Best Practices for Determining the Traffic Matrix in IP Networks

Apricot 2005 - Kyoto, Japan

Tutorial, Monday February 21, 2005
16:00-17:30

Thomas Telkamp, *Cariden Technologies, Inc.*
Stefan Schnitter, *T-Systems*

Presenters and Contributors

- Presenters:
  - Thomas Telkamp, *Cariden Technologies, Inc.*
  - Stefan Schnitter, *T-Systems*

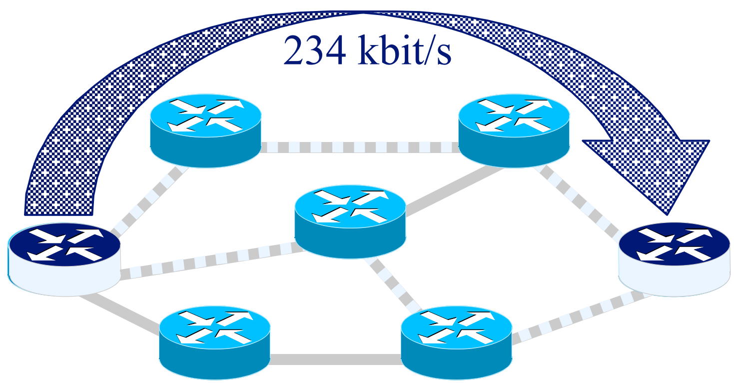
- Contributors:
  - Benoit Claise, *Cisco Systems, Inc.*
    - Cisco NetFlow
  - Tarun Dewan, *Juniper Networks, Inc.*
    - Juniper DCU
  - Mark Pommrehn, *Adlex, Inc.*
    - Adlex NetFlow collector deployment
  - Mikael Johansson, *KTH*
    - Traffic Matrix Estimation

Agenda

- Introduction
  - Traffic Matrix Properties
- Measurement in IP networks
  - NetFlow
  - NetFlow Deployment Case-Study
  - DCU (Juniper)
  - BGP Policy Accounting
- MPLS Networks
  - RSVP based TE
  - LDP
    - Data Collection
    - LDP deployment in Deutsche Telekom
- Traffic Matrices in Partial Topologies
- Estimation Techniques
  - Theory
  - Example Data
  - Case-Study
- Summary

Traffic Matrix

- Traffic matrix: the amount of data transmitted between every pair of network nodes
  - Demands
  - “end-to-end” in the core network
- Traffic Matrix can represent peak traffic, or traffic at a specific time
- Router-level or PoP-level matrices



Determining the Traffic Matrix

- Why do we need a Traffic Matrix?
  - Capacity Planning
    - Determine free/available capacity
    - Can also include QoS/CoS
  - Resilience Analysis
    - Simulate the network under failure conditions
  - Network Optimization
    - Topology
      - Find bottlenecks
    - Routing
      - IGP (e.g. OSPF/IS-IS) or MPLS

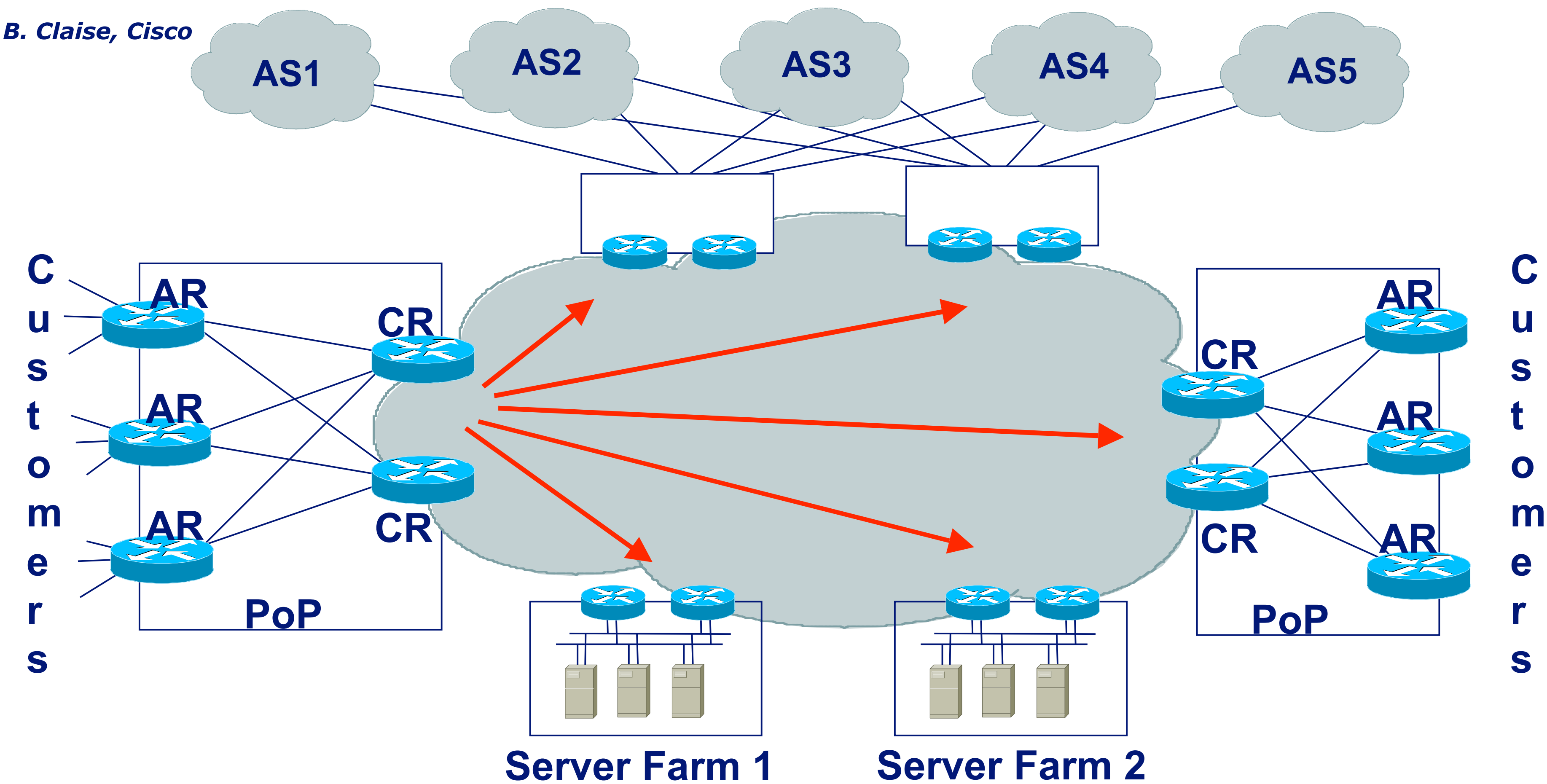
Types of Traffic Matrices

- Internal Traffic Matrix
  - PoP to PoP matrix
    - Can be from core (CR) or access (AR) routers
  - Class based

- External Traffic Matrix
  - PoP to External AS
    - BGP
    - Origin-AS or Peer-AS
      - Peer-AS sufficient for Capacity Planning and Resilience Analysis
    - Useful for analyzing the impact of external failures on the core network (capacity/resilience)

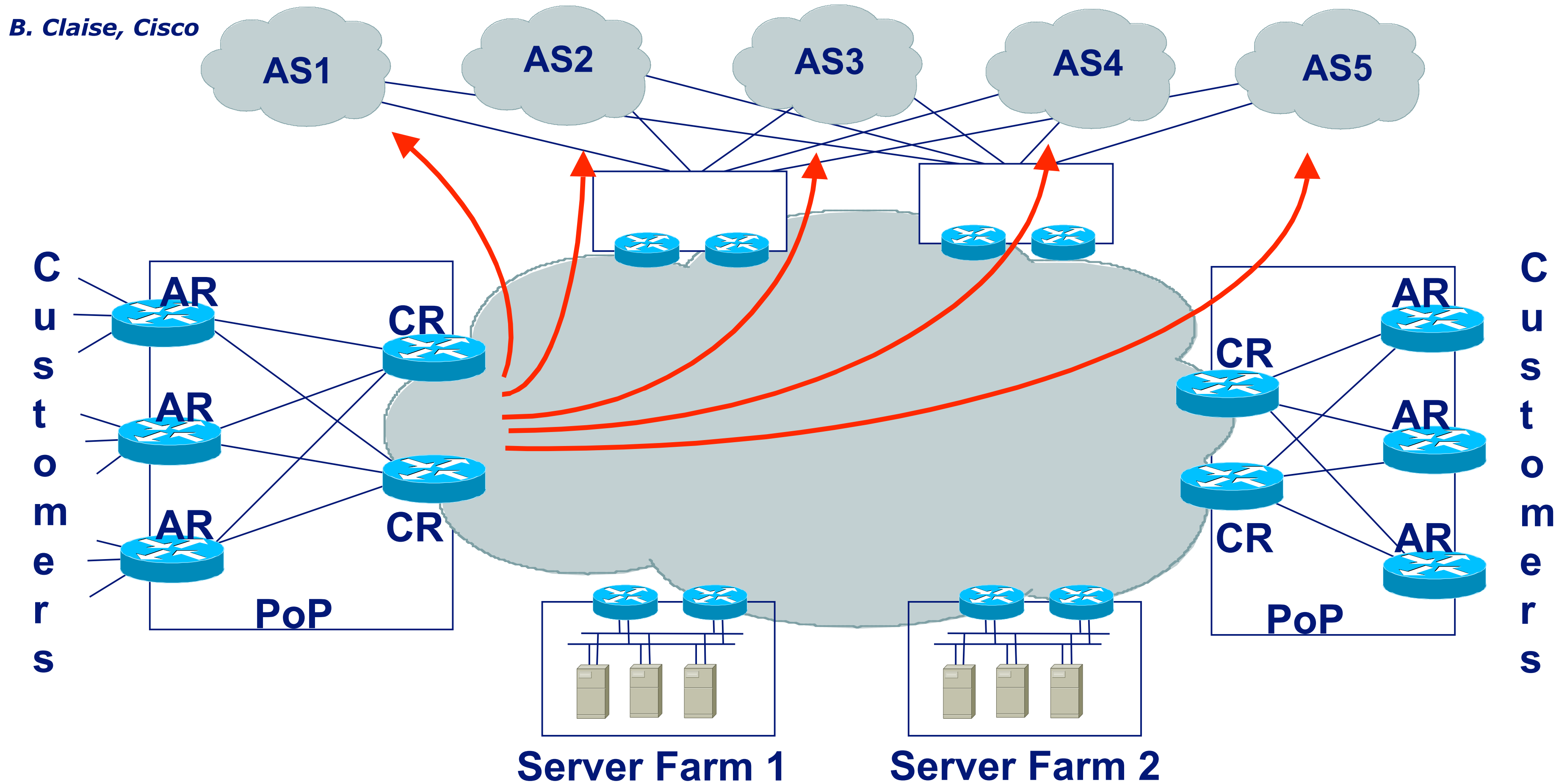
Internal Traffic Matrix

B. Claise, Cisco



"PoP to PoP", the PoP being the **AR** or **CR**

External Traffic Matrix



From "PoP to BGP AS", the PoP being the **AR** or **CR**

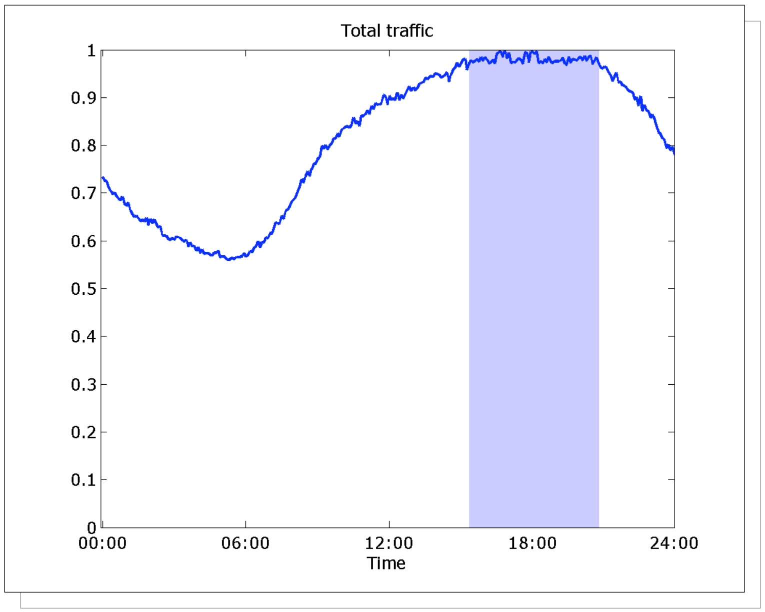
The external traffic matrix can influence the internal one

