Multimedia Communications

Multimedia Technologies & Applications

Prof. Dr. Abdulmotaleb El Saddik

Multimedia Communications Research Laboratory

School of Information Technology and Engineering
University of Ottawa

Ottawa, Ontario, Canada

elsaddik @ site.uottawa.ca

abed @ mcrlab.uottawa.ca

www.site.uottawa.ca/~elsaddik www.el-saddik.com

Outline

- Quality of Media vs. Quality of Service
- QoS Layered Model for Multimedia Systems
- QoS Parameters
- Types of Services
- QoS Intervals
- Negotiation
  - Bilateral peer-to-peer negotiation
  - Triangular negotiation for information exchange
- IP QoS Networks
  - Resource ReSerVation Protocol (RSVP)
  - DiffServ
- Rate-Based Scheduling Disciplines

www.site.uottawa.ca/~elsaddik www.el-saddik.com

Quality of Media vs. Quality of Service

- Quality.... with respect to a Quality "Parameter" ("Measure"?)
- Example Audio QoM (huge set exists in professional audio):
  - Frequency Spectrum (linear amplification? ...)
  - Signal2Noise Ratio SNR (noise, click, ...)
- Example Video QoM
  - spatial / temporal resolution
  - SNR
- intuitively spoken:
  - QoM: something like "HiFi Audio" (16..20k Hz)
  - QoS: something like "Hi bandwidth" (1Gbps)
- but: if application delivers HiFi Audio to the user, it does so as a "service“

→ therefore in the remainder: QoS

www.site.uottawa.ca/~elsaddik www.el-saddik.com

Quality of Service (QoS)

ISO standard defines QoS as a concept for specifying how "good" the offered networking services are.

QoS elements:

- How to define QoS
  - Parameter set
- How to determine QoS
  - Negotiation procedure
- How to ensure QoS
  - Reflection in appropriate network access mechanisms (scheduling)

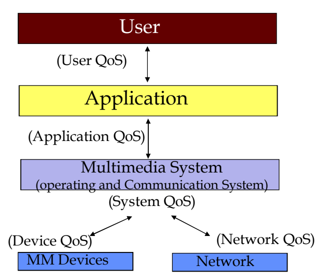
QoS has different implications in different fields:

- Operating system / Resource scheduling
- File system organization
- Compression
- Communication system support
- Media synchronization
- ...
- User Interface

www.site.uottawa.ca/~elsaddik www.el-saddik.com

4

QoS Layered Model for Multimedia Systems



www.site.uottawa.ca/~elsaddik www.el-saddik.com

5

QoS - Layer Model

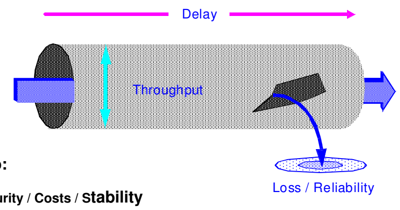
- Perception QoS
  - Tolerable Synchronization Drift
  - Visual Perceptibility
- Application QoS parameters describe requirements for:
  - Media quality (media characteristics, end-to-end delay)
  - Media relations (media conversion, media synchronization)
- System QoS
  - CPU Rate / Usage
  - Available Memory
- Communication QoS
  - Packet Size / Rate
  - Bandwidth
  - End-to-End-Delay
- Device QoS
  - Seek / Data Transfer Rate
  - Sample Rate / Resolution

www.site.uottawa.ca/~elsaddik www.el-saddik.com

6

QoS Parameters - Example Transport System

- Common parameters concerning the Transport System are:
  - **Bandwidth /Throughput**
  - **Delay (Latency)**
  - **Jitter**
  - **Order**
  - **Loss (Reliability)**



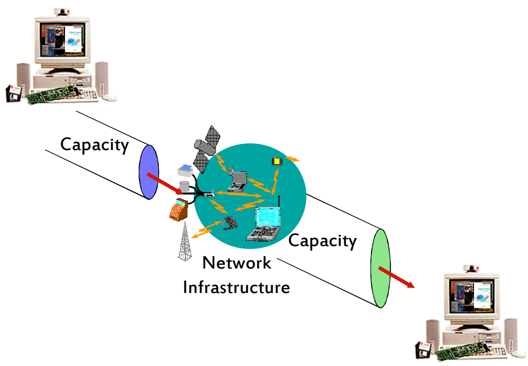
but also:

