



# Best Practices in Network Planning and Traffic Engineering

Nanog 52 – Denver, CO  
12 April, 2011



**(Clarence Filsfils – Cisco Systems)**

**Arman Maghbouleh – Cariden Technologies**

**Paolo Lucente – pmacct**

# Outline

- Objective / Intro [CF\*]
- Traffic Matrix [CF]
  - pmacct [PL]
- Network Planning [AM]
- Optimization/Traffic Engineering [AM]
- Planning for LFA FRR [CF]
- IP/Optical Integration [Skipped]
- A final example [AM]
- Conclusion & References

\* CF not in Denver, AM presenting slightly modified CF material.



# Introduction & Objective



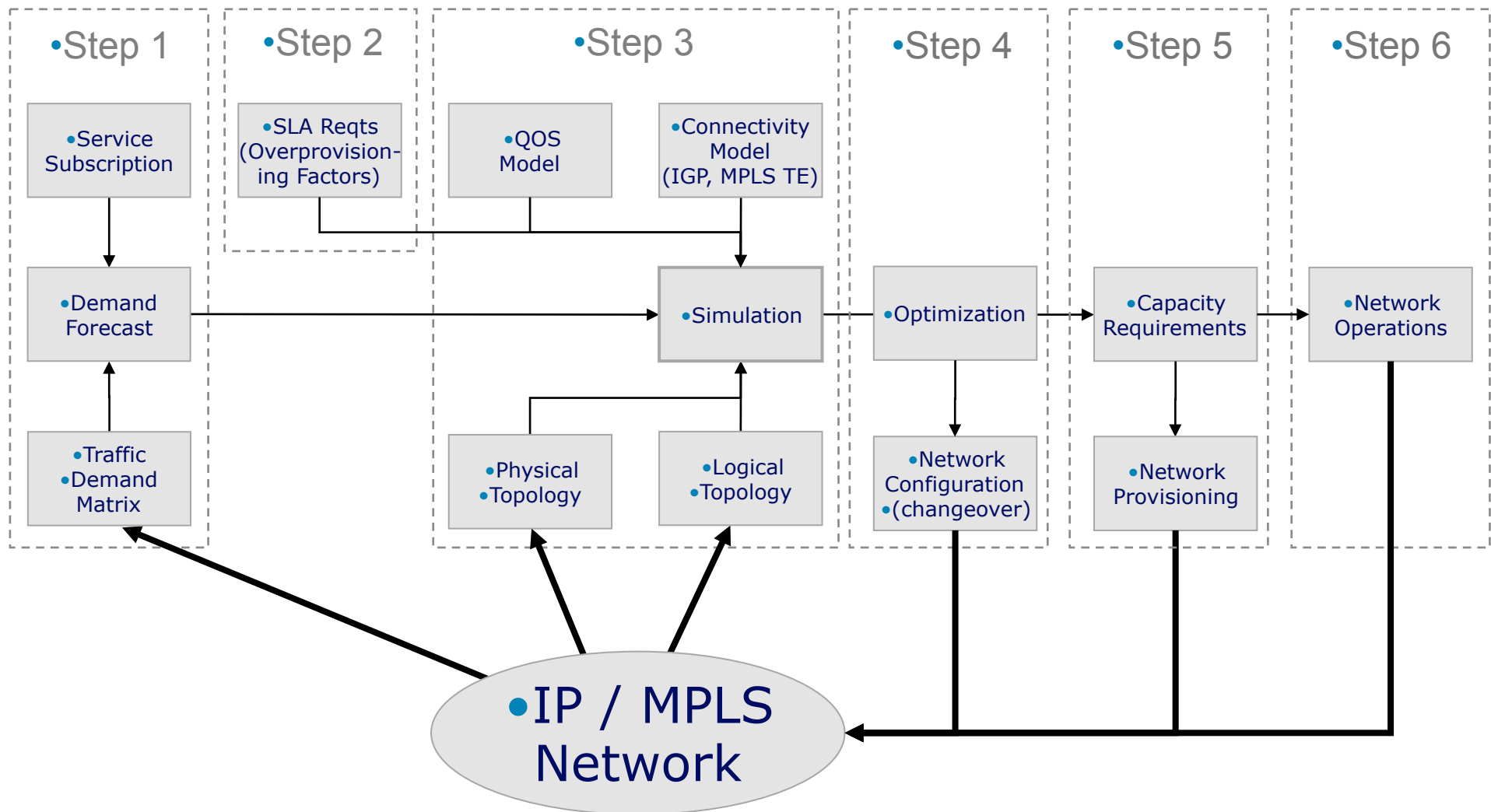
# Objective

- SLA enforcement
  - expressed as loss, latency and jitter availability targets
- How is SLA monitored
  - PoP to PoP active probes
  - Per-link or per-class drops
- How to ensure
  - **Ensure that capacity exceeds demands frequently enough to achieve availability targets**
  - Highlight: catastrophic events (multiple non-SRLG failures) may lead to “planned” congestion. The planner decided not to plan enough capacity for this event as the cost of such a solution outweighs the penalty. A notion of probability and risk assessment is fundamental to efficient capacity planning.

# Basic Capacity Planning

- Input
  - Topology
  - Routing Policy
  - QoS policy per link
  - Per-Class Traffic Matrix
- Output
  - Is Per-class Per-link OPF < a target threshold (e.g. 85%)?  
OPF: over-provisioning factor = load/capacity
- If yes  
then be happy  
else either modify inputs  
or the target output threshold  
or accept the violation

# Overall Process





# Topology/Routing



# Infrastructure

- Sources of information
  - ISIS/OSPF LS Database -> L3 connectivity
  - Configs -> Network as designed
  - SNMP/Show -> Network as operating
- Easy
  - IGP topology
  - Interface counters
- Non-Trivial
  - RSVP and LDP
  - Multicast
  - Qos
- Challenges
  - Port channel mapping
  - Optical topology
  - Physical infrastructure
  - IPv6 (lack of differentiated counters)
  - Multidomain services (across multiple areas, networks)
  - Multitopology Routing
  - L2 Infrastructure
  - External devices (Media gateways, server farms)
  - External effects (e.g., Akamai traffic shift)



# Routing Policy – Primary Paths

- ISIS/OSPF
  - Simple: Dijkstra based on link costs
- Dynamic MPLS-TE
  - Complex because non-deterministic
- Static MPLS-TE
  - Simple: the planning tool computes the route of each TE LSP  
“Simple” from a planning viewpoint but less flexible (higher opex) operationally. There is no free lunch.
- Multicast
- BGP

# Routing Policy – Backup Paths

- ISIS/OSPF – Routing Convergence
  - Simple: Dijkstra based on link costs
- ISIS/OSPF - LFA FRR
  - Complex: the availability of a backup depends on the topology and the prefix, some level of non-determinism may exist when LFA tie-break does not select a unique solution
- Dynamic MPLS-TE – Routing Convergence
  - Complex because non-deterministic
- Dynamic MPLS-TE – MPLS TE FRR via a dynamic backup tunnel
  - Complex because the backup LSP route may not be deterministic
- Dynamic MPLS-TE – MPLS TE FRR via a static backup tunnel
  - Moderate: the planning tool computes the backup LSP route but which primary LSP's are on the primary interface may be non-deterministic
- Static MPLS-TE – MPLS TE FRR via static backup tunnel
  - Simple: the planning tool computes the route of each TE LSP (primary and backup) (reminder... there is a trade-off to this simplicity.

# QoS policy per-link

- Should be made simple
  - network-wide BW allocation policy
  - should rarely change
  - rarely customized per link for tactical goals

# Over-Provision Factor (Utilization Threshold)

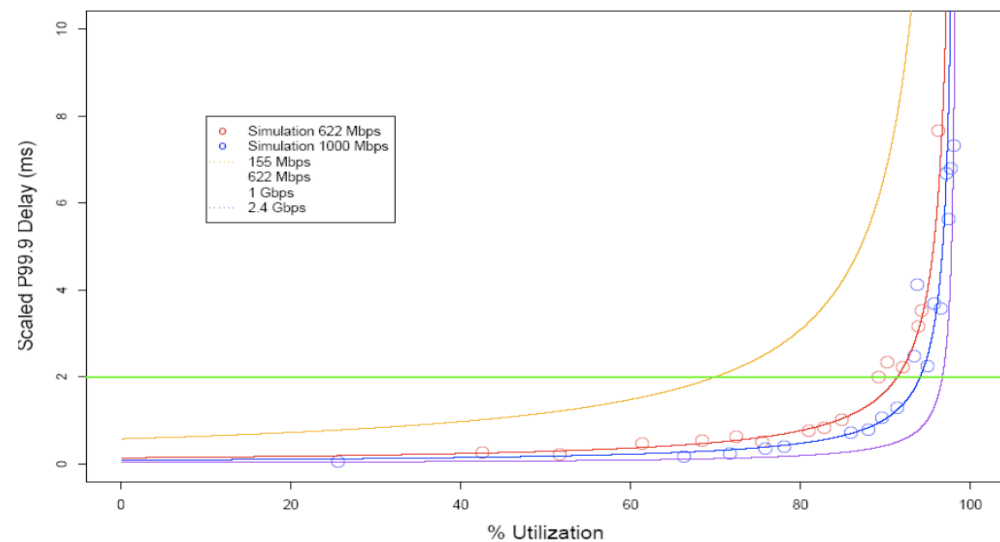
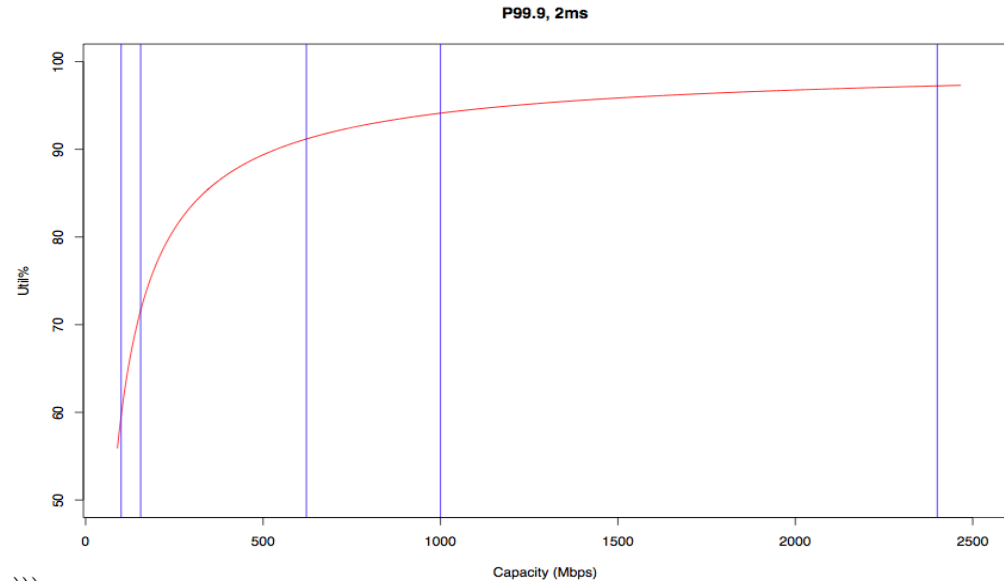
- Area of continued research
- Common agreement: Most services (including voice) tolerate up to [80-90%] utilization when underlying capacity is >1Gbps
  - as long as traffic is a large mix of independent flows

# Over-Provision Factor – Research

- Bandwidth Estimation for Best-Effort Internet Traffic
  - Jin Cao, William S. Cleveland, and Don X. Sun
  - [Cao 2004]
- Data:
  - BELL, AIX, MFN, NZIX
- Best-Effort Delay Formula:

$$\log_2(u) = o + (o_c + o_{\tau\delta}) \log_2(c) + o_{\tau\delta} \log_2(\gamma_b \delta) + o_{\omega}(-\log_2(-\log_2(\omega))).$$

- Similar queueing simulation results [Telkamp 2003/2009]:



# Digression – Role of QoS

- Link = 10Gbps, Load 1 is 2Gbps, Load 2 is 6Gbps
- Class1 gets 90%, Class2 gets 10%, work-conservative scheduler
- Utilization (Class1) =  $2/9 = 22\%$  <<<< 85% (no risk!)
- Utilization (Class2) =  $6/8 = 75\%$  and actually even worse if Class1 gets more loaded than expected. Much closer to the 85% target and hence much more risky!
- But fine because the availability target for Class2 is looser than Class1 (eg. 99% vs 99.999%)
- QoS allows to create excellent OPF for the Tightest-SLA classes at the expense of the loose-SLA classes.
- More details in [Filsfils and Evans 2005] and in [Deploy QoS]

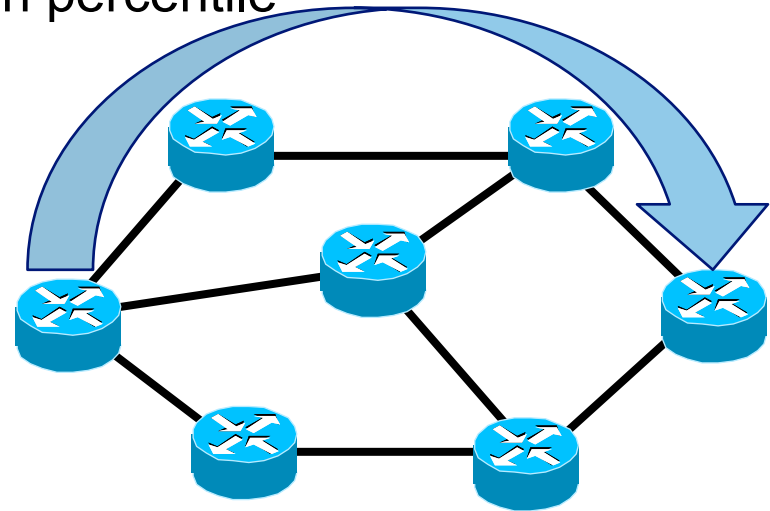


# Traffic Matrix



# Traffic Demand Matrix

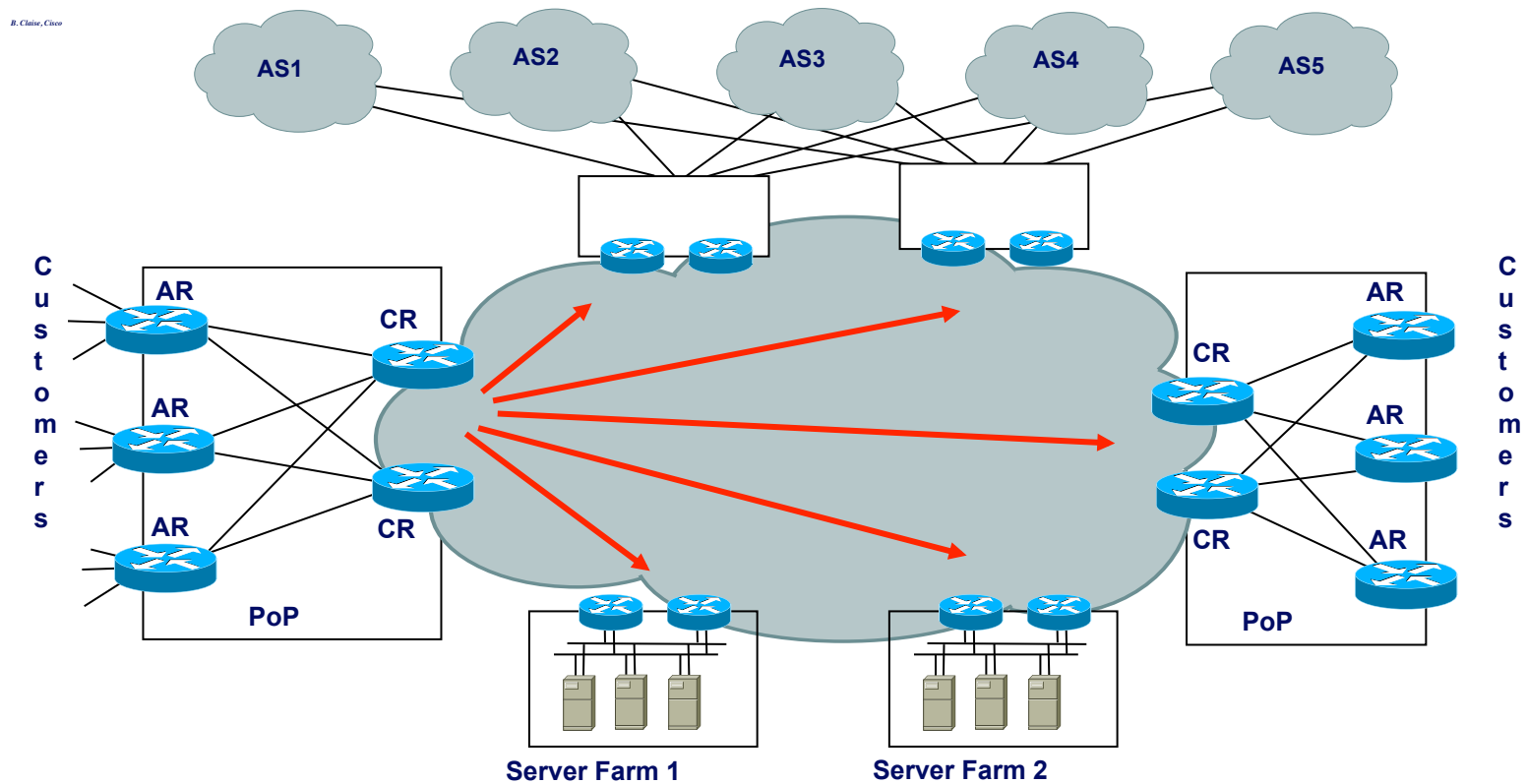
- Traffic demands define the amount of data transmitted between each pair of network nodes
  - Typically per Class
  - Typically peak traffic or a very high percentile
  - Measured, anticipated, or estimated/deduced





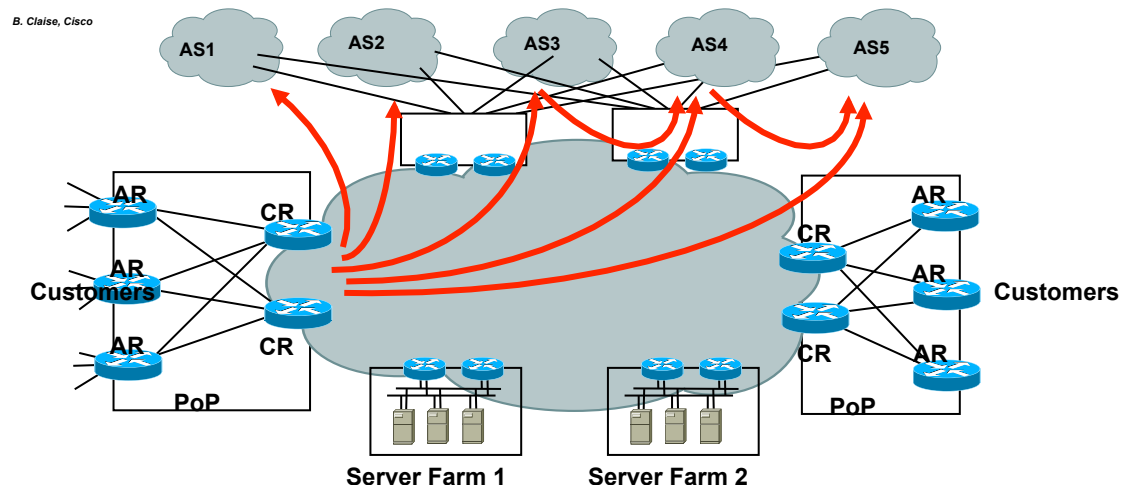
# Internal Traffic Matrix

- POP to POP, CR-to-CR, or AR-to-AR



# External Traffic Matrix

- Router (AR or CR) to External AS or External AS to External AS (for transit providers)
- Useful for analyzing the impact of external failures on the core network
- Peer-AS sufficient for capacity planning and resilience analysis, See RIPE presentation on peering planning [Telkamp 2006]



# Measuring the Traffic Matrix

## Flows

- NetFlow
  - v5

Resource intensive for collection and processing

Non-trivial to convert to Traffic Matrix
  - v9

BGP NextHop Aggregation scheme provides almost direct measurement of the Traffic Matrix

CoS ready

Not universal support

CF method of choice
  - Inaccuracies

Stats can clip at crucial times

NetFlow and SNMP timescale mismatch

## MPLS LSPs

- LDP
  - $O(N^2)$  measurements

Missing values  
(expected when making tens of thousands of measurements)

Can take minutes  
(relevant for quick response, TE)
  - Internal matrix only
  - Not per class
  - Inconsistencies in vendor implementations
- RSVP-TE
  - Requires a full mesh of TE tunnels
  - Internal matrix only
  - Issues with  $O(N^2)$ : missing values, time, ...
  - Not per-class information (unless meshes per class, rare today)

## Difference of Opinion: CF/AM\*

AM][CF



## Stand back-to-back but look different directions

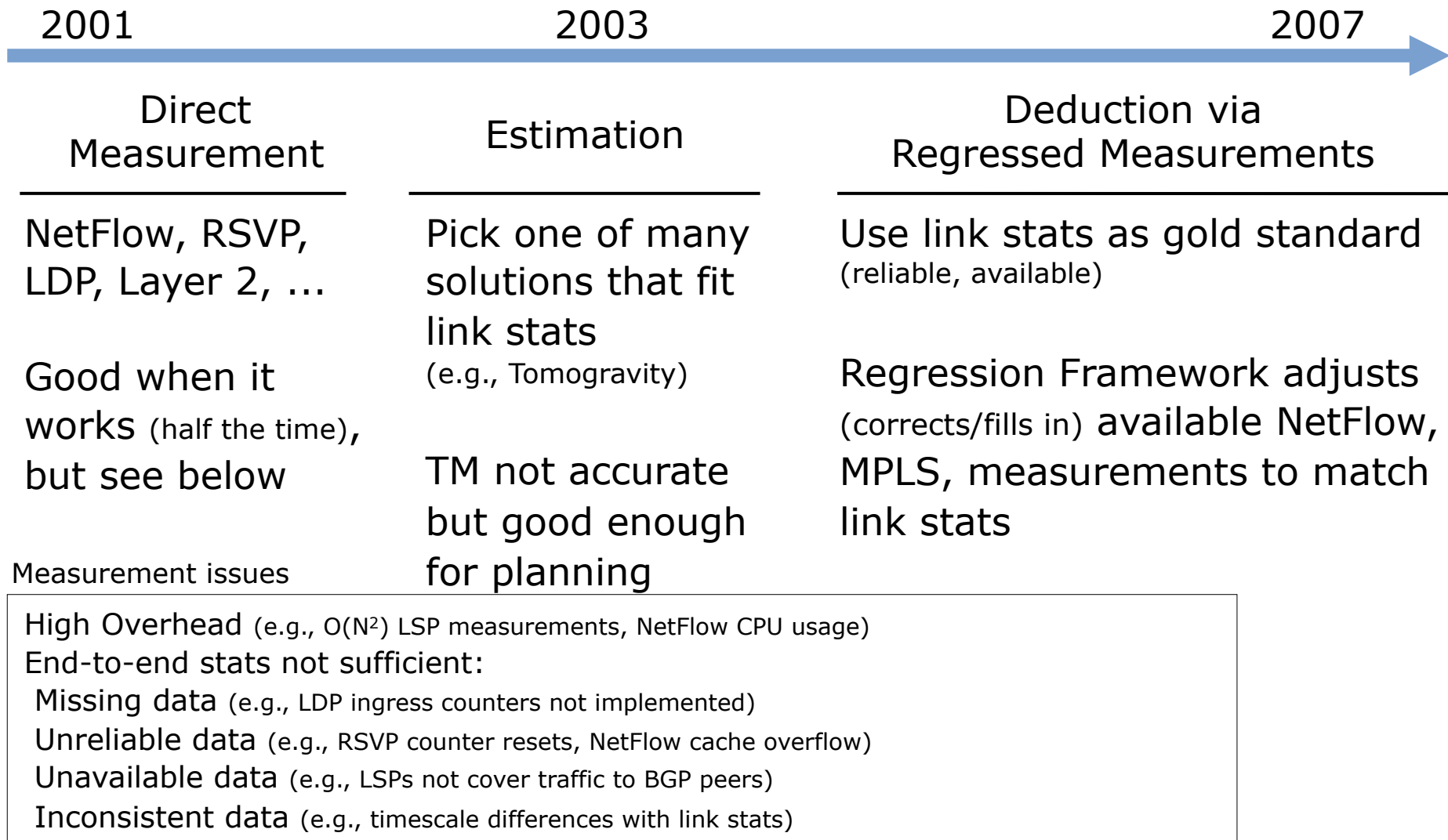
## AM (beaten by reality)

- Looking at what was bought ten years ago (e.g. 7600 sup1)
- Process must work day-in-day-out even after the engineer has moved on
- Giant ships turn slowly (even when the captain knows the way)