Traffic Matrix Properties

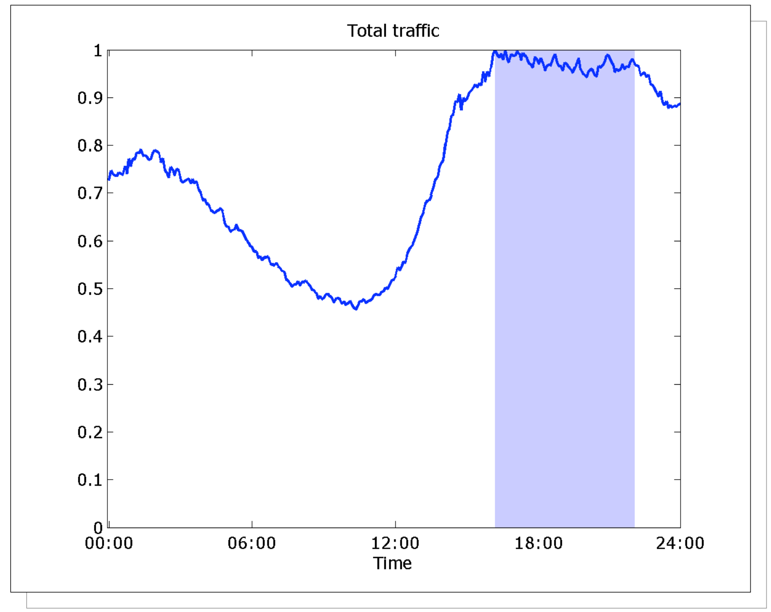
- Example Data from Tier-1 IP Backbone
  - Measured Traffic Matrix (MPLS TE based)
  - European and American subnetworks
  - 24h data
  - See [1]
- Properties
  - Temporal Distribution
    - How does the traffic vary over time
  - Spatial Distribution
    - How is traffic distributed in the network?
  - Relative Traffic Distribution
    - "Fanout"

Total traffic and busy periods

European subnetwork



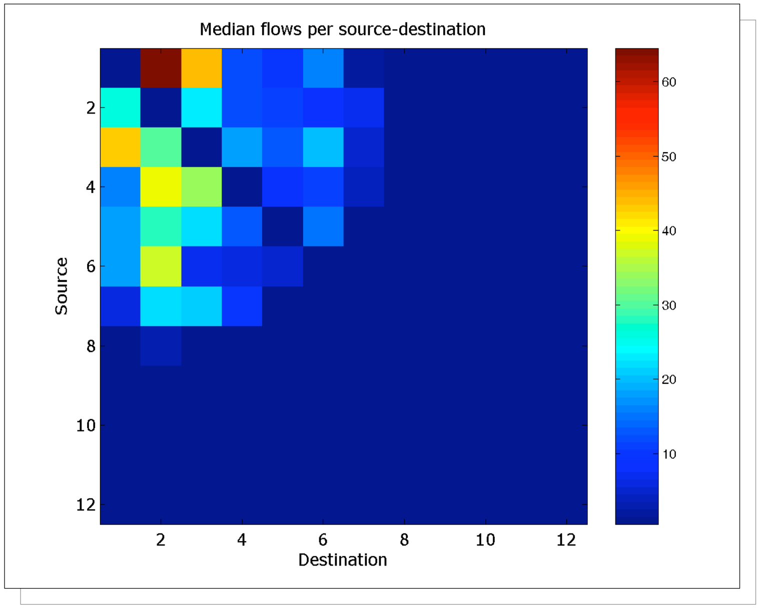
American subnetwork



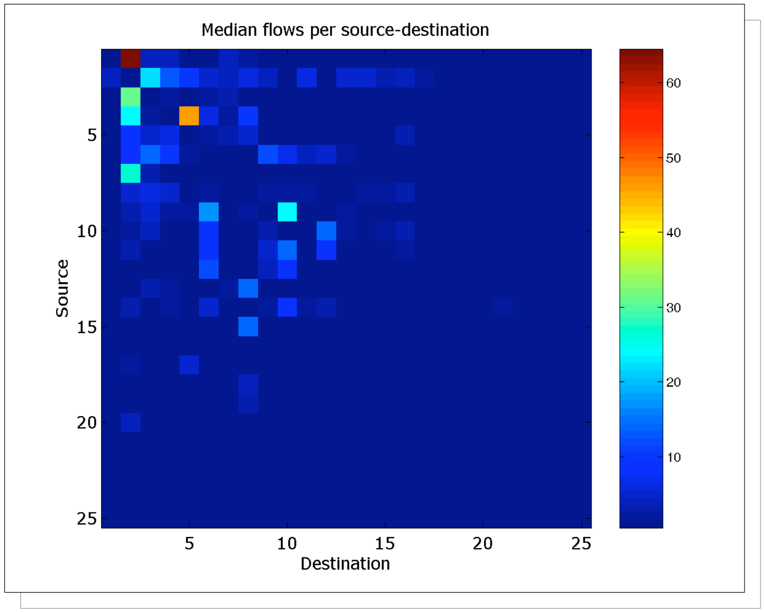
Total traffic very stable over 3-hour busy period

Spatial demand distributions

European subnetwork



American subnetwork

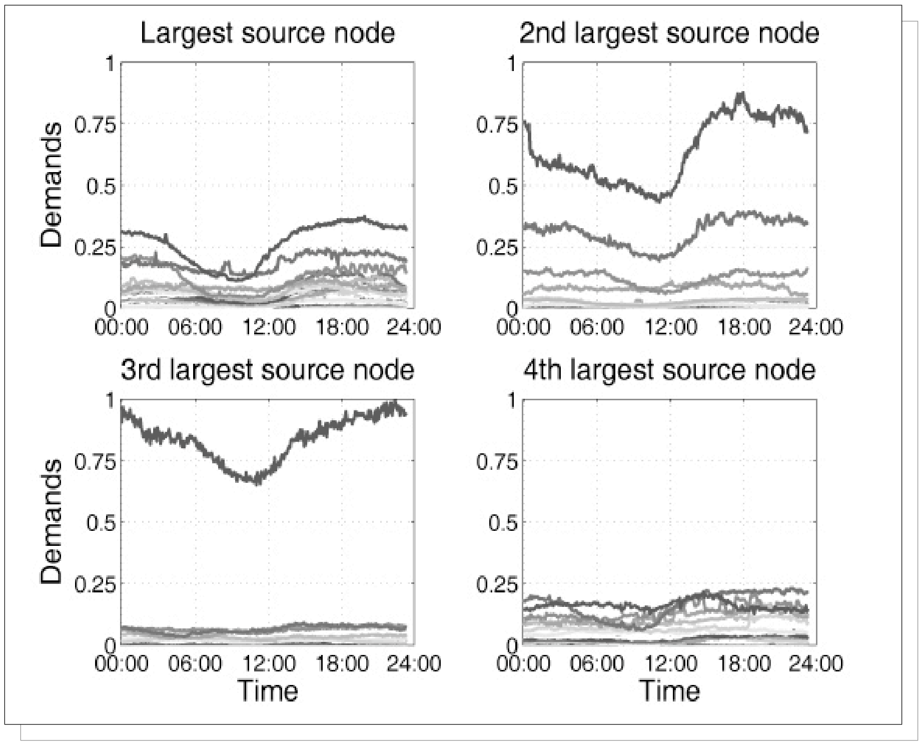


Few large nodes contribute to total traffic (20% demands – 80% of total traffic)

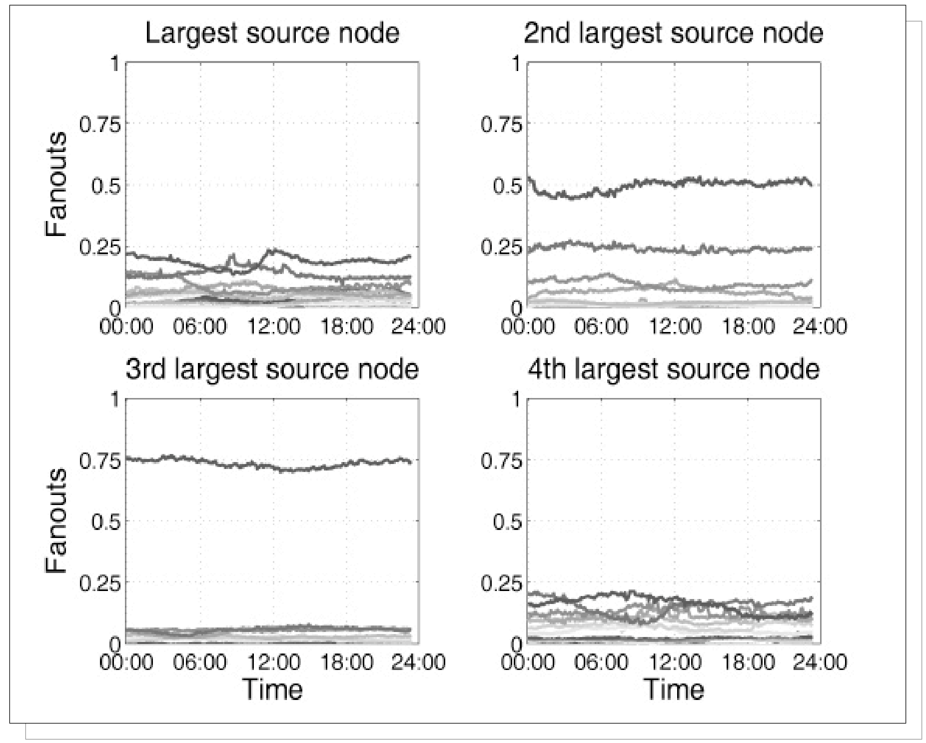
Fanout factors

Fanout: relative amount of traffic (as percentage of total)

Demands for 4 largest nodes, USA



Corresponding fanout factors



Fanout factors much more stable than demands themselves!