Security / Costs / Stability (Resilience)

www.site.uottawa.ca/~elsaddik www.el-saddik.com

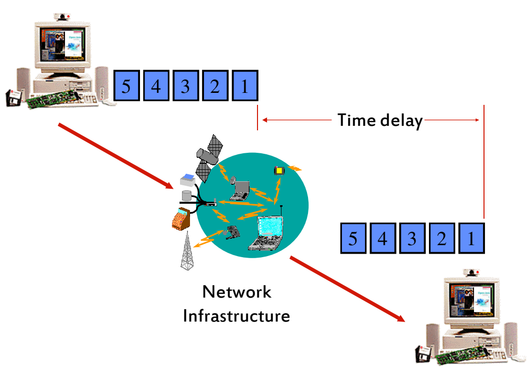
7

Bandwidth



8

Latency (Delay)



9

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

Sources of Delay

Propagation Delay

- Delay through a physical medium

Link Speed

- Data transfer determined by link bit rate

Queuing

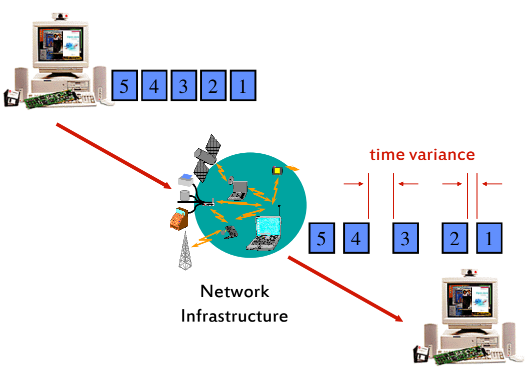
- Time spent in router queues

Hop Count

- Each router or switch adds queuing delay

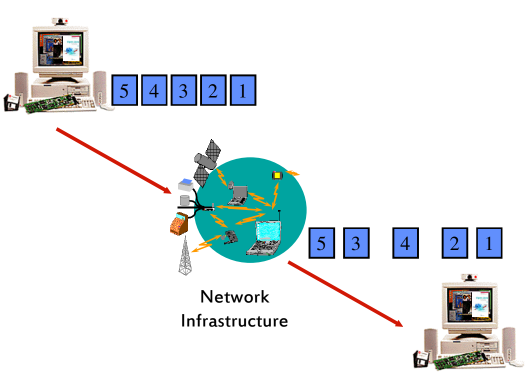
www.site.uottawa.ca/~elsaddik
www.el-saddik.com

Jitter

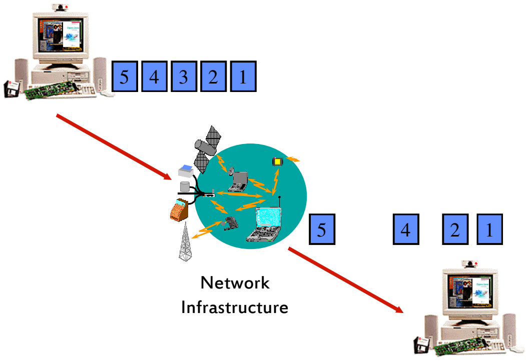


www.site.uottawa.ca/~elsaddik
www.el-saddik.com

Order (Re-Sequencing)



Loss



13

Example QoS Parameters (cont.)

- Delay:
  - Maximum end-to-end delay for transmission of one packet
  - Delay jitter = maximum variance of transmission
- Throughput:
  - Maximum long-term rate = maximum amount of data units transmitted per time interval
  - (e.g. packets or bytes per second)
    - Maximum burst size
    - Maximum packet size
- Loss:
  - Sensitivity class: ignore / indicate / correct losses
  - Loss rate = maximum number of losses per time interval
  - Loss size = maximum number of consecutively lost packets

14

QoS parameters and types of service

<u>note</u>: QoS parameters often subject to statistical process

→ mean, min, max, distribution, variance, ...

