Internet QoS: A definable goal?

Scott Bradner
Harvard University
sob@harvard.edu

Quality of Service (QoS)

- the ability to define or predict the performance of systems on a network
  note: predictable may not mean "best"
- unfair allocation of resources under congestion conditions
- long-time SNA feature
  - SNA as QoS example has problems
    - session-oriented, manual configuration
- pundits want QoS, some purists are not sure
  do you want to block an emergency phone call?

Quality of Service, What Is It?

- the ability to define or predict the performance of systems on a network
- note: predictable may not mean "best"
- **unfair allocation of resources** under congestion
- long-time SNA feature
  - SNA as QoS example has problems
    - session-oriented, manual configuration

Applications

- elastic application
  - wait for data to show up
  - functions, with some negative implications, under adverse network conditions
  - e.g. email, file transfer, telnet, ...
- real-time applications
  - playback applications
    - buffer data to eliminate network jitter
      - e.g. RealAudio, RealVideo
  - interactive applications
    - max interaction time - e.g. people
    - e.g. telephone calls

Playback Applications

- creates client-end buffer to store data
- start playback some after buffer fills to some level
  - may have to adjust buffer size on the fly
    - if too much jitter seen in incoming data
    - can cause disjunction in playback
- playback rate should be same as original sample rate
- can include timing in data packets
  - to control playback rate

Interactive Applications

- max latency determined by some external constraint
  - e.g. human systems
    - max RTT for voice interaction 300 - 400 msec
    - otherwise talk over each other
- smaller buffer at receiver
- data that is too late is useless

IP & QoS

- original goal in IP - TOS bits - RFC 791
  - "provides an indication of the abstract parameters of the quality of service desired"
  - "guide the selection of the actual service parameters when transmitting a datagram through a particular network"
  - intended to be used only within a single network
- "expected to be used to control ... routing and queuing algorithms" (RFC1122)
- "precedence is a scheme for allocating resources in a network based on the importance of different traffic flows" (RFC 1812)

qos - 7

Where is QoS Needed?

- where there are not enough resources
  - "resources" include time
- OK if can send all data within required time
- QoS is what do you do when you need controls

qos - 8

Traditional Service Quality Agreements

- *service level agreement (SLA)*
- big in the glass house world
- some pundits think SLAs will solve all Internet problems
- contract between network provider and users
  - defines service level and cost for that service level
  - can include
    - response time (average & maximum)
    - availability percentages
    - number of active sessions
    - throughout rates

qos - 9

Example Traditional SLA

Application Name:	CICSP01
Est. Volume:	10,000 trans/day
Est. # users:	1000
Maximum outage time:	
Lines:	30 min.
FEP:	15 min.
Clusters:	30 min.
Recovery Procedures:	
Lines:	1) modem testing
	2) FEP Port testing
	3) Matrix switching
	4) dial back-up services
FEP:	matrix switching to direct lines to backup FEP
Clusters:	1) IML of failed cluster
	2) inactive/activate of failed resource
	3) contact field support
Availability:	92-98%
Accessibility:	98.1%
Serviceability:	
Av. response time @ peak periods:	4 seconds
Transmission volume @ peak periods:	4000 transactions per hour

SLAs & Internet

- desire to sign SLAs with ISPs
- hard to get useful guarantees
  - datagram networks do not lend it self to guarantees
  - reliability to where?
  - latency to where? what time of day?
- even if on same ISP could be a remote site problem
- some ISPs will give discounts for "outages"
  - as long as they could do something about it
  - careful definitions of outage types
- ANX may be an example

Token Buckets

- underlying mechanism for many QoS technologies
- buffer of tokens
  - token added to buffer at fixed rate (discarded on overflow)
  - need token to transmit a packet
    - subtract a token for each packet transmitted
    - if no tokens in bucket, have to wait for new token before transmitting
- allows bursts to be transmitted but throttles long-term data rate

