Service Peering and BGP for Interdomain QoS Routing

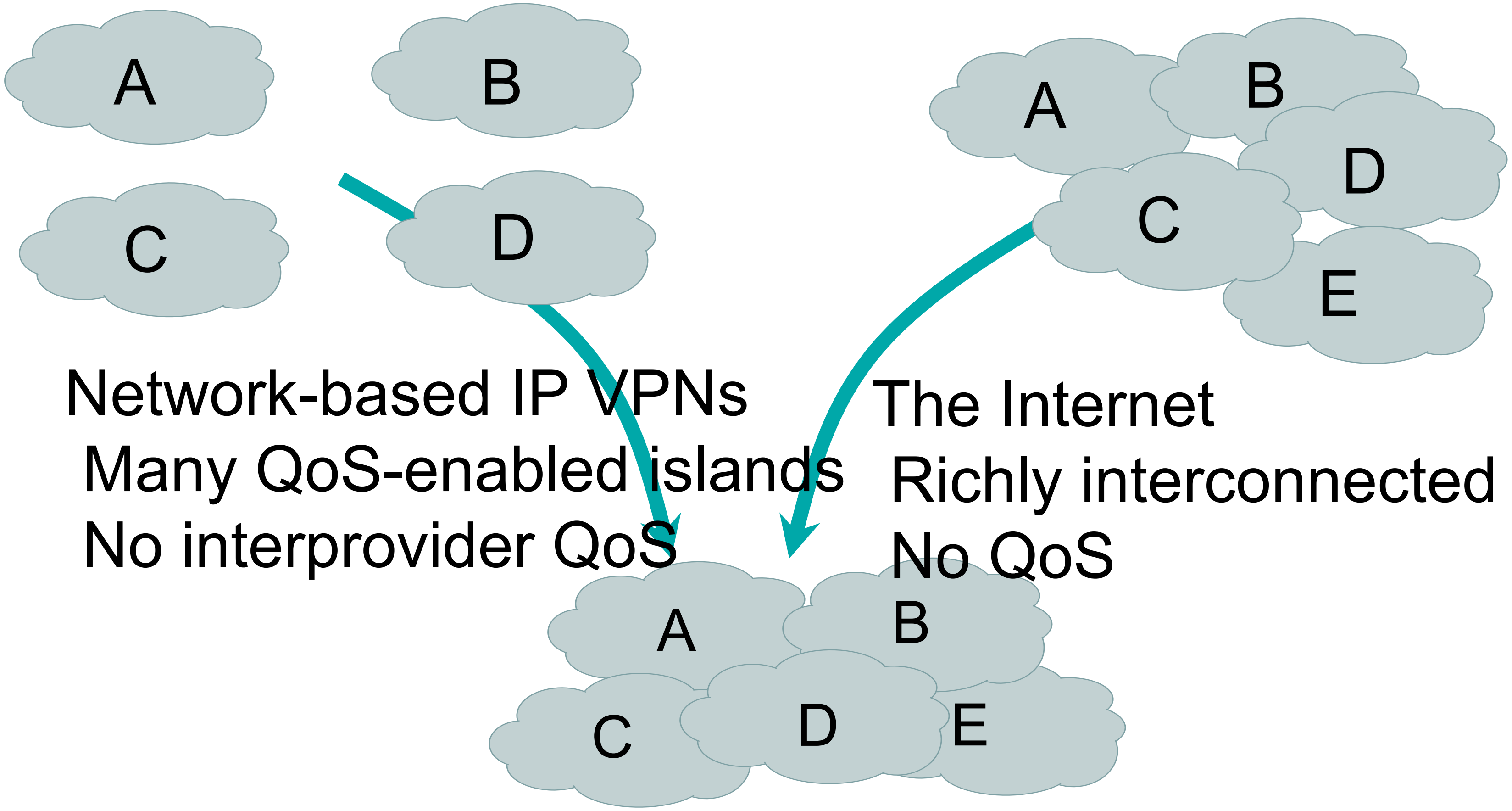
David Ward, John Scudder

mailto:dward@cisco.com

mailto:jgs@cisco.com

Motivation

Introduction — VPNs, the Internet, & Nirvana



Network-based IP VPNs

