Traffic Engineering Beyond MPLS

Apricot 2004 Tutorial
February 24, 2004
Kuala Lumpur, Malaysia

Arman Maghbouleh
Cariden Technologies, Inc.
arman @ cariden.com

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

Carrier IP Backbone Engineering Models

Simple

- Emphasis on Scalability
- Low Overhead Protocols
  - Pure IP
  - No CoS
  - 50% Upgrade

Dynamic

- Emphasis on Smart Network
- Service-Aware Protocols
  - MPLS CSPF
  - Diffserv/–TE

Controlled

- Emphasis on Asset Utilization
- Optimize Offline
  - Static Explicit MPLS/ATM PVC

Simple++

- Pure IP for scalability
- Capacity Planning/TE for QoS (CoS for insurance)
- Metric-Based Offline TE for Control

Goals

- Investigate Assumptions Behind Models
  - Dynamic
    - Internet traffic is highly variable and bursty.
  - Simple
    - Capital expenditures not significant.
  - Controlled
    - Shortest path first protocols do not provide enough levers of control.
  - Simple++
    - Smart Network Engineering vs. Smart Networks

- Demonstrate Simple++

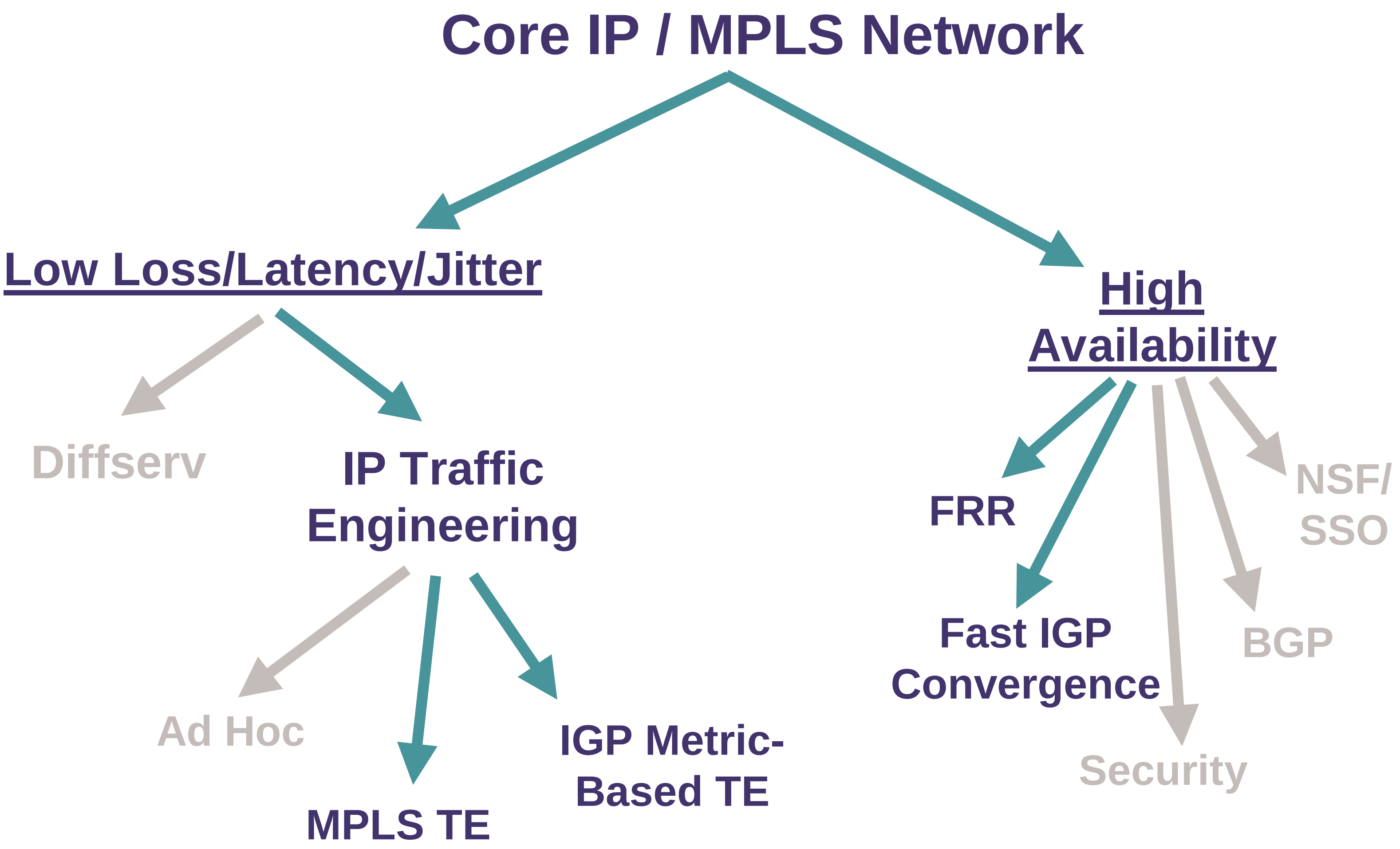
Summary

- Traffic Characteristics
  - Long term is smooth and predictable
  - Uncorrelated microbursts
  - High utilization with little delay at high capacities
  - Little need for dynamic routing or queue management
- Simple++
  - Traffic Matrix (Measure, or Estimate)
  - Capacity plan based on failure simulation
  - TE without Layer 2 Overlay
    - Computer-Aided Metric-Based TE ≈ as Efficient of Theoretical Optimum (though more scalable)
- Multiple Routes to High Availability
  - Fast Reroute
  - Fast Convergence

MPLS TE Aspects

- Covered Here
  - Efficient Use of Assets
  - QoS
  - Fast Reroute
- Not Covered Here
  (less backbone relevance)
  - Admission Control
  - Route Pinning

What is Covered



Core IP / MPLS Network

Low Loss/Latency/Jitter

Diffserv

IP Traffic Engineering

Ad Hoc

MPLS TE

IGP Metric-Based TE

High Availability

FRR

Fast IGP Convergence

Security

BGP

NSF/SSO

Agenda

I. Traffic Characterization

II. Traffic Matrices

III. TE Introduction

IV. Metric-Based TE

V. Convergence

Traffic Characterization

I. Traffic Characterization

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

II. Traffic Matrices

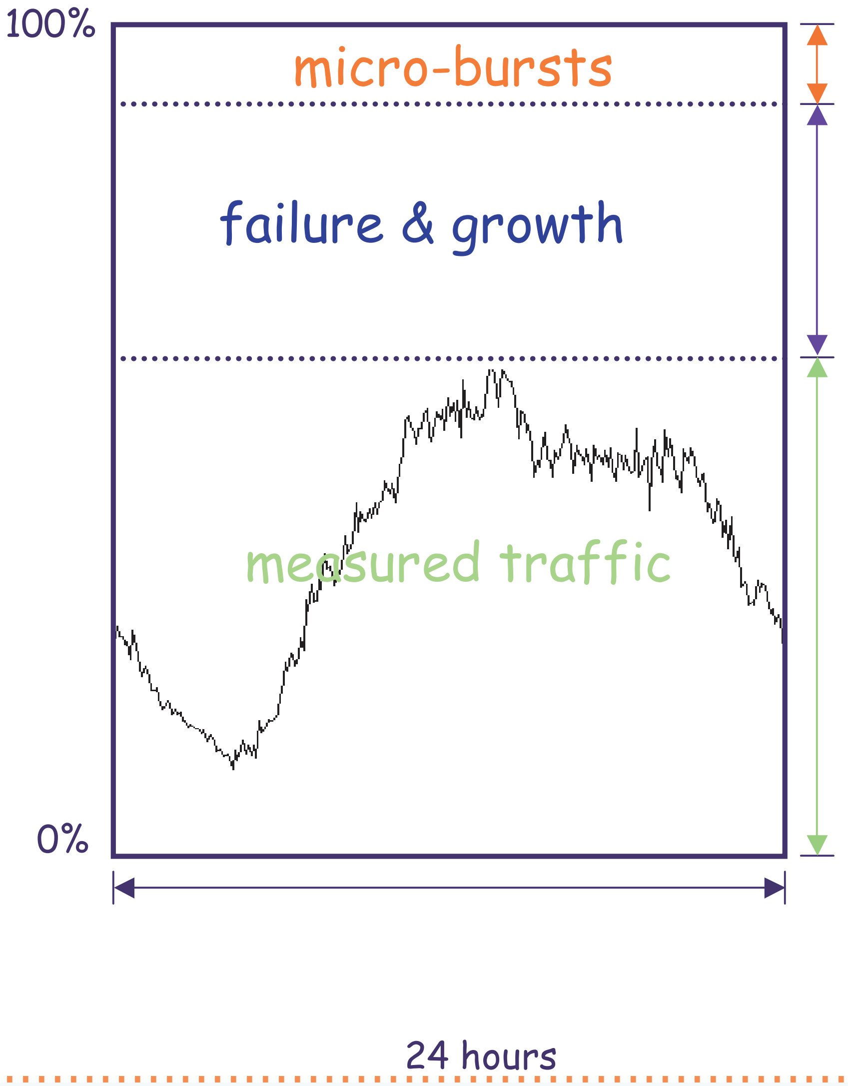
III. TE Introduction

IV. Metric-Based TE

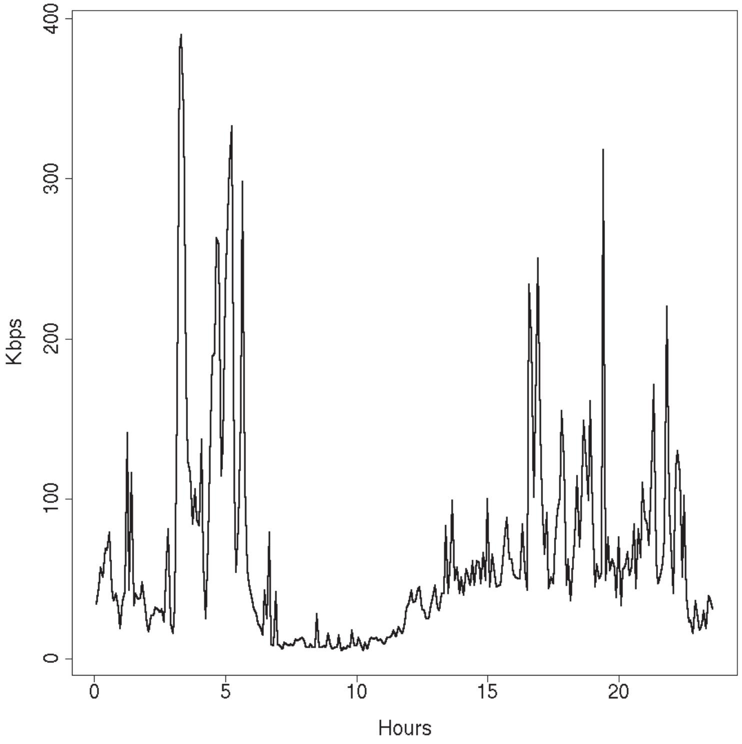
V. Convergence

Traffic Characterization

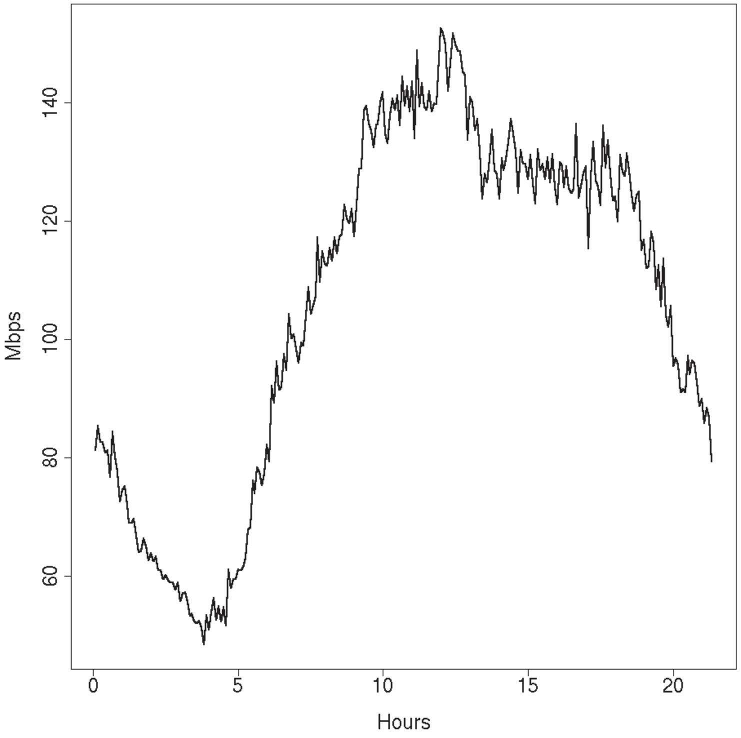
- Long-Term
  - Measured Traffic
    - E.g. P95 (day/week)
  - Accommodate failure and growth
- Short-Term
  - Critical scale for queuing
  - Determine over-provisioning factor that will prevent queue buildup against micro-bursts



High- vs. Low-Bandwidth Demands



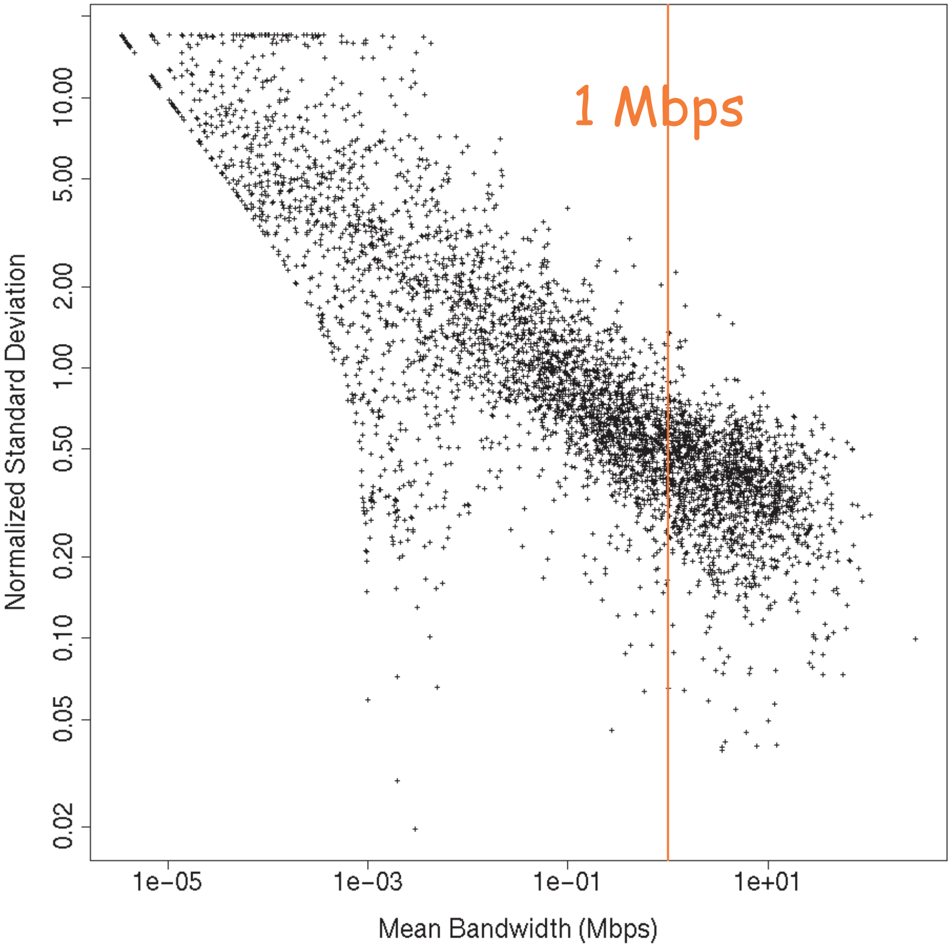
Cleveland -> Denver
Mean=64Kbps, Max=380Kbps
P95=201Kbps, Std. dev.=66Kbps



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 [5]
- High-bandwidth demands seem well-behaved
- 97% of traffic is carried by the demands larger than 1 Mbps
  (20% of the demands!)



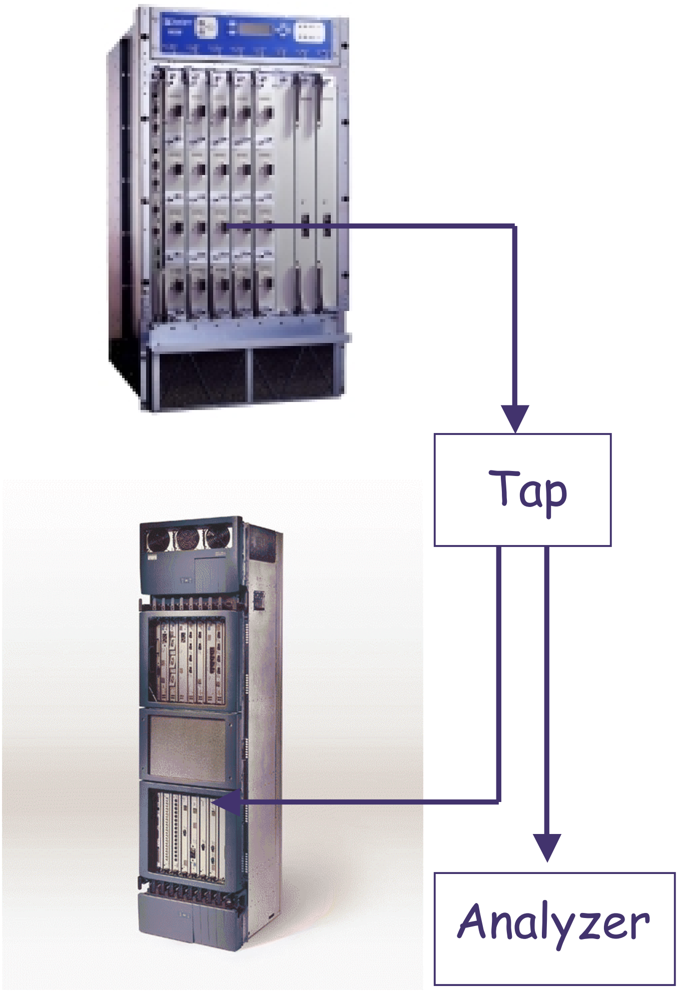
Long Term Traffic Summary

- 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

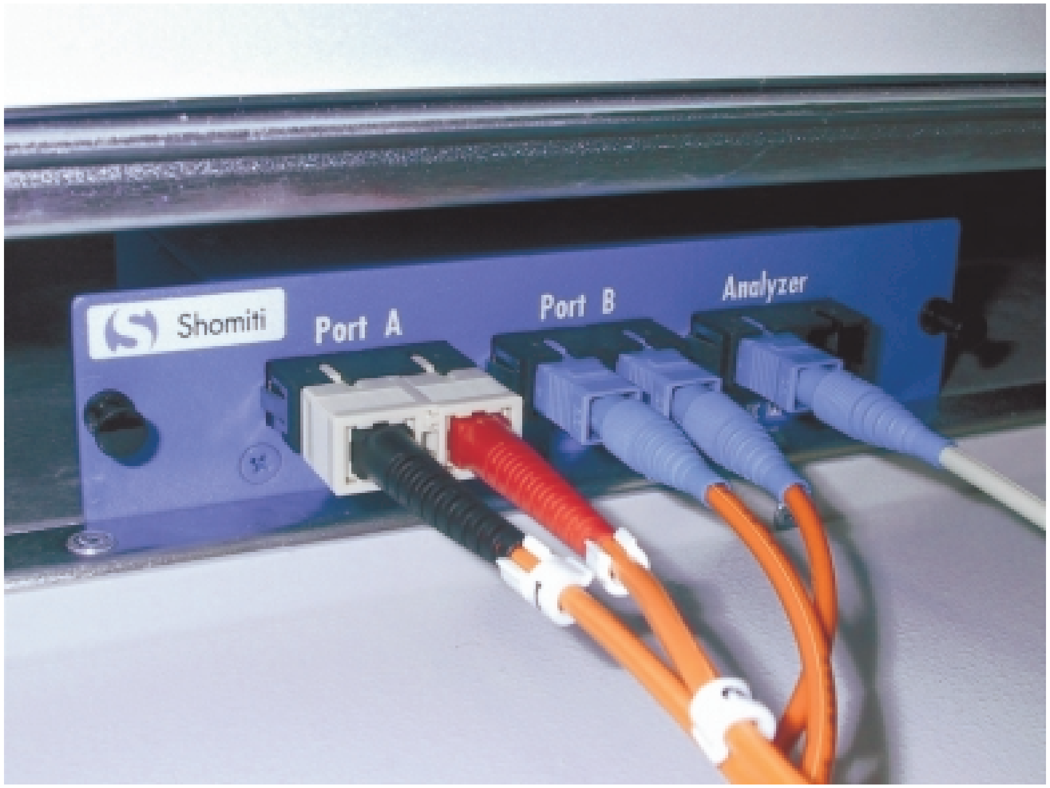
- Investigate burstiness within 5-min intervals
- Critical timescale for queuing, like 1ms or 5ms
- Analyze statistical properties
- Only at specific locations
  - Complex setup
  - A lot of data

Fiber Tap (Gigabit Ethernet)



Tap

Analyzer

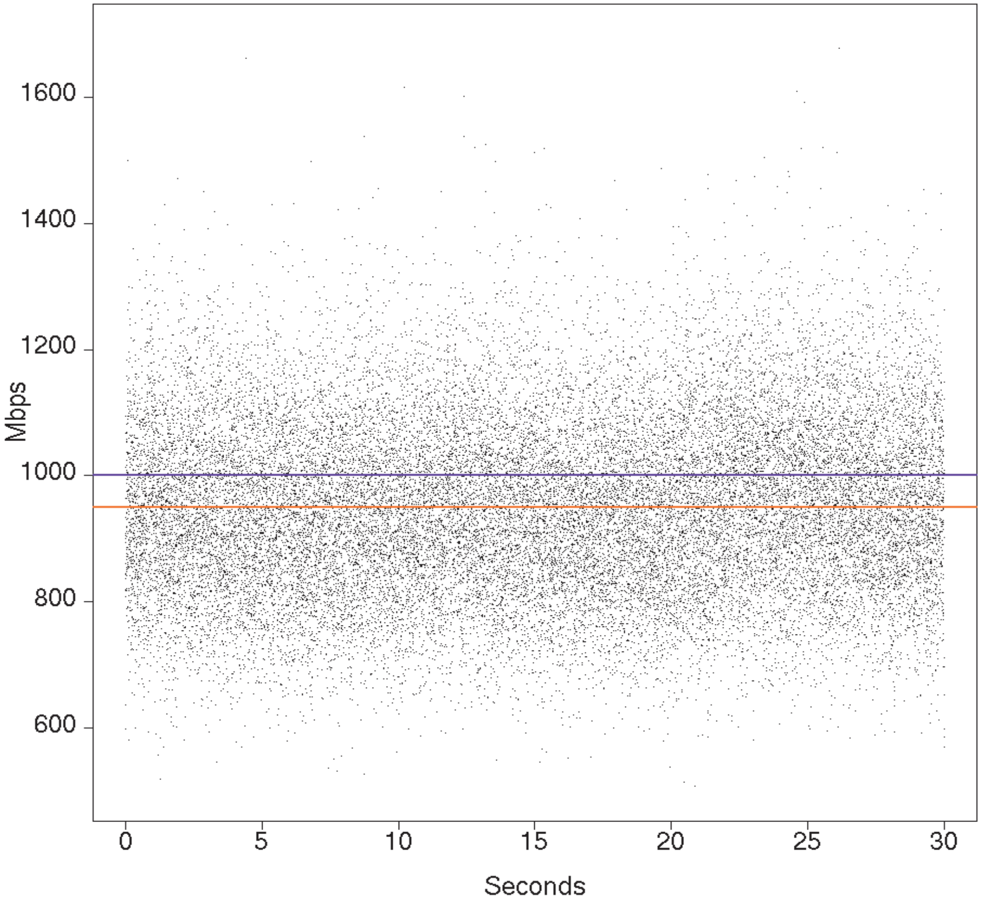


Raw Results
30 sec of data, 1ms scale

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

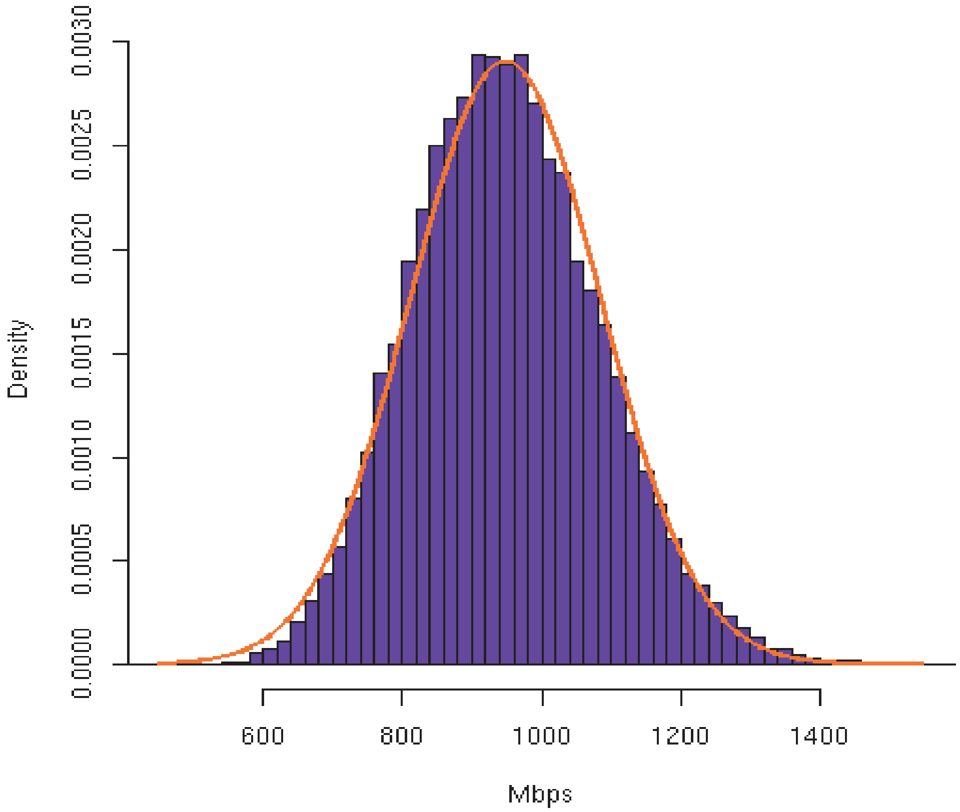
- 95-percentile: 1183 Mbps
- 5-percentile: 737 Mbps