Traffic Matrix Collection

- Data is collected at fixed intervals
  - E.g. every 5 or 15 minutes
- Measurement of *Byte Counters*
  - Need to convert to rates
  - Based on measurement interval
- Create Traffic Matrix
  - Peak Hour Matrix
    - 5 or 15 min. average at the peak hour
  - Peak Matrix
    - Calculate the peak for every demand
    - Real peak or 95-percentile

Collection Methods

- NetFlow
  - Routers collect “flow” information
  - Export of raw or aggregated data
- DCU
  - Routers collect aggregated destination statistics
- MPLS
  - LDP
    - Measurement of LDP counters
  - RSVP
    - Measurement of Tunnel/LSP counters
- Estimation
  - Estimate Traffic Matrix based on Link Utilizations

NetFlow based Methods

NetFlow

- A “Flow” is defined by
  - Source address
  - Destination address
  - Source port
  - Destination port
  - Layer 3 Protocol Type
  - TOS byte
  - Input Logical Interface (ifIndex)
- Router keeps track of Flows and usage per flow
  - Packet count
  - Byte count

NetFlow Versions

- Version 5
  - the most complete version
- Version 7
  - on the switches
- Version 8
  - the Router Based Aggregation
- Version 9
  - the new flexible and extensible version

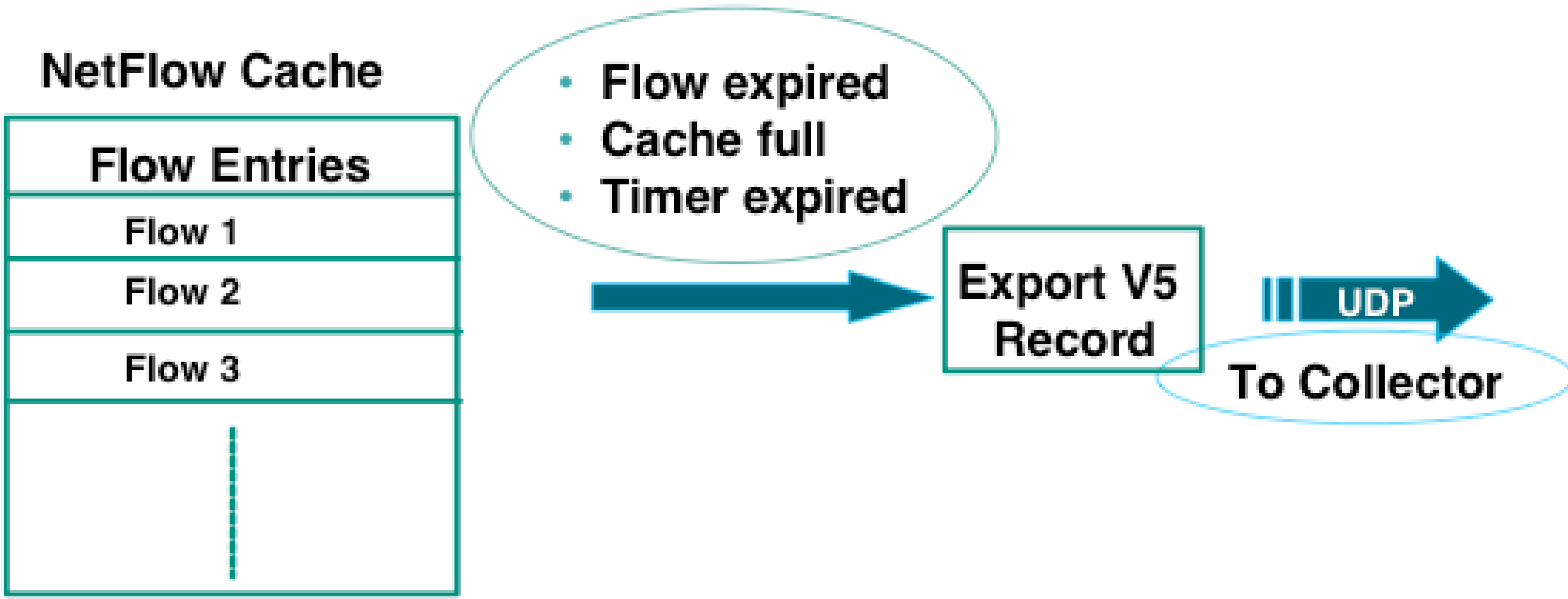
- Supported by multiple vendors
  - Cisco
  - Juniper
  - others

NetFlow Export

- A Flow is exported when
  - Flow expires
  - Cache full
  - Timer expired
- Expired Flows are grouped together into “NetFlow Export” UDP datagrams for export to a collector
  - Including timestamps
- UDP is used for speed and simplicity
- Exported data can include extra information
  - E.g. Source/Destination AS

NetFlow Export

B. Claise, Cisco



NetFlow Deployment

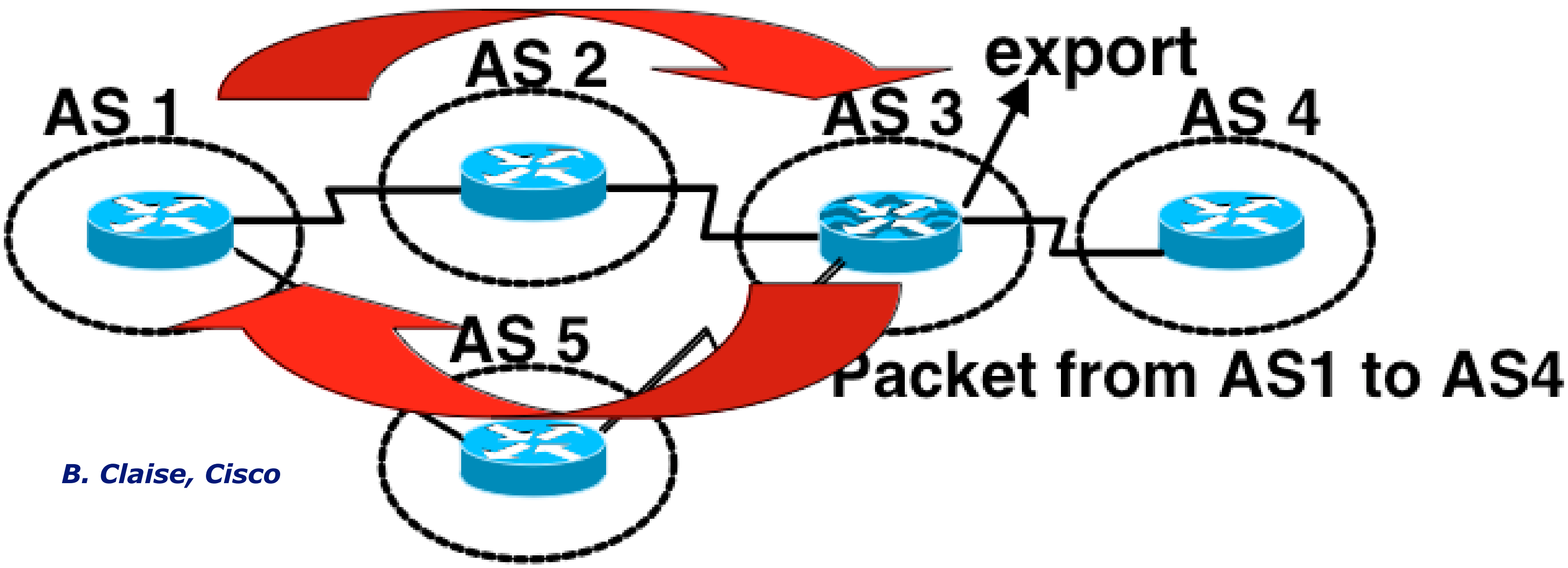
- How to build a Traffic Matrix from NetFlow data?
  - Enable NetFlow on all interfaces that source/sink traffic into the (sub)network
    - E.g. Access to Core Router links (AR->CR)
  - Export data to central collector(s)
  - Calculate Traffic Matrix from Source/Destination information
    - Static (e.g. list of address space)
    - BGP AS based
      - Easy for peering traffic
      - Could use “live” BGP feed on the collector
    - Inject IGP routes into BGP with community tag

BGP Passive Peer on the Collector

- Instead of exporting the peer-as or destination-as for the source and destination IP addresses for the external traffic matrix:
  - Don't export any BGP AS's
  - Export version 5 with IP addresses or version 8 with an prefix aggregation
- A BGP passive peer on the NetFlow collector machines can return all the BGP attributes:
  - source/destination AS, second AS, AS Path, BGP communities, BGP next hop, etc…
- Advantages:
  - Better router performance – less lookups
  - Consume less memory on the router
  - Full BGP attributes flexibility

NetFlow: Asymetric BGP traffic

- Origin-as
  - Source AS1, Destination AS4
- Peer-as
  - Source AS5, Destination AS4 WRONG!
- Because of the source IP address lookup in BGP



B. Claise, Cisco

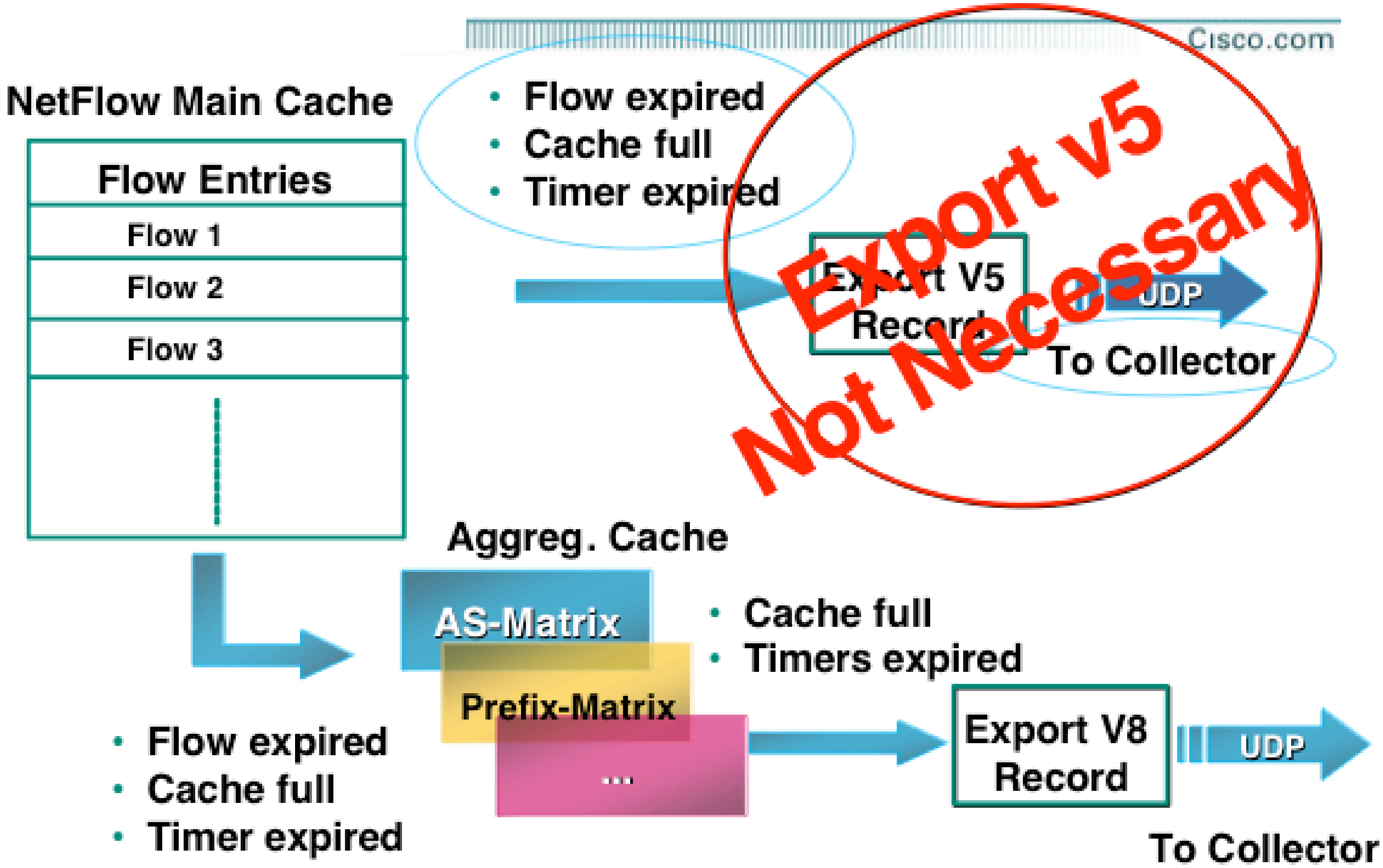
NetFlow Version 8

- Router Based Aggregation
- Enables router to summarize NetFlow Data
- Reduces NetFlow export data volume
  - Decreases NetFlow export bandwidth requirements
  - Makes collection easier