- Guaranteed Service
  - values or intervals of QoS parameters
    - deterministic (at any time)
    - statistical (consider a time interval or certain propability)
    - → QoSmin <= P <= QoSmax
- Predictable Service
  - based on past network behavior
  - QoS parameters are estimates of past behavior
    - consider history
    - from the very beginning of calculation
  - "if it was like that in the last ..., you can rely on ..."
- Best Effort Service
  - no or just partial guarantees

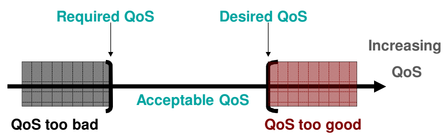
→ most of current network protocols have best effort services

15

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

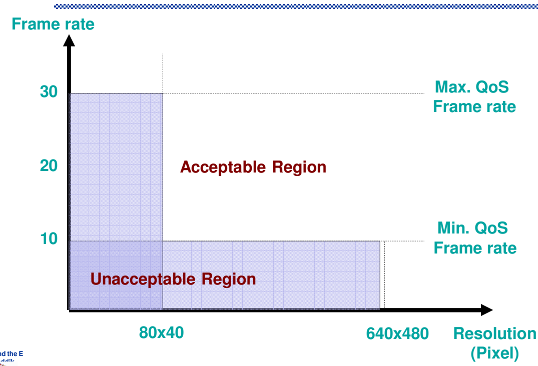
QoS Intervals

- Parameter values result in
  - acceptable regions
  - inacceptable regions
- of QoS
- note:
  - below required QoS level - no reasonable service
  - above required QoS level - unnecessary resource consumption / costs



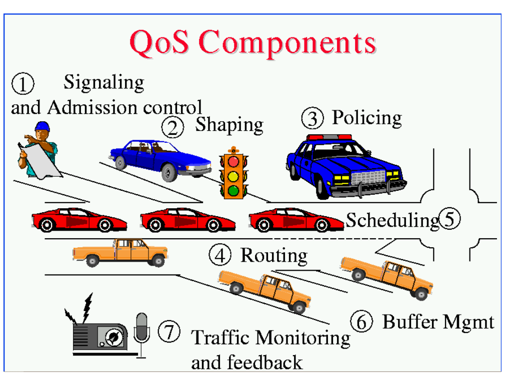
www.site.uottawa.ca/~elsaddik
www.el-saddik.com

QoS Intervals: multidimensional intervals



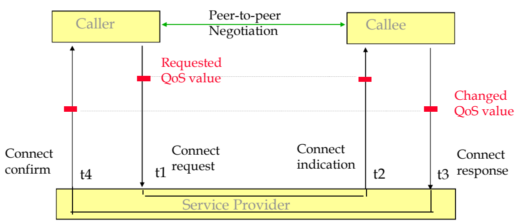
www.site.uottawa.ca/~elsaddik
www.el-saddik.com

QoS



Bilateral peer-to-peer negotiation

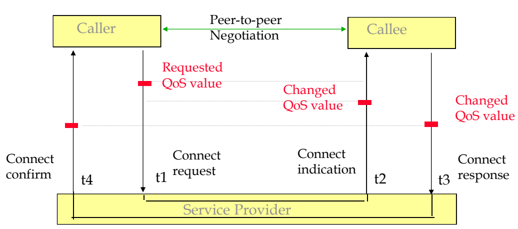
- Negotiation between two peers, with service provider not allowed to modify QoS values



www.site.uottawa.ca/~elsaddik www.el-saddik.com

Triangular negotiation for information exchange

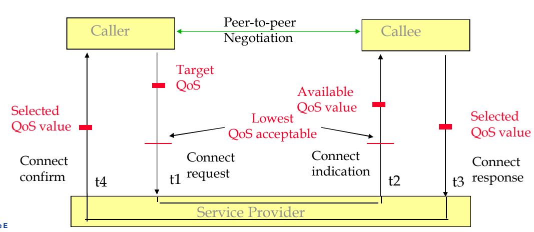
- caller introduces into request average value of QoS parameter
- value can be changed by service provider/callee
- at the end of negotiation, all parties have the same QoS parameter value



www.site.uottawa.ca/~elsaddik www.el-saddik.com

Triangular negotiation for bounded target

- caller specifies target QoS (average value) and lowest QoS acceptable;
- service provider not allowed to change the min. QoS value (connection may be rejected, however), but may change the target value;
- callee makes the final selection of target QoS value and confirms to caller.



www.site.uottawa.ca/~elsaddik www.el-saddik.com

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

Stream Protocol -2 (ST-2)

Internet Stream Protocol, Version 2 (ST-2):

- **Internet RFC 1190 (Oct.1990)**
- **"Full" protocol with data handling and control messages**
- **Connection-oriented substitute for IP:**
  - **Multicast support**
  - **Resource reservation support**
- **Main abstractions:**
  - **Routing tree from one source to multiple targets**
    - **Created during connection establishment**
    - **Originally only sender-initiated connection setup**
  - **"Flow Specification" describes QoS parameters**