- (around 250 packets per 1ms interval)



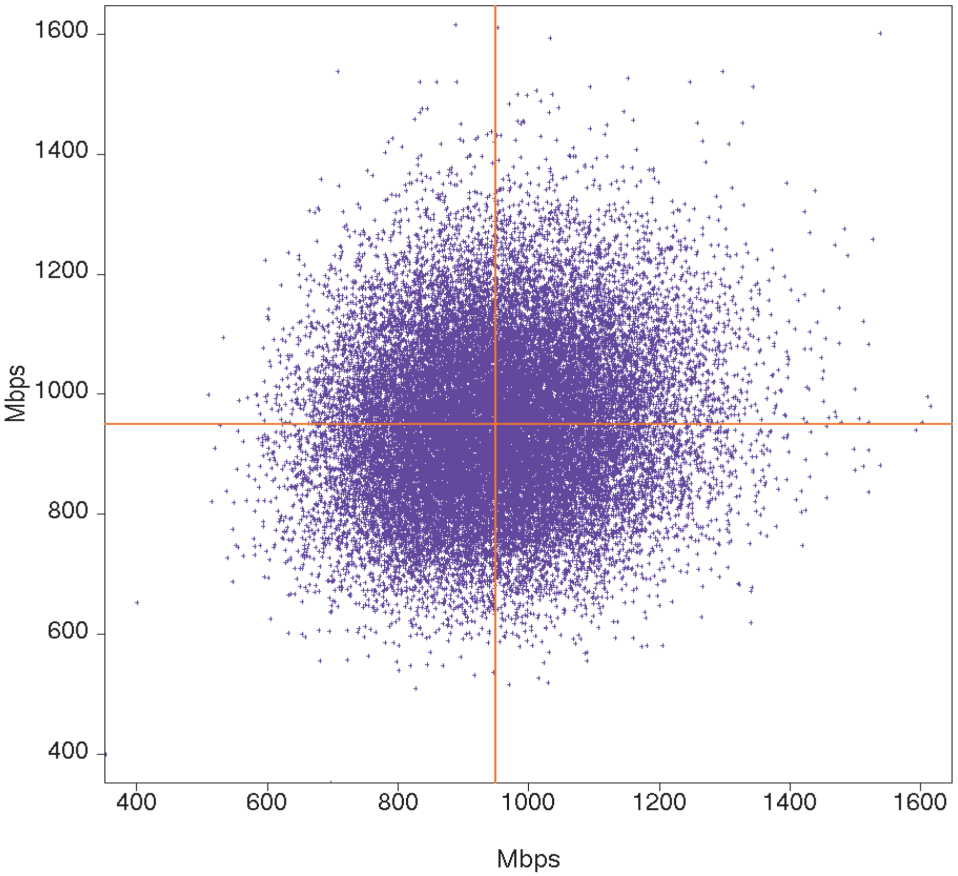
Traffic Distribution Histogram (1ms scale)

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



Autocorrelation, Lag Plot (1ms scale)

- 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

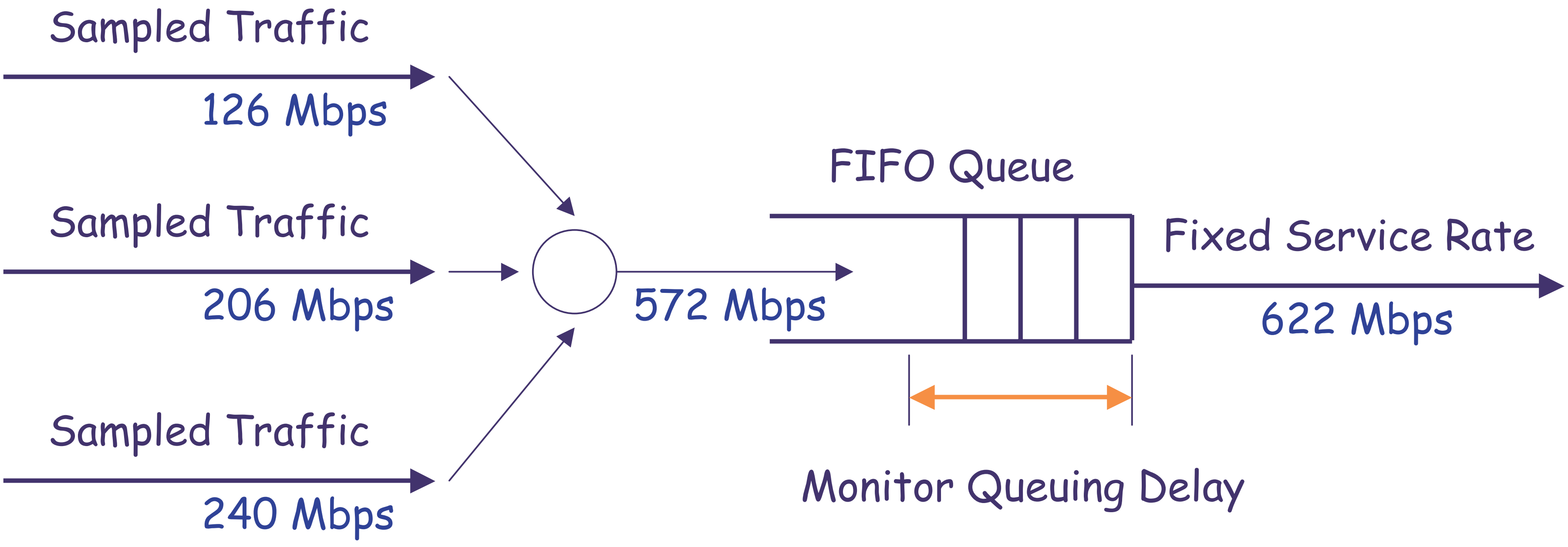
- **M/M/1**
- Markovian
  - Poisson-process
  - Infinite number of sources
- "Circuits can be operated at over 99% utilization, with delay and jitter well below 1ms" [2] [3]

- **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



Sampled Traffic

126 Mbps

Sampled Traffic

206 Mbps

Sampled Traffic

240 Mbps

572 Mbps

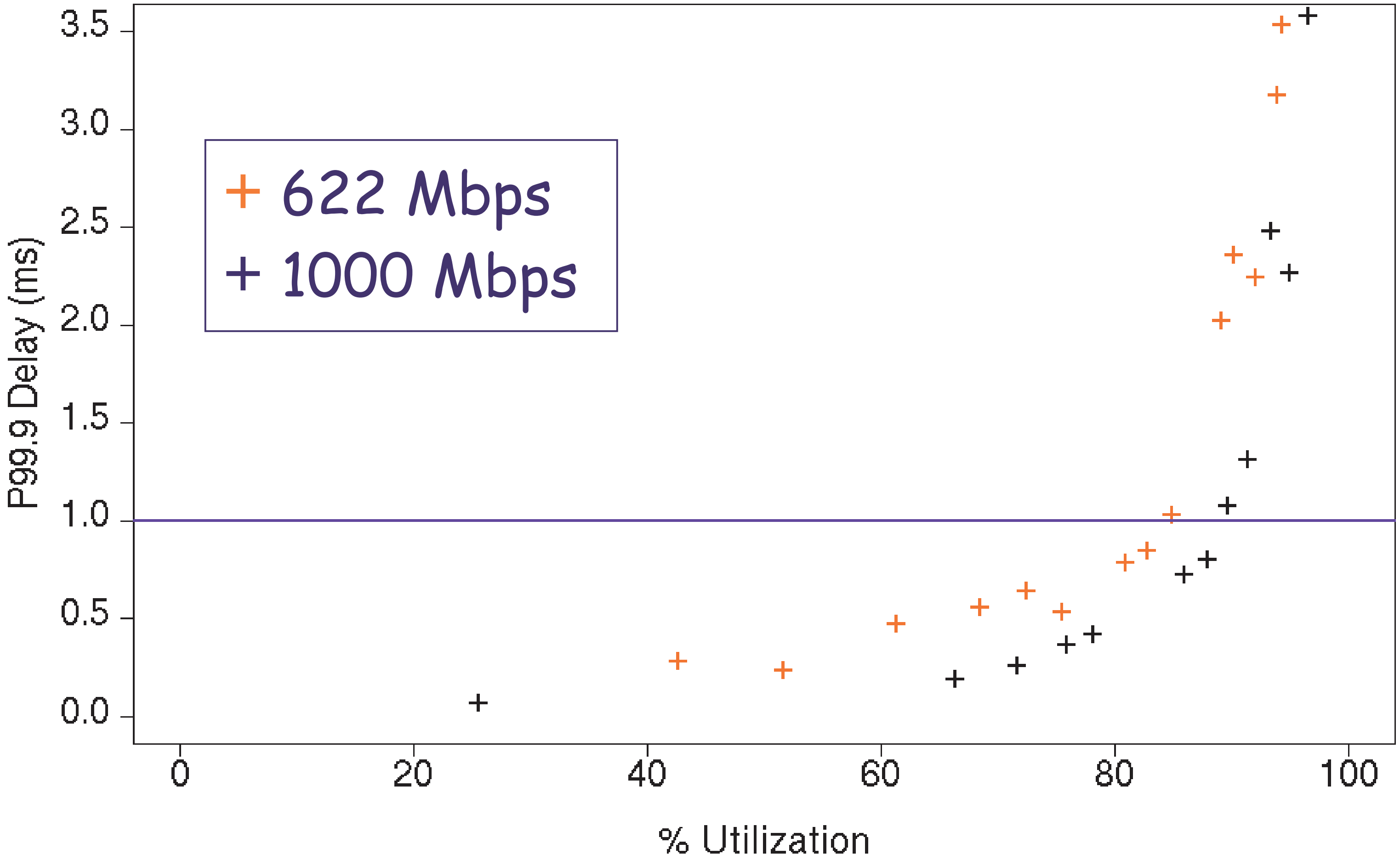
FIFO Queue

Fixed Service Rate

622 Mbps

Monitor Queuing Delay

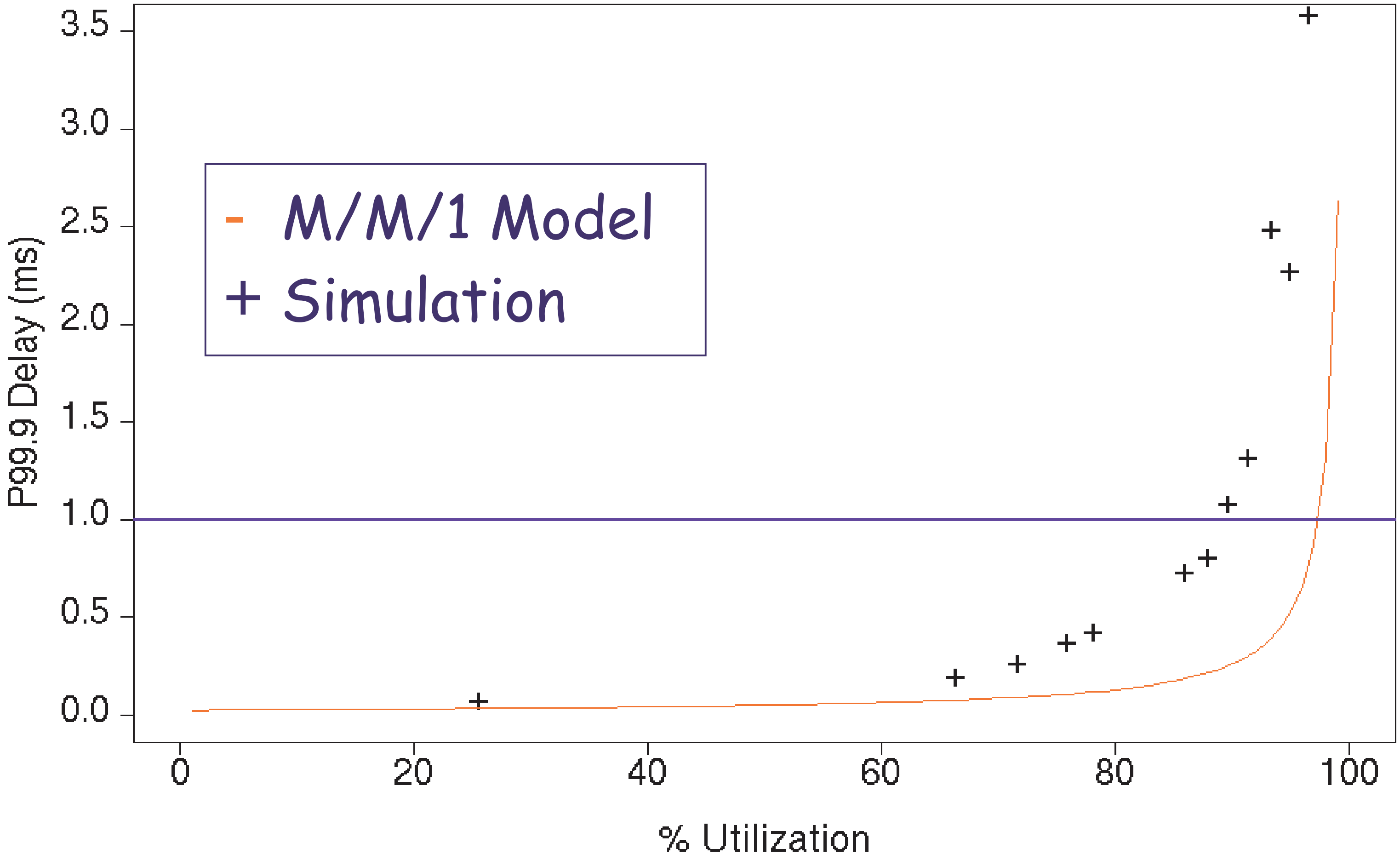
Queuing Simulation: Results



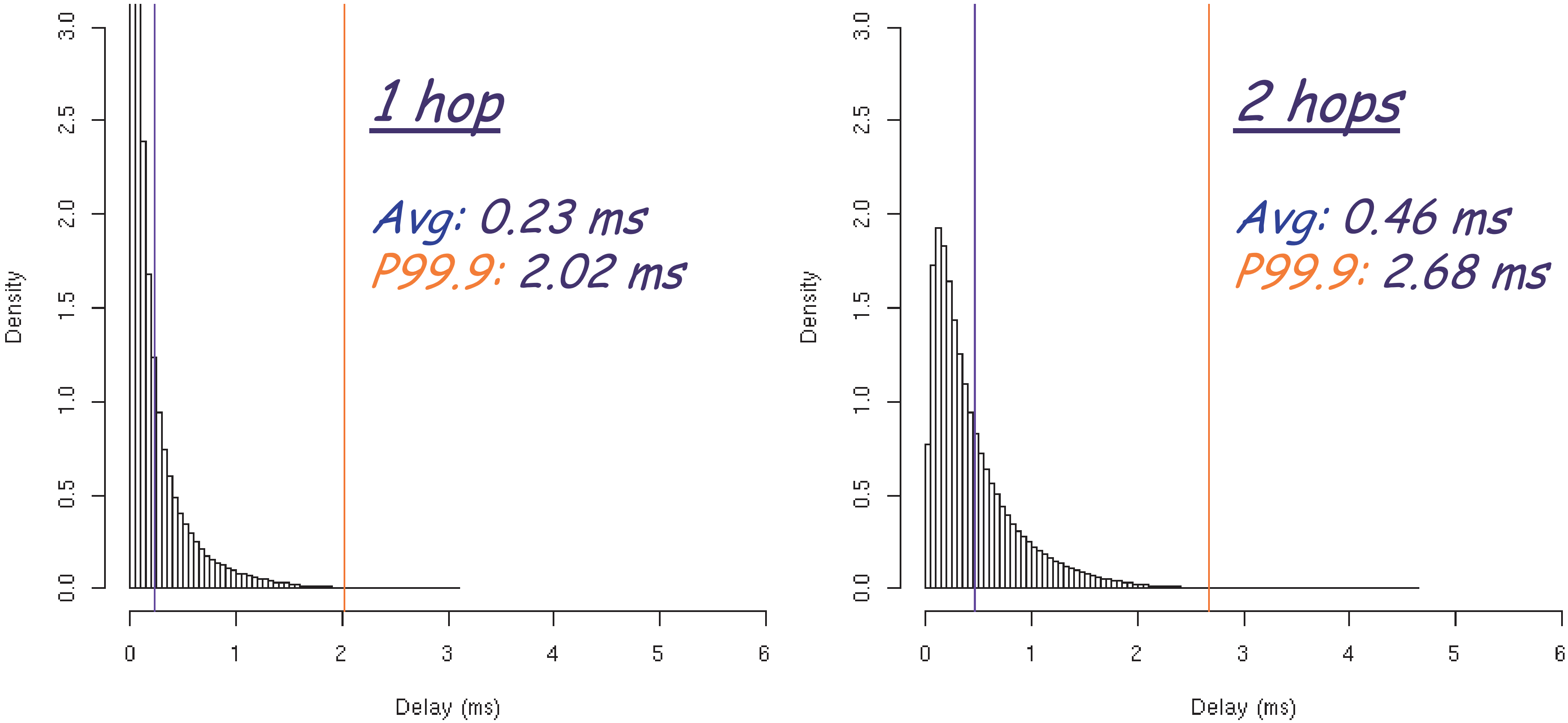
Queuing Simulation Results

- 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 Queueing



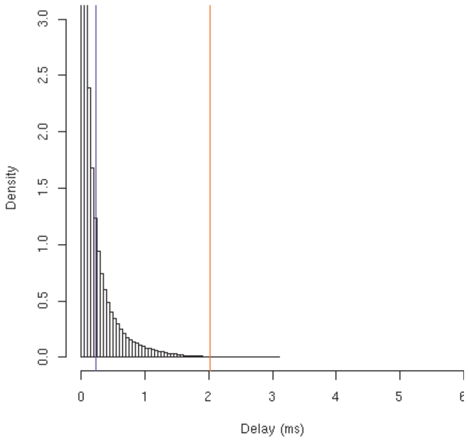
1 hop

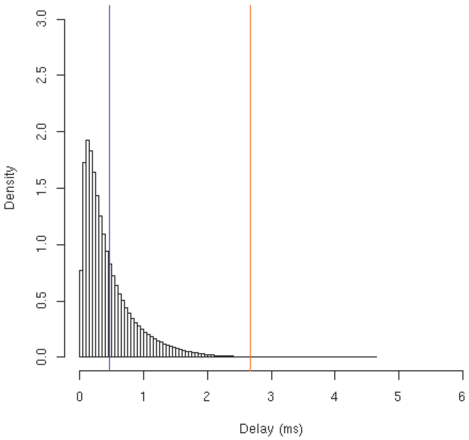
Avg: 0.23 ms
P99.9: 2.02 ms

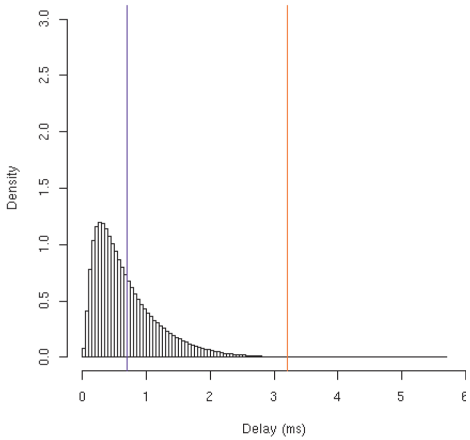
2 hops

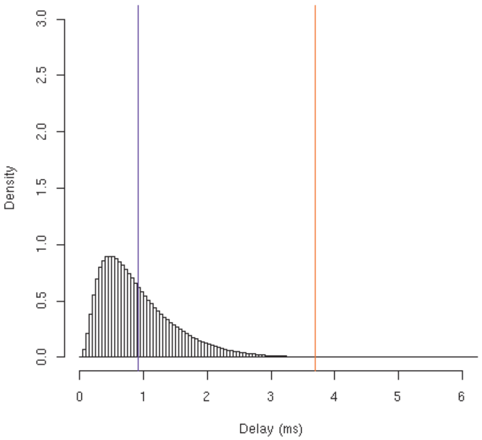
Avg: 0.46 ms
P99.9: 2.68 ms

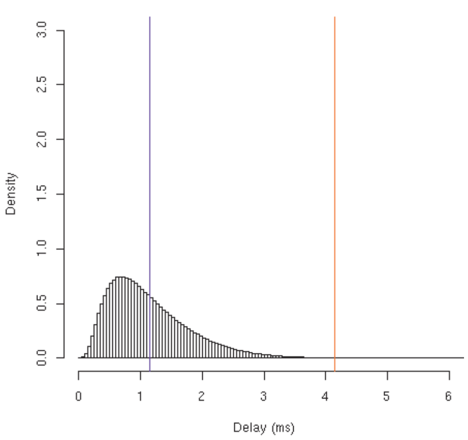
Multi-hop Queueing (1-8 hops)