Leaky Bucket

- older mechanism
- buffer for data
  - data transmitted out of buffer at fixed max rate
  - buffer skipped if no data to be transmitted
  - data lost if buffer overflows
- provides space for burst but does not pass burst on
  - evens out flow
  - adds latency on burst

QoS Types

- predictive
  - architect network based on observed loads
  - can also police input loads
- flow based
  - reserve bandwidth through network for an execution of an application
  - keep track of reservation in each network device in path
- non flow based
  - mark packets to indicate class
  - process differently in network based on marking

Predictive QoS

- QoS in most current datagram networks
- "just" make network "big" enough
- reasonable on a LAN or campus network
- no actual guarantees
- hard to do for WAN
- tends to provide cycles of quality
  - over build for need
  - need catches up and passes capacity
  - over build for new need

Throw Bandwidth at Problem

- with “enough” bandwidth QoS can be easy
  - enough means much more than peaks
  - e.g., gigabit Ethernet for 1 video stream
- still might have to sequence data onto link
  - if bursty traffic

Flow Based QoS

- per flow reservations
- per flow guarantees
- per flow state kept in network
- e.g. ATM
- scaling issues
- IETF per-flow QoS work
  - inteserv - link level mechanisms
  - RSVP - signaling

Flow Based QoS

- ATM QoS
- IP-based QoS
- mixed

ATM QoS

- set up virtual circuit across network
  - defined QoS for each VC
- basic QoS is to control:
  - absolute cell latency from source to destination
  - variation in cell latency
- different requirements for broadcast vs. interactive

Integrated Services (Int-Serv)

- architecture for supporting real-time applications over the Internet Protocols and the Internet
- guaranteed delay bounds
  - absolute upper bound of delay
- link sharing
  - set maximum shares of a link
- predictive real-time service
  - stable delay
- overview - Informational RFC 1633

Integrated Services, contd.

- assume desire to use the Internet as common infrastructure for real-time and non-real-time communication
- two defined services
  - guaranteed
  - controlled-load

Integrated Services, contd.

- basic parts
  - admission control - determines if new flow can be added to existing load - policy and capacity question
  - classifier - determines class of incoming packet
  - packet scheduler - queues packets for transmission
    - reorders output queue
    - also requires an estimator to measure properties of outgoing packet stream
  - packet discarder
    - discard “excess” traffic
- not just traffic prioritization on a link

Integrated Services, contd.

- priority be itself is not enough
  - if too much high-priority traffic, prioritization does not help
  - need separate request process
    - not accepted if it would overload link / system
- requires flow-specific state in routers
  - change in basic Internet model
  - use soft state - can change on path change
    - vs. hard state - (set at start, release at end)
- may require request & flow authentication
- basically controls time-of-delivery of packets
  - absolute & variance

Int-Serv, Resource-Sharing

- multi-entity link-sharing
  - split one link between organizations
- multi-protocol link-sharing
  - split link between protocols (IP, SNA, IPX etc)
  - can help deal with different congestion responses
- multi-service sharing
  - application-based
  - e.g. limit amount of web use

Guaranteed Quality of Service

- deliver guaranteed delay and bandwidth service
  - guarantee = mathematically provable
- all nodes in path must cooperate
- only deals with max queuing delay
  - minimum delay not controlled
  - transmission delay fixed by nature
  - looks like an end-to-end wire per flow
- assumes edge policing
- includes reshaping at merge points in network

Controlled-Load Service

- looks like "unloaded network element"
- requires "active admission control"
- little delay over time scales longer than flow's burst time
- little congestive loss over time scales longer than flow's burst time

