Bob's QOS Overview

An overview of QOS Concepts & Technology

www.TrainingMagic.com
(800) 531-6573

QoS – Objectives

- Define QoS
- Distinguish between QoS and CoS
- Explain the QoS techniques used by Frame Relay and ATM
- Describe the ATMF service classifications
- List the IntServ suite of protocols
- Name 4 queuing algorithms
- Distinguish between RED and WRED

Quality of Service

- **What is Quality of Service?**
- **Many definitions**
  - **Bandwidth**
  - **Delay**
  - **Variable Delay**
  - **Packet Loss**
  - **Reliability**
  - **Security**

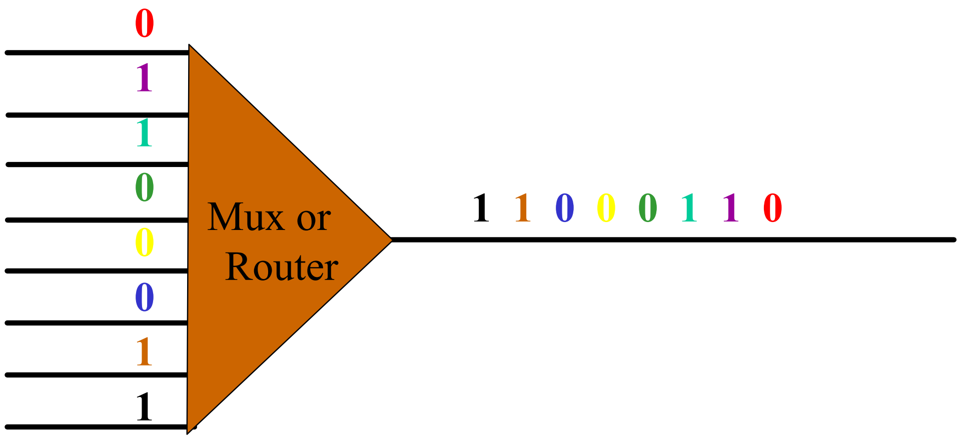
Quality of Service

QoS is defined as providing a satisfactory user experience within a company's Web environment (response time/navigation ease). QoS is measured by the availability and performance of systems and networks supporting end-user applications. Although there is no consensus as to what the metrics of QoS are, most definitions include at least some of the following: bandwidth, delay, variable delay, packet loss, reliability, and security.

Why Quality of Service?

- **Variable delay is the most frequently cited Quality of Service variable**
- **Applications sensitive to variable delay:**
  - **VoIP**
  - **Streaming audio/video**
  - **Video conferencing**
  - **Internet gaming**

Serialization Delay

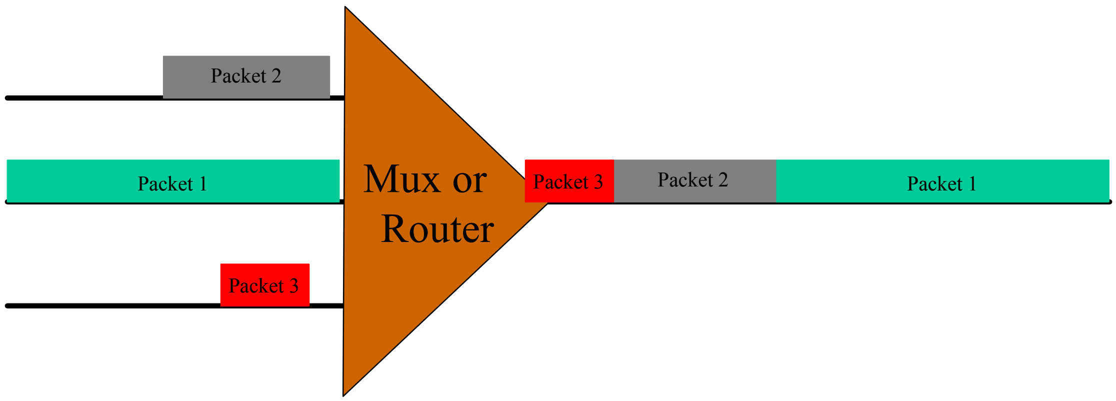


Serialization delay is due to the delay converting the parallel bit stream used inside of equipment to the serial form used to interconnect equipment.

Serialization Delay

Serialization delay is the time it takes to actually transmit the packet. Depending on bit-rate it may be a significant portion of overall delay (for a modem user connected at 33600 baud upstream it would take about 450ms to transmit a 1500 bytes packet [each byte actually takes 10 bits to transmit counting start and stop bits; on the other hand modems have hardware compression that reduces serialization delay somewhat]), or may be negligible (it only takes 3us to transmit a 4000 bytes packet on a 100Mbps connection). In ARPANET days serialization delay was a significant delay component, since there were typically several low bit-rate links between remote hosts. In modern world where high bit-rate connections are becoming the norm serialization delay is becoming more and more irrelevant. In any case, for given packet size and path, serialization delay is constant (this may not be completely true due to hardware compression).

Queuing Delay



Queuing delay is caused by packets waiting for availability of the output interface.

Queuing Delay

Queuing delay is the time a packet spends in router or multiplexer queues. In the case of a multiplexer, this delay tends to be fixed. However, the delay time spent n routers depends naturally on queue lengths: for unloaded network it would be negligible; for network that is heavily congested it is usually the main delay component. It is the most variable delay component in a typical modern network.