## CF (man of vision)

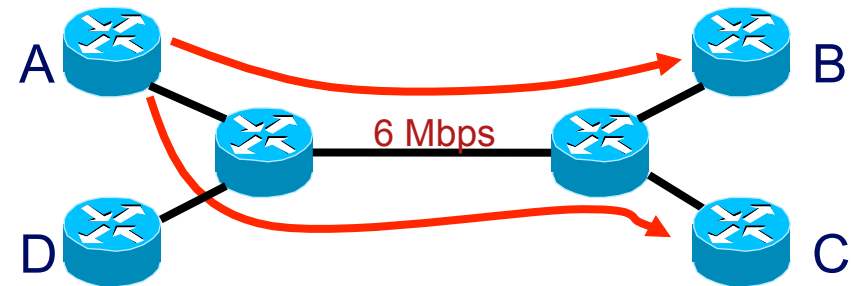
- Looking at what is on the market & coming (e.g., ASR 99M)
- Do neat things with smart people once or twice
- We can make changes, improve things  
(AM snide comment: CF has not dealt with IT:-)

# IP Traffic Matrix Possibilities\*



# Demand Estimation

- Goal: Derive Traffic Matrix (TM) from easy to measure variables
- Problem: Estimate point-to-point demands from measured link loads
- Underdetermined system:
  - N nodes in the network
  - O(N) links utilizations (known)
  - O(N<sup>2</sup>) demands (unknown)
  - Must add additional assumptions (information)
- Many algorithms exist:
  - Gravity model
  - Iterative Proportional Fitting (Kruithof's Projection)
  - ... etc
- Estimation background: network tomography, tomogravity\*, etc.
  - Similar to: Seismology, MRI scan, etc.
  - [Vardi 1996]
  - \* [Zhang et al, 2004]



y: link utilizations

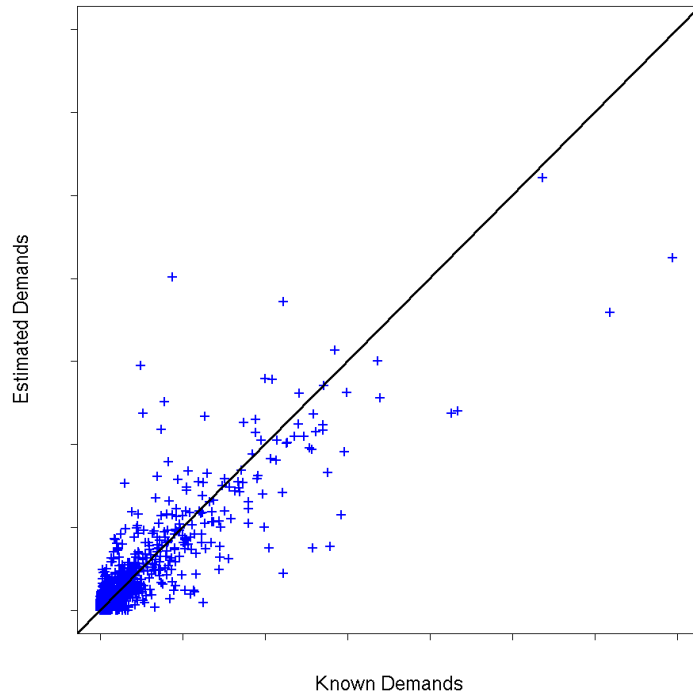
A: routing matrix

x: point-to-point demands

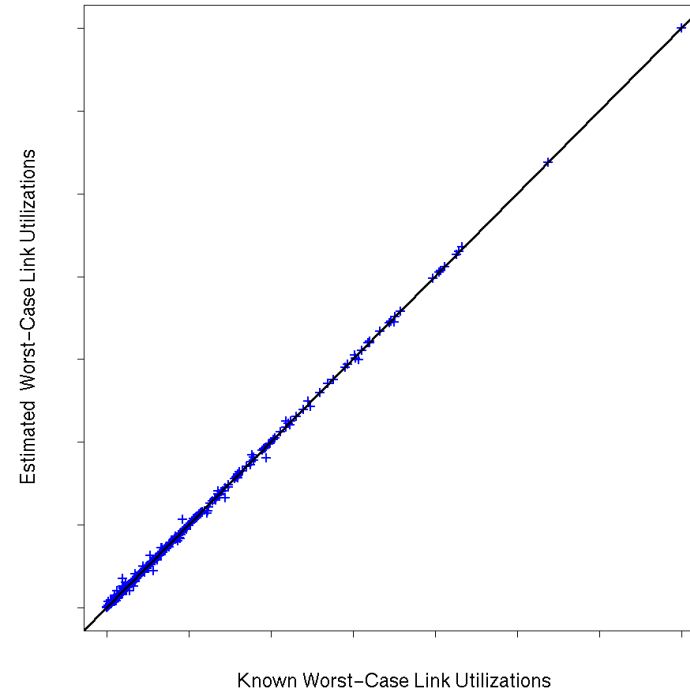
Solve:  $y = Ax$  -> In this example:  $6 = AB + AC$

Calculate the **most likely** Traffic Matrix

# Demand Estimation Results



- Individual demand estimates can be inaccurate

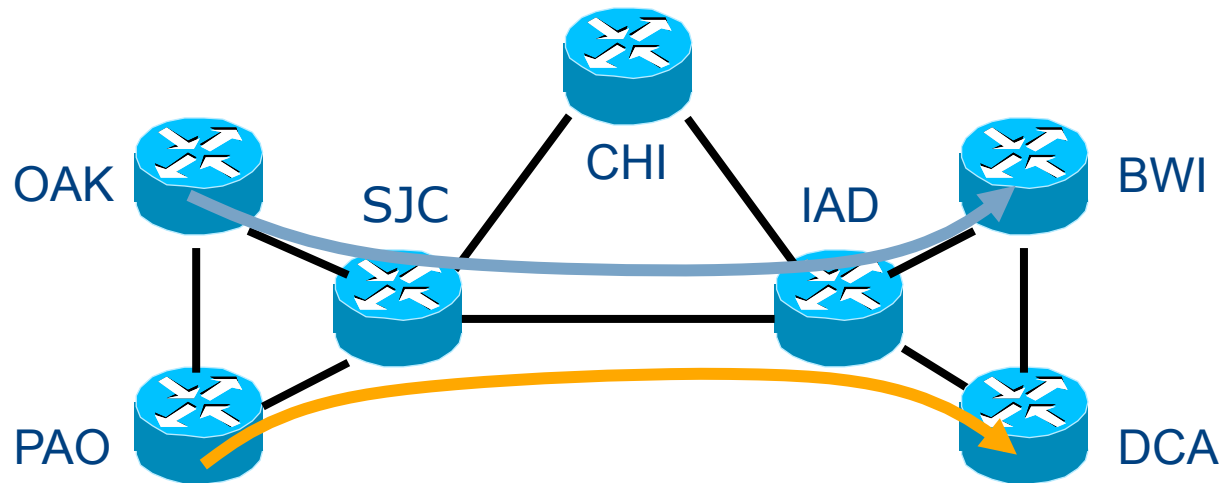


- Using demand estimates in failure case analysis is accurate

*See also [Zhang et al, 2004]: “How to Compute Accurate Traffic Matrices for Your Network in Seconds”*

*Results show similar accuracy for AT&T IP backbone (AS 7018)*

# Estimation Paradox Explained

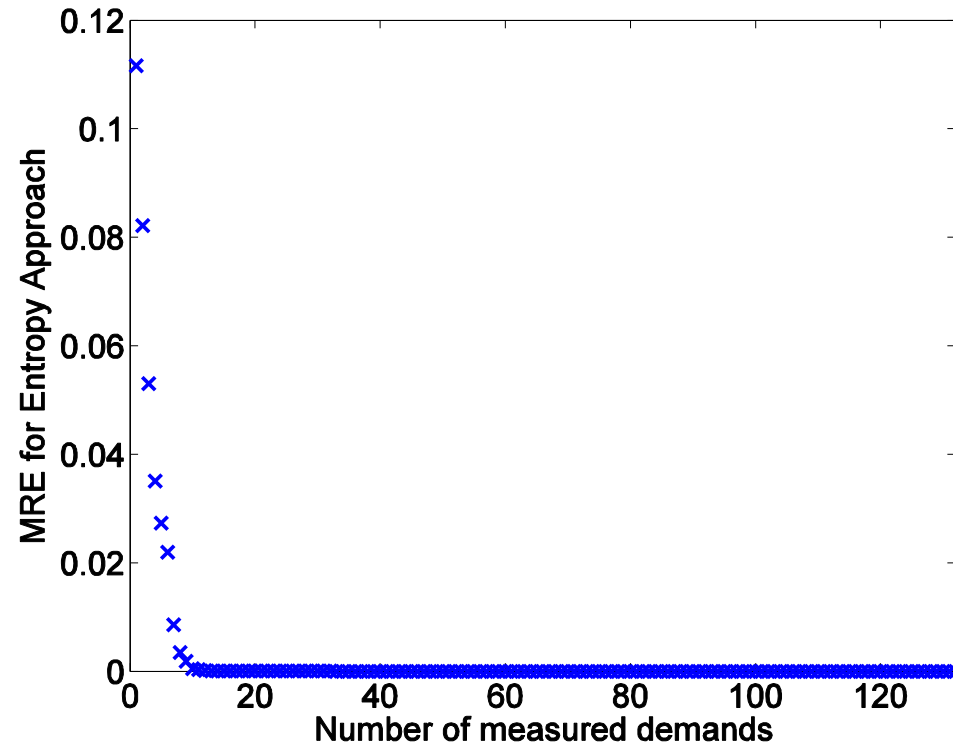


- Hard to tell apart elements
  - OAK->BWI, OAK->DCA, PAO->BWI, PAO->DCA, similar routings
- Are likely to shift as a group under failure or IP TE
  - e.g., above all shift together to route via CHI under SJC-IAD failure



# Role of Netflow, LSP Stats,...

- Estimation techniques can be used in combination with end-to-end measurements
  - E.g. NetFlow or partial MPLS mesh
- Can significantly improve TM estimate accuracy with just a few measurements [Gunner et al]



# Regressed Measurements

- Interface counters remain the most reliable and relevant statistics
- Collect LSP, Netflow, etc. stats as convenient
  - Can afford partial coverage (e.g., one or two big PoPs)
  - more sparse sampling (1:10000 or 1:50000 instead of 1:500 or 1:1000)
  - less frequent measurements (hourly instead of by the minute)
- Use regression (or similar method) to find TM that conforms primarily to interface stats but is guided by NetFlow, LSP stats, etc.

# Regressed Measurements Example

- Topology discovery done in real-time
  - LDP measurements rolling every hour
  - Interface measurement every 2 minutes
  - Regression\* combines the above information
  - Robust TM estimate available every 5 minutes
- 
- (See the DT LDP estimation for another approach for LDP\*\*)

•\*Cariden's Demand Deduction™ in this case( <http://www.cariden.com>)

\*\* Schnitter and Horneffer (2004)

# Forecasted Traffic Matrix

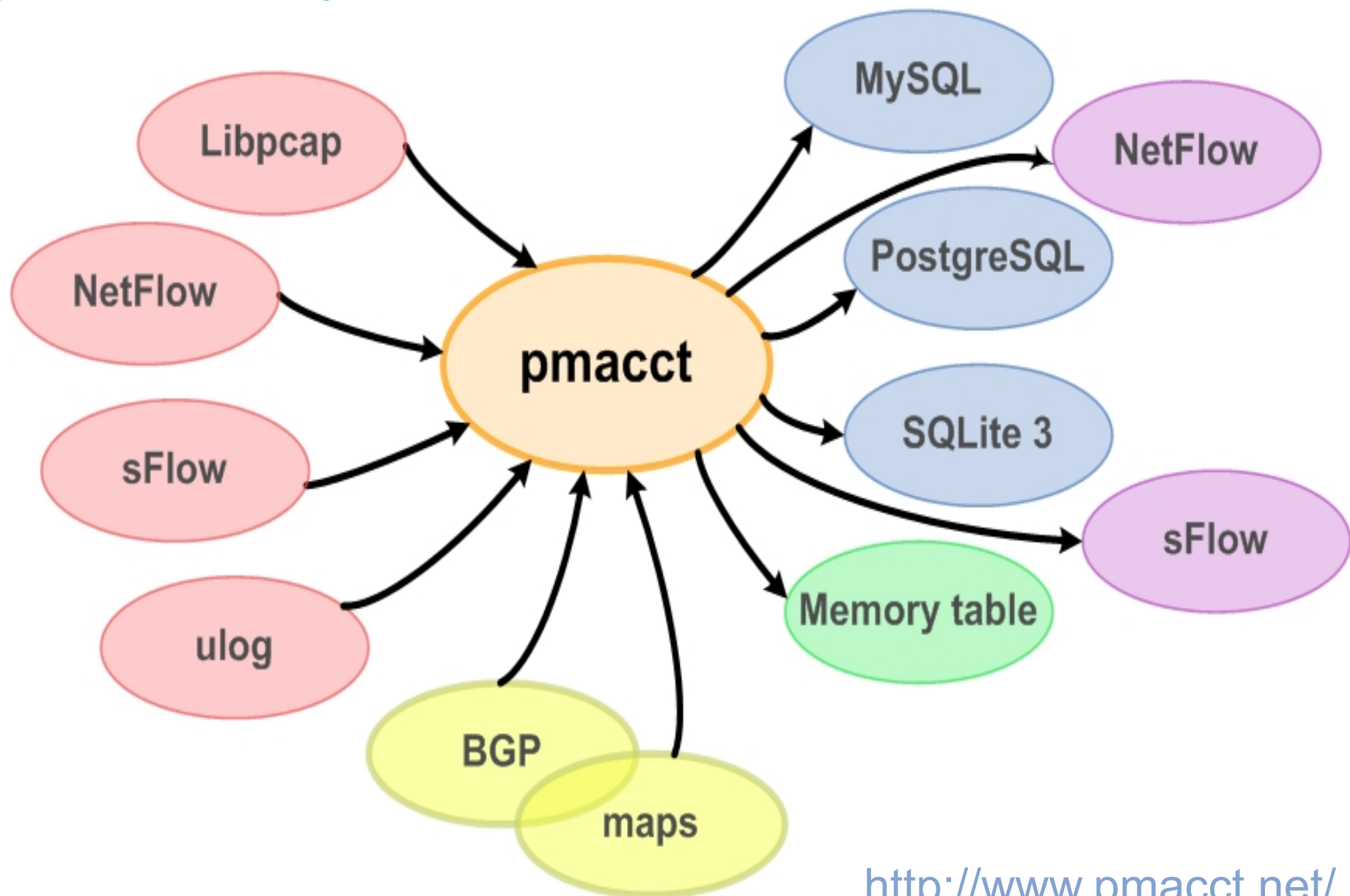
- DWDM provisioning has been slow up to now
  - this will change, see later
- Capacity Planning needs to anticipate growth to add bandwidth ahead of time
  - the slow DWDM provisioning is one of the key reasons why some IP/MPLS networks look “not hot” enough
- Typical forecast is based on compound growth
- Highlight: planning is based on the forecasted TM based on a set of collected TM's



**pmacct**



# pmacct is open-source, free, GPL'ed software



<http://www.pmacct.net/>

# The BGP peer who came from NetFlow (and sFlow)

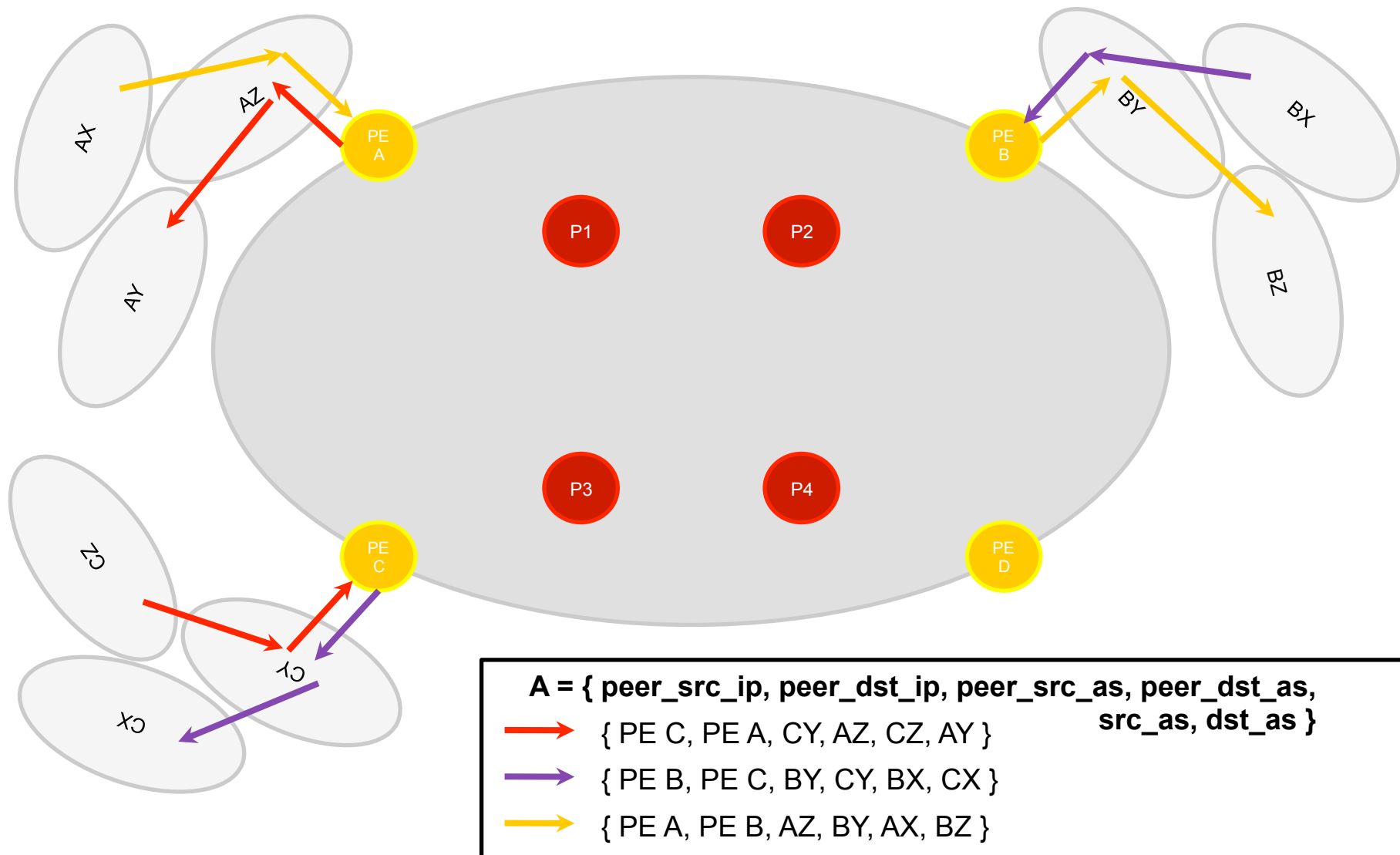
- pmacct introduces a Quagga-based BGP daemon
  - Implemented as a parallel thread within the collector
  - Maintains per-peer BGP RIBs
- Why BGP at the collector?
  - Telemetry reports on forwarding-plane
  - Telemetry should not move control-plane information over and over
- Basic idea: join routing and telemetry data:
  - Telemetry agent address == BGP source address/RID

# Telemetry export models for capacity planning and TE

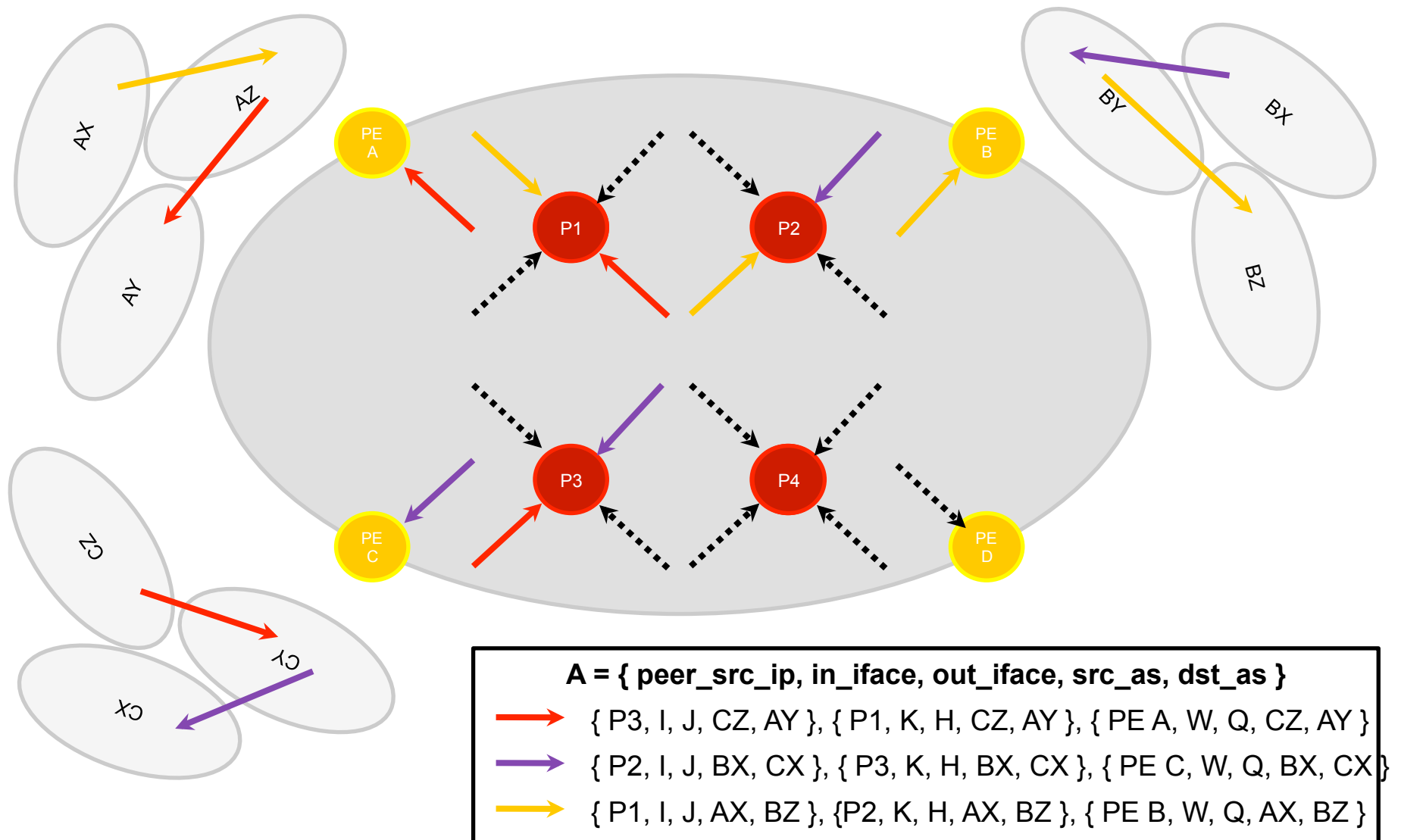
- PE routers: ingress-only at edge interfaces + BGP:
  - Traffic matrix for end-to-end view of traffic patterns
  - Borders (customers, peers and transits) profiling
  - Coupled with IGP information to simulate and plan failures (**strategic solution**)
- P, PE routers: ingress-only at core interfaces:
  - Traffic matrices for local view of traffic patterns
  - No routing information required
  - **Tactical solution** (the problem has already occurred)



# PE routers: telemetry ingress-only at edge interfaces + BGP illustrated

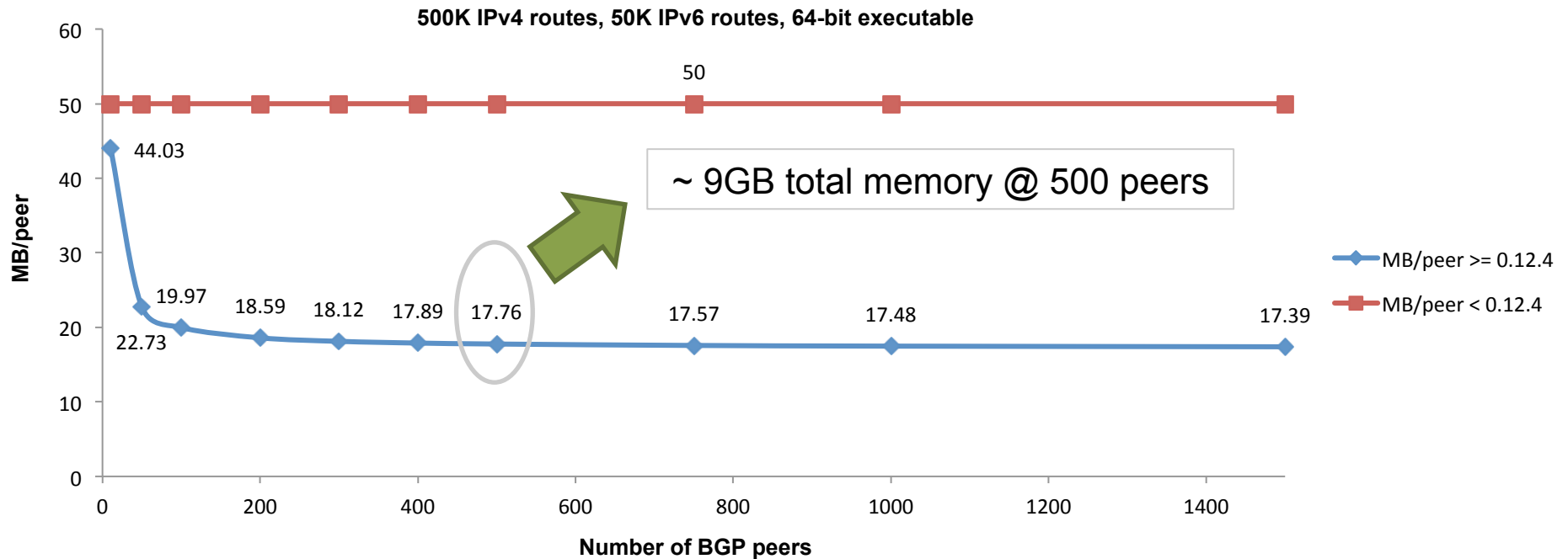


# P, PE routers: telemetry ingress-only at core interfaces illustrated



# Scalability: BGP peering

- The collector BGP peers with all PEs
- Determine memory footprint (below in MB/peer)



