# Evolution of Quality of Service in IP Networks

Kathleen Nichols
knichols@ieee.org

March 30, 2004



## Outline

- Background and history
- Differentiated Services approach
- Services
- Examples

## Why “evolution”?

- The focus of early Internet work was robust connectivity. QoS was not an important topic of research or focus
- Early ideas were not fleshed out into an architecture
- An approach in the early 90’s had an architecture at odds with that of the Internet so did not get deployed
- Current approach more compatible with Internet architecture but purposely set up to be “evolutionary”
- Specific solutions are still evolving



## Quality of Service (QoS) for IP Networks

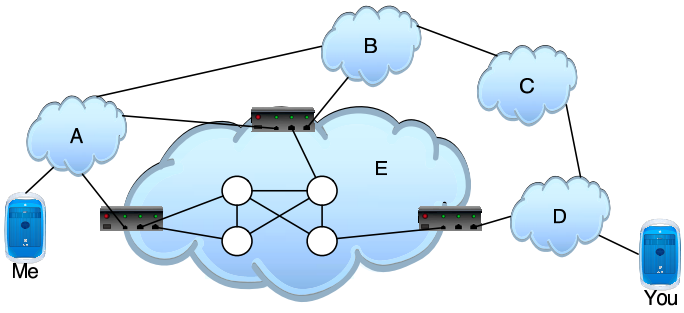
Adding differential QoS to a network can’t and doesn’t add bandwidth. If some packets get better treatment, others will get worse treatment.

A workable QoS architecture is one that provides a framework to manage this unfairness according to current policy governing the network.

Though the verification of differential QoS relies on an ability to quantify the treatment a particular packet can expect as it transits a network and to measure that quantity,

Who gets better service? and Who decides? are critically important questions for any viable QoS architecture

## The Environment for IP QoS






## QoS in Internet Pre-History

Paul Baran invented much of the foundation of packet switching while at RAND in the early 1960's. (www.rand.org/publications/RM/baran.list.html)

In RM-3638-PR (1964), a Communications Control Console (or Priority Control Console) is described and it makes wonderful reading for perspective about the role of QoS control (or allocation of resources). Significantly, Baran states:

> "The console does not seek to supplant human judgment. It simply provides an automated facility to instantaneously implement the human executive decision."

which clearly puts the technical implementation in its place as a way to realize the human policy decisions

## Early Internet QoS

The Internet Engineering Task Force (IETF) makes standards for the Internet and publishes them as RFCs available at http://www.ietf.org/rfc.

RFC 791 (1981) defined a Type of Service (TOS) octet in the IP header with a 4-bit TOS field:

> "The Type of Service provides an indication of the abstract parameters of the quality of service desired. These parameters are to be used to guide the selection of the actual service parameters when transmitting a datagram through a particular network. "

Which is notably vague on both what the service parameters might be as well as what functions might be needed in the forwarding path.



## Early QoS (continued)

The TOS octet had a 3-bit Precedence field:

> "The Network Control precedence designation is intended to be used within a network only. The actual use and control of that designation is up to each network. The Internetwork Control designation is intended for use by gateway control originators only. If the actual use of these precedence designations is of concern to a particular network, it is the responsibility of that network to control the access to, and use of, those precedence designations."

No architecture developed, though another effort was made in (obsolete) RFC 1349. Directives like "maximize reliability" or "maximize monetary cost" were attached to certain bit-patterns in the TOS octet, but this did not translate to an architecture.

## Integrated Services (IntServ)

Interest grew from the early experiments with Internet audio and video.

Overview (RFC1633), published in 1994, identified both “real-time” and “controlled link sharing” as useful. Additional standards in 1997.

Though RFC1633 noted the need for such forwarding path primitives as classification and packet scheduling, it made the contention that flow-level admission control and resource reservation were *required* and that:

> “there is an inescapable requirement for routers to be able to reserve resources, in order to provide special QoS for specific user packet streams, or "flows". This in turn requires flow-specific state in the routers, which represents an important and fundamental change to the Internet model. The Internet architecture [h]as been founded on the concept that all flow-related state should be in the end systems.”



## IntServ Model breaks the Internet

Further, IntServ proposes an admission control model very similar to telephony:

> “Admission control is invoked at each node to make a local accept/reject decision, at the time a host requests a real-time service along some path through the Internet.”

This was so at odds with the Internet’s control of resources and ability to scale that it has not been adopted. Fortunately, there is a better way.