- Still needs the main (version 5) cache
- When a flow expires, it is added to the aggregation cache
  - Several aggregations can be enabled at the same time
- Aggregations:
  - Protocol/port, AS, Source/Destination Prefix, etc.

NetFlow: Version 8 Export

B. Claise, Cisco

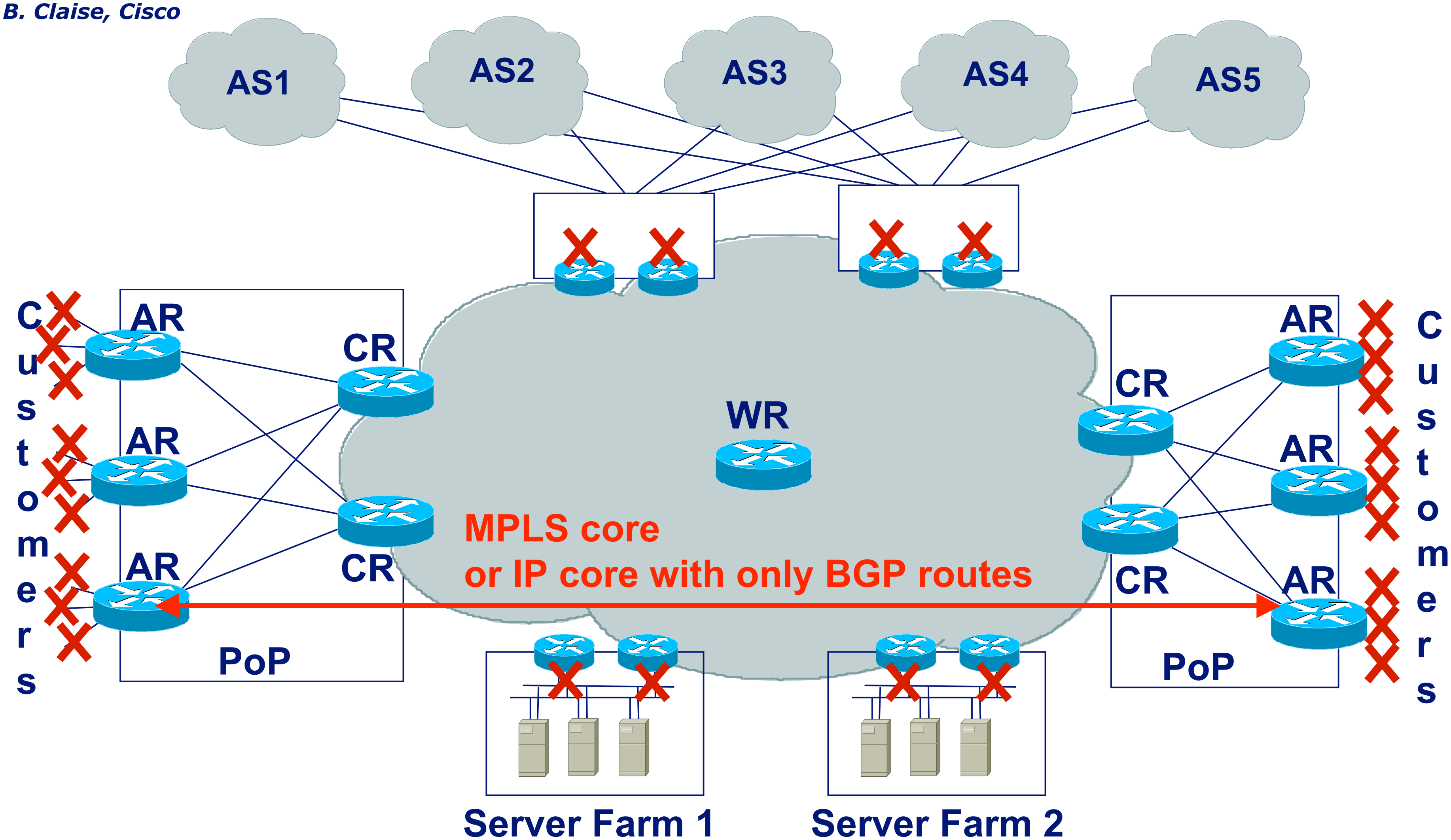


BGP NextHop TOS Aggregation

- New Aggregation scheme
  - Only for BGP routes
    - Non-BGP routes will have next-hop 0.0.0.0
- Configure on Ingress Interface
- Requires the new Version 9 export format
- Only for IP packets
  - IP to IP, or IP to MPLS

BGP NextHop TOS Aggregation

B. Claise, Cisco

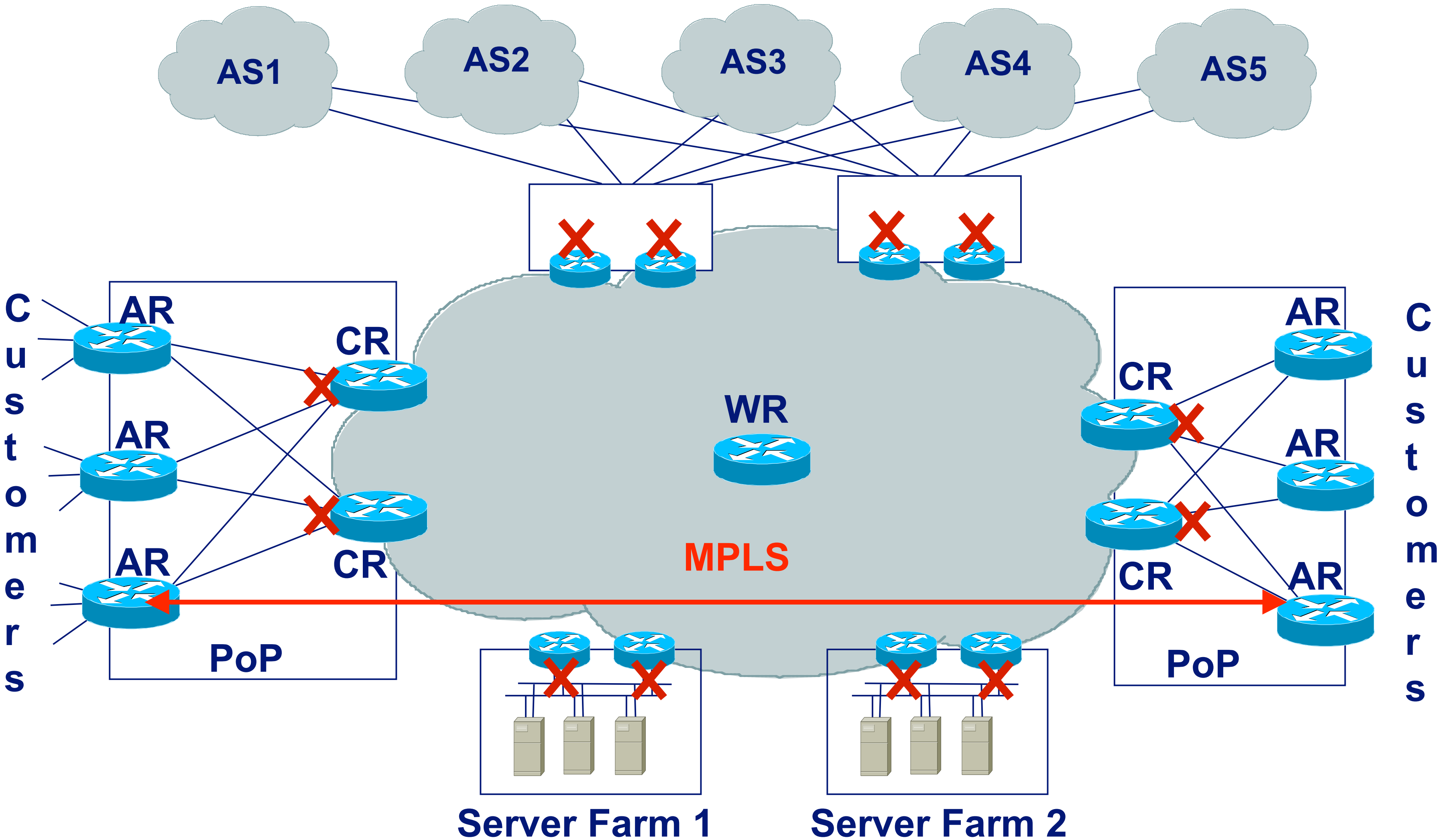


MPLS aware NetFlow

- Provides flow statistics per MPLS and IP packets
  - MPLS packets:
    - Labels information
    - And the V5 fields of the underlying IP packet
  - IP packets:
    - Regular IP NetFlow records
- Based on the NetFlow version 9 export
  No more aggregations on the router (version 8)
- Configure on ingress interface
- Supported on sampled/non sampled NetFlow

MPLS aware NetFlow: Example

B. Claise, Cisco



NetFlow Summary

- Building a Traffic Matrix from NetFlow data is not trivial
  - Need to correlate Source/Destination information with routers or PoPs
- “origin-as” vs “peer-as”
  - Asymetric BGP traffic problem
- BGP NextHop aggregation comes close to directly measuring the Traffic Matrix
  - NextHops can be easily linked to a Router/PoP
  - BGP only
- NetFlow processing is CPU intensive on routers
  - Use Sampling
    - E.g. only use every 1 out of 100 packets
    - Accuracy of sampled data

NetFlow Summary

- Various other features are available
- Ask vendors (Cisco, Juniper, etc.) for details on version support and platforms
- For Cisco, see Benoit Claise's webpage:
  - http://www.employees.org/~bclaise/