# Scalability: aggregation and temporal grouping

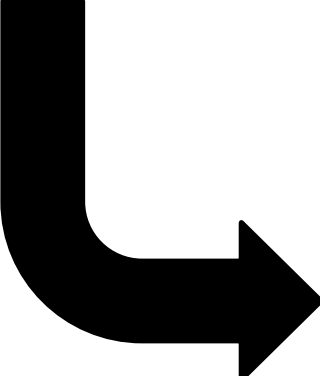
- Flexible spatial and temporal aggregation is:
  - Essential element to large-scale sustainability
  - Original idea underlying pmacct

xacctd.conf:

...

aggregate: peer\_src\_ip, peer\_dst\_ip, peer\_src\_as, peer\_dst\_as, src\_as, dst\_as

sql\_history: 5m



```
acct_5mins_%Y%m%d_%H (  
  id int(4) unsigned NOT NULL AUTO_INCREMENT,  
  as_src int(4) unsigned NOT NULL,  
  as_dst int(4) unsigned NOT NULL,  
  peer_as_src int(4) unsigned NOT NULL,  
  peer_as_dst int(4) unsigned NOT NULL,  
  peer_ip_src char(15) NOT NULL,  
  peer_ip_dst char(15) NOT NULL,  
  packets int(10) unsigned NOT NULL,  
  bytes bigint(20) unsigned NOT NULL,  
  stamp_inserted datetime NOT NULL,  
  stamp_updated datetime DEFAULT NULL,  
  [ ... ] );
```

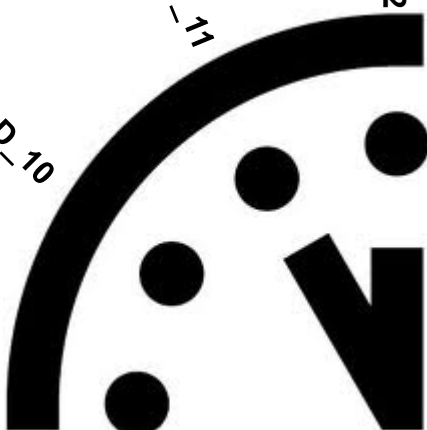


acct\_5mins\_YYYYMMDD\_09

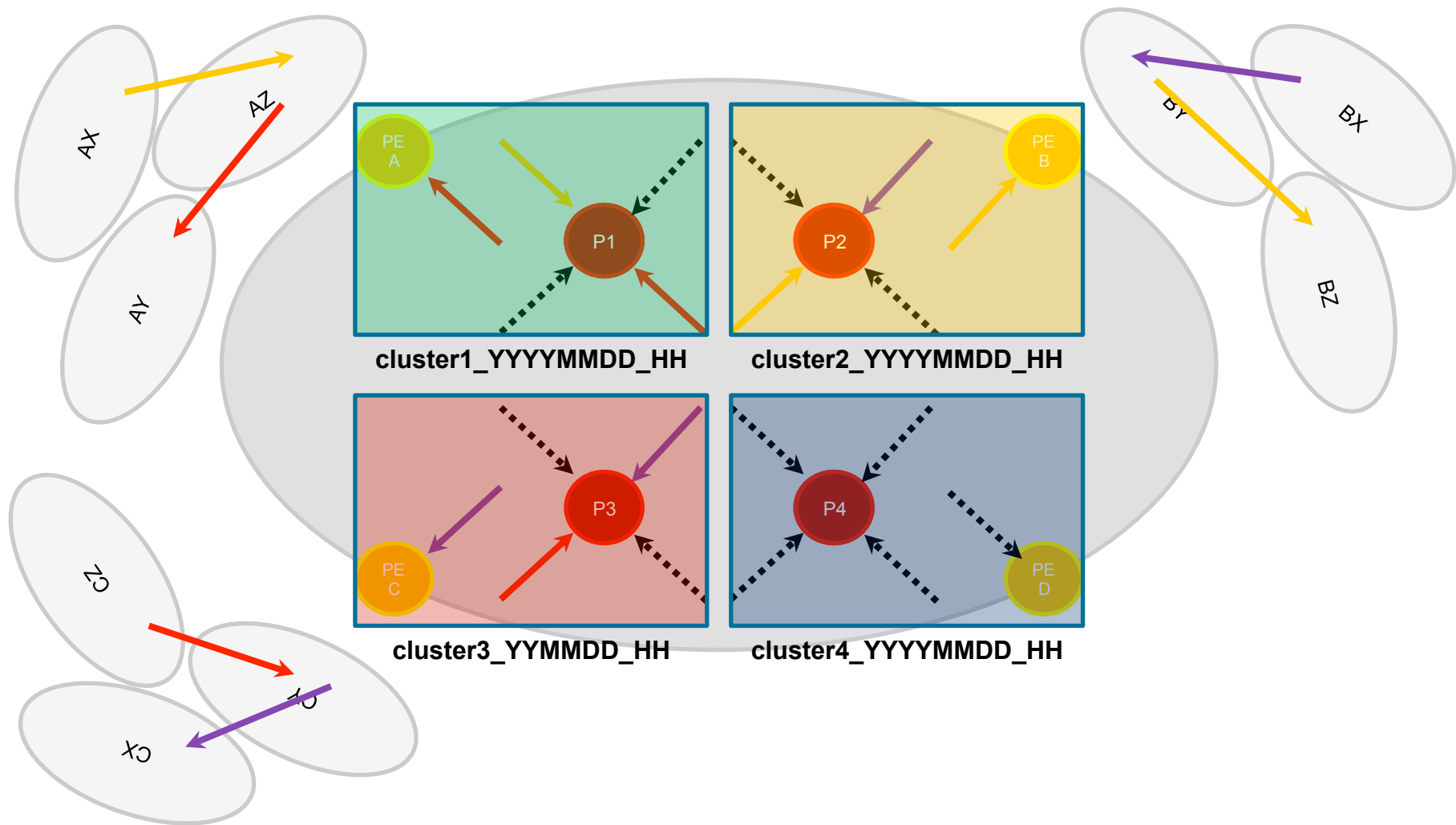
acct\_5mins\_YYYYMMDD\_10

acct\_5mins\_YYYYMMDD\_11

acct\_5mins\_YYYYMMDD\_12



# Scalability: spatial grouping



# Still on scalability

- A single collector might not fit it all:
  - Memory: can't store all BGP full routing tables
  - CPU: can't cope with the pace of telemetry export
  - Divide-et-impera approach is valid:
    - Assign routing elements (telemetry and BGP) to collectors
    - Assign collectors to RDBMSs; or cluster the RDBMS.
- Matrices can get big, but can be reduced:
  - Keep smaller routers out of the equation
  - Filter out specific services/customers on dense routers
  - Focus on relevant traffic direction (ie. upstream if CDN, downstream if ISP)
  - Increase sampling rate

# Downloading traffic matrices

- Strategic CP/TE solution traffic matrix:

```
SELECT peer_ip_src, peer_ip_dst, peer_as_src, peer_as_dst, bytes,  
stamp_inserted  
FROM <table>  
WHERE stamp_inserted = < today | last hour | last 5 mins >  
[ GROUP BY ... ];
```

- Tactical CP/TE solution traffic matrix  $k$  ( $1 \leq k \leq N$ ,  
 $N = \#$  observed interfaces):

```
SELECT peer_ip_src, iface_in, iface_out, as_src, as_dst, bytes,  
stamp_inserted  
FROM <table>  
WHERE peer_ip_src = < Pi | PEj > AND  
      iface_in = k AND  
      stamp_inserted = < today | last hour | last 5 mins >  
[ GROUP BY ... ];
```

# Further information

- [http://www.pmacct.net/lucente\\_pmacct\\_uknof14.pdf](http://www.pmacct.net/lucente_pmacct_uknof14.pdf)
  - AS-PATH radius, Communities filter, asymmetric routing
  - Entities on the provider IP address space
  - Auto-discovery and automation
- [http://www.pmacct.net/building\\_traffic\\_matrices\\_n49.pdf](http://www.pmacct.net/building_traffic_matrices_n49.pdf)  
[http://www.pmacct.net/pmacct\\_peering\\_epf5.pdf](http://www.pmacct.net/pmacct_peering_epf5.pdf)
  - Building traffic matrices to support peering decisions
- <http://wiki.pmacct.net/OfficialExamples>
  - Quick-start guide to setup a NetFlow/sFlow+BGP collector instance, implementation notes, etc.



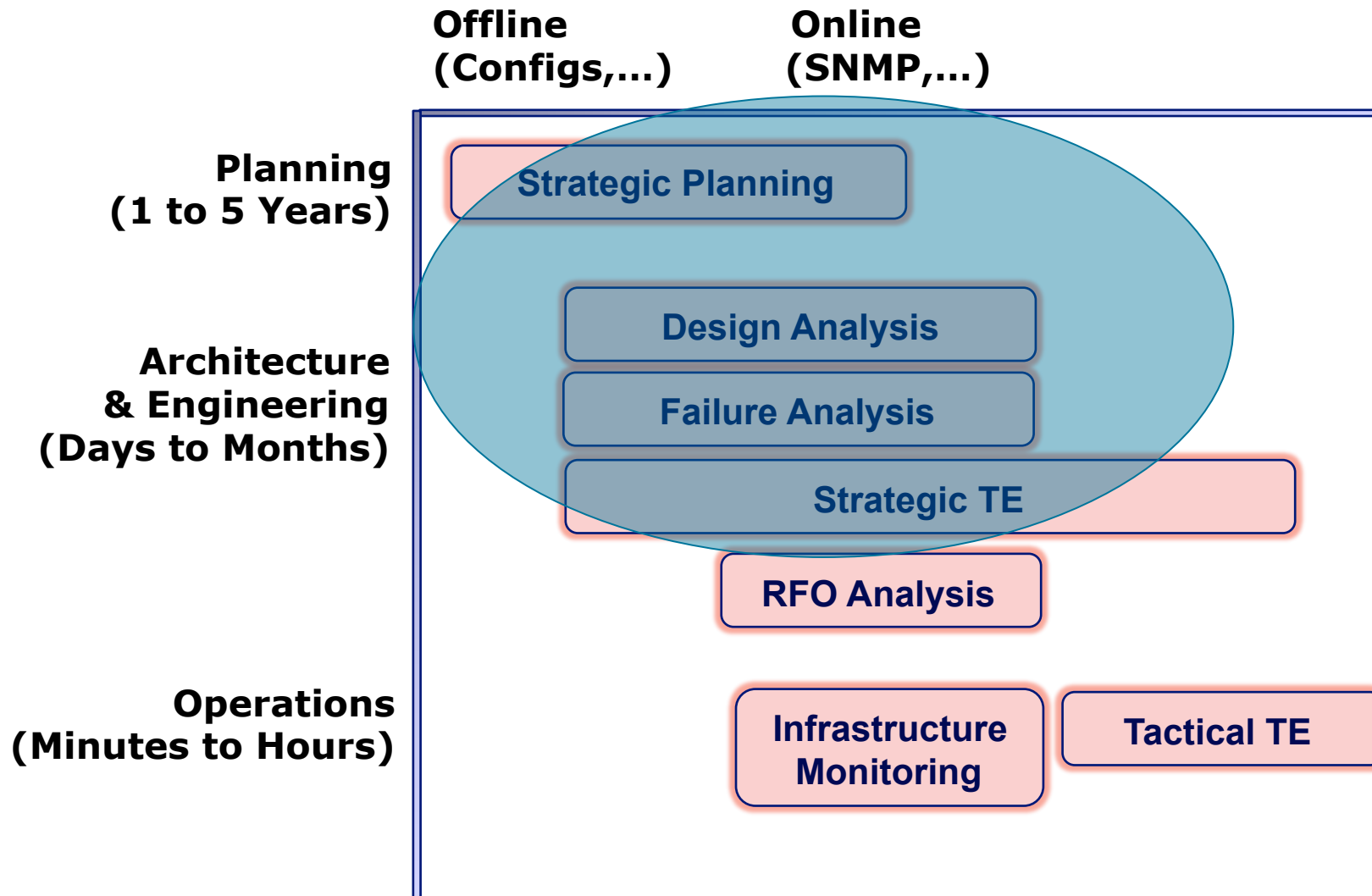


# Network Planning



## Emphasis on Backbone Planning

# Comprehensive Traffic Management

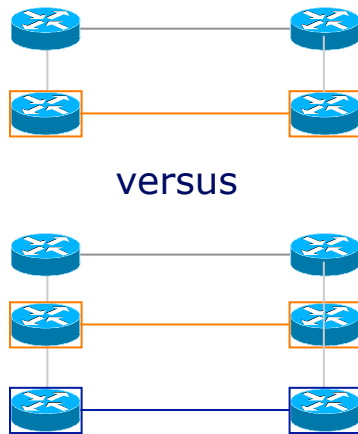


# Common & Wasteful (Core Topologies)

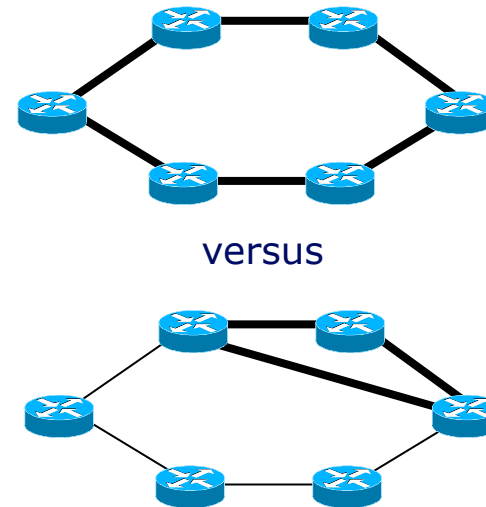


- Link capacity at each ladder section set as twice traffic in that section
- 1:1 protection: 50% of infrastructure for backup
- Ring is upgraded en masse even if one side empty
- Hard to add a city to the core, bypasses (express links) avoided because of complexity
- 1:1. And some infrastructure lightly used

# N:1 Savings



- **1:1 Protection**  
\$100 carrying capacity requires \$200 expenditure
- **2:1**  
\$100 carrying capacity requires \$150 expenditure
- 15%-20% in practice
- E.g. national backbone costing \$100M (capex+opex) saves \$15M-\$20M



- Instead of upgrading all elements upgrade the bottleneck
- Put in express route in bottleneck region
- 10%-20% savings are common

# N:1 Costs

- Physical diversity not present/cheap
  - However, usually present at high traffic points (e.g., no diversity in far away provinces but yes in capital regions)
- Engineering/architecture considerations
  - E.g., how effectively balance traffic
- Planning considerations
  - ➔ Subject of this section

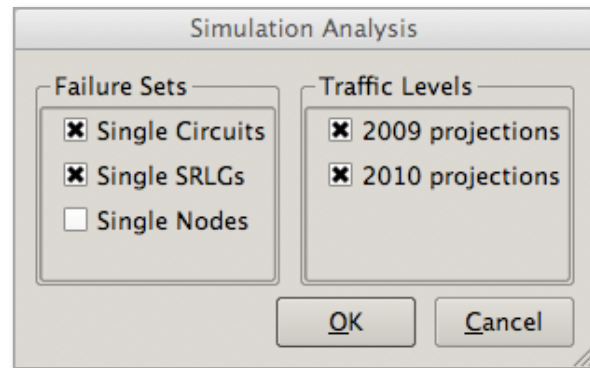
# Planning Methodologies

- Monitoring per link statistics doesn't cut it
- Planning needs to be topology aware
- Failure modes should be considered
- Blurs old boundaries between planning, engineering and operations

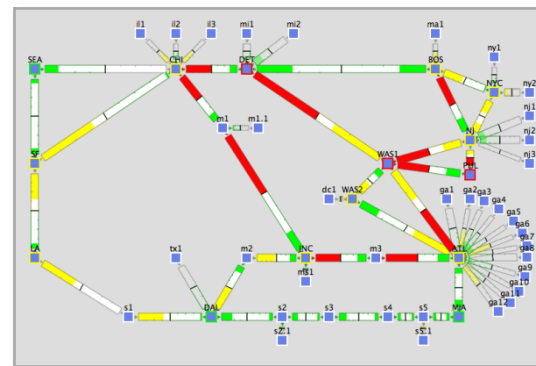
# Failure Planning

**Scenario:** Planning receives traffic projections, wants to determine what buildout is necessary

Simulate using external traffic projections

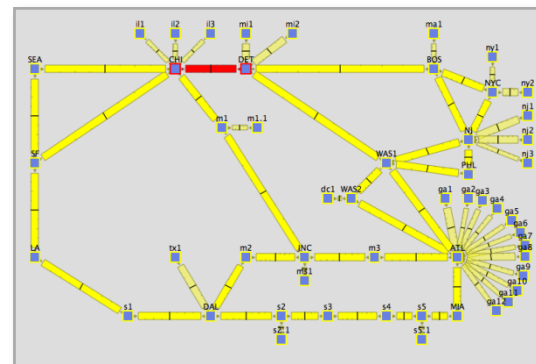


Worst case view



Potential congestion under failure in **RED**

Failure impact view



Failure that can cause congestion in **RED**

Perform  
topology What-If  
analysis

# Topology What-If Analysis

**Scenario:** Congestion between CHI and DET

•Add new circuit

•Specify parameters

•Congestion relieved

The screenshot displays a network topology analysis tool interface. On the left, a network map shows nodes CHI, DET, m1, m1.1, dc1, WAS2, and WAS1. A red line between CHI and DET indicates congestion. A 'New circuit' dialog box is open, with the 'Insert' tab selected. The 'Circuit...' option is highlighted in the menu. The dialog box contains the following fields and values:

- Circuit Name: CHI-WAS1
- Capacity: 2488
- Delay: 8.73
- OSPF Area: 0.0.0.0
- SRLGs: 0
- Protected: ☐
- Active: ☒
- Node A: Site: CHI, Node: cr1.chi
- Node B: Site: , Node:
- Interface A: Name: , IP Address: , IGP Metric: Shortest Distance, Description:
- Interface B: Name: , IP Address: , IGP Metric: SH, Description:

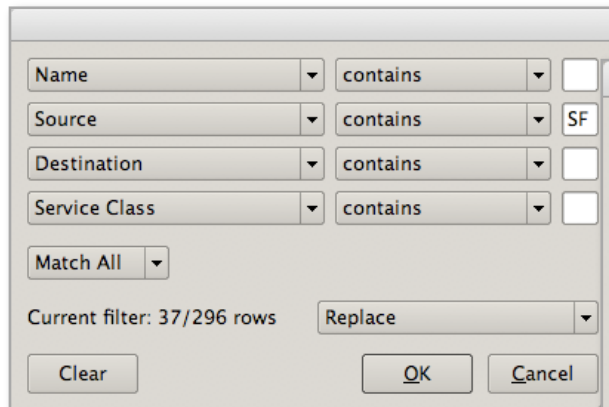
On the right, a smaller version of the network map shows the same topology, but the congestion between CHI and DET has been relieved, indicated by a green line.



# Evaluate New Services, Growth,...

**Scenario:** Product marketing expects 4 Gbps growth in SF based on some promotion

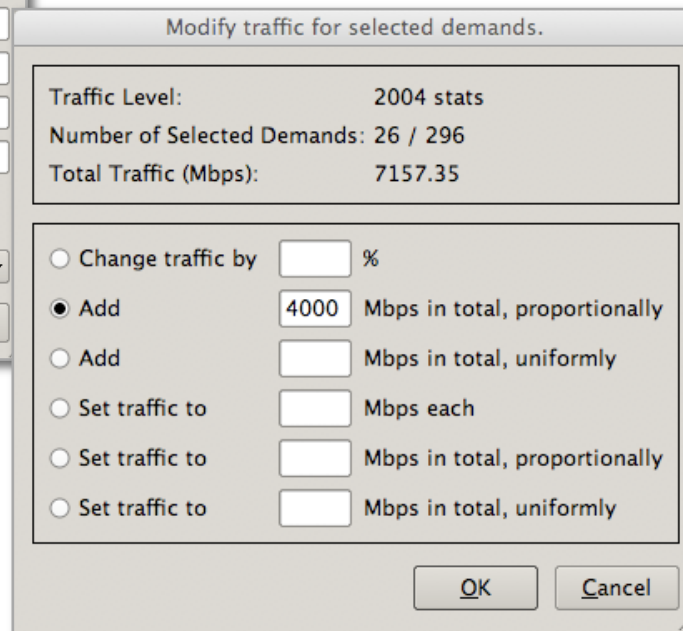
- Identify flows for new customer



Filter dialog box with the following fields:

- Name: contains
- Source: contains
- Destination: contains
- Service Class: contains
- Match All: (checked)
- Current filter: 37/296 rows
- Buttons: Clear, OK, Cancel

- Add 4Gbps to those flows



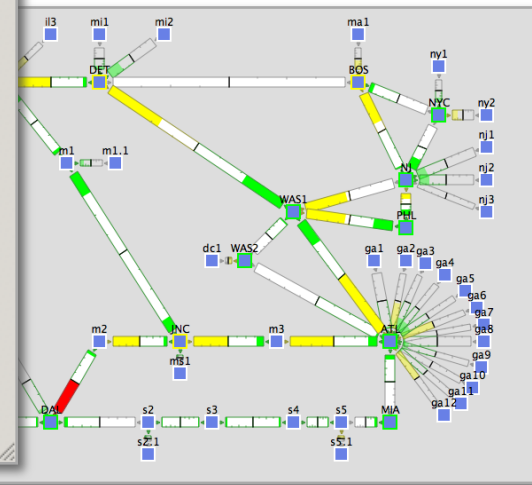
Modify traffic for selected demands.