© elsaddik

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

IP Networks - *QoS*

- **QoS is about managing congestion while making optimal use of resources**
  - **RSVP:** ***Uses signaling and per-flow info***
  - **Diffserv:** ***differential queuing, aggregate traffic***
  - **Media Scaling**
  - **Bandwidth Control**
  - **Congestion Control**

© elsaddik

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

IP Networks - Integrated Services

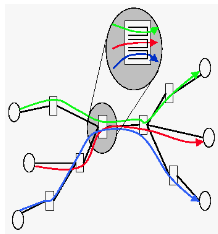
- **Enhance IP's service model**
  - **Old model: single best-effort service class**
  - **New model: multiple service classes including best-effort class and QoS classes**
- **Key architecture difference**
  - **Old model: stateless**
  - **New model: per flow states maintained at routers**
    - **Used for admission and scheduling**
    - **Setup by signaling protocol**

© elsaddik

www.site.uottawa.ca/~elsaddik www.el-saddik.com

IP Networks - Integrated Services

- Each flow has a fixed path
- Routers along the path maintain the state of the flows
- Relies on resource reservation



25 Beyond the E

www.site.uottawa.ca/~elsaddik www.el-saddik.com

IP QoS Networks - *IntServ*

- RSVP (ReSource reserVation Protocol)
  - Path of guaranteed bandwidth
  - Reservations on a per flow basis
- Resources are managed per-flow
- Controlled Load Service, Guaranteed Service
- Every forwarder examines, classifies and polices every packet
- Advantage:
  - QoS guaranteed
- Disadvantage:
  - Scalability a problem

26 Beyond the E

www.site.uottawa.ca/~elsaddik www.el-saddik.com

Resource ReSerVation Protocol (RSVP):

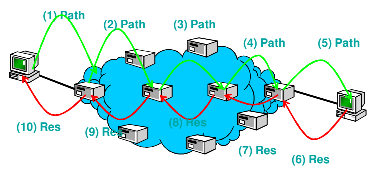
- Version 1 of RSVP has been specified in RFC 2205
- Contains only protocol elements for control, not data transfer
- Companion protocol to IP:
  - Controls how IP sends a packet
  - Resource reservation support
- Main abstractions:
  - IP multicast routing tree from source(s) to multiple targets
  - Receiver-initiated reservation
  - Filtering provides for:
    - Heterogeneous receivers
    - Different reservation styles
  - Concentrates on resource reservation only
- Allows aggregations using filters
  - resource allocations for sum of sources
- New sender-oriented RSVP version

27 Beyond the E

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

RSVP

- Creating and maintaining reservation state:
  - Source:
    - Sends PATH message
  - Receiver:
    - Joins multicast group
    - Determines own QoS requirements
    - Sends RESERVATION message
  - Periodic refresh of 'soft state' via transmission of
    - PATH messages
    - RESERVATION message
  - Source not restricted from transmitting data at any time
    - Packets may go across unreserved routes
    - User data forwarding protocol has to be aware of relation between a packet and reserved resources

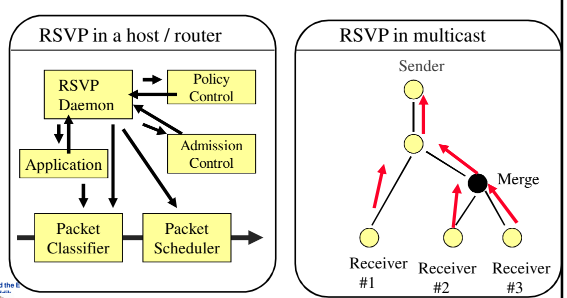


28

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

RSVP & QoS

Reservation Protocol: A host uses RSVP to request a specific Quality of Service from the network.



29

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

IP QoS Networks - *DiffServ*

- **Differentiated Services (Diff-Serv)**
- **Traffic classification, policing or conditioning at border or edge nodes**
- Per Hop Behaviour based QoS
  - **Packets allocated a specific Class of Service (CoS)**
  - **Determines relative forwarding priority or drop sequence**
  - **Different amounts of bandwidth allocated to different classes**
- Advantage:- Lower router overheads
- Disadvantage:- QoS cannot be guaranteed