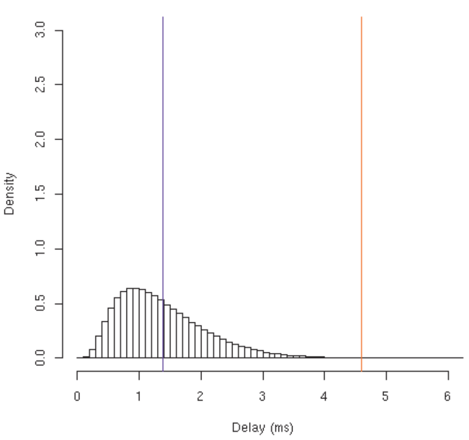


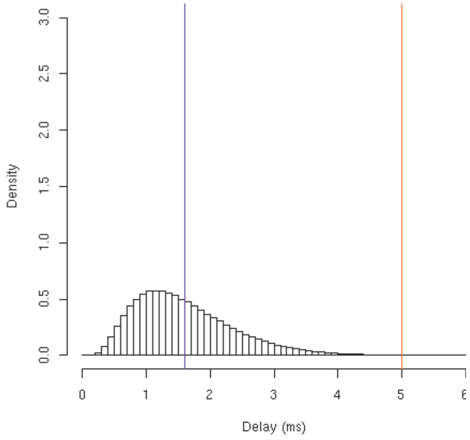


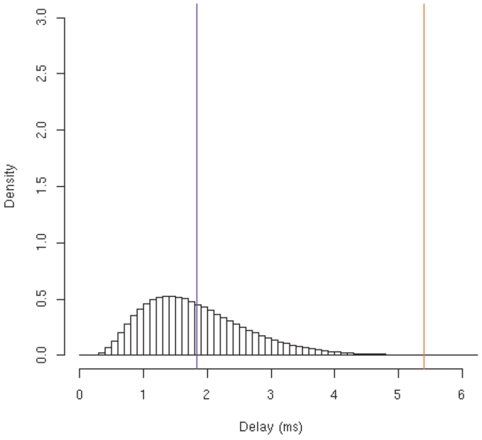












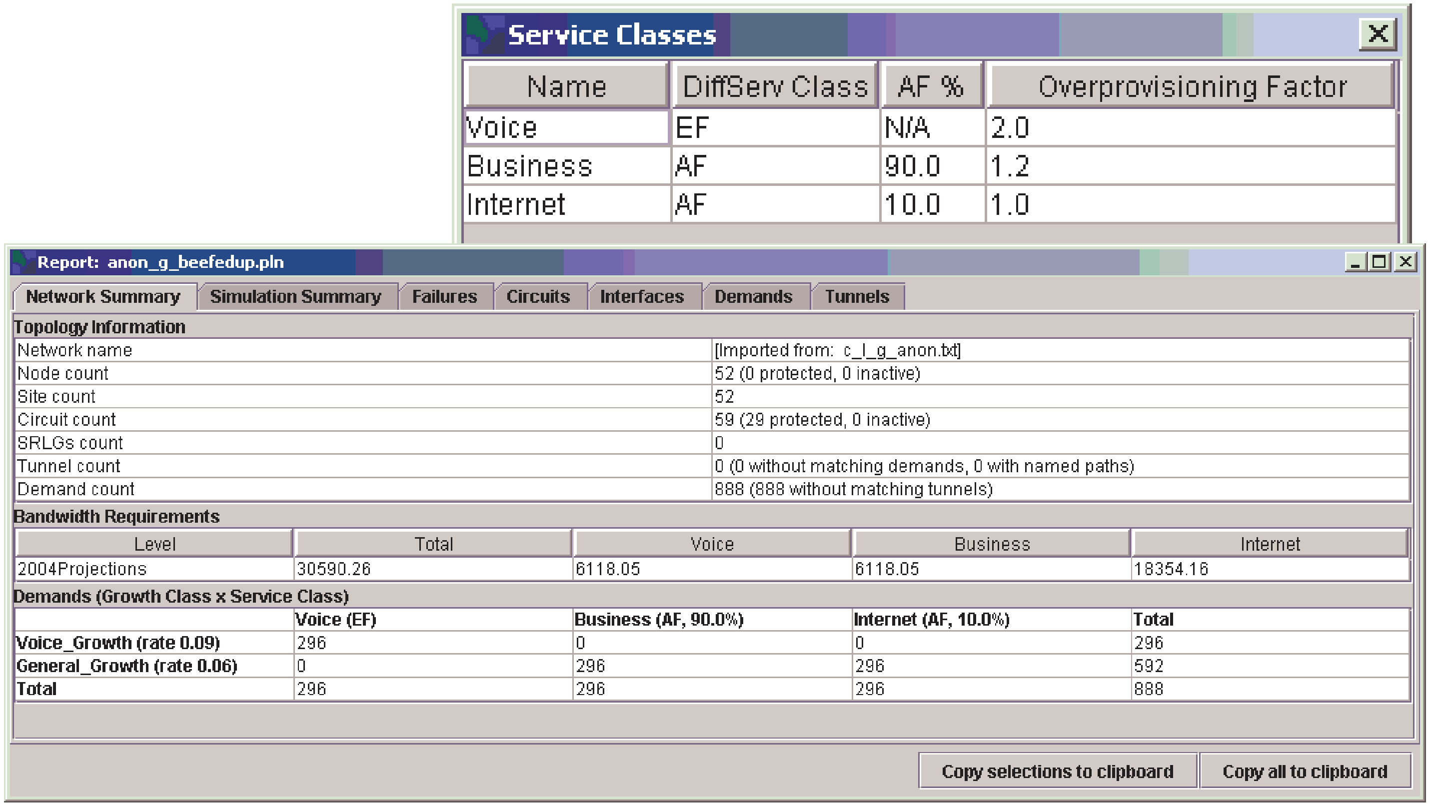
Queueing: Summary

- Queueing 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

Role of Backbone CoS

- Insurance for Issues Beyond Planning
  - Denial of Service Attacks
  - Catastrophic Failure (e.g., earthquake, terrorist attack)
- Traffic Separation Under Massive Load
  - Coarse-grained service types
  - ATM-style queue management not necessary with high speed links
- (See example in the demo section)

COS Example



Name	DiffServ Class	AF %	Overprovisioning Factor
Voice	EF	N/A	2.0
Business	AF	90.0	1.2
Internet	AF	10.0	1.0

Topology Information

Network name	[Imported from: c_l_g_anon.txt]
Node count	52 (0 protected, 0 inactive)
Site count	52
Circuit count	59 (29 protected, 0 inactive)
SRLGs count	0
Tunnel count	0 (0 without matching demands, 0 with named paths)
Demand count	888 (888 without matching tunnels)

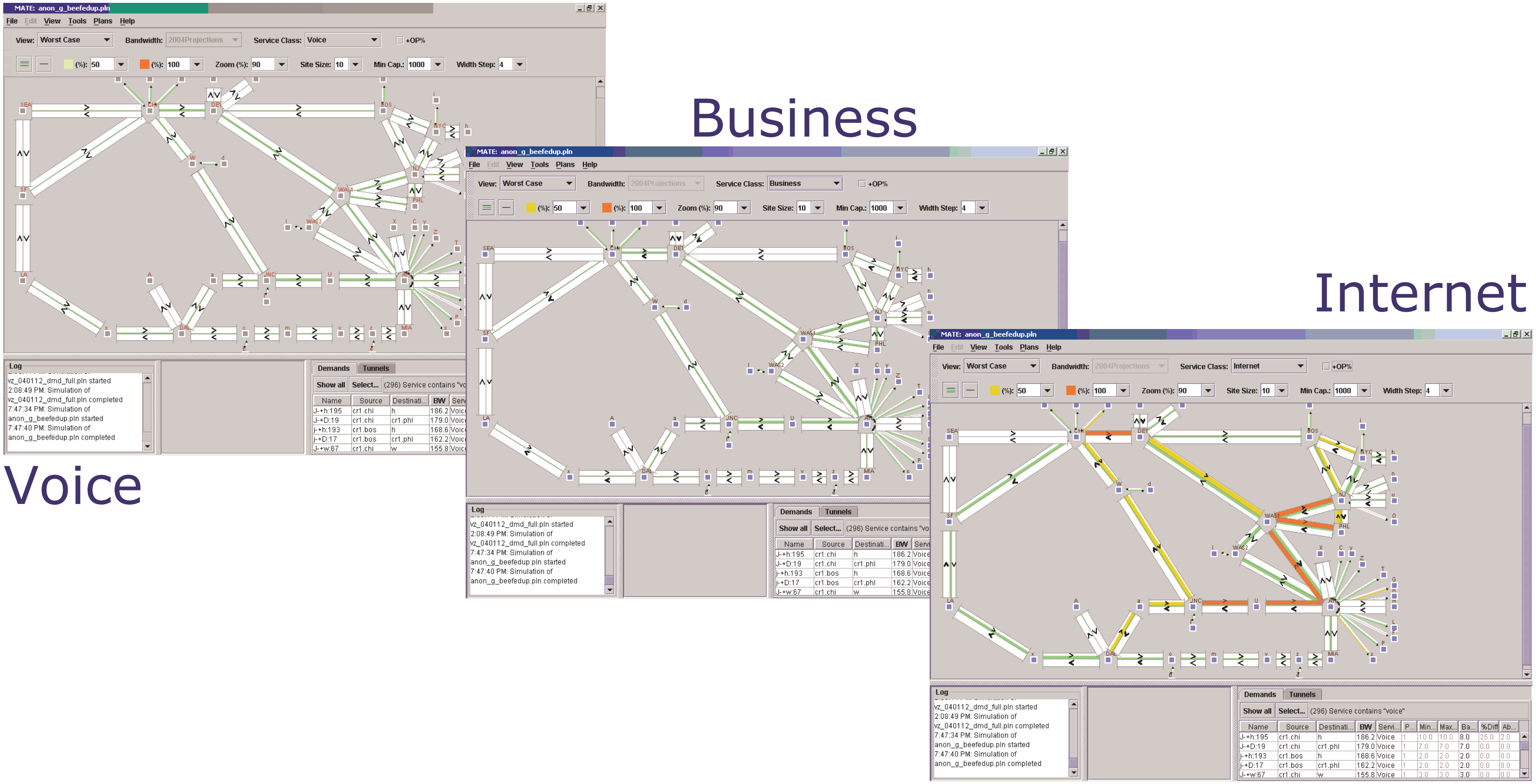
Bandwidth Requirements

Level	Total	Voice	Business	Internet
2004Projections	30590.26	6118.05	6118.05	18354.16

Demands (Growth Class x Service Class)

	Voice (EF)	Business (AF, 90.0%)	Internet (AF, 10.0%)	Total
Voice_Growth (rate 0.09)	296	0	0	296
General_Growth (rate 0.06)	0	296	296	592
Total	296	296	296	888

Worst-Case Failure per Class



Voice

Business

Internet

MATE: anon_g_beefedup.pln

View: Worst Case | Bandwidth: 2004Projections | Service Class: Voice | ☐ +OP%

View: Worst Case | Bandwidth: 2004Projections | Service Class: Business | ☐ +OP%

View: Worst Case | Bandwidth: 2004Projections | Service Class: Internet | ☐ +OP%

(%): 50 | (%): 100 | Zoom (%): 90 | Site Size: 10 | Min Cap.: 1000 | Width Step: 4

```
vz_040112_dmd_full.pln started
2:08:49 PM: Simulation of
vz_040112_dmd_full.pln completed
7:47:34 PM: Simulation of
anon_g_beefedup.pln started
7:47:40 PM: Simulation of
anon_g_beefedup.pln completed
```

(296) Service contains "voice"

Name	Source	Destinati...	BW	Servi...	P...	Min...	Max...	Ba...	%Diff	Ab...
J-+h:195	cr1.chi	h	186.2	Voice	1	10.0	10.0	8.0	25.0	2.0
J-+D:19	cr1.chi	cr1.phl	179.0	Voice	1	7.0	7.0	7.0	0.0	0.0
j-+h:193	cr1.bos	h	168.6	Voice	1	2.0	2.0	2.0	0.0	0.0
j-+D:17	cr1.bos	cr1.phl	162.2	Voice	1	2.0	2.0	2.0	0.0	0.0
J-+w:67	cr1.chi	w	155.8	Voice	1	3.0	3.0	3.0	0.0	0.0

Traffic Characterization Summary

- Long Term Traffic Patterns
  - Smooth for big (relevant) flows
  - Predictable Trends
- Millisecond Time Scale
  - Uncorrelated
  - Not Self-Similar
- High Utilization, Little Delay on High Speed Backbone Links
- QoS via Capacity Planning
  - CoS insurance for failure of capacity planning/TE

Traffic Matrices

I. Traffic Characterization

II. Traffic Matrices

- Measurement Methods
- Estimation Methods

III. TE Introduction

IV. Metric-Based TE

V. Convergence

Core traffic matrix

- Options
  - Full mesh of TE tunnels and Interface MIB
  - NetFlow BGP Next Hop TOS Aggregation
  - NetFlow MPLS Aware
  - MPLS LSR MIB
  - BGP Policy Accounting
  - Interface MIB and 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

Core traffic matrix

- NetFlow
  - MPLS aware Netflow
    - Provides flow statistics per MPLS and IP packets
    - FEC implicitly maps to BGP next hop / egress PE
  - NetFlow BGP Next Hop TOS Aggregation
    - v9 includes accounting based upon BGP next hop NetFlow
- MPLS LSR MIB
  - MPLS-LSR-MIB mirrors the Label Forwarding Information Base (LFIB)
  - FEC implicitly maps to BGP next hop / egress PE

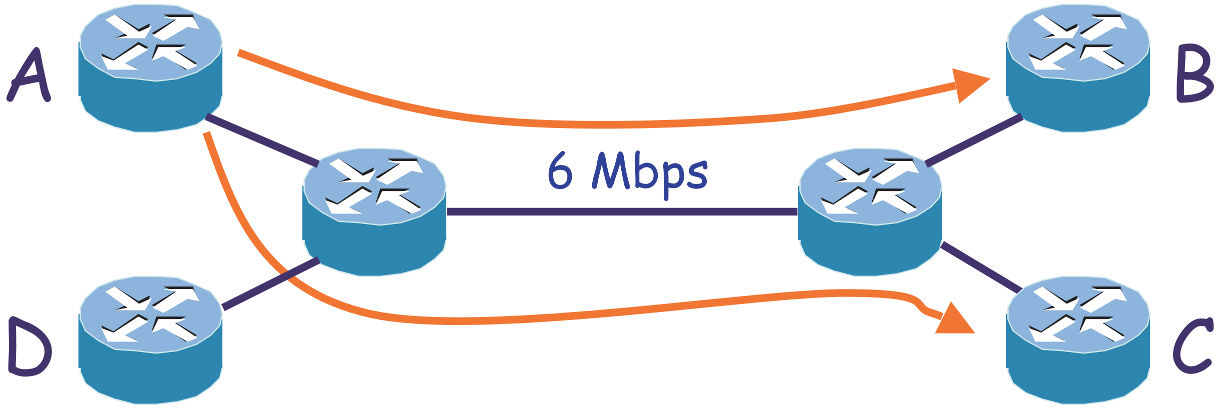
Core traffic matrix

- 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 above methods 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)
  - $O(N^2)$ demands (unknown)

Example

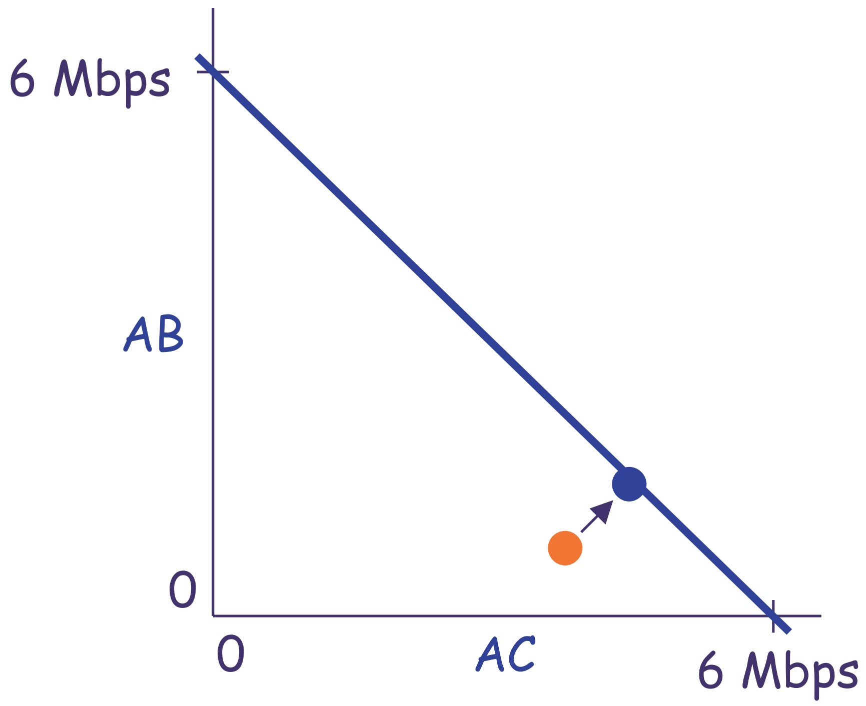


y: link utilizations
A: routing matrix
x: point-to-point demands

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

Example

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%*.

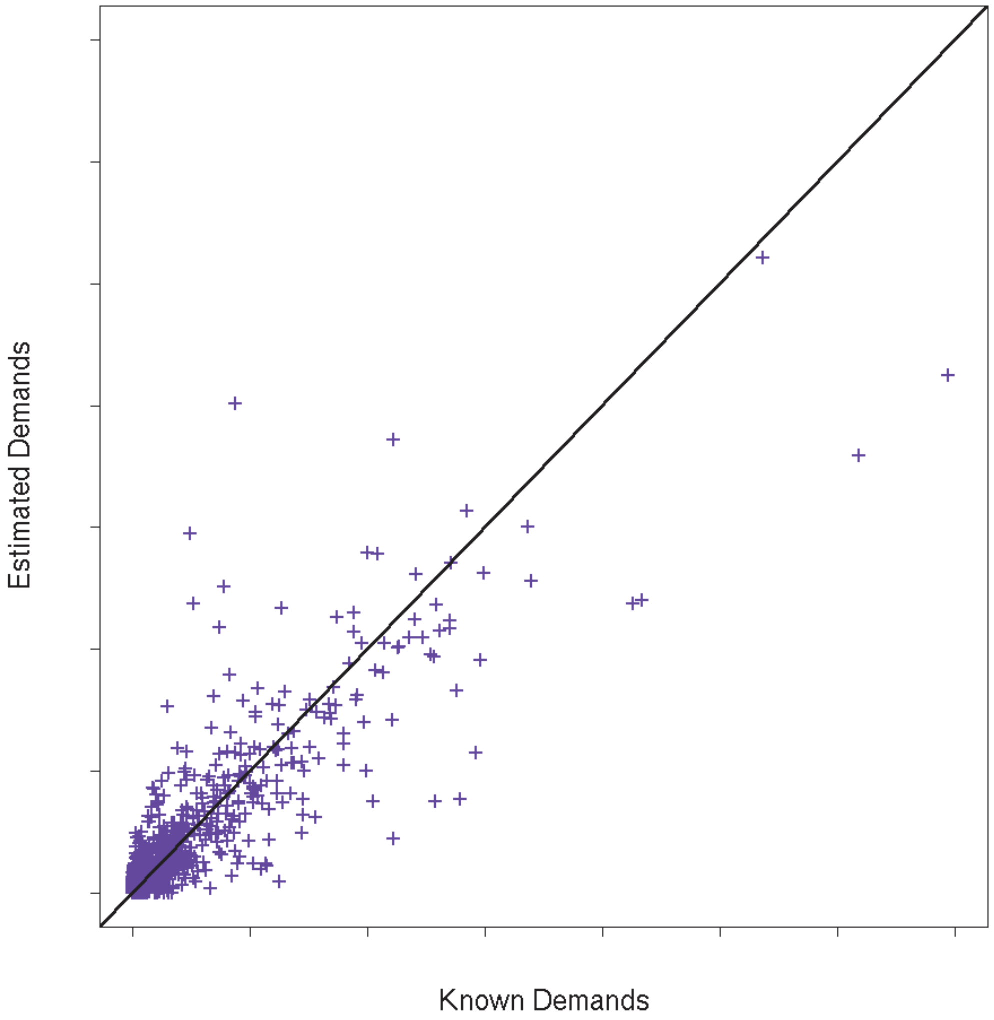
AB = 1 Mbps and AC = 4 Mbps

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

Real Network: Estimated Demands

Cariden Demand Deduction Tool

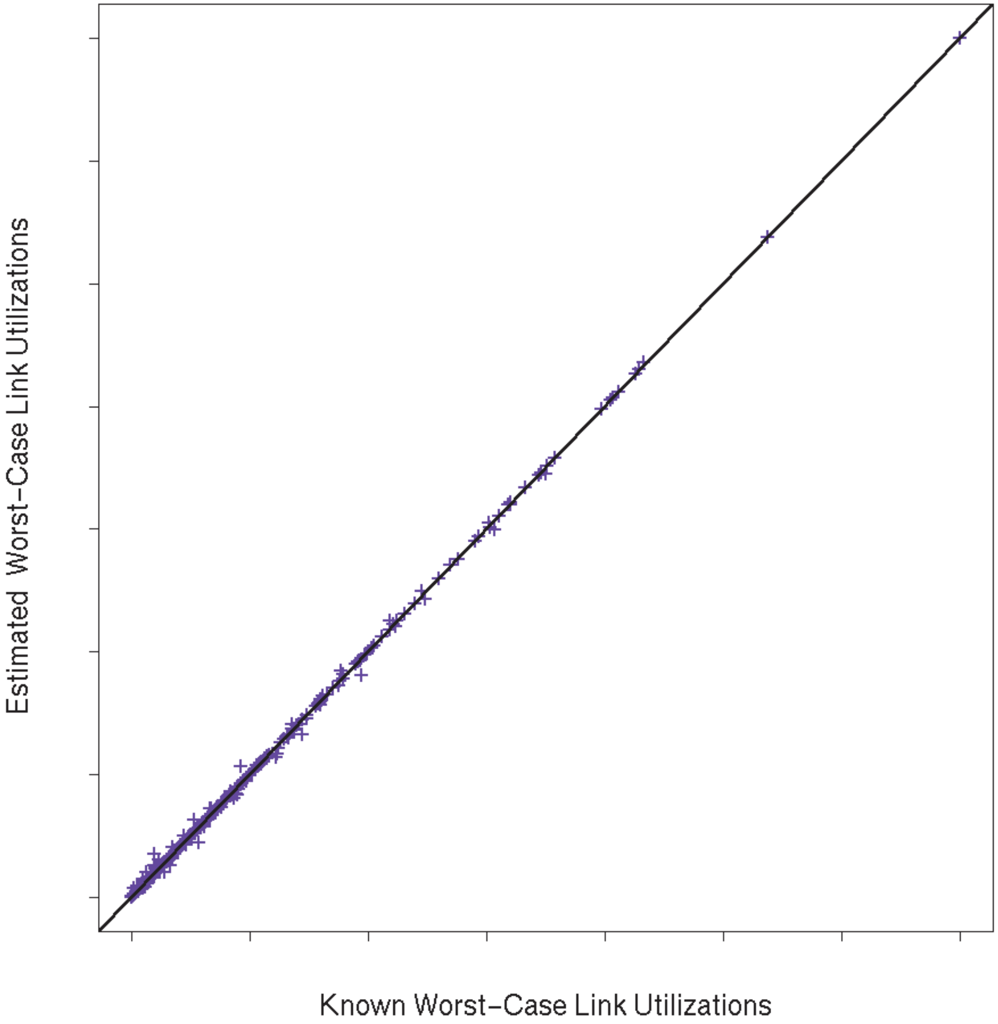
GBLX Network



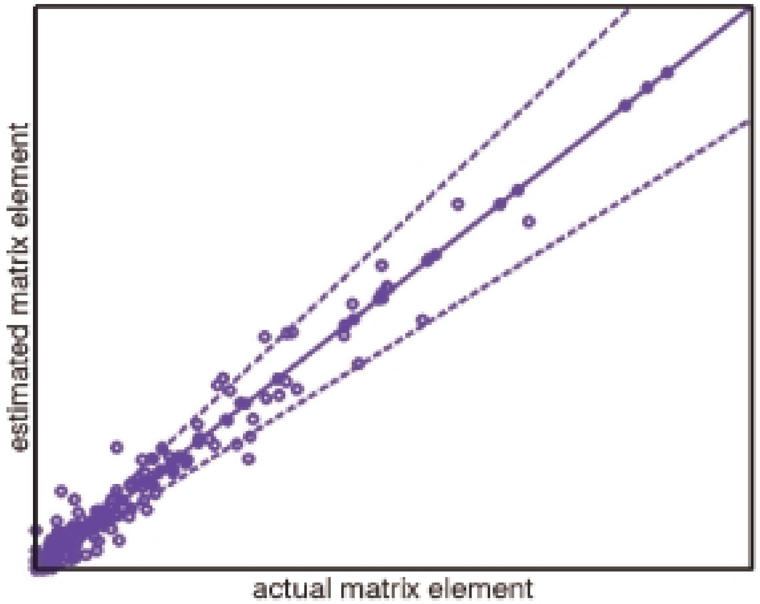
Estimated Link Utilizations!

