Practical Strategies for IP Traffic Engineering and Enhancing Core network Availability

RIPE 48 EOF Tutorial

Monday, May 3rd 2004

Amsterdam

John Evans
Cisco Systems, Inc.
joevans @ cisco.com

Alan Gous
Cariden Technologies, Inc.
alan @ cariden.com

IP Traffic Engineering and Core Network Availability

- We can consider MPLS Traffic Engineering in terms of its potential applications

    1. Bandwidth Optimisation

        Making efficient use of bandwidth, a.k.a. offload routing, a.k.a. traffic engineering

    2. Improving service availability

        Faster recovery around failures, i.e. using FRR

    3. Admission control
    4. Route pinning

- #1 and #2 have become synonymous with MPLS TE and will be covered in this session
- #3 and #4 are not covered in this session

IP Traffic Engineering and Core Network Availability

- **In this session we will …**

  Consider the theory behind traffic engineering in general

  Analyse some of the benefits, limitations, and deployment considerations of MPLS TE in the context of IP traffic engineering and engineering core network availability

  Give consideration to possible alternate approaches for IP traffic engineering and engineering core network availability

Agenda

I. IP TE Introduction

- The problem
- The objective
- Limitations
- Approaches

II. Traffic Characterization

III. Traffic Matrices

IV. Approaches for TE

V. High availability options

Defining Traffic Engineering

- **Network engineering**

  Manipulating your network to suit your traffic

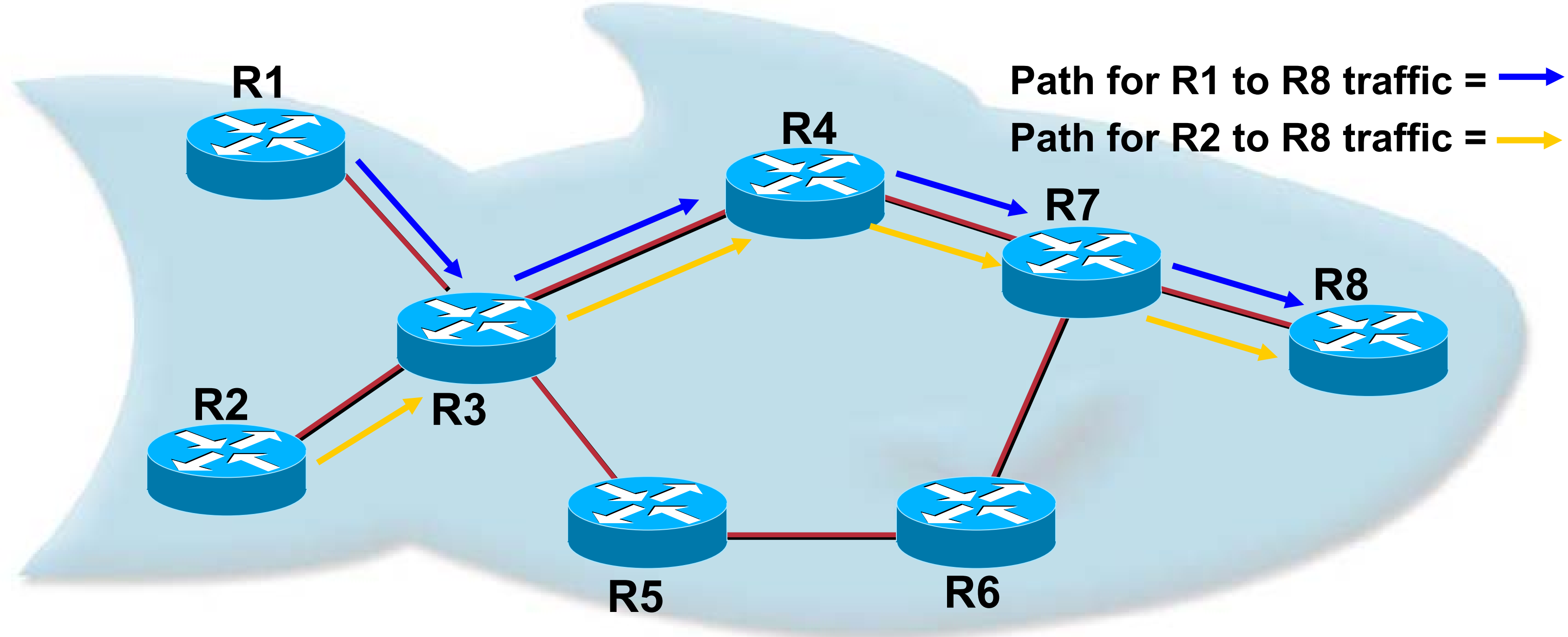
- **Traffic engineering**

  Manipulating your traffic to suit your network

- **Clearly network engineering and traffic engineering are linked**

- **We will consider only traffic engineering**

IP Traffic Engineering: The Problem



- Conventional IP routing uses pure destination-based forwarding
- Conventional IGP path computation is selected based upon a simple additive metric
    - Bandwidth availability is not taken into account
- Some links may be congested while others are underutilized

Traffic Engineering: The Problem

- **The traffic engineering problem can be defined as an optimization problem**

  Definition – "optimization problem": **A computational problem in which the objective is to find the best of all possible solutions**

  ➔ **Given a fixed topology and a fixed source-destination matrix of traffic to be carried, what routing of flows makes *most effective* use of aggregate or per class (Diffserv) bandwidth?**

  ➔ **How do we define *most effective* … ?**

IP Traffic Engineering: The objective

- **What is the primary optimization objective?**

  Either …

  minimizing maximum utilization in normal working (non-failure) case

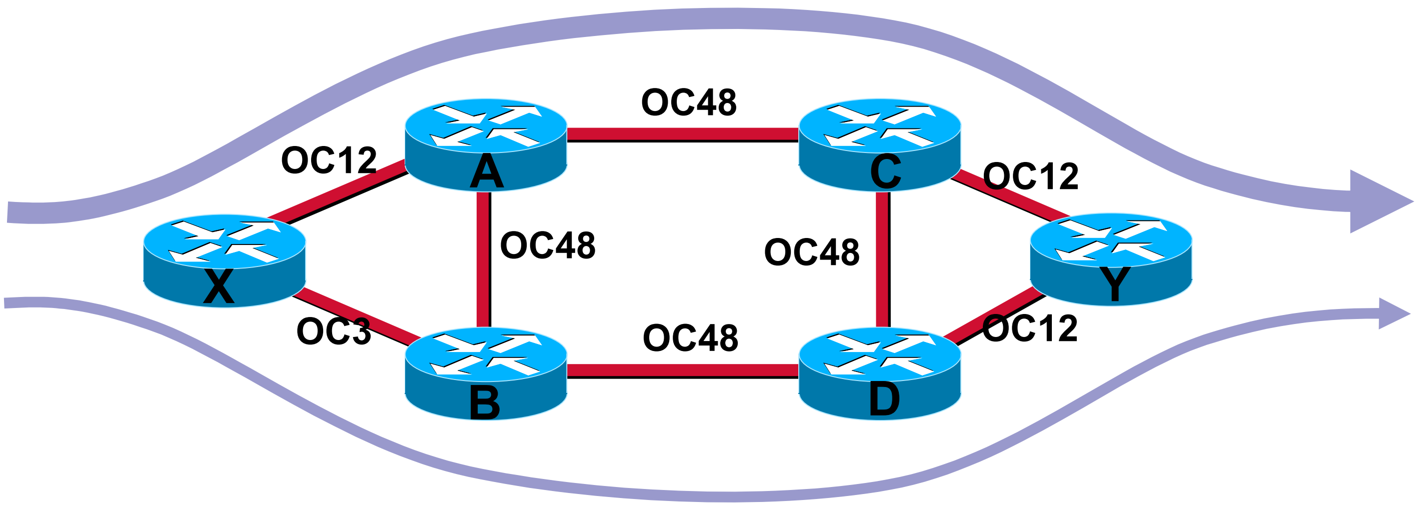
  Or …

  minimizing maximum utilization under single element failure conditions

- **Understanding the objective is important in understanding where different traffic engineering options can help and in which cases more bandwidth is required**

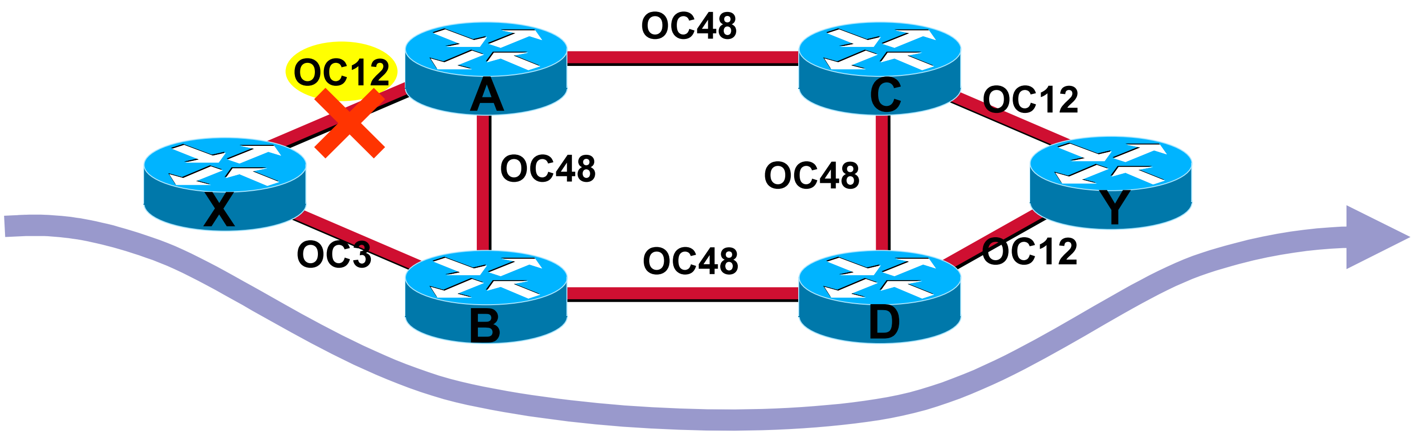
  Other optimization objectives possible: e.g. minimize propagation delay, apply routing policy …

Working Case Optimisation



- In this asymmetrical topology, if the demands from X➔Y > OC3, traffic engineering can help to distribute the load when all links are working

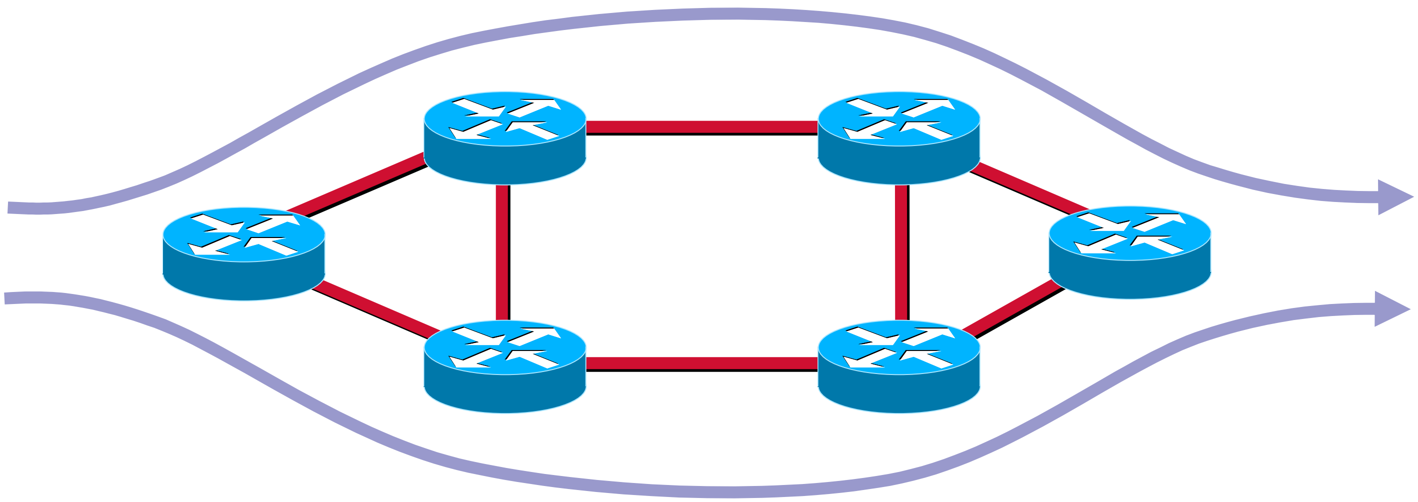
Failure Case Optimisation



- **However, in this topology when optimization goal is to minimize bandwidth for single element failure conditions, if the demands from X➔Y > OC3, TE cannot help ➔ must upgrade link X-B**

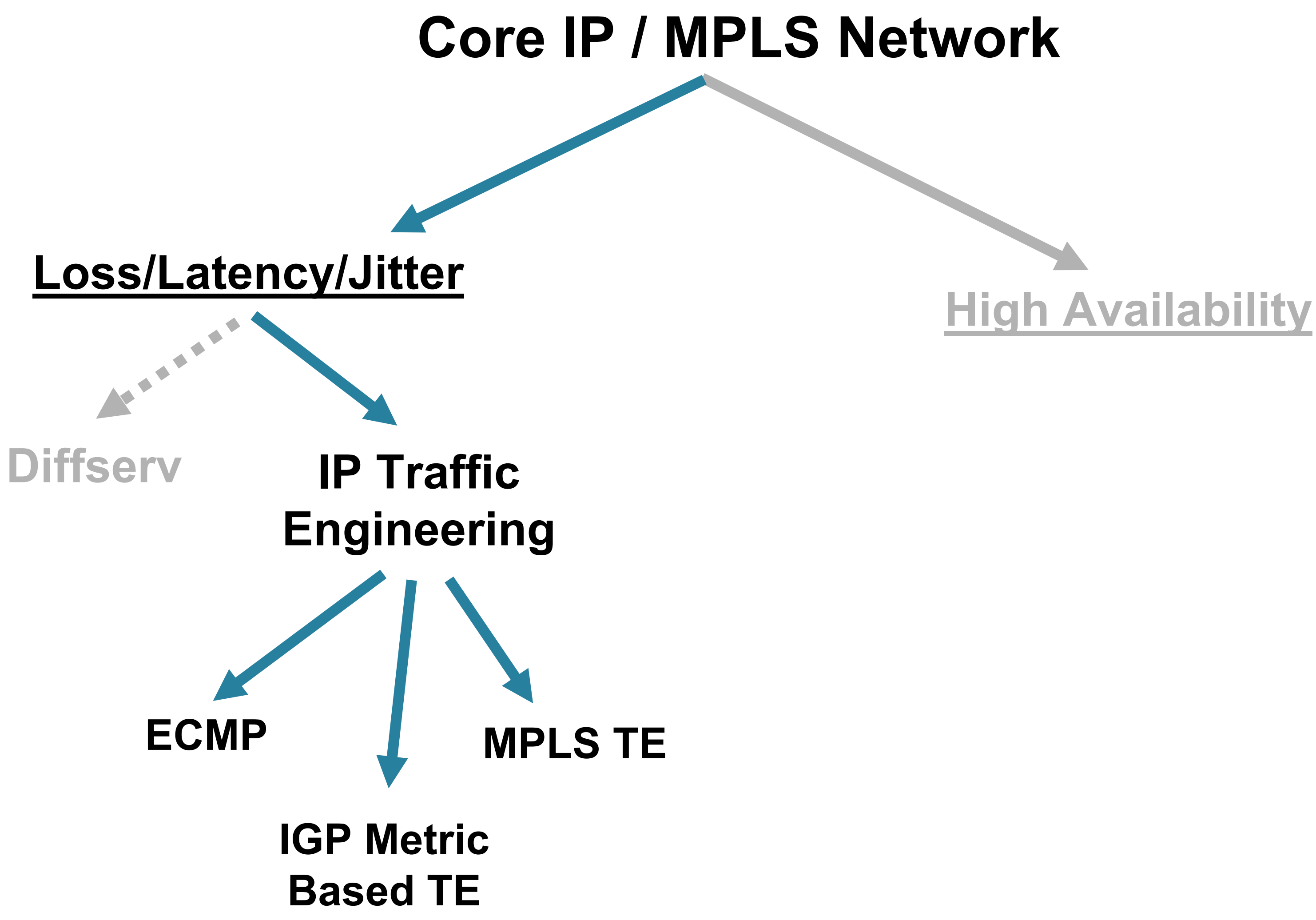
  this is a simply a problem of capacity provisioning, not a problem of traffic engineering

Traffic Engineering Limitations



- **TE cannot create capacity**
  - **e.g. "V-O-V" topologies allow no scope strategic TE if optimizing for failure case**
    - **Only two directions in each "V" or "O" region – no routing choice for minimizing failure utilization**
- **Other topologies may allow scope for TE in failure case**
  - **As case study later demonstrates**

Options for IP Traffic engineering



Tactical versus Strategic

- **Tactical TE**
  - Ad hoc approach aimed at fixing current problems
  - Short term operational/engineering process
  - Configured in response to failures, traffic changes
- **Strategic TE**
  - Systematic approach aimed at cost savings
  - Medium term engineering/planning process
  - Configure in anticipation of failures, traffic changes
    - Resilient metrics, or
    - Primary and secondary disjoint paths, or
    - Dynamic tunnels, or …

Traffic Engineering: The Benefit

- The more effective use of backbone bandwidth potentially allows:

    Either …

        higher SLA targets (lower loss, lower delay) to be offered with the existing backbone bandwidth

    Or …

        the existing SLA targets to be achieved with less backbone bandwidth or with delayed time to bandwidth upgrades

- Either way, the benefit is one of cost saving to the provider
- To quantify this, it is important to understand the relationship between bandwidth and QOS

Agenda

Cisco.com Cariden.com

I. IP TE Introduction

II. Traffic Characterization

- Long term (minutes +)
- Short term (milliseconds)

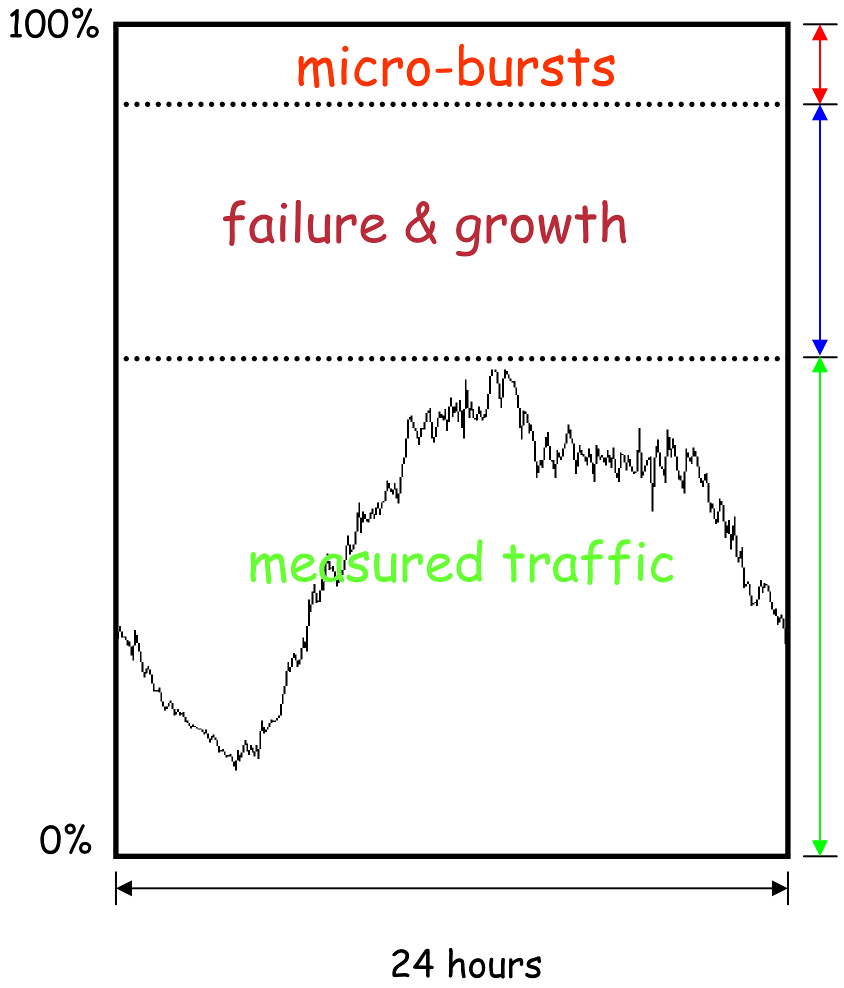
III. Traffic Matrices

IV. Approaches for TE

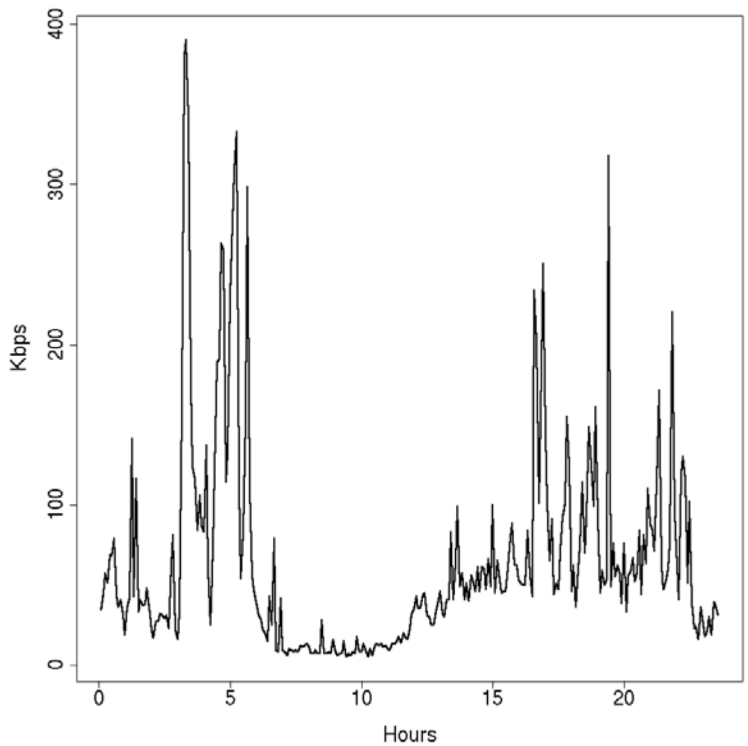
V. High availability options

Traffic Characterisation

- **Long-Term**
  - **Measured Traffic**
    - **E.g. P95 (day/week)**
  - **Accommodate failure and growth**
- **Short-Term**
  - **Critical scale for queuing**
  - **Determine over-provisioning factor that will prevent queue buildup against micro-bursts**



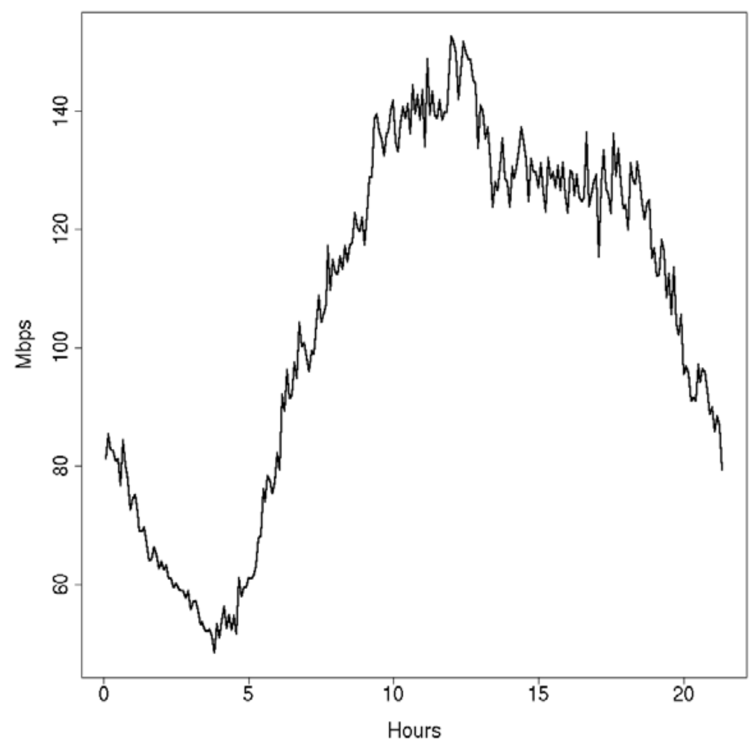
High- vs. Low-Bandwidth Demands



Cleveland -> Denver

Mean=64Kbps, Max=380Kbps

P95=201Kbps, Std. dev.=66Kbps



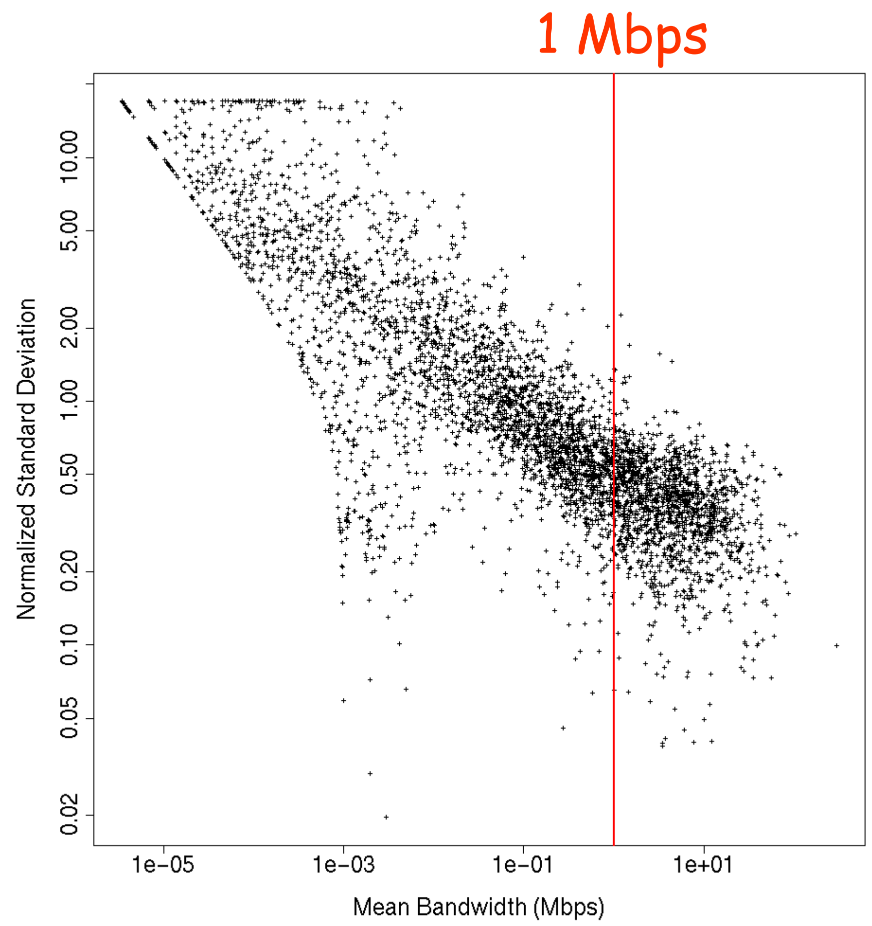
Washington D.C. -> Copenhagen

Mean=106Mbps, Max=152Mbps

P95=144Mbps, Std. dev=30Mbps

Variance vs. Bandwidth

- **Around 8000 demands between core routers**
- **Relative variance decreases with increasing bandwidth**
- **High-bandwidth demands seem well-behaved**
- **97% of traffic is carried by the demands larger than 1 Mbps (20% of the demands!)**