NetFlow Case-Study

Deployment Scenario

- NetFlow deployment in a large ISP network ("ISP X") using Adlex FlowTracker
  - Traffic Engineering Analysis (TEA)
- Goal is to obtain an accurate Traffic Matrix
  - Router to Router matrix
- Internal Traffic sources/sinks
  - typically blocks of customer address space in PoPs, such as such as broadband access devices (DSL or Cable Modem termination systems, dedicated corporate Internet access routers, dial NASes, etc).
- External traffic sources/sinks
  - typically public or private peering links (eBGP connections) in peering centers or transit PoPs

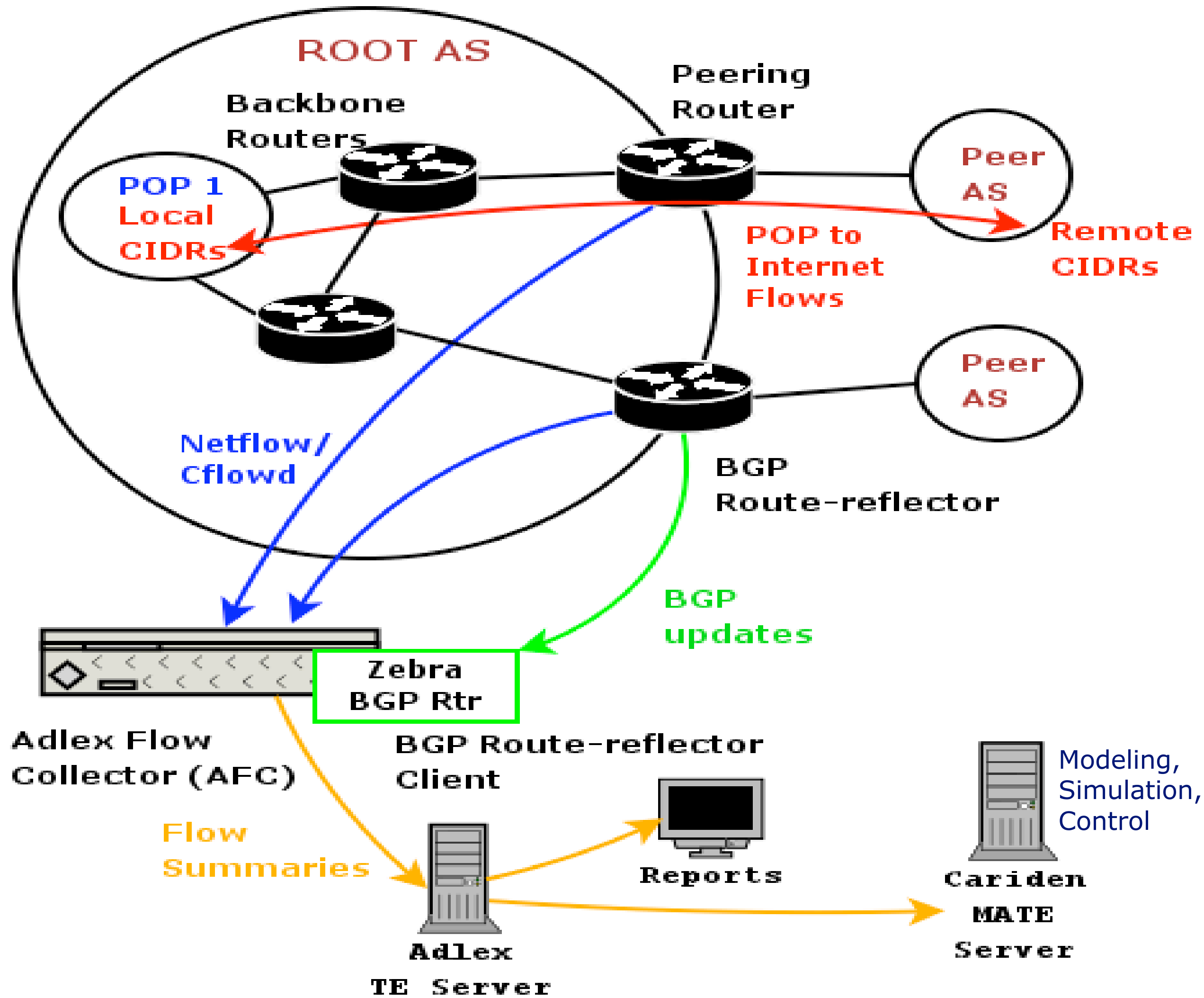
Associating Traffic with Routers

- Customer routes in each PoP are advertised into iBGP from the IGP
  - by each of the two backbone routers in each PoP
  - with the backone router's loopback address as the BGP Next Hop IP address for each of the local routes in the PoP
- The Adlex TEA system can pick them up from the BGP table via an integrated Zebra software router component in the Adlex Flow Collector (AFC)
- ISP uses Version 5 Netflow with Adlex Flow Collectors that are BGP Next-Hop aware at the local (PoP) and external (Internet) CIDR level

ISP X Phase 1: Internet Traffic

- Enable NetFlow on all interfaces on eBGP peering routers
  - Flows are captured at the Internet border, from the peering routers, as they pass through peering routers to/from Internet eBGP peers
- Adlex Traffic Engineering Report Server:
  - Retreives and aggregates summarized flows from multiple AFCs
  - Exports daily traffic matrix CSV files to Cariden MATE
    - For Modeling, Simulation and Control
- Hourly router-router values actually contain the the highest 15-minute average bandwidth period within that whole hour (4 periods/hour)
  - provides sufficient granularity to get near daily peak values between routers or PoPs

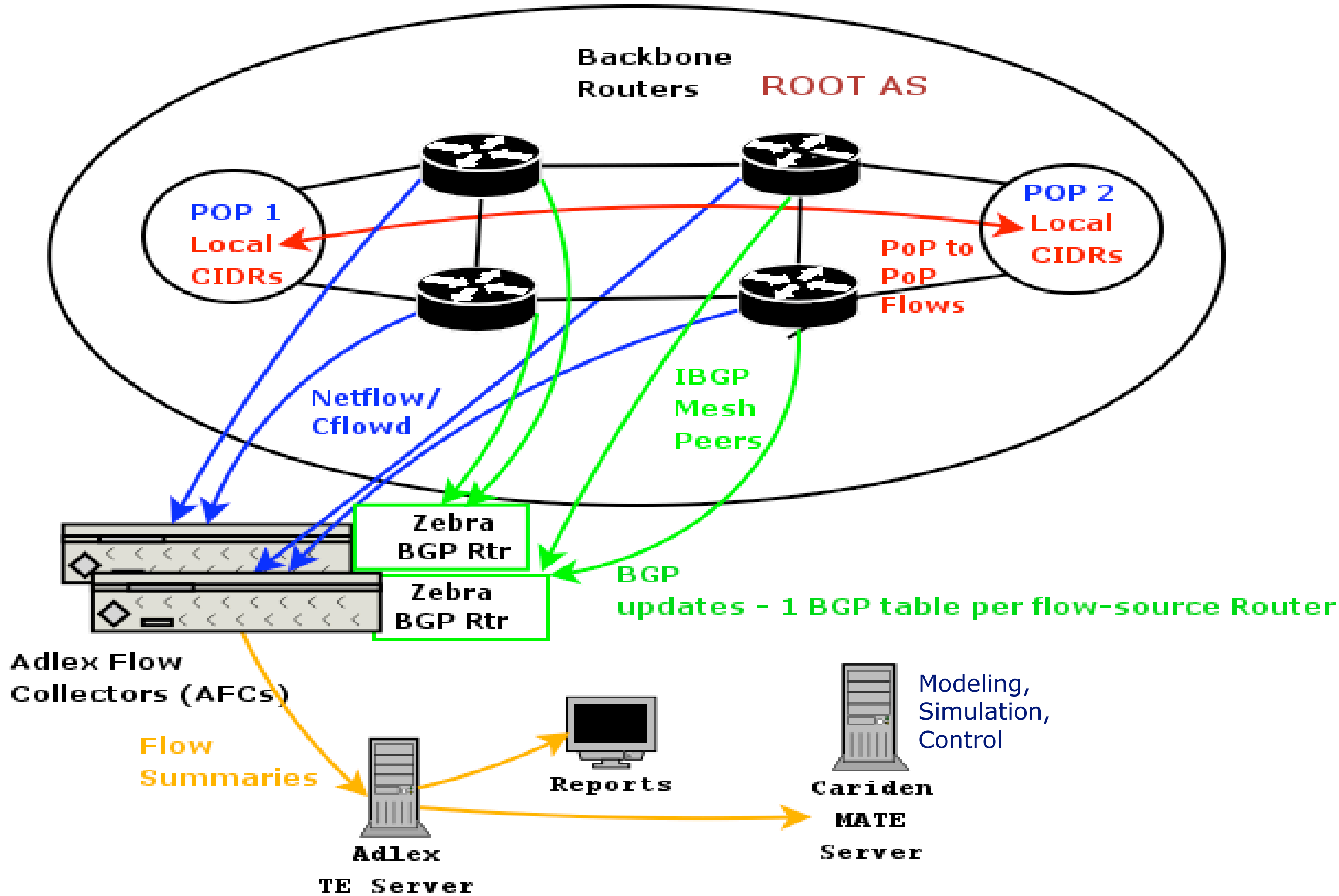
ISP X Phase 1: Internet Traffic



ISP X Phase 2: PoP-to-PoP Traffic

- Ingress-only Netflow exported from PoP-facing interfaces on Backbone routers
  - Enables capturing data flowing between POPs
- Flow assignment accuracy is optimized if each router that exports flows has those flows analyzed according to its own BGP table
  - Thus the traffic collection and analysis system must process a BGP table per-router
- BGP table per backbone router and per peering router

ISP X Phase 2: PoP-to-PoP Traffic



Example Results

Abbreviated header showing hourly columns:

```
  # Time Router A IP,Time Router B
IP,09/01/04 12:00:00 AM Max Bits/s Router A-
>B(bps),09/01/04 01:00:00 AM Max Bits/s
Router A->B(bps), 09/01/04 02:00:00 AM Max
Bits/s Router A->B(bps)
```

Abbreviated data showing source & dest router IPs and hourly max-15-minute values:

```
63.45.173.83,173.27.44.02,2639.64453125,2858
.09765625,15155.2001953125,10594.986328125,2
1189.97265625,8747.2353515625,104866.703125,
136815.5,31976.107421875,12642.986328125,851
0.578125,6489.88427734375,8192.0
```