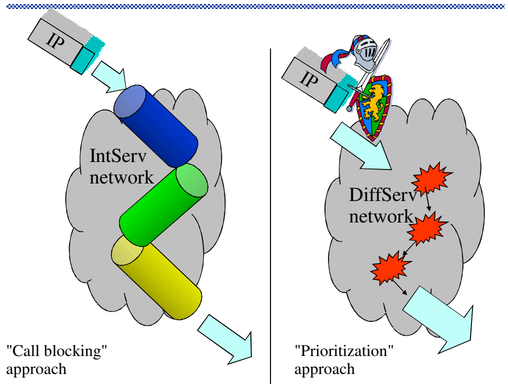
30

Combining RSVP & Diffserv

- IntServ more suited to access network
- DiffServ more suited to core network
- Good QoS system
  - Uses both methods
  - IntServ confined to edge of network
- How it works
  - Sender sends PATH message to diffserv network ingress router
  - PATH is carried transparently through diffserv network
  - Receiver responds with RESV
  - RESV follows PATH trail back towards sender
  - Diffserv ingress router determines whether to admit RSVP request based on mapping from requested service type to diffserv service level and capacity at that service level, per SLA (Service Level Agreement)

31

IntServ vs. DiffServ



32

QoS Translation

- **Human Interface-Application QoS: Tuning Service**
  - **Graphical User Interface (GUI) for user input of desired application QoS and output of negotiated QoS**
- **Application QoS-System QoS**
  - **translation maps application requirements into system QoS parameters (e.g., "high-quality" lip-synchronization user requirement is mapped to few msec audio-video "skew" QoS parameters)**
- **System QoS-Network QoS**
  - **maps system QoS (e.g., transport packets end-to-end delay) into underlying network QoS parameters (e.g, in ATM, cell end-to-end delay) and vice versa**

33

Media Scaling

- Scaling means sub-sampling a data stream and only presenting a fraction of its original contents
  - Adjusting media stream according to network bandwidth
- can be done either at the source or at the destination
- Two methods
  - Transparent scaling:
    - transport system scales-down the media, by dropping some portion of stream; must identify portions dropped.
  - Non-transparent scaling:
    - requires some interaction of transport system with the upper layers;
    - modification of media stream before it is presented to transport layer;
    - typically requires modification of some parameters in the coding algorithms, or even re-coding of a stream that was previously encoded in a different format.

www.site.uottawa.ca/~elsaddik www.el-saddik.com Beyond the E

Scaling and reservation

- Relation between reservation and scaling
  - Reservation:
    - Available in more and more networks
    - Should be utilized by higher layers
  - Scalability:
    - Complementary technique
    - Feasible for existing networks
    - Makes solution more stable through dynamic adjustment
    - Requires appropriate media encoding

www.site.uottawa.ca/~elsaddik www.el-saddik.com Beyond the E

Video Scaling Methods

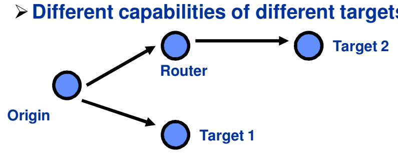
- Multitude of scaling dimensions
  - Temporal
    - reduces frame rate
  - Spatial
    - reduces pixel number
  - Frequency
    - reduces number of DCT coefficients
  - Amplitude
    - reduces color depth for each pixel, e.g. by coarser quantization of DCT coefficients
  - Color space
    - number of colors, luminance/chrominance
- Usability of scaling dimension depends on:
  - Coding method
  - System support (knowledge about data stream structure)

www.site.uottawa.ca/~elsaddik www.el-saddik.com Beyond the E

Multiple Target Scenario

Problem:

- Different throughput to different targets
- Different capabilities of different targets



Solutions:

- Worst for all ("socialist approach")
- Best for all ("capitalist approach")
  - Not all connections go to all targets

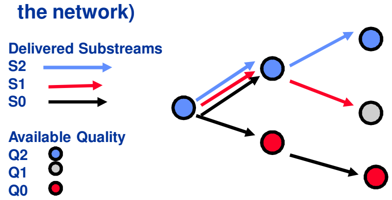
www.site.uottawa.ca/~elsaddik www.el-saddik.com

37

Heterogeneous Receiver- Filtering

Targets require different quality levels:

