Optimizing Converged Cisco Networks (ONT)

Quality of Service (QoS) Overview

http://www.INE.com

Network Metrics Overview

- Bandwidth
  - How much traffic you can send at a time
    - e.g. pipe “width”
  - Measured in bps
- Delay
  - How long it takes traffic to get there
    - e.g. pipe “length”
  - Measured in ms
- Jitter
  - Variation in delay
    - e.g. different pipe “lengths” along path
  - Measured in ms
- Loss
  - Traffic dropped in transit
    - e.g. leaky pipes ☺

Data vs. Voice Requirements

- Data
  - Generally higher bandwidth requirements
  - Larger packet sizes
  - Can tolerate delay and jitter
  - Can tolerate loss
    - i.e. TCP retransmission
  - Can tolerate network reconvergence time
- Voice
  - Lower bandwidth requirements
  - Smaller packet sizes
  - Cannot tolerate delay and jitter
  - Cannot tolerate loss
    - Too late to retransmit once dropped
  - Highly sensitive to network reconvergence time

Bandwidth Issues

- Bandwidth is a physical function of the circuit
- Lowest bandwidth circuit along the path becomes the "bottleneck"
  - e.g. FastEthernet LAN to T1 WAN
- There is never enough bandwidth
  - e.g. TCP by nature will take 100% of pipe
- How can we deal with this other than buying larger circuits?

Delay Issues

- Time that traffic takes to arrive can be broken down into…
- Processing delay
  - How long did the device take to make a decision?
    - Software and hardware function
    - e.g. CPU, CEF, ACLs, etc.
- Queueing delay
  - How long did the traffic have to wait in line?
    - Software and hardware function

Delay Issues (cont.)

- Serialization delay
  - How long did it take to put the frame on the wire?
    - Hardware interface driver function only
- Propagation delay
  - How long did it take to get there?
    - Electrical / optical / wireless function
- How can we deal with this other than buying faster hardware?

Loss Issues

- Packet loss usually occurs due to "buffer overflow"
  - Router has no more space to hold the traffic
  - i.e. "congestion"
- Packet loss tends to have a snowball effect
  - e.g. "global TCP synchronization"
    - More on this later…
- Some loss is acceptable, but how can we minimize it?

The Solution: QoS

- Quality of Service
- Method of preferring services for particular types of traffic while sacrificing others
  - e.g. guaranteed delivery vs. best effort
- Goal is to make the network more predictable
  - e.g. I know my worst case delay for VoIP is 100ms
  - e.g. I know that FTP should get at least 5Mbps

QoS Classification

- To offer different services, we must know who is out there and what they need
- "Classification" is used to separate the services and their needs, such as…
  - VoIP – minimum delay
  - Database replication – guaranteed delivery
  - Email – best effort
- Without proper classification, entire QoS model falls apart

QoS Enforcement

- Once we know what the needs are, the network must enforce our "policy"
- QoS policy defines how network devices actually deal with the traffic, such as…
  - Prioritization
  - Queueing
  - Shaping
  - Policing
  - Etc.

QoS Models

- QoS implementation can be generalized into three “models”
  – Best Effort
  – Integrated Services (IntServ)
  – Differentiated Services (DiffServ)

Best Effort

- Effectively no QoS
- All traffic is given the same classification (priority)
- Very simple implementation
  – e.g. First-In-First-Out (FIFO) queueing

IntServ QoS

- RFC 1633 *Integrated Services in the Internet Architecture: an Overview*
- Goal was that QoS would be “integrated” into the applications and the network at the same time
- Example
  - Host application asks the network for 64kbps of low delay service
  - Devices in the network transit path check for available resources and grant or deny the request
- Uses RFC 2205’s *Resource Reservation Protocol* (RSVP) for the actual implementation

IntServ QoS Problems

- Assumes application support
  - Do we really want the end devices controlling the network’s resources?
- Assumes network maintains the state
  - Do we really want to know about every flow in the network?
- Does not scale, and is rarely implemented with few exceptions
  - e.g. MPLS Traffic Engineering (MPLS TE)
- In Cisco IOS RSVP only makes reservation in the “control plane” not the “data plane”

Copyright © 2009 Internetwork Expert, Inc
www.INE.com

DiffServ QoS

- RFC 2475 *An Architecture for Differentiated Services*
- Goal is that the network itself classifies flows (differentiates between them) and enforces the policy
- Classification plus the policy defines the Per Hop Behavior (PHB)
  - e.g. each router makes an independent decision
- Uses fields in the IP header (amongst other things) to define the PHB
  - RFC 2474 *Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers*

DiffServ QoS Problems

- Assumes correct classification
  - What stops a host from marking all its traffic as priority?
- Assumes network enforces the policy
  - What methods can we use to actually guarantee the services?
- Despite these, DiffServ is more scalable and the preferred implementation

QoS Overview Q&A