(The signalling protocol for IntServ, RSVP, ReSource reServation Protocol in RFC 2750, has been adapted for a range of uses.)

## Differentiated Services

Diffserv's policy model is based on an Internet made up of independently administered domains, each of which is connected to at least one other

QoS is provisioned on a network for a traffic aggregate, not for a flow

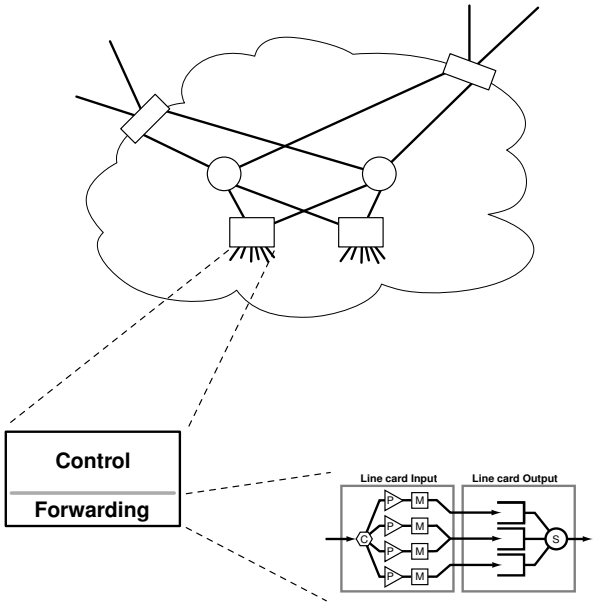
Flows (or any other useful traffic partition) are admitted into a particular aggregate at the edge of the network.

Standards action (RFC 2474) redefines TOS octet: a 6-bit DSCP field (bits 0-5) is used to identify a packet's traffic aggregate.

The forwarding path treatment required at each node (per-hop behavior or PHB) internal to a network is indexed by the DSCP



## Fitting QoS to the Internet



## QoS Control Plane

Routing is the control plane function for packet forwarding

- it configures a forwarding table that is used by packet forwarding to determine a packet's output interface from its packet header
- within a domain (intradomain routing) there are several options
- between domains (interdomain routing) there is one, BGP, but this evolved over time

Diffserv's control plane configures a behavior table used by QoS-enabled packet forwarding to determine an output queue. Expect multiple methods of configuring the behavior table within a network and QoS between networks to evolve slowly.



## QoS Forwarding Path Primitives

Classifier (C): takes apart (input) packet stream. Two types: a multiple field (MF) classifier or a behavior aggregate (BA) classifier. MF classifiers filter on an arbitrary range of IP header fields. Must be capable of running at line rate.

Policer (P): enforces the rules governing packet substreams. Contain meters used to measure a traffic stream against a traffic profile. Packets which do not conform are forwarded to a particular policing action which might include dropping or re-marking (to become part of a different aggregate).

Marker (M): propagates QoS information about the packet downstream. Particular forwarding treatments are determined by the "mark" that appears in the packet's DSCP field. Must be capable of writing a six-bit DSCP into the packet's TOS octet at forwarding rate.

## More QoS Forwarding Path Primitives

Queues (Q): *isolate* traffic aggregates from each other; more queues at an output interface enable more isolated traffic aggregates. Large number might be useful at some boundaries, but when provisioning aggregates for an entire domain, unlikely that more than a small number will be practical.

Sharing/Shaping (S): constructs an (output) packet stream based on local policy and downstream agreements. Selection of next queue to transmit. Delivering a fixed bandwidth, independent of other traffic, requires time-based queue service (traffic shaping, e.g. appendix of WO0030307A1 at www.delphion.com). Delivering relative link shares requires a WRR variant. A CBQ scheduler allows a range of flexible mechanisms.

These are configured to deliver a particular PHB.



## Traffic Conditioning

- functions that can be applied to a behavior aggregate, application flow, or other operationally useful subset of traffic, e. g., routing updates
- used to enforce agreements between domains
- used to condition traffic to receive a differentiated service within a domain by marking packets with the appropriate DSCP and by monitoring and altering the temporal characteristics of the aggregate where necessary
- may alter the temporal characteristics of a behavior aggregate to conform to some requirements of a particular domain
- includes metering, policing, shaping, and marking

## Per-Domain Behaviors (PDB): network-level QoS specs

- the behavior experienced by a particular set of packets as they cross a Diffserv cloud.
- specifies both the forwarding path treatment and the edge rules for its traffic aggregate
- is characterized by specific metrics that quantify the treatment a set of packets with a particular DSCP will receive as it crosses a domain

These metrics should follow the general framework of those that are provided in SLAs, but should permit a wider range of guarantees

PDBs include Lower Effort (RFC 3662) and Virtual Wire (in revision). Status of Assured Rate PDB is unknown.