- New compression schemes (e.g.,MPEG-2) offer hierarchical coding
- Data stream can be divided into sub-streams
- Only desired information part is forwarded
- Not interesting part is removed by filters (in the network)



www.site.uottawa.ca/~elsaddik www.el-saddik.com

38

Resource Admission Service

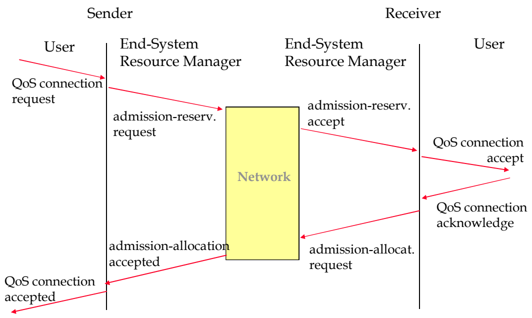
- Follows QoS specification at each layer;
- Checks availability of shared resources, through admission tests
  - schedulability test (CPU sched., packet sched. for given QoS)
  - spatial test (buffer allocation for QoS guarantees)
  - link bandwidth test (throughput guarantees)
- Based on admission tests, creates either "reserve" message with admitted QoS values or "reject" message when min. QoS cannot be satisfied.

www.site.uottawa.ca/~elsaddik www.el-saddik.com

39

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

Resource Reservation/Allocation Protocol



40
Beyond the E

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

Resource Management During Multimedia Transmission

- **Rate Control**
  - New rate-based flow control and service disciplines
  - Provide with a min. service rate, independent of traffic characteristics of other connections
  - Provide throughput, delay, jitter, and loss-rate guarantees
- **End-to-End Error Control**
  - End-to-end data integrity
- **Resource Monitoring**
  - Monitoring of resource utilization during MM connection
- **Resource Adaptation**
  - Dynamic change of QoS parameters

41
Beyond the E

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

Flow Control of Bursty Traffic

- **End-to-end window-based, e.g., TCP**
- **Rate-based: source rate is adjusted with feedback from the destination and the network**
- **Credit-based: hop-by-hop window-based**
  - Pre-allocated buffers per connection

42
Beyond the E

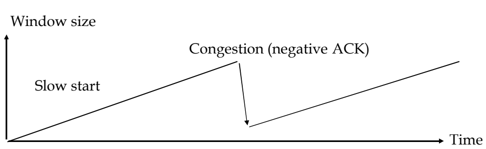
End-to-End Window-Based

- "Best effort" used in TCP
- The size of the window determines the number of outstanding PDUs without positive acknowledgement
  - Example: for window size k, PDU with sequence number n cannot be sent before the sender receives a positive ACK for PDU n-k
  - PDU = protocol data unit

43

End-to-End Window-Based in TCP

"Slow start" mechanism in TCP, controls window size in order to avoid congestion



- "Slow start" is in fact shaping the bursty data sources

44

Rate-Based Scheduling Disciplines

- Fair Queuing
- Virtual Clock
- Delay Earliest-Due-Date (Delay EDD)
- Jitter Earliest-Due-Date (Jitter EDD)
- Stop-and-Go
- Hierarchical Round Robin (HRR)

45

Fair Queueing

- If N channels share an output trunk, then each gets 1/N of the bandwidth
- If a channel uses less than its share, then the portion saved is shared among the rest equally
- Mechanism can be achieved by *Bit-by-bit Round Robbin (BR)* among the channels, but is inefficient:
  - one bit from each queue that has a packet in it
- Fair queueing emulates BR:
  - each packet is given a finish number, which is the round number at which the packet would have received service, should the server have used BR (finish number calculated given bit rate and no.of active connections)
  - packets are served in the order of that round number (increasing)
  - channels can be given different fractions of bandwidth, by assigning them weights, which correspond to the number of bits of service the channel receives per round of BR service.

Weighted Fair Queuing

- Generalization of Fair Queueing
- For VC (Video conferencing) / sessions with different rates
- One round-robbin cycle gives service proportional to the session rate
  - e.g., twice the rate gets twice the number of bits in BR
- Delay is inversely proportional to VC / session rate
- Currently proposed by IETF INT-SERV working group for the Internet guaranteed QoS
- It has been proven that if traffic shaped at the end of the network with a token-leaky bucket and then WFQ in all nodes, then end-to-end delay is bounded

Virtual Clock

- This discipline emulates *Synchronous Time Division Multiplexing (STDM)* in an asynchronous network
- Each switch has a LOCAL TIMER: Virtual Clock
- A virtual transmission time is allocated to each packet (time stamp)
  - time at which the packet would have been transmitted, if the server would actually be doing STDM.
  - time-stamp equal to max of real-time and deadline
- Virtual Clock Time is used to determine
  - transmission priority
    - low rate connections have low priority
- Per VC /session queueing and scheduling
- Virtual Clock, coupled with admission control and resource management, provides deterministic bandwidth but not delay guarantees.

