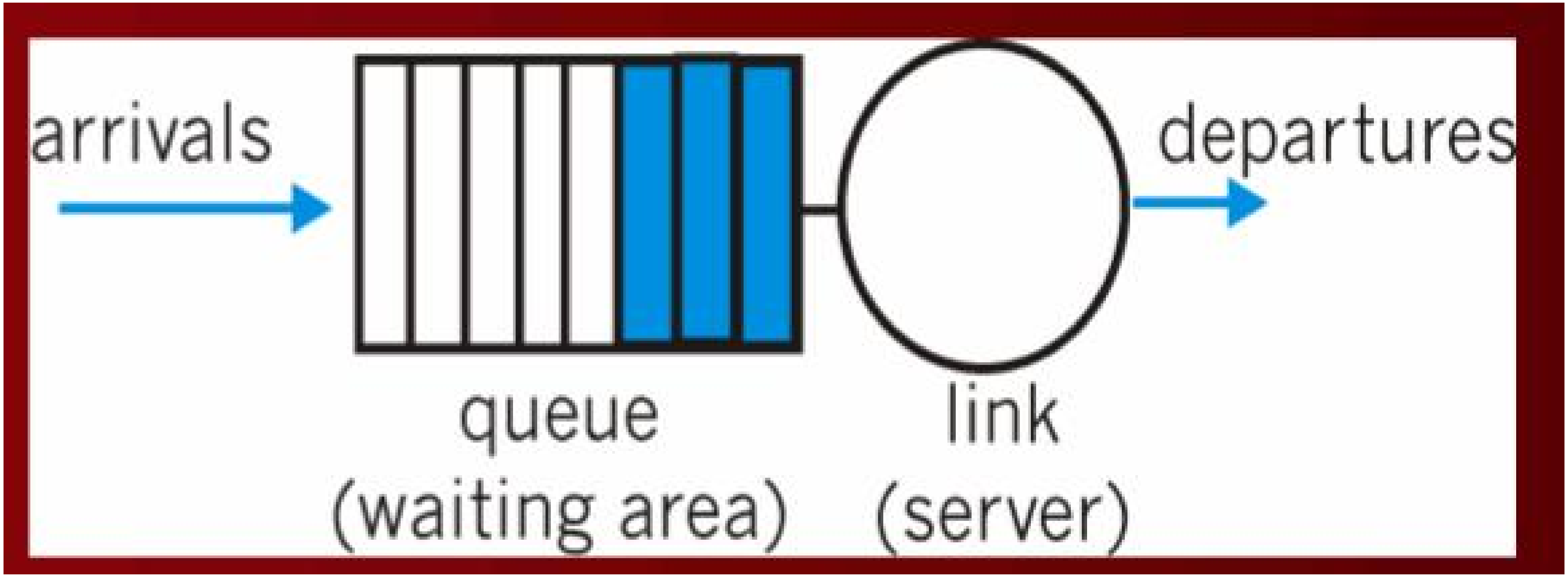
- Scheduling
  - Determines the order in which traffic is processed for transmission
  - Applied throughout a network

# Resource control

- Queuing
  - Storing packets (cell/frames etc) while they wait for processing
  - May use one of the following queue mechanisms
    - First In First Out (FIFO)
    - Class Based Queuing (CBQ)
    - Weighted Fair Queuing (WFQ)
    - Random Early Detect (RED)
    - Weighted RED (WRED)