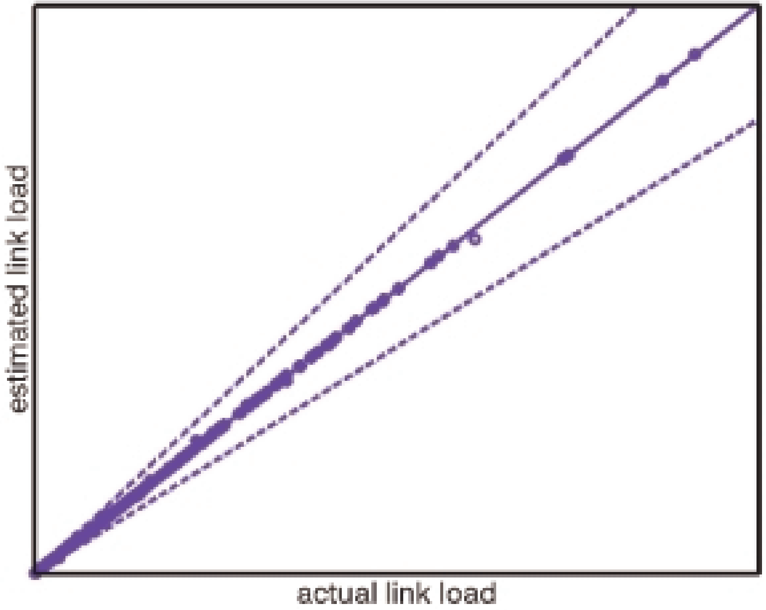
Cariden Demand Deduction Tool

GBLX Network



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

Traffic Matrix Summary

- Existing Options
  - MPLS
  - Netflow
- New Options
  - Netflow BGP Next Hop Aggregation
  - Estimation Based on Link Utilization
- Individual Demand Estimation can be inaccurate
- Estimated Link Utilizations very Accurate

TE Introduction

I. Traffic Characterization

II. Traffic Matrices

III. TE Introduction

- Objectives
- Payback
- Limitations
- Relation to Network Design

IV. Metric-Based TE

V. Convergence

IGP Traffic Engineering

- Manipulate Internal Routing
  - SPF Metrics (OSPF/IS-IS Metrics/Costs/Weights)
  - Explicit Routes

- Minimize Maximum Utilization
  - Normal (Non-Failure) Conditions
  - Single-Element Failure Conditions (typical)

  \+ Latency, Policy Constraints

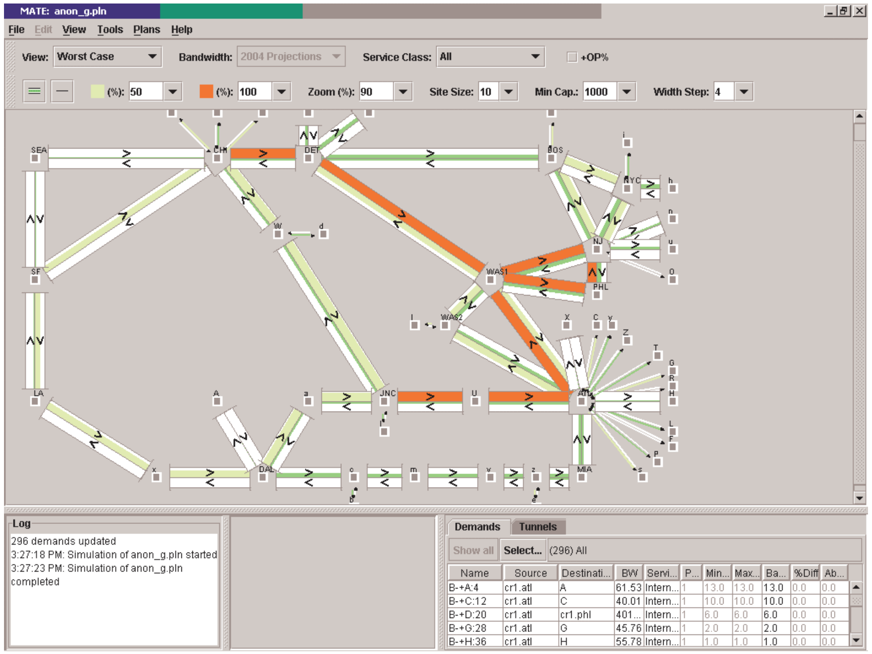
- Given
  - Topology
  - Source-Destination Traffic Matrix

Strategic versus Tactical

- Strategic TE (focus of this presentation)
  - Aimed at $ 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 …

- Tactical TE
  - Aimed at Fixing Problems
  - Short Term Operational/Engineering Process
  - Configure in Response to Failures, Traffic Changes

Strategic TE Payback



View: Worst Case Bandwidth: 2004 Projections Service Class: All ☐ +OP%

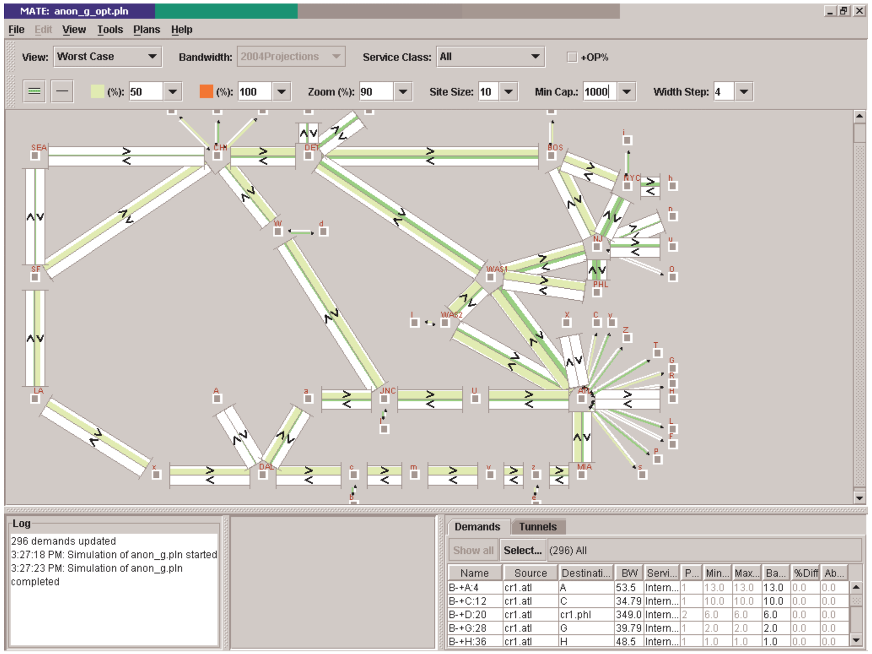
(%): 50 (%): 100 Zoom (%): 90 Site Size: 10 Min Cap.: 1000 Width Step: 4

```
296 demands updated
3:27:18 PM: Simulation of anon_g.pln started
3:27:23 PM: Simulation of anon_g.pln
completed
```

(296) All

Name	Source	Destinati...	BW	Servi...	P...	Min...	Max...	Ba...	%Diff	Ab...
B-+A:4	cr1.atl	A	61.53	Intern...	1	13.0	13.0	13.0	0.0	0.0
B-+C:12	cr1.atl	C	40.01	Intern...	1	10.0	10.0	10.0	0.0	0.0
B-+D:20	cr1.atl	cr1.phl	401...	Intern...	1	6.0	6.0	6.0	0.0	0.0
B-+G:28	cr1.atl	G	45.76	Intern...	1	2.0	2.0	2.0	0.0	0.0
B-+H:36	cr1.atl	H	55.78	Intern...	1	1.0	1.0	1.0	0.0	0.0

Without TE



View: Worst Case Bandwidth: 2004Projections Service Class: All ☐ +OP%

(%): 50 (%): 100 Zoom (%): 90 Site Size: 10 Min Cap.: 1000 Width Step: 4

```
296 demands updated
3:27:18 PM: Simulation of anon_g.pln started
3:27:23 PM: Simulation of anon_g.pln
completed
```

(296) All

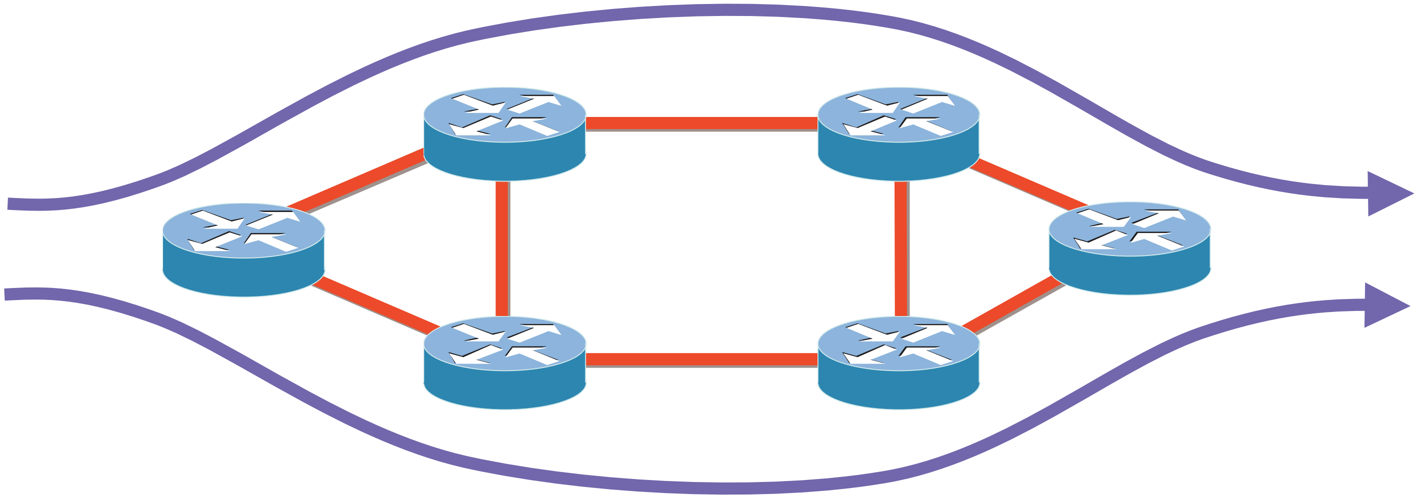
Name	Source	Destinati...	BW	Servi...	P...	Min...	Max...	Ba...	%Diff	Ab...
B-+A:4	cr1.atl	A	53.5	Intern...	1	13.0	13.0	13.0	0.0	0.0
B-+C:12	cr1.atl	C	34.79	Intern...	1	10.0	10.0	10.0	0.0	0.0
B-+D:20	cr1.atl	cr1.phl	349.0	Intern...	2	6.0	6.0	6.0	0.0	0.0
B-+G:28	cr1.atl	G	39.79	Intern...	1	2.0	2.0	2.0	0.0	0.0
B-+H:36	cr1.atl	H	48.5	Intern...	1	1.0	1.0	1.0	0.0	0.0

With TE

- Real Example
  - Delay 6 OC-192 Circuits for a year
    (17 circuits under 50% upgrade policy)
  - Capital + Operational Savings ≈ $1M/OC-192/year

TE Limitations

- Cannot Create Capacity
  - Bottlenecks need capacity not TE
- Limited by Topology
  - E.g., V-O-V topologies allow no Strategic TE
    Only two directions in each “V” or “O” region
    One taken under normal, other under failure
    No routing choice for minimizing failure utilization



TE versus Design Diagnostic

- Proxy for Optimal $/bit Calculation
- Calculate Maximum Link Utilization

	Current Routing	Multicommodity Flow
No Failure	A	C
Worst-Case Failure	B	D

- C/D ≈ 1/2 -> Design Limits Efficiency
  C/D ≈ 3/4 -> Efficient Design
- A»C or B»D -> Inefficient Routing
  A≈C or B≈D -> Efficient Routing

Metric-Based TE

I. Traffic Characterization

II. Traffic Matrices

III. TE Introduction

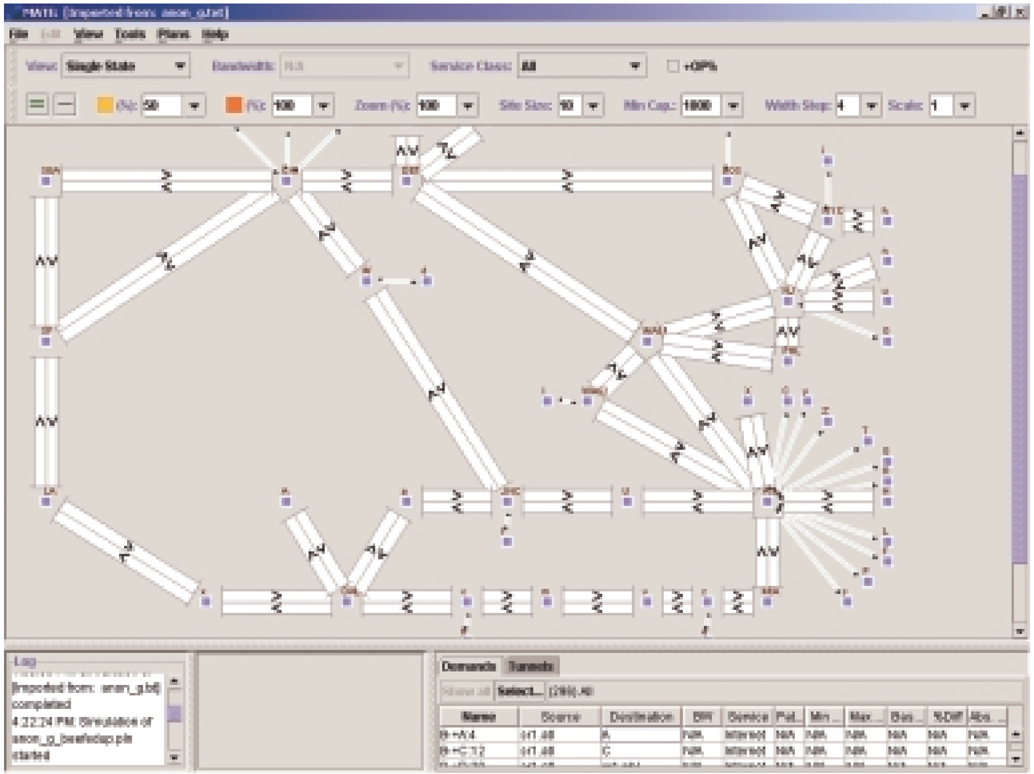
IV. Metric-Based TE

- Case Study
- Performance Evaluation
- Comparison to MPLS TE

V. Convergence

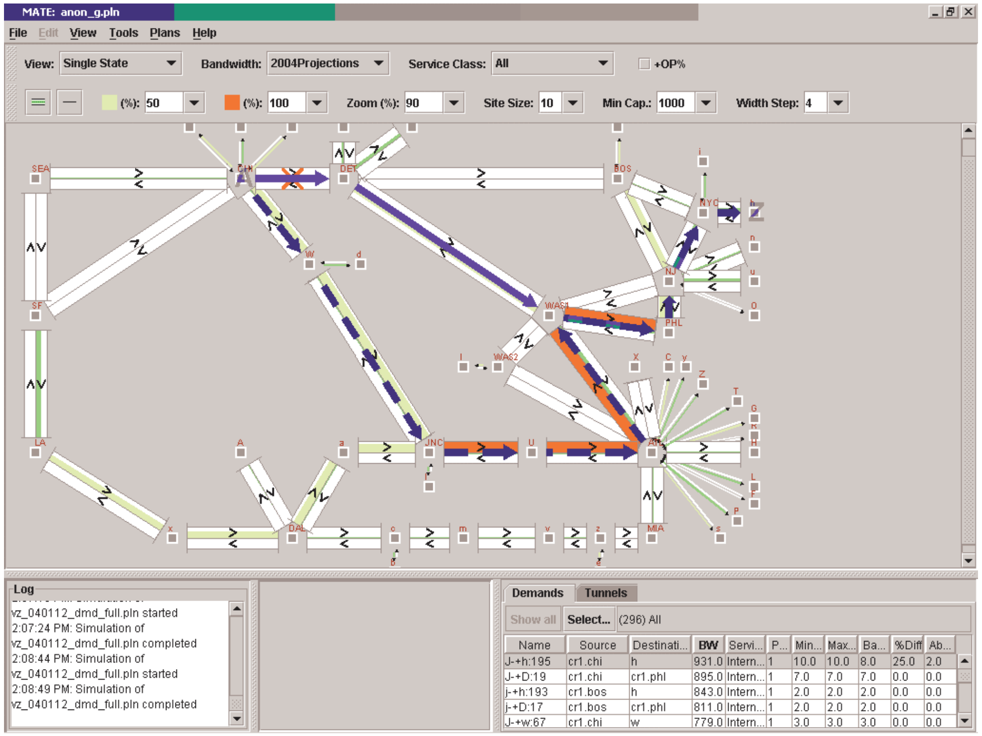
Case Study

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



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



View: Single State
Bandwidth: 2004Projections
Service Class: All
☐ +OP%
🟨 (%): 50
🟧 (%): 100
Zoom (%): 90
Site Size: 10
Min Cap.: 1000
Width Step: 4

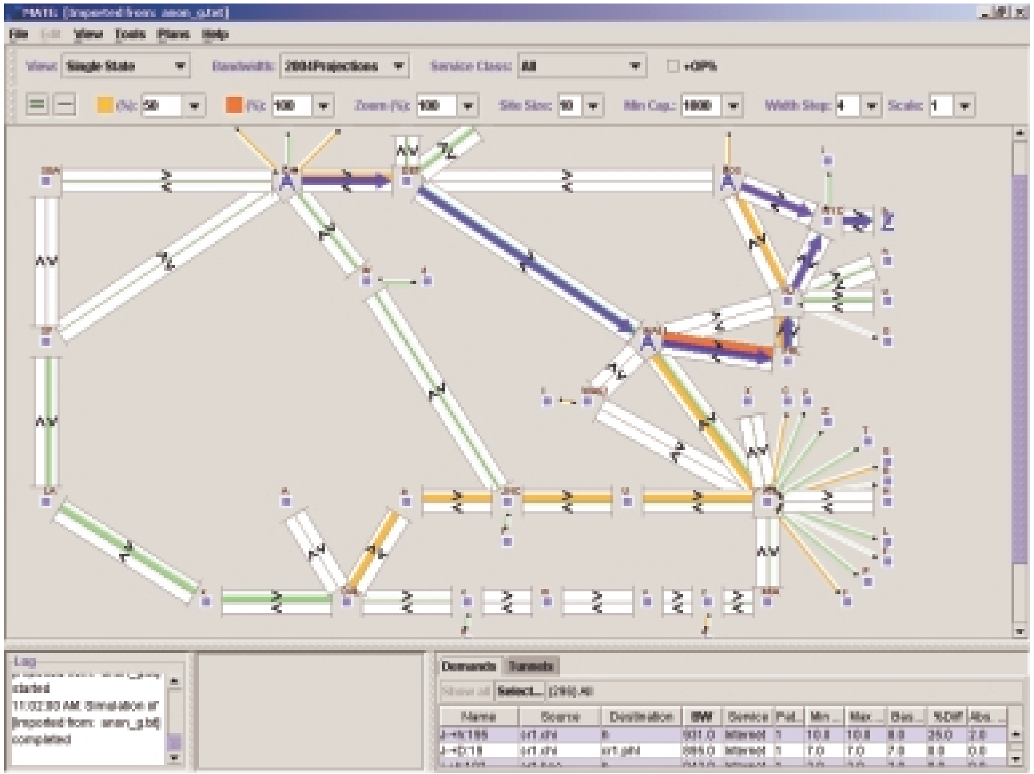
```
vz_040112_dmd_full.pln started
2:07:24 PM: Simulation of
vz_040112_dmd_full.pln completed
2:08:44 PM: Simulation of
vz_040112_dmd_full.pln started
2:08:49 PM: Simulation of
vz_040112_dmd_full.pln completed
```

(296) All

Name	Source	Destinati...	BW	Servi...	P...	Min...	Max...	Ba...	%Diff	Ab...
J-+h:195	cr1.chi	h	931.0	Intern...	1	10.0	10.0	8.0	25.0	2.0
J-+D:19	cr1.chi	cr1.phl	895.0	Intern...	1	7.0	7.0	7.0	0.0	0.0
j-+h:193	cr1.bos	h	843.0	Intern...	1	2.0	2.0	2.0	0.0	0.0
j-+D:17	cr1.bos	cr1.phl	811.0	Intern...	1	2.0	2.0	2.0	0.0	0.0
J-+w:67	cr1.chi	w	779.0	Intern...	1	3.0	3.0	3.0	0.0	0.0

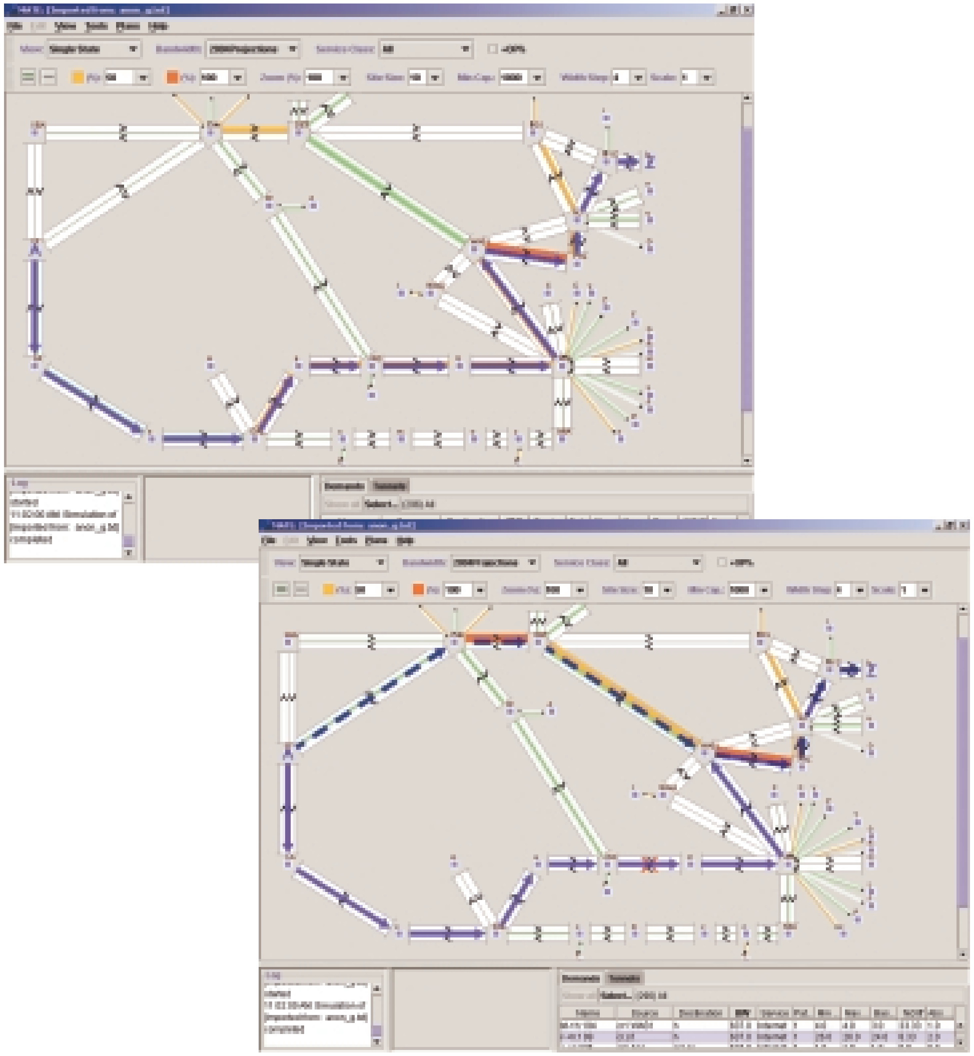
Traffic Overview

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



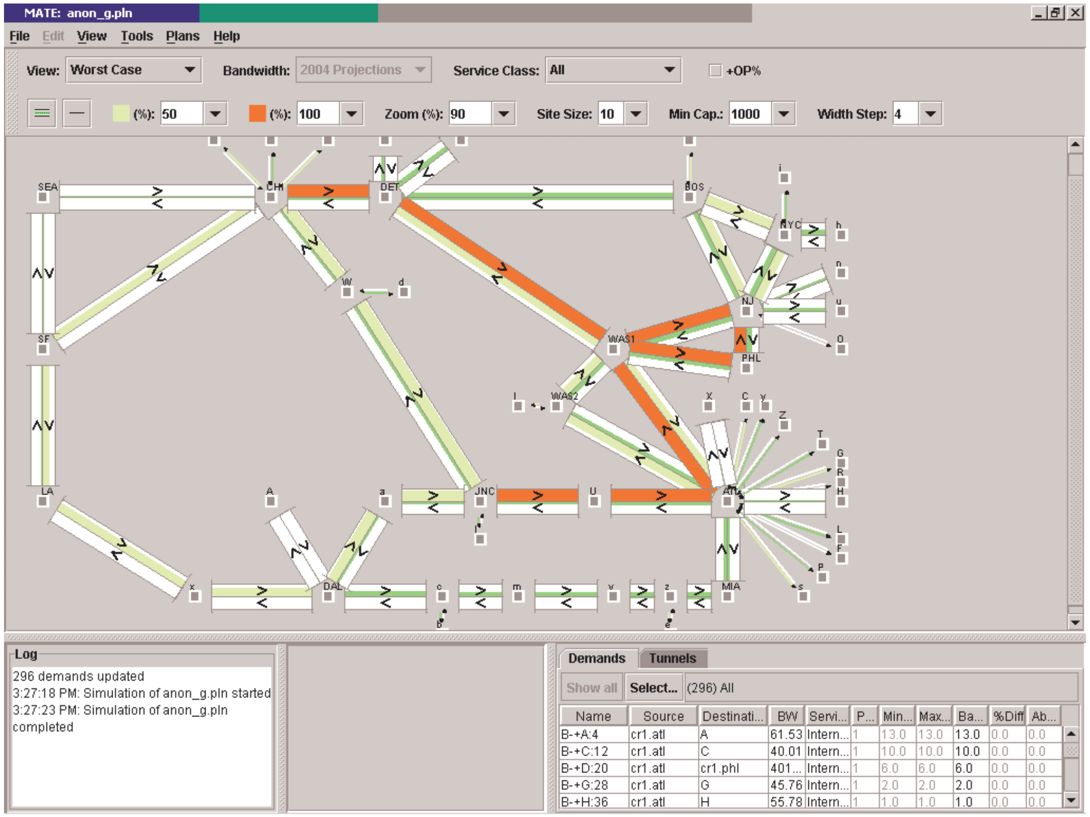
Manual Attempt at Metric TE

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



Worst Case Failure View

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



View: Worst Case

Bandwidth: 2004 Projections

Service Class: All

☐ +OP%

🟩 (%): 50

🟧 (%): 100

Zoom (%): 90

Site Size: 10

Min Cap.: 1000

Width Step: 4

Log

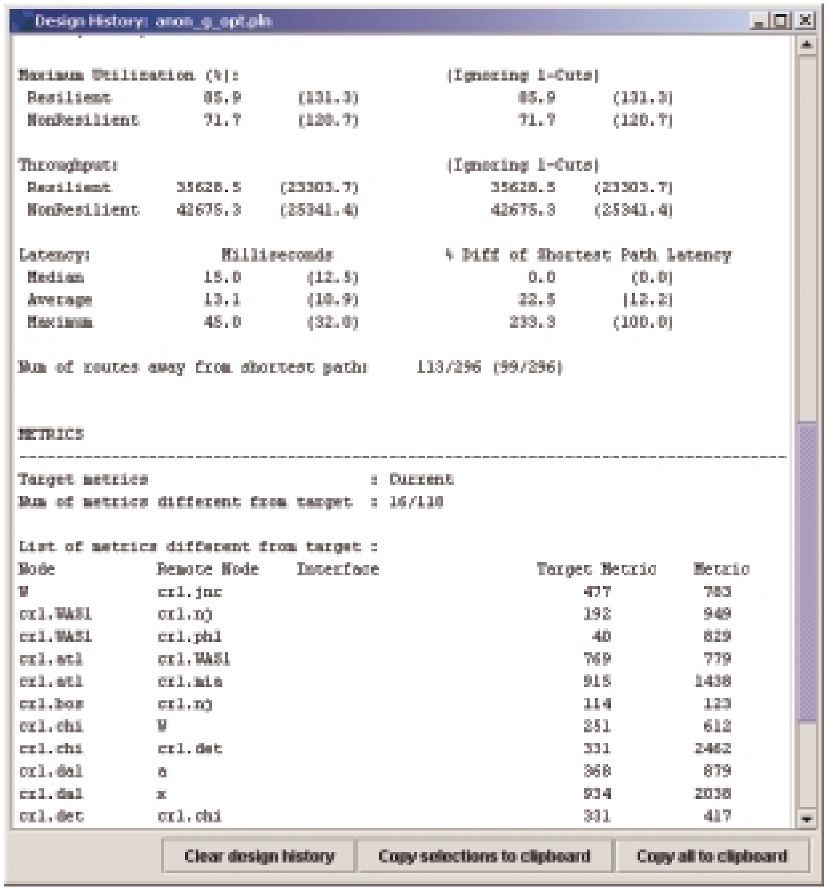
```
296 demands updated
3:27:18 PM: Simulation of anon_g.pln started
3:27:23 PM: Simulation of anon_g.pln
completed
```