Delay Earliest-Due-Date (Delay EDD)

- Delay EDD is an extension of Earliest Deadline First (EDF) scheduling
- Server negotiates a service contract with each source
  - Contract states that if source obeys a peak and average sending rate, then the server provides bounded delay
- Key lies on assignment of deadlines to packets
  - Server sets a packet's deadline to time at which it should be sent, if it had been received according to contract
  - this is actually the expected arrival time added to the delay bound at the server
- Delay EDD can assure each channel a guaranteed delay bound, by reserving bandwidth at the peak rate.

Jitter Earliest-Due-Date (Jitter EDD)

- Jitter EDD extends Delay EDD to provide delay-jitter bounds
- Each hop has a jitter bound, j, and queueing deadline per connection, b
- At each hop packet can be delayed up to b-j
  - and must be transmitted by the deadline b
- After served by a server, a packet gets stamped with the difference between its deadline and actual finishing time
- A regulator at the entrance of the next switch holds the packet for this period, before scheduling it, thus providing min. and max. delay guarantees
  - i.e., next hop compensates for queueing jitter of previous hop
- Packet suffers max. delay at all nodes, except the last

Stop-and-Go

- Preserves "smoothness" of traffic
- Traffic viewed as frames of length T bits
- At each time frame, only packets that have arrived at the server at the previous time frame are sent
- Delay and delay-jitter are bounded
  - at the expense of a delay T for packets that arrive at the current frame
- Multiple frame sizes can improve the performance of Stop-and-Go.

Hierarchical Round Robin (HRR)

- HRR server has several service levels, each providing RR service to a fixed number of slots
- Certain number of slots at a given level are allocated to a channel and server cycles through slots at each level
- *Frame Time* at a level, is the time of servicing all slots at that level
- HRR gives each level a constant share of bandwidth
- "Higher" levels get more bandwidth than "lower" levels (frame time smaller at higher levels)
- Max. delay bound provided to channels allocated at a level

52

Rate Control at Boundaries

- One of the main causes of congestion is that traffic is bursty
- Traffic shaping tries to manage congestion by forcing the packets to be transmitted at a more predictable rate
- Traffic shaping is about regulating the average *rate* (and burstiness) of data transmission
- When the circuit is setup, the user and carrier agree on a shape for that circuit

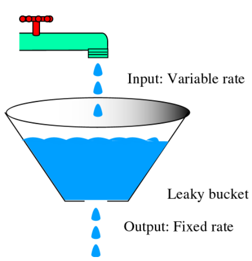
IDEA

- keep network simple
  - complexity at the boundaries
- setting-up different types of connections (with different QoS) for
  - data
  - voice
  - video
- bandwidth reservation for peak and/ or average rate
- transmission starts after setup is completed

53

Leaky Bucket Algorithm

- Essentially what we want to do is to provide a consistent, even flow of traffic
- Think of a bucket with a hole in the bottom, or a leaky faucet, no matter how much water enters the bucket the output flow is constant
- That's the idea behind the leaky bucket algorithm

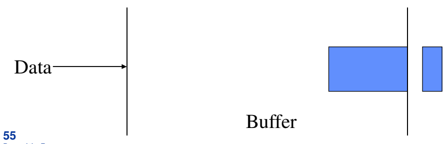


54

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

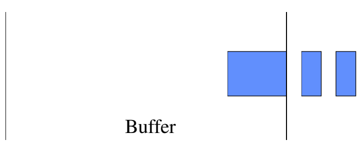
The Leaky Bucket

- Variable input rate
- Data output at constant rate
- Size of buffer '= bucket' is important
  - If bucket overflows input data discarded
- Suitable for multimedia such as TV, audio
- In practice the bucket is a finite queue that outputs at a finite rate.
- The bucket has a finite buffer, danger of overflow!



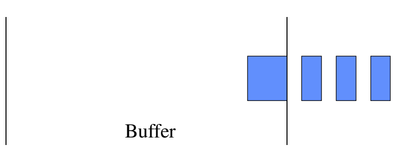
The Leaky Bucket

- Variable input rate
- Data output at constant rate
- Size of buffer '= bucket' is important
  - If bucket overflows input data discarded
- Suitable for multimedia such as TV, audio
- In practice the bucket is a finite queue that outputs at a finite rate.
- The bucket has a finite buffer, danger of overflow!