Many QoS-enabled islands

No interprovider QoS

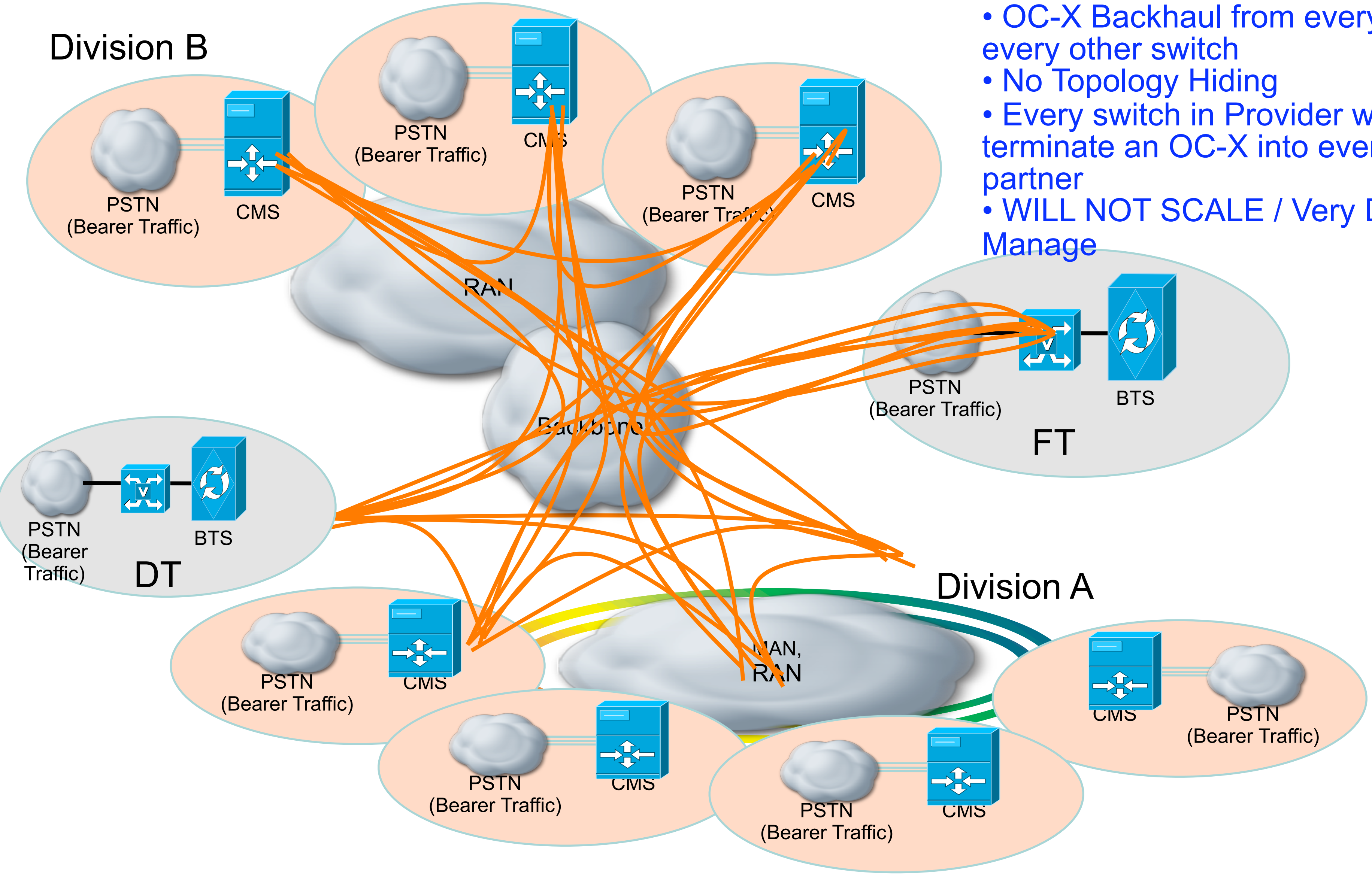
The Internet

Richly interconnected providers

No QoS

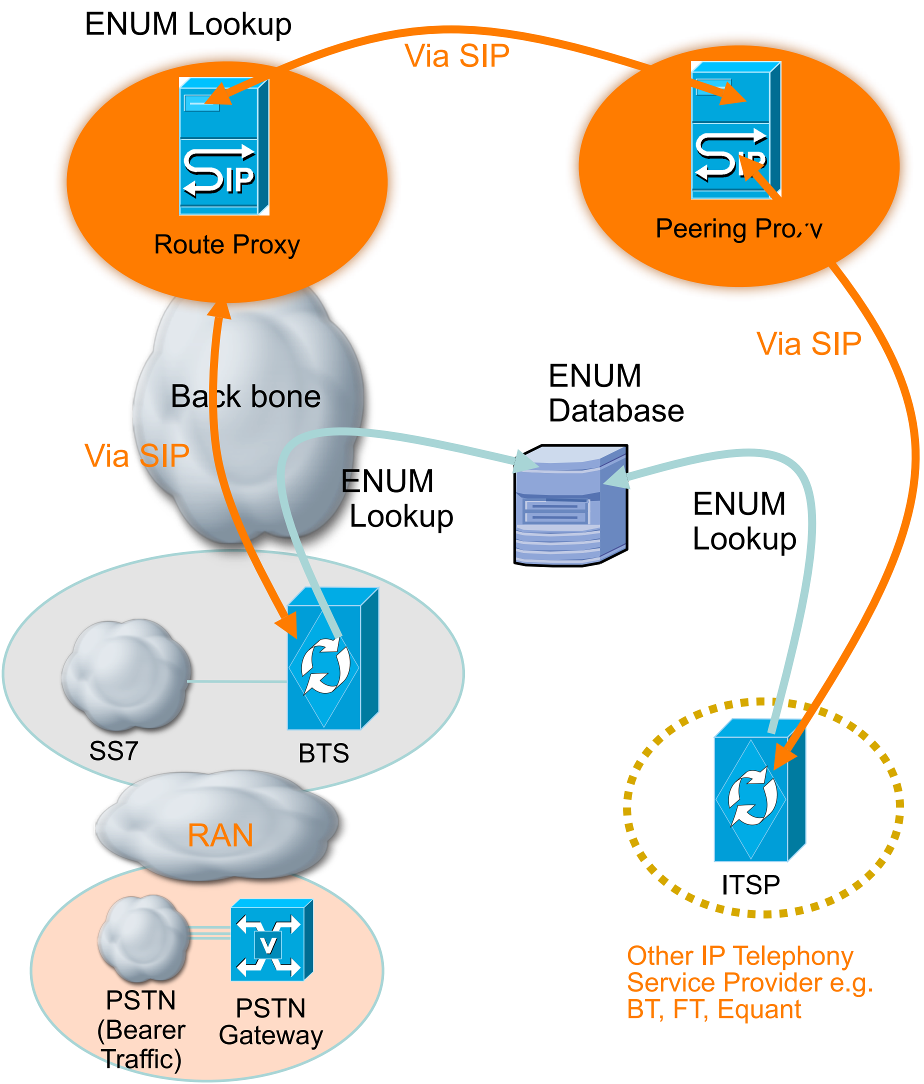
Nirvana: Richly connected AND QoS-enabled

Why Hierarchical Network with IP Peering is necessary…



- OC-X Backhaul from every switch to every other switch
- No Topology Hiding
- Every switch in Provider would have to terminate an OC-X into every peering partner
- WILL NOT SCALE / Very Difficult to Manage