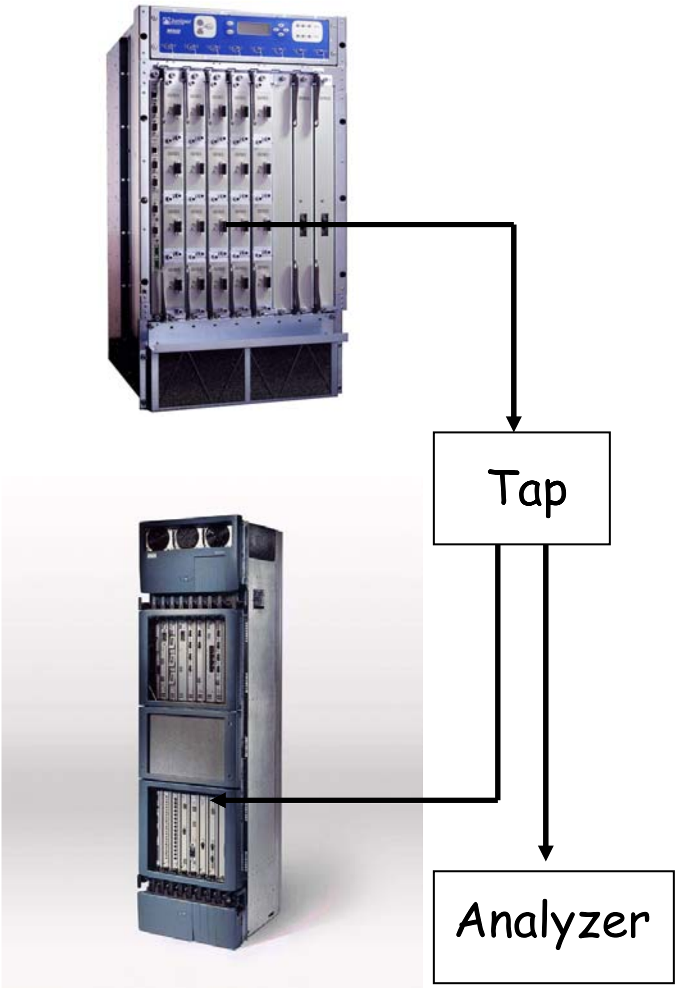
Long Term Traffic Summary

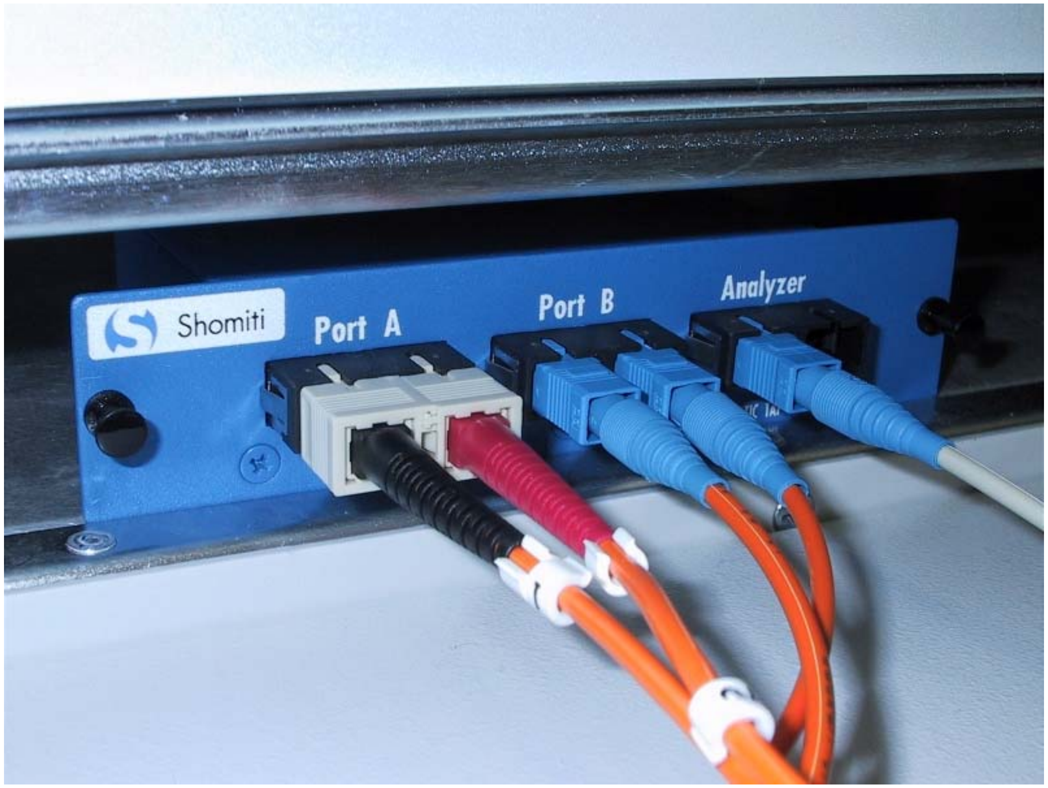
- **Most traffic carried by (relatively) few big demands**
- **Big aggregated demands are well-behaved (predictable) during the course of a day and across days**
- **Little motivation for dynamically changing routing during the course of a day**

Short-term Traffic Characterization

- Investigate burstiness within 5-min intervals
- Critical timescale for queuing, like 1ms or 5ms
- Analyze statistical properties
- Only at specific locations
    - Complex setup
    - A lot of data

Fiber Tap (Gigabit Ethernet)



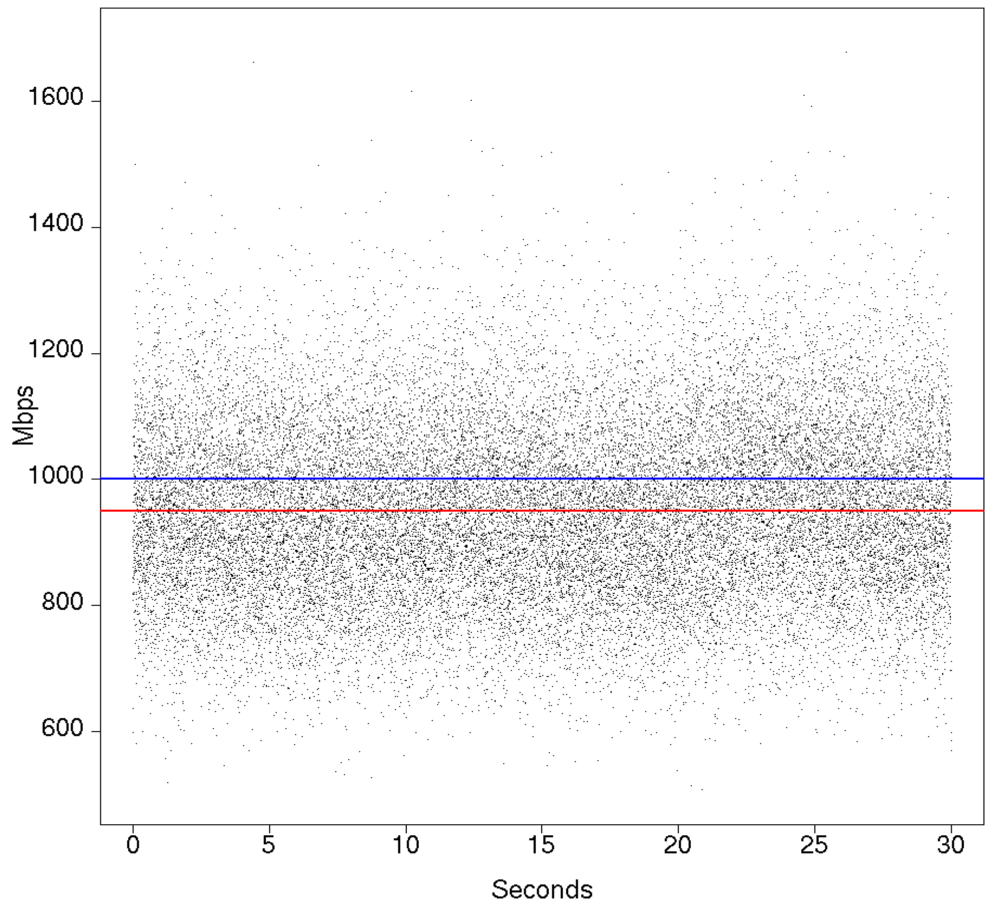


Raw Results
30 sec of data, 1ms scale

- Mean = 950 Mbps
- Max. = 2033 Mbps
- Min. = 509 Mbps

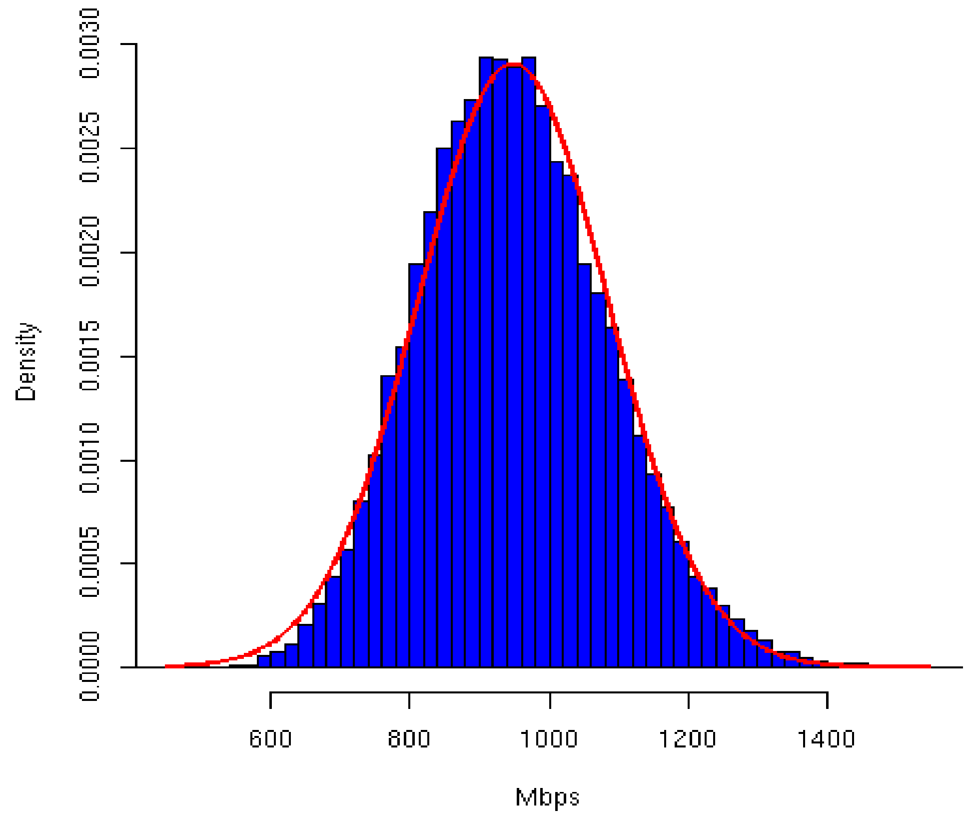
- 95-percentile: 1183 Mbps
- 5-percentile: 737 Mbps

- (around 250 packets per 1ms interval)



Traffic Distribution Histogram (1ms scale)

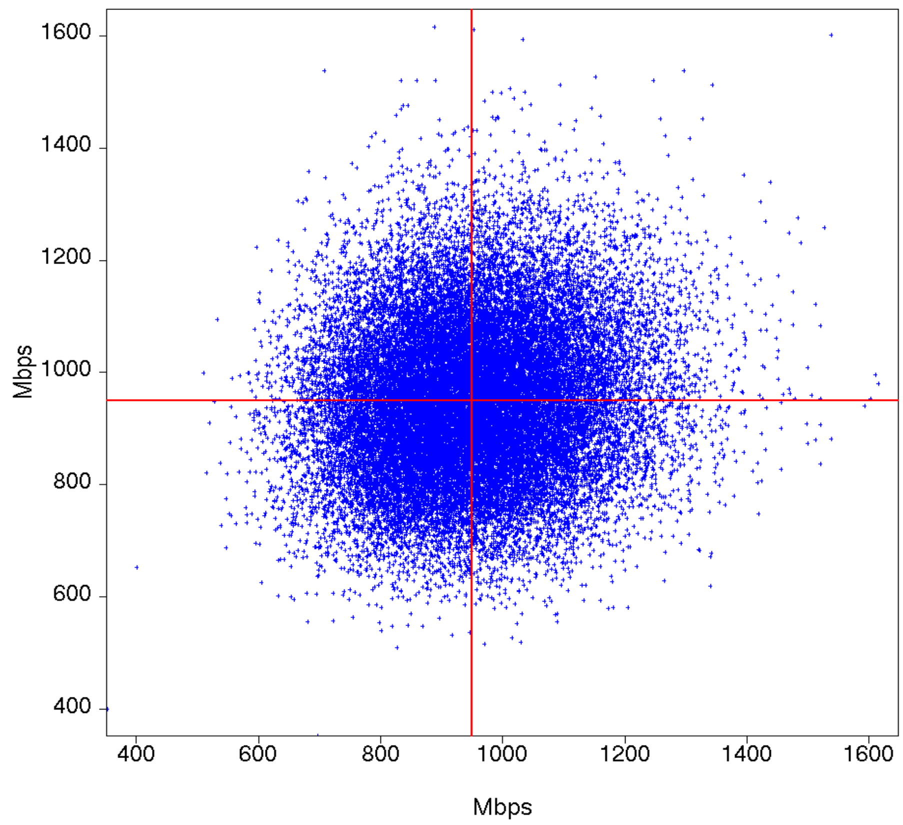
- **Fits normal probability distribution very well (Std. dev. = 138 Mbps)**
- **No Heavy-Tails**
- **Suggests small overprovisioning factor**



Autocorrelation, Lag Plot (1ms scale)

Cisco.com Cariden.com

- **Scatterplot for consecutive samples**
- **Are periods of high usage followed by other periods of high usage?**
- **Autocorrelation at 1ms is 0.13 (=uncorrelated)**



Traffic: Summary

- **Long Term Traffic Patterns**
  - Smooth for big (relevant) flows
  - Predictable Trends
  - Less motivation for dynamic routing
- **Millisecond Time Scale**
  - Uncorrelated
  - Not Self-Similar Long-term well-behaved traffic
  - Less headroom required for QoS as circuit capacity increases

Theoretical Models

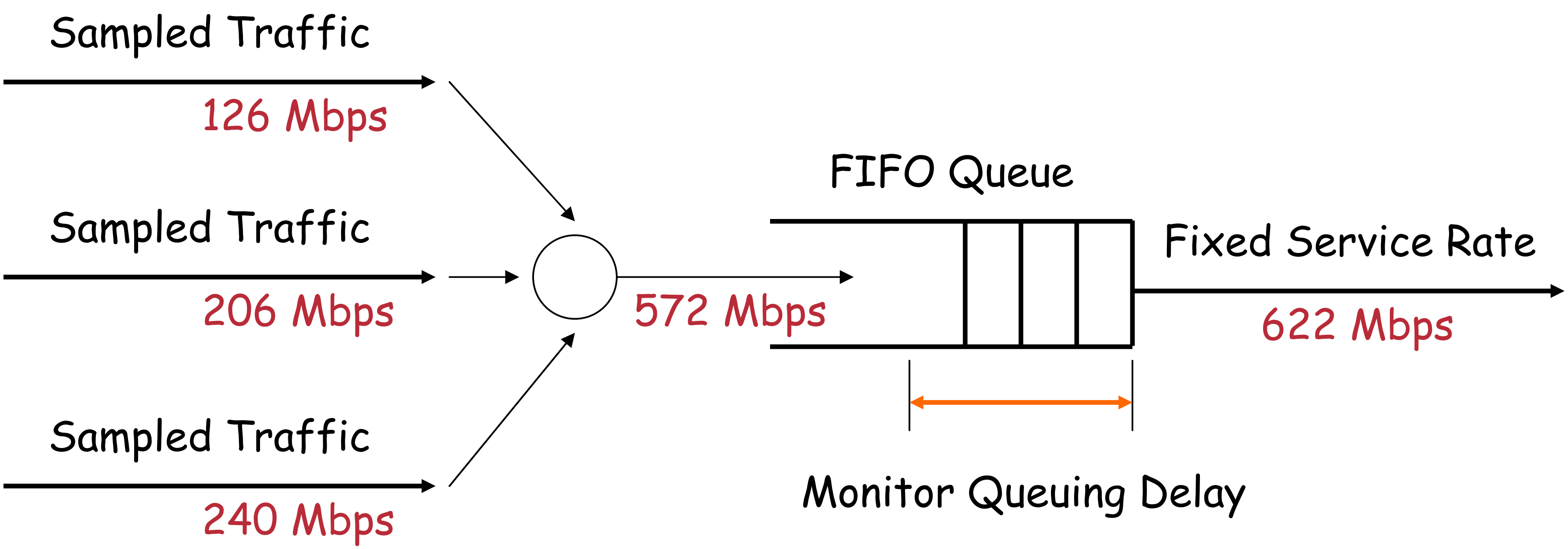
- **M/M/1**
- Markovian
  - Poisson-process
  - Infinite number of sources
- “Circuits can be operated at over 99% utilization, with delay and jitter well below 1ms” [2] [3]

- **Self-Similar**
- Traffic is bursty at many or all timescales
- “Scale-invariant burstiness (i.e. self-similarity) introduces new complexities into optimization of network performance and makes the task of providing QoS together with achieving high utilization difficult” [4]
- (Various reports: 20%, 35%, …)

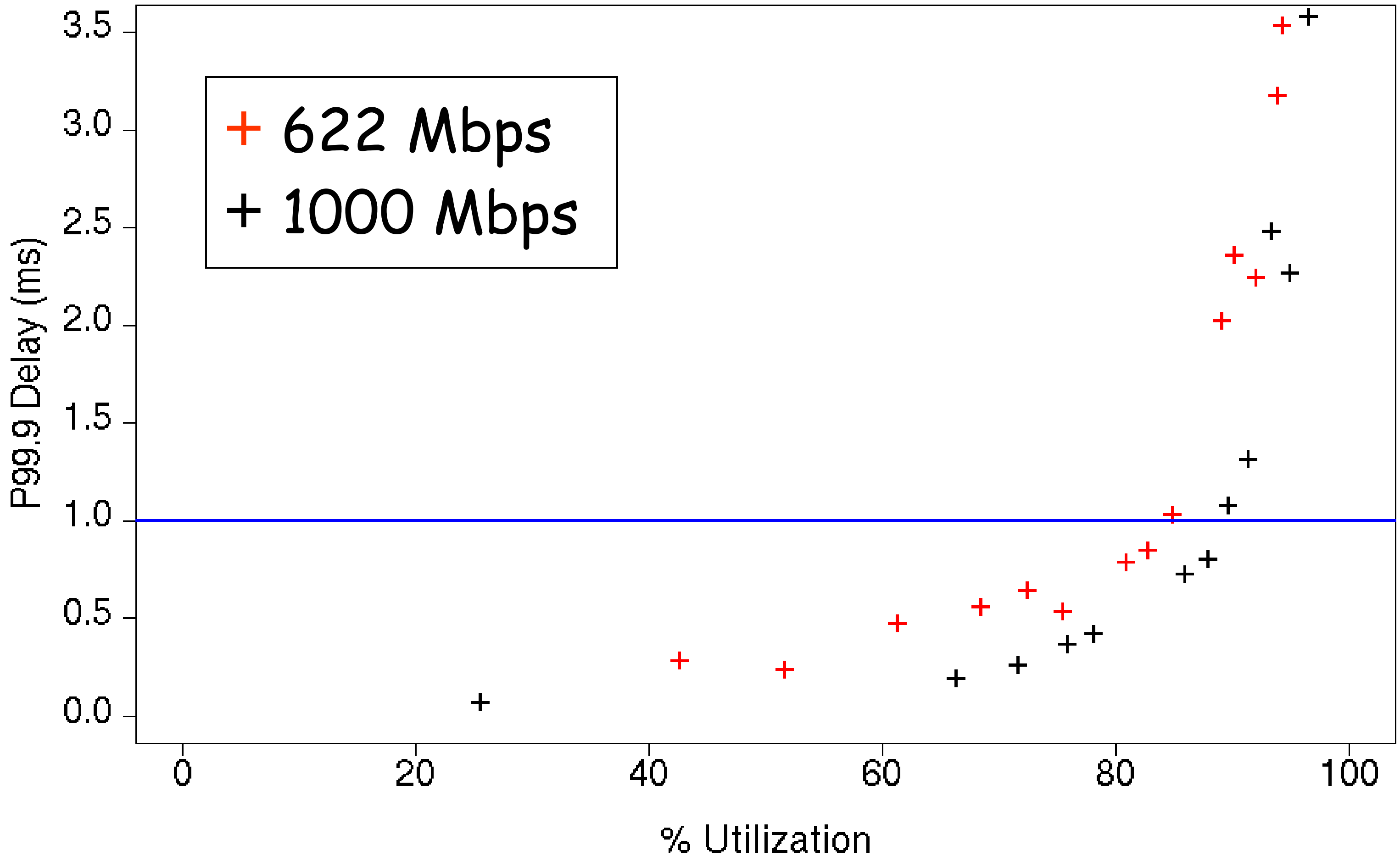
Empirical Simulation

- Feed multiplexed sampled traffic data into FIFO queue
- Measure amount of traffic that violates the delay bound

Example: 92% Utilization



Queuing Simulation: Results



Queuing Simulation Results

- **1 Gbps (Gigabit Ethernet)**

  1-2 ms delay bound for 999 out of 1000 packets (99.9-percentile):

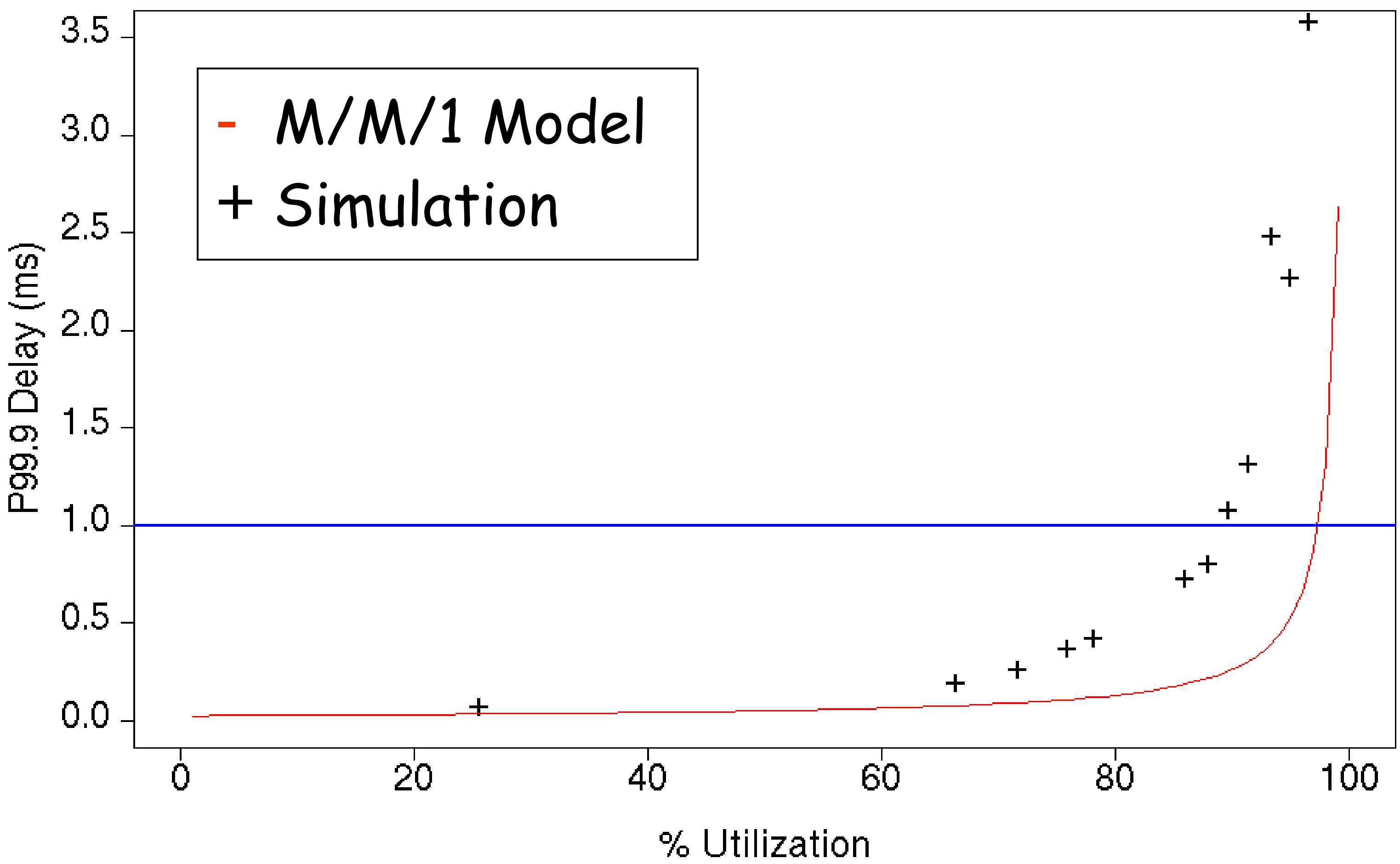
  90%-95% maximum utilization

- **622 Mbps (STM-4c/OC-12c)**

  1-2 ms delay bound for 999 out of 1000 packets (99.9-percentile):