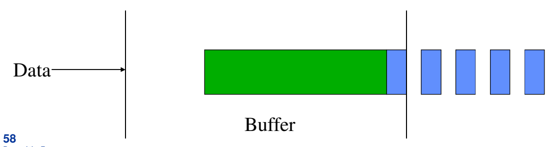
The Leaky Bucket

- Variable input rate
- Data output at constant rate
- Size of buffer '= bucket' is important
  - If bucket overflows input data discarded
- Suitable for multimedia such as TV, audio
- In practice the bucket is a finite queue that outputs at a finite rate.
- The bucket has a finite buffer, danger of overflow!



The Leaky Bucket

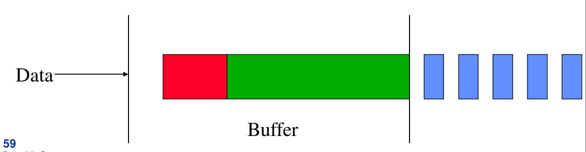
- Variable input rate
- Data output at constant rate
- Size of buffer '= bucket' is important
  - If bucket overflows input data discarded
- Suitable for multimedia such as TV, audio
- In practice the bucket is a finite queue that outputs at a finite rate.
- The bucket has a finite buffer, danger of overflow!



Beyond the E
www.site.uottawa.ca/~elsaddik
www.el-saddik.com

The Leaky Bucket

- Variable input rate
- Data output at constant rate
- Size of buffer '= bucket' is important
  - If bucket overflows input data discarded
- Suitable for multimedia such as TV, audio
- In practice the bucket is a finite queue that outputs at a finite rate.
- The bucket has a finite buffer, danger of overflow!



Beyond the E
www.site.uottawa.ca/~elsaddik
www.el-saddik.com

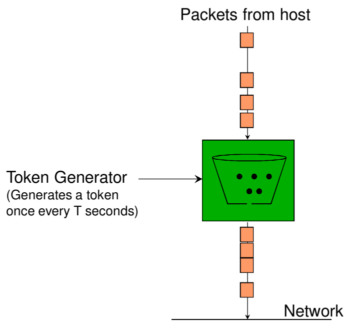
Token Bucket Algorithm

- In some cases, we may want to allow short bursts of packets to enter the network without smoothing them out
- For this purpose we use a token bucket, which is a modified leaky bucket

- The token bucket algorithm allows the output to speed up somewhat when large bursts arrive
- The bucket holds tokens instead of packets
- Tokens are generated and placed into the token bucket at a constant rate
  - In regular intervals tokens are thrown into the bucket
- If there is a ready packet, a token is removed from the bucket, and the packet is send.
- If there is no token in the bucket, the packet cannot be send.
- The token bucket has a fixed size (maximum capacity), so when it becomes full, subsequently generated tokens are discarded

Beyond the E
www.site.uottawa.ca/~elsaddik
www.el-saddik.com

Token Bucket



61

End-to-End Error Control

- Error Detection
- Error Correction
  - Go-back-N Retransmission
  - Selective Retransmission
  - Partially Reliable Streams
  - Forward Error Correction (FEC)
  - Priority Channel Coding
  - Slack Automatic Repeat Request (S-ARQ)

62

Resource Monitoring

- Important part of resource management in networks and end systems
- End-user mode
  - requests status report about resources
  - supervisor function, observing QoS parameters
- Network mode
  - reports status of different nodes in a MM connection
- Monitoring can add overhead during MM transmission, thus it should be flexible
  - use many optional variables
  - able to be turned on and off

63

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

Resource Adaptation

- Two important goals for dynamic change of QoS
  - notification of change and QoS re-negotiation
  - resource adaptation to accommodate changed QoS
- User request for re-negotiation
- Host-system request for re-negotiation/ change
- Network request for re-negotiation/ change
- Network adaptation
  - dynamic re-routing mechanism
- Source adaptation
  - Rate control using network feedback
  - Source traffic shaping
  - hierarchical coding

64
Beyond the E

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

Resource De-allocation

- Following completion of Multimedia call, the connection should be released and reservation of resources cancelled;
- Closing process should not disrupt other media flows in the network;
- Two types of resource de-allocation procedures:
  - Sender requests closing of MM call;
  - Receiver requests closing of MM call.

65
Beyond the E

www.site.uottawa.ca/~elsaddik
www.el-saddik.com

66
Beyond the E