Service Provider Peering – Via SIP



Local Switch - DN2 Table Lookup

Local Switch - ENUM Dip

Call forwarded to Route Proxy

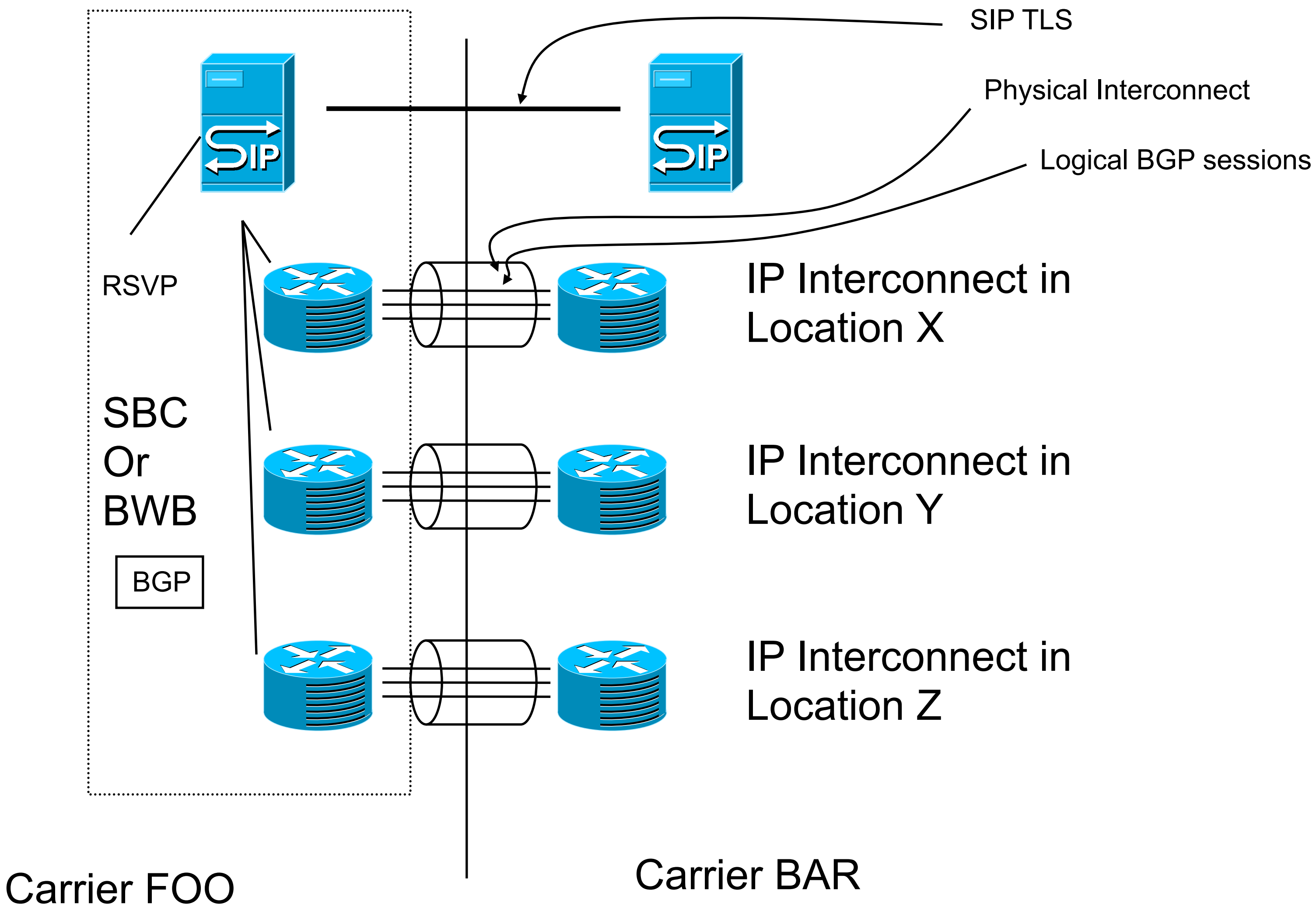
Route Lookup, call forwarded to Peering Proxy