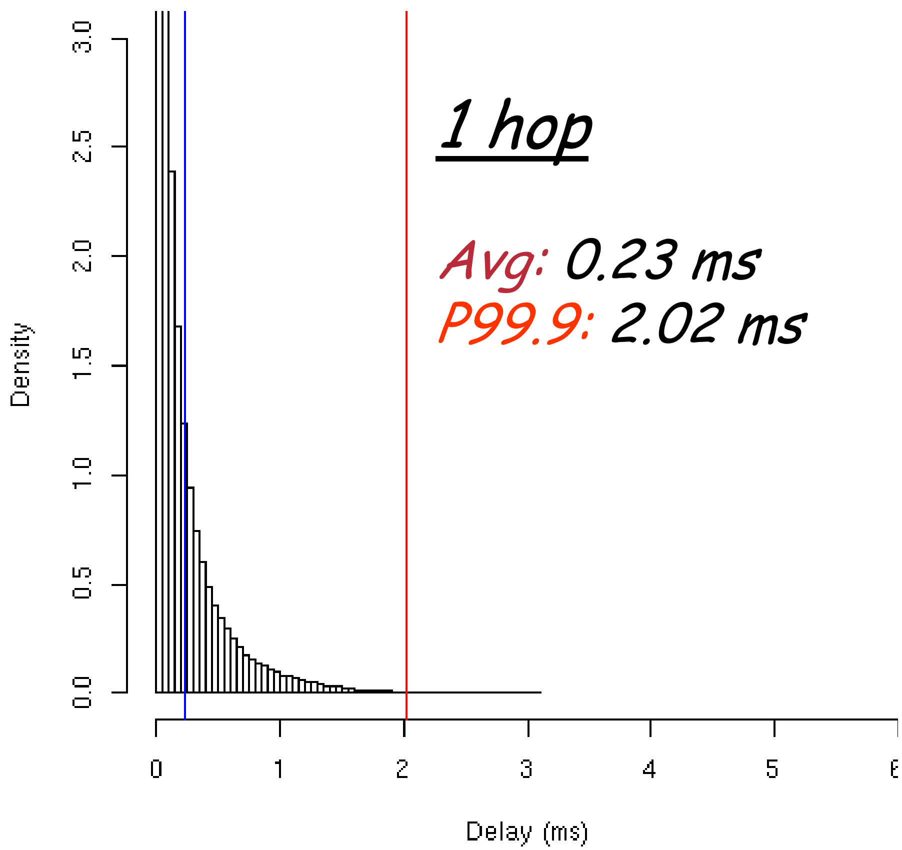
  85%-90% maximum utilization

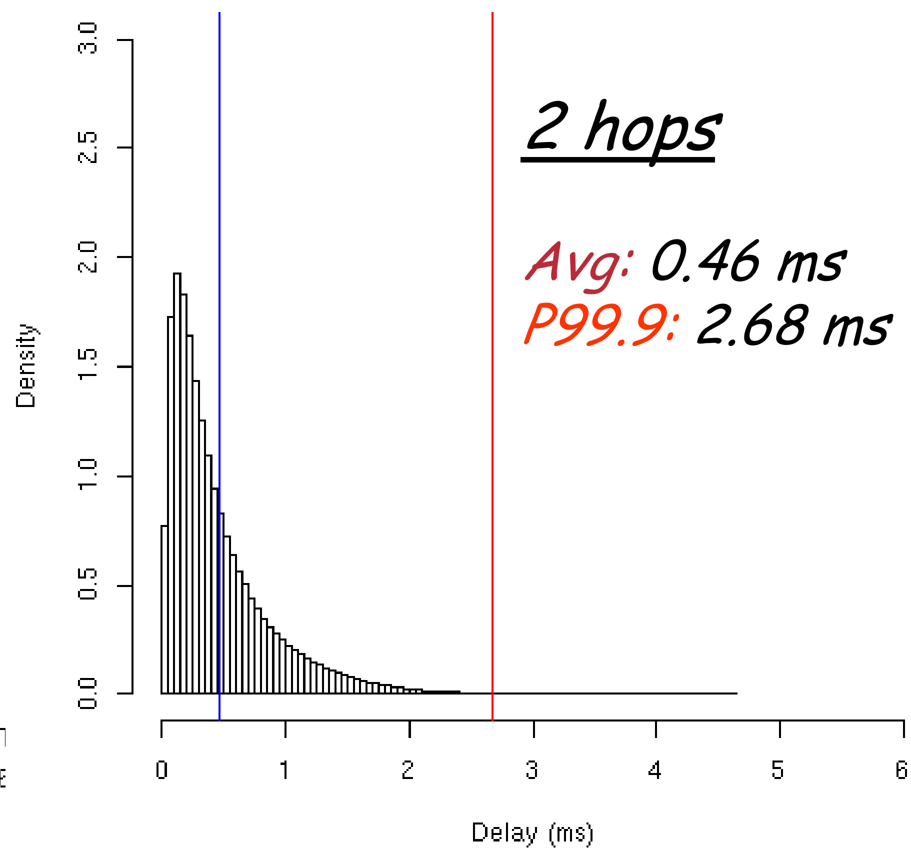
Theory vs. Simulation (1Gbps)



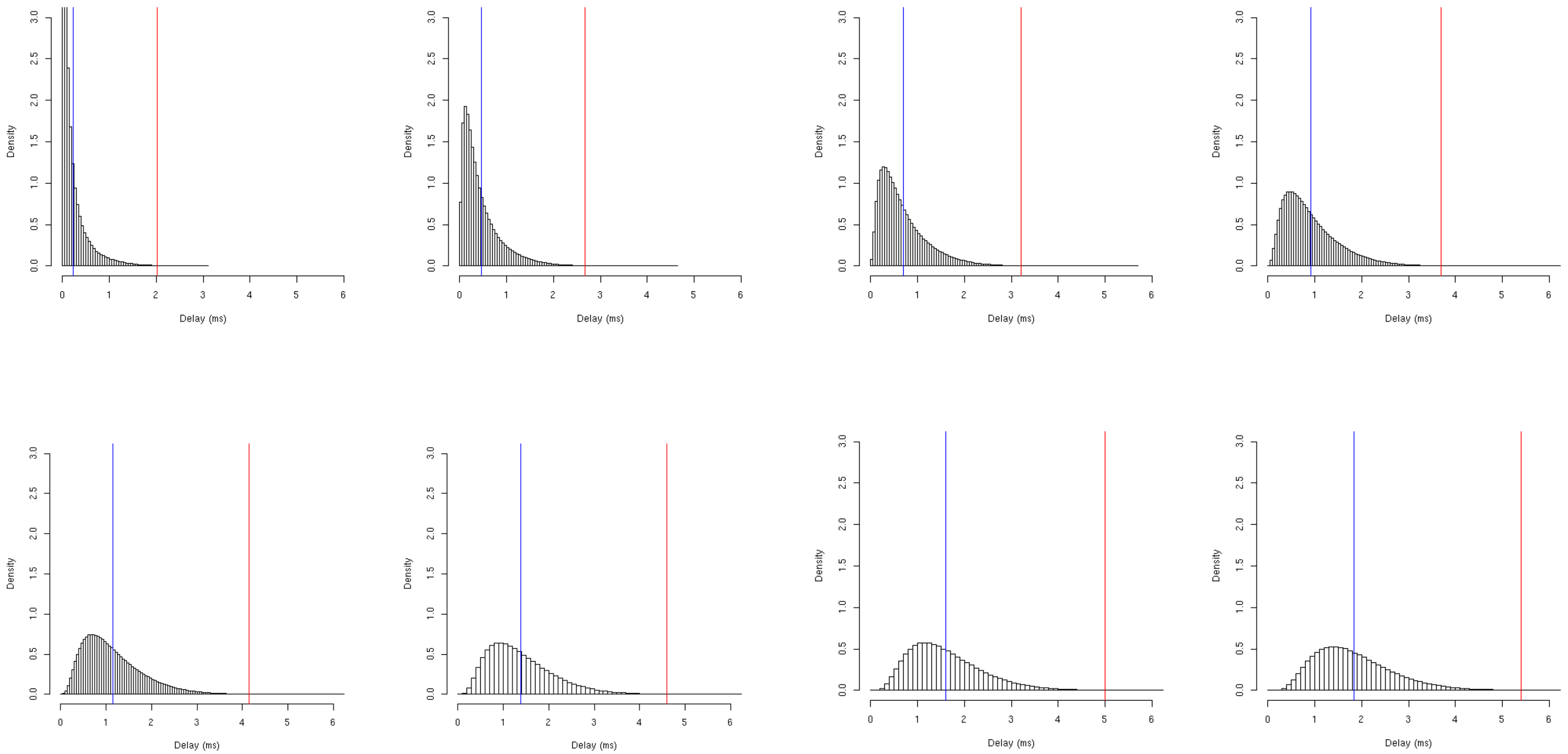
Multi-hop Queuing

Cisco.com Cariden.com





Multi-hop Queuing (1-8 hops)



Queuing: Summary

- **Queuing Simulation:**

  622Mbps, 1Gbps (backbone) links

  overprovisioning percentage in the order of 10% is required to bound delay/jitter to less than 1-2 ms

  Lower speeds (≤155Mpbs)

  overprovisioning factor is significant,

  Higher speeds (2.5G/10G)

  overprovisioning factor becomes very small

  P99.9 multi-hop delay/jitter is not additive

Agenda

I. IP TE Introduction

II. Traffic Characterization

III. Traffic Matrices

- Measurement methods
- Estimation methods

IV. Approaches for TE

V. High availability options

IP Traffic Engineering: The inputs

- Need a minimum set of information in order to determine what (if any) benefit each of the options can give
    - Core traffic demand matrix
        - The matrix of ingress to egress traffic demands
        - Enables trending and “what-if” scenarios
    - Core topology (logical and physical)
        - Mapping traffic matrix to the topology allows quantitative comparison

Core traffic matrix

- **Number of options available for deriving the core traffic demand matrix**
  - Measurement methods
    - Full mesh of TE tunnels and Interface MIB
    - NetFlow – BGP Next Hop TOS Aggregation
    - NetFlow – MPLS aware netflow
    - MPLS LSR MIB
    - BGP Policy Accounting
  - Demand estimation

Core traffic matrix

- **Full mesh of TE tunnels and Interface MIB**
  - Tunnel interface stats provide bandwidth usage between all entry and exit points on core
  - Data collected via SNMP from headend Router
  - Requires full mesh of TE tunnels
  - No support for per-CoS routing into tunnels yet
- **NetFlow**
  - NetFlow BGP Next Hop TOS Aggregation
    - v9 includes accounting based upon BGP next hop

Core traffic matrix

- NetFlow (contd.)
    - MPLS aware netflow
        - Provides flow statistics for MPLS labelled packets
        - FEC implicitly maps to BGP next hop / egress PE
- MPLS LSR MIB
    - MPLS-LSR-MIB mirrors the Label Forwarding Information Base (LFIB)
    - FEC implicitly maps to BGP next hop / egress PE

Core traffic matrix

- **BGP Policy Accounting**

  Allows accounting for IP traffic differentially by assigning counters based on:

  BGP community-list (included extended)

  AS number

  AS-path

  destination IP address

- **For more details on core traffic matrix options see:**

  Benoit Claise, Traffic Matrix: State of the Art of Cisco Platforms, Intimate 2003 Workshop in Paris, June 2003, http://www.employees.org/~bclaise/

Demand Estimation

- **Problem:**

  Estimate point-to-point demands from measured link loads

- **Network Tomography**

  Y. Vardi, 1996

  Similar to: Seismology, MRI scan, etc.

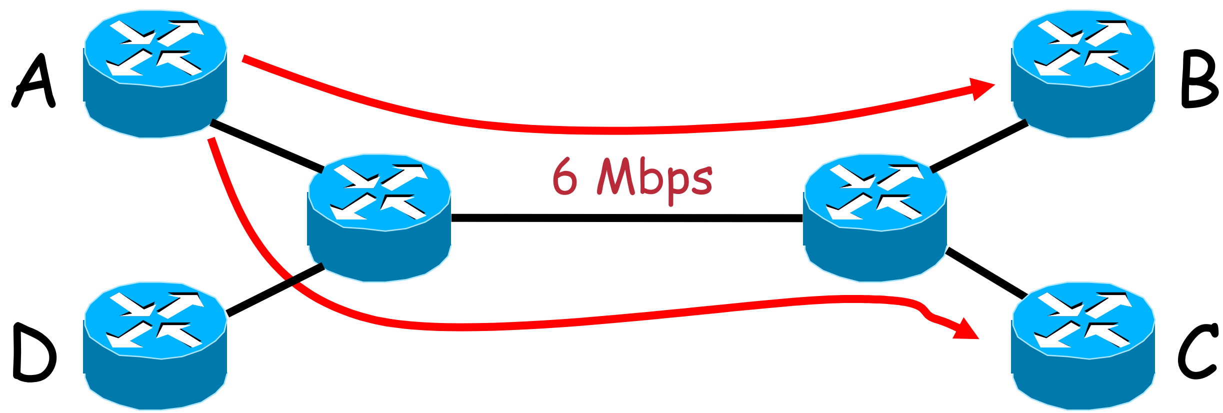
- **Underdetermined system:**

  N nodes in the network

  O(N) links utilizations (known)

  O(N2) demands (unknown)

Demand Estimation: Example



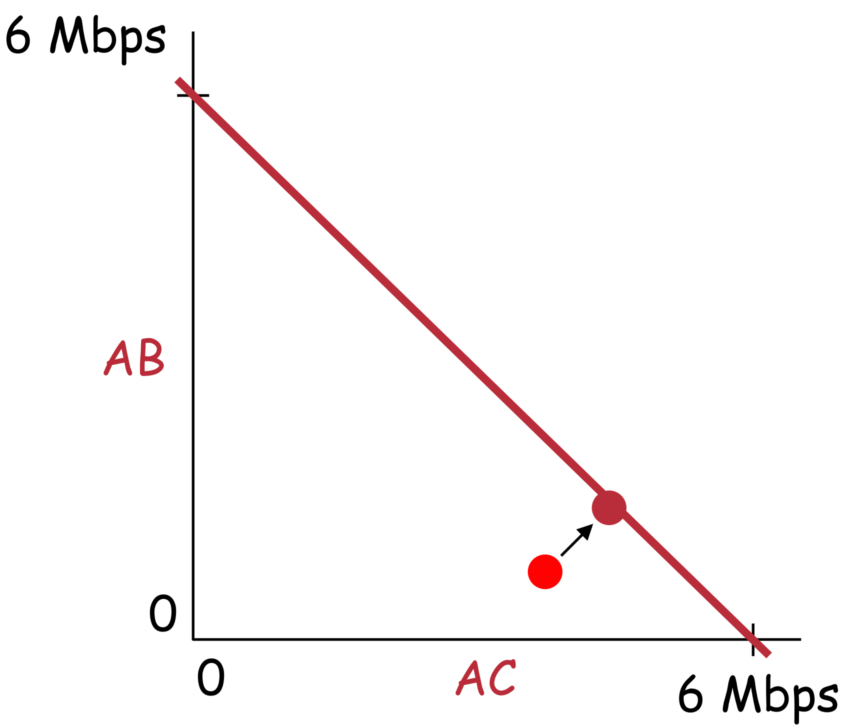
y: link utilizations
A: routing matrix
x: point-to-point demands

Solve: $\underline{y = Ax}$ -> *In this example:* $\underline{6 = AB + AC}$

Demand Estimation: Example

Cisco.com Cariden.com

Solve: y = Ax -> *In this example: 6 = AB + AC*



Additional information

E.g. Gravity Model (every source sends the same percentage as all other sources of it's total traffic to a certain destination)

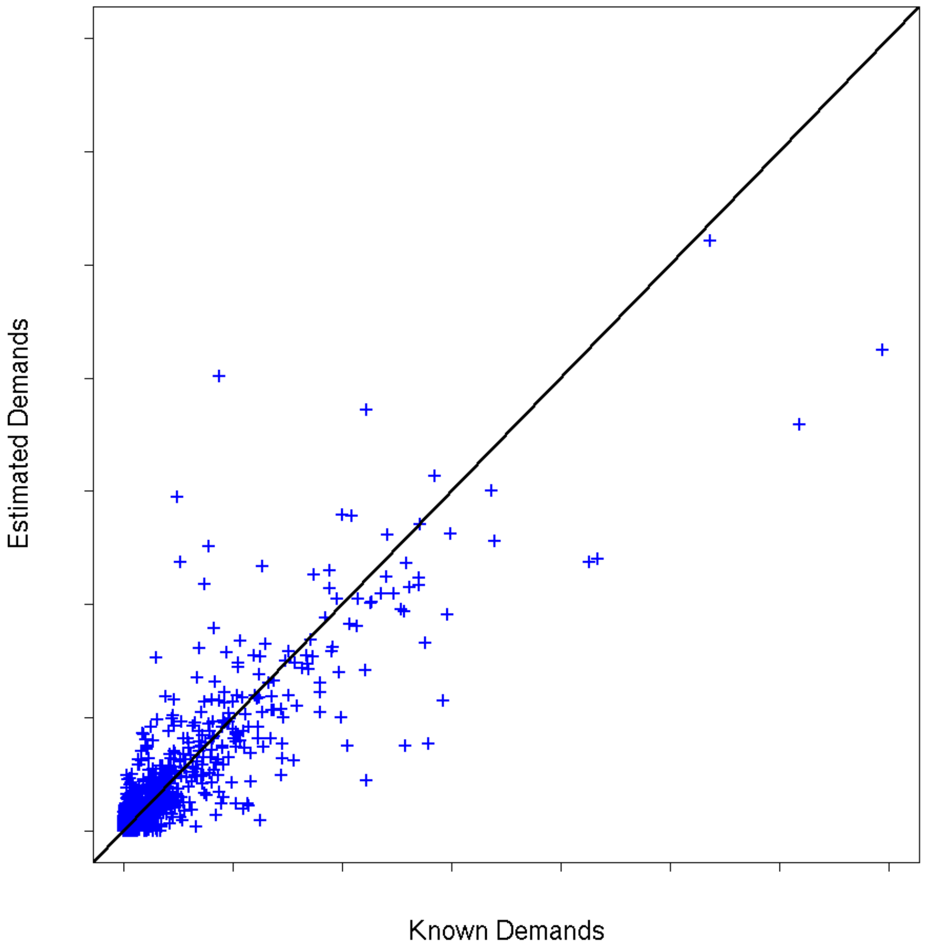
Example: Total traffic sourced at Site A is *50Mbps.*
Site B sinks *2%* of total network traffic, C sinks *8%.*

AB = 1 Mbps and AC = 4 Mbps

Final Estimate: AB = 1.5 Mbps and AC = 4.5 Mbps

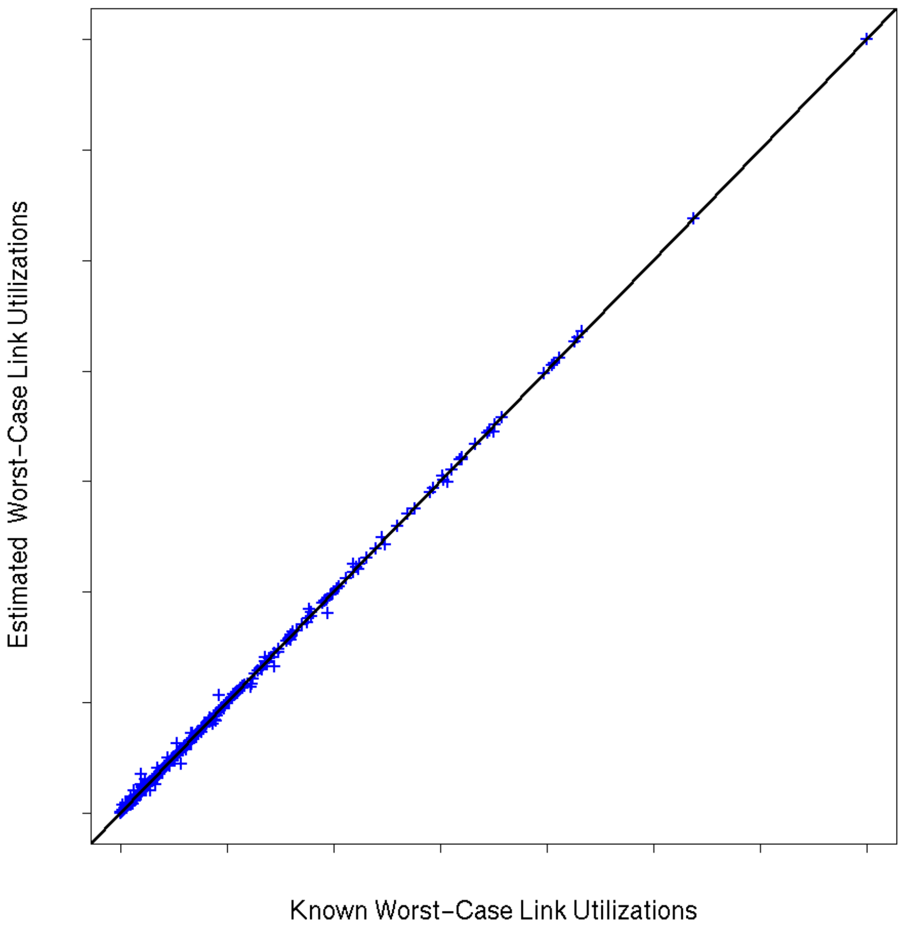
Real Network: Estimated Demands

- **Cariden Demand Deduction Tool**
- **GBLX Network**

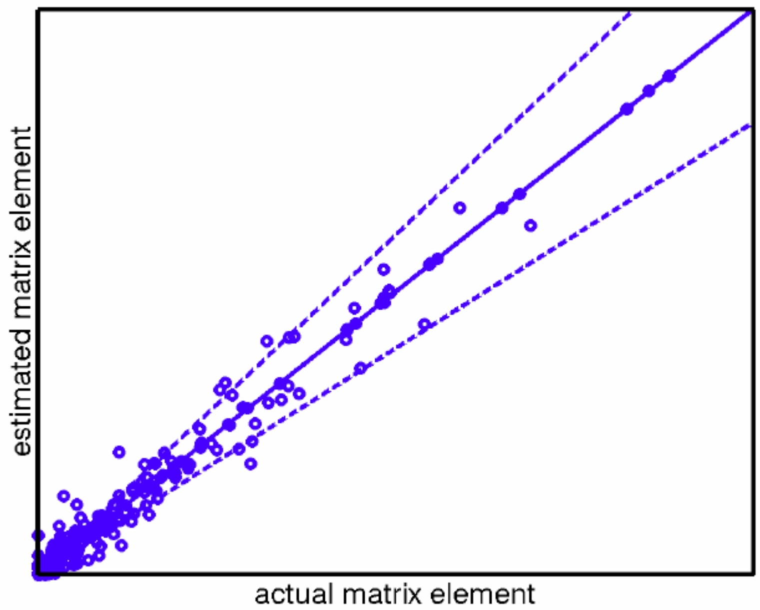


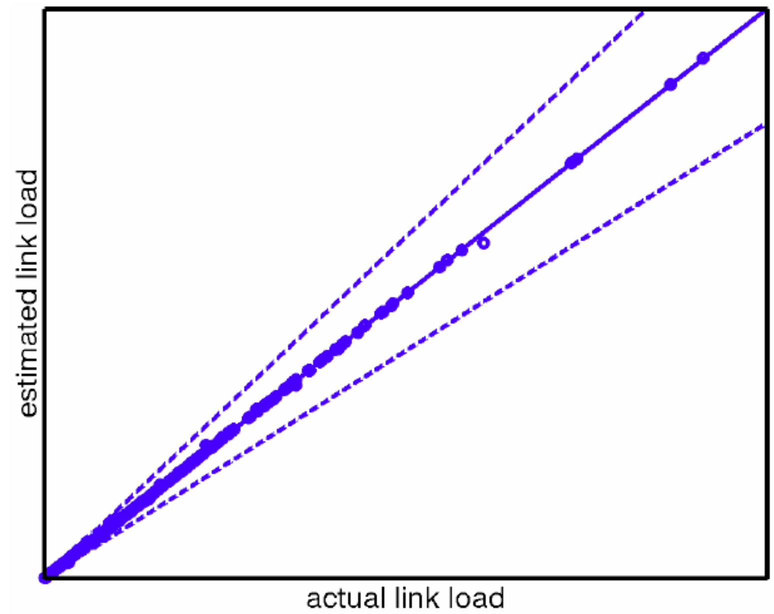
Predicted Link Utilizations!

- **Cariden Demand Deduction Tool**
- **GBLX Network**



Demand Estimation: AT&T Labs Procedure





- **NANOG 29: "How to Compute Accurate Traffic Matrices for Your Network in Seconds"**

    Implemented on AT&T IP backbone (AS 7018)

    Hourly traffic matrices for > 1 year (in secs)

    Used in reliability analysis, capacity planning, TE

Demand Estimation: Results

- **Individual demands:**

    Can be inaccurate.

- **Estimated worst-case link utilizations:**

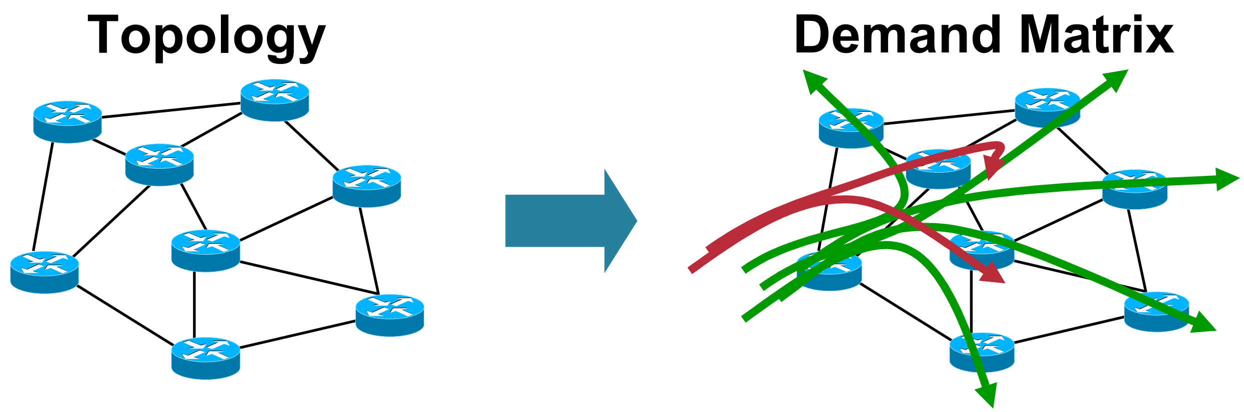
    Accurate!

- **Explanation:**

    Multiple demands on the same path indistinguishable, but their sum is known

    If these demands fail-over to the same alternative path, the resulting link utilizations will be correct

Core Capacity Planning



- **Map core traffic matrix to topology**
- **Simulate for link, node and SRLG failures**
  - **Can add a traffic growth factor if required**
- **On a per class basis if Diffserv deployed**
- **Enables:**
  - **Comparison of different TE approaches**
    - **Optimal distribution is defined as multi commodity flow**
  - **Forecasting of which links need upgrading when**
  - **Understand of if topology should be changed**

Agenda

I. IP TE Introduction

II. Traffic Characterization

III. Traffic Matrices

IV. Approaches for IP TE

- ECMP
- IGP Metric Based TE
- MPLS TE

V. High availability options

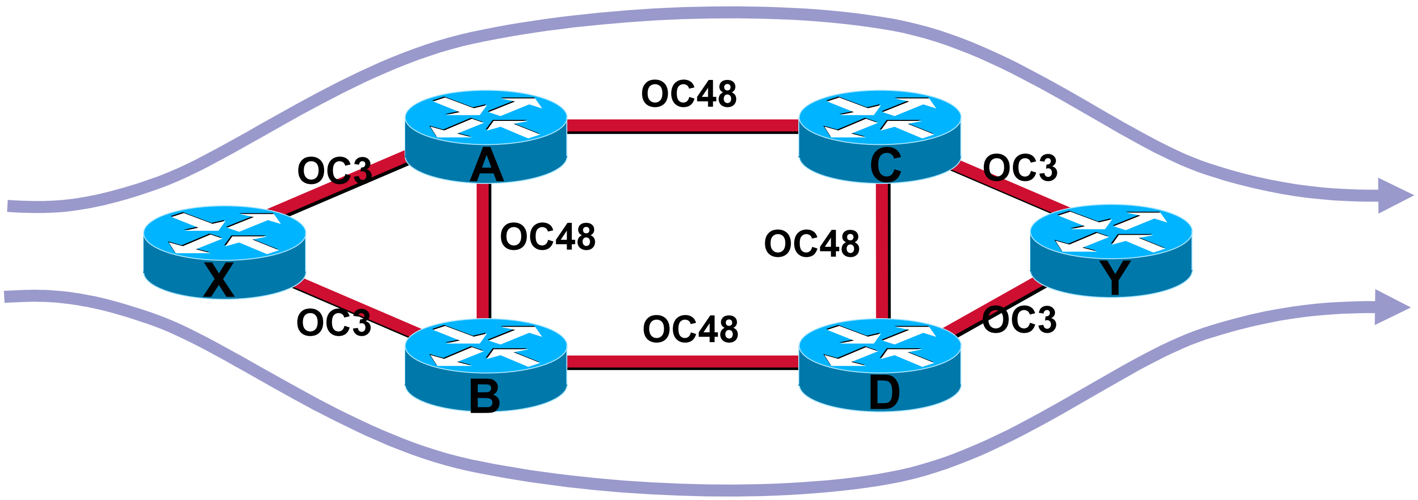
Equal Cost Multi-path (ECMP)

- CEF supports load-balancing across IGP equal cost paths on a per-destination or per-packet basis using ECMP
  - Generally per-destination load balancing is recommended to avoid impact that packet re-ordering can have on applications such as video and TCP
    - "Effect of Packet Reordering in a Backbone Link on Applications Throughput", Michael Laor, Lior Gendel, IEEE Network Magazine, September 2002
- Up to 8 equal cost IGP paths are supported in Cisco Express Forwarding (CEF)
  - Each is stored as a separate CEF adjacency

Equal Cost Multi-path (ECMP)

- Hash is performed on received packets to determine which one of the paths should be used for the packet
  - Hash is function of: source addr, dest addr, source port, dest port, with randomisation seeded by router ID
  - show ip cef <prefix> displays the path share
- Load balancing across equal cost paths achieved for general distributions of addresses and ports
- For more details on CEF load balancing search CCO for “Document ID: 18285”

Equal Cost Multi-path (ECMP)



- Where both topology and traffic demands are symmetrical, IGP ECMP load balancing may be sufficient with default metrics
  - Likely little benefit from IGP metric-based TE or MPLS TE
  - Many new networks have been designed in this way

Agenda

I. IP TE Introduction

II. Traffic Characterization

III. Traffic Matrices

IV. Approaches for IP TE

- ECMP
- IGP Metric Based TE
- MPLS TE

V. High availability options

IGP metric-based traffic engineering

Cisco.com Cariden.com

- **Has seen recent increase in interest**

B. Fortz, J. Rexford, and M. Thorup, "Traffic Engineering With Traditional IP Routing Protocols", IEEE Communications Magazine, October 2002.