## Services and E2E QoS in the Diffserv Framework

- Services are built by adding rules to govern traffic aggregates:
  - ➥ initial packet marking
  - ➥ how particular aggregates are treated at boundaries
  - ➥ temporal behavior of aggregates at boundaries
- Different user- visible services can share the same aggregate
- Services must be sensible and quantifiable under aggregation
- Start by defining general issues for delivery of QoS services in a network

## Considerations for QoS in a Network

- Within a network cloud, QoS is allocated according to some locally determined set of rules (policy)
- Almost all the complex forwarding path work is confined to the boundaries of clouds and derives directly from the rules

➥ Might be possible to apply an “off-the-shelf” PDB

- Rules might not be symmetric across a boundary
- QoS information exchanged between networks is confined to boundaries and, where networks have different owners, covered by bilateral agreements



## Issues in Allocating QoS

Assigning QoS by application is a non-starter

- File transfers might be low-priority to Marketing and critical for Accounting
- VoIP might be critical for University administration and low-priority for University dorms

Preferential network access is resource that should be doled out between organizations according to overall organizational budget

- Marketing uses its local policy to allocate its chunk between its users
- Diffserv permits indicating QoS level with a DSCP in each packet; determination of packet’s DSCP happens at the edge and can be based on who/what/when/how much?

## Organizational Policy and Technical Requirements

- Technical specifications of the network and the desired special characteristics determine *how much* total special traffic is feasible
- Organization's policies determine *who* gets *what* share of special treatment and (possibly) *when*
- This policy information is used to (possibly) evaluate requests for special service from autheticated users and to configure the network's classifiers and traffic conditioners to control access of packets to special treatment
- Organizational allocation of the special treatment resource may be on a monthly, quarterly, or annual basis. Configuration of the network's boundaries may be static for that period or dynamic, in response to requests which are audited against those allocation policies

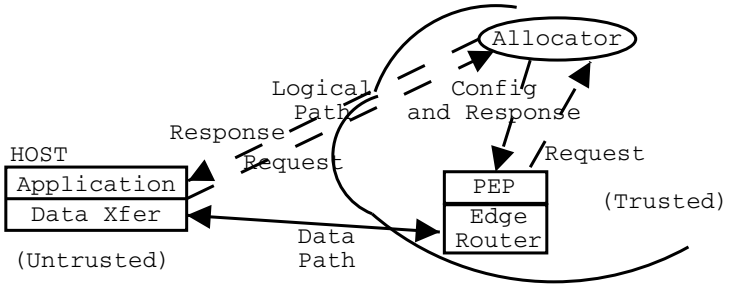


## Controlling the Boundaries

- A repository of of policy is needed to keep track of priorities and limits on QoS allocations for individual users, projects, and/ or departments.
- An entity needs to receive requests for QoS, consult and update the database, and send configuration information to the routers where indicated.
- This entity is sometimes called a bandwidth broker (BB) (V. Jacobson, RFC2638).
- BB is part of the network infrastructure and must authenticate requests from users. Some information can also be configured.
- Intradomain policy decisions and implementations remain up to each domain.

## Components of Intradomain QoS Allocation

Requests, responses, and configuration are control path




**Config** data sets classifiers to check for permitted packet header fields (can be *any* required signature, e.g. MAC, src IP, dst IP, DSCP) and sets traffic conditioners to enforce characteristics on data flows