Peering Proxy – Call forwarded to terminating Partner Peering Proxy

Voice bearer path setup between originating MTA and terminating CPE

No relation to routing topology

Architectural Reference Model: *Integrate*



Biggest Issues

CAC — See RSVP proxy and SIP proxy integration and network state check precondition

Interdomain TE Guarantees — MIT Consortium

Interdomain QoS and Routing — BGP

A Plethora of Fora

IETF

- Inter-AS TE and VPNs progressing
- IPPM for Measurement
- No current group for interprovider QoS (MAVS forming)
- Protocol definitions today are inadequate

ITU

- Has done some work in the past (e.g. Y.1711)
- Could probably do it all in the future

MPLS Frame Relay Alliance

- Started work in this area - MPLS centric view

IPSphere

MIT “Communications Futures Program”

MIT CFP - Led by Dave Clark

Feedback from many network operators, enterprises that are involved was:

- We need a multivendor forum
- Don’t want to go to IETF yet
- IPSphere is not sufficiently working on extensions (aka marketing)

MIT CFP was an existing framework

- http://cfp.mit.edu
- Willing to host a group on interprovider QoS - first meeting October 2004
- http://cfp.mit.edu/qos/slides.html - agenda, slides & agreements from 2nd meeting (Jan 2005)

MIT CFP

Currently working on a whitepaper that roughly follows the IDQ approach

- Numerous service provider co-authors + Cisco + Juniper
- Could become basis for an IETF submission: MAVS? and IDR work

BGP

BGP Functionality

What can BGP do?

- Find routes which (purport to) support a given QoS e2e

What can’t BGP do?

- Treat QoS as anything other than opaque
- Signal dynamic path characteristics (e.g., instantaneous loss or delay)

BGP for QoS Routing

BGP well-suited to carrying multiple classes of routing information

Consider QoS as a distinct class of routes

- Service classes, metrics, etc are opaque — BGP simply signals reachability

Small number of classes = tractable problem

Issues

BGP multiplexes all routing information onto a single session

- Undesirable fate-sharing between classes of routes
- Not possible to prioritize different classes of routes (on Rx side anyway)

BGP converges slowly in some cases

No means of carrying multiple routes for same NLRI

- For service separation
- For QoS

Some Solutions

Multisession to fix fate-sharing

Convergence

- Withdraw routes more efficiently
- Advertise more backup routes

Several options to distinguish multiple routes

- New AFI/SAFI
- Distinct session per QoS
- Agree upon and exchange QoS markings

Solution Assumptions

Must have opaque semantics for QoS bits on either side of AS boundary

- On link across boundary may administratively configure marking
- Re-mark at borders

May want to have distinct logical links for each QoS class OR multiplex QoS classes across one link

Want to have minimal changes to protocol for ease of deployment

Some Solutions

Multisession to fix fate-sharing

Convergence

- Withdraw routes more efficiently
- Advertise more backup routes

Several options to distinguish multiple routes

- New AFI/SAFI
- Distinct session per QoS
- Agree upon and exchange QoS markings

Multisession BGP

Moves multiplexing to transport layer (where it belongs)

No requirement for multiple loopbacks

Minimal configuration (for default behavior)

Support for multiplexing selected AFI/SAFI (”grouping”)

Easy to comprehend, manage and configure a new BGP peering session

Multisession High Availability

Multiple sessions can…

- Terminate on different processes (fault isolation)
- Terminate on different processors (performance isolation)
- Be serviced in priority order
  - Normal BGP session must be serviced FIFO

Relevance of Multisession BGP to QoS

Classes of routes can be divided by service class (gold/silver/bronze, etc)

Once divided, fault isolation, performance, prioritization can be applied

Issue is no administrative marking across AS boundaries

- Complete human intervention

Some Solutions

Multisession to fix fate-sharing

Convergence

- Withdraw routes more efficiently
- Advertise more backup routes

Several options to distinguish multiple routes

- New AFI/SAFI
- Distinct session per QoS
- Agree upon and exchange QoS markings

Withdraw for Multiple Destinations

AKA “Aggregate Withdraw”