D. Lorenz, A. Ordi, D. Raz, and Y. Shavitt, "How good can IP routing be?", DIMACS Technical, Report 2001-17, May 2001.

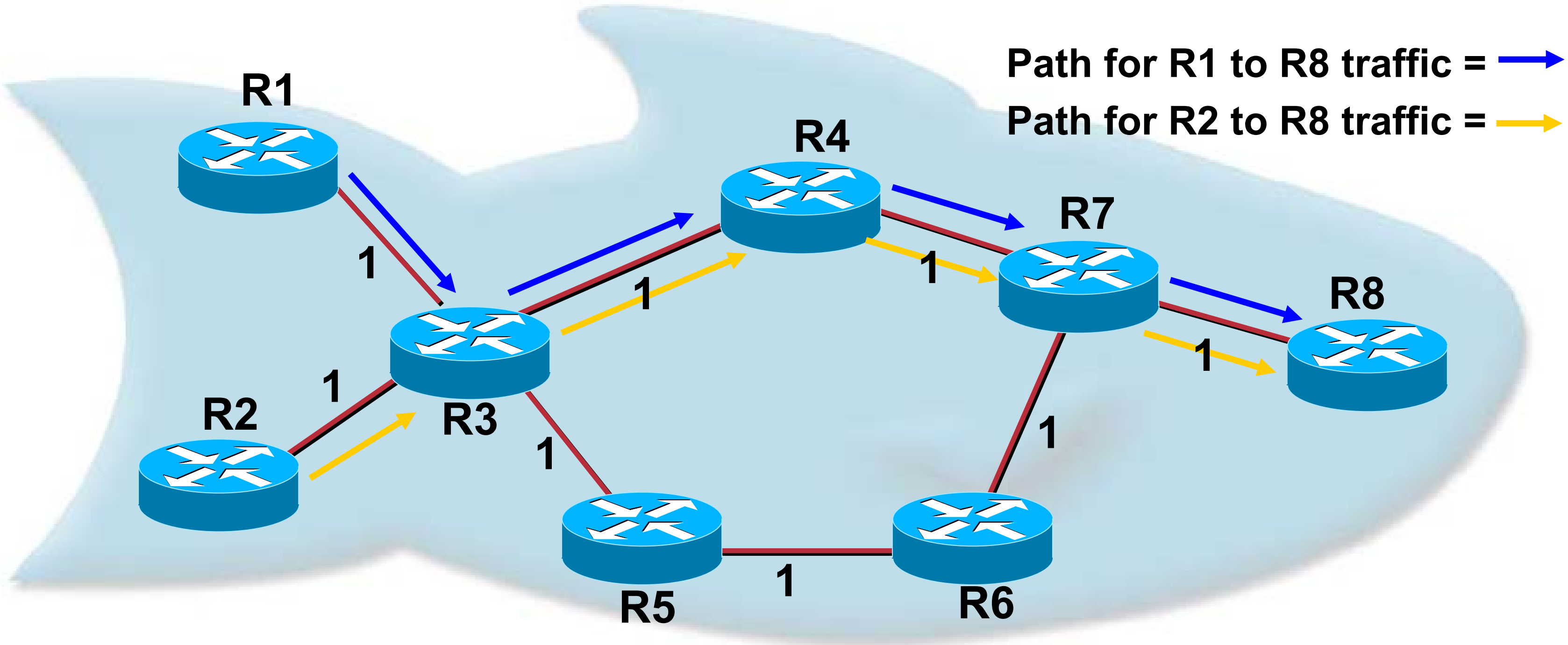
L. S. Buriol, M. G. C. Resende, C. C. Ribeiro, and M. Thorup, "A memetic algorithm for OSPF routing" in Proceedings of the 6th INFORMS Telecom, pp. 187188, 2002.

M. Ericsson, M. Resende, and P. Pardalos, "A genetic algorithm for the weight setting problem in OSPF routing" J. Combinatorial Optimization, volume 6, no. 3, pp. 299-333, 2002.

W. Ben Ameur, N. Michel, E. Gourdin et B. Liau. Routing strategies for IP networks. Telektronikk, 2/3, pp 145-158, 2001.

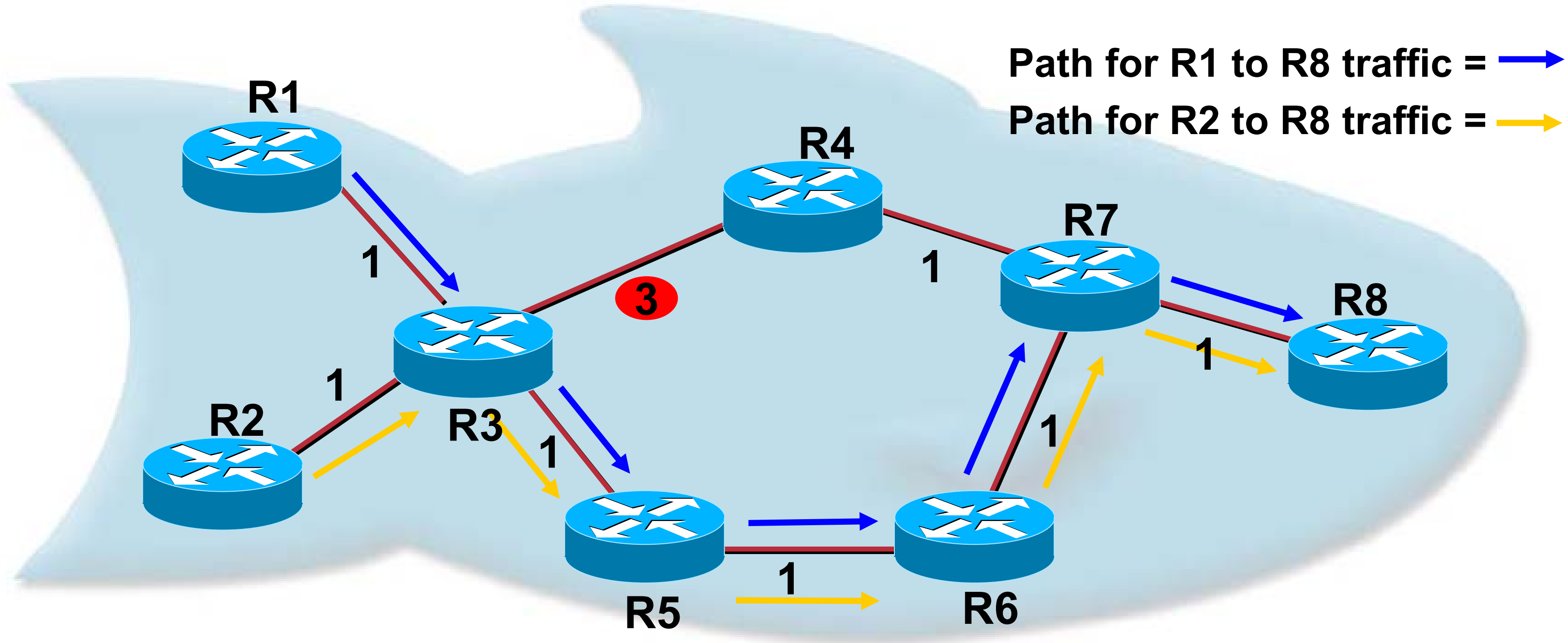
…

IP Traffic Engineering: The Problem



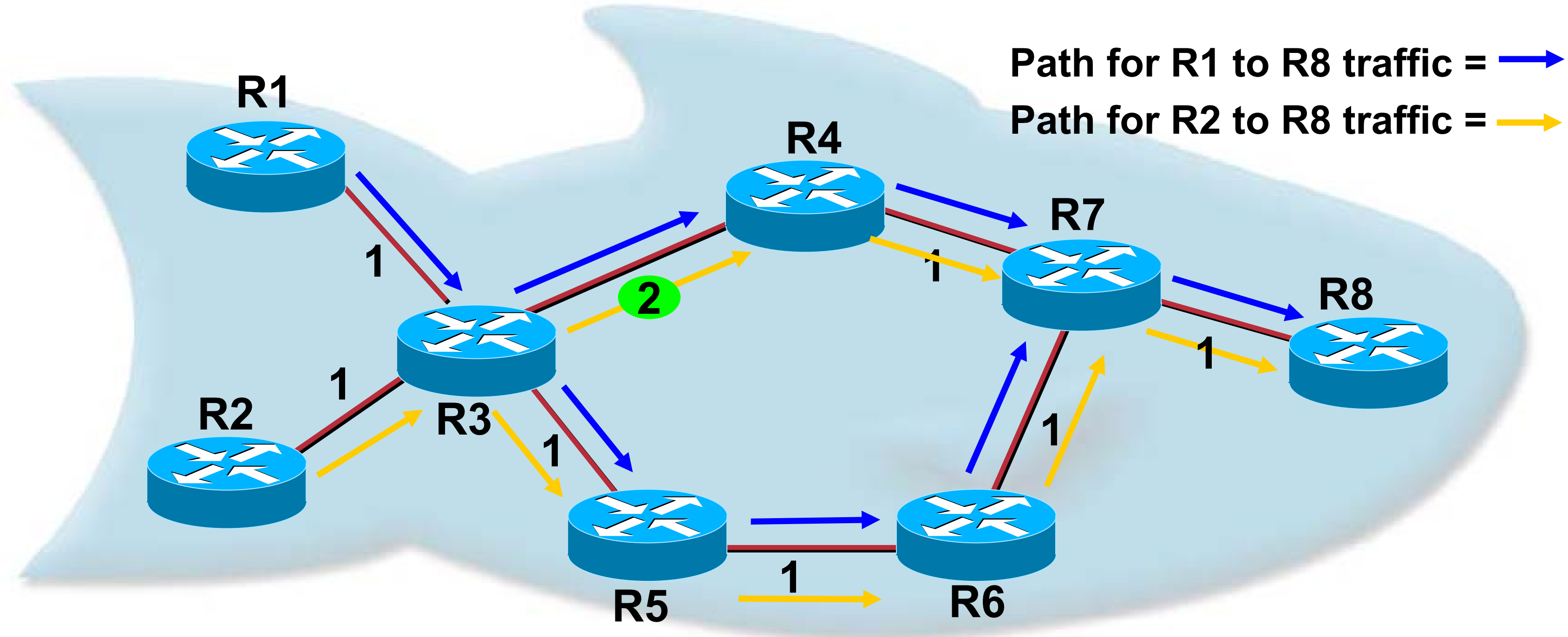
… but changing the link metrics will just move the problem around the network?

IGP metric-based traffic engineering



… but changing the link metrics will just move the problem around the network?

IGP metric-based traffic engineering



- …the mantra that tweaking IGP metrics just moves problem around is a generalisation which may not always be true in practise

  Note: IGP metric-based TE can use ECMP

IGP Metric Based TE: Deployment Strategies

IP Traffic Engineering

ECMP

IGP Metric Based TE

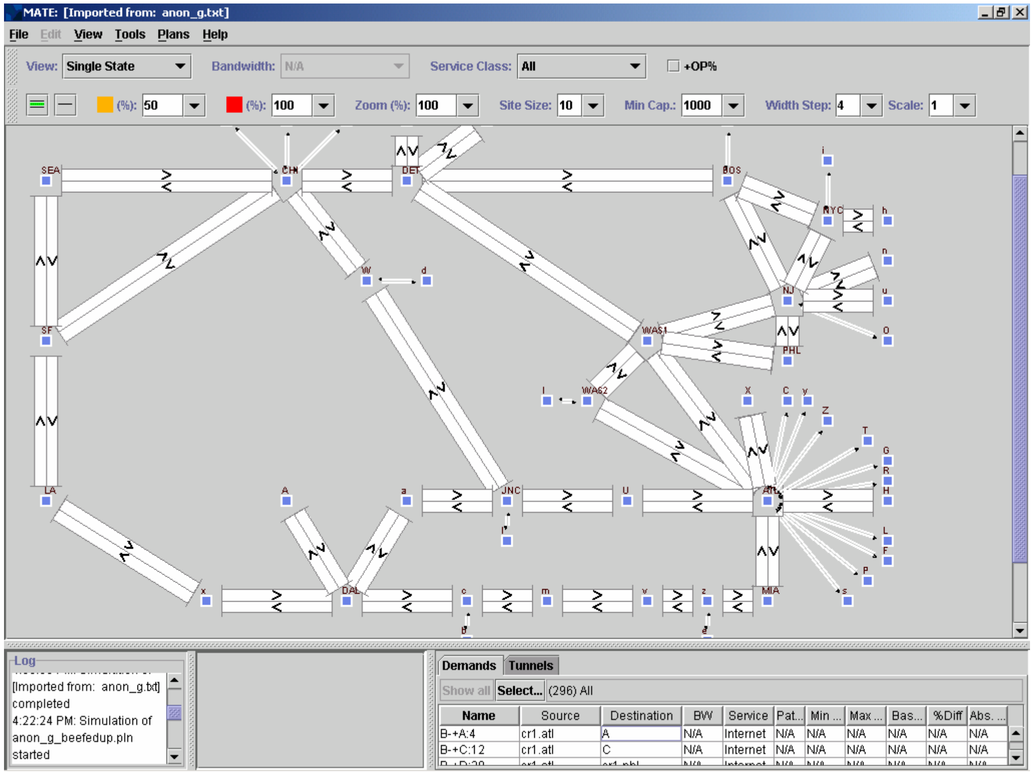
MPLS TE

Tactical:
Individual link metrics tweaked (manually) ad hoc to relieve specific points of congestion

Strategic:
Holistic approach with a defined optimization goal – computer-aided tools can ease the task

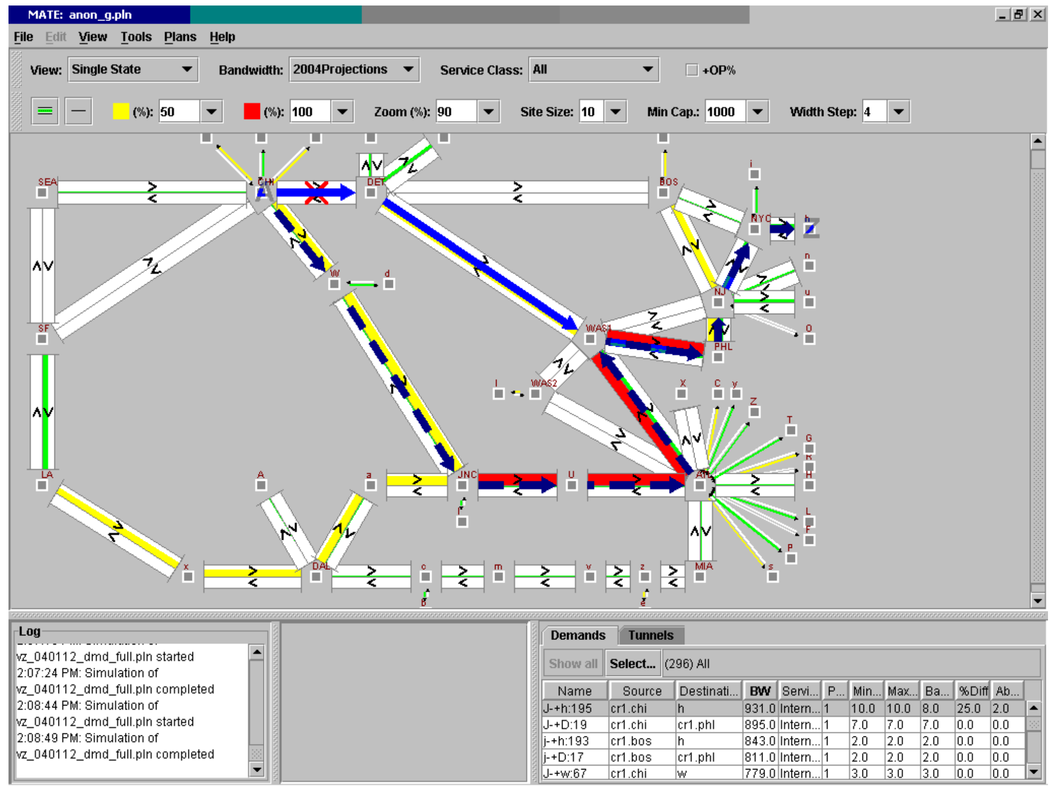
IGP metric-based traffic engineering: Case study

- Proposed OC-192 U.S. Backbone
- Connect Existing Regional Networks
- Anonymized (by permission)
- Live Demo (Some Stills)



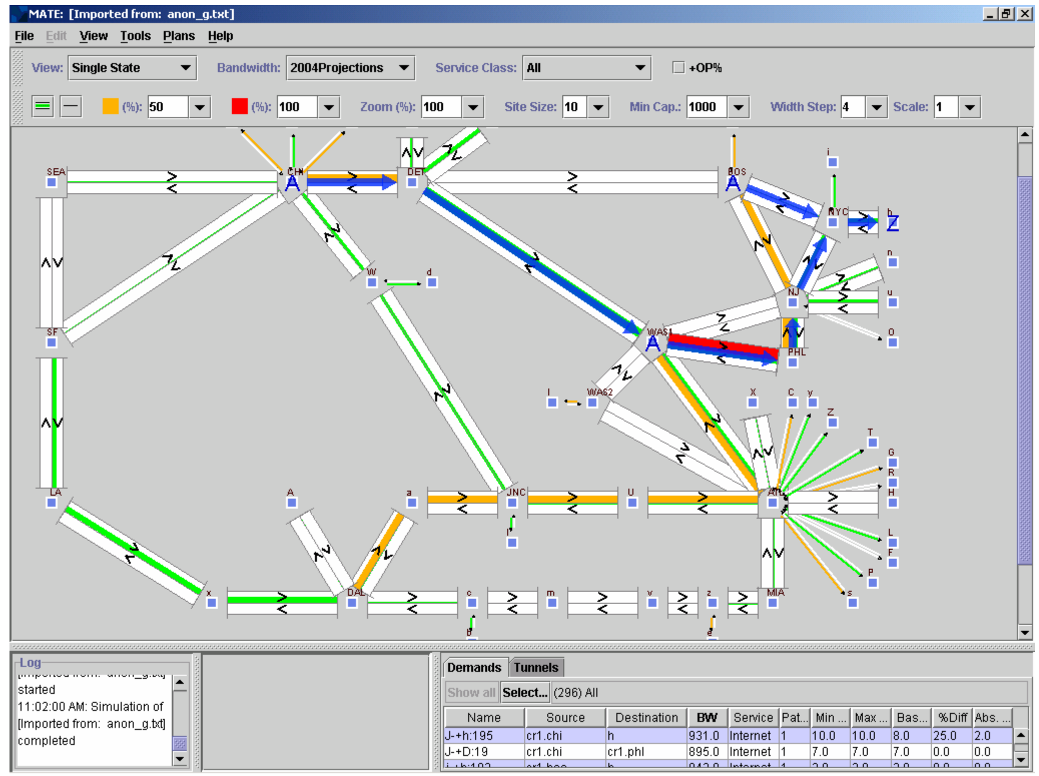
Metric TE Case Study:
Plot Legend

- Squares ~ Sites (PoPs)
- Routers in Detail Pane (not shown here)
- Lines ~ Physical Links
  - Thickness ~ Speed
  - Color ~ Utilization
    - Yellow ≥ 50%
    - Red ≥ 100%
- Arrows ~ Routes
  - Solid ~ Normal
  - Dashed ~ Under Failure
- X ~ Failure Location



Metric TE Case Study: Traffic Overview

- **Major Sinks in the Northeast**
- **Major Sources in CHI, BOS, WAS, SF**
- **Congestion Even with No Failure**



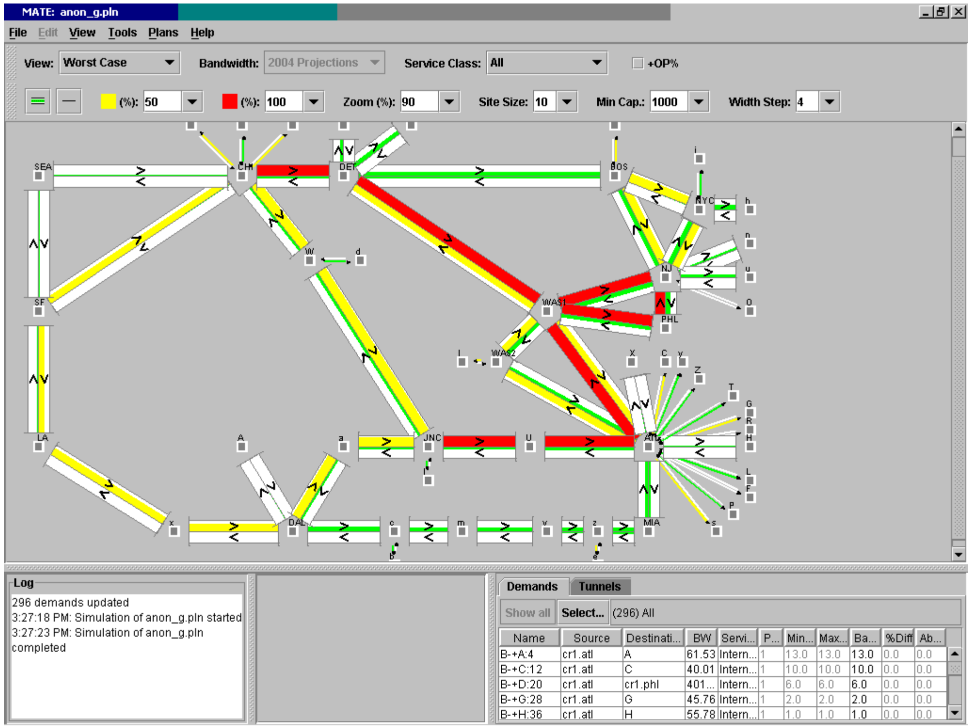
Metric TE Case Study:
Manual Attempt at Metric TE

- **Shift Traffic from Congested North**
- **Under Failure traffic shifted back North**

Metric TE Case Study: Worst Case Failure View

Cisco.com Cariden.com

- **Enumerate Failures**
- **Display Worst Case Utilization per Link**
- **Links may be under Different Failure Scenarios**
- **Central Ring+ Northeast Require Upgrade**