## Flow of Requests and Configuration

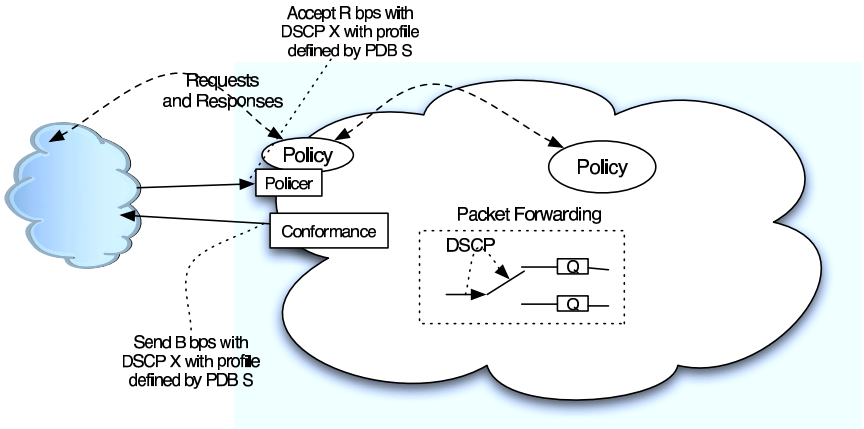
- Requests can come from many sources: network admins, applications, hosts, trusted signals (RSVP?)
- Requests must include requestor's identifying information as well as identifying info for the flow, microflow, or behavior aggregate for which the request is intended
- Responses directly from allocator across untrusted boundary can contain "cookies"
- Agnostic about signalling
- Configuration can come from static or signalled requests and can be done manually or by a BB or other QoS agent

## Allocation at Work

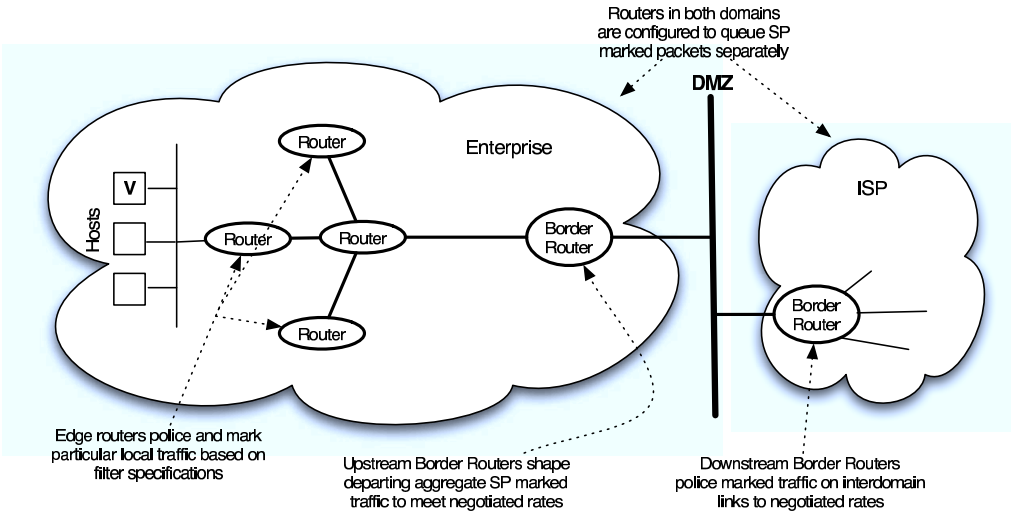
- A host (or other entity) launches a request toward the intended destination (RSVP-like)
- At a Policy Enforcement Point, recognized as a request and sent to BB's allocator
- Allocator checks policy database, credentials, time of day, etc and sends back "yes" or "no" to PEP (or host). If "yes", gives DSCP and policer information for PEP to enforce, host to conform
- Enforcement can be set by PEP to classify on DSCP and IP dst or any MF classification as deep as MAC address, polices on given rate and burst
- Cryptographically sign the "yes" message if it goes from allocator to host to PEP



## Connecting Networks: Evolution to Come

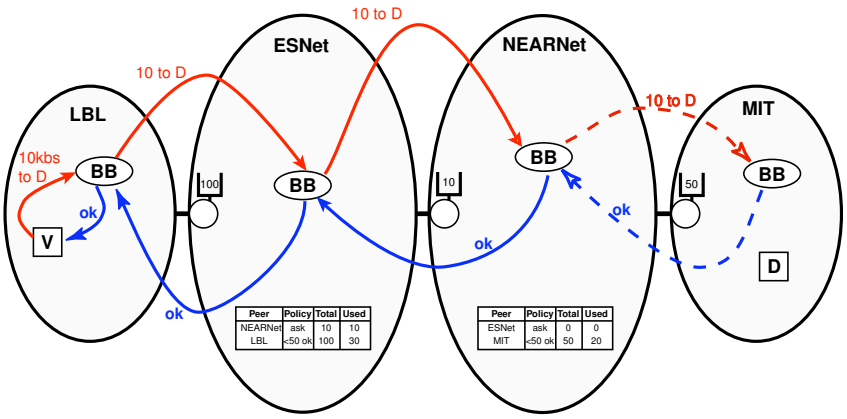


# Connecting an Enterprise to an ISP with Differential QoS






# Dynamically Connecting Network QoS




(From RFC2638)

V asks LBL's BB for a specific level of service. The request is passed along between BBs. May or may not require a signal to MIT.