(296) All

Name	Source	Destinati...	BW	Servi...	P...	Min...	Max...	Ba...	%Diff	Ab...
B-+A:4	cr1.atl	A	61.53	Intern...	1	13.0	13.0	13.0	0.0	0.0
B-+C:12	cr1.atl	C	40.01	Intern...	1	10.0	10.0	10.0	0.0	0.0
B-+D:20	cr1.atl	cr1.phl	401...	Intern...	1	6.0	6.0	6.0	0.0	0.0
B-+G:28	cr1.atl	G	45.76	Intern...	1	2.0	2.0	2.0	0.0	0.0
B-+H:36	cr1.atl	H	55.78	Intern...	1	1.0	1.0	1.0	0.0	0.0

Cariden Metric TE

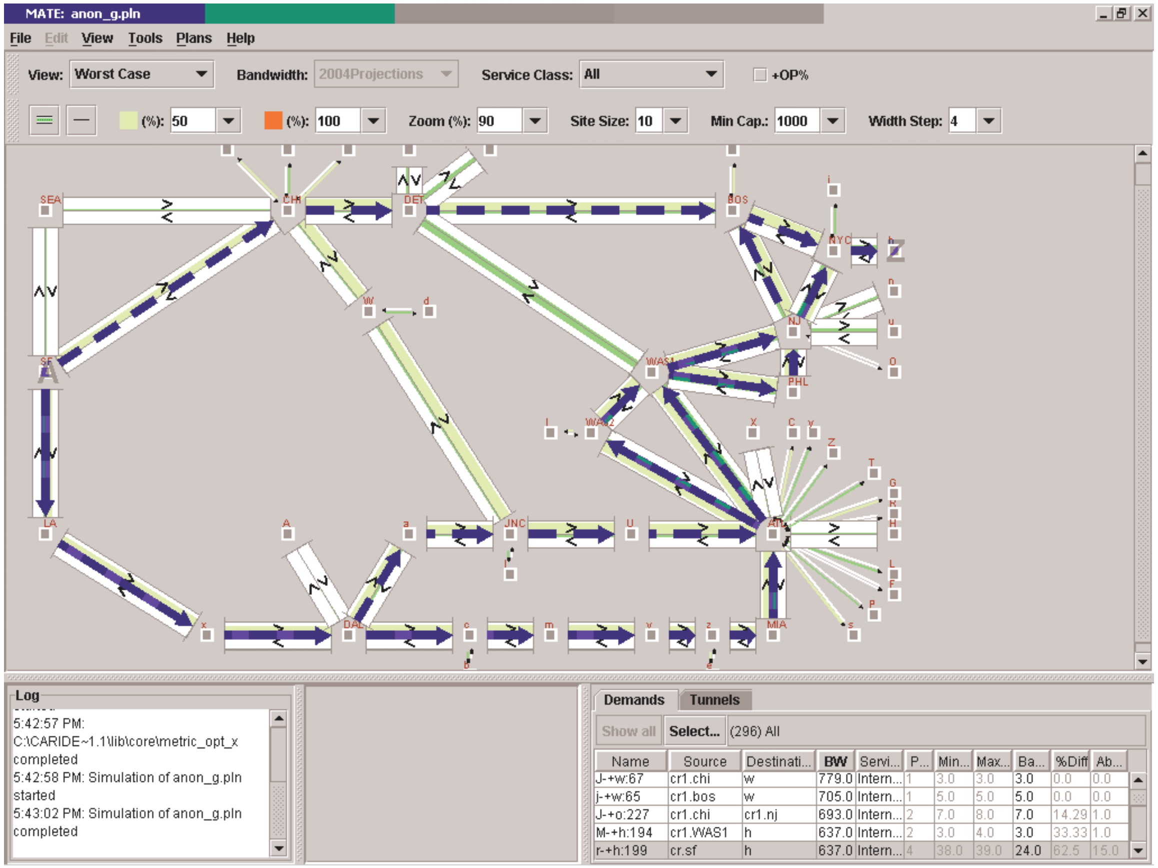
- Change 16 metrics
- Remove congestion
  - Normal (121% -> 72%)
  - Worst case link failure (131% -> 86%)



```
Maximum Utilization (%):                             (Ignoring 1-Cuts)
 Resilient           85.9        (131.3)                       85.9       (131.3)
 NonResilient        71.7        (120.7)                       71.7       (120.7)

Throughput:                                          (Ignoring 1-Cuts)
 Resilient        35628.5      (23303.7)                    35628.5     (23303.7)
 NonResilient     42675.3      (25341.4)                    42675.3     (25341.4)

Latency:                  Milliseconds               % Diff of Shortest Path Latency
 Median              15.0         (12.5)                        0.0         (0.0)
 Average             13.1         (10.9)                       22.5        (12.2)
 Maximum             45.0         (32.0)                      233.3       (100.0)

Num of routes away from shortest path:      113/296 (59/296)


METRICS
----------------------------------------------------------------------------------
Target metrics                               : Current
Num of metrics different from target         : 16/118

List of metrics different from target :
Node              Remote Node     Interface                 Target Metric      Metric
W                 crl.jnc                                             477         783
crl.WAS1          crl.nj                                              192         949
crl.WAS1          crl.phl                                              40         829
crl.atl           crl.WAS1                                            769         779
crl.atl           crl.mia                                             915        1438
crl.bos           crl.nj                                              114         123
crl.chi           W                                                   251         612
crl.chi           crl.det                                             331        2462
crl.dal           a                                                   368         879
crl.dal           x                                                   934        2038
crl.det           crl.chi                                             331         417
```

New Routing Visualization

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



View: Worst Case | Bandwidth: 2004Projections | Service Class: All | ☐ +OP%

(%): 50 | (%): 100 | Zoom (%): 90 | Site Size: 10 | Min Cap.: 1000 | Width Step: 4

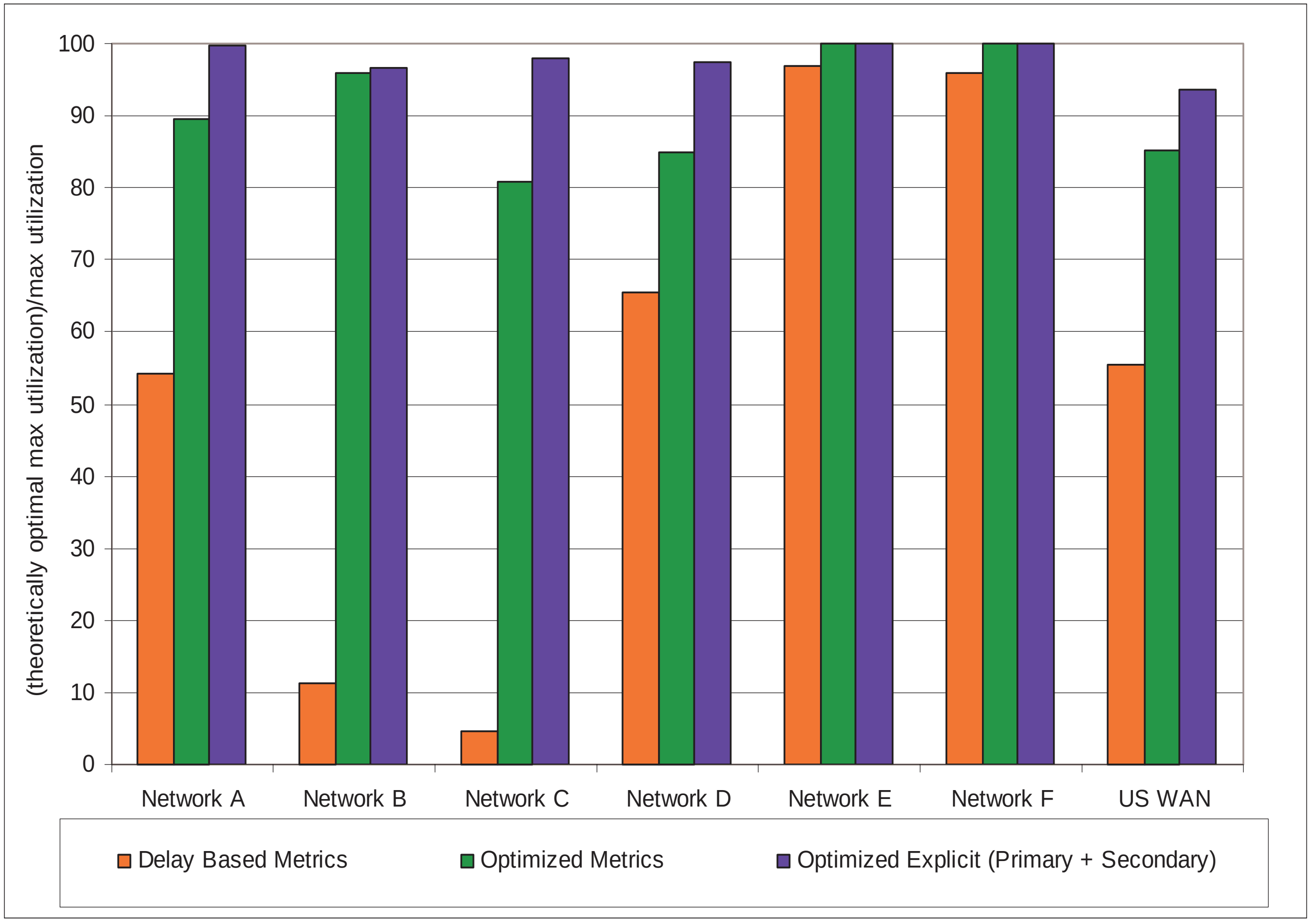
```
5:42:57 PM:
C:\CARIDE~1.1\lib\core\metric_opt_x
completed
5:42:58 PM: Simulation of anon_g.pln
started
5:43:02 PM: Simulation of anon_g.pln
completed
```

(296) All

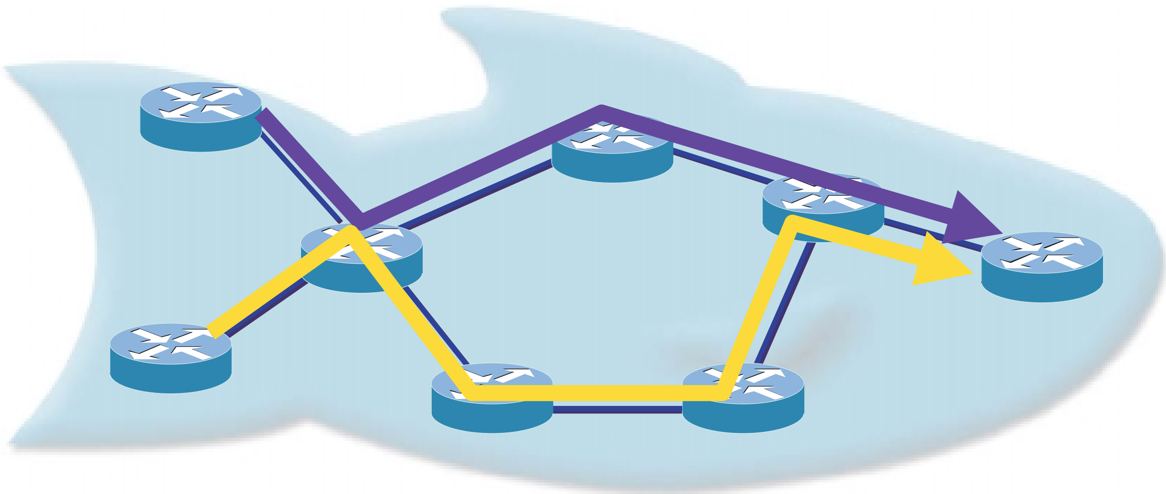
Name	Source	Destinati...	BW	Servi...	P...	Min...	Max...	Ba...	%Diff	Ab...
J-+w:67	cr1.chi	w	779.0	Intern...	1	3.0	3.0	3.0	0.0	0.0
j-+w:65	cr1.bos	w	705.0	Intern...	1	5.0	5.0	5.0	0.0	0.0
J-+o:227	cr1.chi	cr1.nj	693.0	Intern...	2	7.0	8.0	7.0	14.29	1.0
M-+h:194	cr1.WAS1	h	637.0	Intern...	2	3.0	4.0	3.0	33.33	1.0
r-+h:199	cr.sf	h	637.0	Intern...	4	38.0	39.0	24.0	62.5	15.0

Metric-Based TE Evaluation

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



MPLS TE

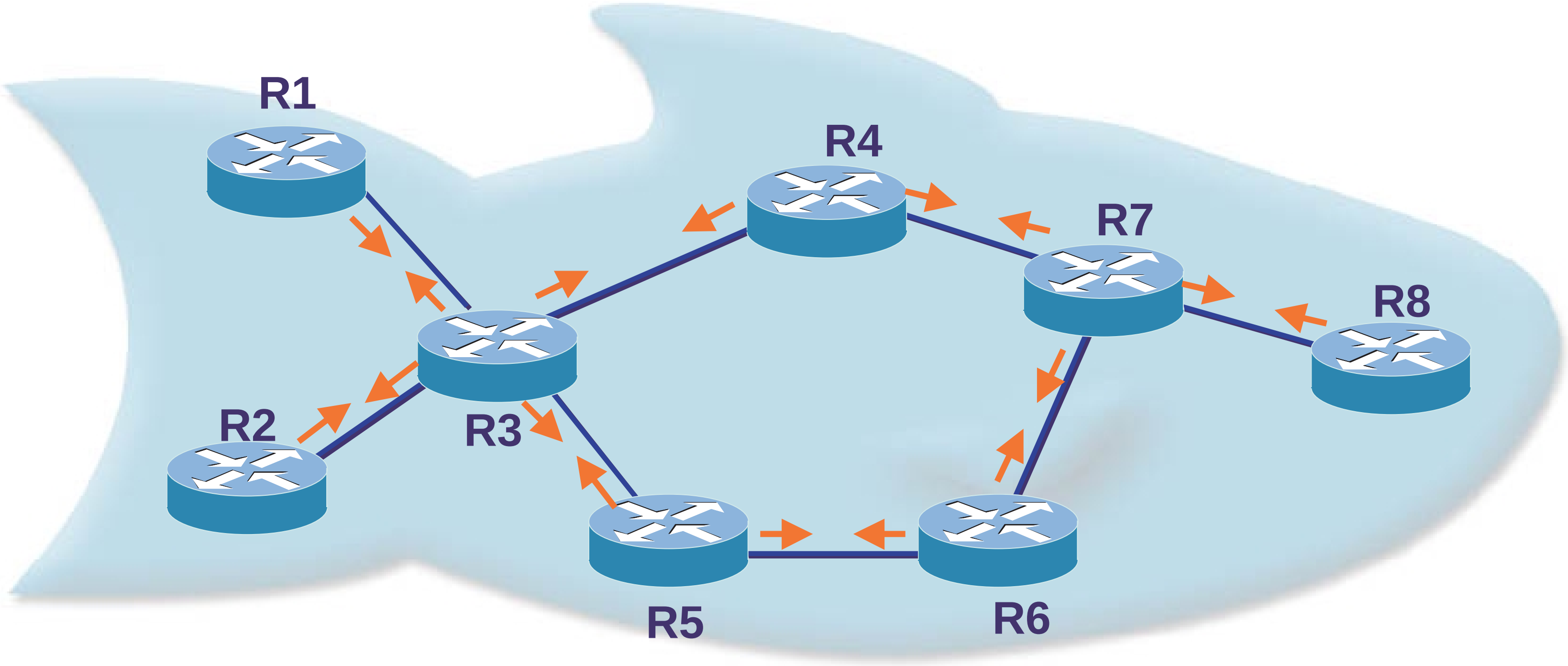


- MPLS Traffic Engineering gives us an “explicit” routing capability (a.k.a. “source routing”) at Layer 3
  - Lets you use paths other than IGP shortest path
  - Allows unequal-cost load sharing
- MPLS TE label switched paths (termed “traffic engineering tunnels”) are used to steer traffic through the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

MPLS TE Components – Refresher

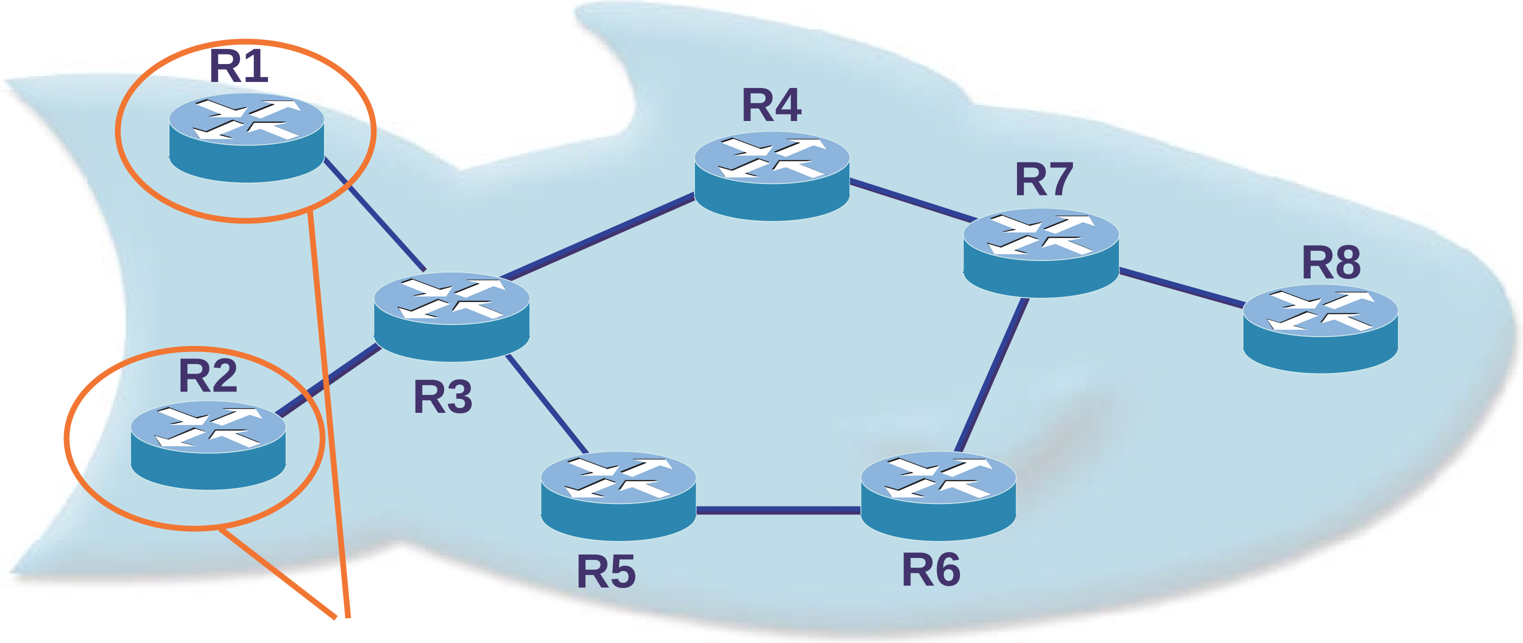
- Resource / policy information distribution
- Constraint based path computation
- RSVP for tunnel signaling
- Link admission control
- LSP establishment
- TE tunnel control and maintenance
- Assign traffic to tunnels

MPLS TE Components (1)



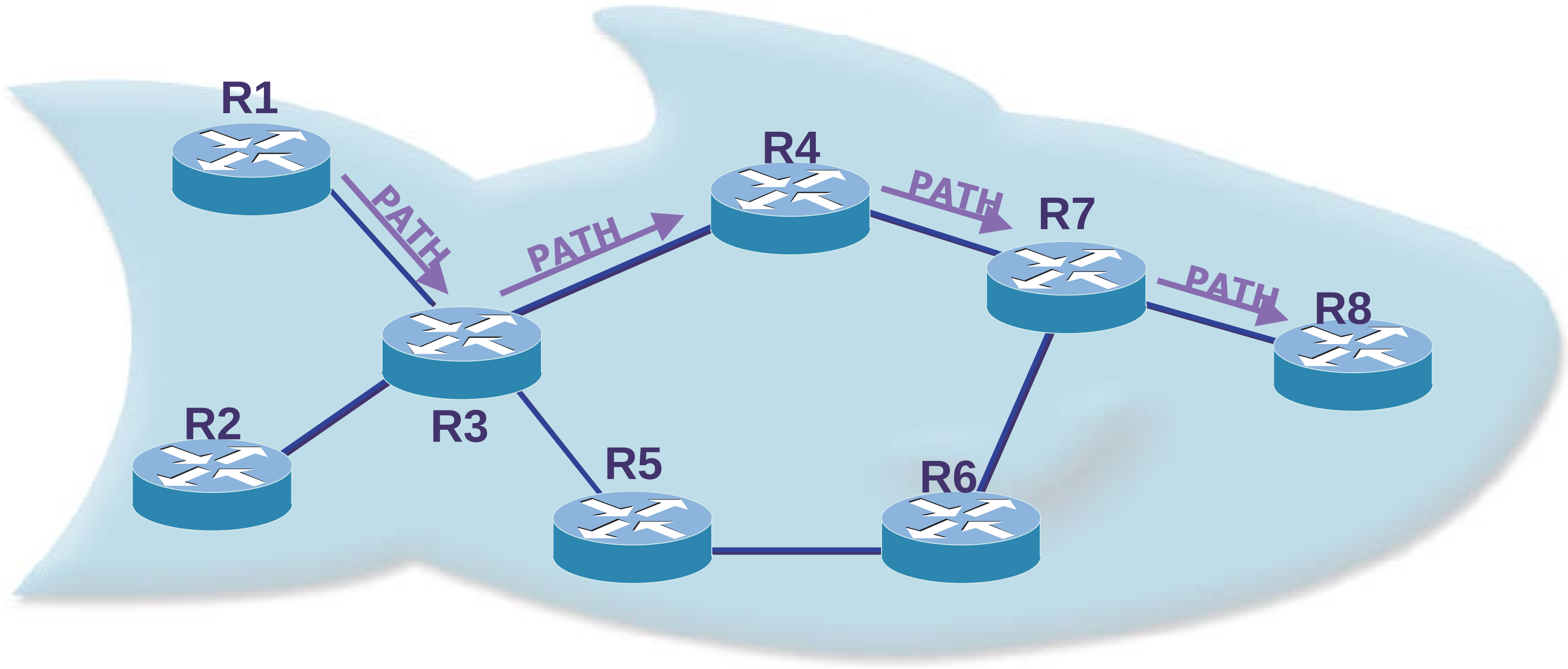
- Resource / policy information distribution
  - OSPF / IS-IS extensions are used to advertise “unreserved capacity” and administrative attributes per link