Another type of delay not mentioned in the slide is Propagation delay. Propagation delay is the time it takes the physical signal to traverse the path. Although This delay is usually fairly constant if there are no route changes

Solving the QoS Problem

- **Four current approaches**
  - **Reduce delay by faster processing**
  - **Reduce delay by prioritizing sensitive data**
  - **Equalize delay by delaying faster traffic**
  - **Infinite Bandwidth**
- **Four example technologies**
  - **Frame Relay**
  - **ATM**
  - **RSVP**
  - **DWDM/Gigabit Ethernet**

Solving the QoS Problem

A great deal of work has been done on the issue Quality of Service. One approach that reduces the variable delay experienced by a network packet is to speed up the processing of the routers/switches in a network. Both frame relay, and wire line speed routers, reduce the ammount of processing time necessary to move a packet through the node. Another category of solutions involve giving some packets priority in the data based in the type of data being processed. In other words, if a queue contains both voice and HTML packets, then the node will process the voice data first. An older approach to equalizing delay between various packets is to SLOW DOWN faster packets until the slower packets catch up. This approach is most often used in various inverse-multiplexing applications. Recently, many have put forth the theory that if you throw enough bandwidth at an application (effectively infinite bandwidth) that there would be no variable delay.

Frame Relay Characteristics

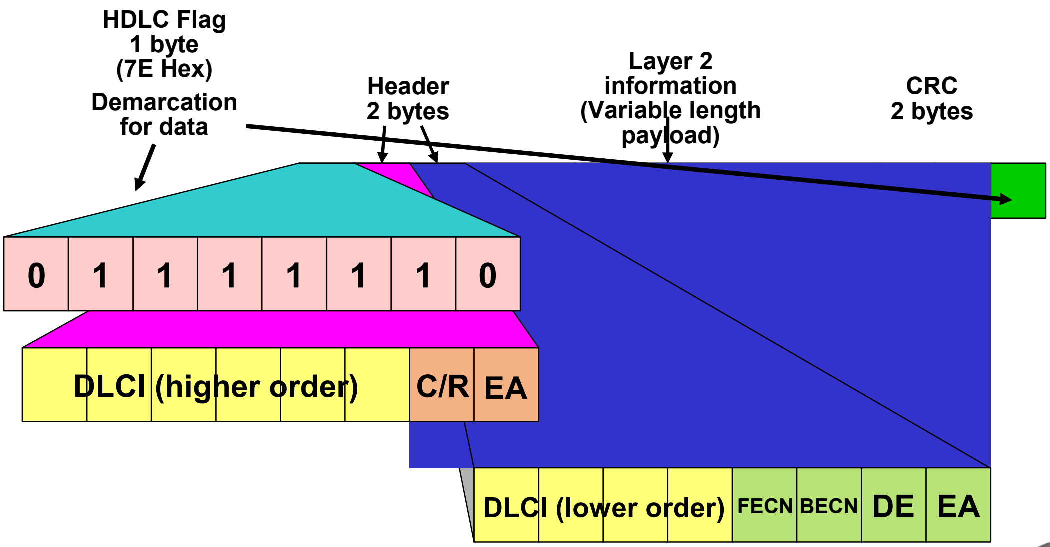
- **Earliest fast packet technology**
- **Reduces delay**
  – **Move routing to L2 (hardware processing vs. software)**
  – **No error correction**
  – **Reduced Processing**
- **Still subject to delay**
  – **Variable packet size**
  – **Queuing delay (congestion)**
- **No QoS labels/Prioritization**

Frame Relay Characteristics

Frame relay uses a variable-length framing structure, which, depending on user data, ranges from a few to more than a thousand characters.

Why frame relay is faster previous versions of packet switching? Frame relay eliminates Level 3, thereby significantly reducing processing overhead. In fact, a frame switch in a frame relay network is capable of beginning retransmission of a frame to its next destination before the entire frame has even arrived. Also, Frame Relay makes no attempt at error correction. However, Frame Relay is not a panacea. Queuing delay can still be a problem for Frame Relay. Small frames can wait for relatively long time periods while larger frames are being transmitted. Also, when too much traffic is sent to a node the delay to exit that node can increase (this condition is known as congestion.

Frame Relay Data Structure



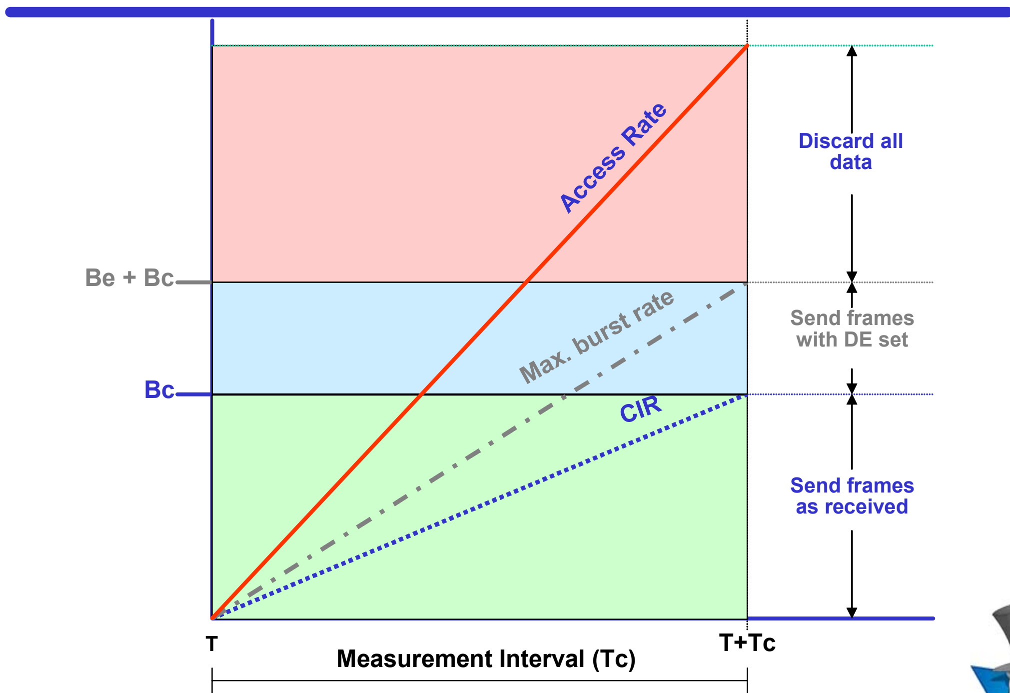
FR Class Of Service Parameters

- **Committed Information Rate (CIR)**
  - **guaranteed data rate (bits per second)**
- **Committed Burst Size (Bc)**
  - **max. agreed amount of data (bits) over interval Tc (seconds)**
  - **frames will have DE bits cleared (0)**
- **Excess Burst Size (Be)**
  - **max. uncommitted amount of data (bits), in excess of Bc, over interval Tc (seconds)**
  - **frames will have DE bits set to 1**