Traffic Level: 2004 stats  
Number of Selected Demands: 26 / 296  
Total Traffic (Mbps): 7157.35

☐ Change traffic by [ ] %  
☒ Add 4000 Mbps in total, proportionally  
☐ Add [ ] Mbps in total, uniformly  
☐ Set traffic to [ ] Mbps each  
☐ Set traffic to [ ] Mbps in total, proportionally  
☐ Set traffic to [ ] Mbps in total, uniformly

Buttons: OK, Cancel

- Simulate results



- Congested link in **RED**



# Optimization/ Traffic Engineering



# Network Optimization

- Network Optimization encompasses network engineering and traffic engineering
  - Network engineering
    - Manipulating your network to suit your traffic
  - Traffic engineering
    - Manipulating your traffic to suit your network
- Whilst network optimization is an optional step, all of the preceding steps are essential for:
  - Comparing network engineering and TE approaches
  - MPLS TE tunnel placement and IP TE

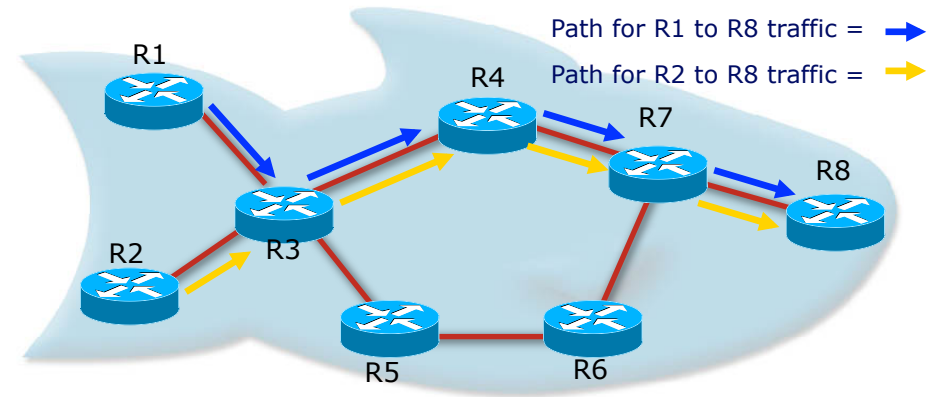
# Network Optimization: Questions

- What optimization objective?
- Which approach?
  - IGP TE or MPLS TE
- Strategic or tactical?
- How often to re-optimize?
- If strategic MPLS TE chosen:
  - Core or edge mesh
  - Statically (explicit) or dynamically established tunnels
  - Tunnel sizing
  - Online or offline optimization
  - Traffic sloshing

# IP Traffic Engineering: The Problem

- Conventional IP routing uses pure destination-based forwarding where path computation is based upon a simple additive metric

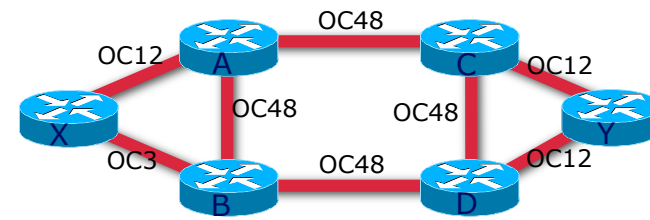
- Bandwidth availability is not taken into account



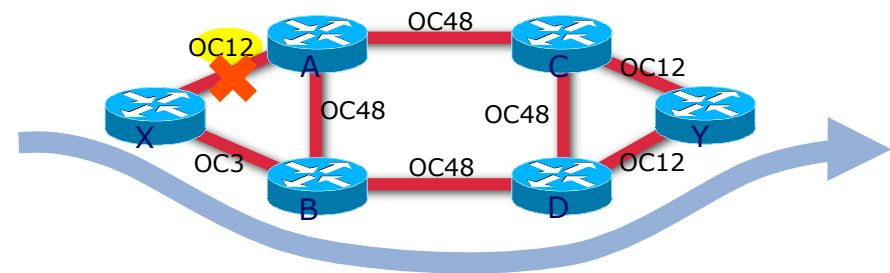
- Some links may be congested while others are underutilized
- The traffic engineering problem can be defined as an optimization problem
  - Definition – “*optimization problem*”: A computational problem in which the objective is to find the best of all possible solutions
    - ➔ Given a fixed topology 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 ... ?
    - ➔ Maximum Flow problem [MAXFLOW]

# 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 ...
- Ultimate measure of success is cost saving

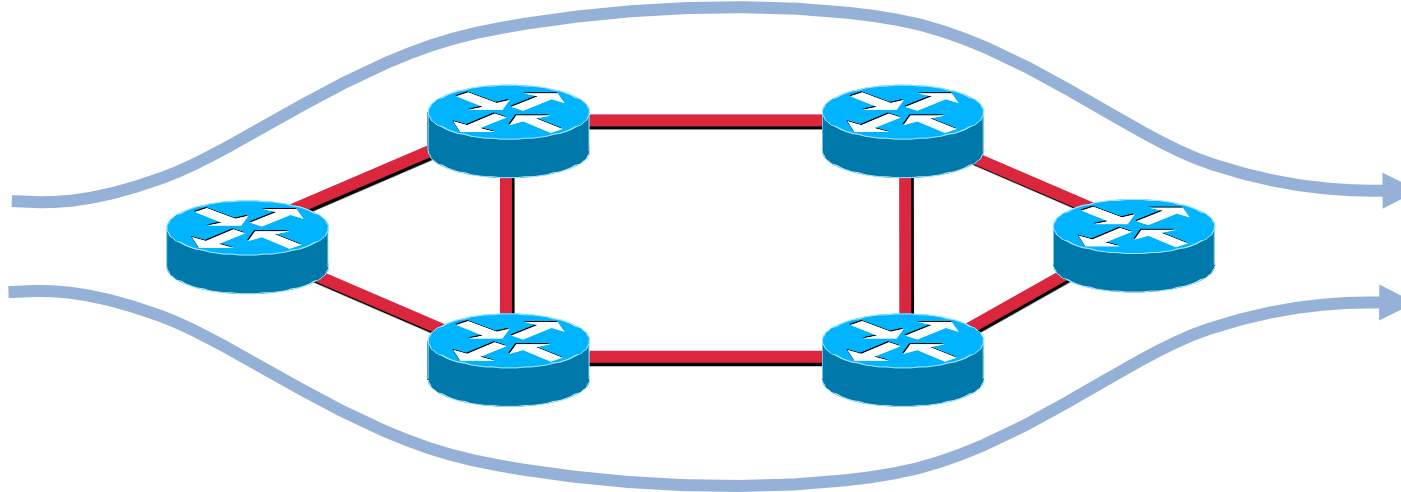


- In this asymmetrical topology, if the demands from  $X \rightarrow Y > OC3$ , traffic engineering can help to distribute the load when all links are working



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

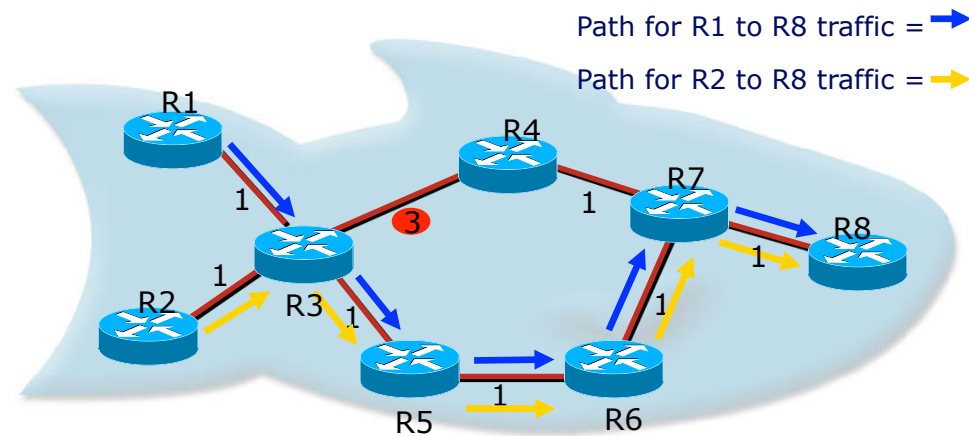
# Traffic Engineering Limitations



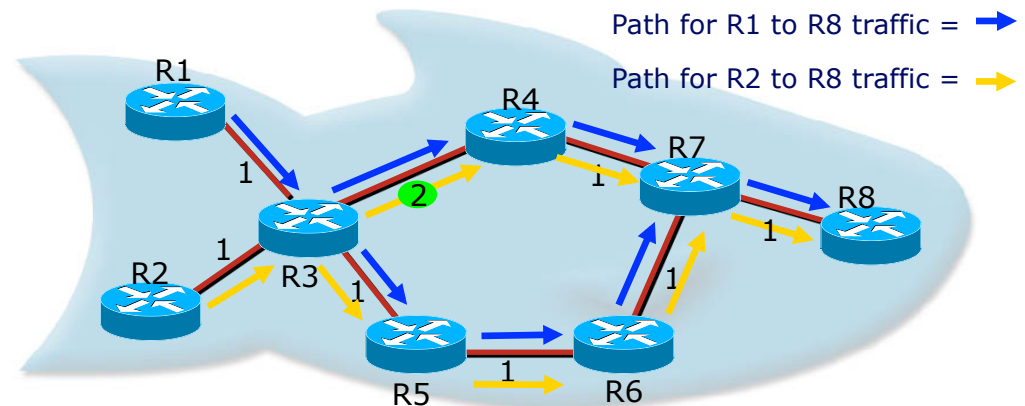
- TE cannot create capacity
  - e.g. “V-O-V” topologies allow no scope strategic TE if optimizing for failure case
    - Only two directions in each “V” or “O” region – no routing choice for minimizing failure utilization
- Other topologies may allow scope for TE in failure case
  - As case study later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

# IGP metric-based traffic engineering

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



- ... the mantra that tweaking IGP metrics just moves problem around is not generally true in practise
  - Note: IGP metric-based TE can use ECMP



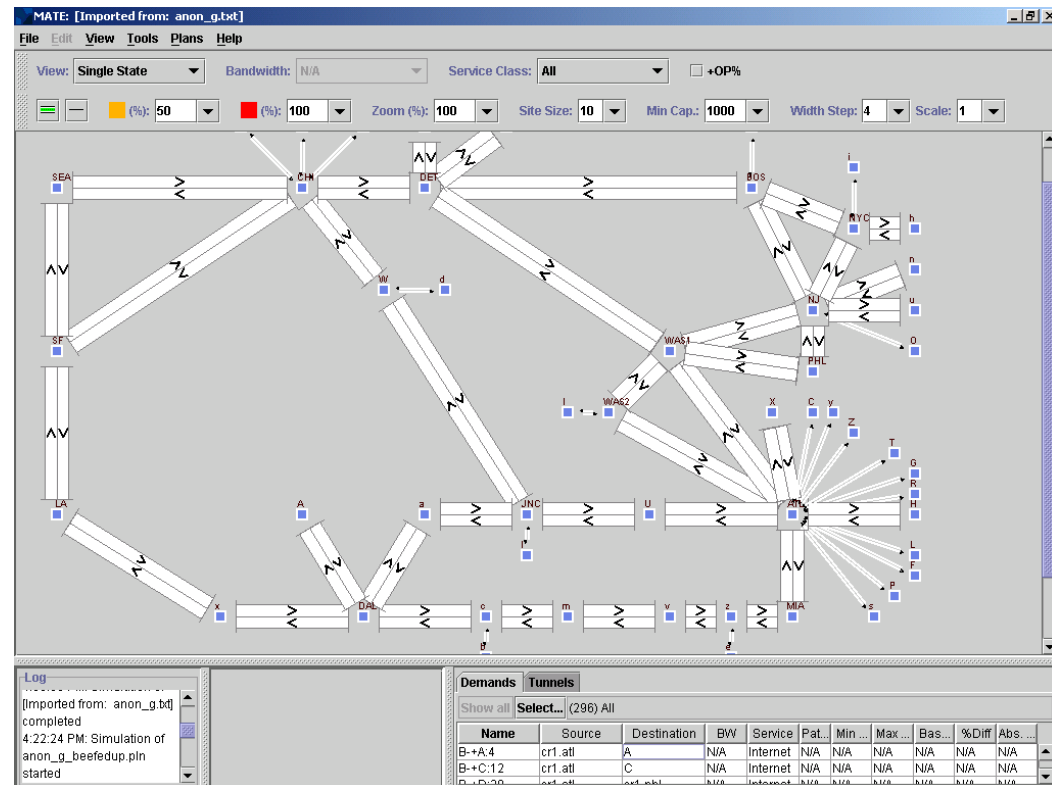


# IGP metric-based traffic engineering

- Significant research efforts ...
  - 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. Liao. Routing strategies for IP networks. Teletronikk, 2/3, pp 145-158, 2001.
  - ...

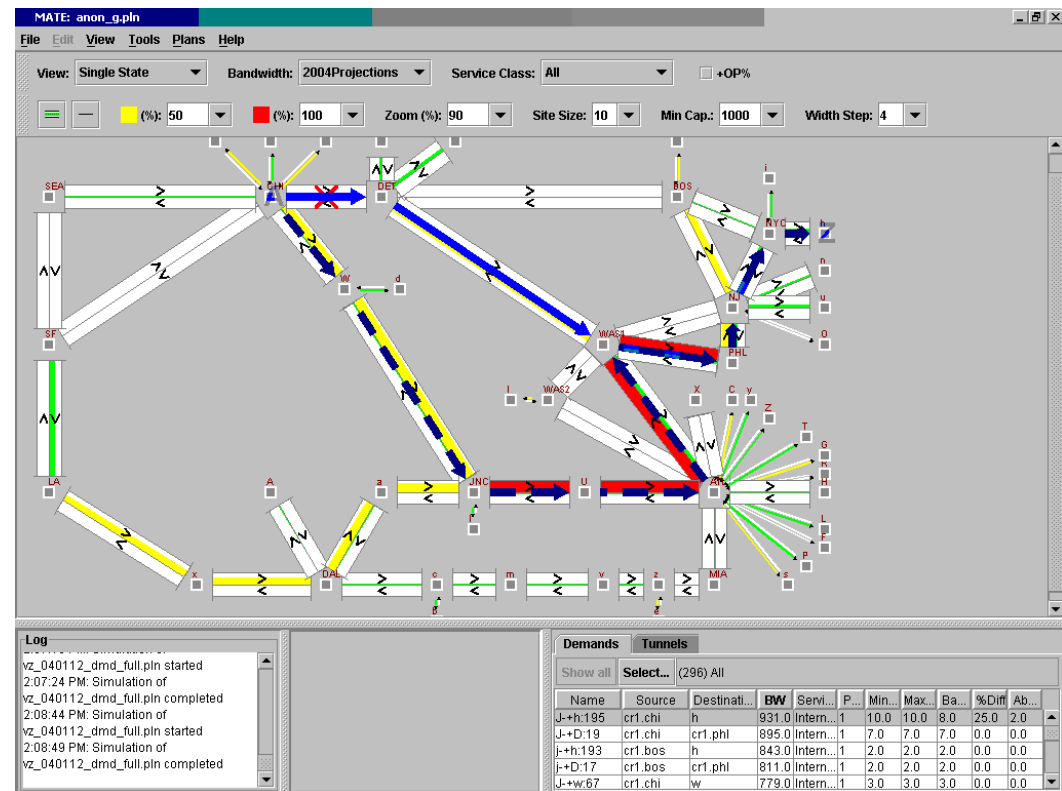
# IGP metric-based traffic engineering: Case study

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



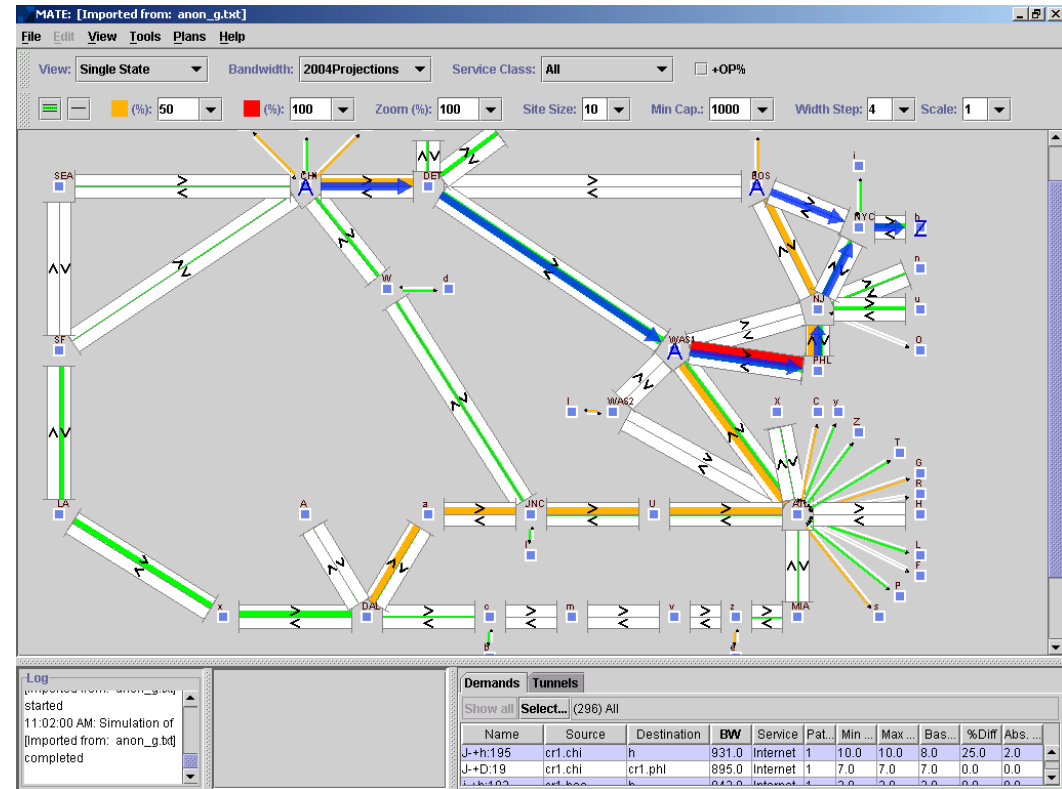
# Metric TE Case Study: Plot Legend

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



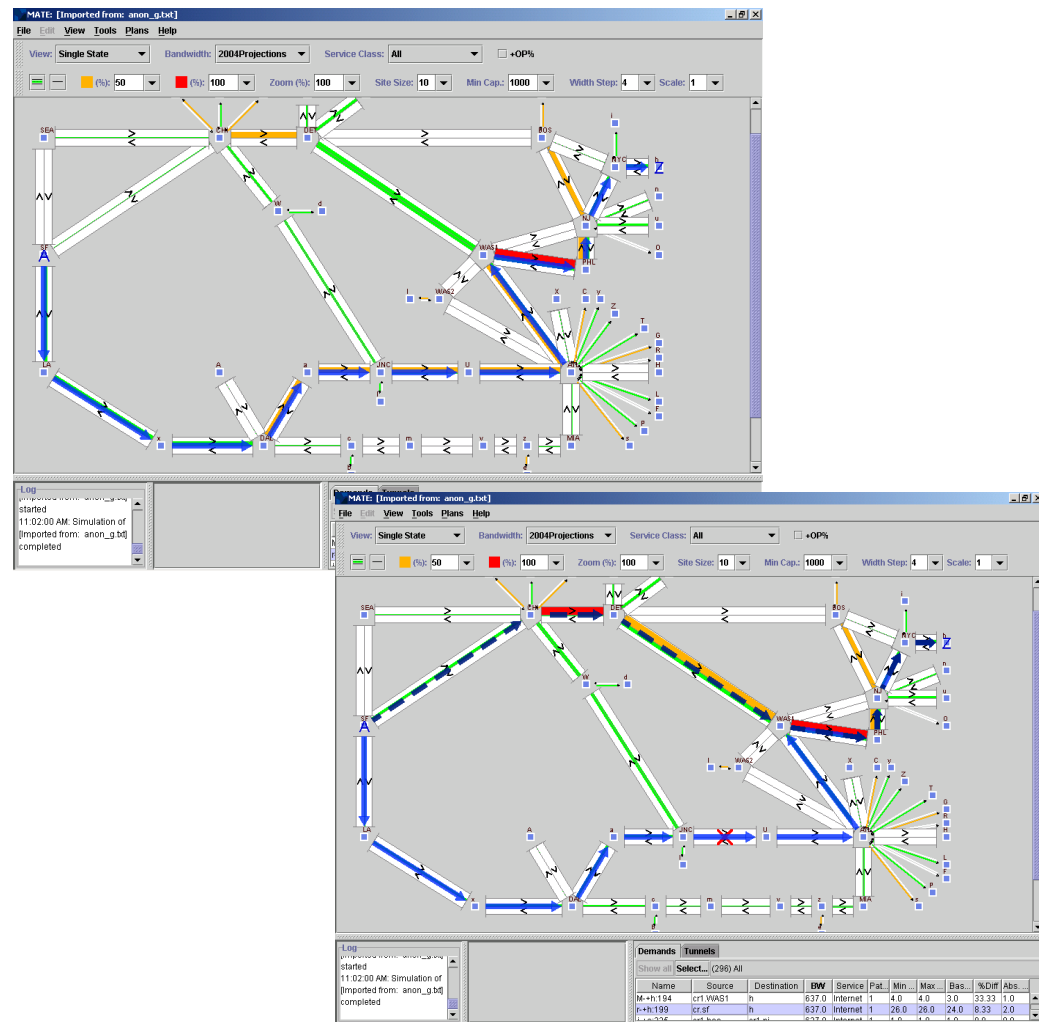
# Metric TE Case Study: Traffic Overview

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



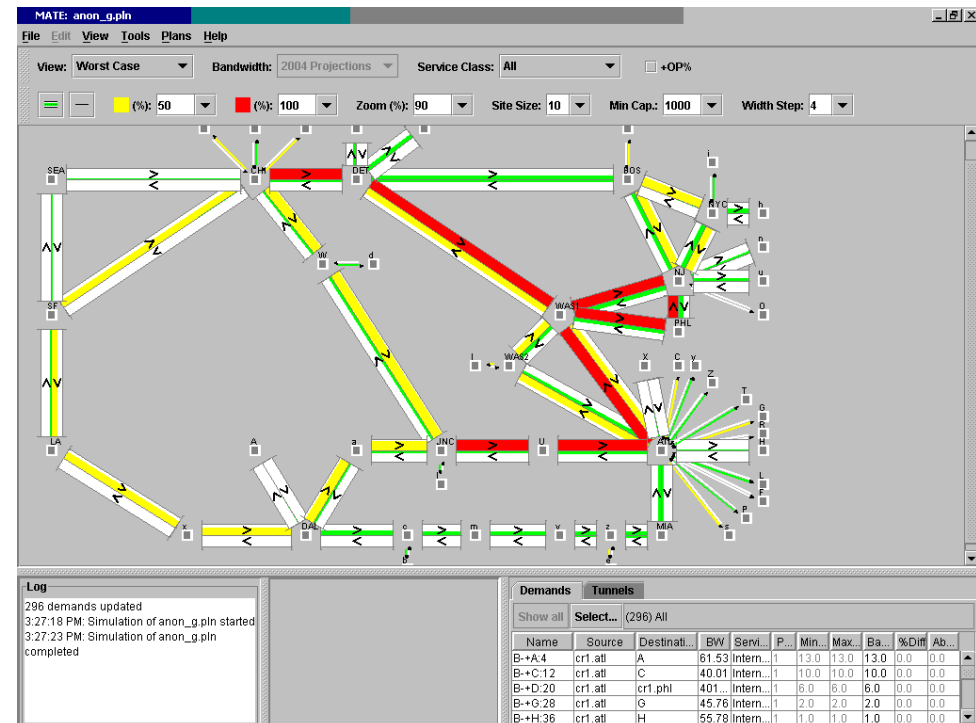
# Metric TE Case Study: Manual Attempt at Metric TE

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



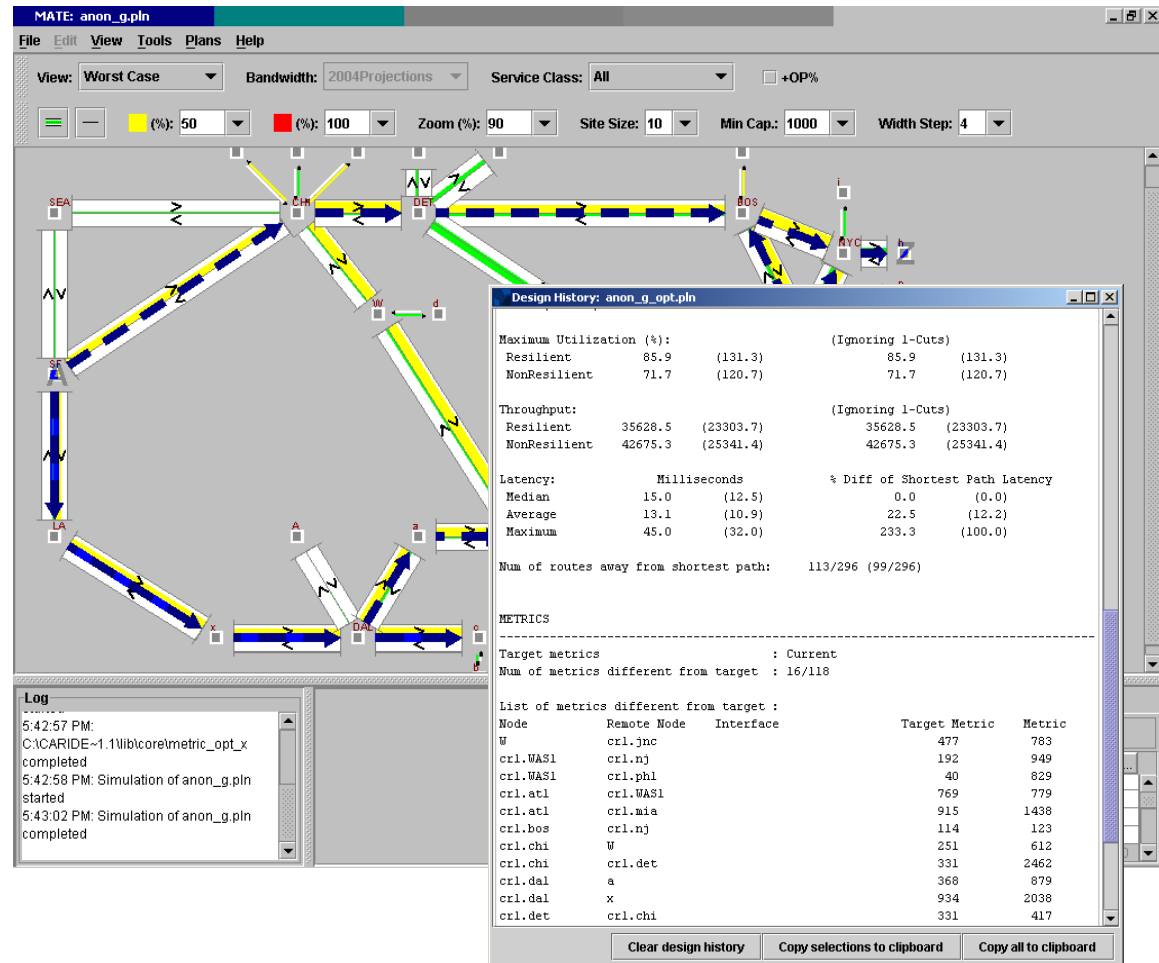
# Metric TE Case Study: Worst Case Failure View (Before)