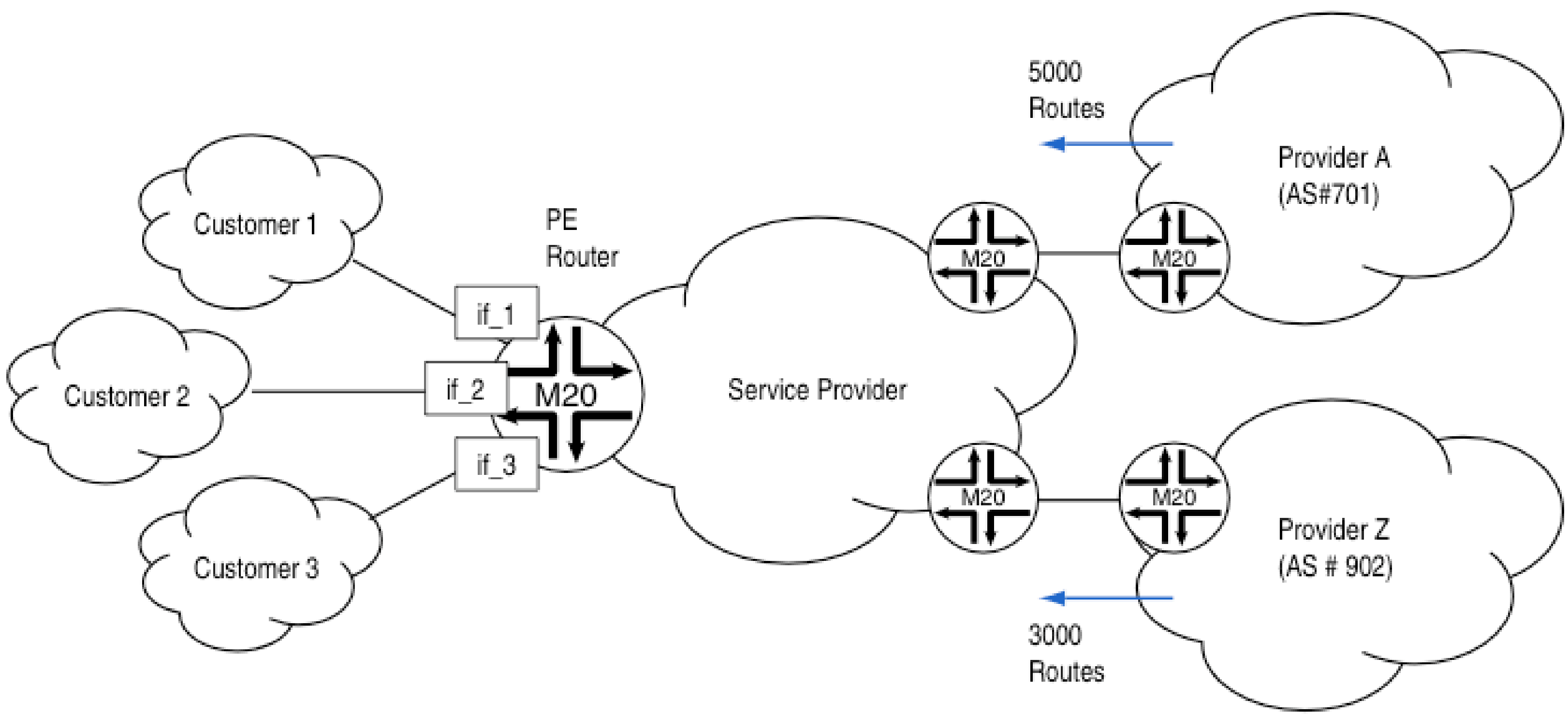
Destination Class Usage (DCU)

Destination Class Usage (DCU)

- Juniper specific!
- Policy based accounting mechanism
  - For example based on BGP communities
- Supports up to 16 different traffic destination classes
- Maintains per interface packet and byte counters to keep track of traffic per class
- Data is stored in a file on the router, and can be pushed to a collector
- But…
- *16 destination classes is in most cases too limited to build a useful full Traffic Matrix*

DCU Example

- Routing policy
  - associate routes from provider A with DCU class 1
  - associate routes from provider B with DCU class 2
- Perform accounting on PE



BGP Policy Accounting

BGP Policy Accounting

- Accounting traffic according to the route it traverses
- Account for IP traffic by assigning counters based on:
  - BGP community-list
  - AS number
  - AS-path
  - destination IP address
- 64 buckets
- Similar to Juniper DCU

MPLS Based Methods

MPLS Based Methods

Two methods to determine traffic matrices:

- Using RSVP-TE tunnels
- Using LDP statistics

Some comments on DT's practical implementation

Traffic Matrices in partial topologies

T-Systems

RSVP-TE in MPLS Networks

RSVP-TE (RFC 3209) can be used to establish LSPs

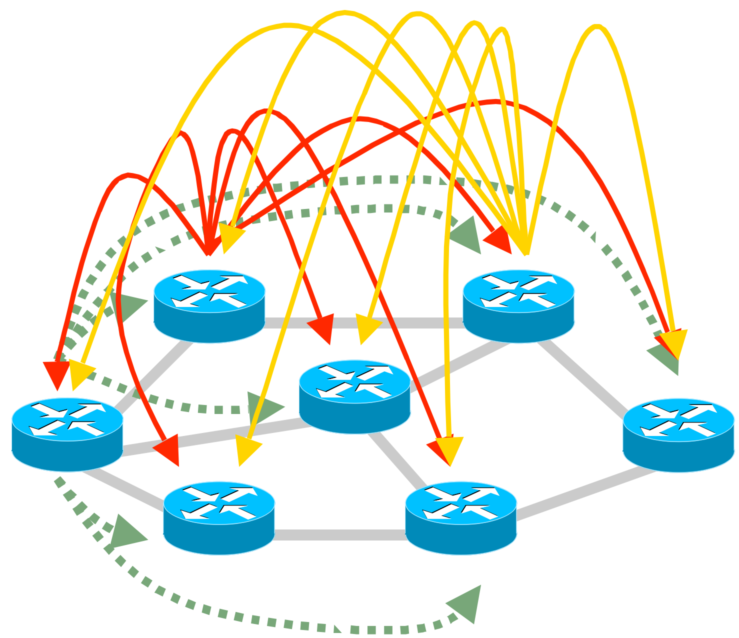
Example (IOS)

```
interface Tunne99
 description RouterA => RouterB
 tag-switching ip
 tunnel destination 3.3.3.3
 tunnel mode mpls traffic-eng
 tunnel mpls traffic-eng priority 5 5
 tunnel mpls traffic-eng bandwidth  1
 tunnel mpls traffic-eng path-option 3 explicit identifier 17
 tunnel mpls traffic-eng path-option 5 dynamic
!
ip explicit-path identifier 17 enable
 next-address 1.1.1.1
 next-address 2.2.2.2
 next-address 3.3.3.3
!
```

RSVP-TE in MPLS Networks
How to Obtain the Traffic Matrix (TM)?

Explicitly routed Label Switched Paths (TE-LSP) have associated byte counters;

A full mesh of TE-LSPs enables to measure the traffic matrix in MPLS networks directly;



RSVP-TE in MPLS Networks
Pro's and Con's

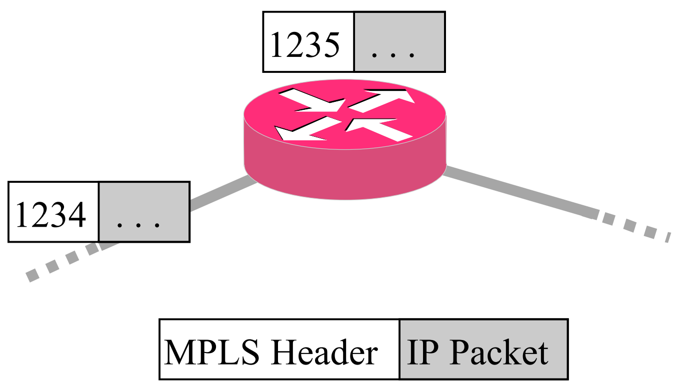
Advantage: Method that comes closest a traffic matrix measurement.

Disadvantages:

- A full mesh of TE-LSPs introduces an additional routing layer with significant operational costs;
- Emulating ECMP load sharing with TE-LSPs is difficult and complex:
  - Define load-sharing LSPs explicitly;
  - End-to-end vs. local load-sharing;
- Only provides Internal Traffic Matrix, no Router/PoP to peer traffic

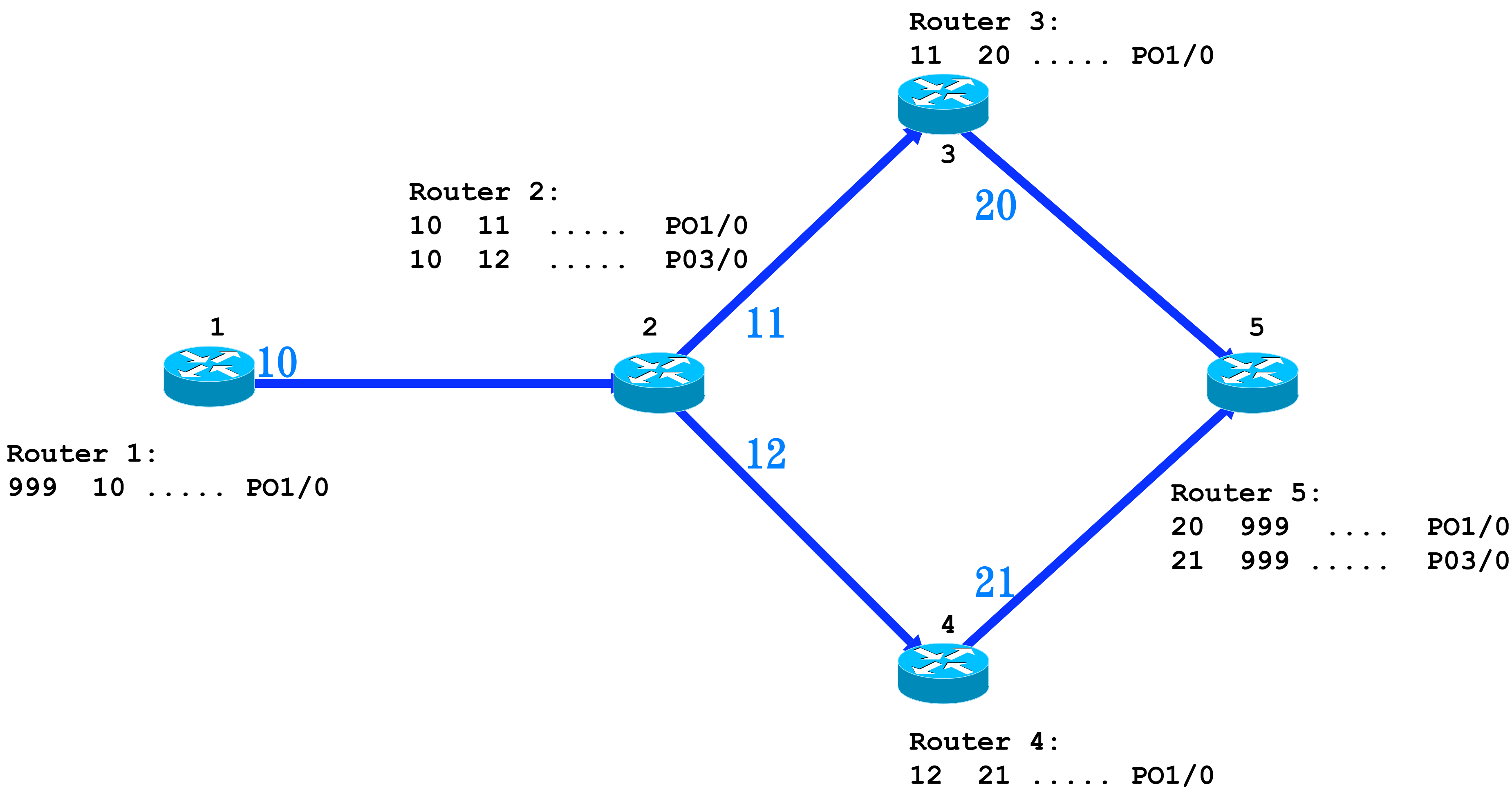
Traffic matrices with LDP statistics

- In a MPLS network, LDP can be used to distribute label information;
- Label-switching can be used without changing the routing scheme (e.g. IGP metrics);
- Many router operating systems provide statistical data about bytes switched in each *forwarding equivalence class* (FEC):