FR Class Of Service Parameters

CIR - The Committed Information Rate (CIR) for a DLC is the rate in Kb/s that the network commits to transfer user payload under normal conditions. **Bc** - The committed burst (Bc) is the maximum number of bits of user data that the network commits to transfer during the TC at the CIR under normal conditions.All frames in excess of the Bc are sent with the DE bit set. The DE bit marks frames that can be discarded when congestion has increased past an acceptable level. **Be** - The excess burst (Be) is the maximum number of bits in excess of the Bc that the network will attempt to transfer over the Tc under normal conditions.All frames in excess of the Be are discarded.

FR CoS Graph



FR CoS Graph

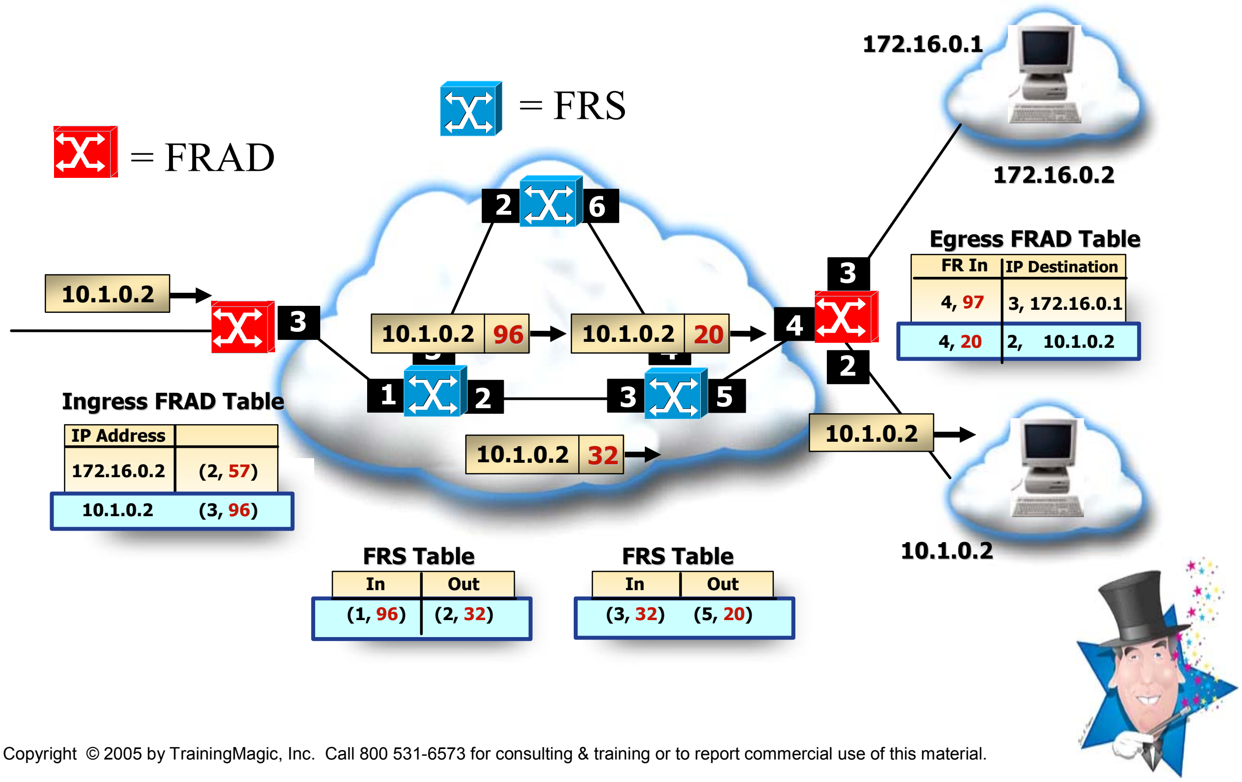
Tc - Tc is not configurable. It is calculated as Bc/CIR and is the time interval over which you may transfer Bc bits of committed data or Bc+Be bits of uncommitted data.
Maximum Burst Size - Maximum burst size is the number of cells that can be received at the peak cell rate. This allows a burst of cells to arrive at a rate higher than the sustainable cell rate. If the burst is larger than anticipated, the additional cells are either tagged or dropped.

www.TrainingMagic.com
(800) 531-6573

FR Congestion

- **None**
  - **all traffic experiences a relatively fixed amount of delay**
- **Mild**
  - **delay through the network is increased - no frames dropped due to congestion**
- **Severe**
  - **delay increased — some frames dropped due to congestion**
- **Absolute**
  - **all new frames dropped until buffer space becomes available**