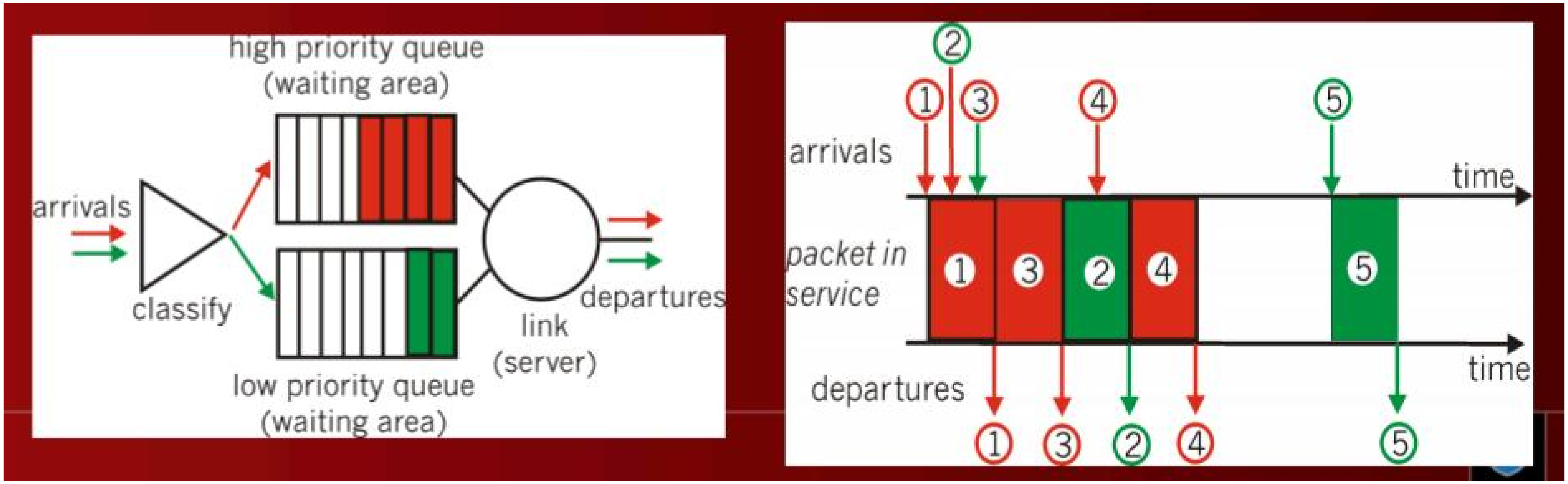
# Resource control

- FIFO (first in first out) scheduling: send in order of arrival to queue



# Resource control

- Priority scheduling: transmit highest priority queued packet
- Multiple classes, with different priorities
  - class may depend on marking or other header info, e.g. IP source/dest, port numbers, etc.



# Scheduling Policies: still more

- Round-robin scheduling:
  - multiple classes
  - cyclically scan class queues, serving one from each class (if available)

# Scheduling Policies: still more

- Round-robin scheduling:
  - Round-robin (RR) is one of the algorithms employed by process and network schedulers in computing.
  - Time slices (also known as time quanta) are assigned to each process in equal portions and in circular order, handling all processes without priority (also known as cyclic executive).
  - RR scheduling is simple, easy to implement, and starvation-free. RR scheduling can be applied to other scheduling problems, such as data packet scheduling in computer networks.
  - The name of the algorithm comes from the round-robin principle known from other fields, where each person takes an equal share of something in turn.

# Scheduling Policies: still more

- Round-robin scheduling:
  - In best-effort packet switching and other statistical multiplexing, RR scheduling can be used as an **alternative** to first-come first-served queuing.
  - A multiplexer, switch, or router that provides RR scheduling has a separate queue for every data flow, where a data flow may be identified by its source and destination address. The algorithm allows every active data flow that has data packets in the queue to take turns in transferring packets on a shared channel in a periodically repeated order.
  - The RR scheduling is work-conserving, meaning that if one flow is out of packets, the next data flow will take its place. Hence, the scheduling tries to prevent link resources from going unused.

# Scheduling Policies: still more

- Round-robin scheduling:
  - Round-robin scheduling results in max-min fairness if the data packets are equally sized, since the data flow that has waited the longest time is given scheduling priority.
  - It may not be desirable if the size of the data packets varies widely from one job to another.
  - A user that produces large packets would be favored over other users.
  - If guaranteed or differentiated QoS is offered, and not only best-effort communication, deficit round-robin (DRR) scheduling, weighted round-robin (WRR) scheduling, or weighted fair queuing (WFQ) may be considered.