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



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

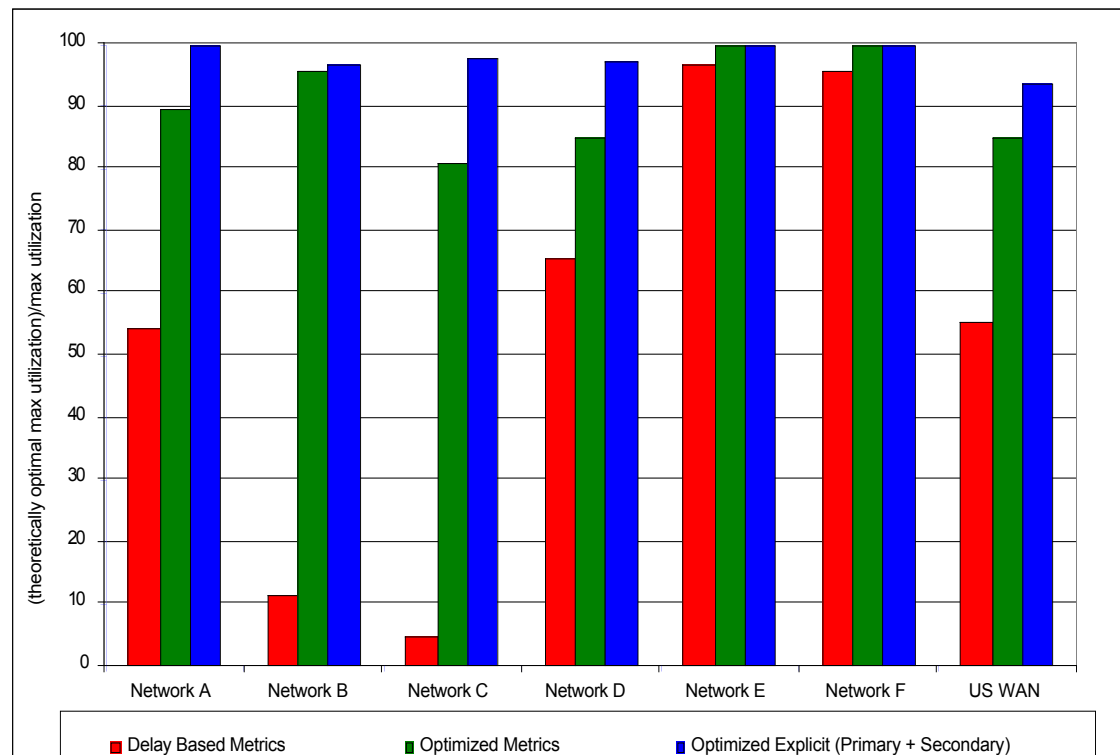


# Metric TE Case Study: Performance over Various Networks

- See: [Maghbouleh 2002]
- Study on Real Networks

- Single set of metrics achieves 80-95% of theoretical best across failures
- Optimized metrics can also be deployed in an MPLS network

■ e.g. LDP networks





# MPLS TE deployment considerations

- 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
- 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

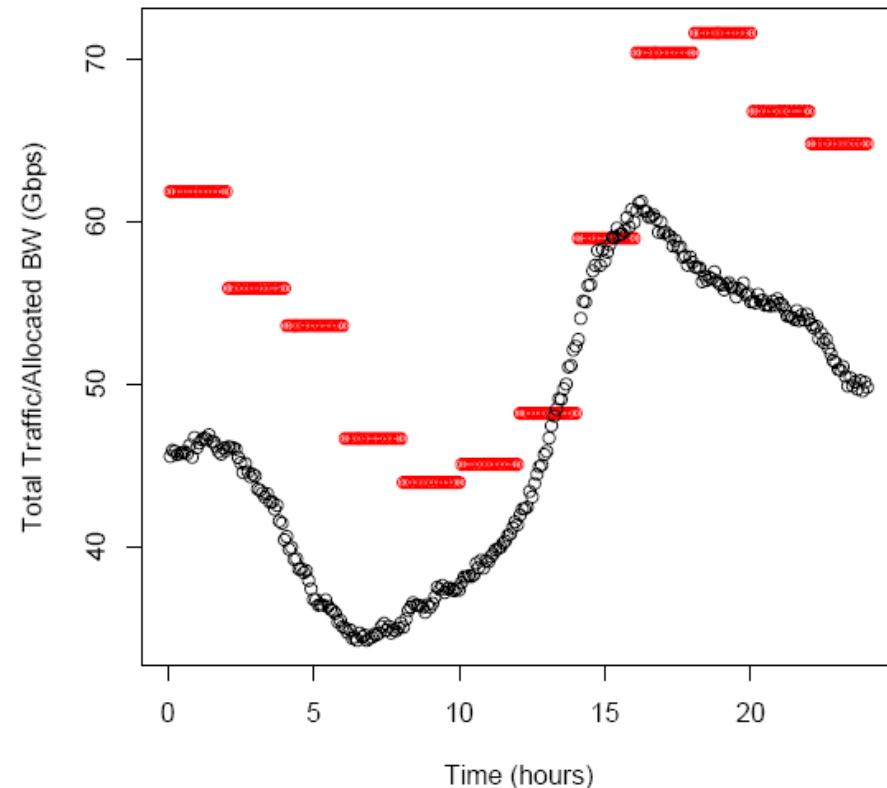
# Tunnel Sizing

- Tunnel sizing is key ...
  - Needless congestion if actual load  $\gg$  reserved bandwidth
  - Needless tunnel rejection if reservation  $\gg$  actual load
    - Enough capacity for actual load but not for the tunnel reservation
- 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

# Tunnel Sizing

- Online vs. offline sizing:
  - Online sizing: autobandwidth
    - Router automatically adjusts reservation (up or down) based on traffic observed in previous time interval
    - Tunnel bandwidth is not persistent (lost on reload)
    - Can suffer from “bandwidth lag”
  - Offline sizing
    - Statically set reservation to percentile (e.g. P95) of expected max load
    - Periodically re-adjust – not in real time, e.g. daily, weekly, monthly

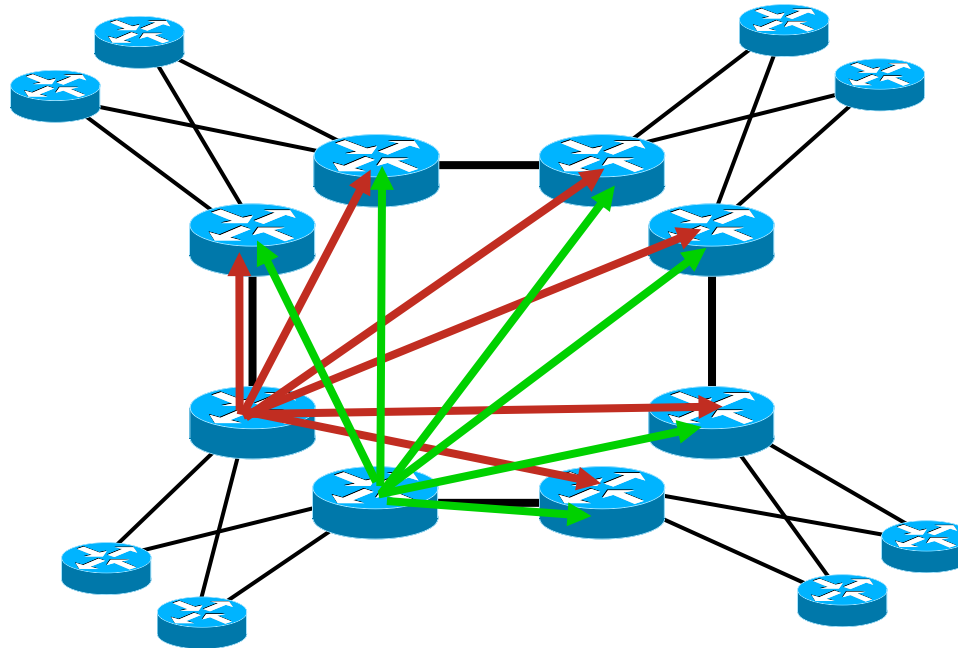
“online sizing: bandwidth lag”



# Tunnel Sizing

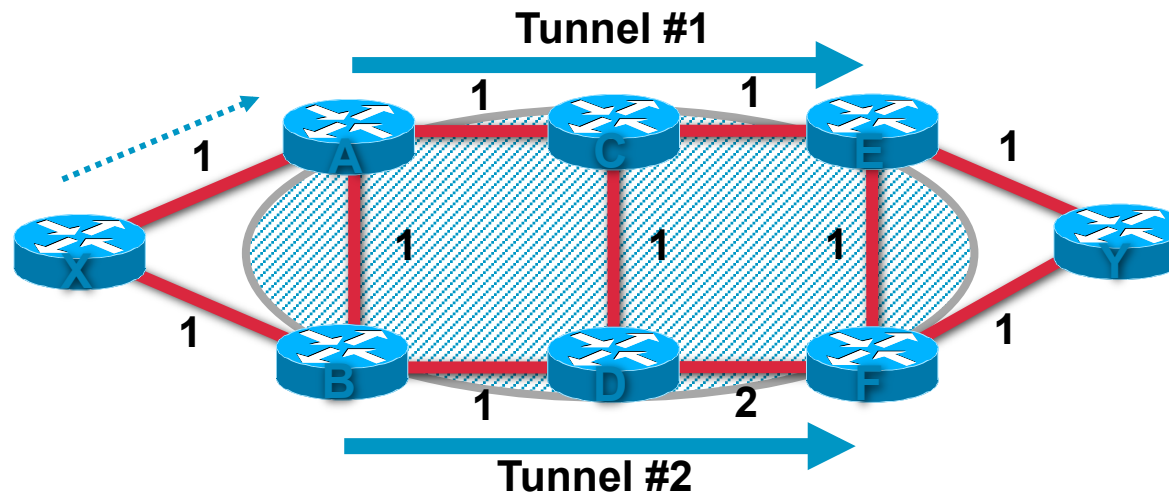
- When to re-optimize?
  - Event driven optimisation, e.g. on link or node failures
    - Won't re-optimize due to tunnel changes
  - Periodically
    - Tunnel churn if optimisation periodicity high
    - Inefficiencies if periodicity too low
    - Can be online or offline

# Strategic Deployment: Core Mesh



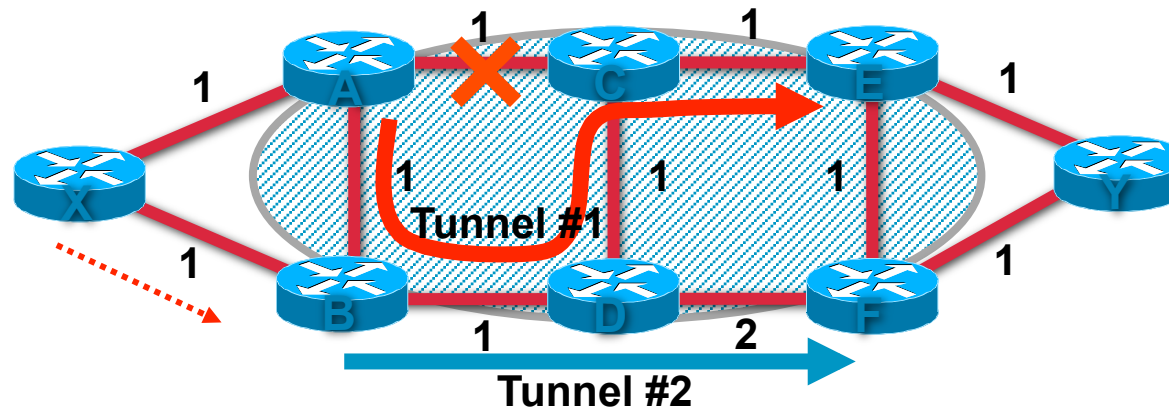
- Reduces number of tunnels required
- Can be susceptible to “traffic-sloshing”

# Traffic “sloshing”



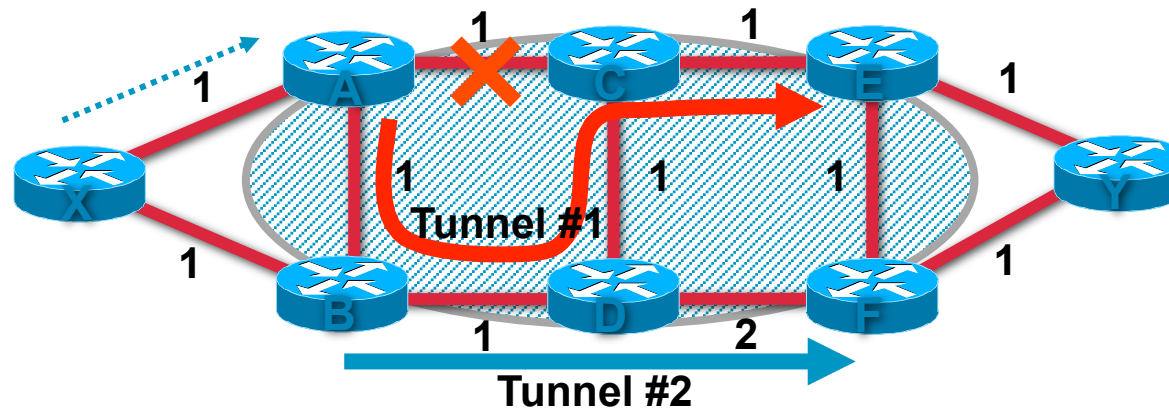
- In normal case:
  - For traffic from X → Y, router X IGP will see best path via router A
  - Tunnel #1 will be sized for X → Y demand
  - If bandwidth is available on all links, Tunnel from A to E will follow path A → C → E

# 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”

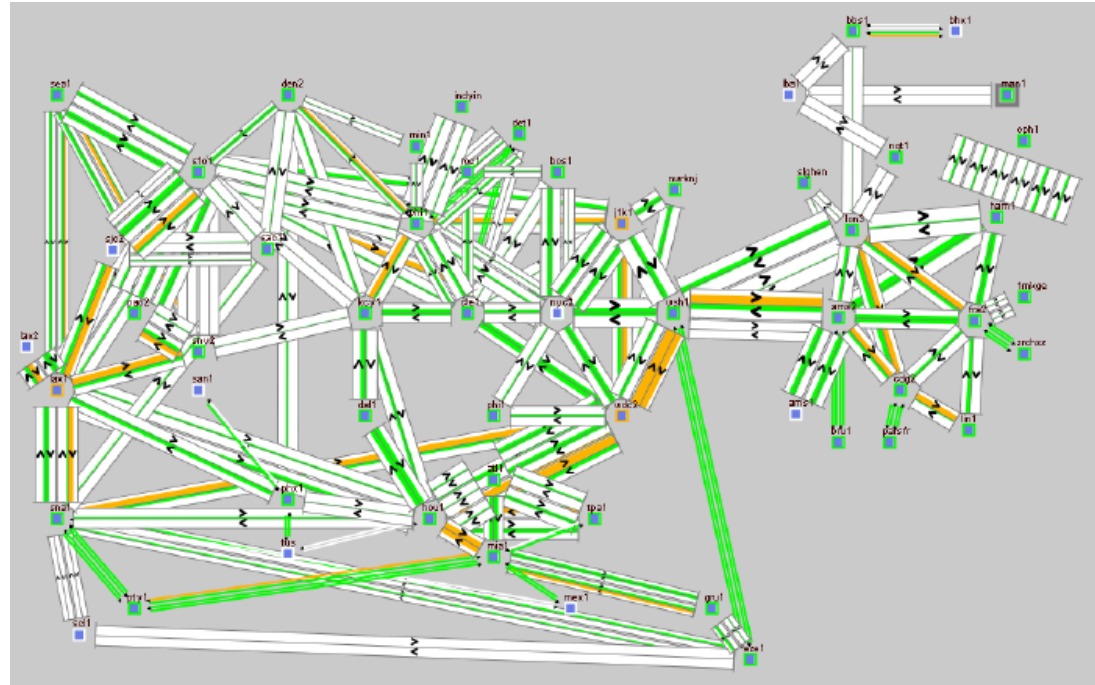


- Forwarding adjacency (FA) could be used to overcome traffic sloshing
  - Normally, a tunnel only influences the FIB of its head-end and other nodes do not see it
  - With FA the head-end advertises the tunnel in its IGP LSP
    - Tunnel #1 could always be made preferable over tunnel #2 for traffic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



# TE Case Study 1: Global Crossing\*

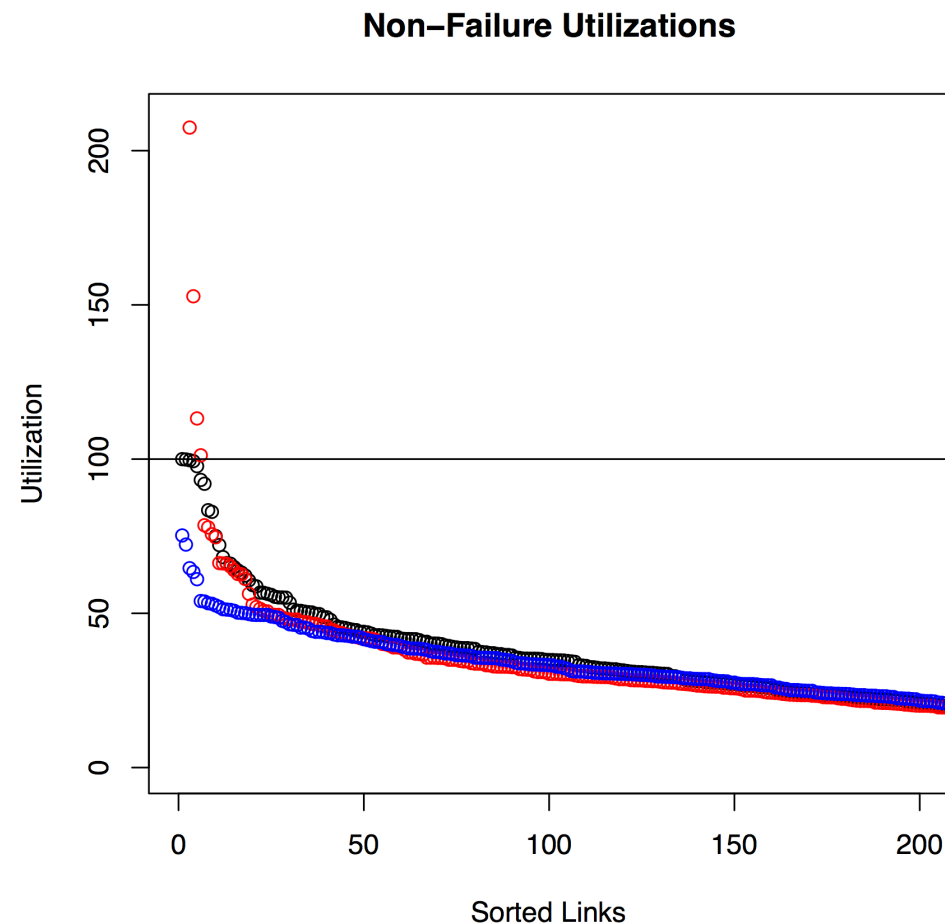
- Global IP backbone
  - Excluded Asia due to migration project
- MPLS TE (CSPF)
- Evaluate IGP Metric Optimization
  - Using 4000 demands, representing 98.5% of total peak traffic
- Topology:
  - highly meshed



(\*) Presented at TERENA Networking Conference, June 2004

# TE Case Study 1: Global Crossing

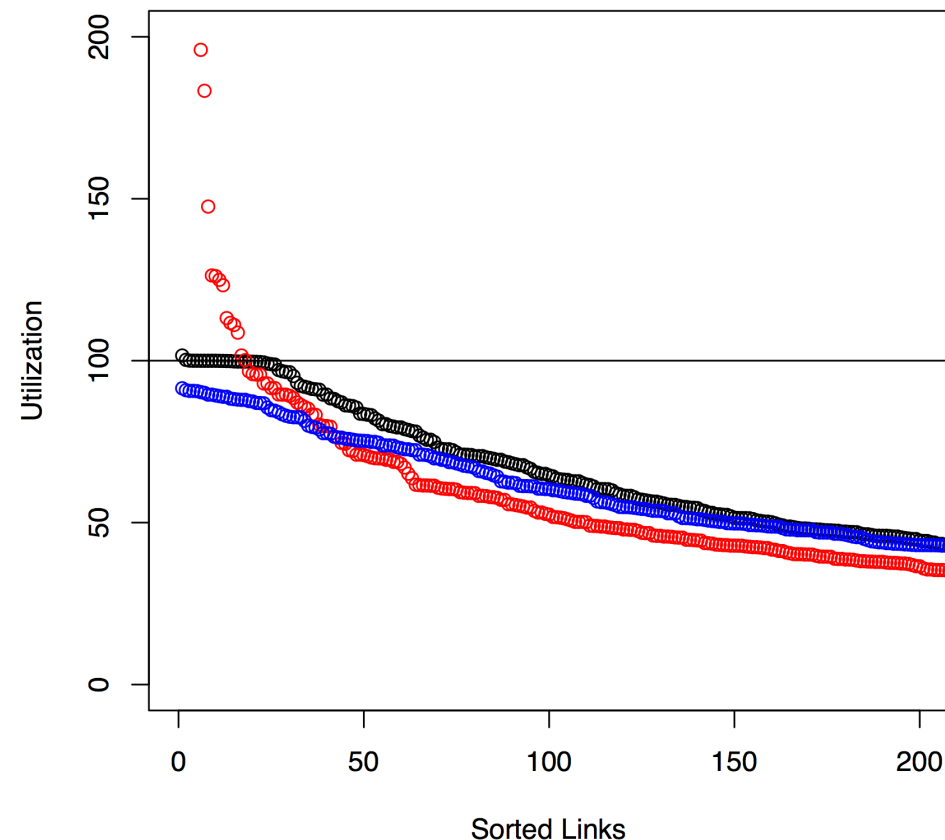
- Comparison:
  - Delay-based Metrics
  - MPLS CSPF
  - Optimized Metrics
- Normal Utilizations
  - no failures
- 200 highest utilized links in the network
- Utilizations:
  - Delay-based: **RED**
  - CSPF: **BLACK**
  - Optimized: **BLUE**



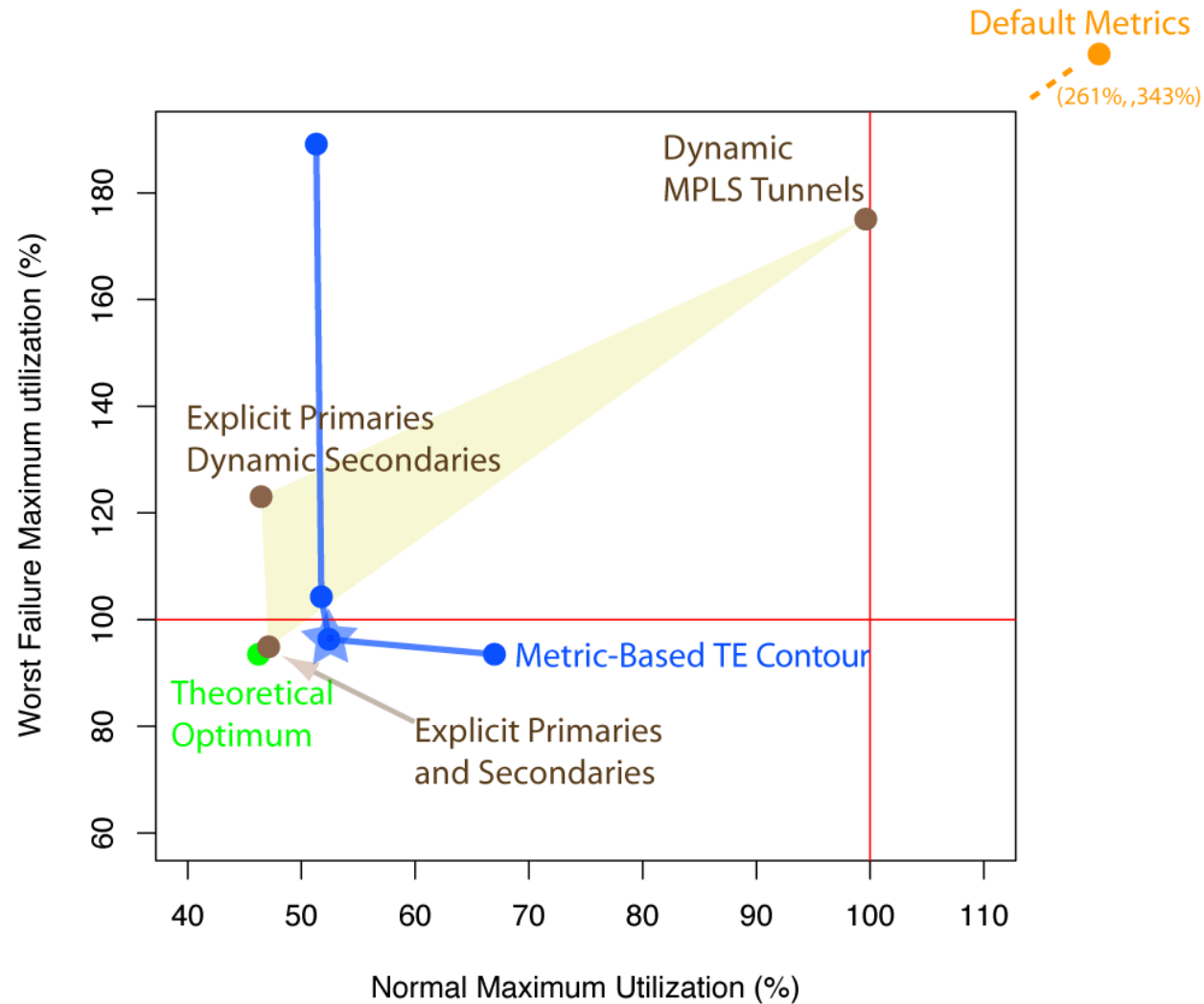
# TE Case Study 1: Global Crossing

- Worst-Case Utilizations
  - single-link failures
  - core network
  - 263 scenarios
- Results:
  - Delay-based metrics cause congestions
  - CSPF fills links to 100%
  - Metric Optimization achieves <90% worst-case utilizations