Frame Relay PVC



ATM

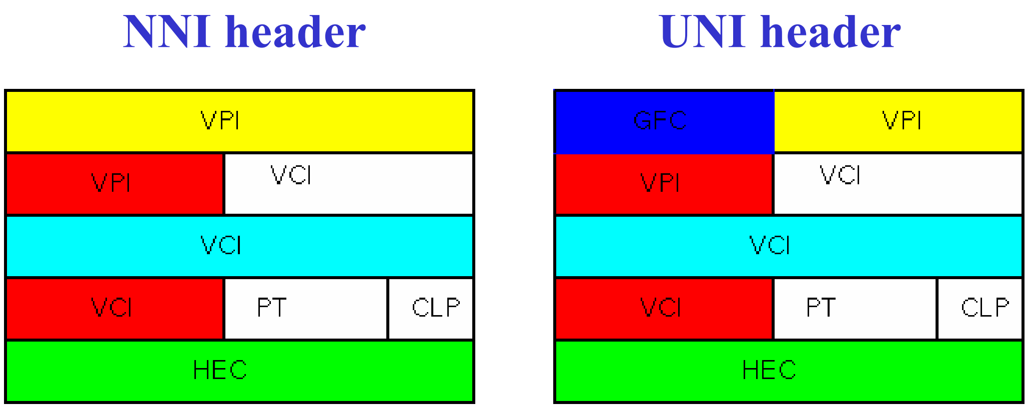
- **Virtual Circuit technology**
- **Addresses queuing delay with fixed cells**
  - **Cells are 53 bytes (5 byte header, 48 byte payload.)**
- **Provides QOS guarantees**
- **Single network for all traffic types**

ATM

ATM is based on transmitting data in fixed, small cells. ATM standards define the format of the cells, how data is mapped into these cells is, as well as the type of services that can be supported.By using small, fixed-sized transmission units, it is possible to minimize the variation in delay experienced by a cell traversing a network. Small, fixed-sized entities also make the design of switches, buffer allocation schemes, and multiplexing much simpler; in addition, it is possible to build switches that can process more than one cell simultaneously. Like other fast packet switching schemes, most of the error detection and correction is performed on an end-to-end basis by higher layer protocols. When combined with prioritization by traffic type, ATM can support a wide range of applications within a single network.

In ATM systems, error control within the network is minimal. An ATM cell is 53 octets in length and comprises Header and Payload fields. The cell Header contains information that is needed on a node-to-node basis, such as addressing and congestion notification, while the cell Payload contains service-specific information that is used on a user-to-user basis.

ATM Header



<u>ATM Header</u>

GFC—4 bits of *generic flow control* that are used to provide local functions, such as identifying multiple stations that share a single ATM interface. The GFC field is typically not used. The GFC field is not present in the NNI header.

VPI—8 bits (12 bits in the NNI header) of *virtual path identifier* that is used, in conjunction with the VCI, to identify the next destination of a cell as it passes through a series of switch routers on its way to its destination.

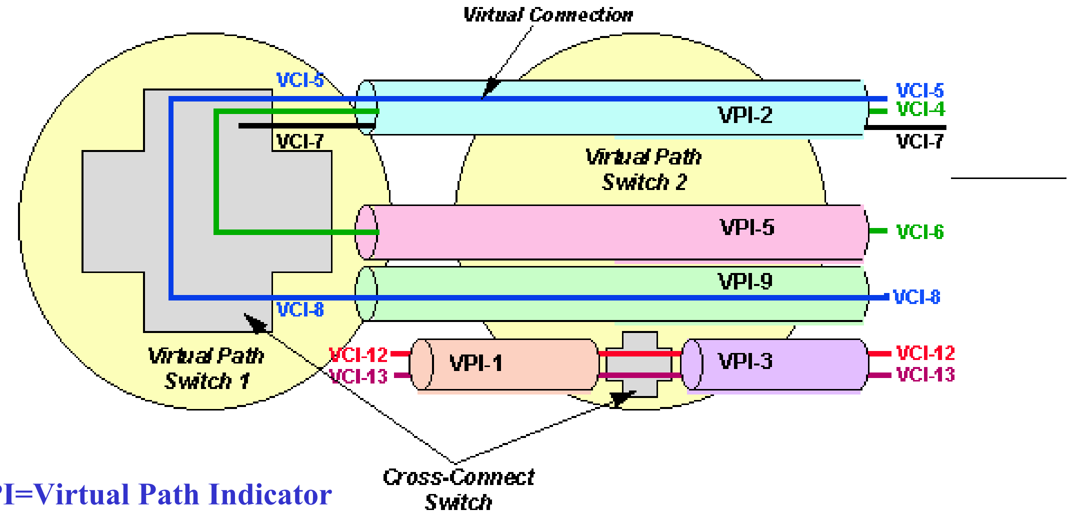
VCI—16 bits of *virtual channel identifier* that is used, in conjunction with the VPI, to identify the next destination of a cell as it passes through a series of switch routers on its way to its destination.

PT—3 bits of *payload type*. The first bit indicates whether the cell contains user data or control data. If the cell contains user data, the second bit indicates congestion, and the third bit indicates whether the cell is the last in a series of cells that represent a single AAL5 frame.

CLP—1 bit of *congestion loss priority* that indicates whether the cell should be discarded if it encounters extreme congestion as it moves through the network.

HEC—8 bits of *header error control* that are a checksum calculated only on the header itself.