BGP enhancement for single message withdraw

Use associated community for all related prefixes

Withdraw the community in one message and all prefixes are withdrawn

- Examples:
  - Withdraw all routes for a given QoS
  - Withdraw all routes for a given border router

To be discussed in IDR WG

Some Solutions

Multisession to fix fate-sharing

Convergence

- Withdraw routes more efficiently
- Advertise more backup routes

Several options to distinguish multiple routes

- New AFI/SAFI
- Distinct session per QoS
- Agree upon and exchange QoS markings

Aside — Route Reflectors

Route reflectors are used in IBGP to be able to scale "Full Mesh" requirement

- Adds server that can select the 'best path' from a number of clients and reflect it back to clients

Can be deployed in a hierarchy

Easily fits model of scaling QoS and even having an RR per service

In some topologies, converge slower

- Due to hiding of available backup routes
- Therefore convergence time may not meet QoS SLAs

Advertise Extra Backup Routes

ADD_PATH proposal discussed in IDR

- Advertise multiple paths for same prefix without new paths implicitly replacing previous ones.
- General purpose mechanism

Identify backup path at each RR

- Then propagate using ADD_PATH
- Increases state in network
- But eliminates transient black holes — “instantaneously” switch to backup path

Some Solutions

Multisession to fix fate-sharing

Convergence

- Withdraw routes more efficiently
- Advertise more backup routes

Several options to distinguish multiple routes

- New AFI/SAFI
- Distinct session per QoS
- Agree upon and exchange QoS markings

More to do

Inter-domain convergence an active topic!

Some Solutions

Multisession to fix fate-sharing

Convergence

- Withdraw routes more efficiently
- Advertise more backup routes

Several options to distinguish multiple routes

- New AFI/SAFI
- Distinct session per QoS
- **Agree upon and exchange QoS markings**

Aside — Service Separation Within the Network

Today: Path followed by packet is based on destination address

Today: Statically configured Policy Based Routing – path followed based on attributes such as DSCP etc

Problem Statement: How to dynamically use multiple paths to a given destination based on traffic types?

What are MTR, TE, VRs?: Service Separation

Adding another dimension to destination based routing –

- Class-specific next-hops, class specific VRFs, class specific tunnels…

End Goal:

- To influence the path that certain types of traffic would take (to reach a given destination) based on attributes such as DSCP, Application Type etc.
- Traffic Separation across network infrastructure

Conceptual View of Service Separation

Creation of multiple topologies

- Logical path that traffic will take across the given network
- VR, TE, MTR means that each topology will route/forward a subset of the traffic as defined by the classification criteria

Mapping of traffic to a topology—topology selection

- Determine which traffic (based on classification criteria) is subject to topology specific forwarding