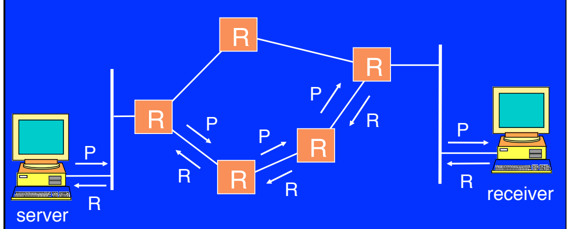
RSVP

- Resource ReServation Protocol (RSVP)
- implementation of INTSRV reservation process
- can be used to set aside resources for a specific application along a communications path
- can transfer the requests to a new path if rerouted
- may make use of QoS-active links
  - like ATM if there

RSVP Features

- unicast & multicast
- simplex (one direction per reservation)
- receiver-oriented
- maintains “soft state” in routers
- uses underlying existing routing protocols
- transports (opaque to RSVP) control & policy info
- can work through non-RSVP routers
- supports both IPv4 & IPv6

RSVP Process

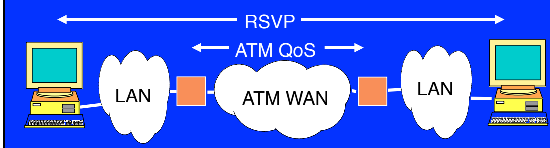


RSVP - Process, cont.

- using admission control, router
  - will accept reservation request if enough capacity
    - record reservation and forward **resv** to next-hop
  - if not - send **resverr** to previous hop
- state refreshed periodically with new messages
  - entry removed on timeout
- periodic refresh deals with reroute

Mixed QoS

- since only sure end-to-end technology is IP must use mixed QoS if want to use ATM QoS
- use IP signaling (like RSVP) to control link-level QoS (like ATM) when present



Mixed QoS, contd.

- create VC when needed for a path across ATM cloud
- can not change ATM QoS so must create new VC if path QoS chnages - then remove old VC
- map intserv QoS parameters to ATM parameters
  - RFCs published but ongoing work in IETF
    - RFC 2379 - RFC 2382

Flow Based QoS Issues

- scaling issues - per flow state an issue
- authorization (policy) issues - who says "OK"
- accounting issues - how to bill user
- security issues - theft / denial of service
- advanced reservations *very* hard
- good for long flows (video, audio, large file transfers)
  - flow setup cost must be low when averaged over flow length
- many mice on the Internet

Policy

- need to be able to say who can make reservations
- can be absolute
  - yes to Bill, no to Sally
- can be relative
  - Sally more important than Joe if limited resources
- can preempt
  - Fred can preempt Bill
- can be checked at various places in network
- part of general AAA problem

RSVP Admission Policy

- IETF working group
- separate policy server
- design is to have a router ask a policy server what to do when a reservation request is received
  - too much complexity to add to router
  - may be under different management
- router passes RSVP resev info to policy server
  - gets back hints about what to do
  - accept / reject / priority

RSVP Aggregation

- attempt to reduce per-flow messages in RSVP network
- aggregate path & reservation messages going between adjacent routers
  - or between ingress & egress routers
- multiple proposals

Flow Lengths in the Internet

from cic nets' Chicago hub

IP Flow Switching Cache, 16384 active flows, 0 inactive
132159644 added, 124468367 replaced, 4892577 timed out, 2782316 invalidated
statistics cleared 270640 seconds ago

Protocol	Total Flows	Flows /Sec	Packets /Flow	Bytes /Pkt	Packets /Sec	Active(Sec) /Flow	Idle(Sec) /Flow
TCP-Telnet	5222464	19.2	40	89	785.3	32.9	17.3
TCP-FTP	2087345	7.7	6	87	47.9	7.3	22.7
TCP-FTPD	1275958	4.7	95	390	449.5	21.9	23.6
TCP-WWW	83916123	310.0	9	304	2944.5	5.4	20.9
TCP-SMTP	14106833	52.1	8	173	448.9	6.4	21.6
TCP-X	94849	0.3	81	176	28.6	24.1	17.8
TCP-other	16095661	59.4	38	274	2290.8	20.9	21.5
UDP-TFTP	339	0.0	1	207	0.0	2.3	21.0
UDP-other	5059444	18.6	11	217	208.4	9.4	26.0
ICMP	4201689	15.5	2	83	46.0	5.2	26.8
IGMP	39809	0.1	30	398	4.4	48.2	29.4
IPINIP	9431	0.0	1808	254	63.0	147.1	18.6
GRE	32811	0.1	594	204	72.0	62.1	18.8
IP-other	909	0.0	3	223	0.0	1.2	31.8
Total:	132143665	488.2	15	260	7389.7	0.0	0.0