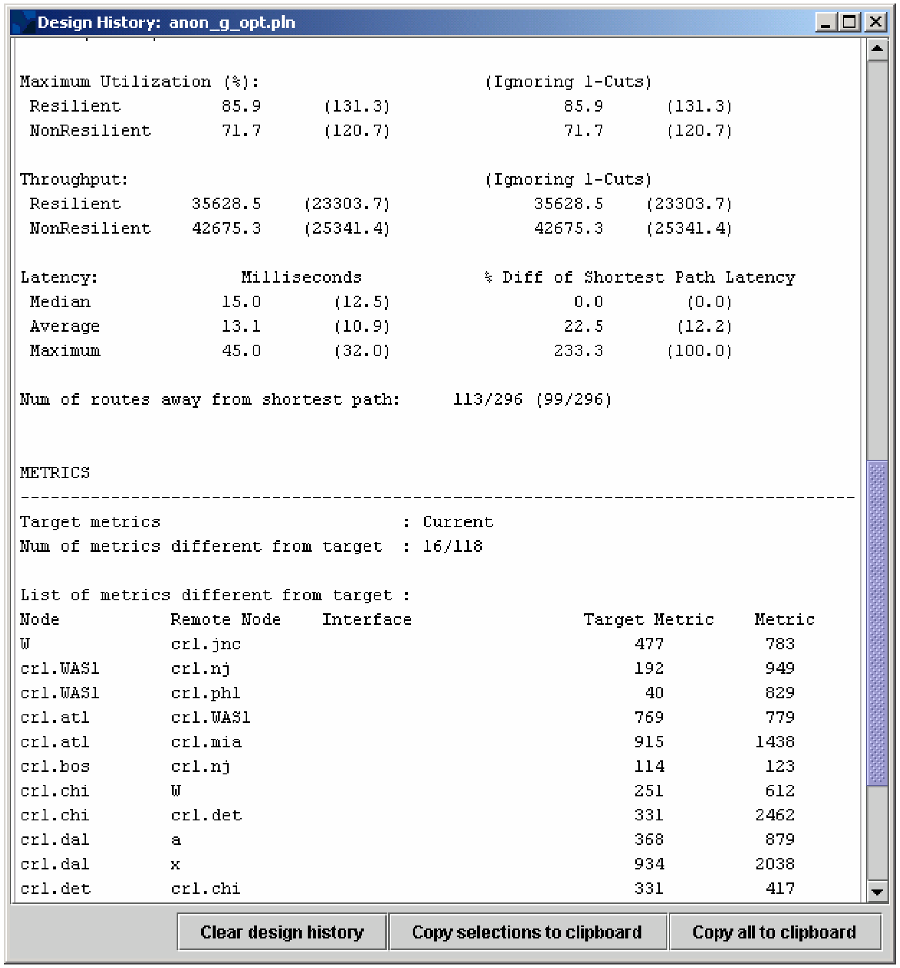


Metric TE Case Study: Cariden Metric TE

- **Change 16 metrics**
- **Remove congestion**

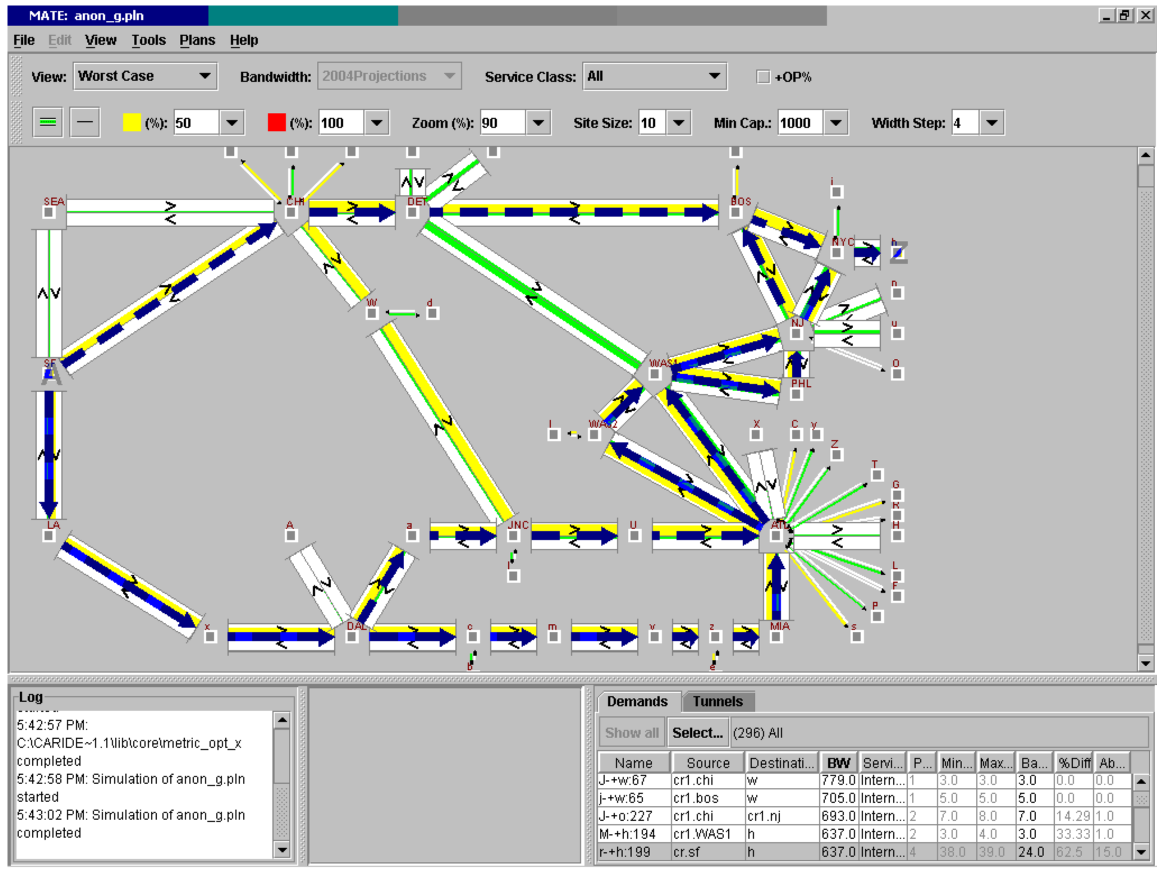
  Normal (121% -> 72%)

  Worst case link failure (131% -> 86%)



Metric TE Case Study:
New Routing Visualisation

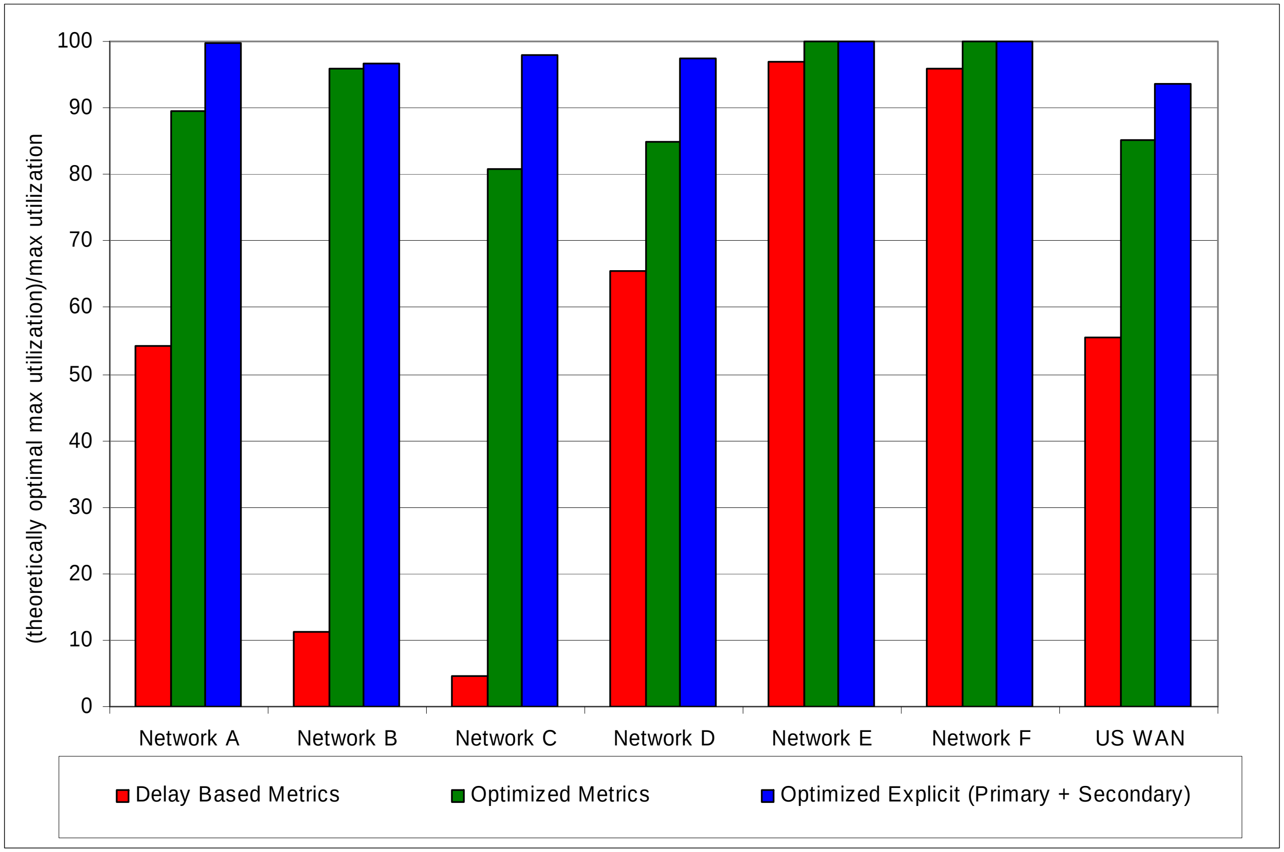
- **ECMP in congested region**
- **Shift traffic to outer circuits**
- **Share backup capacity: outer circuits fail into central ones**



Metric TE Case Study: Performance over Various Networks

Cisco.com Cariden.com

- See:
  NANOG 27
  APRICOT ‘04
- Study on Real Networks
- Single Set of Metrics Achieve 80-95% of Theoretical Best across Failures



Agenda

I. IP TE Introduction

II. Traffic Characterization

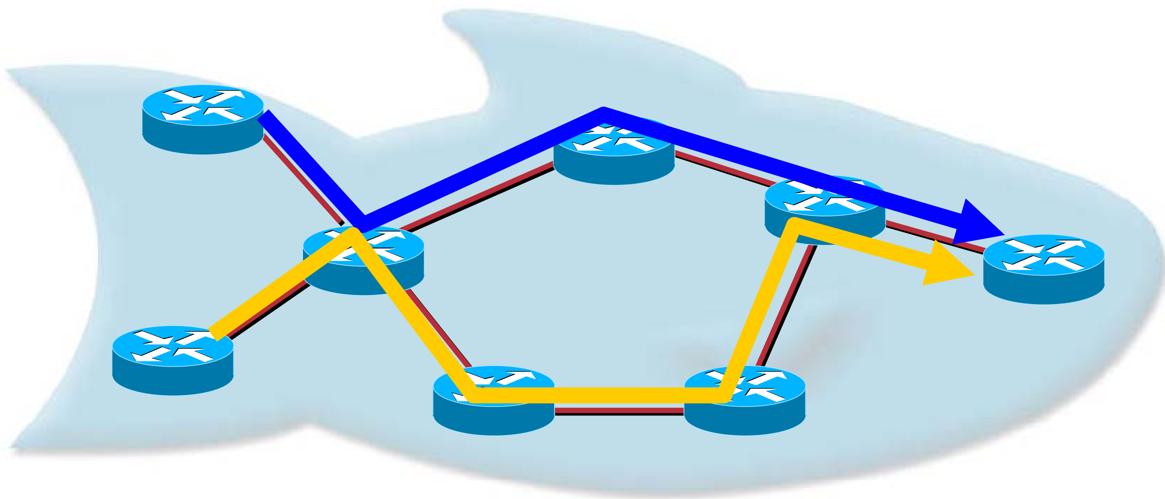
III. Traffic Matrices

IV. Approaches for IP TE

- ECMP
- IGP Metric Based TE
- MPLS TE

V. High availability options

MPLS Traffic Engineering

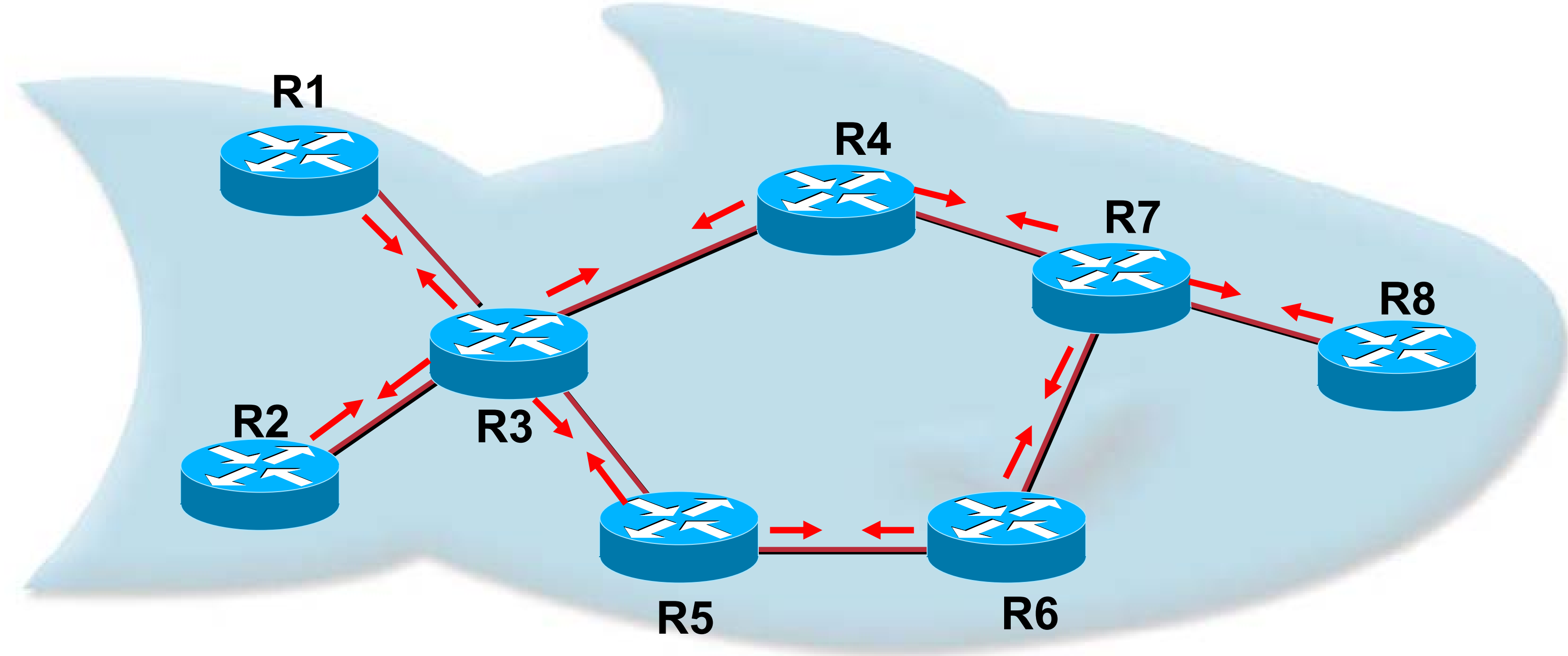


- **MPLS Traffic Engineering gives us an “explicit” routing capability (a.k.a. “source routing”) at Layer 3**
- **Lets you use paths other than IGP shortest path**
- **Allows unequal-cost load sharing**
- **MPLS TE label switched paths (termed “traffic engineering tunnels”) are used to steer traffic through the network**

MPLS TE Components – Refresher

1) Resource / policy information distribution
2) Constraint based path computation
3) RSVP for tunnel signaling
4) Link admission control
5) LSP establishment
6) TE tunnel control and maintenance
7) Assign traffic to tunnels

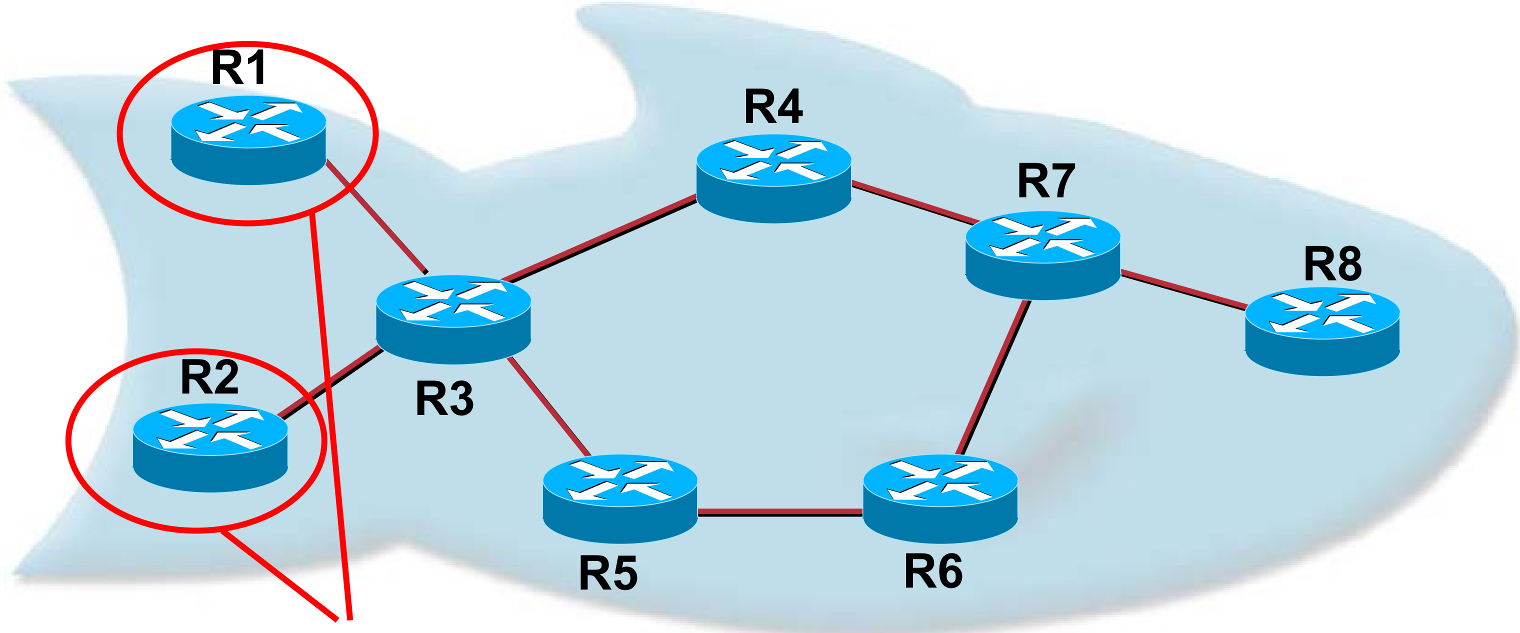
MPLS TE Components (1)



- **Resource / policy information distribution**

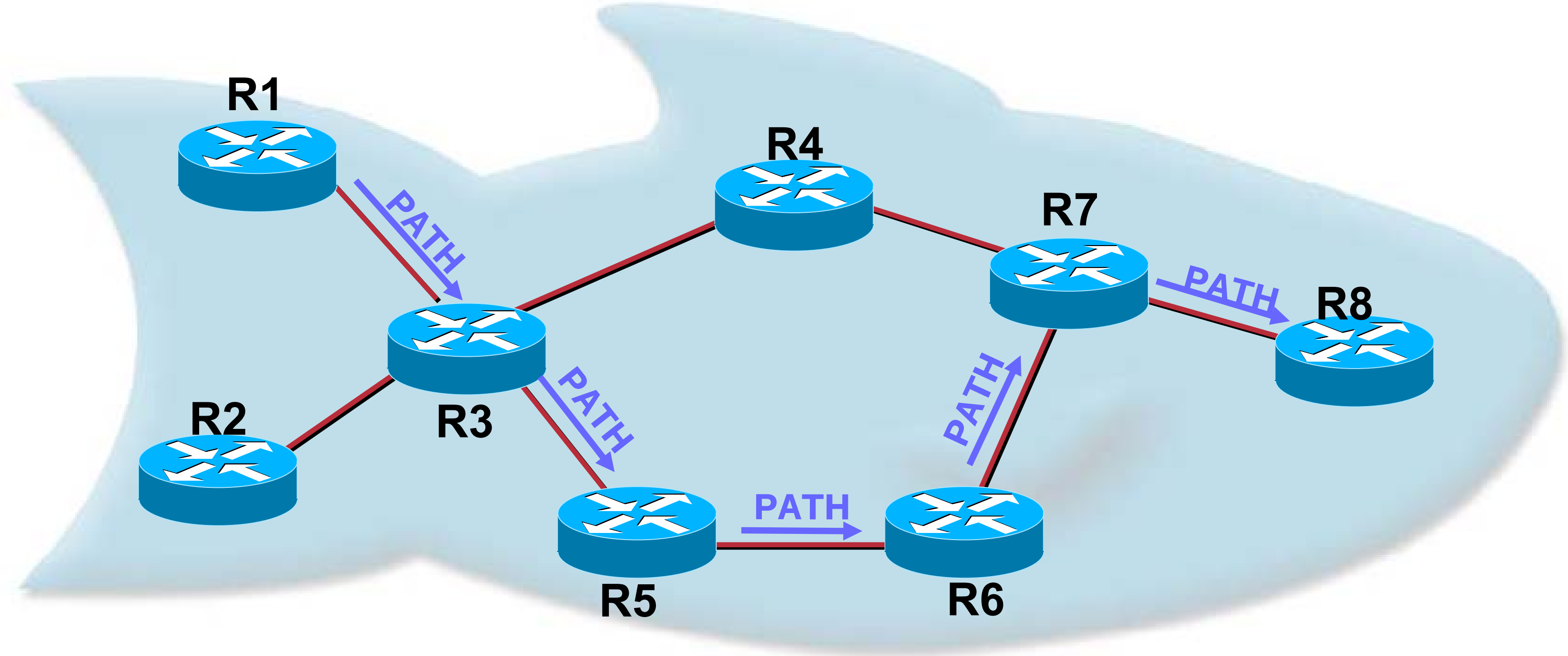
  OSPF / IS-IS extensions are used to advertise "unreserved capacity" and administrative attributes per link

MPLS TE Components (2)



- **Constraint based path computation**
  - **Constraints (required bandwidth and policy) are specified for a TE "tunnel"**
  - **Constraint based routing – PCALC on head-end routers calculates best path that satisfies constraints based upon the received topology and policy information**
    - **prune unsuitable links from the topology and pick shortest path on the remaining topology**

MPLS TE Components (3)

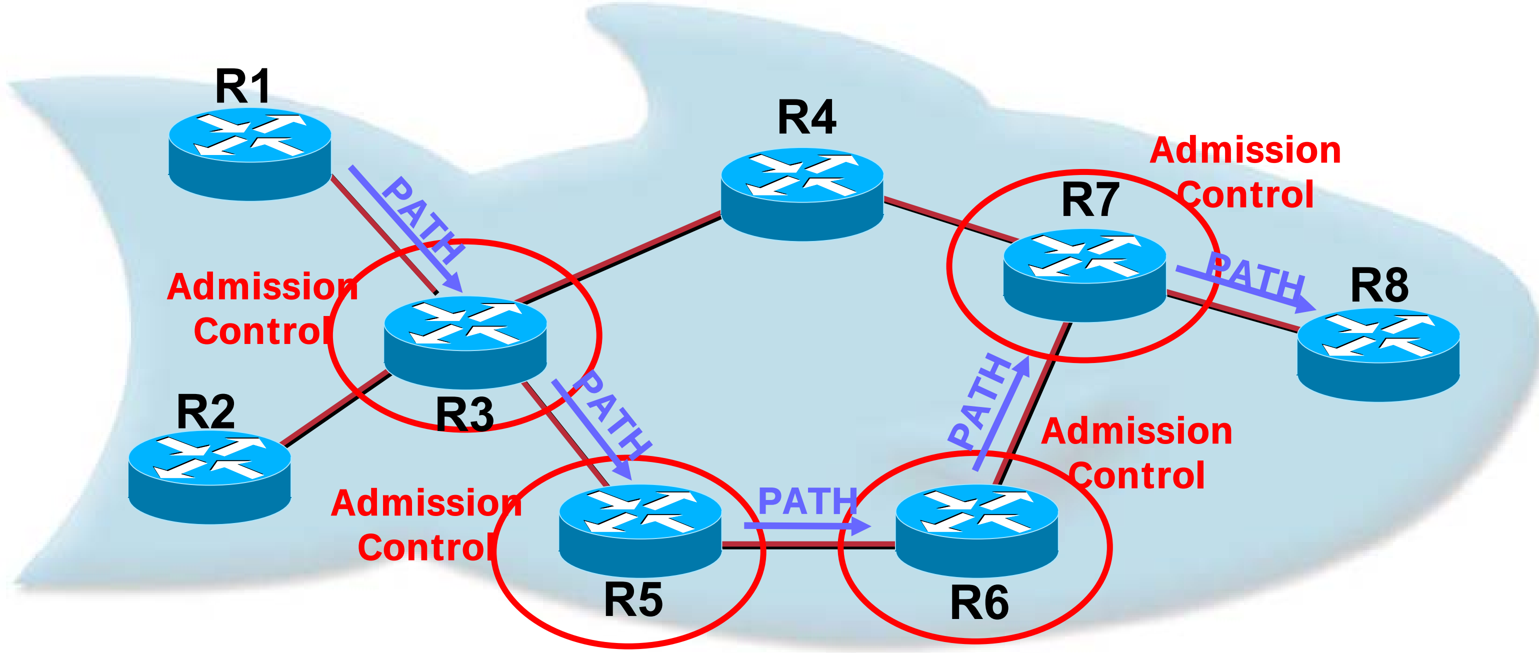


- **RSVP for Tunnel Signaling**

  Output of constraint based routing is an explicit route used by RSVP (with extensions) for tunnel signaling

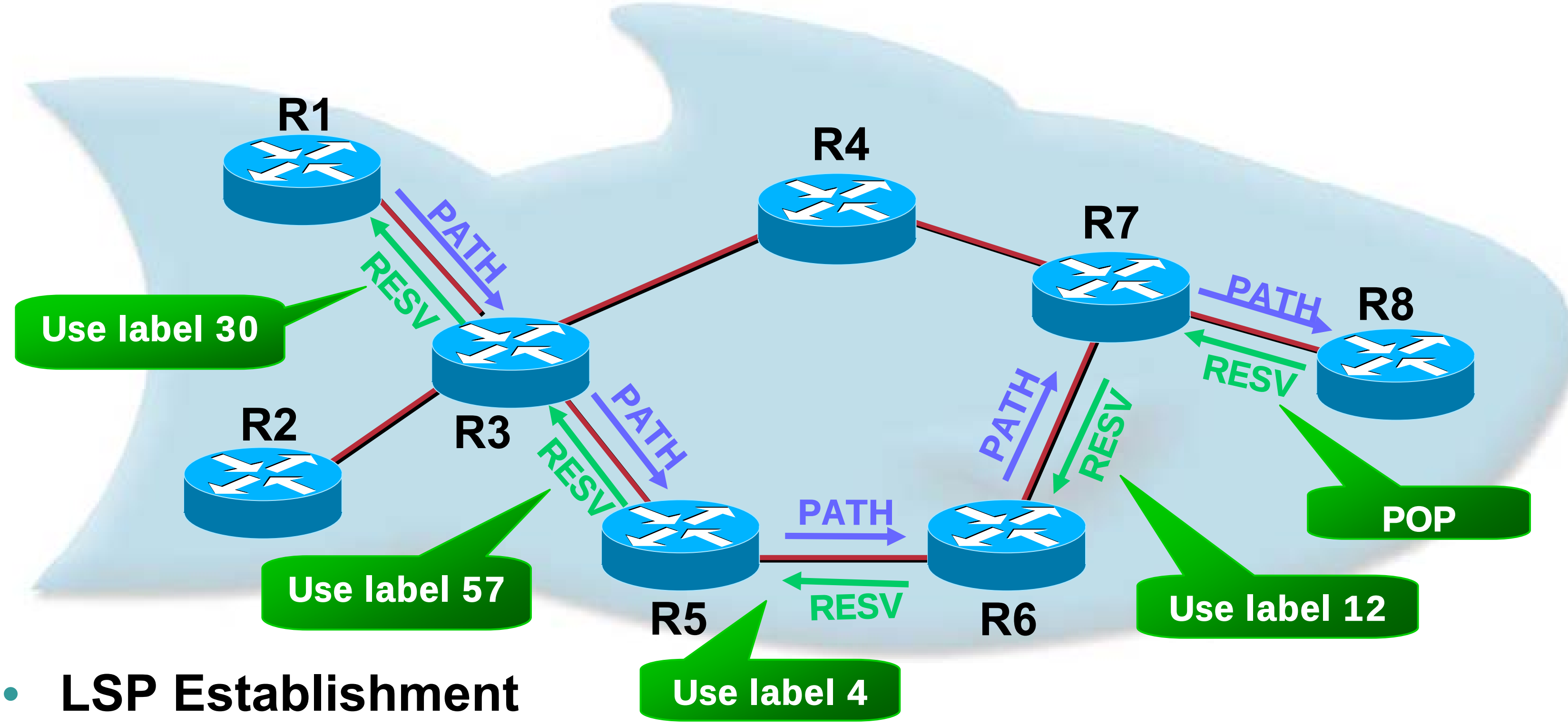
  ERO = R1➔R3➔R5➔R6➔R7➔R8

MPLS TE Components (4)



- **Link admission control**
  - At each hop – determines if resources are available
    - If Admission Control fails, send PathError
    - May tear down (existing) TE LSPs with a lower priority
    - Triggers IGP information distribution when resource thresholds are crossed

MPLS TE Components (5)