MPLS TE Components (2)



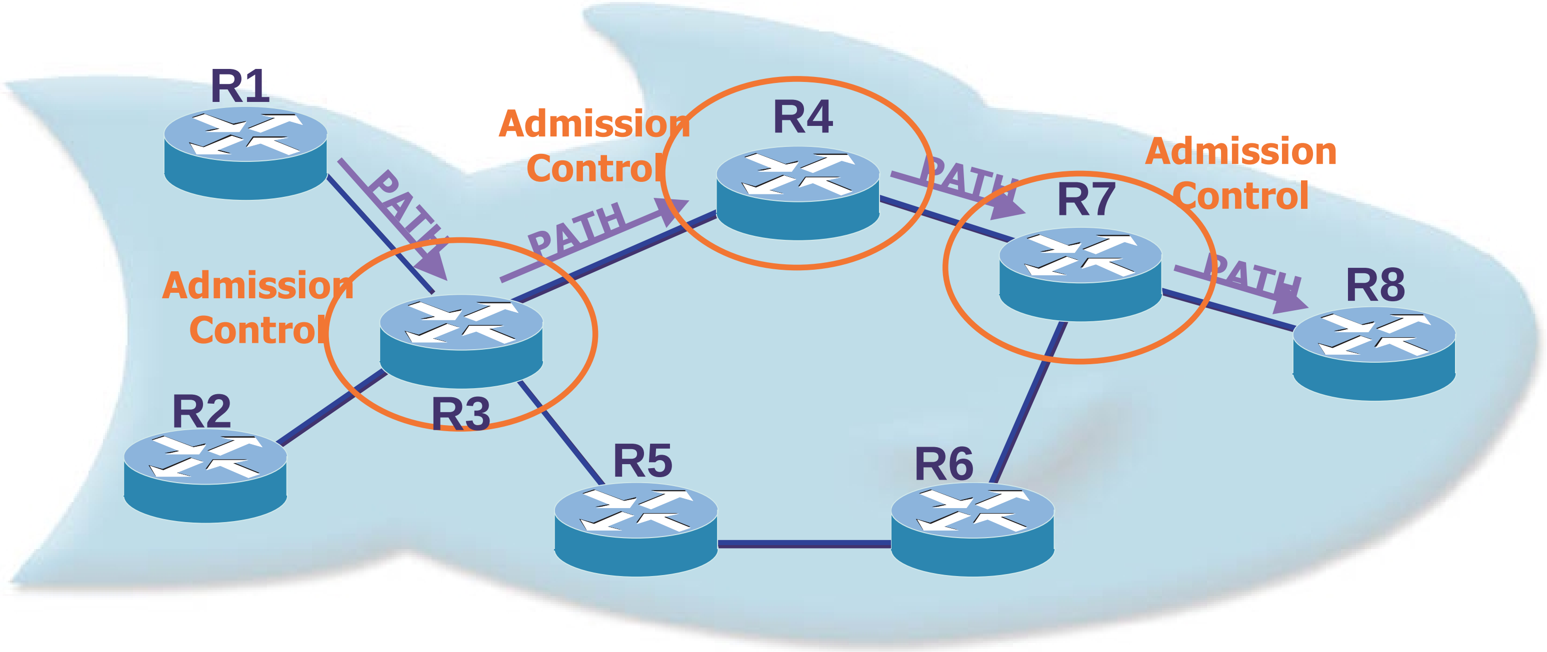
- Constraint based path computation
  - Constraints (required bandwidth and policy) are specified for a TE “tunnel”
  - Constraint based routing – PCALC on head-end routers calculates best path that satisfies constraints based upon the received topology and policy information
    - prune unsuitable links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

MPLS TE Components (3)



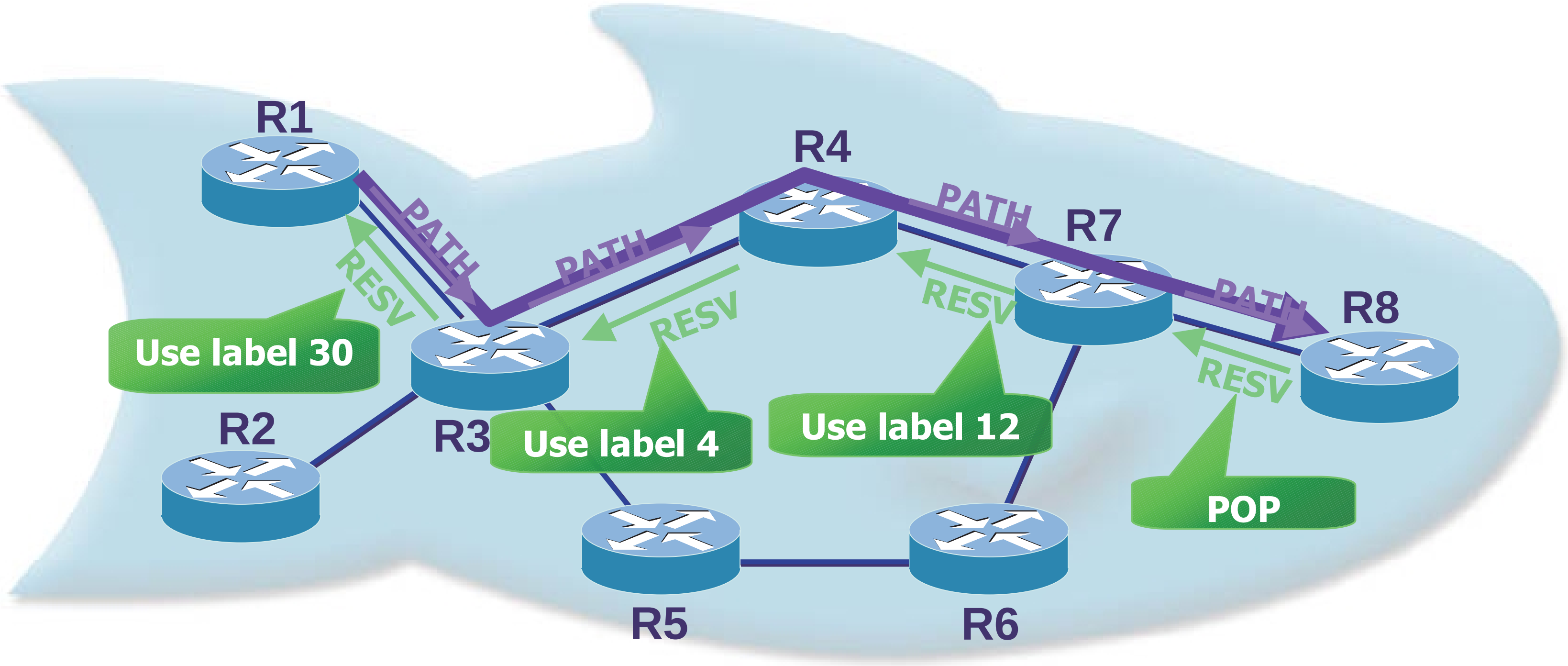
- RSVP for Tunnel Signaling
  - Output of constraint based routing is an explicit route used by RSVP (with extensions) for tunnel signaling
    - ERO = R1->R3->R4->R7->R8

MPLS TE Components (4)



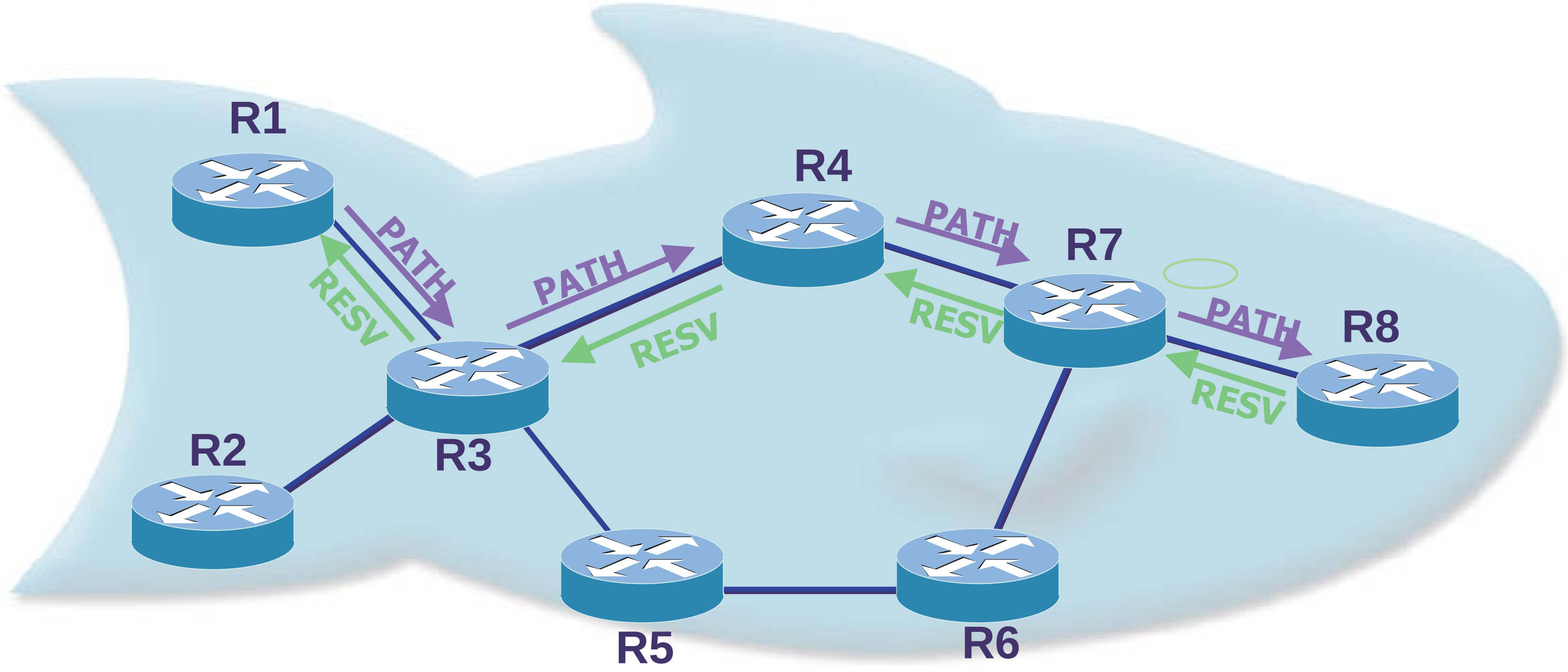
- Link admission control
  - At each hop – determines if resources are available
    - If Admission Control fails, send PathError
    - May tear down (existing) TE LSPs with a lower priority
    - Triggers IGP information distribution when resource thresholds are crossed

MPLS TE Components (5)



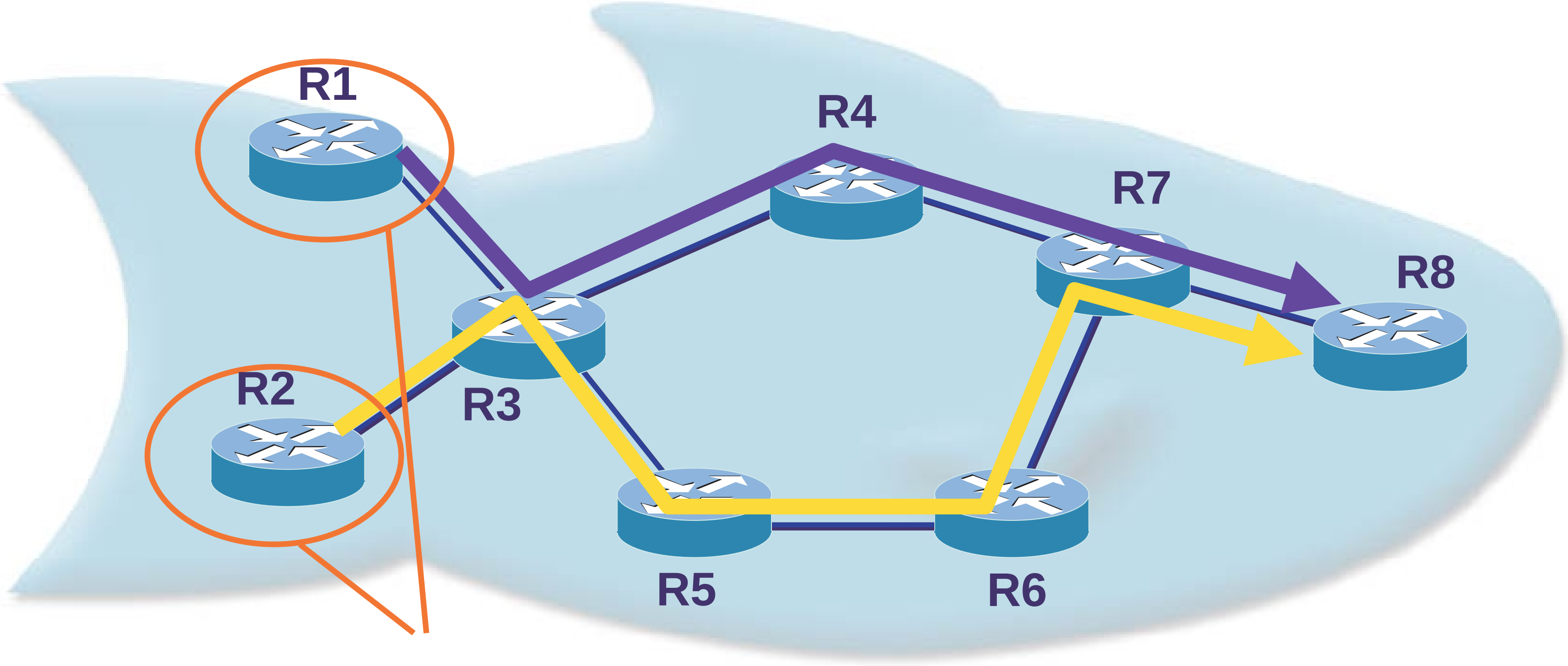
- LSP Establishment
  - RESV confirms bandwidth reservation and distributes labels
    - downstream on demand label allocation
  - MPLS used for forwarding – overcomes issues of IP destination based forwarding

MPLS TE Components (6)



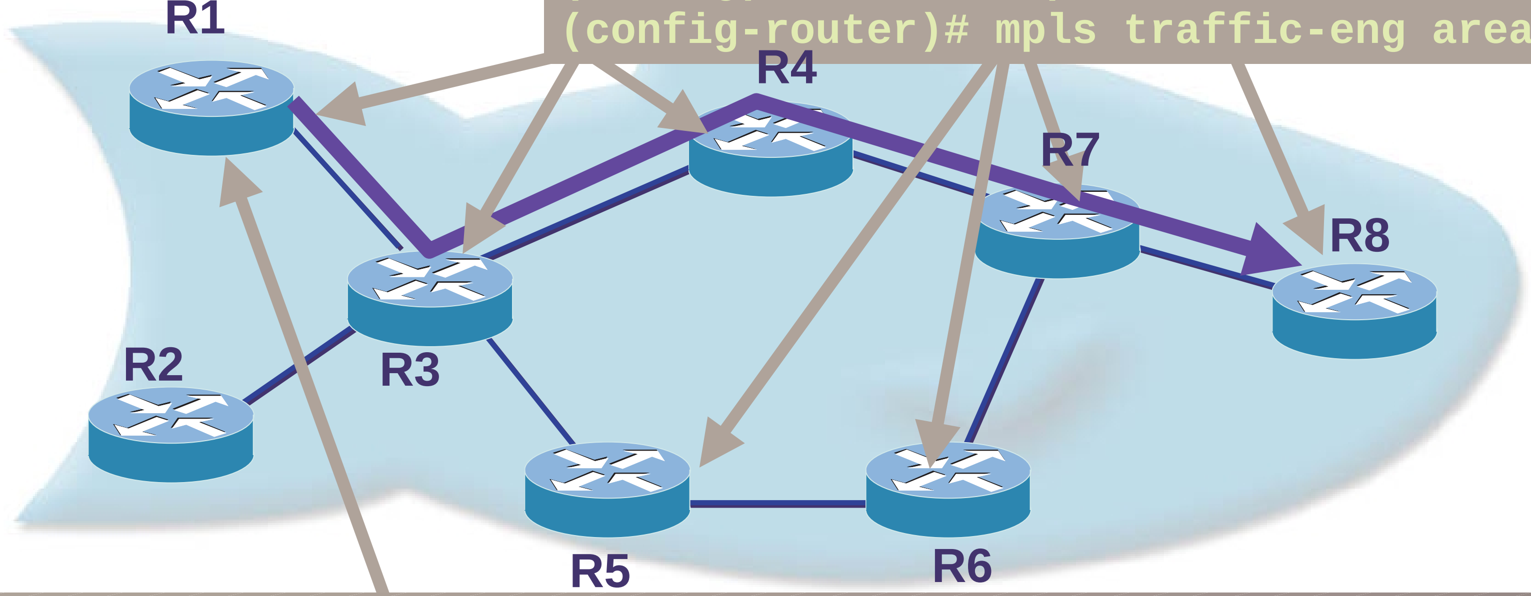
- TE tunnel control and maintenance
  - Periodic RSVP PATH/RESV messages maintain tunnels

MPLS TE Components (7)



- Assign traffic to tunnels
  - Head-end routers assign traffic to tunnels using:
    - Static routing, Autoroute or PBR