Inteserv / RSVP AS (RFC 2208)

- good stuff but some issues
  - SNA / DLSW is a good application
- scalability
  - high-bandwidth backbones not appropriate now
- security
  - needs better
- policy control
  - needs some

Non Flow Based Qos

- packet headers are "marked" at edge of network
  - precedence bits most common place to mark
- one or more bits used
  - two (priority and best effort) or more levels
- different mechanisms proposed
  - drop priority
  - queue selector - WFQ on queues
- contract with ISP, contract between ISPs
  - a problem if too much traffic for destination
- new (unproven) ideas
- creates N predictive Vnets on same Pnet

Non Flow Based QoS, contd.

- 1st model = “sender pays”
  - “receiver pays” will come later
- can use long or short term QoS contracts with ISP
  - dynamic requests for more bandwidth
- better scaling than per flow QoS
- easier authentication, authorization and accounting
- still much research needed

Non Flow Based Qos in the IETF

- Differentiated Services working group in IETF
- does not replace intserv /RSVP
- to define class-based QoS
  - replace earlier definition of use of TOS byte
- define behaviors not services
- may look at traffic shapers & packet markers
- must understand security issues

IETF Diffserv WG

- rename IP TOS Byte to “DS Field”
- components
  - mark bits in DS Field at network “edge”
  - routers in net use markings to determine packet treatment
  - conditioning marked packets at network boundaries
- deals with flow aggregates
- DS Field may change in flight
  - some disagreement - what about end-to-end?
- note! - diffserv not guaranteed service
  - does not know “destination”

IETF Diffserv WG, contd.

- base RFC published as a proposed standard
  - backward compatible with the IP precedence bits
  - old TOS bit meanings not supported
- deals with flow aggregates
- DS Field a codepoint
  - points to a Per Hop Behavior through a configurable mapping table
- unknown codepoint must be treated like best-effort
  - codepoints xxxxx0 - assigned by standards action
  - codepoints xxxx11 - experimental and local
  - codepoints xxxx01 - currently experimental and local

DS Byte

- rename TOS byte to be Differentiated-Services (DS) Field
- use to designate behaviors
  - not services to "customer"
  - build services from behaviors
- format

PHB	CU

PHB per-hop behavior
CU currently unused

PHB

- PHB = 000000 default (best effort)
- PHB = xxx000 ordered priority handling
  - backward compatible with precedence bits
- other proposals in process
  - EF - expedited forwarding
  - AF - assured forwarding group

Expedited Forwarding (EF)

- one PHB
- strict policing at edges
  - to ensure no overload in network
- produces a guaranteed service
- requires system to coordinate edge policing
  - proposal for a "Bandwidth Broker"

Expedited Forwarding, contd.

- departure rate of traffic must equal or exceed a configurable rate
- measured over any time interval equal or longer the time it takes to send one MTU sized packet at the configured rate
  - e.g. if configured rate = 1Mbps, time to average over is 12 msec (12, 080 bits)

Assured Forwarding Group (AF)

- set of PHBs
  - 4 sets of 3 PHBs
  - organized as 4 queues, each with 3 levels of drop precedence
    - traffic must be forwarded based on precedence - not absolute priority
  - no specific ordering between classes
- can be used to provide frame-relay like services
- assured rather than guaranteed
- depends on edge policing & marking
  - can remark drop precedence in net

AF, contd.