# Scheduling Policies: still more

- Weighted Fair Queuing:
  - generalized Round Robin
  - each class gets weighted amount of service in each cycle

# Scheduling Policies: still more

- Weighted Fair Queuing:
  - Weighted Fair Queueing (WFQ) is a scheduling algorithm used in computer networks to manage traffic in a fair and efficient manner.
  - The primary goal of WFQ is to allocate bandwidth to different traffic flows based on weights, ensuring that high-priority flows or those with more significant bandwidth needs get more resources while maintaining fairness.
  - WFQ is often used in the context of packet-switched networks, where packets from different flows (i.e., sources or destinations) are transmitted over shared network links.
  - The algorithm works by simulating an idealized system in which packets are processed in the order of their "**virtual finish times**", which are determined by their weights and the arrival times of the packets.

# Scheduling Policies: still more

- Key concepts of Weighted Fair Queuing:
  - **Flow**: A flow is a stream of packets from a specific source to a specific destination (often identified by a flow ID or combination of source/destination IP and port).
  - **Weight**: Each flow is assigned a weight that dictates the proportion of bandwidth it should receive compared to other flows. A flow with a higher weight will receive more bandwidth.
  - **Virtual Time**: The virtual time is used to simulate when each packet would finish being transmitted in an idealized system. It is a key factor in determining the packet's position in the queue.

# Scheduling Policies: still more

- Characteristics of Weighted Fair Queuing:
  - **Fairness**: It ensures fair allocation of bandwidth based on weights. If all flows have equal weights, the algorithm guarantees equal bandwidth distribution.
  - **Priority**: Flows with higher weights get a proportionally larger share of the bandwidth.
  - **Real-time Allocation**: WFQ effectively handles real-time traffic and can prioritize delay-sensitive packets while maintaining fairness for other flows.
  - **Complexity**: WFQ is computationally expensive because it requires maintaining and sorting virtual finish times for each packet in the queue. However, it's more efficient than simpler methods like First-Come-First-Served (FCFS) in achieving fair allocation.

# Service Level Agreements

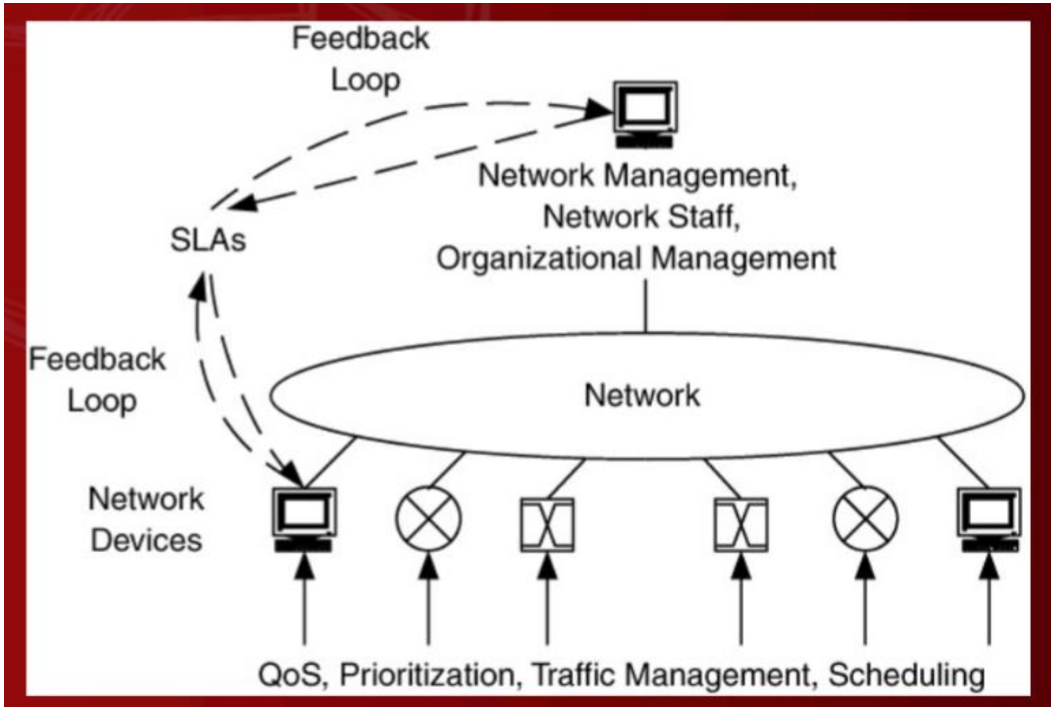
- Typically formal contracts between a provider and a user
- Define the provider's responsibility to user and extent of accountability
- SLA performance elements may include
  - Data rate
  - Burst tolerance
  - Upstream/downstream
  - Delay and RMA metrics
- Can be further defined as upstream and down stream