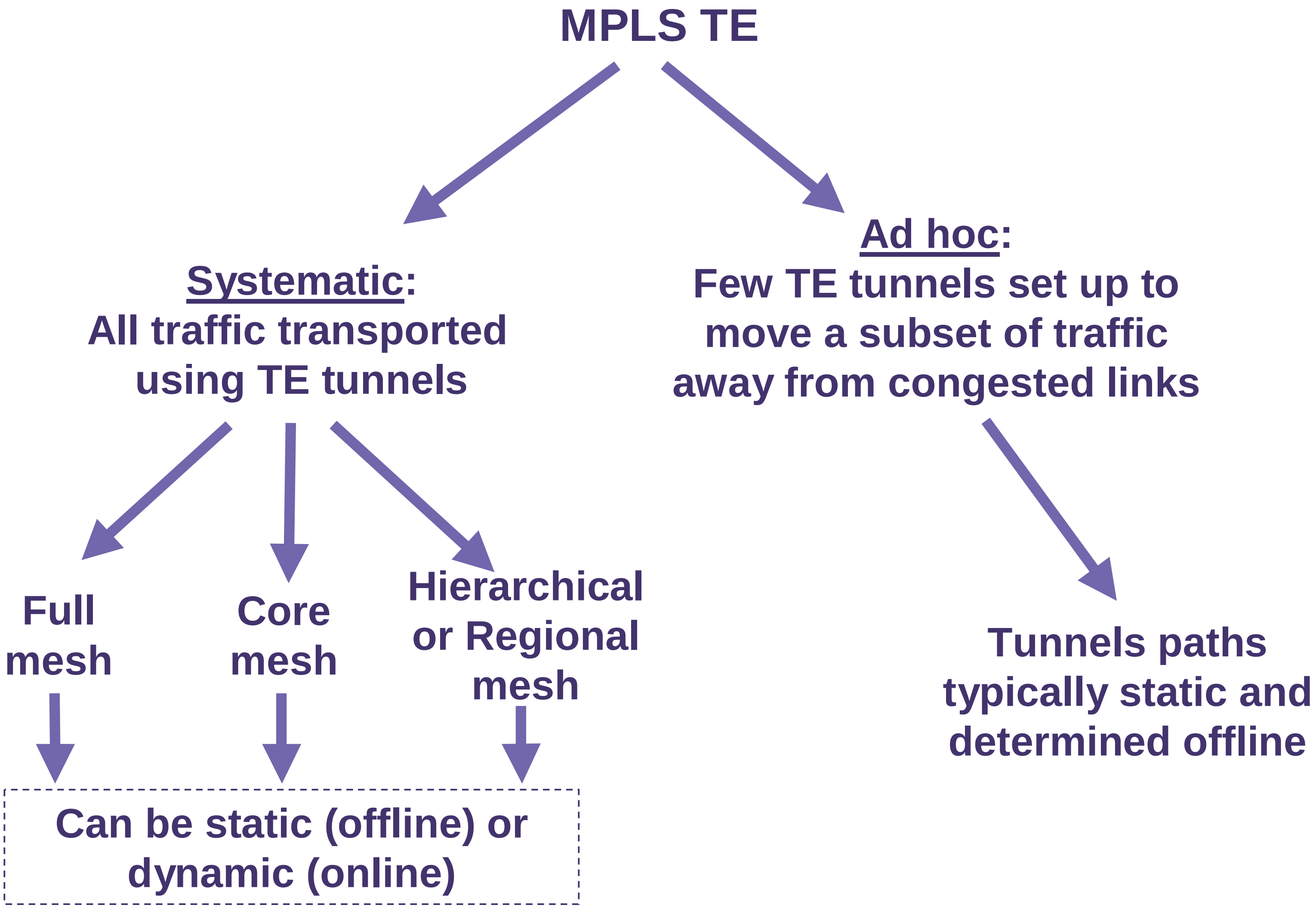
MPLS TE Components: Minimum Config



```
(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

Systematic:
All traffic transported
using TE tunnels

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

Full
mesh

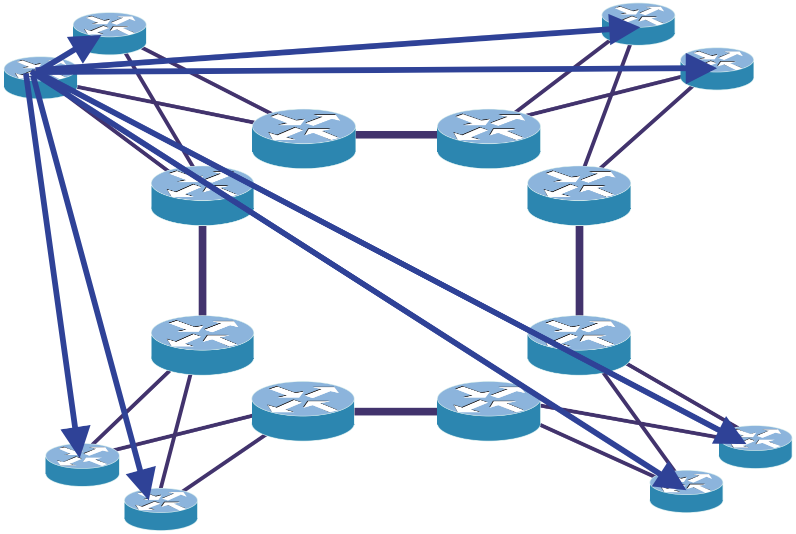
Core
mesh

Hierarchical
or Regional
mesh

Can be static (offline) or
dynamic (online)

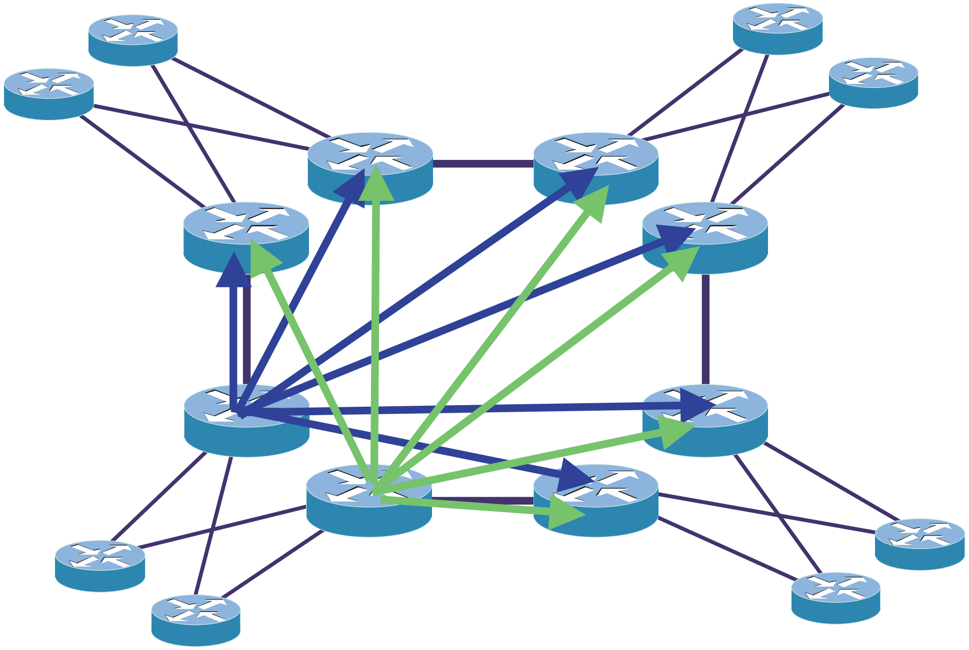
Tunnels paths
typically static and
determined offline

Systematic Deployment: Full Mesh



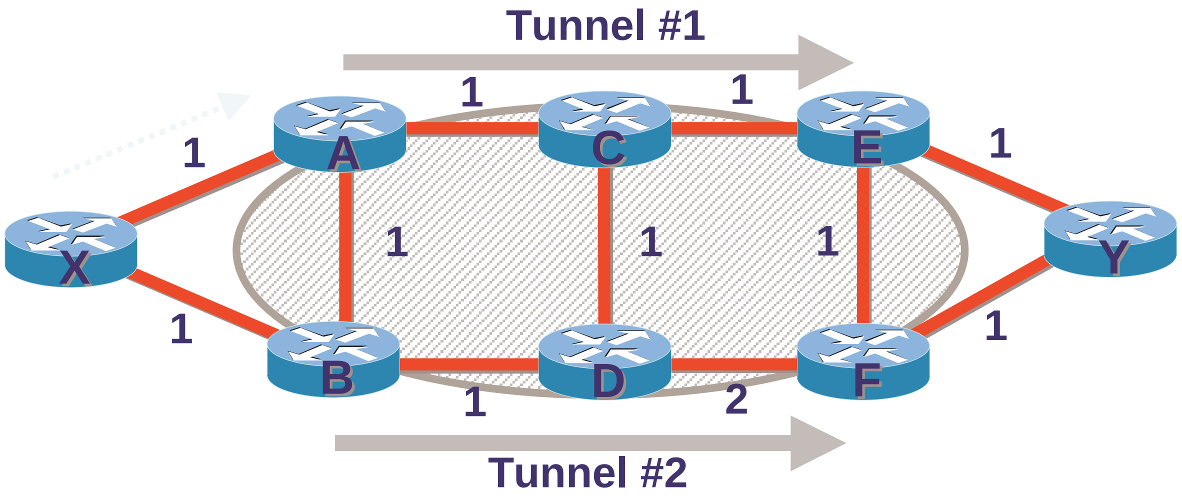
- Requires n * (n-1) tunnels, where n = # of head-ends
- Reality check: largest TE network today has ~100 head-ends
  - ➔ ~9,900 tunnels in total
  - ➔ max 99 tunnels per head-end
  - ➔ max ~1,500 tunnels per link
- Provisioning burden may be eased with AutoTunnel Mesh

Systematic Deployment: Core Mesh



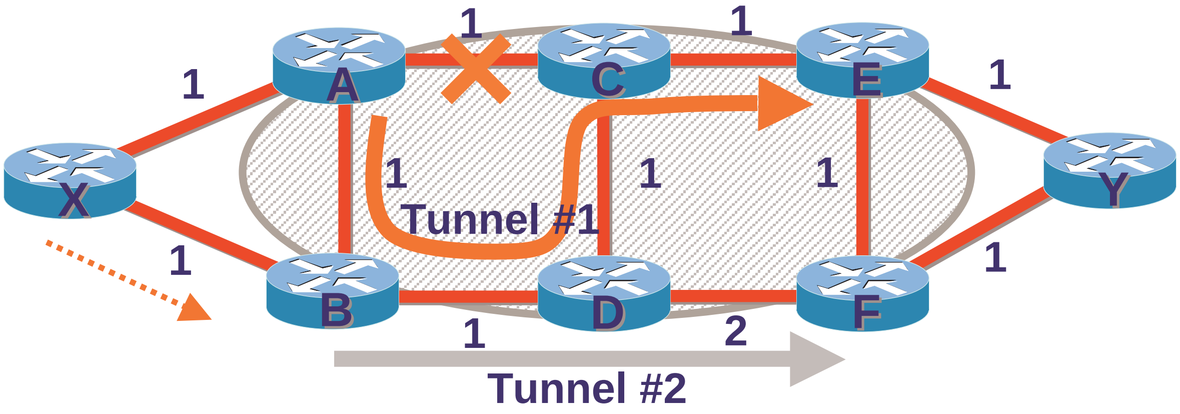
- Reduces number of tunnels required
- Can be susceptible to “traffic-sloshing”

Traffic "sloshing"



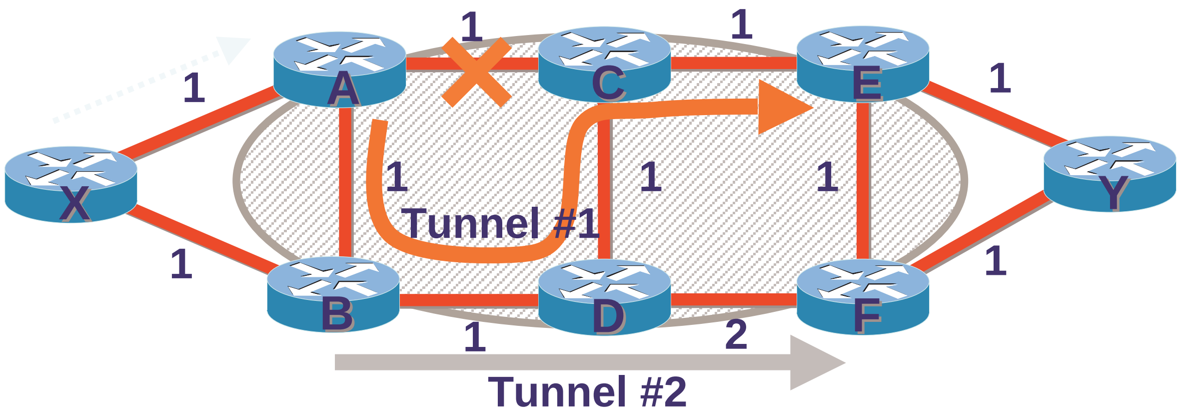
- In normal case:
  - For traffic from X ➔ Y, router X IGP will see best path via router A
  - Tunnel #1 will be sized for X ➔ Y demand
  - If bandwidth is available on all links, Tunnel from A to E will follow path A ➔ C ➔ E

Traffic "sloshing"



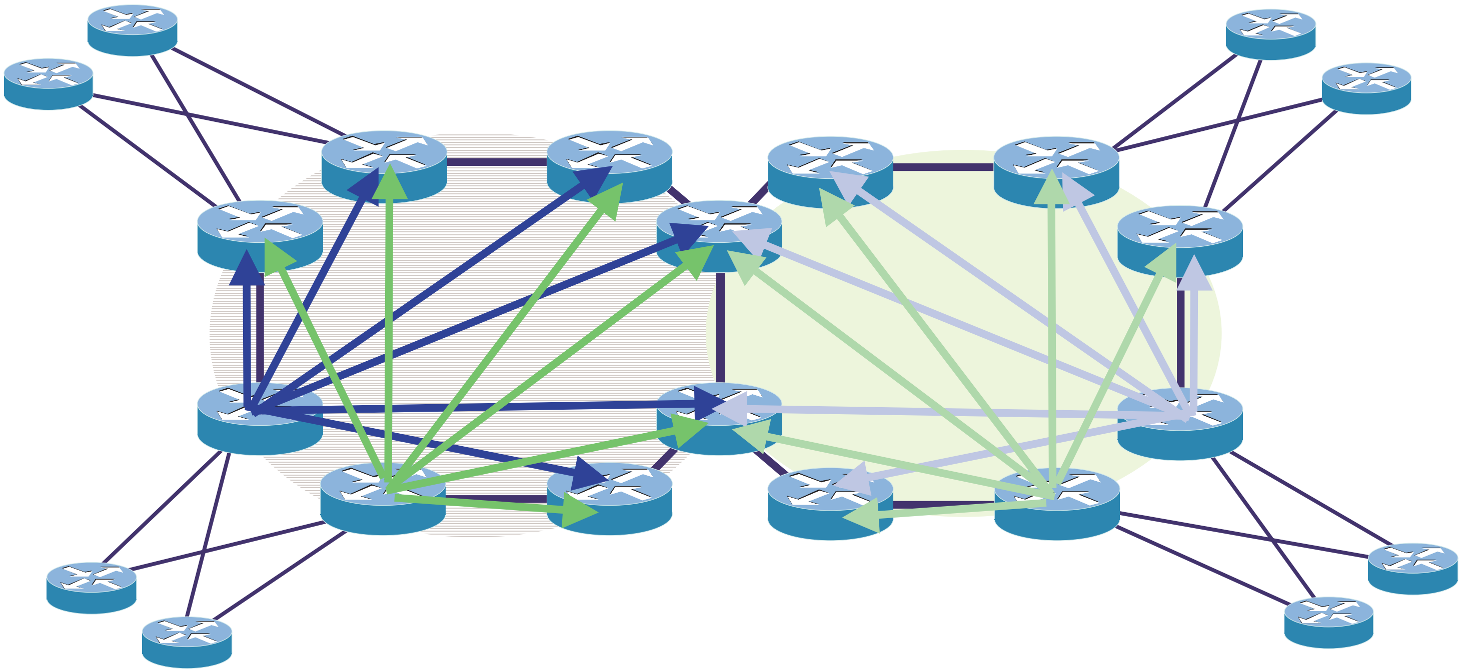
- In failure of link A-C:
  - For traffic from X ➔ Y, router X IGP will now see best path via router B
  - However, if bandwidth is available, tunnel from A to E will be re-established over path A ➔ B ➔ D ➔ C ➔ E
  - Tunnel #2 will not be sized for X ➔ Y demand
  - Bandwidth may be set aside on link A ➔ B for traffic which is now taking different path

Traffic “sloshing”

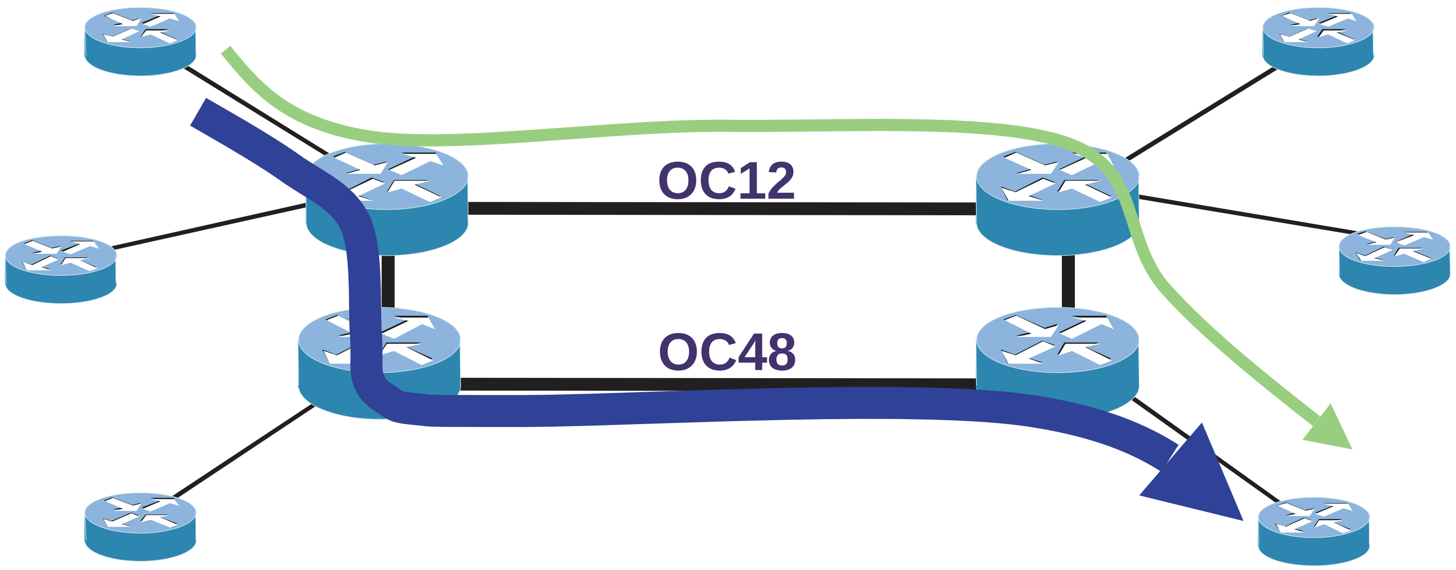


- Forwarding adjacency could be used to overcome traffic sloshing
  - Normally, a tunnel only influences the FIB of its head-end
    - other nodes do not see it
  - With Forwarding Adjacency the head-end advertises the tunnel in its IGP LSP
    - Tunnel #1 could always be made preferable over tunnel #2 for traffic from X ➔ Y

Hierarchical or Regional Mesh



Ad hoc Deployment



- Explicit path configured on head-end for each tunnel to offload traffic from congested links
- Can be useful when faced with:
  - Unexpected traffic demands
  - Long bandwidth lead-times

MPLS TE deployment considerations

- Systematic (strategic) or ad hoc (tactical) deployment
- Statically (explicit) or dynamically established tunnels
  - If dynamic – must specify bandwidths for tunnels
    - Otherwise defaults to IGP shortest path
  - Dynamic tunnels introduce indeterminism
    - Can be addressed with explicit tunnels or prioritisation scheme – higher priority for larger tunnels
- Tunnel sizing and how often to re-optimise?

Tunnel Sizing

- Tunnel sizing is key …
  - Needless congestion if actual load exceeds expected max (even by a little bit)
  - Needless tunnel rejection if reservation > actual
    - Enough capacity for actual but not for the tunnel reservation
    - Traffic reverts to SPF, which is presumably set for latency not for traffic distribution
- … as is the relationship of tunnel bandwidth to QoS
  - Actual heuristic will depend upon dynamicism of tunnel sizing

Tunnel Sizing

- Static (offline) Sizing
  - Statically set reservation to percentile of expected max load (e.g. P95)
  - Periodically readjust – not in real time

Tunnel Sizing

- Dynamic (online) Sizing: autobandwidth
  - Router automatically adjusts reservation (up or down) potentially in near real time based on traffic observed in previous time slot:
    1. Monitor the 5 min average counter (as in show interface command)
    2. keep track of the largest 5 min average over a configurable interval
    3. re-adjusting the tunnel bandwidth based upon the largest 5 min average for that interval
    4. After the interval has expired, the largest 5 min average is cleared (set to 0)
  - Tunnel churn if autobandwidth periodicity high
    - Tunnels de-establish and establish needlessly during the day as links fill up
  - Tunnel bandwidth not persistent

Pipes, Hoses, and Tunnels

Pipe Services

- Point-to-point commodity
  - Defined ICR and ECR between two specified points
- TE bandwidth based upon sold ICR / ECR
- Less Risk of Traffic-Tunnel Size Mismatch

Hose Services

- Point-to-multipoint commodity
  - Defined ICR and ECR to cloud
- TE bandwidth based upon monitored load
- More Risk of Traffic-Tunnel Size Mismatch

- Always OK to use Offline Explicit or Metric-Based TE

TE Summary

- Strategic TE important to resilience and cost savings
- Computer-Aided Metric-Based TE is a new option
- MPLS TE has many deployment considerations
- Metric-Based TE close to theoretical optimum, even under failure conditions

Convergence

I. Traffic Characterization

II. Traffic Matrices

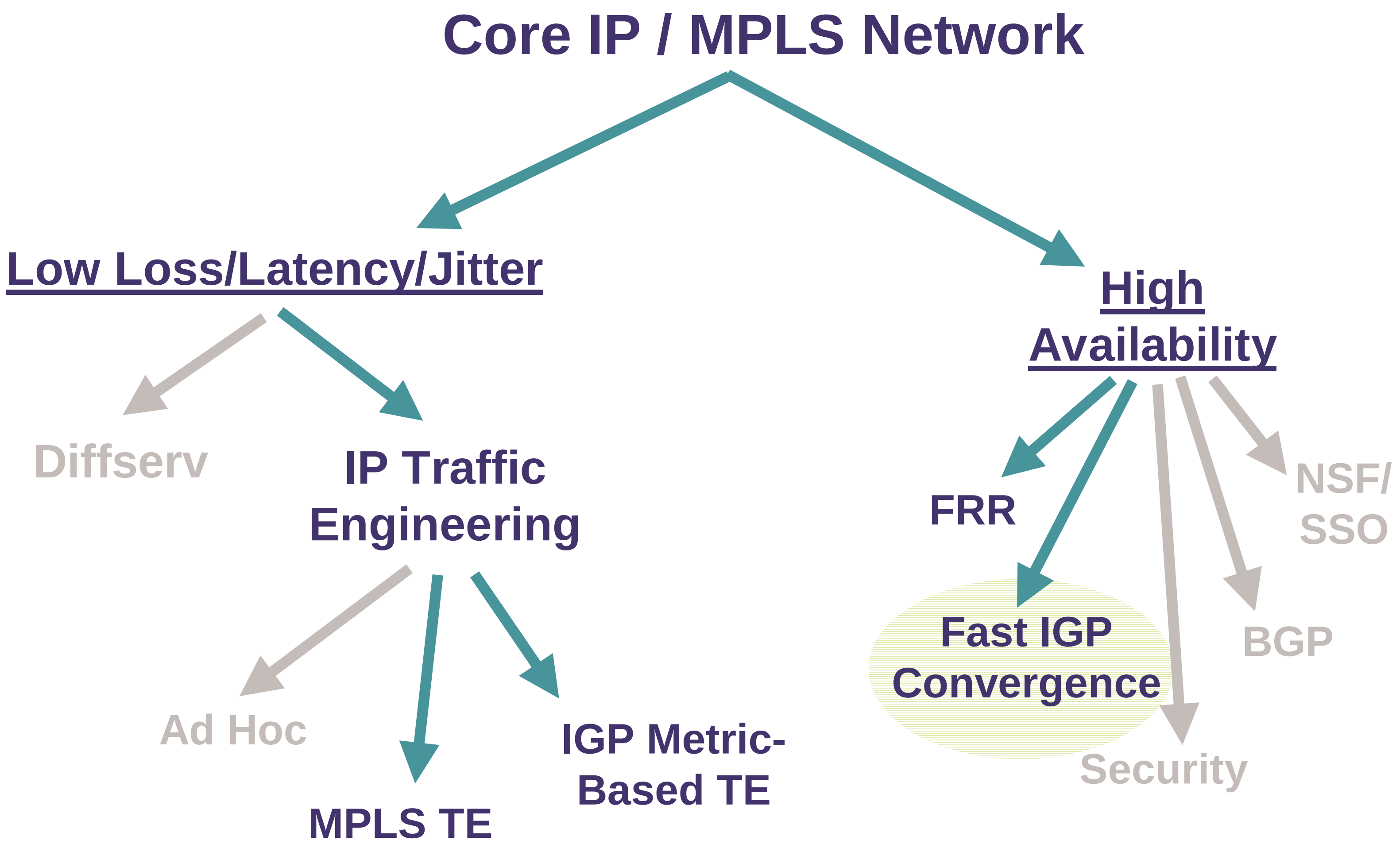
III. TE Introduction

IV. Metric-Based TE

V. Convergence

- Fast SPF Convergence
- Fast Reroute

Options for IP Traffic engineering



Core IP / MPLS Network

Low Loss/Latency/Jitter

Diffserv

IP Traffic Engineering

Ad Hoc

MPLS TE

IGP Metric-Based TE

High Availability

FRR

Fast IGP Convergence

Security

BGP

NSF/ SSO

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 conference session on “Fast IGP Convergence”, Wednesday 25 February 16:00-16:30

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 a maximum bound for loss of connectivity

MPLS TE Fast Reroute (FRR)

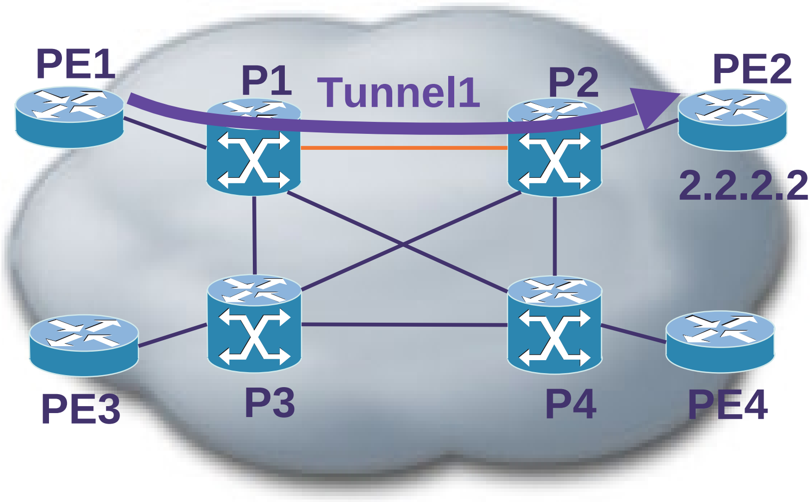
- If …
  - recovery around failures is needed in few 100s of ms
  - or time to reroute around a failure needs to be more deterministic
- Then …
  - MPLS TE fast reroute is required
- MPLS TE FRR is faster and more deterministic than IGP convergence

MPLS TE FRR link/node protection

- FRR uses local detection and protection at the point of failure
  - Use POS for rapid detection
  - Fast local protection at the point of failure: in ms
  - 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

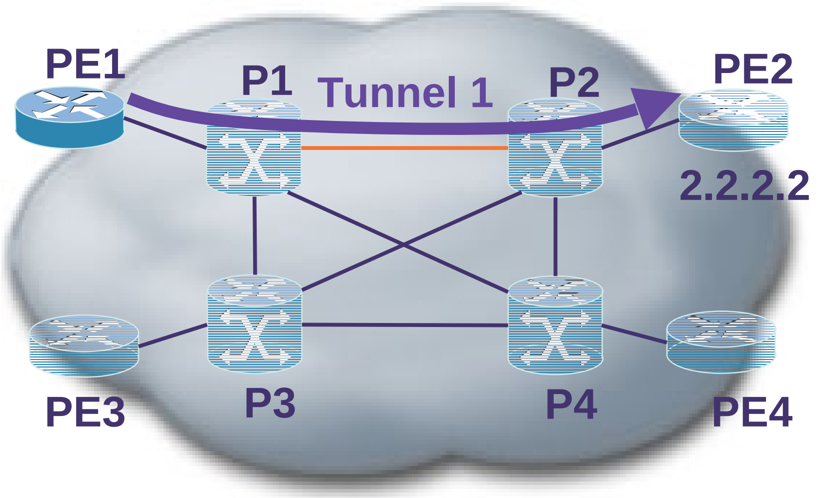


Resilience Strategy: two pronged approach

- FRR allows for temporary protection of TE LSPs affected by a link/node 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 (100 milliseconds)
    - May be sub-optimal
  - Path restoration
    - Repair made at the head-end
    - An optimized long term repair
    - Slower—O (seconds)

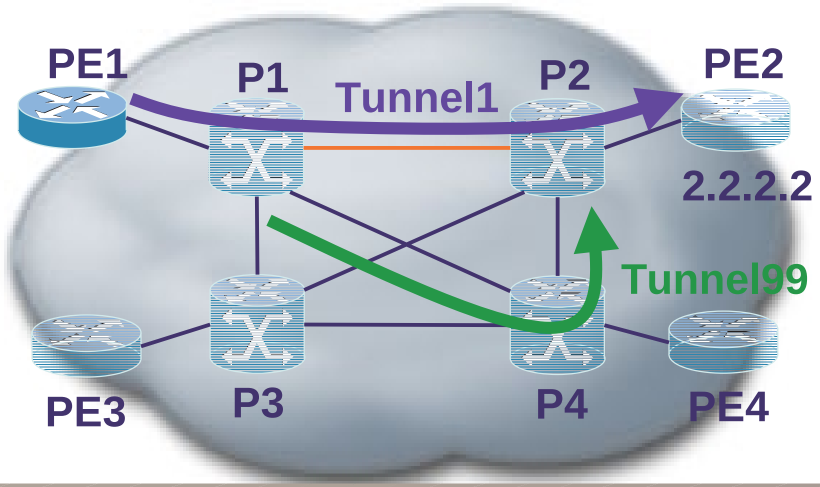
FRR Refresher (1)

- 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

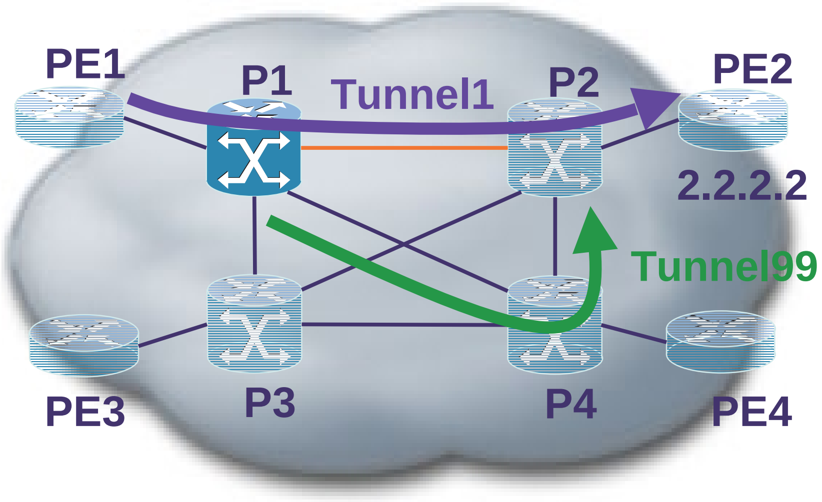


- Explicitly routed back-up Tunnel99 is configured on P1 to P2 via P4
- No “tunnel mpls traffic-eng autoroute announce” !
  - –The back-up tunnel MUST only be used when a failure occurs