Worst-Case Utilizations for Single Link Failures



# TE Case Study 2: Deutsche Telekom\*

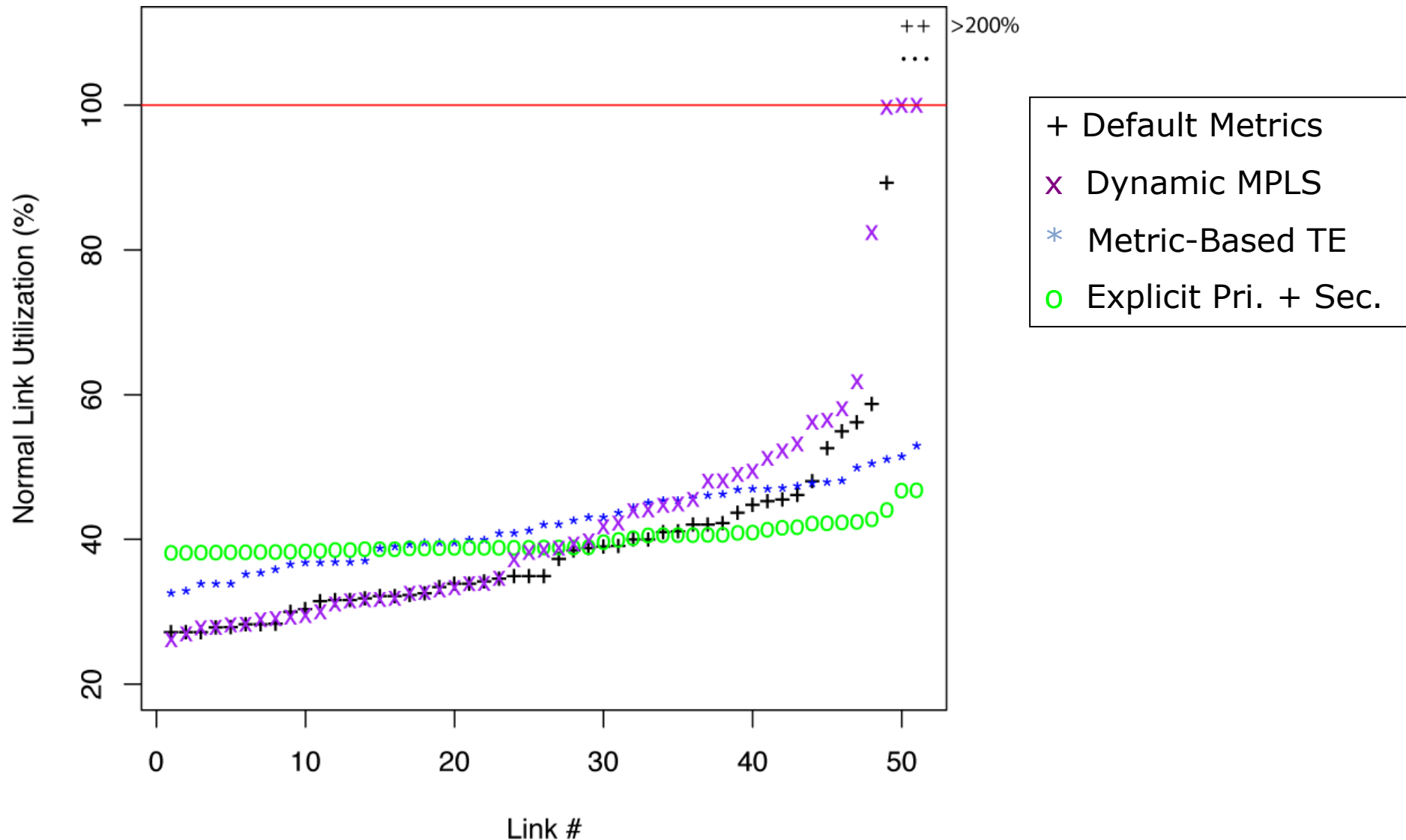


(\*) Presented at Nanog 33, by Martin Horneffer (DT)

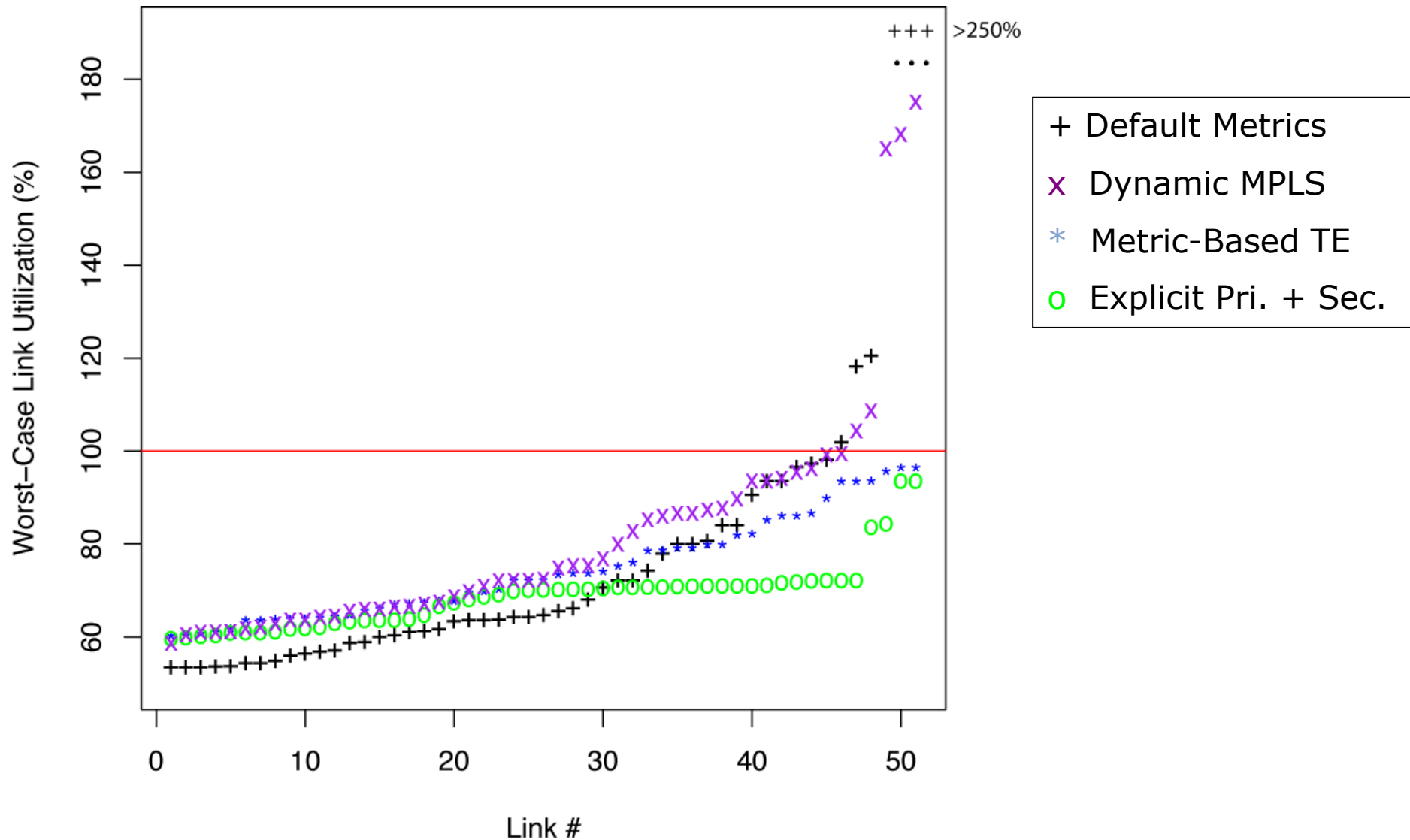
# TE Case Study 3

- Anonymous network...
- TE Options:
  - Dynamic MPLS
    - Mesh of CSPF tunnels in the core network
    - “Sloshing” causes congestion under failure scenarios
  - Metric Based TE
  - Explicit Pri. + Sec. LSPs
  - Failures Considered
    - Single-circuit, circuit+SRLG, circuit+SRLG+Node
    - Plot is for single-circuit failures

# Top 50 Utilized Links (normal)



# Top 50 Utilized Links (failures)



# Traffic Engineering Experiences

- Some meshing in the topology required to save costs
- Metric TE
  - Simple to deploy
  - Requires uniform capacities across parallel paths
- MPLS TE
  - Dynamic tunnels
    - Very resilient and efficient
    - Tunnel mesh and sizing issues, non deterministic
  - Explicit tunnels
    - Very efficient
    - Requires complex solutions to deploy

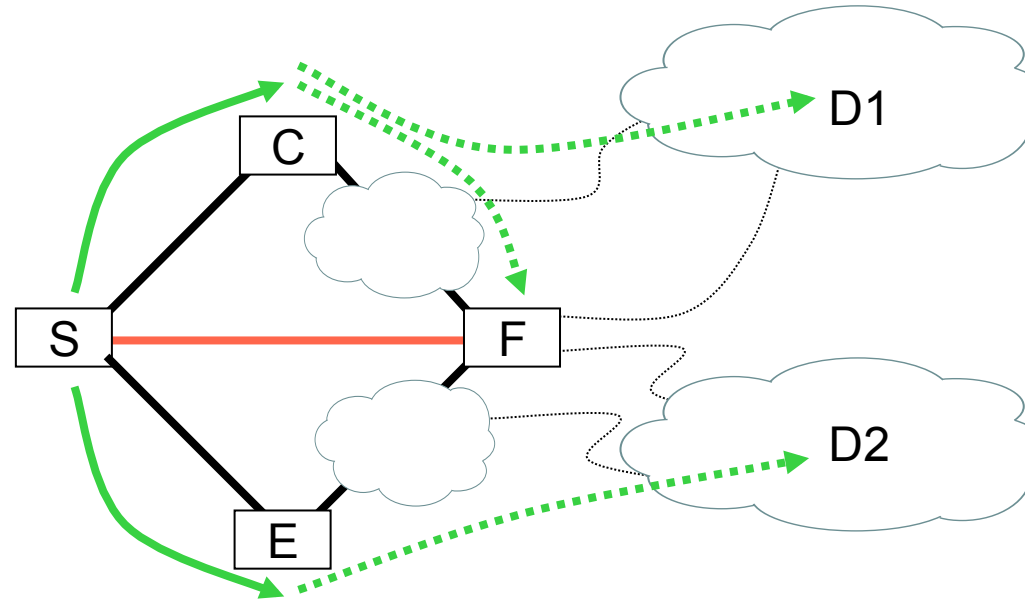




# Planning for LFA FRR



# Per-Prefix LFA Algorithm



- For IGP route D1, S's primary path is link SF.
- S checks for each neighbor N ( $\neq F$ ) whether  $ND1 < NS + SD1$  (Eq1)
  - “does the path from the neighbor to D1 avoid me?”
  - If so, it is a loop-free alternate (LFA) to my primary path to D1

# One backup path per primary path

- Default tie-break
  1. Prefer primary over secondary
  2. Prefer lowest backup path metric
  3. Prefer linecard disjointness
  4. Prefer node disjointness
- CLI to customize the tie-break policy
  - Default is recommended. Simplicity.

# Benefits

- Simple
  - the router computes everything automatically
- <50msec
  - pre-computed and pre-installed
  - prefix-independent
  - Leverage IOS-XR Hierarchical dataplane FIB
- Deployment friendly
  - no IETF protocol change, no interop testing, incremental deployment

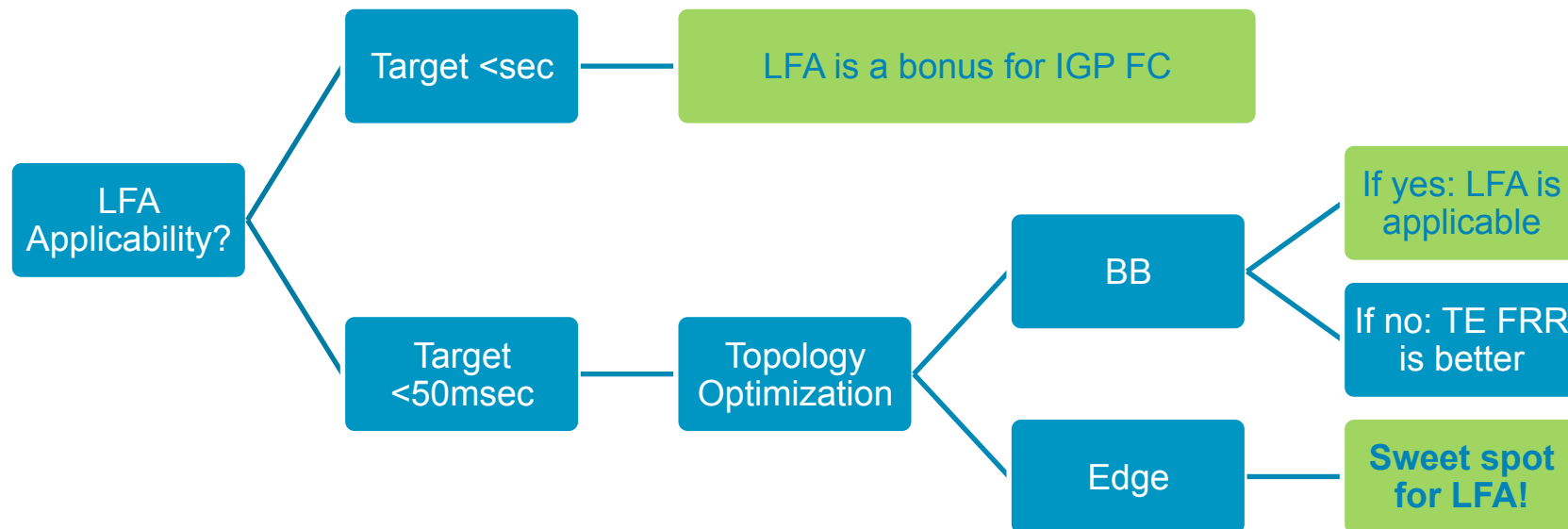
# Benefits

- Good Scaling
- No degradation on IGP convergence for primary paths
- Capacity Planning
- Node Protection (Guaranteed or De Facto)
  - an LFA can be chosen on the basis of the guaranteed-node protection
  - simulation indicate that most link-based LFA's anyway avoid the node (ie. De Facto Node Protection)

# Constraints

- Topology dependent
  - availability of a backup path depends on topology
  - Is there a neighbor which meets Eq1?

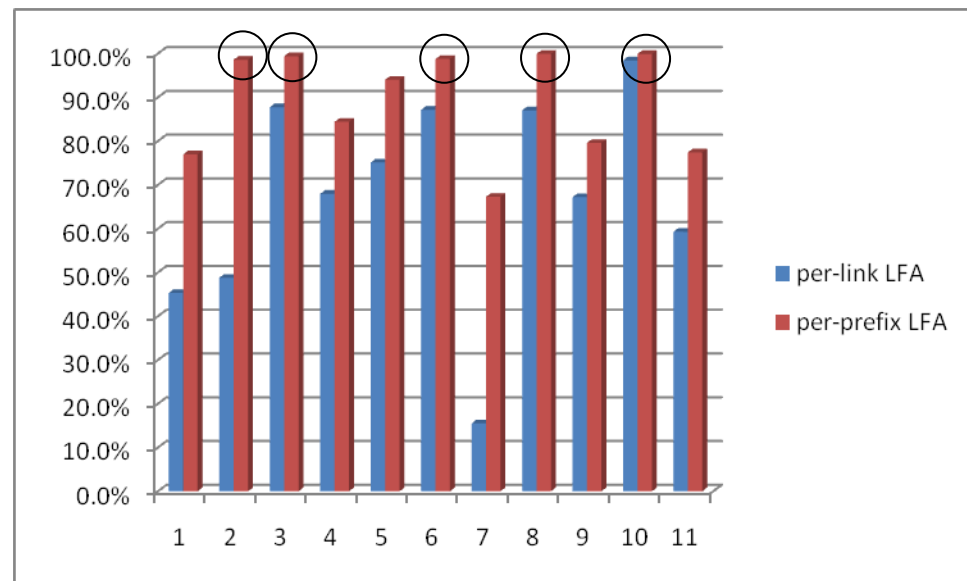
# Deployment



draft-ietf-rtgwg-lfa-applicability-00

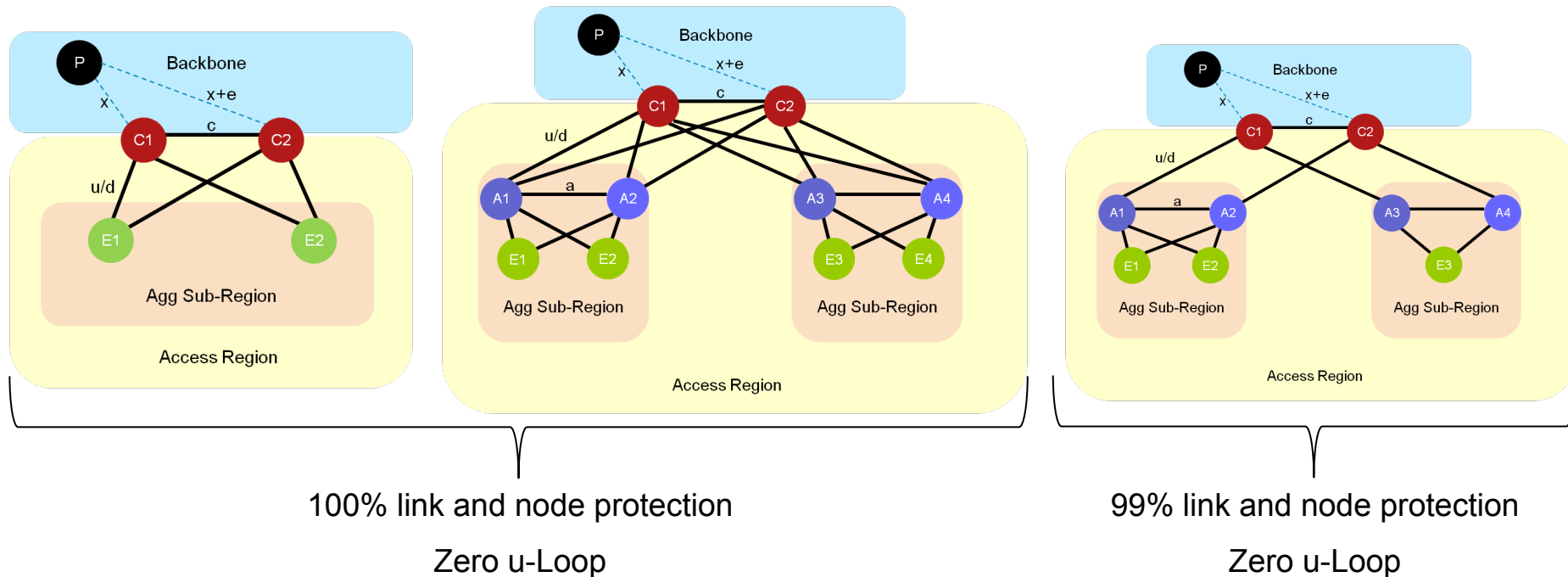
# Backbone Applicability

- Based on ~10 SP backbone topologies
  - Link LFA: 70% of the links are protected
  - Prefix LFA: 94% of the prefixes across all links are protected
- Some SP's selected LFA FRR for the backbone
  - implies a tight process to plan the topology
  - needs tools such as Cariden MATE
  - 5 topologies are well above 95% protection
  - Per-Prefix LFA is likely selected for its better coverage





# Access/Aggregation Topologies



- Assuming a few IGP metric rules described in draft-filsfils-lfa-applicability-00

- A reference to consult if interested
- Slight modification to slide 17. The solution will be called “Remote LFA” and an ietf draft should be released in the next weeks.

### LFA (Loop-Free Alternates) Case Studies in Verizon's MPLS Network

Ning So, Verizon Inc., [ning.so@verizonbusiness.com](mailto:ning.so@verizonbusiness.com)  
 Fengman Xu, Verizon Inc., [fengman.xu@verizonbusiness.com](mailto:fengman.xu@verizonbusiness.com)  
 Connie Chen, WANDL Inc., [connie@wandl.com](mailto:connie@wandl.com)  
 Tony Lin, WANDL Inc., [thl@wandl.com](mailto:thl@wandl.com)

[www.mpls2010.com](http://www.mpls2010.com)



MPLS 2010

### Coverage Case Studies for Various Networks

Topology	#links/#nodes	Per-link LFA Coverage	Per-prefix LFA Coverage
T1	4.2	49%	73%
T2	2.1	22%	73%
T3	6.5	51%	68%
T4	3.0	36%	70%
T5	5.7	45%	69%
T6	1.7	42%	72%
T7	2.5	56%	90%

- Coverage studies performed for various carriers and service providers.
- Topologies varied in terms of link-to-node ratio, protocols (RSVP-TE FRR in inner core and LDP in outer core), flat and hierarchical topologies.
- Per-prefix LFA coverage also computed and included for comparison.



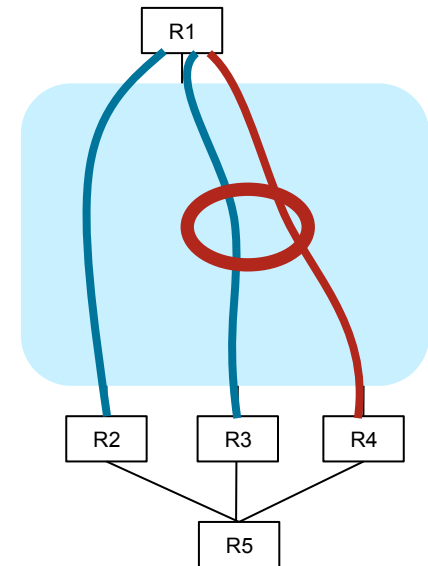


# IP/Optical Integration



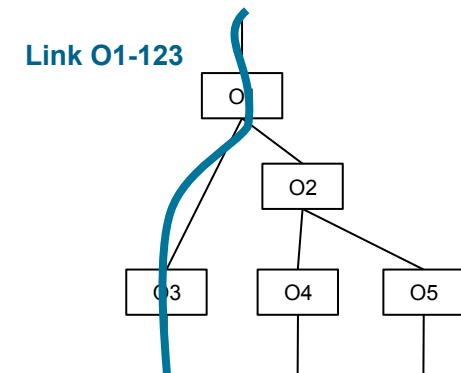
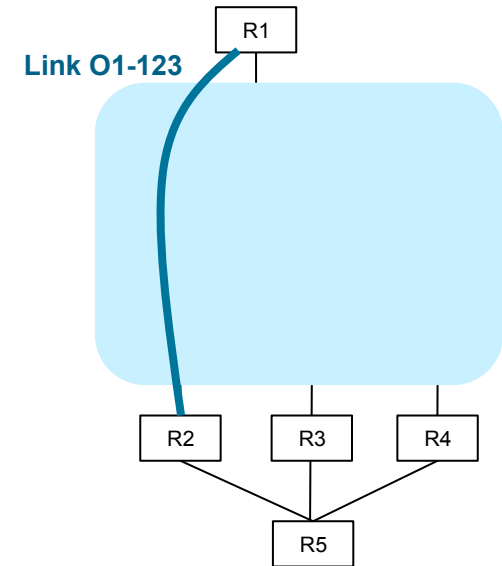
# SRLG

- To backup R1R4, R2 or R3?
- R2: disjoint optical path!