InLabel	OutLabel	Bytes	FEC	OutInt
1234	1235	4124	10.10.10.1/32	PO1/2
…	…	…		…

Traffic matrices with LDP statistics
Use of ECMP load-sharing

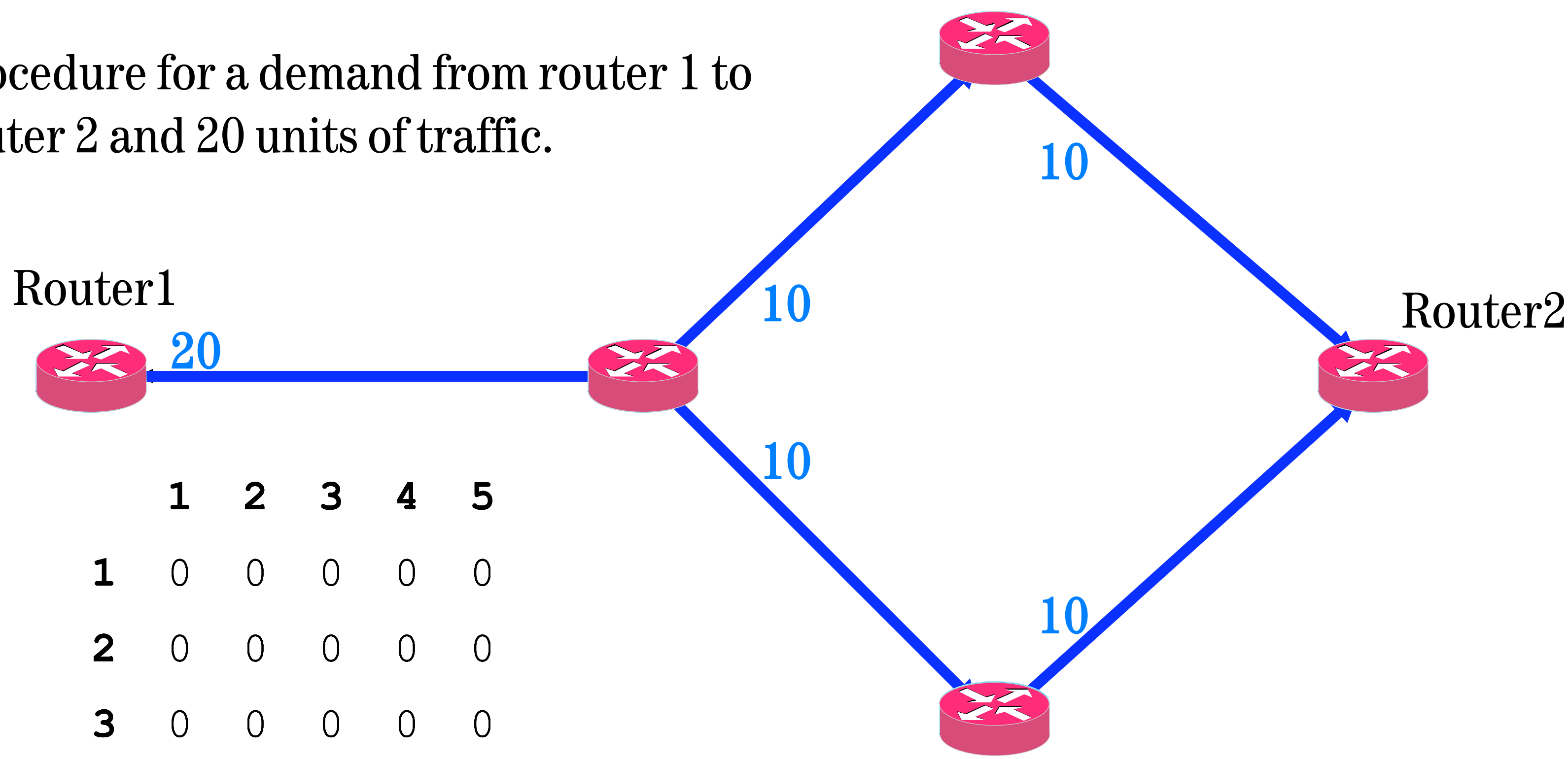


Traffic matrices with LDP statistics

- The given information allows for a forward chaining;
- For each router and FEC a set of residual paths can be calculated (given the topology and LDP information)
- From the LDP statistics we gather the bytes switched on each residual path;
- Problem: It is difficult to decide whether the router under consideration is the beginning or transit for a certain FEC;
- Idea: For the traffic matrix *TM*, add the paths traffic to *TM(A,Z)* and subtract from *TM(B,Z)*. [4]

Traffic matrices with LDP statistics
Example

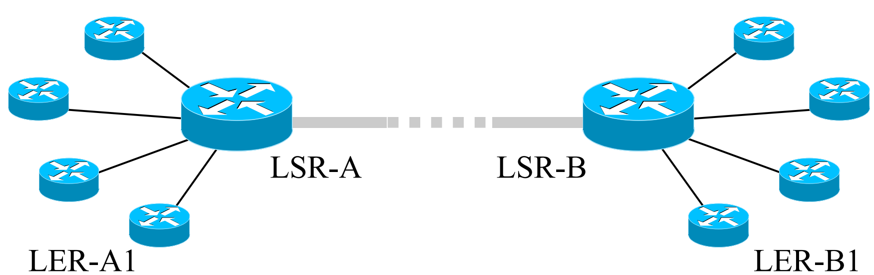
Procedure for a demand from router 1 to router 2 and 20 units of traffic.



	1	2	3	4	5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	20	-2010010100			0

Practical Implementation
Cisco's IOS

- LDP statistical data available through "show mpls forwarding" command;
- Problem: Statistic contains no ingress traffic (only transit);
- If separate routers exist for LER- and LSR- functionality, a traffic matrix on the LSR level can be calculated
- A scaling process can be established to compensate a moderate number of combined LERs/LSRs.



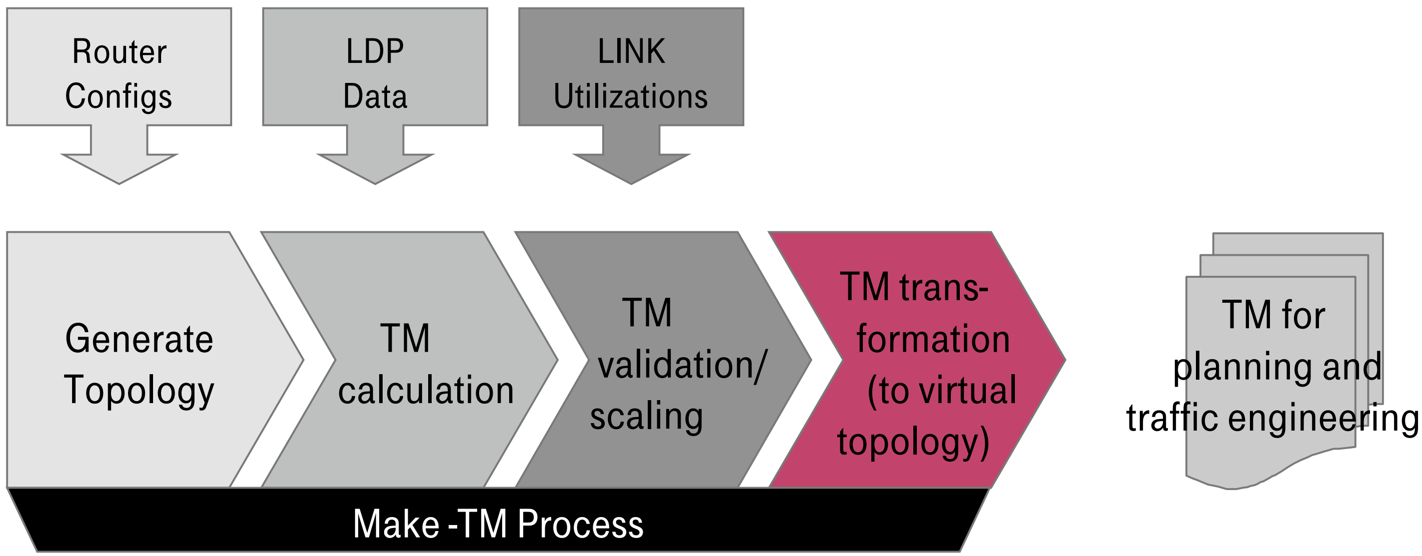
Practical Implementation
Juniper's JunOS

- LDP statistical data available through "show ldp traffic-statistics" command;
- Problem: Statistic is given only per FECs and not per outgoing interface;
- As a result one cannot observe the branching ratios for a FEC that is split due to load-sharing (ECMP);
- Assume that traffic is split equally;
- Especially for backbone networks with highly aggregated traffic this assumption is met quite accurately.

Practical Implementation
Results

- The method has been successfully implemented in Deutsche Telekom’s global MPLS Backbone;
- A continuous calculation of traffic matrices (15min averages) is accomplished in real-time for a network of 180 routers;
- The computation requires only one commodity PC;
- No performance degradation through LDP queries;
- Calculated traffic matrices are used in traffic engineering and network planning.

Practical Implementation
Deployment Process



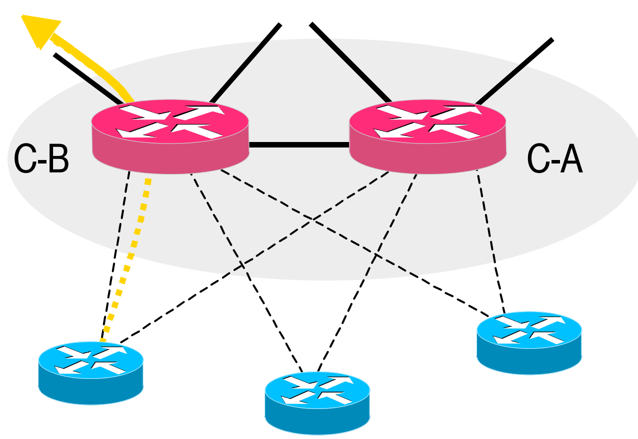
Conclusions for LDP method

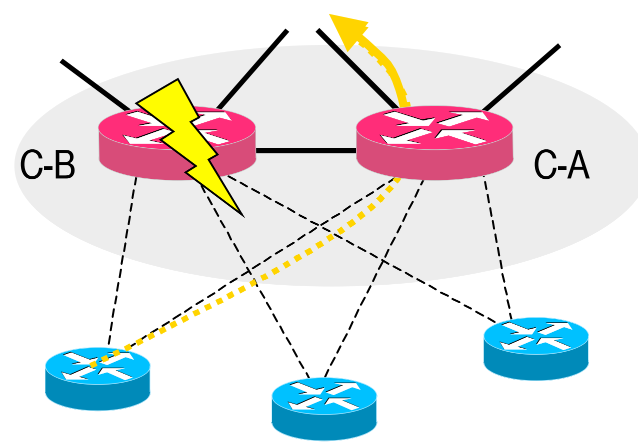
- Our method can be implemented in a multi-vendor network;
- It does not require the definition of explicitly routed LSPs;
- It allows for a continuous calculation;
- There are some restrictions concerning
  - vendor equipment;
  - network topology.
- See Ref. [4]