VPIs & VCIs



VPI=Virtual Path Indicator

VCI=Virtual Circuit Indicator

VPIs & VCIs

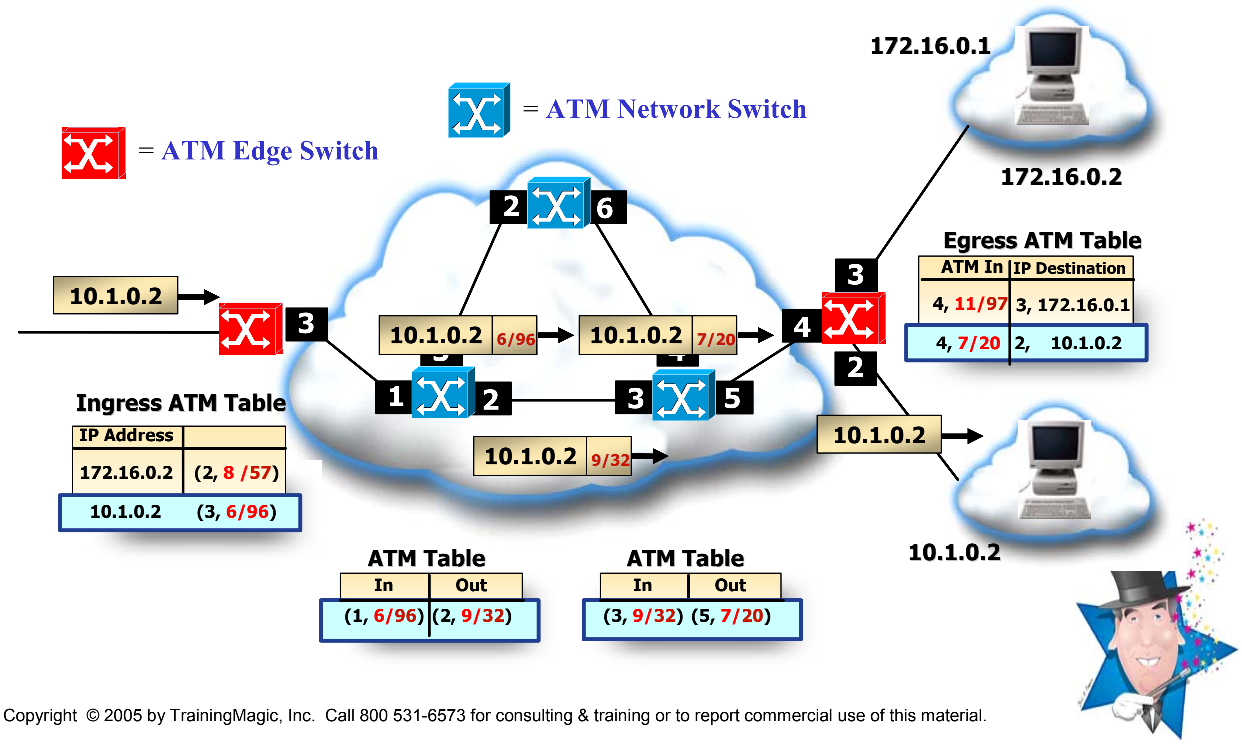
Virtual paths are be used to group specific traffic types together services which require a specific quality of service real time services. The VPI field at the UNI is 8 bits and will allow for 256 Virtual Paths to be defined. The VPI field at the NNI is 12 bits which resolves 4096 possible Virtual paths. The VPI is of local significance only.

The VCI field in the ATM Cell header provides an identifier for user channels The Virtual Channel Identifier field in the ATM Cell header is 16 bits at both the UNI and NNI, this allows a theoretical maximum of 65,535 ATM Virtual Channel Connections in any ATM Virtual Path.

The combined ATM address (VPI & VCI) supports a maximum of 16 million connections on a UNI link and 268 million connections on an NNI link

www.TrainingMagic.com
(800) 531-6573

ATM PVC



ATM Payload

ATM Cell Payload

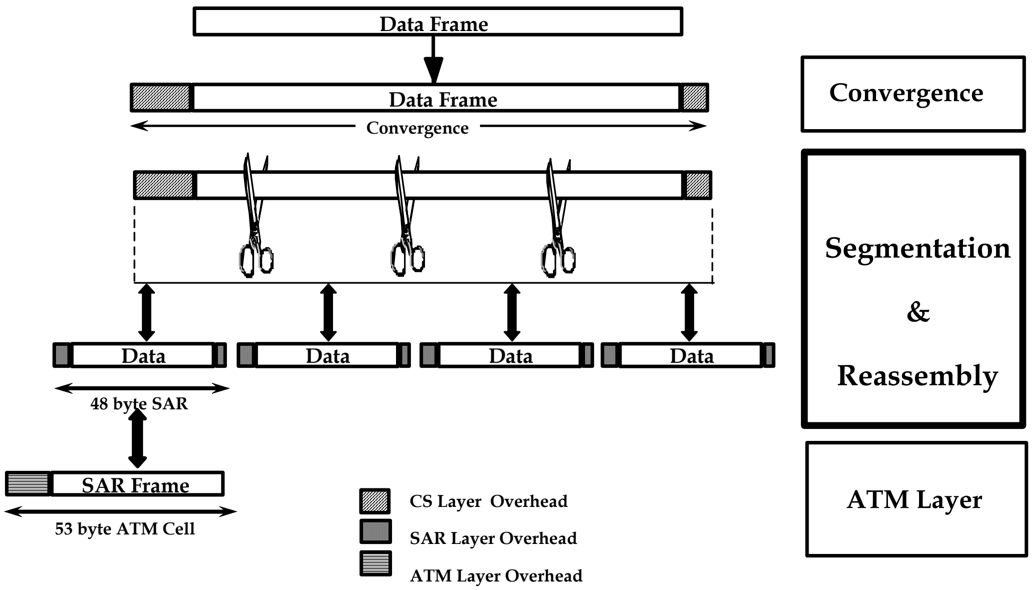
- **Payload is produced by ATM Adaptation Layer**
- **Payload & ATM Header are decoupled.**
- **Overall throughput depends on AAL used.**

<u>ATM Payload</u>

A number of AAL adaptations have been defined, specifically, four types which are listed below:

AAL Type	**Typical Application**
AAL-1	Circuit Emulation, N-ISDN, Voice over ATM
AAL-2	Packetized Voice, VBR Voice and Video
AAL-3/4	SMDS
AAL 5	Bridge/Routed data/LAN/Frame Relay, LANE, PNNI, Video, SMDS

Segmentation & Reassembly (SAR)



Segmentation & Reassembly (SAR)

The Segmentation and Reassembly Layer is the lower of two sublayers (Convergence Sublayer (CS) and SAR) that make up the ATM Adaptation Layer. The SAR is responsible for mapping data from the AAL Convergence Sublayer into the cell payloads of an ATM cell stream.