# Circuit ID

- Multi-Layer Planning optimization requires mapping circuits between L3 and L0 topologies
- Circuit ID acts as glue between L3 topology and underlying L0 topology
- Other applications:
  - troubleshooting
  - disjointness



# SRLG and Circuit ID Discovery

- Current: retrieve info from optical NMS and map the SRLG's to L3 topology. Labor intensive.
- Near future: automated discovery from the router L3 control plane thanks to L3/L0 integration

# Fasted DWDM provisioning

Does OTN really make sense for  
Core Router Bypass?

Gerstel, Batchellor, Spraggs, Filsfils  
Cisco Systems

MPLS 2010



## What really help is faster DWDM provisioning

- ❑ Slow DWDM provisioning leads to provision spare bandwidth a long time in advance and hence the perception that IP core utilization is not high enough
- ❑ Fully reconfigurable DWDM layer play a significant role solving this issue





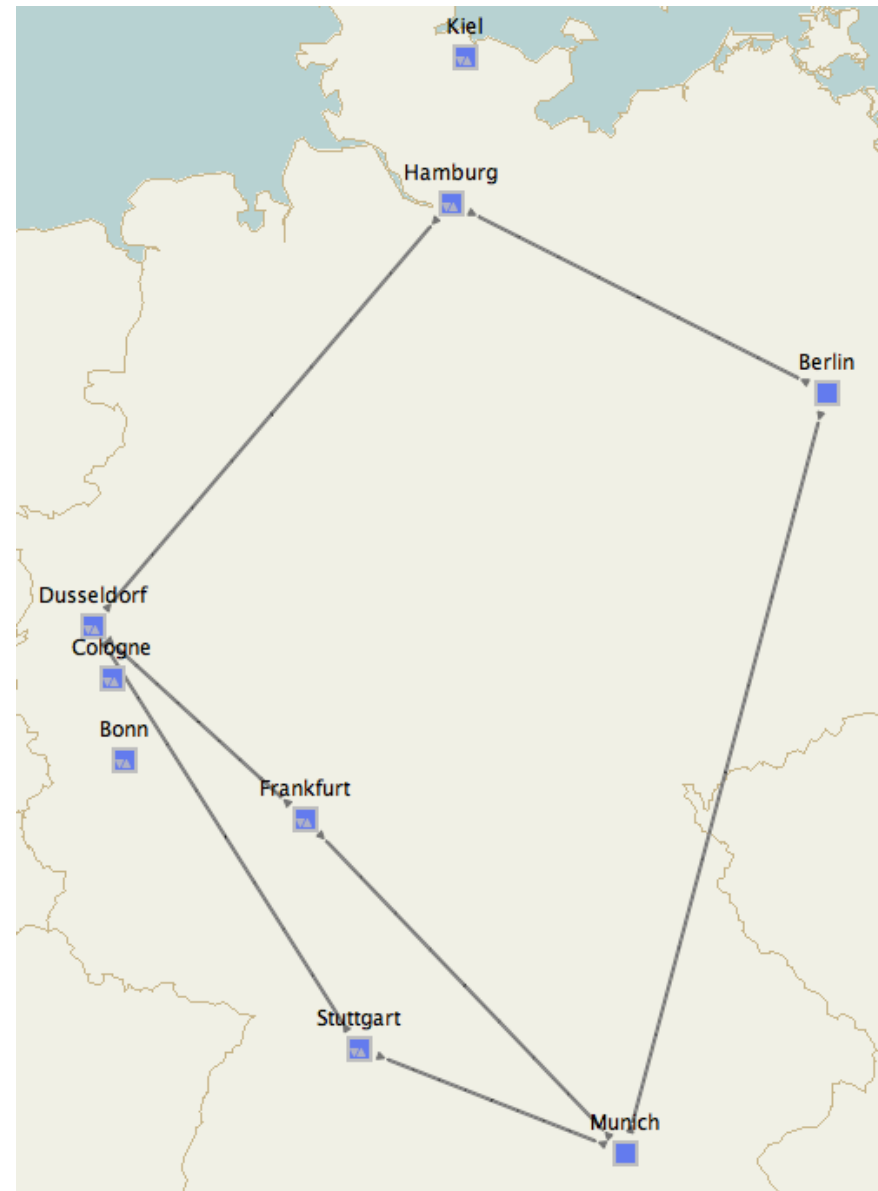
## A final example





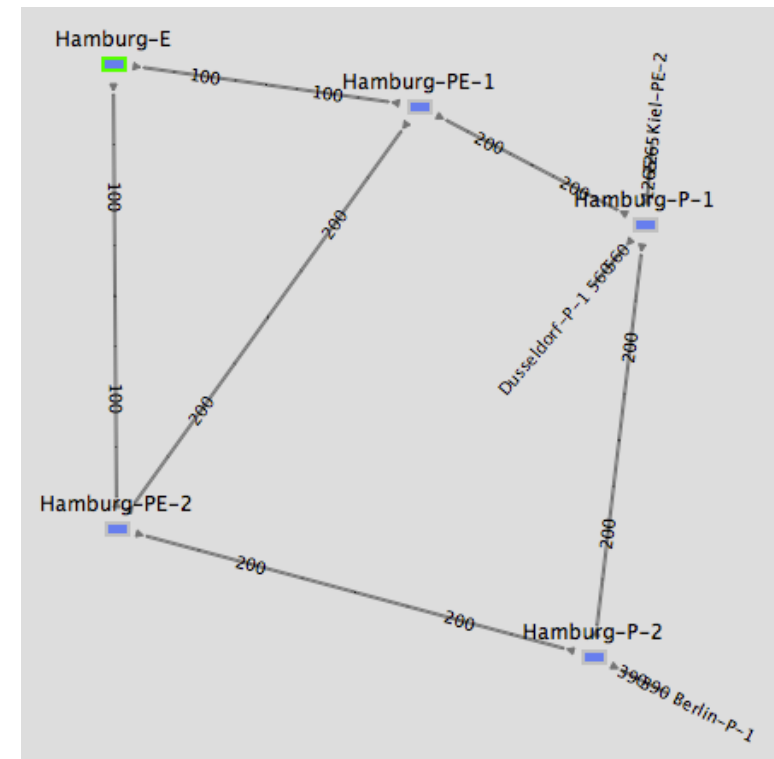
# Network Design

- For Mobile backbone
  - Fictional (German) topology
- IP over optical
- Projected Traffic Matrix
- Objectives:
  - Cost effective
  - Low delay
  - IPFRR LFA coverage
- Topology:
  - IP/Optical
  - 6 core sites



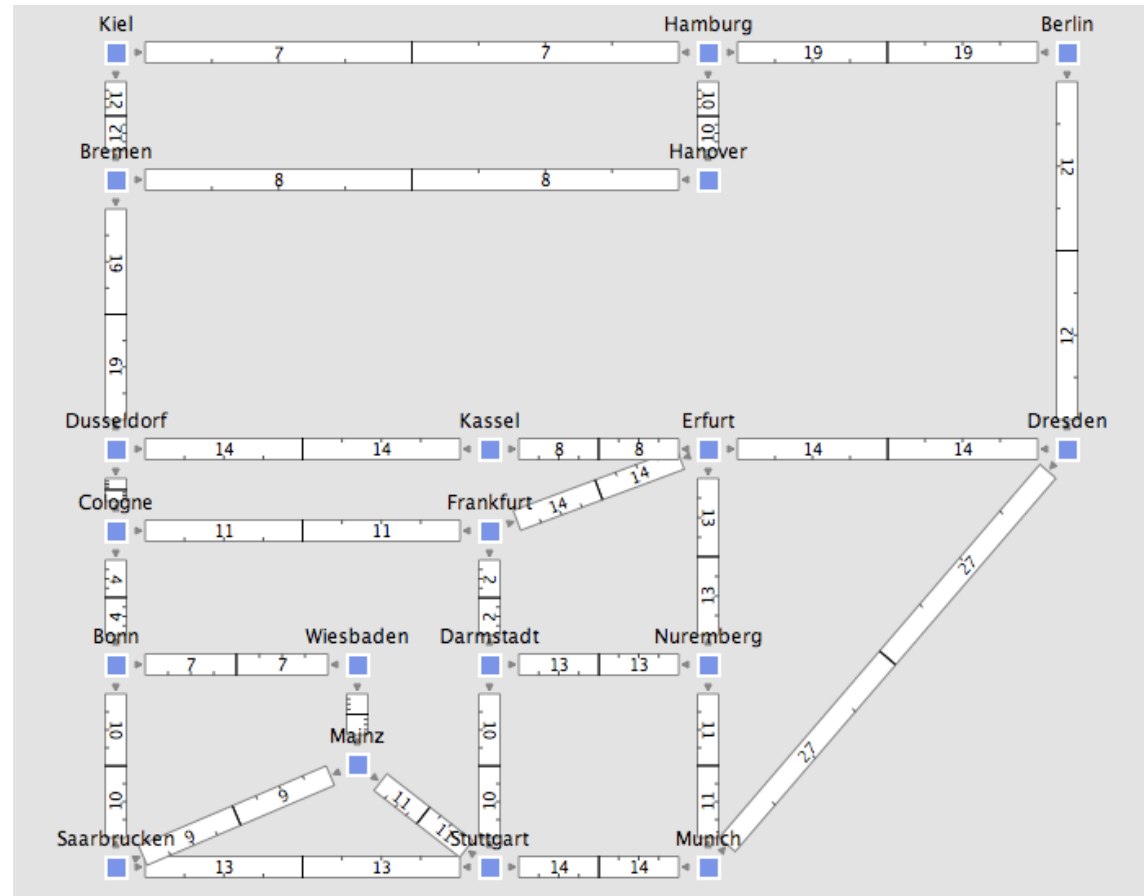
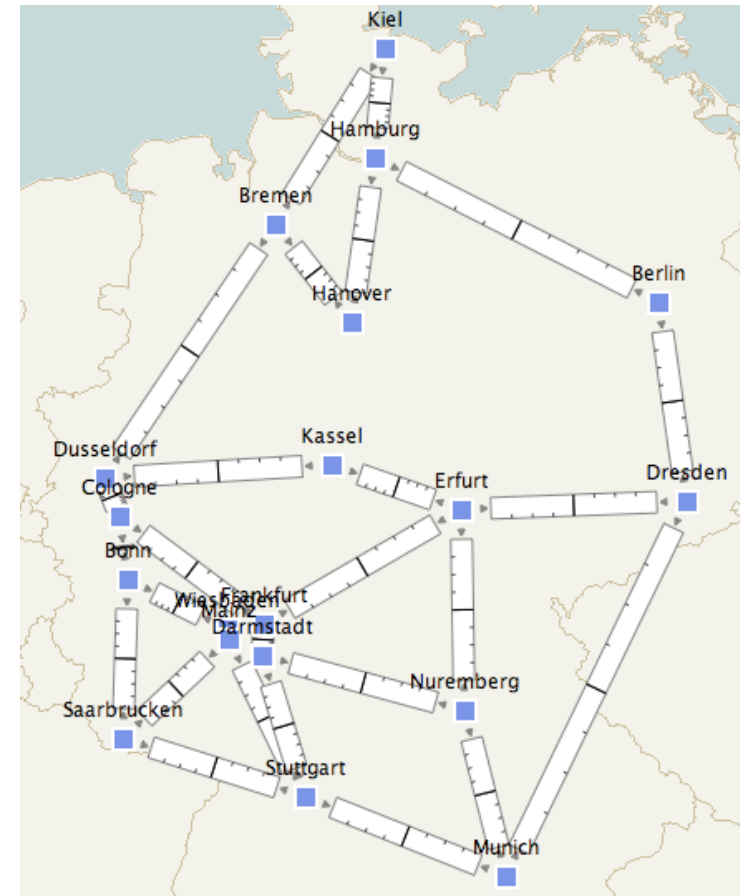
# Base Network Design

- Optical Design Rules:
  - Core links over shortest delay *diverse* optical path
  - Remote PE's homes into the closest P, and second closest P over *diverse* path
- IP Design Rules
  - 2 P-routers in core sites, 2 PE-routers in all sites
  - E(dge)-routers represent traffic sources (behind PE's)
  - Lowest Delay routing:
    - IGP metrics inter-site links:  $10 * \text{delay}$



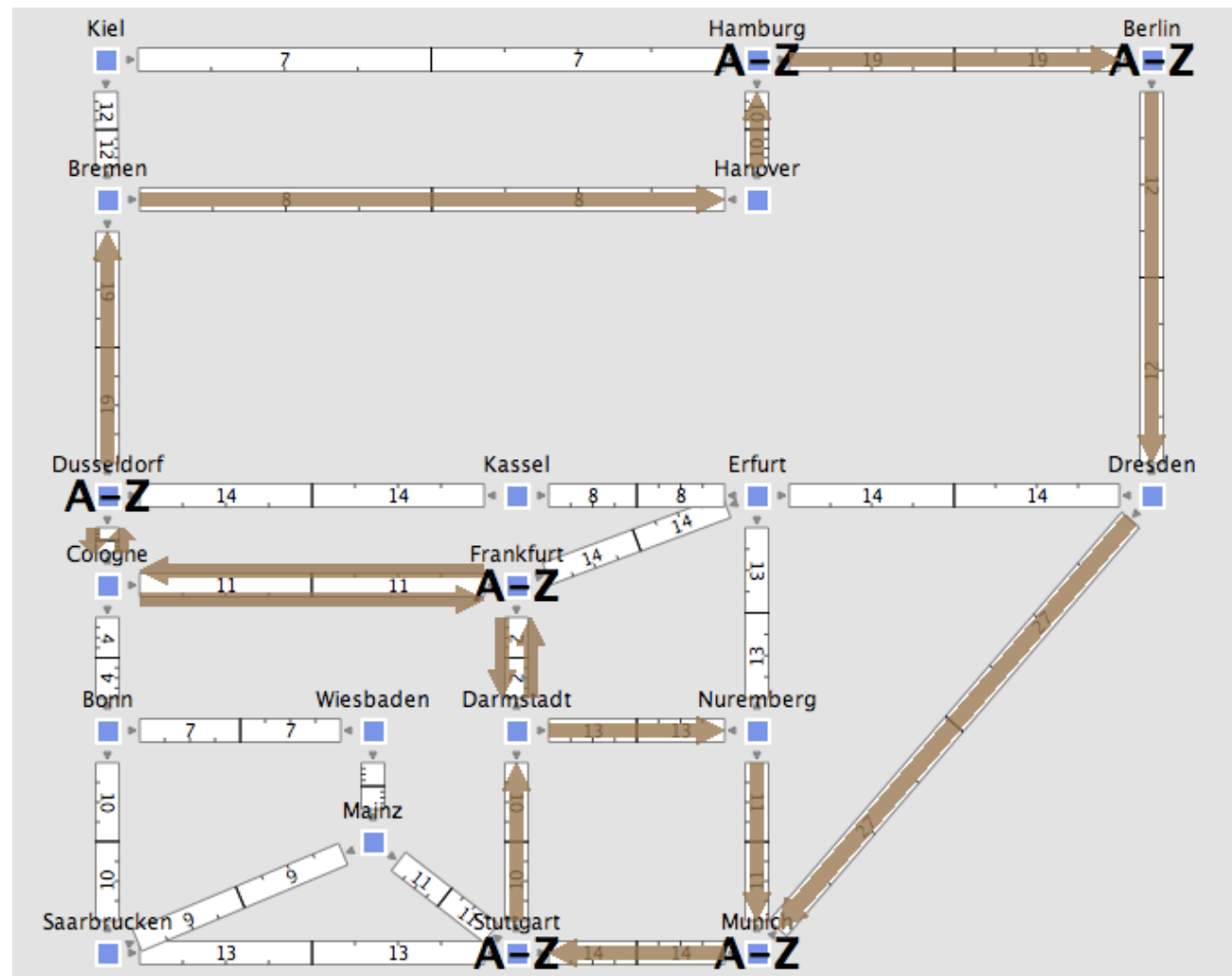
IGP metrics intra-site according to  
'draft-filsfils-rtgwg-lfa-applicability-00'

© 2015 Pearson Education, Inc. or its affiliate(s). All rights reserved.

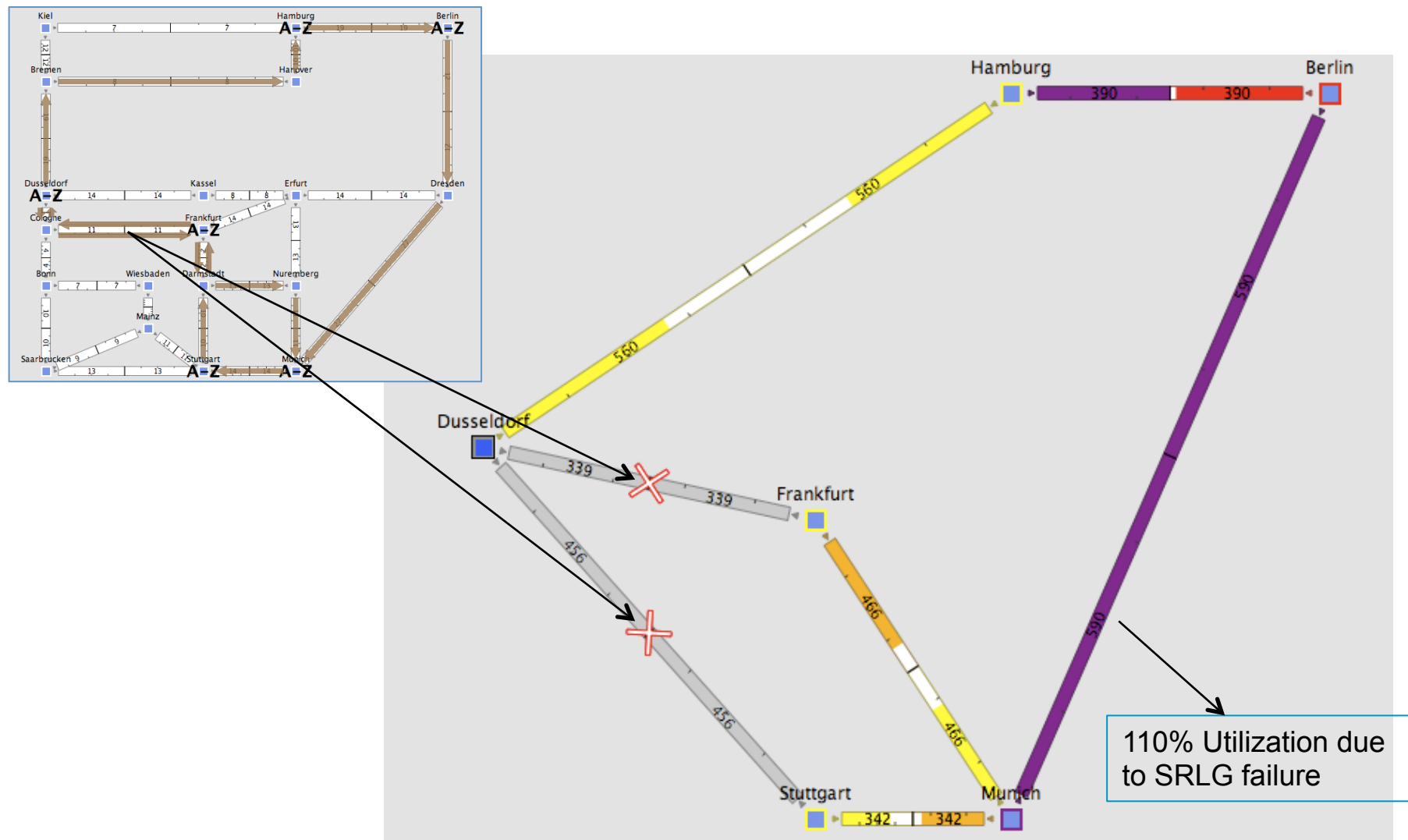


# Circuit routing over optical network

- 6 core sites
- IP circuits routed over shortest delay paths
- Note: fiber used for more than one circuit around Frankfurt

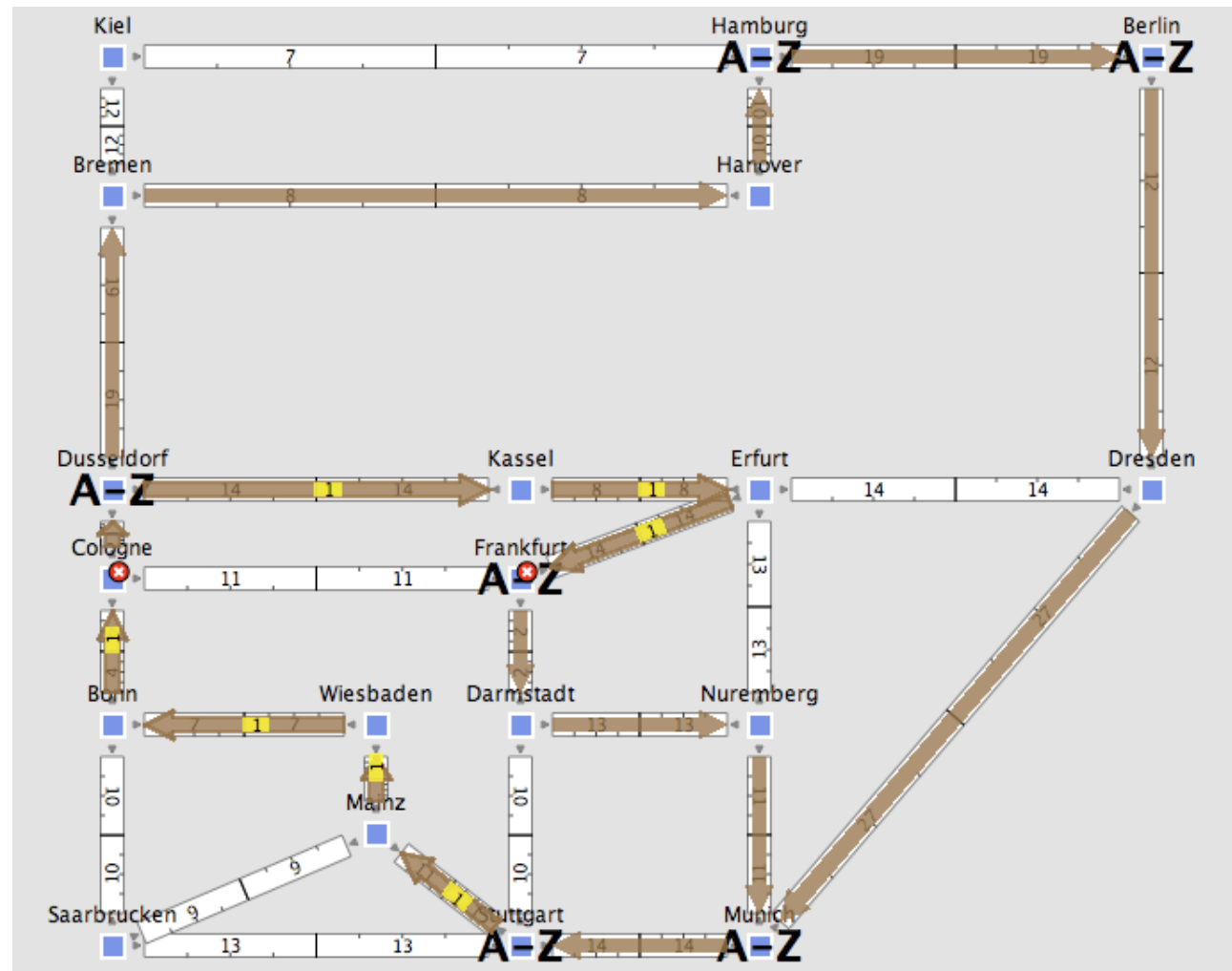


# SRLGs on IP layer



# Create diverse routing on optical layer

- Move *Dusseldorf-Stuttgart* away from *Frankfurt*
- Move *Dusseldorf-Frankfurt* away from *Cologne*