- **LSP Establishment**

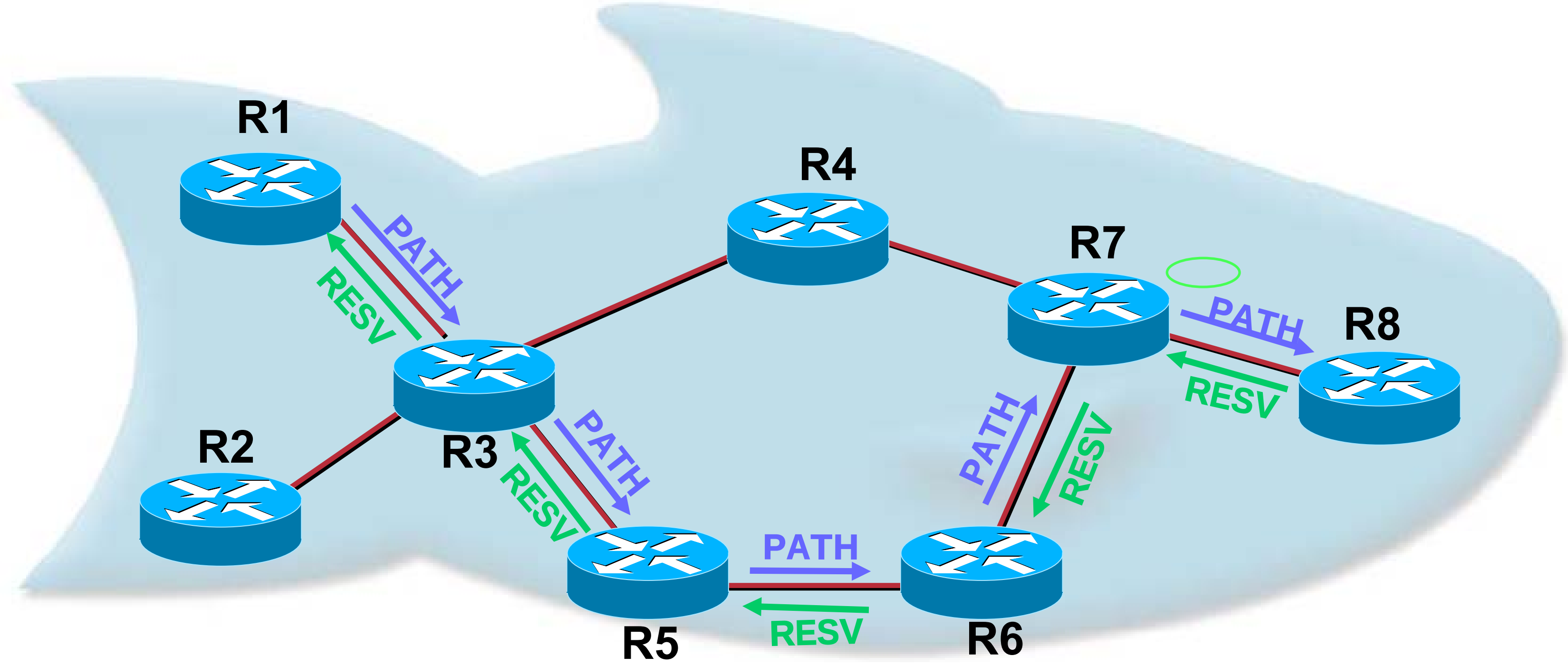
  RESV confirms bandwidth reservation and distributes labels

  ➔ downstream on demand label allocation

  MPLS used for forwarding – overcomes issues of IP destination based forwarding

MPLS TE Components (6)

Cisco.com Cariden.com

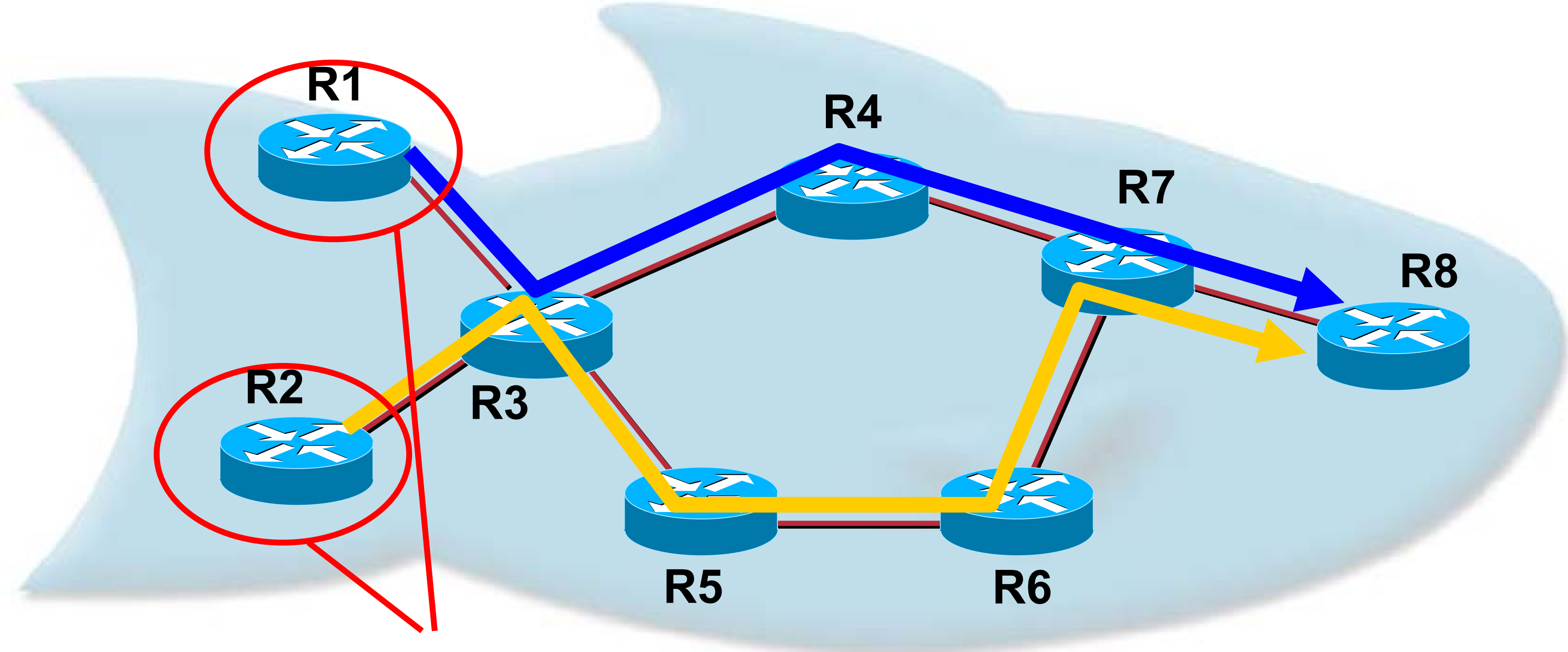


- **TE tunnel control and maintenance**

  Periodic RSVP PATH/RESV messages maintain tunnels

  Unlike tunnel set up, tunnel refresh messages are independent and asynchronous

MPLS TE Components (7)



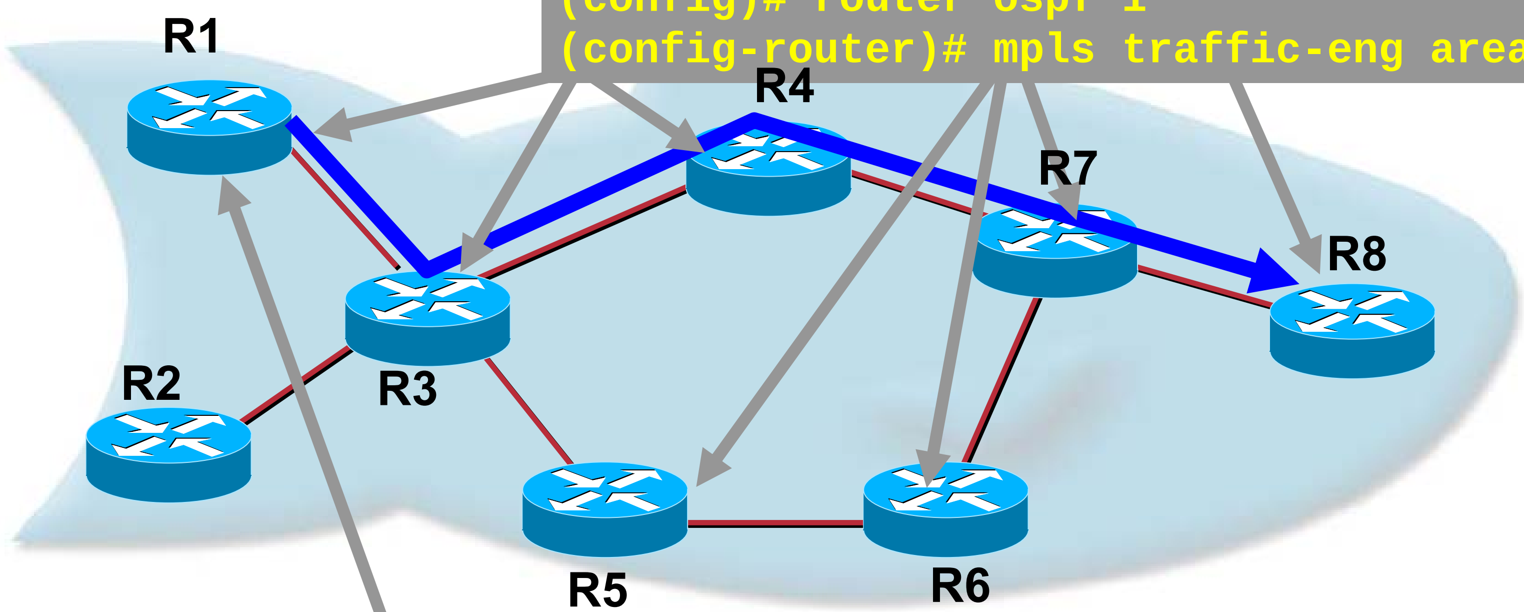
- **Assign traffic to tunnels**

  Head-end routers assign traffic to tunnels using:

    Static routing, Autoroute or PBR

MPLS TE Components: Minimum Config


```
(config-if)# mpls traffic-eng tunnels
(config-if)# ip rsvp bandwidth 150000 150000
(config)# router ospf 1
(config-router)# mpls traffic-eng area 0
```




```
(config)# interface tunnel 1
(config-if)# ip unnumbered Loopback0
(config-if)# tunnel destination 24.1.1.1
(config-if)# tunnel mode mpls traffic-eng
(config-if)# tunnel mpls traffic-eng priority 0 0
(config-if)# tunnel mpls traffic-eng path-option 1 dynamic
(config-if)# tunnel mpls traffic-eng autoroute announce
```

MPLS TE Deployment Strategies

MPLS TE

Strategic: All traffic transported using TE tunnels

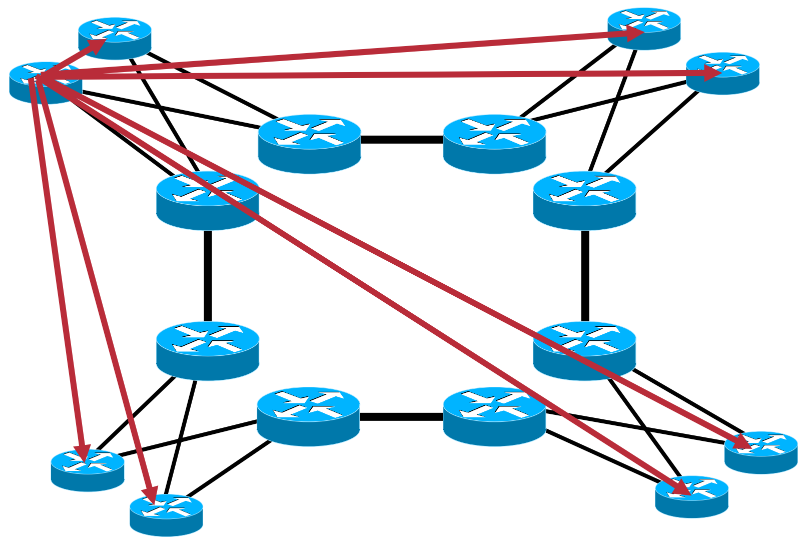
- **Full mesh**
- **Core mesh**
- **Regional mesh**

Can be static (offline) or dynamic (online)

Tactical: Few TE tunnels set up to move a subset of traffic away from congested links

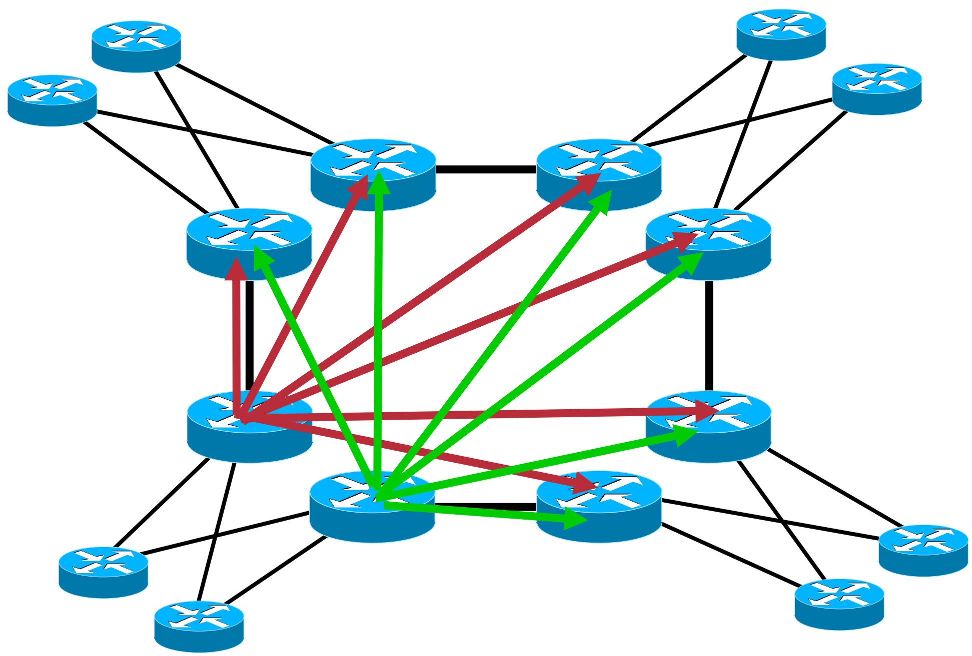
Tunnels paths typically static and determined offline

Strategic Deployment: Full Mesh



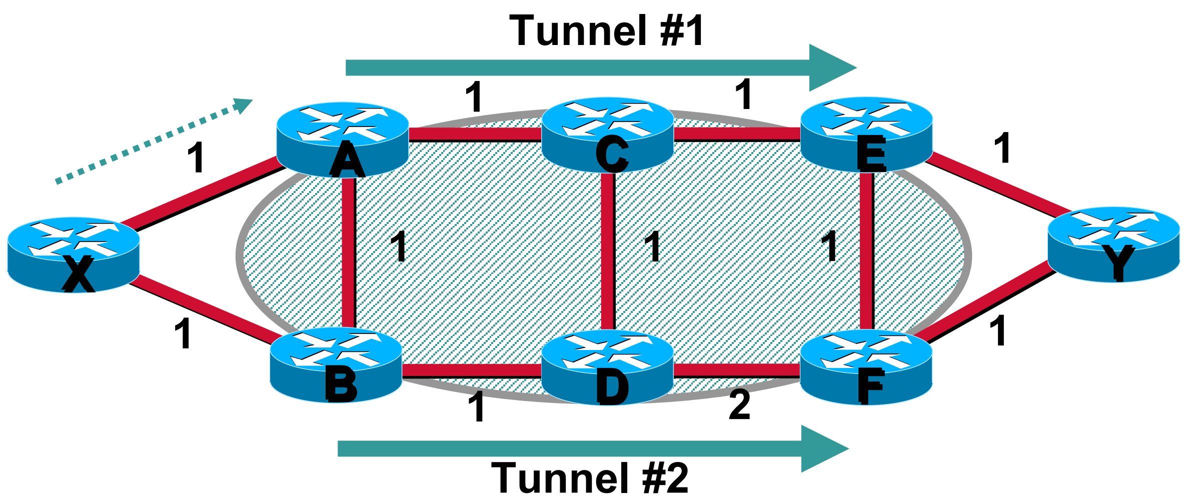
- Requires n * (n-1) tunnels, where n = # of head-ends
- Reality check: largest TE network today has ~100 head-ends
    - ➔ ~9,900 tunnels in total
    - ➔ max 99 tunnels per head-end
    - ➔ max ~1,500 tunnels per link
- Provisioning burden may be eased with AutoTunnel Mesh Groups

Strategic Deployment: Core Mesh



- **Reduces number of tunnels required**
- **Can be susceptible to “traffic-sloshing”**

Traffic "sloshing"



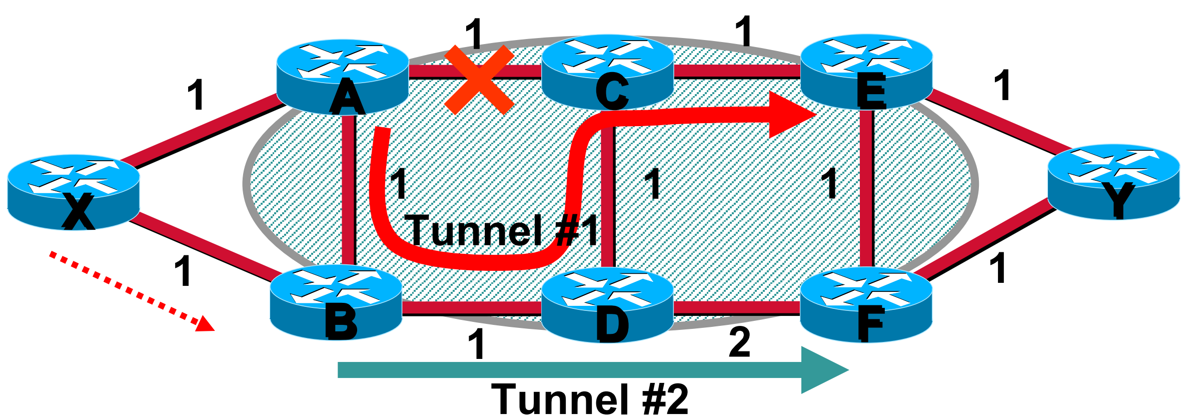
- **In normal case:**

  For traffic from X ➔ Y, router X IGP will see best path via router A

  Tunnel #1 will be sized for X ➔ Y demand

  If bandwidth is available on all links, Tunnel from A to E will follow path A ➔ C ➔ E

Traffic “sloshing”



- **In failure of link A-C:**

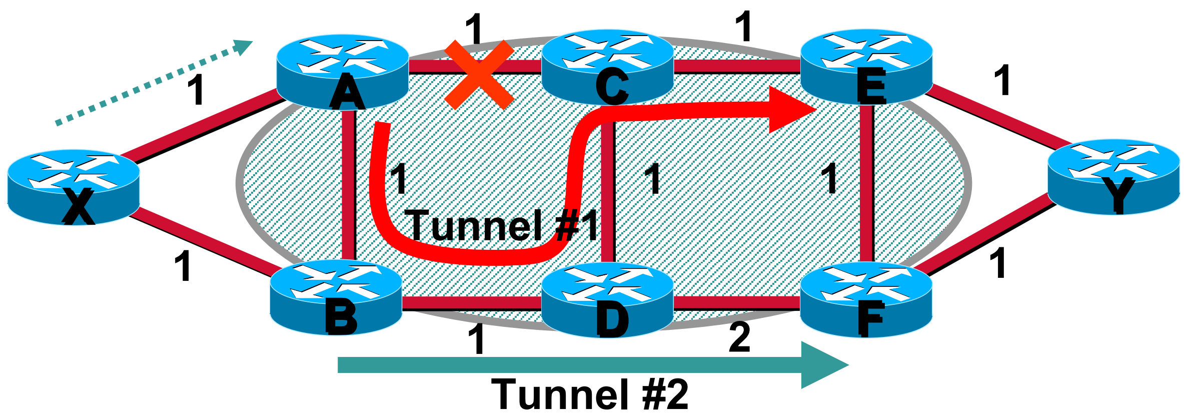
  For traffic from X ➔ Y, router X IGP will now see best path via router B

  However, if bandwidth is available, tunnel from A to E will be re-established over path A ➔ B ➔ D ➔ C ➔ E

  Tunnel #2 will not be sized for X ➔ Y demand

  Bandwidth may be set aside on link A ➔ B for traffic which is now taking different path

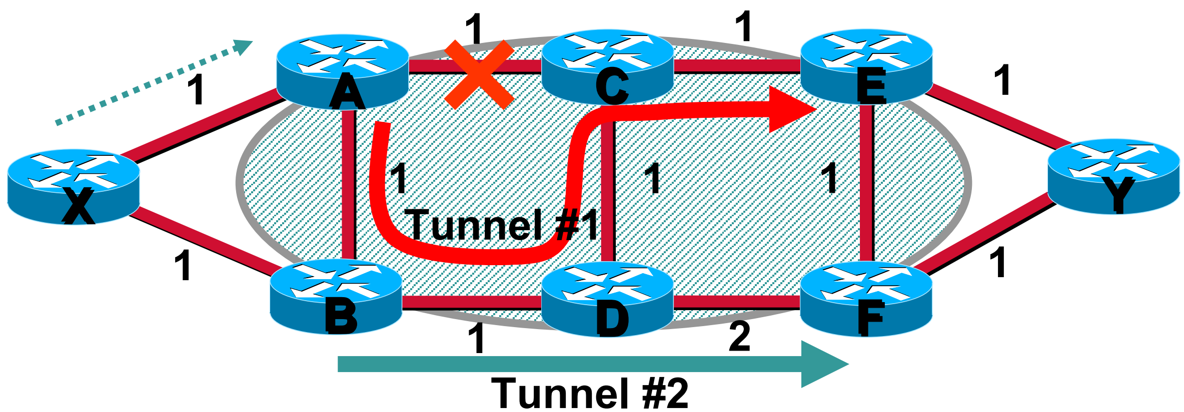
Traffic “sloshing”



- **Forwarding adjacency (FA) could be used to overcome traffic sloshing**
  - **Normally, a tunnel only influences the FIB of its head-end and other nodes do not see it**
  - **With FA the head-end advertises the tunnel in its IGP LSP**
    - **Tunnel #1 could always be made preferable over tunnel #2 for traffic from X ➔ Y**
- **Holistic view of traffic demands (core traffic matrix) and routing (in failures if necessary) is necessary to understand impact of TE**

Traffic "sloshing"

Cisco.com Cariden.com



- **Forwarding adjacency could be used to overcome traffic sloshing**

  Normally, a tunnel only influences the FIB of its head-end

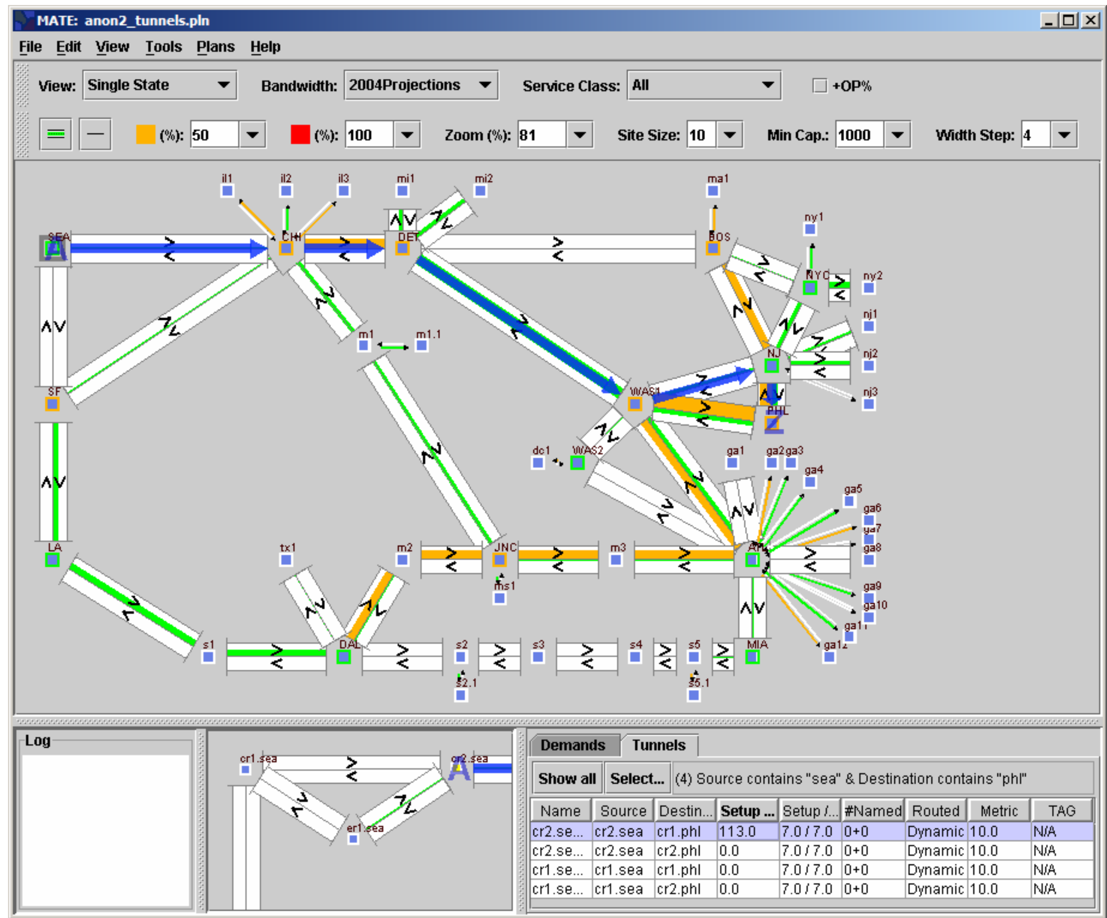
  other nodes do not see it

  With Forwarding Adjacency the head-end advertises the tunnel in its IGP LSP

  Tunnel #1 could always be made preferable over tunnel #2 for traffic from X → Y

Traffic “sloshing”: A Real Example (I)