AAL 1 Overview

- Used for ITU-T “Class A” traffic
- Connection Oriented Service
- Constant Bit Rate

<u>AAL 1 Overview</u>

• SN - Sequence Number of cell stream (4 bits)

- CSI - Convergence Sub-Layer Indication
- SC - Sequence Count (mod 8)**• SNP - Sequence Number Protection - Error Detection & Correction**
- CRC - Cyclic Redundance Check (for SN field)
- EPB - Even Parity Bit (for SN & CRC fields)

AAL 3/4

- **Class C / D**
  - –**Functions Connectionless or Connection Oriented Service**
  - –**Variable Bit Rate**
  - –**End-To-End Timing Not Required (Unsynchronized)**
- **Most Complex CS & SAR Functions**
- **High CS & SAR Protocol Overhead**

<u>AAL 3/4</u>

AAL3/4 was the first of the ATM data adaptations and allows for the transmission of both connectionless and connection oriented data via Asynchronous Transfer Mode networks.

At the Common Part Convergence Sublayer the AAL3/4 format includes the following fields.

ST - Segment Type - 2 bits

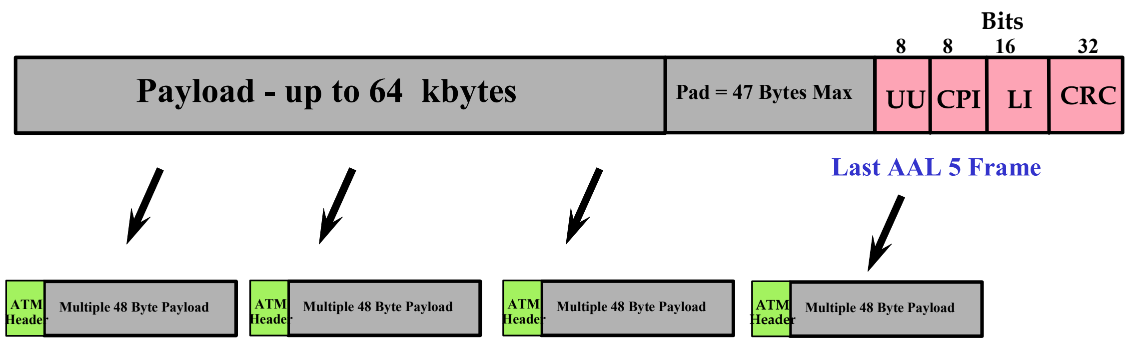
-	Beginning of Message (BOM)	10
-	Continuation of Message (COM)	00
-	End of Message (EOM)	01
-	Single Segment Message(SSM)	11

SN - Sequence Number - 4 bits
MID - Multiplexing Identifier - 10 bits
LI - Length Indicator (CS data in payload) - 6 bits

AAL 5

- **Any CoS except CBR/VBR-RT**
  - **ABR**
  - **UBR**
  - **VBR-nrt**
- **End-To-End Timing Not Required**
  - **Unsynchronized**
- **Designed for LAN Protocols**
  - **Variable Payload (0 - 65,000 bytes)**
- **Developed As A Simplified AAL 3 / 4**

AAL 5 PDU



UU = User to User

CPI = Common Part Indicator (undefined)

LI= Length indicator

CRC=Cyclic Redundancy Check

AAL 5 PDU

The structure of the AAL 5 CS PDU is as follows:

CS Payload: Can contain up to a 65,535 byte user data structure.
Pad: Used to byte align the content of the user data structure plus eight control bytes to an equal number of 48 bytes. Therefore, the pad field can contain from 0 to 47 bytes of padding.

Control: Two bytes allocated for control information include CPCS-UU - Common Part CS (User-To-User Indication) and CPI - Common Part Indicator. The functions of fields have yet to be defined. The CPI field is usually set to all 0s and the CPCS-UU unused.

Length Indicator: The bytes used to indicate the length of the user data structure.

ATM Traffic Descriptors

PCR — **Peak Cell Rate - Maximum Cell throughput in a given time. Once the value is reached cells are discarded**

SCR — **Sustainable Cell Rate -Connection rate conforming to the average ATM cell rate. Useful for VBR connections (which are bursty by nature)**

BT — **Burst Tolerance - Traffic parameter that reflects a time scale in which cell rate fluctuations can be tolerated.**

CDVT — **Cell Delay Variation Tolerance - Defines the min/max amount of delay a connection can tolerate between the arrival of any two cells**

CLR — **Cell Loss Ratio - the ratio of the lost cells to the total number of transmitted cells.**

www.TrainingMagic.com
(800) 531-6573

ATM Forum Service Classifications

<table>
<tr><td rowspan="2">ATM QoS</td><td></td><td>VBR-RT</td><td>VBR-NRT</td><td>UBR</td></tr>
<tr><td colspan="2">CBR</td><td colspan="2">ABR</td></tr>
<tr><td>Timing Relationship</td><td colspan="2">Required</td><td colspan="2">Not Required</td></tr>
<tr><td>Bit Rate</td><td>Constant</td><td colspan="3">Variable</td></tr>
<tr><td>Connection Mode</td><td colspan="3">Connection-Oriented</td><td>Connectionless</td></tr>
<tr><td>Associated AAL</td><td>AAL 1</td><td>AAL 2</td><td>AAL 3/4
AAL 5</td><td>AAL 3/4</td></tr>
</table>

ATM Forum Service ClassificationsCBR (Constant Bit Rate) is designed for real-time applications requiring tightly constrained delay and delay variation such as voice and video applications. For CBR, the following ATM attributes are specified: PCR/CDVT(peak cell rate/cell delay variation tolerance) and Cell Loss Rate.