# Service Level Agreements

| Network Service Description for My Enterprise | | | |
|---|---|---|---|
| **Service Levels:** | **Capacity Performance** | **Delay Performance** | **Reliability Performance** |
| Basic Service | As Available (Best Effort) | As Available (Best Effort) | As Available (Best Effort) |
| Silver Service | 1.5 Mb/s (Bidirectional) | As Available (Best Effort) | As Available (Best Effort) |
| Gold Service | 10 Mb/s (Bidirectional) (Burst to 100 Mb/s) | Max 100-ms Round-Trip (Between Specified Points) | As Available (Best Effort) |
| Platinum Service | 100/10 Mb/s Up/Down (Burst to 1 Gb/s) | Max 40-ms Round-Trip (Between Specified Points) | 99.999% Uptime (User-Server) |

Figure 8.9: Example of enterprise SLA.

- If used correctly, SLAs can form an important link in communicating network performance between users, staff and management



# Policies

- Policies complete the framework for performance for a network by coupling the high level management view of how the network should perform with mechanisms to implement at the network devices (QoS) and feedback loops using SLAs.

# Policies

- Policies describe
  ① what network computing storage or other resources are available to users,
  ② when these resources are available or
  ③ which users are permitted access to certain resources.
- Share some similar attributes to routing and security policies

# Architectural Considerations

- Performance mechanisms should be evaluated against
  1. Requirements of the network
  2. Goals
  3. Type of environment

# Architectural Considerations

- Performance mechanisms that operate on
  - Individual flows should be considered more when traffic flows are more likely to be individual (access networks)
  - Aggregate flows should be considered where traffic flows are more likely to aggregate (interfaces to external networks)

# Architectural Considerations

- Internal Relationships
  - Trade-offs between end-to-end v per-hop scheduling and conditioning of traffic flows,
  - are flows treated individually or aggregated into groups?
- These factors will determine the kinds of performance mechanisms you impose

# Architectural Considerations

- Internal (or External?) Relationships - NW management
  - Once we decide on policies and SLAs this will determine placement of databases for SLAs, including: edge devices
  - Policy Decision points (PDPs); Policy Enforcement Points (PEPs), DiffServ Edge devices

# Architectural Considerations

- External Relationships
  - Performance and Addressing/Routing
    - Closely coupled when performance relies on protocols such as DiffServ or MPLS
    - Access networks and switches aim to provide maximum bandwidth and minimal delay before they need to be routed elsewhere
  - Performance and Network Management (NM)
    - Depends on Network Management to configure, monitor, manage, verify and bill for performance
    - NM links QoS with SLAs and Policies

# Architectural Considerations

- External Relationships
- Performance and Security
  - Security can be the enemy of performance because it
    - intercepts, inspects and controls network access
    - Results in reduced capacity, delays.
    - Capacity can be compensated for but delay is more difficult

# Readings

- This week
  - Chapter 8 Performance Architectures, McCabe
- Additional material
  - *Internet QoS: Architectures and Mechanisms for Quality of Service,* Zheng Wang, Morgan Kaufman, 2001
  - *Queuing Systems Theory:* Volume 1, Leonard Kleinrock, John Wiley and Sons, 1975
  - *RFC's 2474, 2475* [www.ietf.org](http://www.ietf.org)