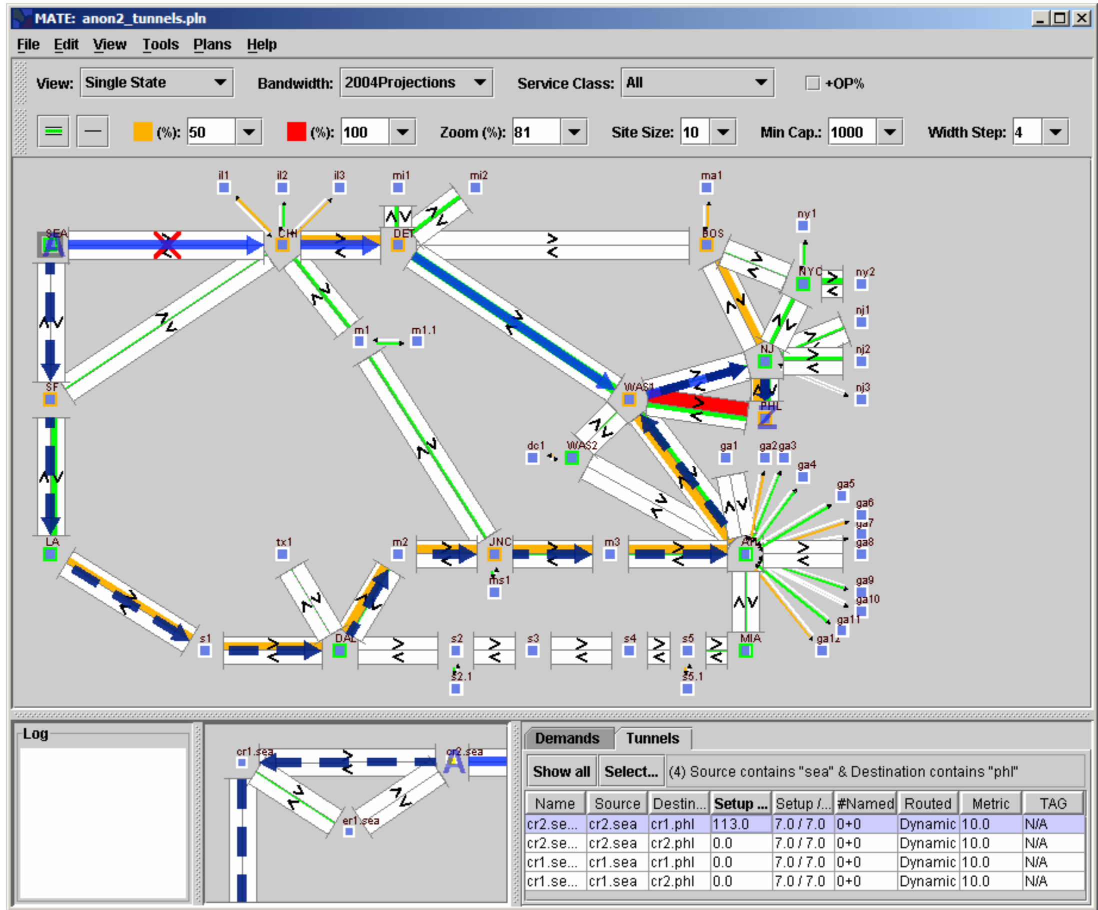
- 2 core routers in SEA
- x2 core routers in PHL
- = 4 tunnels between all pairs
- One of these pairs has the shortest IGP path between them
- So all traffic from SEA-PHL goes on this tunnel



Traffic “sloshing”: A Real Example (II)

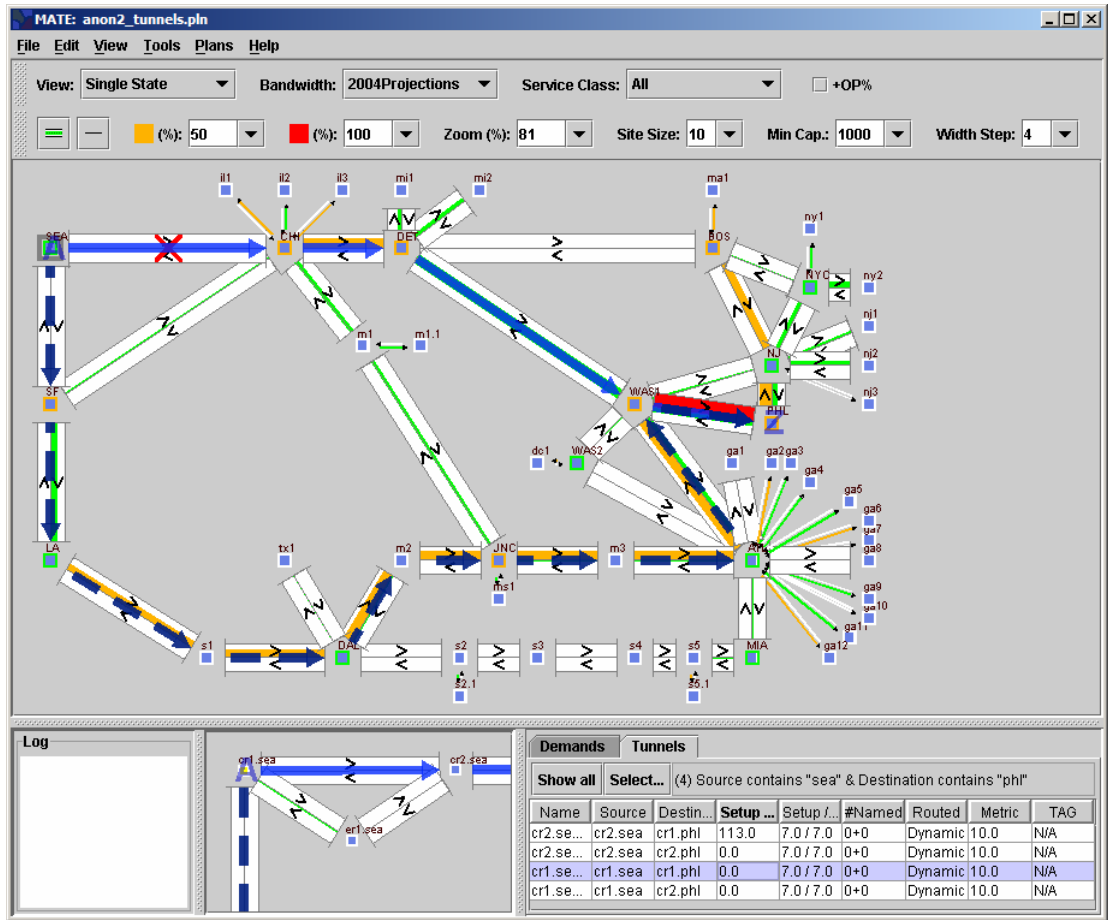
Cisco.com Cariden.com

- **This tunnel reserves enough space for all traffic through it.**
- **So under failure, finds alternate path avoiding congested links**



Traffic “sloshing”: A Real Example (III)

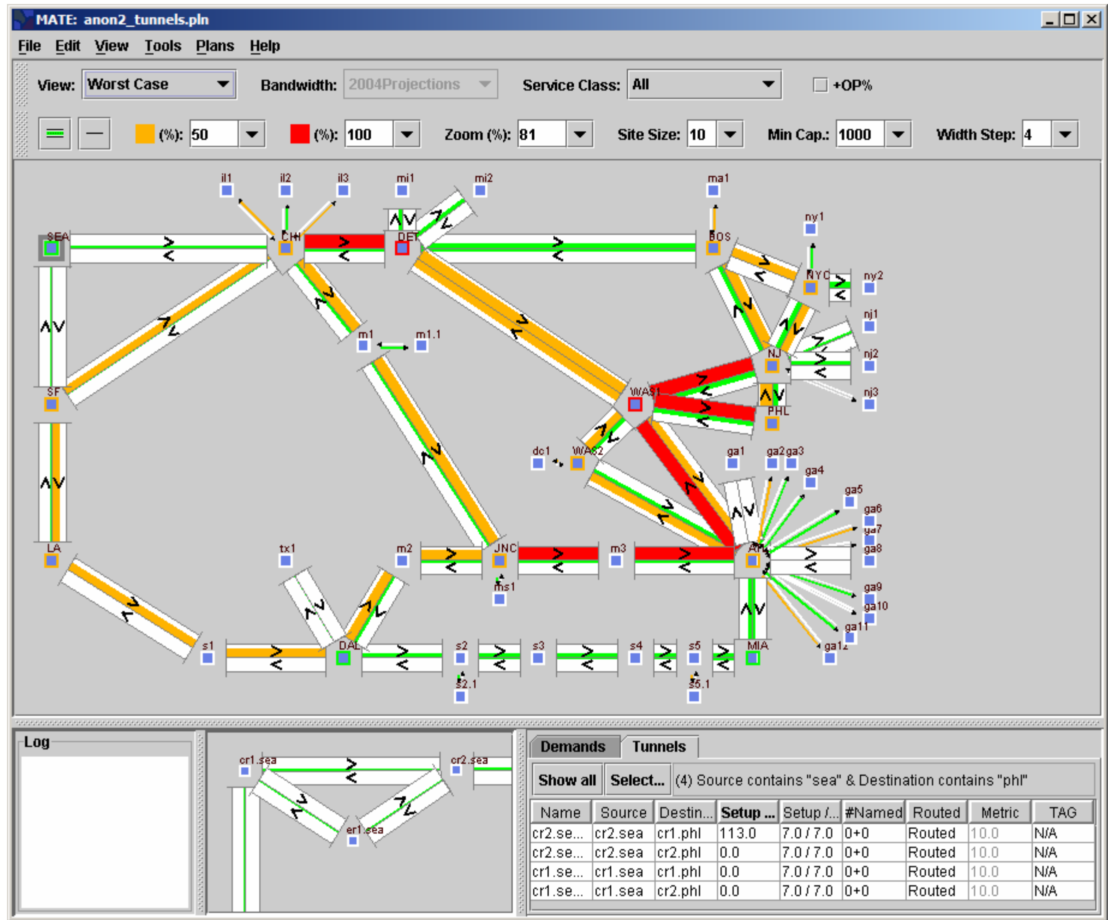
- **BUT, under failure a different pair of core routers is now closest by IGP metric**
- **So traffic “sloshes” to new tunnel**
- **New tunnel has zero bandwidth reserved, so has taken congested path.**
- **Traffic in new tunnel congests network further.**



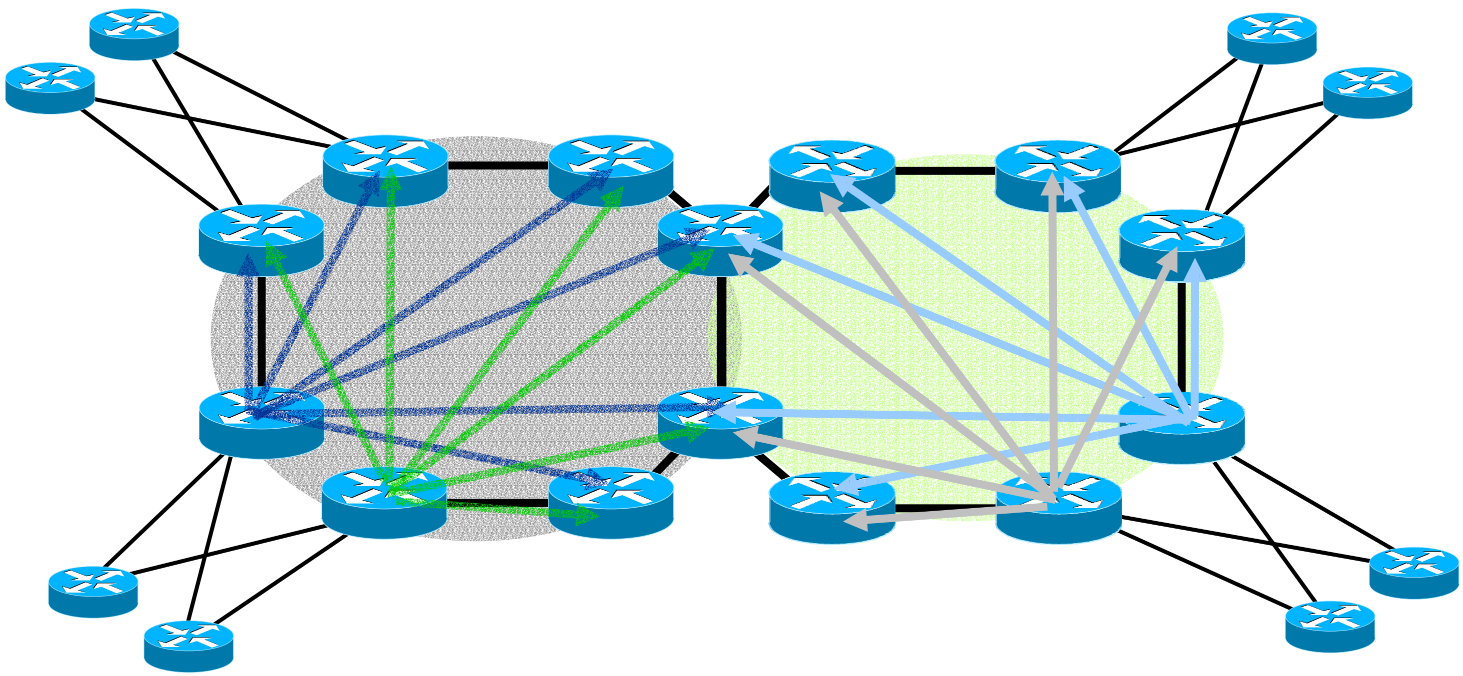
Traffic “sloshing”: A Real Example (IV)

Cisco.com Cariden.com

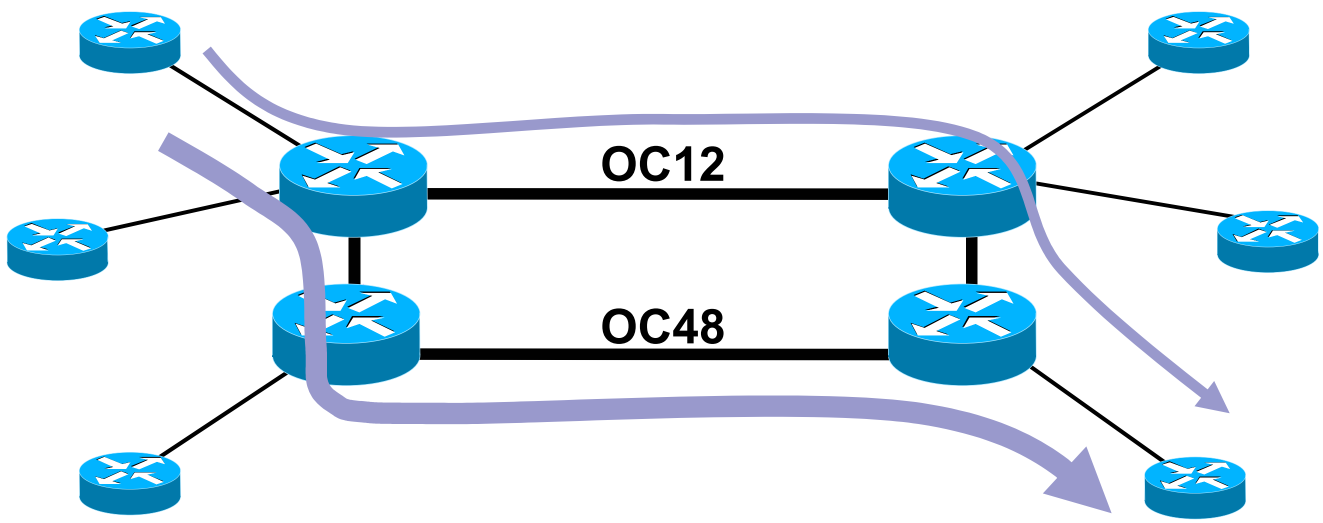
- Worst-case view: “sloshing” causes congestion under failure in many circuits.
- Metric-based optimization on same network. Maximum utilization = 86% under any circuit failure.



Strategic Deployment: Regional (or hierarchical) Mesh



Tactical Deployment



- Explicit path configured on head-end for each tunnel to offload traffic from congested links
- Can use unequal cost load balancing based upon configured bandwidth or load-share ratio
- Can be useful when faced with:
  - Unexpected traffic demands
  - Long bandwidth lead-times

MPLS TE deployment considerations

- **Statically (explicit) or dynamically established tunnels**
  - **Dynamic path option**
    - **Must specify bandwidths for tunnels**
      - **Otherwise defaults to IGP shortest path**
    - **Dynamic tunnels introduce indeterminism and cannot solve "tunnel packing" problem**
      - **Order of setup can impact tunnel placement**
      - **Each head-end only has a view of their tunnels**
      - **Tunnel prioritisation scheme can help – higher priority for larger tunnels**

MPLS TE deployment considerations

- **Statically (explicit) or dynamically established tunnels (contd.)**
  - Static – explicit path option
    - More deterministic, and able to provide better solution to “tunnel packing” problem
      - Offline system has view of all tunnels from all head-ends
    - If strategic approach then computer-aided tools can ease the task of primary tunnel placement

Tunnel Sizing

- Tunnel sizing is key …
    - Needless congestion if actual load exceed reserved bandwidth
    - Needless tunnel rejection if reservation >> actual load
        - Enough capacity for actual load but not for the tunnel reservation
        - Traffic reverts to SPF, which is presumably set for latency not for traffic distribution

Tunnel Sizing

- Actual heuristic for tunnel sizing will depend upon dynamism of tunnel sizing
  - Need to set tunnel bandwidths dependent upon tunnel traffic characteristic over optimisation period
- When to re-optimise?
  - Event driven optimisation, e.g. on link or node failures
    - Won't re-optimise due to tunnel changes
  - Periodically
    - Tunnel churn if optimisation periodicity high
    - Inefficiencies if periodicity too low
    - Can be online or offline

Tunnel Sizing

- Online sizing: autobandwidth
  - Router automatically adjusts reservation (up or down) based on traffic observed in previous time interval:
    - Monitor avg tunnel utilization over a configurable period (5 min by default)
    - Track max avg tunnel utilization over successive periods
    - Readjust tunnel bandwidth to the highest recorded utilization over a longer configurable interval (60 minutes by default)
    - After interval has expired, the max avg. tunnel utilization counter is reset
  - Tunnel bandwidth is not persistent (lost on reload)

Tunnel Sizing

- Offline sizing
  - Statically set reservation to percentile (e.g. P95) of expected max load
  - Periodically readjust – not in real time, e.g. daily, weekly, monthly

Diffserv-aware TE (DS-TE)

- Introduces an additional bandwidth pool to allow separate constraint based routing and admission control for two distinct classes of traffic

  Traffic engineer EF and AF class capacity separately for most efficient use of bandwidth

  Constrain EF utilisation per link for tight VoIP SLA

- Not needed if only want to traffic engineer EF class

  can use conventional traffic engineering for EF class only and allow AF class to use IGP

- Otherwise considerations for DS-TE are the same

Agenda

I. IP TE Introduction

II. Traffic Characterization

III. Traffic Matrices

IV. Approaches for IP TE

V. High availability options

SLA Metrics: Availability

- Can be defined as network or service availability

  Network availability (connectivity)

    defined as the fraction of time the network is available between a specified ingress point and a specified egress point – IP connectivity

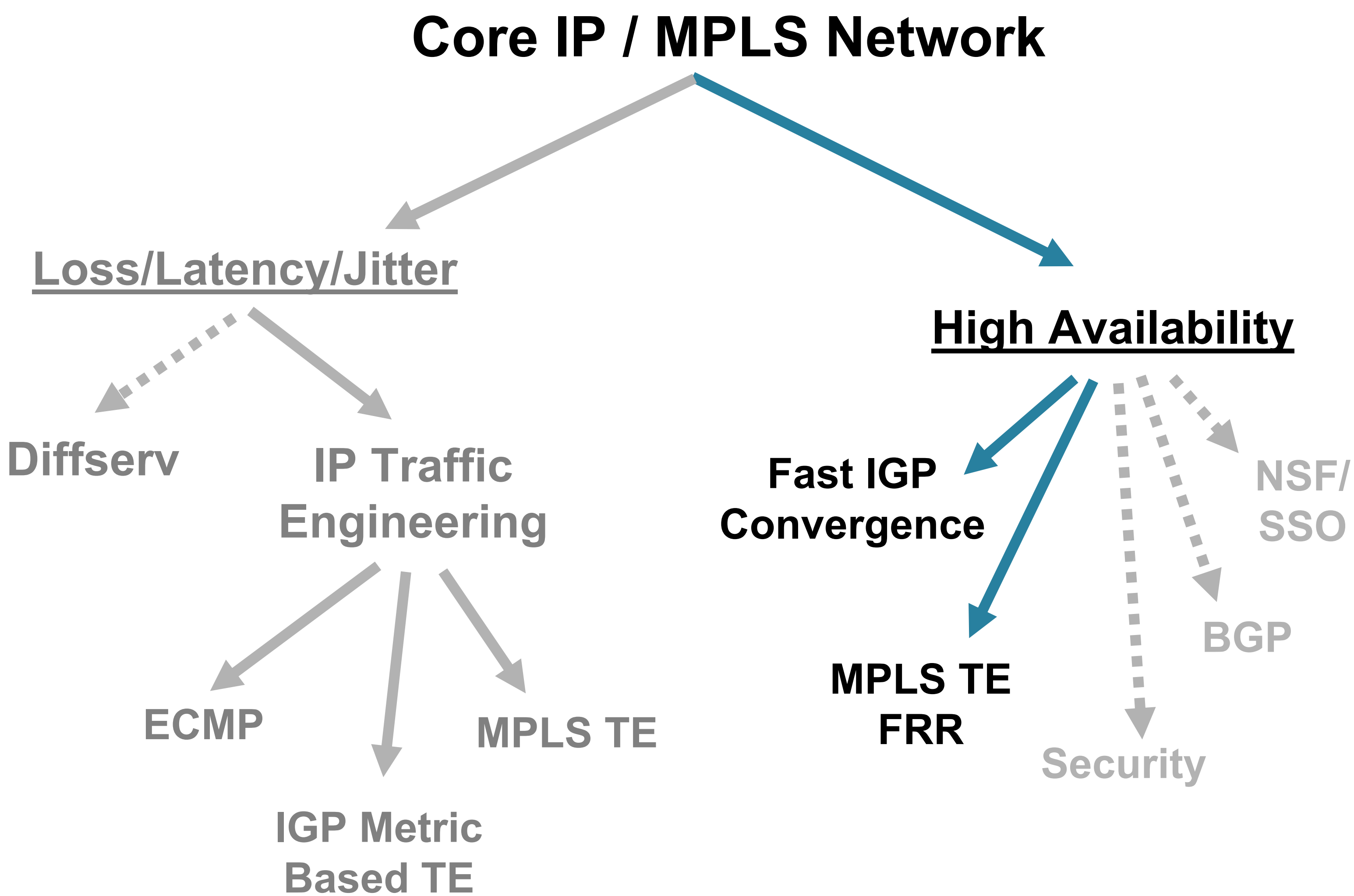
    RFC 2678 – IPPM Metrics for Measuring Connectivity

  Service availability

    Defined as the fraction of time the service is available within the bounds of the defined SLAs

- Service availability is most important to customers

High Availability Options



Agenda

Cisco.com Cariden.com

I. IP TE Introduction

II. Traffic Characterization

III. Traffic Matrices

IV. Approaches for IP TE

V. High availability options

- Fast IGP Convergence
- MPLS TE Fast Reroute (FRR)

IGP fast convergence

- Historical IGP convergence ~ O(10-30s)
  - Focus was on stability rather than fast convergence
- Optimisations to IGPs enable reduction in convergence to <1s for first 500 prefixes in a well designed backbone
  - with no compromise on network stability or scalability
  - where POS links are used - slower for non-POS
- Allows higher availability of service to be offered across all classes of traffic
- For more details see RIPE 47 Routing WG sessions from Clarence Filsfils, Henrik Villför, and Nicolas Dubois et al at

  http://www.ripe.net/ripe/meetings/ripe-47/presentations/index.html#routing

IGP fast convergence

- **IGP convergence time depends upon a number of factors**
  - **Propagation delay – distance from failure detecting node**
  - **Flooding delay – number of hops from failure detecting node to rerouting node**
  - **Number of nodes in the network**
  - **Number of prefixes**
  - **Position of prefixes in terms of order of processing**
- **Hence IGP convergence time is not deterministic**
  - **Difficult to define the worst-case bound for loss of connectivity**

Agenda

I. IP TE Introduction

II. Traffic Characterization

III. Traffic Matrices

IV. Approaches for IP TE

V. High availability options

- Fast IGP Convergence
- MPLS TE Fast Reroute (FRR)

MPLS TE Fast Reroute (FRR)

- **If …**

  recovery around failures is needed in <100ms

  or time to reroute around a failure needs to be more deterministic

- **Then …**

  MPLS TE fast reroute is required

- **MPLS TE FRR is faster and more deterministic than IGP convergence**

MPLS TE FRR link/node protection

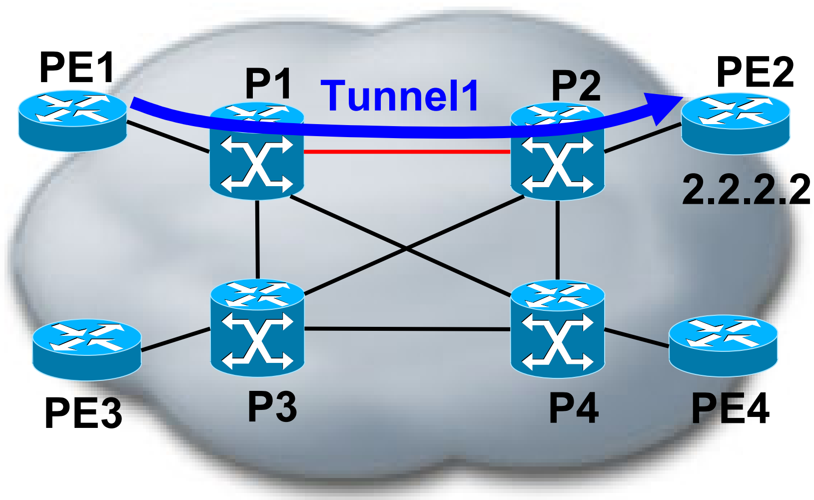
- **FRR uses local detection and protection at the point of failure**
  - POS provides most rapid failure detection
  - Fast local protection at the point of failure
  - No dependency on propagation, flooding etc
  - Uses a pre-established back-up tunnel to protect all appropriate tunnels on a link
    - Uses nested LSPs (stack of labels) – original LSP nested within link protection LSP
  - Switching entries pre-calculated before failure

MPLS TE FRR link protection

- **How to protect Tunnel1 against the failure of the red link?**

  LSP restoration will take a few seconds

- **Using Fast Re-Route (FRR) link protection can ensure restoration in <<1s**



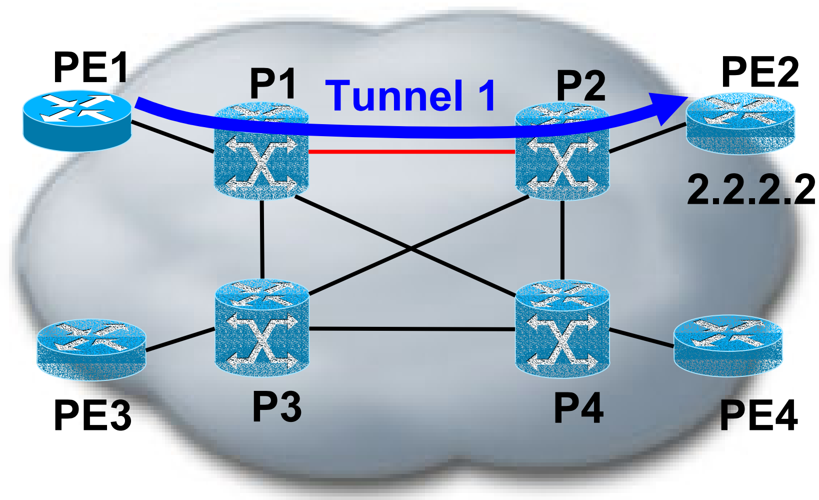
Resilience Strategy: two pronged approach

- **FRR allows for temporarily protecting LSPs affected by a link failure, while their head-end is reoptimizing**
  - Local detection and protection at POF
    - Uses a back-up tunnel to protect all appropriate tunnels on a link
      - Uses nested LSPs (stack of labels) – original LSP nested within link protection LSP
    - Fast—O (few 100s of milliseconds)
    - May be sub-optimal
  - Path restoration
    - Repair made at the head-end
    - An optimized long term repair
    - Slower—O (few seconds)

FRR Refresher (1)