## Some Words about the Examples

- Every application of QoS will have its own idiosyncracies. These examples are more motivational than recipe.
- Intent is to make solutions no more complex than necessary
- In these examples, resources (bandwidth) are **provisioned** for a specific desired QoS, **allocation** parcels out the provisioned resources according to policy, and **resource control** ensures allocation is not violated.
- Allocation and resource control may be static, dynamic or some combination

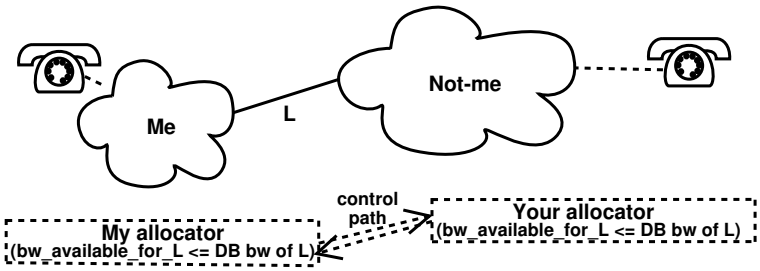


## Voice Example: Inside a Network Cloud

Inside a single cloud, VoIP bandwidth is often trivial and flows are not worth tracking. Provision the network, consider using IP phones that mark DSCP, police if necessary.

- Configure a Delay Bound PHB (RFC 3248) at each network node sufficient to handle all calls of a network (i.e., if all the IP phones are active at once or some percentage determined use Erlang models). Note that 100 Mbps/ 64 Kbps > 1,500 and 100 Mpbs is 10% of 1 Gbps.
- When a call is initiated, check if the destination is within the cloud. If so, just admit the call
- May want to set up classifers/policers on the edge routers to ensure no spoofing. Policing for voice streams (no bursts permitted) should be adequate since not very useful for other kinds of traffic

## VoIP Across Two Network Clouds




Link *L* is the scarce resource

This could apply to a remote office of a company as well as two administratively different clouds

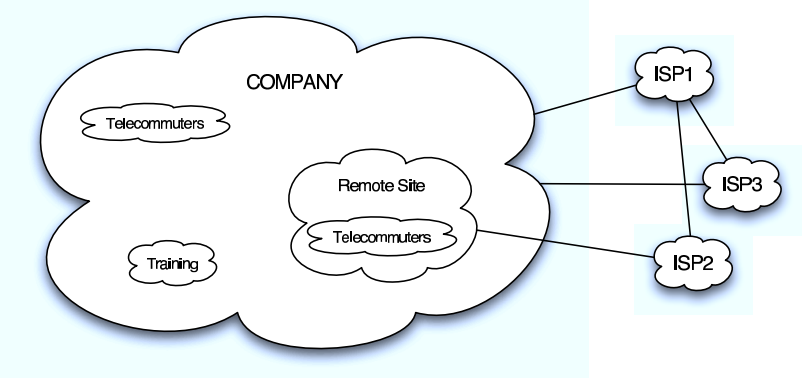


## Voice Example: Across Two Clouds

Locally, track connections only where resources are limited, e.g. link to next cloud.

- Assume the link is configured for a DB rate that sufficient to handle all outside calls most of the time (*bw_ available*). Erlang models from literature.
- When a call is initiated, check destination
- If its **not-me**, check: (*bw_ available - call_ bw*) >= 0?
- If not, refuse call : If yes, signal/ message next cloud and wait for reply
- If reply is positive, *bw_ available -= call_ bw* and proceed

## Network Domains within an AS (mostly trusted boundaries)




Company contains a number of separate domains where QoS allocations can be handled locally and some rules are enforced at the boundaries. Since the domains are all part of the same Autonomous System and have the same owner, some rules at these boundaries might be more “relaxed.”



## Mission Critical: Special Handling for Ordinary IP Traffic

Gets a minimum percentage or bandwidth from each link, even under congestion. No guarantees won’t be congestion in the class. Then ask: What does the organization look like and where are the scarce resources (bottlenecks)?