VBR-rt (Variable Bit Rate-real time) is designed for real-time applications requiring tightly constrained delay and delay variation. Typical applications would be compressed (and therefore bursty) voice and video applications. For VBR-rt, the following ATM attributes are specified: PCR/CDVT, CLR, SCR and BT (sustainable cell rate and burst tolerance) For non-real time VBR, the following attributes are supported: PCR/CDVT ,CLR, SCR and BT.

UBR (Unspecified Bit Rate) is designed for delay-tolerant or non-real-time applications. Sources are expected to transmit non-continuous bursts of cells. UBR service is not guaranteed by mechanisms – it is considered a “best effort service”. However it is expected that resources will be provisioned for UBR service in such a way as to make it usable for some set of applications. For UBR, the following ATM attributes are supported: PCR/CDVT ABR (available bit rate) is designed for applications that have the ability to reduce their information transfer rate if the network requires them to do so. In other words, traffic is controlled by the application as directed by the network.

ITU-T Service Classifications

	Class A	Class B	Class C	Class D
Timing	Real Time		Non-Real Time	
Bit Rate	Constant	Variable		
Mode	Connection Oriented			Connection-less
AAL	1	2	3	4
			3/4	
			5	

ITU-T Service Classification are less commonly used than their ATMF counterparts

ITU-T Service Classifications

CLASS A

Provides a service for constant bit rate, connection oriented data that requires synchronisation from end to end. Services such as these would be circuit emulation of E1/T1 links or standard compressed video over the ATM network.

CLASS B

Provides service for variable bit rate traffic that is connection oriented and requires synchronisation between the end nodes. Services such as these are variable bit rate video and audio.

CLASS C

Provides service for variable bit rate, connection oriented traffic that does not require timing.

CLASS D

Provides service for variable bit rate, connectionless traffic that does not require timing.

www.TrainingMagic.com
(800) 531-6573

ATM Limitations

- **ATM is ignorant of L3 protocols**
  - **IP**
  - **Routing**
  - **Mutlicasting**
- **ATM is not ubiquitous**
  - **Most hosts do not support ATM**
- **Cell tax**
  - **Significant for some applications/networks**
- **No unified management of ATM and L1**
  - **Gigabit Ethernet**
  - **DWDM**

IP Integrated Services

- **IntServ from the IETF**
- **End-to-end (signaled) approach**
- **Suite of Protocols**
  - **RSVP ensures valid request and available resources**
  - **RTP reorders and synchronizes data**
  - **RTCP controls RTP transmission rate at host**

IP Integrated Services

The Internet Engineering Task Force (IETF) has created an end-to-end IP QoS scheme winch involves a suite of protocols. These protocols are referred to as Integrated Services (IntServ). How does IntServ work? In a nutshell, Resource Reservation Protocol (RSVP) manages a users request for a given QoS. When a request is made, RSVP checks the policy control to confirm that the user request is valid (that the user has permission to request the desired service level.) The RSVP admission control then reserves resources along the path of the packet flow. RSVP passes on the request to all intermediate nodes (typically routers). If any router along the path of a flow cannot support the requested QoS, the request is refused. RSVP does not work alone in the IntServe process. Two other protocolos, RPT and RTCP, play a central role. Real-time Protocol (RTP) ensures session integrity by including sequence numbers and time stamping packets. This permits the receiver to reorder and synchronize incoming information. RTP has no data recovery procedure. Real-time Control Protocol (RTCP) controls the rate at which a real-time application sends data.

RSVP

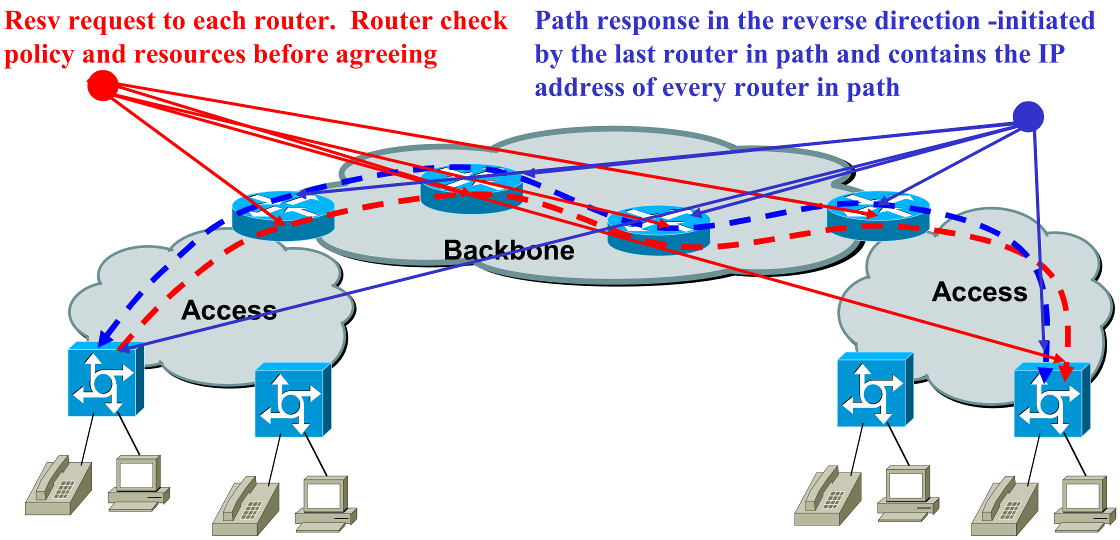
- **What is RSVP?**
  - **Method to request desired QoS across the net**
  - **Switch state establishment protocol (signaling)**
  - **Multicast friendly, receiver-oriented**
  - **Simplex reservations (single direction)**
- **Designed to work with any protocol**
  - **Protocol must provide QoS support**
  - **Examples: ATM, IP with Integrated Services**

RSVP

RSVP messages carrying reservation requests originate at receivers and are passed upstream towards the senders. At each intermediate node, a reservation triggers two general actions: 1.The RSVP passes the request to admission and policy control and the check is executed. 2. A reservation request is propagated upstream towards the appropriate senders. The set of sender hosts to which a reservation request is propagated is called the "scope" of that request.