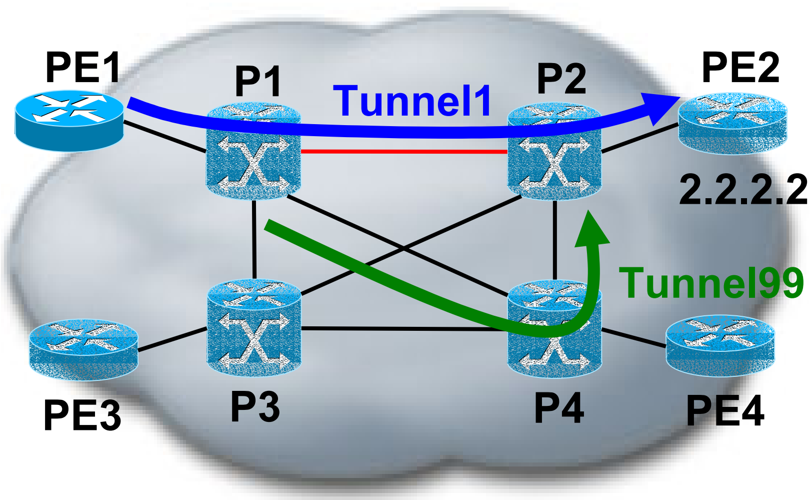
- **Tunnel1 is configured as fast reroutable on headend (PE1)**

  Session_Attribute's Flag = 0x01 in the path message




```
(config)# interface Tunnel1
(config-if)# description VOIP_TUNNEL
(config-if)# ip unnumbered Loopback0
(config-if)# tunnel destination 2.2.2.2
(config-if)# tunnel mode mpls traffic-eng
(config-if)# tunnel mpls traffic-eng priority 0 0
(config-if)# tunnel mpls traffic-eng bandwidth sub-pool 10000
(config-if)# tunnel mpls traffic-eng path-option 1 dynamic
(config-if)# tunnel mpls traffic-eng fast-reroute
```

FRR Refresher (2): Configuration



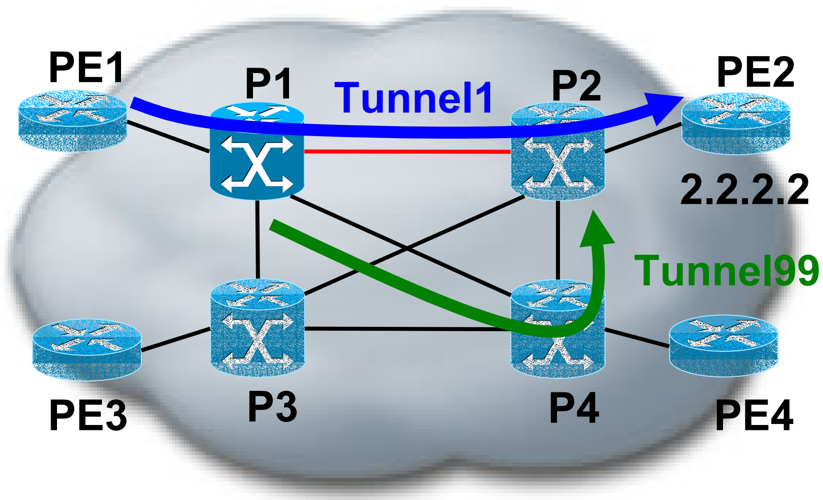
- **Explicitly routed back-up Tunnel99 is configured on P1 to P2 via P4**
- **No "tunnel mpls traffic-eng autoroute announce" !**

  The back-up tunnel MUST only be used when a failure occurs

```
(config)# interface Tunnel99
(config-if)# ip unnumbered Loopback0
(config-if)# tunnel destination 10.0.42.2
(config-if)# tunnel mode mpls traffic-eng
(config-if)# tunnel mpls traffic-eng priority 0 0
(config-if)# tunnel mpls traffic-eng bandwidth 10000
(config-if)# tunnel mpls traffic-eng path-option 1 explicit name tu99
(config-if)# exit
(config-cfg-ip-expl-path)# ip explicit-path name tu99 enable
(config-cfg-ip-expl-path)# next-address 10.0.14.4       ![P4]
(config-cfg-ip-expl-path)# next-address 10.0.42.2       ![P2]
```

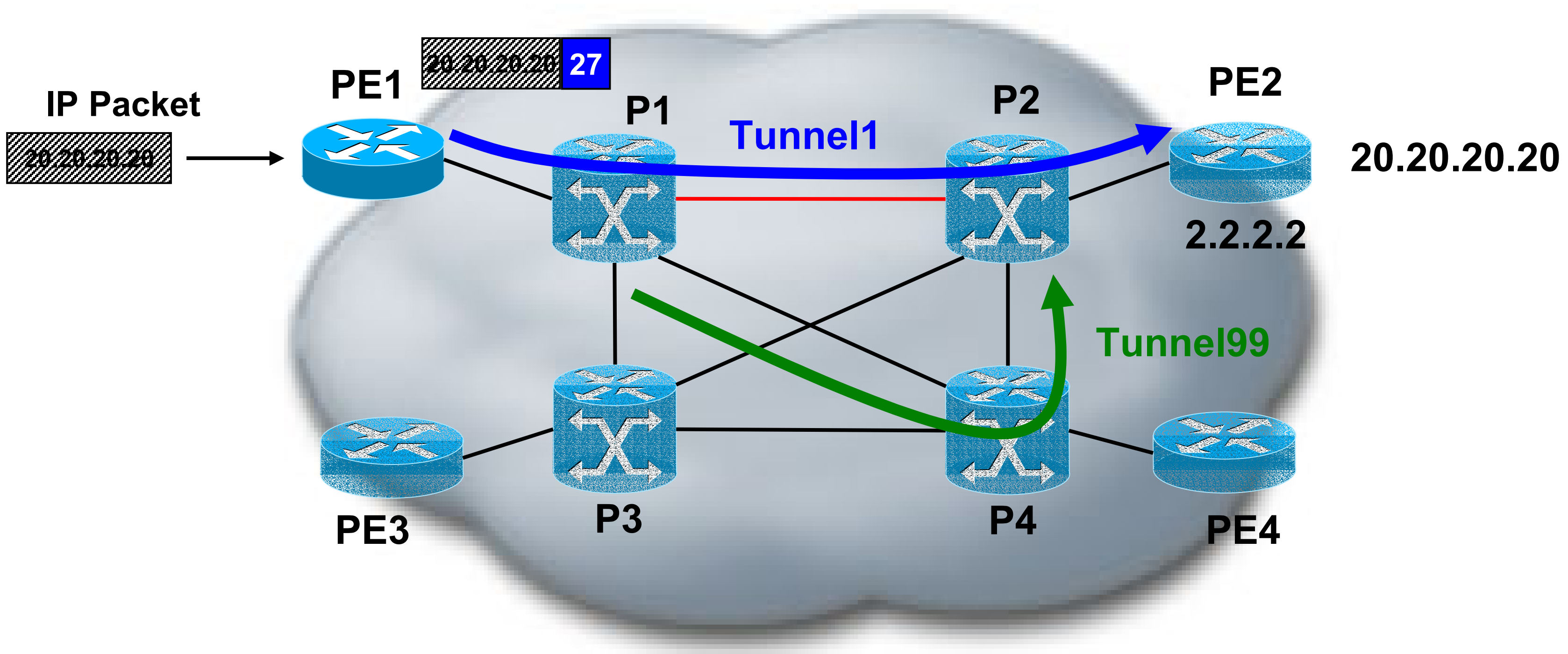
FRR Refresher (3): Configuration

- **On P1 configure Tunnel99 to backup valid tunnels on P1-P2 link**



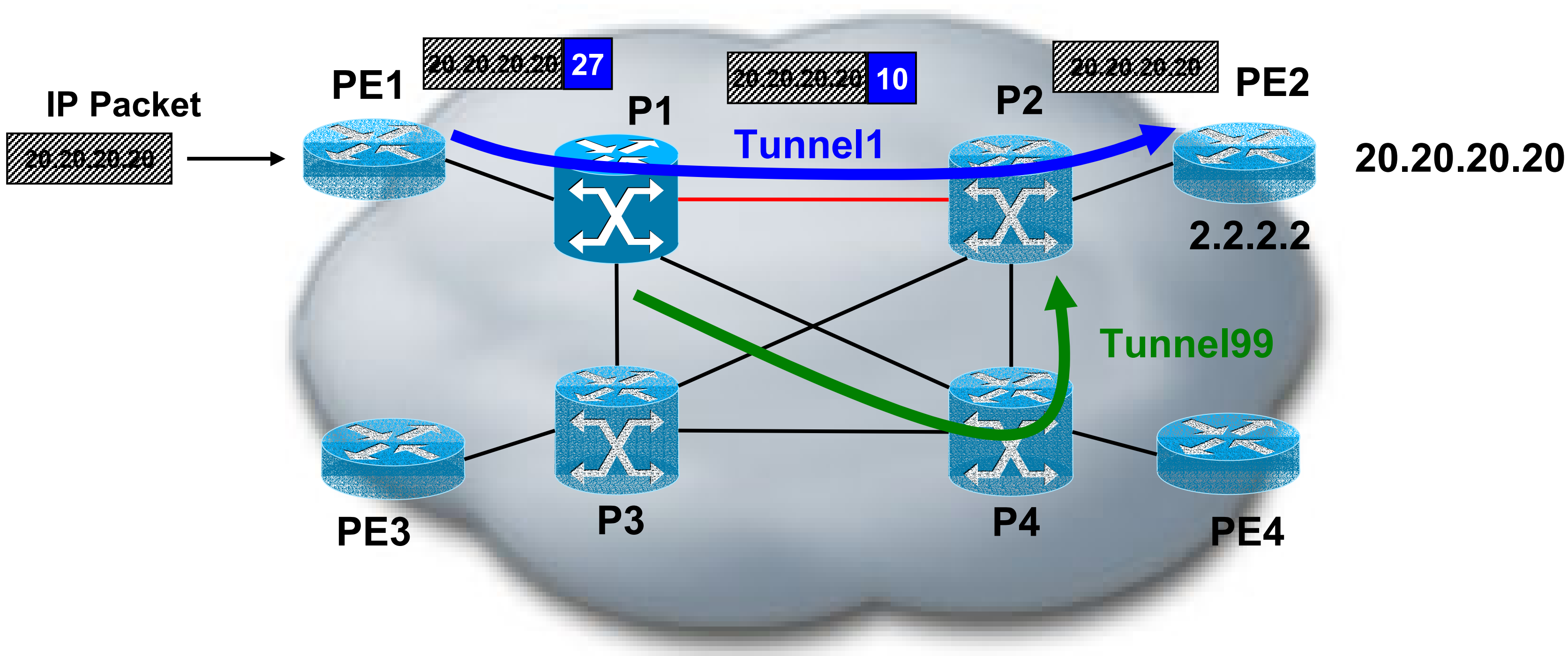

```
(config)# interface POS2/0
(config-if)# description Link to P2
(config-if)# ip address 10.0.12.2 255.255.255.252
(config-if)# mpls traffic-eng tunnels
(config-if)# ip rsvp bandwidth 150000 150000 sub-pool 30000
(config-if)# mpls traffic-eng backup-path Tunnel99
(config-if)# pos ais-shut
```

FRR Refresher (3): before failure



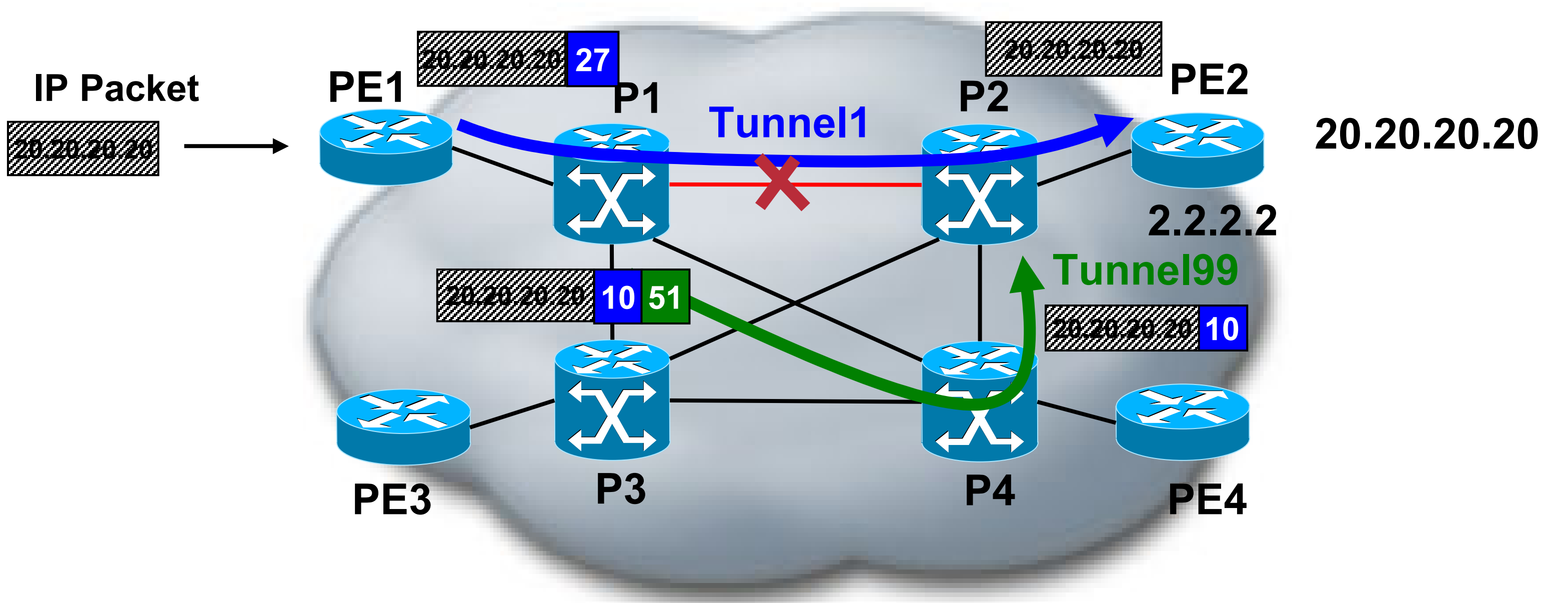

```
PE1# sh tag for 20.20.20.20
     Local  Outgoing    Prefix            Bytes tag  Outgoing    Next Hop
     tag    tag or VC   or Tunnel Id      switched   interface
     28     27          1.1.1.1/32        0          TU1         point2point
```

FRR Refresher (4): before failure




```
P1# sh tag for ...
    Local  Outgoing    Prefix              Bytes tag  Outgoing   Next hop
    tag    tag or VC   or Tunnel Id        switched   interface
    27     10          [T] 1.1.1.1/32      0          POS2/0     point2point
    [T]      Forwarding through a TSP tunnel.
```

FRR Refresher (5): after failure



t1. P1-P2 link fails

t2. Data plane: P1 will immediately swap 27 <-> 10 (as before) and pushes 51 (done for all protected LSPs)

t3. Control Plane registers a link-down event. RSVP PATH_ERR message sent

t4. P4 will do PHP

t5. P2 receives an identical labelled packet as before

Global label allocation

MPLS TE FRR

- Rapid local protection

  1. Link Failure Notification

     POS alarm detection in <10ms

  2. RP updates LFIB

     Replace a swap by a swap-push

  3. LFIB change notified to the linecards

     1 message covers all the entries that need modification

  4. LFIB rewrite

     In parallel – distributed on all the linecards

FRR – why do it?

- **VoIP impact of packet loss:**
  - **Most VoIP packet loss concealment algorithms can interpolate for the loss of 30-40ms of VoIP samples**
    - **Greater loss than this may produce an audible glitch**
  - **If the loss of connectivity lasts for several seconds (dependent on signalling), the phone call may be dropped**
- **FRR allows highest availability of service to be offered**
  - **For voip – reduces possibility audible glitches and prevents calls being dropped due to network failures**

MPLS TE FRR – deployment scenarios

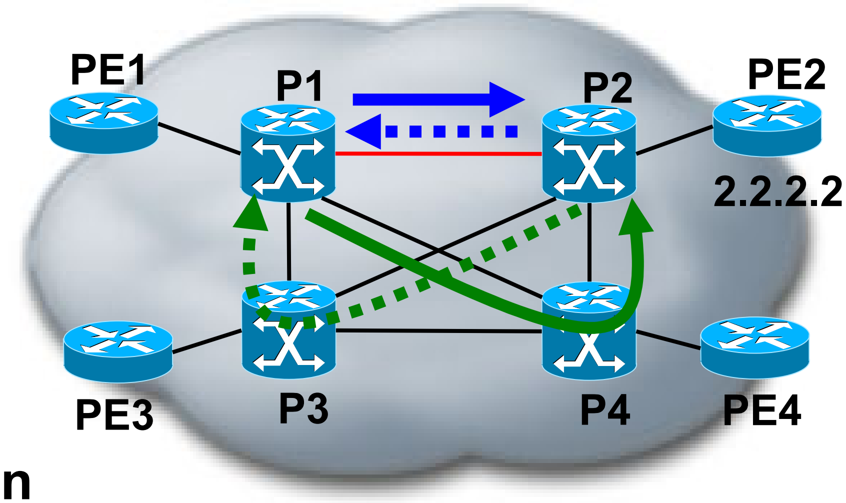
MPLS TE FRR

<u>Systematic</u>:
Deployed to provide complete protection for the failure of every link and/or node

<u>Ad hoc</u>:
Deployed only to protect key components whose failures will have a severe impact on services

MPLS TE FRR – deployment scenarios

- Full mesh of TE tunnels is not needed for systematic approach
- Can instead use 1-hop primary tunnels on every link



  1-hop zero bandwidth tunnel on every link in each direction

  Run autoroute on every tunnel

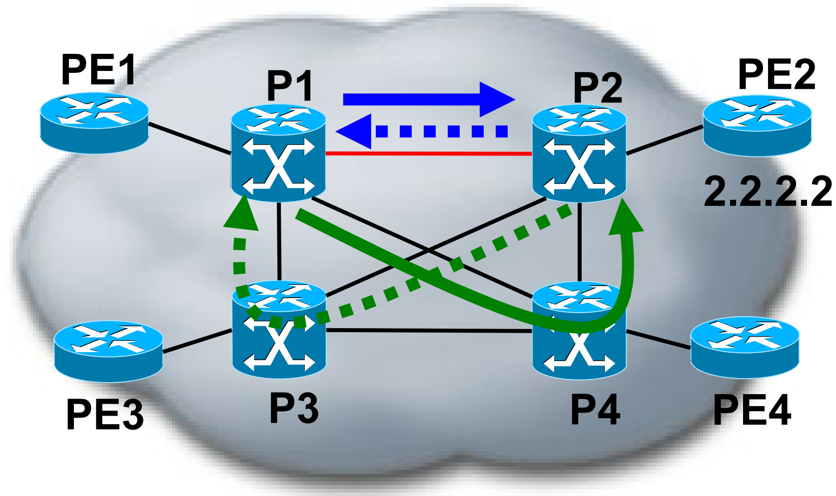
  As tunnels are 1 hop, due to penultimate hop popping, in normal operation:

    no labels are imposed

    packets are not label switched

    traffic follows the IGP shortest path

MPLS TE FRR – deployment scenarios



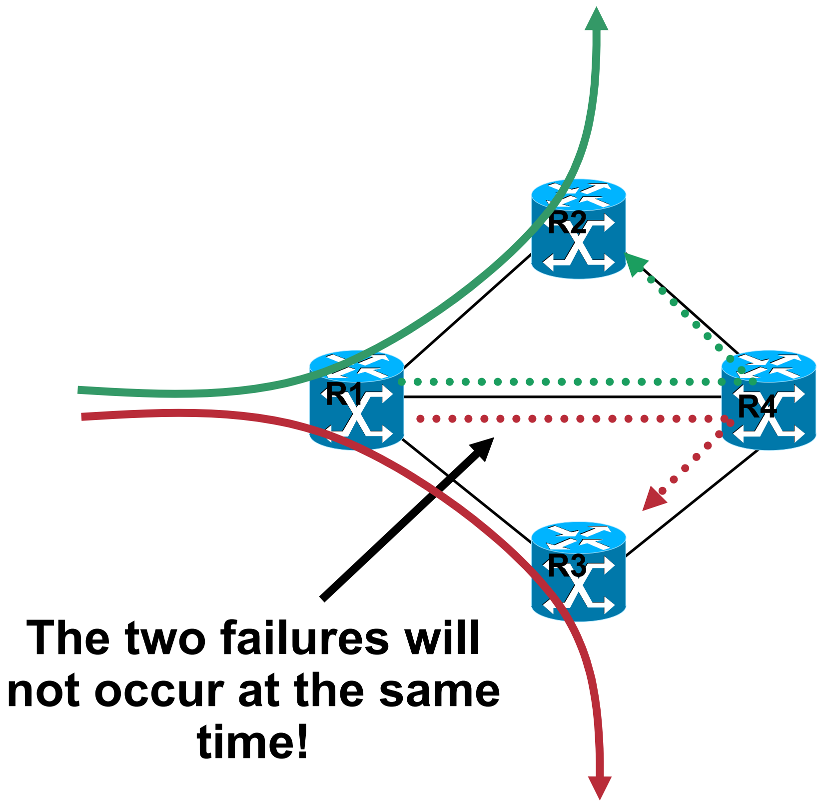
- Allows FRR to be used for link protection without needing a TE full mesh
  - ➔ Backup tunnel protecting every link
  - ➔ Recovery time becomes a function of number of LSPs / prefixes
- Can similarly use 2-hop tunnels to protect every node

  Can use Autotunnel to simplify provisioning of both 1-hop and 2-hop tunnels
- Allows decisions on need for TE for bandwidth optimisation and high availability to be independent

MPLS TE FRR – bandwidth protection

- **Backup tunnels can be configured with non-zero or zero bandwidth**
- **Zero bandwidth backup tunnels provide more efficient use of resources**

  Assuming single element failures



MPLS TE FRR – bandwidth protection

- **With zero bandwidth tunnels some local congestion might occur during rerouting**

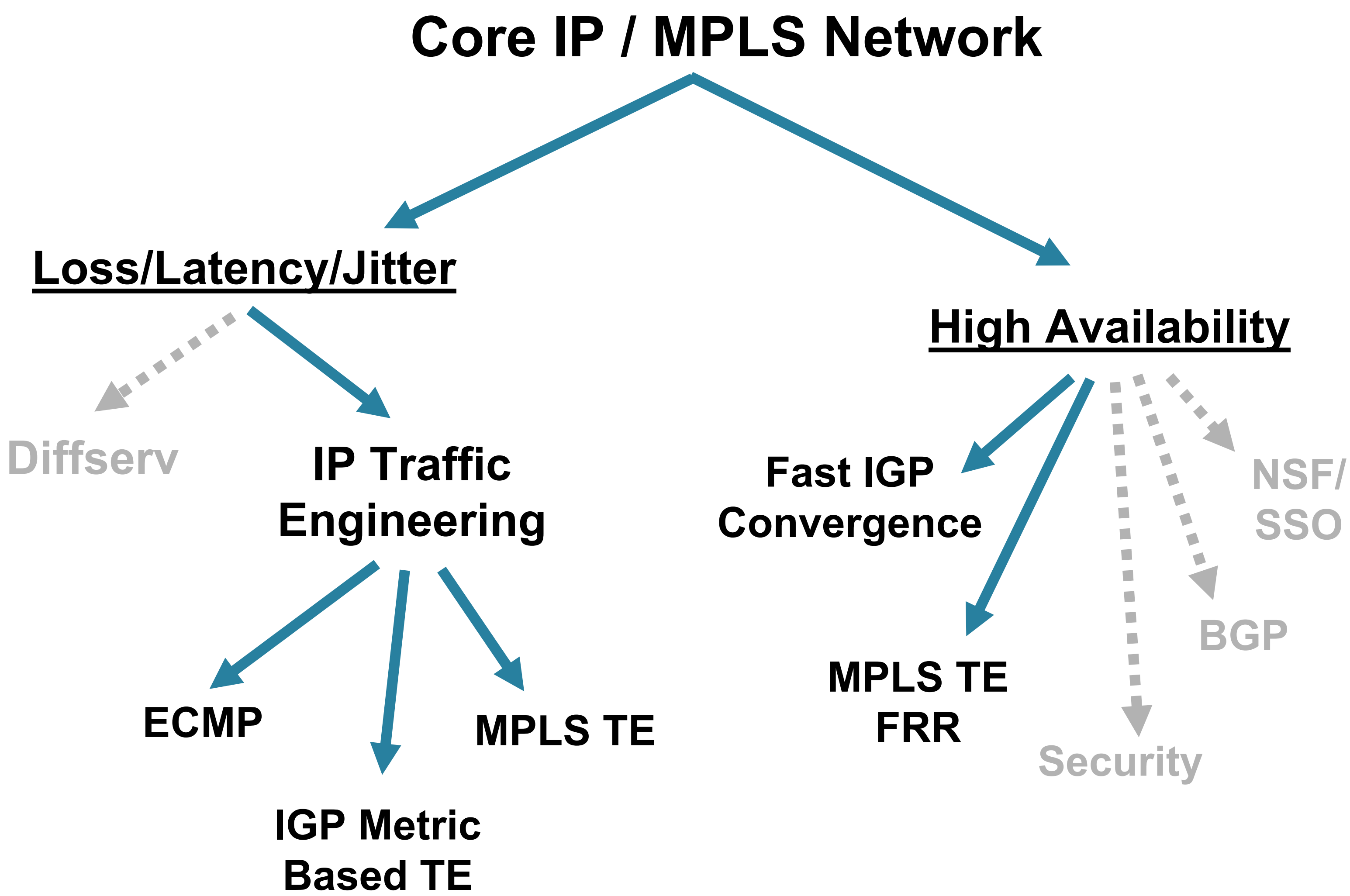
  Conflict between resource efficiency and tight SLA guarantees

  - ➔ **Use Diffserv to mitigate this short-term congestion**
  - ➔ **Use LSP reoptimization to handle the long-term congestion**

- **Simulation/modelling tools such as Tunnel Builder Pro may be useful to figure out more optimal configurations under different link/node failure scenarios**

Summary

IP Traffic engineering and core network availability



IP Traffic Engineering

- **Number of options are available for IP bandwidth optimisation, a.k.a. traffic engineering**
  - **With symmetrical topology and flows ➔ ECMP may be good enough**
  - **With asymmetrical topology or flows ➔ IGP metric optimisation may provide an acceptable solution**
  - **MPLS TE can provide a solution when neither of the above is acceptable**
- **Essential to decide primary TE objective: to optimise for working (normal) case or for single element failure case?**
- **Holistic view of traffic demands (core traffic matrix) and routing (in failures if necessary) is essential to understand benefits of each option, and behaviour of different deployment models**

Core network availability

- **Number of technologies to increase core network availability**
  - **IGP fast convergence**
    - **Where recovery in < ~1-2s is acceptable**
  - **MPLS TE FRR**
    - **Where faster recovery or more determinism is required**
- **Could adopt a hybrid approach**
  - **MPLS TE FRR – to protect key resources or services such as VoIP**
  - **Fast IGP convergence – for everything else**

Summary

- **Decisions on which technologies to deploy for traffic engineering and core network availability can be orthogonal**
- **Important to do your own analysis before deciding on best approach**
  - **Overkill is unnecessary by definition**

Any Questions ?

Cisco.com