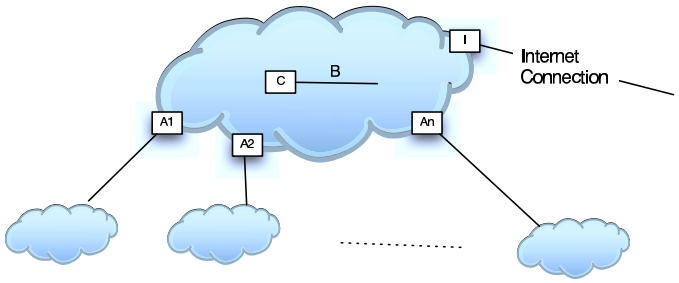
**Case 1:** All agree on what’s special: CVS, HR, inventory transactions.

**Solution:** Make the src/dst addresses of those servers “magic” and either:

1. Classify on those addresses at bottlenecks and steer to special queue
2. Classify on those addresses at edge and mark with a DSCP for special queue

## Case 2: SubNetworks Share a Transit Network

Several (*n*) fairly autonomous networks share a network for transit




**Per-cloud:**

sub-organizations decide what's *mission critical,* handle internally

sub-organizations control access to *special* class which passes through **A*i***



## Case 2 continued

**Organization:**

Decide what's corporate mission critical, check at **A*i*** entry with MF classifier, DSCP mark may be checked or marked here

Decide which sub-organization gets what share of Internet link (or, more generally any bottleneck, **B**) and classify at **I** or (**C**) into queues (marking is not necessary)

Decide which sub-organization gets how much special traffic, check at **A*i*** entry with DSCP classifier, discard or re-mark if exceeds allocation

## All Together

This network could have separate service types mapping to different queues, for example:

- Configure Class Selector PHBs (from RFC 2474) to requirements. (A CBQ or WRR type scheduler like DRR would work.)
  - ➥ voice (10% DB PHB using CS5)
  - ➥ corporate mission critical (e.g. CS4 with a share of 50%)
  - ➥ special (e.g. CS3 with a share of 25%)
  - ➥ best effort (e.g. Default PHB using CS0)
  - ➥ even possible to use a Lower Effort service



## Does Path Matter?

There seems to be a level of uneasiness when the complete path is not specified for packets expecting a certain treatment, thus the attraction of complex approaches that start with fixing the path.

Each application of differential QoS must be examine to determine if the path must indeed be fixed due to the control and forwarding complexity of doing so.

When it is necessary to fix the path, it is preferable to do it with the mechanisms that are closest to the original aims of IP, e.g. use traffic engineering extensions to routing protocols that use DSCP as an additional field to determine next hop as opposed to MPLS which pins entire paths.

## Diffserv in Use

Genuity (before it was purchased) had announced differential QoS offerings. British Telecom is offering Class-of-Service in conjunction with MPLS (http://www.btglobalservices.com/en/products/uk/ipclear/)

Thomas Telkamp of Global Crossing described using Diffserv, provisioned empirically, to create a separate queue for VoIP (bounded delay) so that the delay bounds could be maintained by only overprovisioning that queue, not the entire network. (www.nanog.org/mtg-0210/telkamp.html)

Telecordia has been working on diffserv QoS and Bandwidth Broker architectures as reported in:

"A Simple Admission Control Algorithm for IP Networks", Keith Kim, Petros Mouchtaris, Sunil Samtani, Rajesh Talpade, Larry Wong, Proceedings of International Conference on Networking, July 2001. "A Bandwidth Broker Architecture for VoIP QoS", Keith Kim, Petros Mouchtaris, Sunil Samtani, Rajesh Talpade, Larry Wong, Proceedings of SPIE's International Symposium on Convergence of IT and Communications (ITCom), Colorado, August 2001. "An Integrated IP QoS Architecture - Performance", Byungsuk Kim, Isil Sebuktekin, Proceedings of MILCOM'02, Anaheim, CA, October 2002.



## A DIffserv Architecture

Cable Labs (www.cablelabs.com) has created standards for IP data over cable.

Documents cover a wide range of specifications and include the PacketCable$^{TM}$Interdomain Quality of Service (PKT-SP-IQOS-I01-001128) with its discussion of interdomain differential QoS for that environment

## What Barriers Remain?

- Current economic drivers
- Overspecifying the solution: diffserv is incrementally deployable and evolvable
- Classifier problems: difficulty in configuration, limited capabilities though better technology is deployed some places
- Absence of useful network monitoring tools
- FUD: overcomplicating the solution, etc