- requires RED-like function to drop excess packets
- two thresholds per drop precedence
  - thresholds based on averaged queue depth
  - min thresh - point below which no traffic is dropped
  - max thresh - point above which all traffic is dropped
  - probability of drop increases linearly from 0 at min thresh to 1 at max thresh
- can be used to implement "Olympic" service
  - gold, silver, bronze - with different drop precedence values

CU

- reserved for future
- could be used for congestion experienced

Packet Ordering

- bad idea to reorder packets in a "microflow"
  - single instance of an application-to-application flow of packets which is identified by source address, destination address, protocol id, and source port, destination port (where applicable).
- i.e. don't put packets from the same microflow in different queues

Traffic Conditioners at Edges

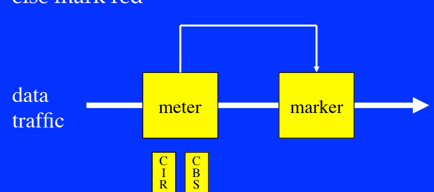
- packet classifiers
  - use fields in packet headers to steer processing
- markers
  - set DS field
- policer
  - monitor traffic & react if profile exceed
    - drop, remark packets
- shapers
  - modify packet flow to control TCP flows

Packet Marker / Remarker

- marks packets based on input conditions
- could be type of traffic
  - web vs. email vs. file transfer
- could be traffic level
  - e.g. “A Three Color Marker”
  - mark packet with AF drop probability based on traffic
  - three parameters
    - Committed Information Rate - CIR
    - Committed Burst Size (CBS)
    - Excess Burst Size (EBS)

Three Color Marker, contd.

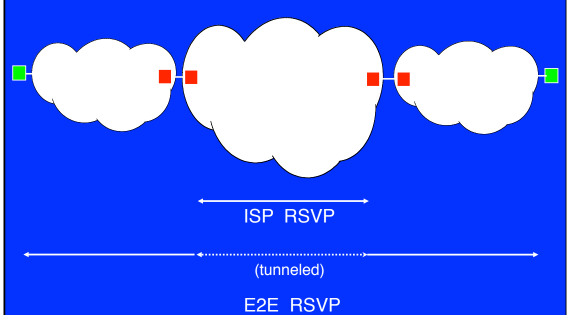
- uses two token buckets - CIR & CBS
  - if incoming traffic fits in CIR bucket - mark green
  - if not fit in CIR but does fit in CBS - mark yellow
  - else mark red



RSVP as signaling

- much thought about using RSVP for signaling between host and “local” marking device
- could also be used in backbone to see if capacity available
  - when to release is a problem
- some see RSVP as a general signaling protocol
  - e.g. MPLS

RSVP Signaling for Diffserv



Bandwidth Brokers

- policy system
- sub-allocate class allocations
- could do dynamic request for allocation
- not a current diffserv work item
  - non-IETF work underway

Policy

- AAA (authentication, authorization & accounting) an issue
  - is there one or more "answer"?
  - major problems in defining problem set
  - is it OK for user X to use service Y?
  - how account for use?
  - . . .

QoS Between ISPs

- both diffserv & RSVP
- hardest problem is policy not technology
  - $$$$

Issues

- policy
  - when to give a busy signal
- end-to-end?
- $$$$
  - what billing info is needed?

A Different View

- is adding bandwidth all that's needed?
- Andrew Odlyzko of AT&T Labs
  - may be cheaper to just throw bandwidth at QoS problem
    - 1 - only a few points of congestion
    - 2 - 80% of data com costs non-transmission
    - 3 - adding QoS complexity will add to other costs
      - labor, management & billing systems etc
    - 4 - local part of data com dominate overall cost
    - 5 - cost of transmission coming down
      - Fortune reports - 99.8 Tbps capacity by 2001 = glut
  - upgrade congested points - cheaper than QoS complexity