There are two fundamental RSVP messages: Path and Resv (reserve).Path - Each receiver host sends a Resv message upstream towards the sender. These messages must follow the exact reverse path the data will use. They create and maintain the reservation state in each node along the path.Resv - Path: Each RSVP sender host transmits a Path message downstream. These store the 'path state' in each node along the way. This path state includes the IP address of the previous hop node that is used to route the Resv message in the reverse direction.

RSVP Signal Flow



Differentiated Services

Diffserv redefines the IPv4 Type Of Service field. The least significant 6 bits of the TOS are used for its service definition. This field is referred to as the Differentiated Services Code Point (DSCP). Of the 6 bits, 5 bits are allocated to the Per-Hop-Behavior (PHB) and 1 bit (known as IN bit) is to mark the packet as in-profile or out-of-the profile. The remaining 2 bits are reserved for future use.

IntServ/RSVP Today

- Most Cisco Routers Support Intserv and RSVP but not enabled
- Windows 98+ has support for RSVP
- Most vendors moving toward implementing DiffServ (Microsoft, Cisco)

Differentiated Services

- Interpret the TOS differently
- Uses a six bit differential code point
- Hop by hop QoS
- Requires major changes to internet
  - Network wide implementation
  - Enforcement of policy by ALL edge routers

Differentiated Services

Diffserv redefines the IPv4 Type Of Service field. The least significant 6 bits of the TOS are used for its service definition. This field is referred to as the Differentiated Services Code Point (DSCP). Of the 6 bits, 5 bits are allocated to the Per-Hop-Behavior (PHB) and 1 bit (known as IN bit) is to mark the packet as in-profile or out-of-the profile. The remaining 2 bits are reserved for future use.

Queuing algorithms

- **Priority Queuing (PQ)**
  - **Each queue is assigned a priority**
  - **High queues are serviced first**
- **Weighted Fair Queuing (WFQ)**
  - **Weights are assigned to each flow based on volume**
  - **Low volume traffic is scheduled first**
- **Weighted Round Robin (WRR)**
  - **Packets are assigned to a queue by CoS**
  - **Works with packets of same length**
- **Deficit Weighted Round Robin (DWRR)**
  - **Similar to WRR**
  - **Works with variable packet length**

Queuing algorithms

Priority Queuing supports some number of queues, usually from high to low. Queues are serviced in strict order of queue priority. The high queue always is serviced first, then the next-lower priority and so on. This mechanism is good for important traffic, but can lead to queue starvation.Fair Queuing first assigns traffic in flows based on the type of traffic. Each flow is assigned to a separate queue, the queues are serviced in a round robin order. Fair Queuing tends to be implemented in software and therefore slow. High volume/High Priority traffic can be experience excessive delay.Weighted Fair Queuing is a flow-based queuing algorithm that schedules low-volume traffic first, while letting high-volume traffic share the remaining bandwidth. A weight is assigned to each flow - lower weights are the first to be serviced. Fair Queuing tends to be implemented in software and therefore slow. WFQ requires large and complex state tables to track flows. Weighted Round Robin classifies packets into service classes (CBR, VBR-rt, VBR-nrt, etc). Packets are then assigned to a queue dedicated to that service class. Queues are serviced in a round robin order. WRR queuing is less than optimal if all of the packets in all of the queues are not the same size.Deficit Weighted Round Robin addresses the limitations of the WRR model by accurately supporting the weighted fair distribution of bandwidth when servicing queues that contain variable-length packets.

Queue Management

- **Random Early Detection (RED)**
  - **Uses TCP congestion control mechanisms**
  - **Discards random packets to avoid sever congestion**
- **Weighted Random Early Detection (WRED)**
  - **From Cisco**
  - **Drops packets based on IP precedence**

Queue Management

Random Early Detection (RED) is a congestion avoidance mechanism that controls the average queue size by communicating to the end hosts via TCP when they should temporarily stop sending packets. If packets continue to be send to the router, RED randomly dropping packets prior to periods of high congestion. Weighted Random Early Detection (WRED), combines the TCP capabilities of the RED algorithm with IP Precedence. This combination provides for preferential traffic handling for higher priority packets. It can selectively discard lower priority traffic when the interface begins to get congested, and provide differentiated performance characteristics for different classes of service.

Infinite Bandwidth

- **Concept of telecom gurus**
- **THE solution for QoS**
  - **Allocate each user infinite bandwidth**
  - **Addresses delay & congestion**
  - **No complex protocols**
  - **Infinite is defined as 10 times use**
- **Pre 2001 concept?**
- **DWDM and Gigiabit Ethernet?**

Infinite Bandwidth

The simplest solution to solve QoS issues would be to provide "infinite" bandwidth to every user. This approach has been advanced by a number of "Telecom/Networking Gurus" including George Gilder. In practice, infinite bandwidth is defined as 10 times the bandwidth needed by the application. Among the many benefits of inflate bandwidth is the ability to solve the QoS problem with out resorting to complex protocols. Is your data experiencing congestion/delay – add a whole lot more bandwidth. In today's networks there will always be some points of congestion. Some kind of traffic aggregation and priority mechanism is necessary to provide QoS. However, this concept may still have a future as technologies like DWDM and Gigabit Ethernet become ubiquitous.