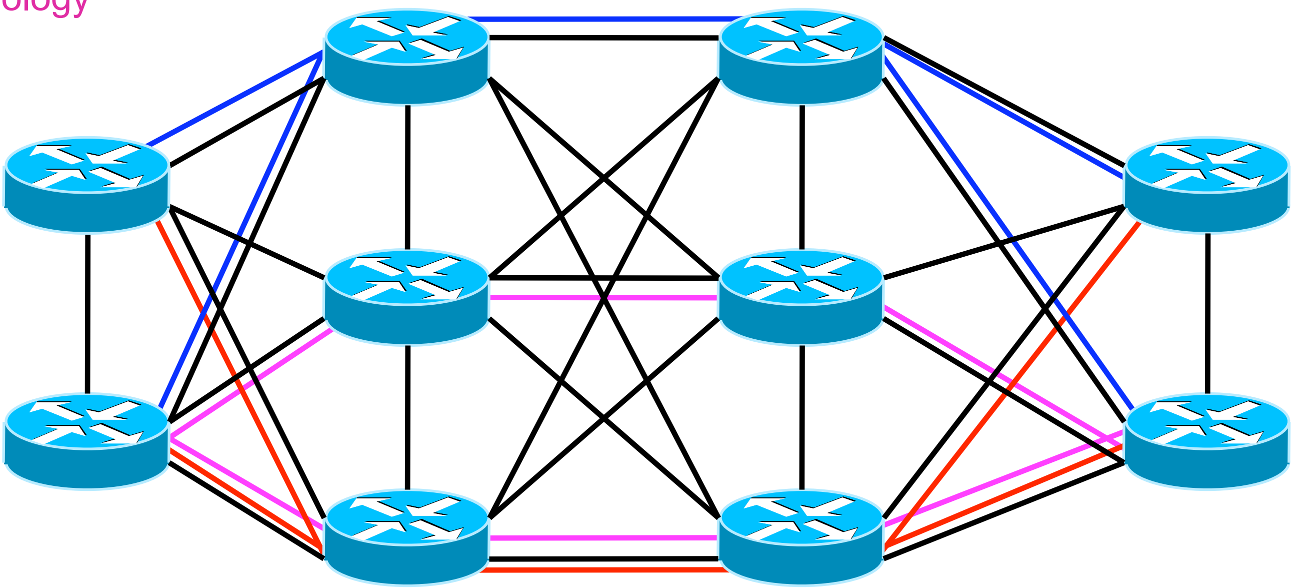
QoS provides per-hop service differentiation within a single path, VR, TE — but MTR provides path-based service differentiation

- Most often QoS policies are congruent with service topologies

Routing by Service — Defining Topologies

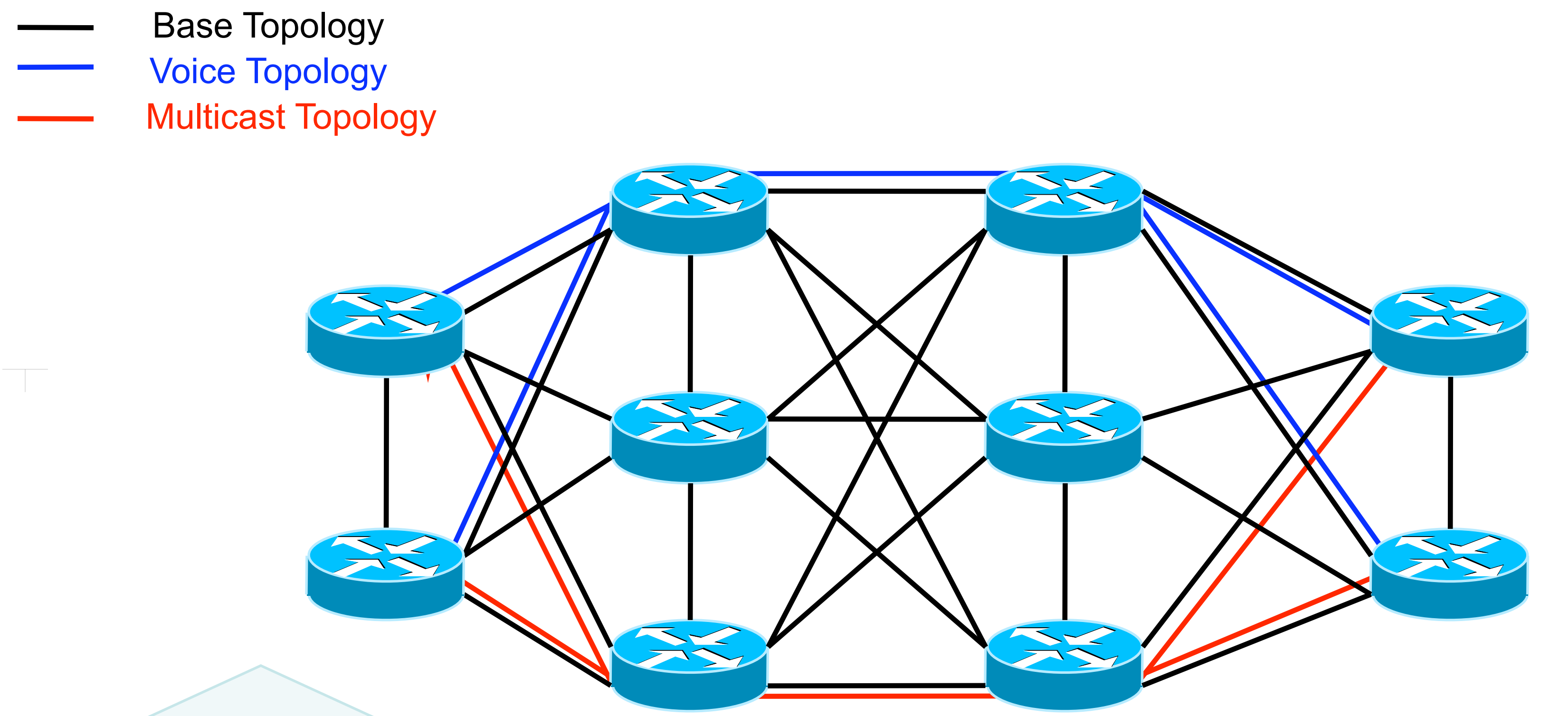
— Base Topology
— Voice Topology
— Multicast Topology
— Video Topology