```
(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]
```

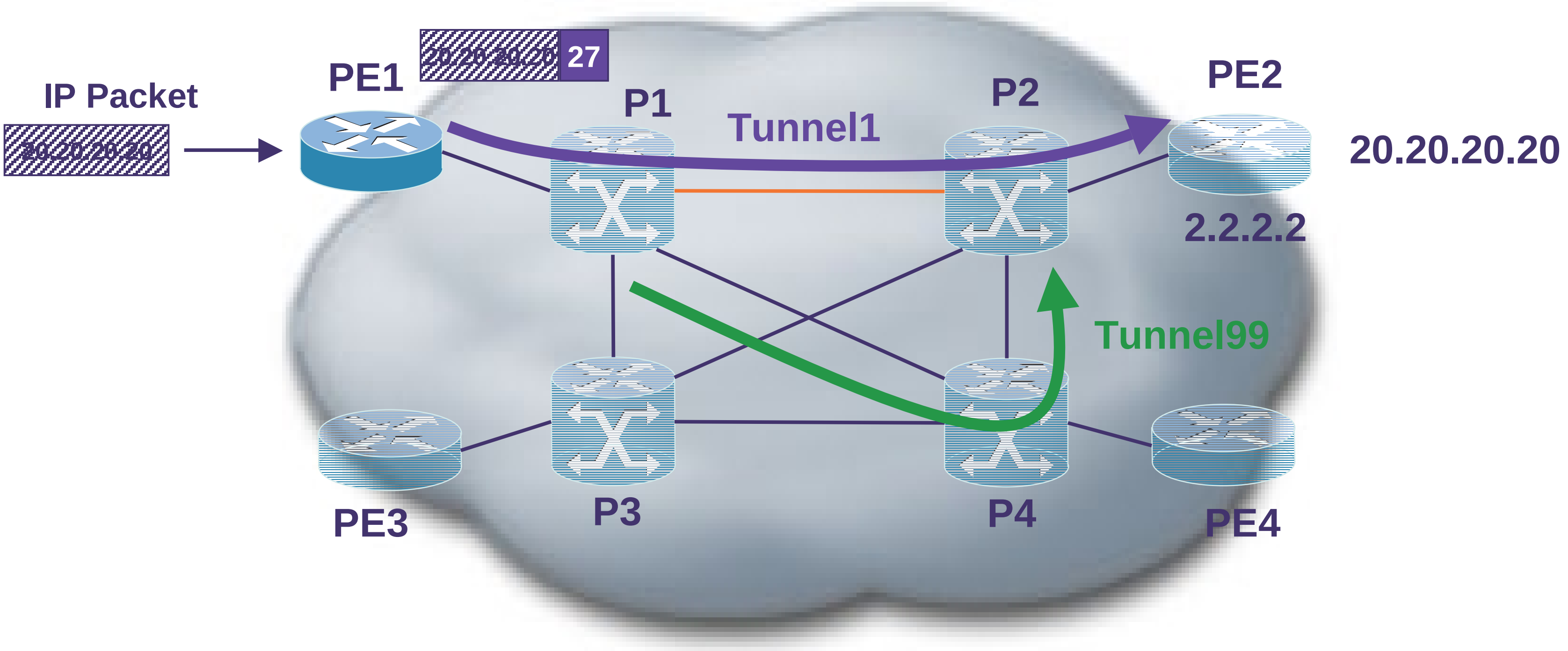
FRR Refresher (3): Configuration

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



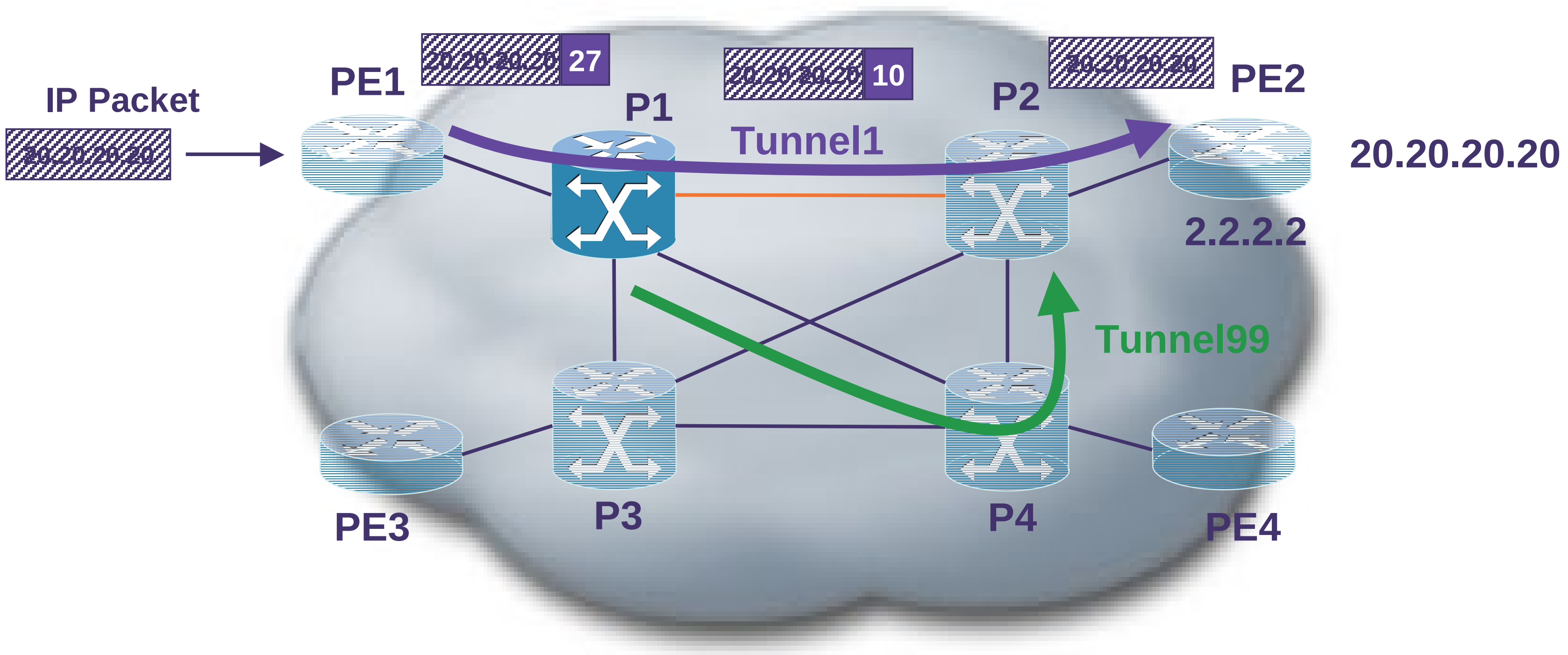
```
(config)# interface POS2/0
(config-if)# description Link to P2
(config-if)# ip address 10.0.12.2 255.255.255.252
(config-if)# mpls traffic-eng tunnels
(config-if)# ip rsvp bandwidth 150000 150000 sub-pool 30000
(config-if)# mpls traffic-eng backup-path Tunnel99
(config-if)# pos ais-shut
```

FRR Refresher (3): before failure



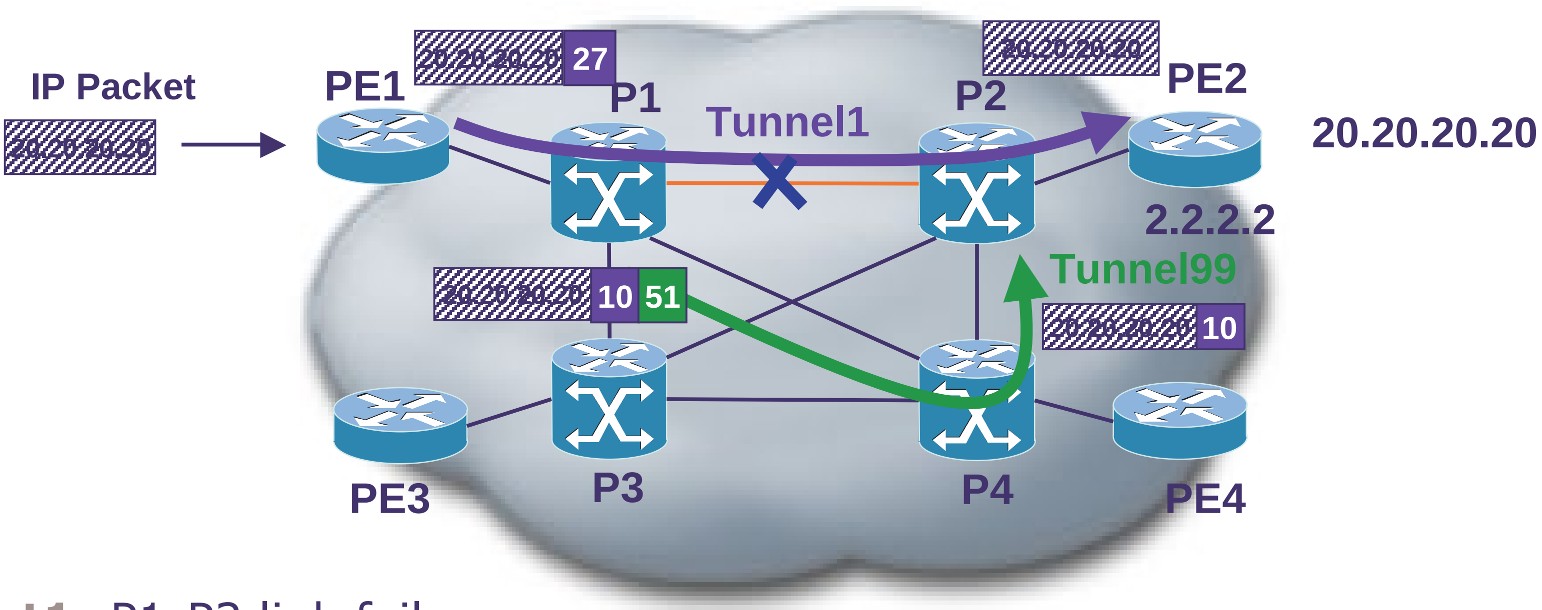
```
PE1# sh tag for 20.20.20.20
     Local  Outgoing    Prefix             Bytes tag  Outgoing    Next Hop
     tag    tag or VC   or Tunnel Id       switched   interface
     28     27          1.1.1.1/32         0          TU1         point2point
```

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

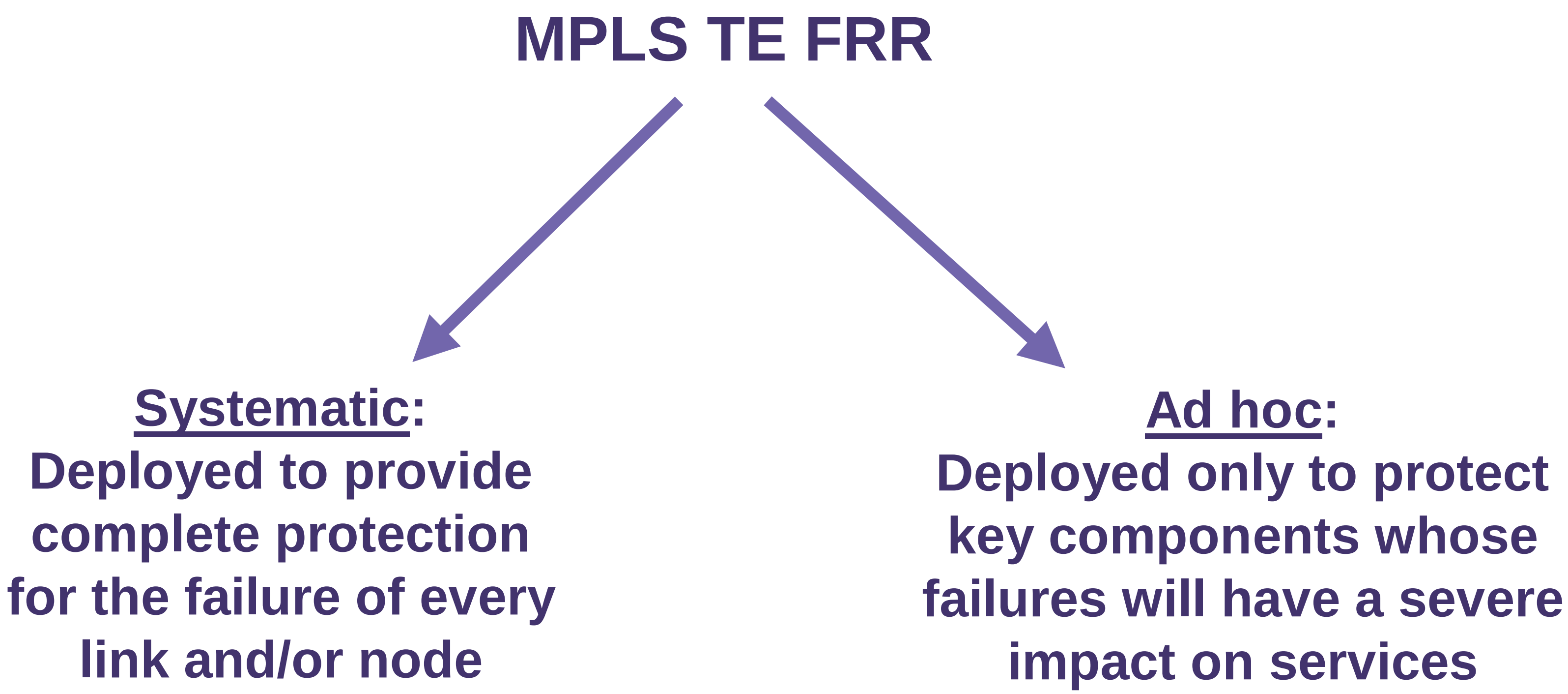
MPLS TE FRR

- 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?

- For telephony users:
  - If the connectivity is lost for >150ms, a glitch may be perceived
    - 150ms equates to at least 2 lost samples for 50ms packetisation interval
  - If the loss of connectivity lasts for several seconds, the phone call may be dropped
- Hence FRR required where very tight SLAs are required
  - Allows highest availability of service to be offered for VoIP class

MPLS TE FRR – deployment scenarios



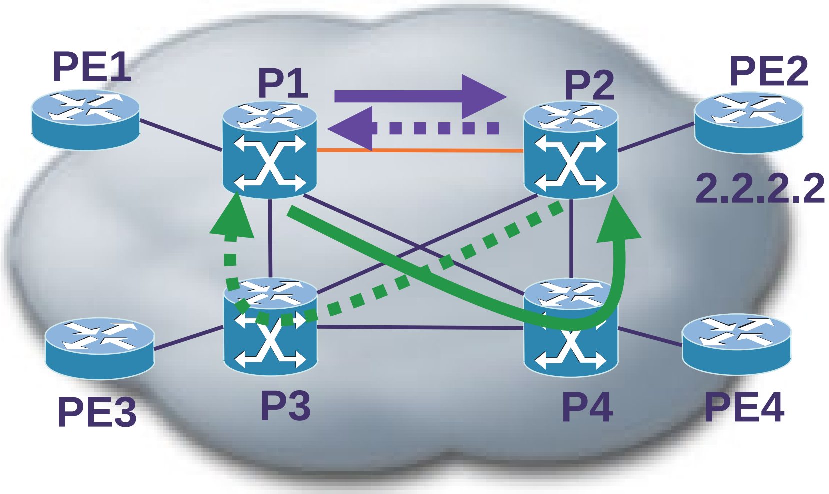
MPLS TE FRR

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

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

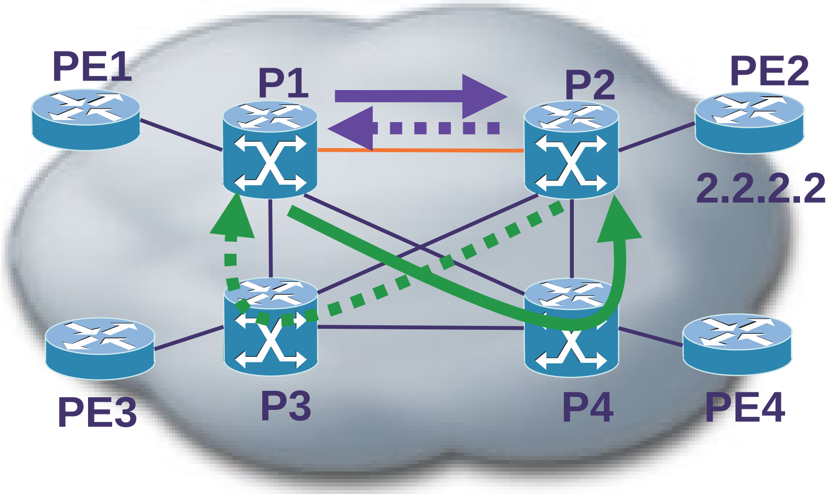
MPLS TE FRR – deployment scenarios

- Full mesh of TE tunnels is not needed for systematic approach
- Can instead use next-hop (NH) tunnels on every link
  - Single hop 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



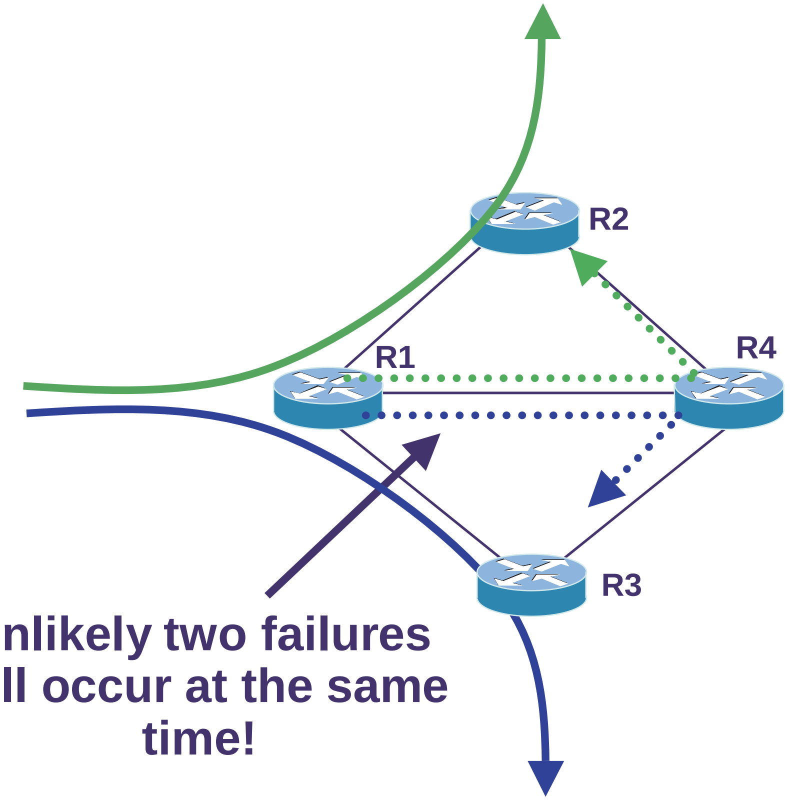
MPLS TE FRR – deployment scenarios

- Allows FRR to be used for link protection without needing a TE full mesh
  - Recovery time becomes a function of number of LSPs / prefixes
- Can similarly use next-next-hop (NNH) tunnels to protect every node
- Allows decisions on need for TE and FRR 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



Unlikely two failures will occur at the same time!

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 may be useful to figure out more optimal configurations under different link/node failure scenarios

Convergence Summary

- Number of technologies to increase core convergence and hence core network availability
  - IGP fast convergence
    - Where recovery in < ~1s 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

- Traffic Characteristics
  - Long term is smooth and predictable
  - Uncorrelated microbursts
  - High utilization with little delay at high capacities
  - Little need for dynamic routing or queue management
- Simple++
  - Traffic Matrix (Measure, or Estimate)
  - Capacity plan based on failure simulation
  - TE without Layer 2 Overlay
    - Computer-Aided Metric-Based TE ≈ as Efficient of Theoretical Optimum (though more scalable)
- Multiple Routes to High Availability
  - Fast Reroute
  - Fast Convergence

Traffic Engineering References

- B. Fortz, J. Rexford, and M. Thorup, “Traffic Engineering With Traditional IP Routing Protocols” in 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.
- Cariden “IGP Traffic Engineering Case Study”, Cariden Technologies, Inc., October 2002.
- B. Fortz and M. Thorup, “Internet traffic engineering by optimizing OSPF weights” in Proceedings of IEEE INFOCOM, March 2000.
- B. Fortz and M. Thorup, “Optimizing OSPF/IS-IS weights in a changing world” IEEE Journal on Selected Areas in Communications, volume 20, pp. 756-767, May 2002.
- 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.

Traffic Characterization References

[1] Steve Casner, Cengiz Alaettinoglu and Chia-Chee Kuan, *A Fine-Grained View of High-Performance Networking*, NANOG 22 http://www.nanog.org/mtg-0105/casner.html

[2] Chris Liljenstolpe, *Design Issues in Next Generation Carrier Networks*, MPLS 2001 Conference

[3] Peter Lothberg, *A View of the Future: The IP-Only Internet*, NANOG 22, http://www.nanog.org/mtg-0105/lothberg.html

[4] Zafer Sahinoglu and Sirin Tekinay, *On Multimedia Networks: Self-Similar Traffic and Network Performance*, IEEE Communications Magazine, January 1999

[5] Robert Morris and Dong Lin, *Variance of Aggregated WebTraffic*, IEEE INFOCOM 2000, Tel Aviv, March 2000, pages 360-366.

[6] Anna Charny, Jean-Yves Le Boudec, *Delay bounds in networks with aggregate scheduling*, April 14 2001.

[7] Thomas Bonald, et al, *Statistical Guarantees for Streaming Flows Using Expedited Forwarding*, INFOCOM 2001.

[8] Roberts *Traffic Theory and the Internet*, IEEE Communications Magazine, January 2001.

[9] Jin Cao, William S. Cleveland, Don X. Sun, *A Statistical Model for Allocating Bandwidth to Best-Effort Internet Traffic*, to appear in Statistical Science, 2004

[10] Chuck Fraleigh, Fouad Tobagi, Christophe Diot, *Provisioning IP Backbone Networks to Support Latency Sensitive Traffic*, Proc. IEEE INFOCOM 2003, April 2003

[11] Cao, J., W.S. Cleveland, D. Lin, D.X. Sun,*Internet Traffic Tends Towards Poisson and Independent as the Load Increases.* In *Nonlinear Estimation and Classification,* New York, Springer-Verlag, 2002