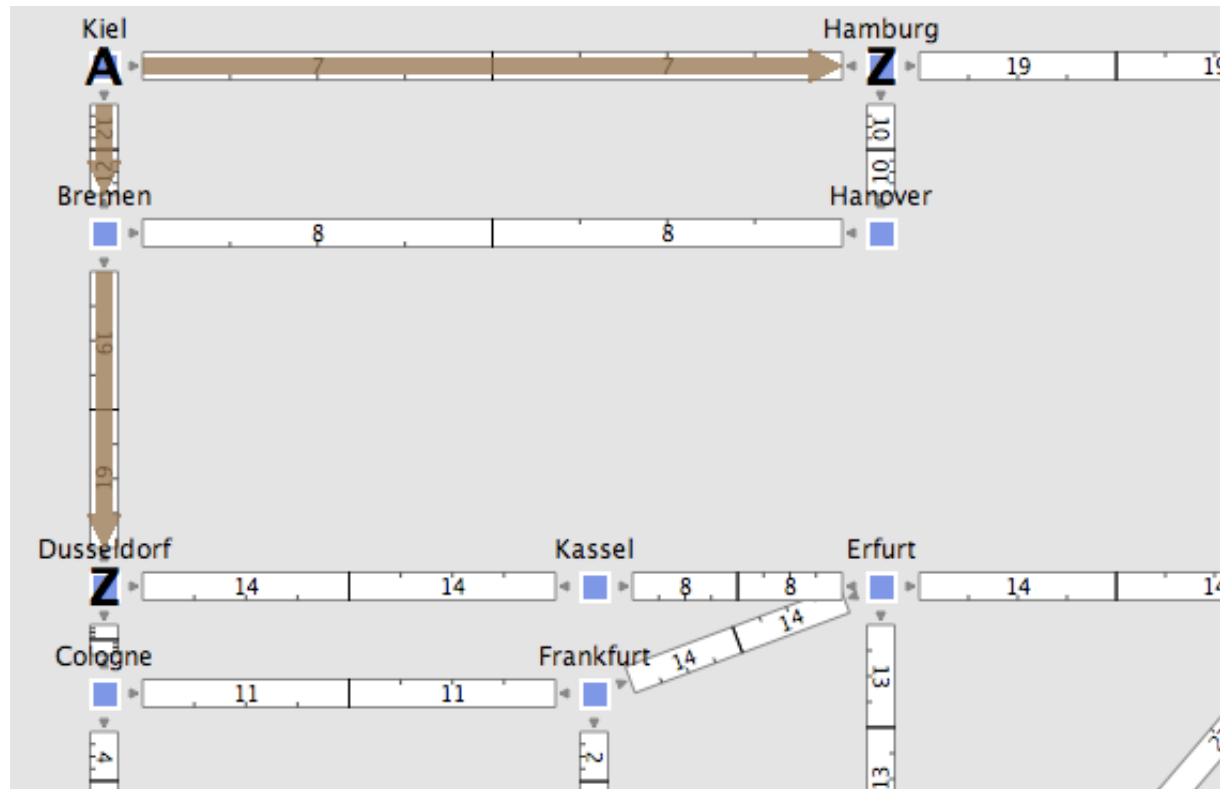


- 1. Kiel

- Closest PE is Hamburg

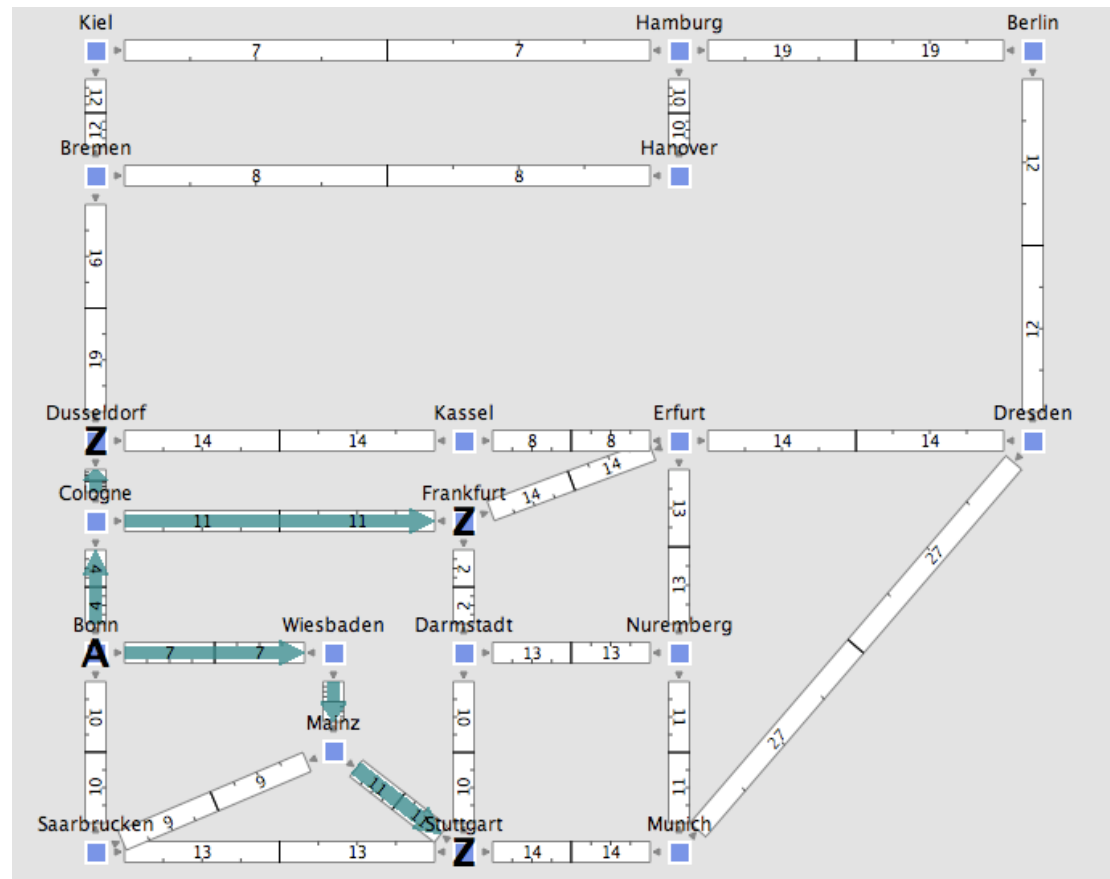
- 2<sup>nd</sup> closest Dusseldorf

- Diverse!



# Add Remote PE's

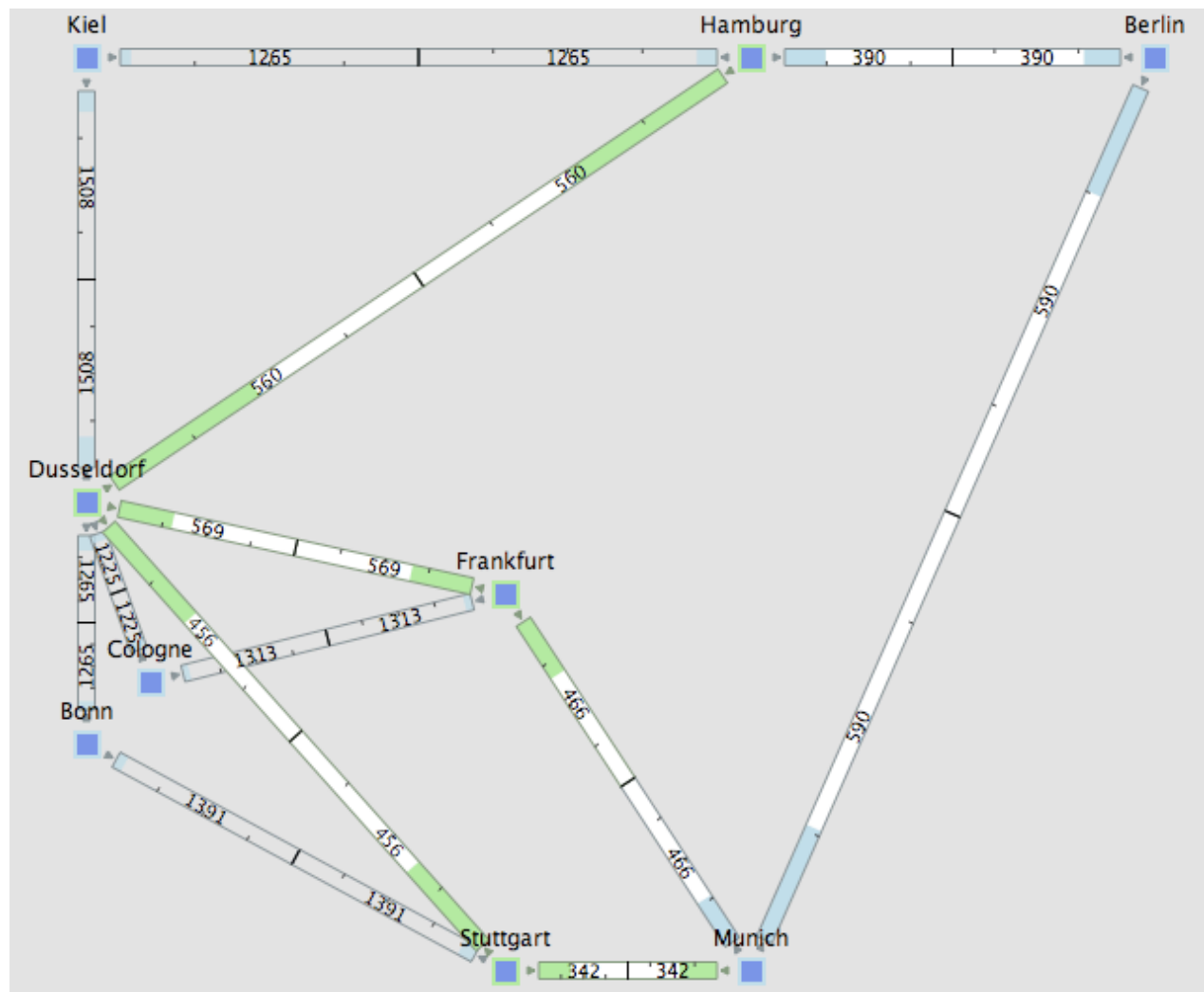
- 2. Bonn
  - Closest PE is Dusseldorf
  - 2<sup>nd</sup> closest Frankfurt: *but not diverse*
  - Excluding the links Bonn-Cologne and Cologne-Dusseldorf, Stuttgart is 2<sup>nd</sup> closest PE
- 3. etc...





# Final IP topology

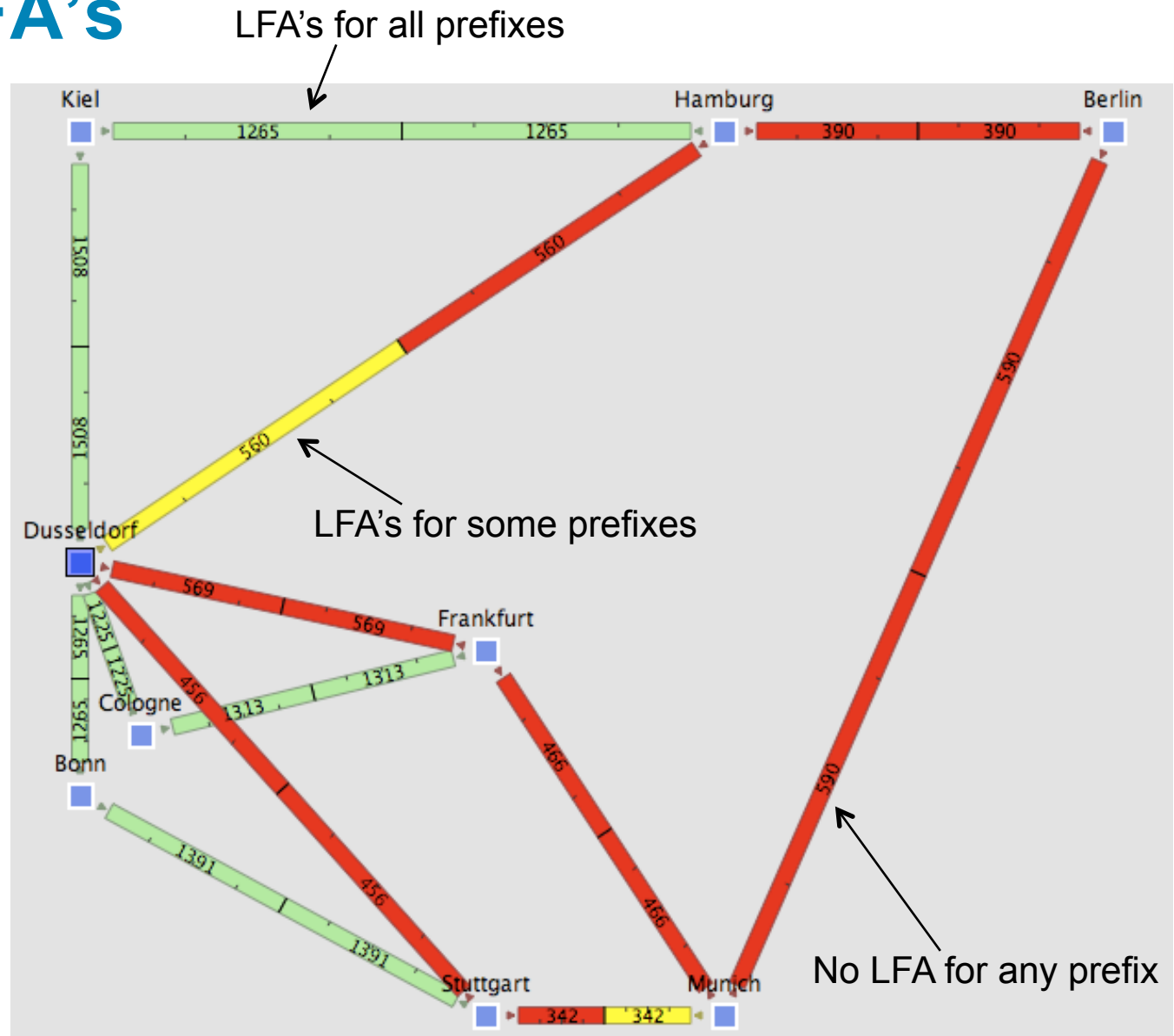
- Highest utilization due to any circuit or SRLG failure is 90%
- Saving of 20% due to diversity of Dusseldorf-Frankfurt and Dusseldorf-Stuttgart



# IPFRR LFA's

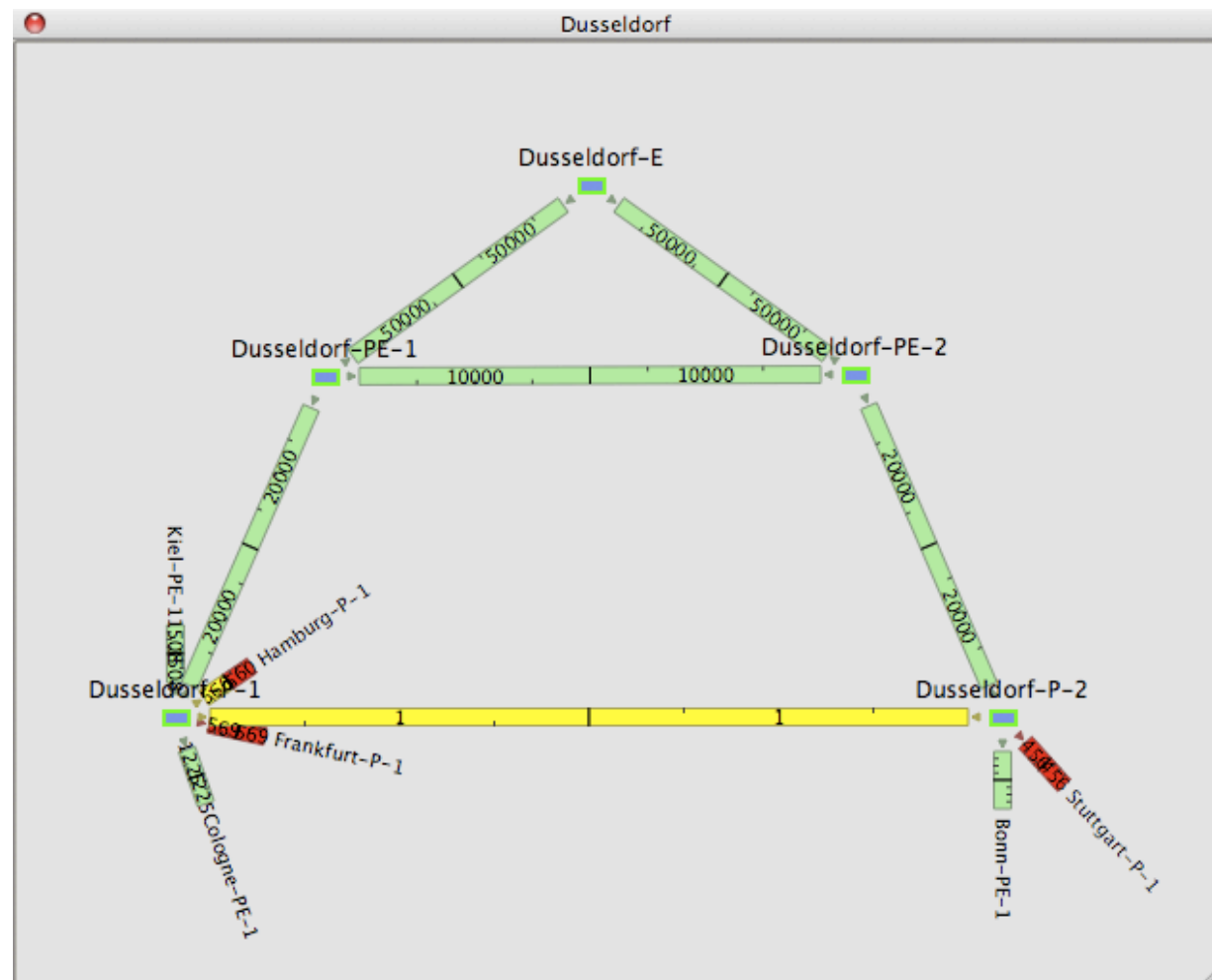
- 75% of interface traffic has an LFA available
- Some inter-site links are not protected due to ring topology

(Results from MATE Add-On.  
Contact Cariden for free  
analysis.)



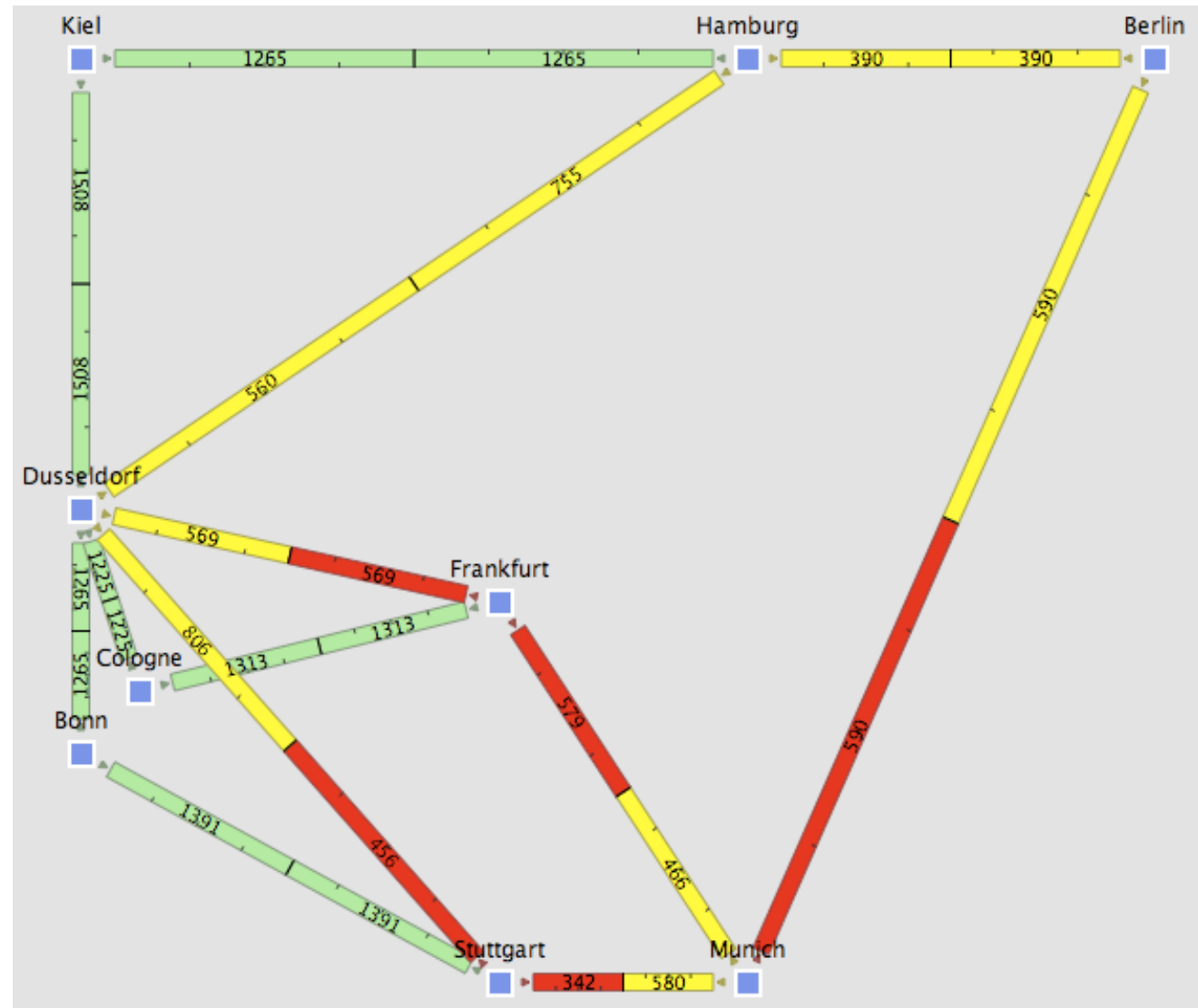
# IPFRR LFA's: site view

- LFA applicability draft section 3.3: Square



# IPFRR LFA's: metric optimization

- IPFRR coverage on core links has improved
- Average delay went up with 0.2 ms





# Conclusion



# Conclusion

Capacity Planning is essential for ensuring SLA within reasonable cost

- Routers provide
  - Traffic Matrix (neftflow)
  - Topology (LSDB)
  - QoS and Routing Policy
  - near-future: IP/Optical integrated data
- Planning tools provide
  - Automated Discovery
  - Traffic Matrix Deduction
  - Simulation and Optimization

Service provider staff must provide the vision and fortitude to “steer the ship” towards best practices.

# References

- [Filsfils and Evans 2005]
  - Clarence Filsfils and John Evans, "Deploying Diffserv in IP/MPLS Backbone Networks for Tight SLA Control", IEEE Internet Computing\*, vol. 9, no. 1, January 2005, pp. 58-65
  - <http://www.employees.org/~jevans/papers.html>
- [Deploy QoS]
  - Deploying IP and MPLS QoS for multiservice networks: theory and practice, By John Evans, Clarence Filsfils
  - [http://books.google.be/books?id=r6121tRwA6sC&pg=PA76&lpg=PA76&dq=book+deploying+qos+sp+filsfils&source=bl&ots=xauvtXLg3X&sig=f1NGddiXrZ\\_FAA3ZbRtoxVDiwPc&hl=en&ei=grDaTL6nBY3CsAOOsoHIBw&sa=X&oi=book\\_result&ct=result&resnum=1&ved=0CБУQ6AEwAA#v=onepage&q&f=false](http://books.google.be/books?id=r6121tRwA6sC&pg=PA76&lpg=PA76&dq=book+deploying+qos+sp+filsfils&source=bl&ots=xauvtXLg3X&sig=f1NGddiXrZ_FAA3ZbRtoxVDiwPc&hl=en&ei=grDaTL6nBY3CsAOOsoHIBw&sa=X&oi=book_result&ct=result&resnum=1&ved=0CБУQ6AEwAA#v=onepage&q&f=false)
- [Telkamp 2003]
  - Thomas Telkamp, "Backbone Traffic Management", Asia Pacific IP Experts Conference (Cisco), November 4th, 2003, Shanghai, P.R. China
  - [http://www.cariden.com/technology/white\\_papers/entry/backbone\\_traffic\\_management](http://www.cariden.com/technology/white_papers/entry/backbone_traffic_management)
- [Vardi 1996]
  - Y. Vardi. "Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data." J.of the American Statistical Association, pages 365–377, 1996.
- [Zhang et al. 2004]
  - Yin Zhang, Matthew Roughan, Albert Greenberg, David Donoho, Nick Duffield, Carsten Lund, Quynh Nguyen, and David Donoho, "How to Compute Accurate Traffic Matrices for Your Network in Seconds", NANOG29, Chicago, October 2004.
  - See also: <http://public.research.att.com/viewProject.cfm?prjID=133/>

# References

- [Gunnar et al.]
  - Anders Gunnar (SICS), Mikael Johansson (KTH), Thomas Telkamp (Global Crossing). “Traffic Matrix Estimation on a Large IP Backbone - A Comparison on Real Data”
  - [http://www.cariden.com/technology/white\\_papers/entry/traffic\\_matrix\\_estimation\\_on\\_a\\_large\\_ip\\_backbone\\_-\\_a\\_comparison\\_on\\_real\\_dat](http://www.cariden.com/technology/white_papers/entry/traffic_matrix_estimation_on_a_large_ip_backbone_-_a_comparison_on_real_dat)
- [Telkamp 2009]
  - “How Full is Full?”, DENOG 2009
  - [http://www.cariden.com/technology/white\\_papers/entry/how\\_full\\_is\\_full](http://www.cariden.com/technology/white_papers/entry/how_full_is_full)
- [Maghbouleh 2002]
  - Arman Maghbouleh, “Metric-Based Traffic Engineering: Panacea or Snake Oil? A Real-World Study”, NANOG 26, October 2002, Phoenix
  - <http://www.cariden.com/technologies/papers.html>
- [Cao 2004]
  - Jin Cao, William S. Cleveland, and Don X. Sun. “Bandwidth Estimation for Best-Effort Internet Traffic”
  - Statist. Sci. Volume 19, Number 3 (2004), 518-543.
- [MAXFLOW]
  - Maximum Flow problem
  - [http://en.wikipedia.org/wiki/Maximum\\_flow\\_problem](http://en.wikipedia.org/wiki/Maximum_flow_problem)