Start with a Base Topology
Includes all routers and all links



- **Define the class-specific topology across a contiguous section of the network**
- Individual links can belong to multiple topologies

Routing by Service —
Traffic Paths



Traffic is marked
at the network edge.
DSCP value is used
to assign traffic to
a topology, pushed into a label,
Lookup in a specific VRF

As traffic traverses the network
it is constrained to its own
class-specific topology

Usage Scenarios

Delay vs. throughput

- Voice to follow paths that are delay sensitive, whereas data can follow paths that have good throughput, but propagation delay/jitter is not that important

Backup links

- Using under utilized (backup) links for batch traffic

Traffic separation

- Using network infrastructure for certain traffic types. For example – incongruent unicast and multicast topologies.

Quarantine Topology

- Forwards all “suspicious traffic” on a separate topology that has security devices and/or to dump it in a “bit bucket”

Basic Forwarding Model/Behavior

Forwarding path

- 1. Classifies packet into service type
- 2. Determines the corresponding class table or VRF
- 3. Looks up the destination address in that table
- 4. Forwarding entry is found for that destination
- 5. Forwards the packet to the next hop or label push

If no forwarding entry within a topology, packet is dropped

- If packet does not match any classifier, it is forwarded on the base topology

Relevance to Interdomain QoS

May want to signal inter-domain services

- May want specific peering or entry/exit points to services

Services topologies most often have congruent QoS semantics

- May want to have orthogonal QoS and service topologies
- May want to have QoS within service topologies

Need to signal internally and externally with BGP

Context AF for BGP

Advertise flexible descriptions of tables (RIBs/FIBs), allow updates targeted to these tables

Context description and ID advertised in Capability

- Extensible description format, currently AFI/SAFI, QoS, Topology

No changes to actual update format

- Existing features which rely on AFI/SAFI pair to describe the target table are backward compatible

Context AF and features

Enables BGP for

- Topologies
- QoS
- Both (QoS policy within a service topology)

Context ID is Opaque

- Does not define local QoS config
- Instead, defines a service

Wrap-Up

What’s left?

Need to signal anything beyond reachability (and AS hop count)?

- **BGP not particularly good for very dynamic data**
  - **BGP not to propagate link attribute info**
- **History teaches that global BGP route selection metrics are difficult to agree on**
- **On the other hand, BGP is pretty good at carrying around bags of data the protocol doesn’t care about**

Summary: What does this architecture provide?

Exchanges QoS and Topology information

- Enabling service differentiation

Follows current BGP configuration, policies and management

- Uses backwards compatible technique - Easy deployment

Allows for fast convergence per service

- Announcing multiple paths per prefix/service
- Withdraw all prefixes in a AF/SAF/topo/QoS in one message

Doesn’t interfere with deployed features or availability mechanisms

Allows for any service separation design: VR, TE, MTR

Cisco Systems ®