Traffic Matrices in Partial Topologies

Traffic Matrices in Partial Topologies

- In larger networks, it is often important to have a TM for a partial topology (not based on every router)
- Example: TM for core network (planning and TE)
- Problem: TM changes in failure simulations
- Demand moves to another router since actual demand starts outside the considered topology (red):



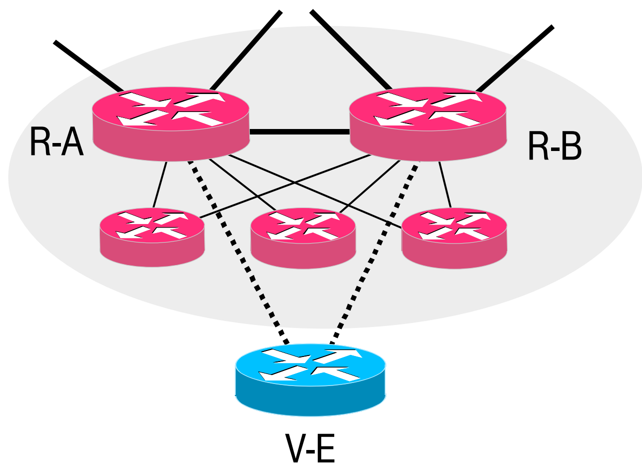


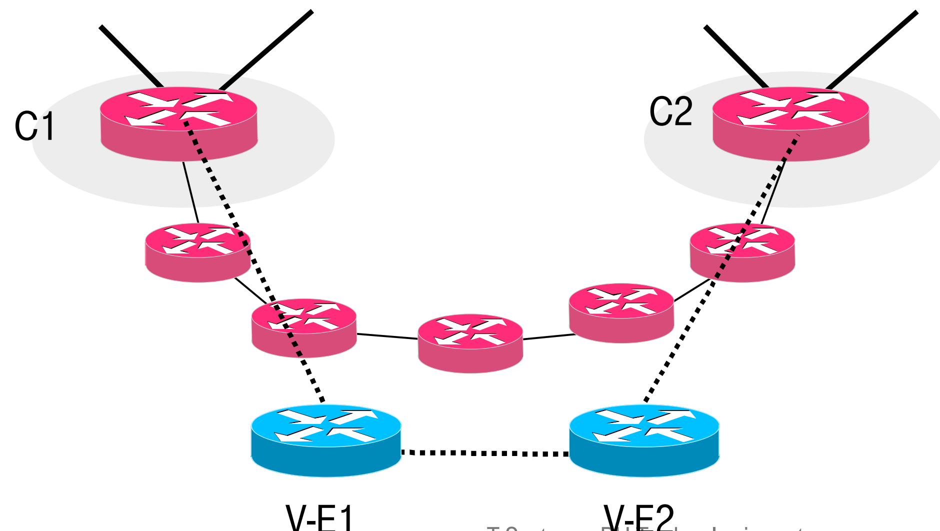
Traffic Matrices in Partial Topologies

- The same problem arises with link failures
- Results in inaccurate failure simulations on the reduced topology
- Metric changes can introduce demand shifts in partial topologies, too.
- But accurate (failure) simulations are essential for planning and traffic engineering tasks

Traffic Matrices in Partial Topologies
Use of Virtual Edge Router in Simulations

- Introduce virtual edge devices as new start-/endpoints for demands
- Map real demands to virtual edge devices;
- Model depends on real topology;
- Tradeoff between simulation accuracy and problem size.





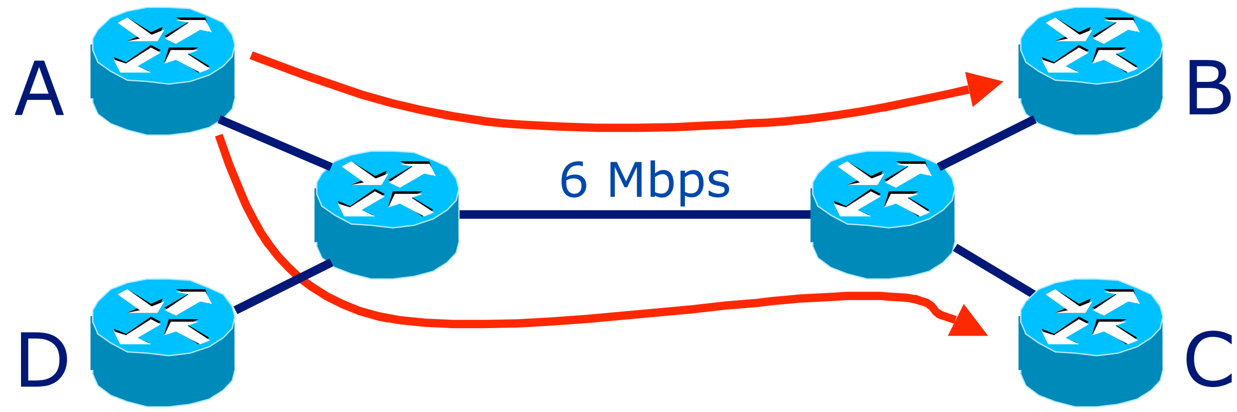
Estimation Techniques

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)*
- Must add additional assumptions (information)

Example

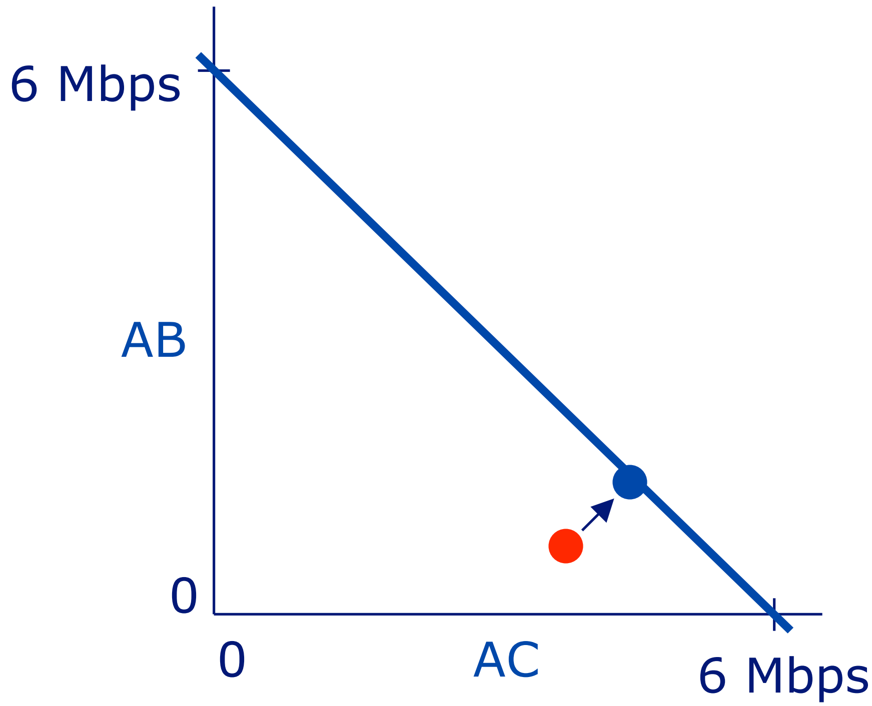


y: link utilizations
A: routing matrix
x: point-to-point demands

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

Example

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



Additional information

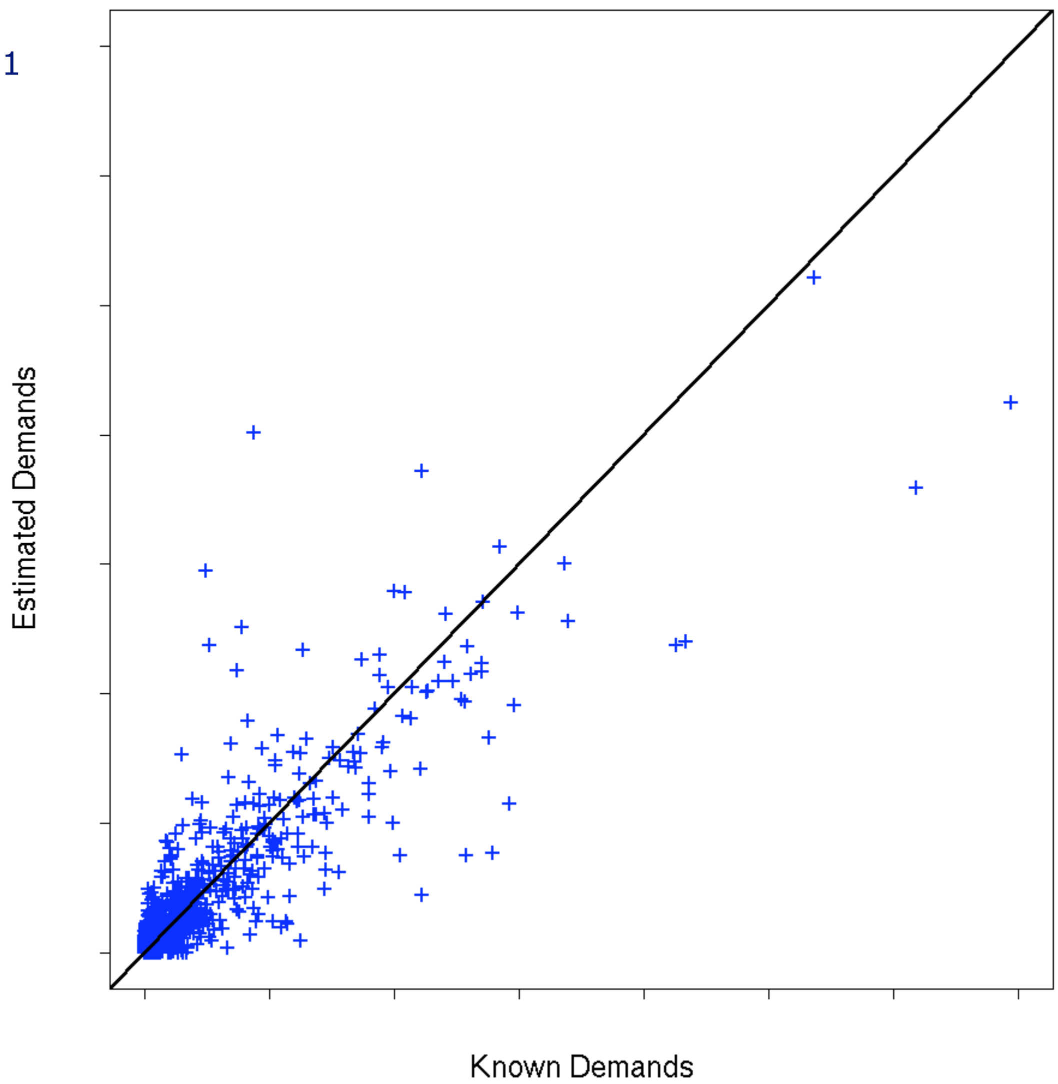
E.g. Gravity Model (every source sends the same percentage as all other sources of it's total traffic to a certain destination)

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

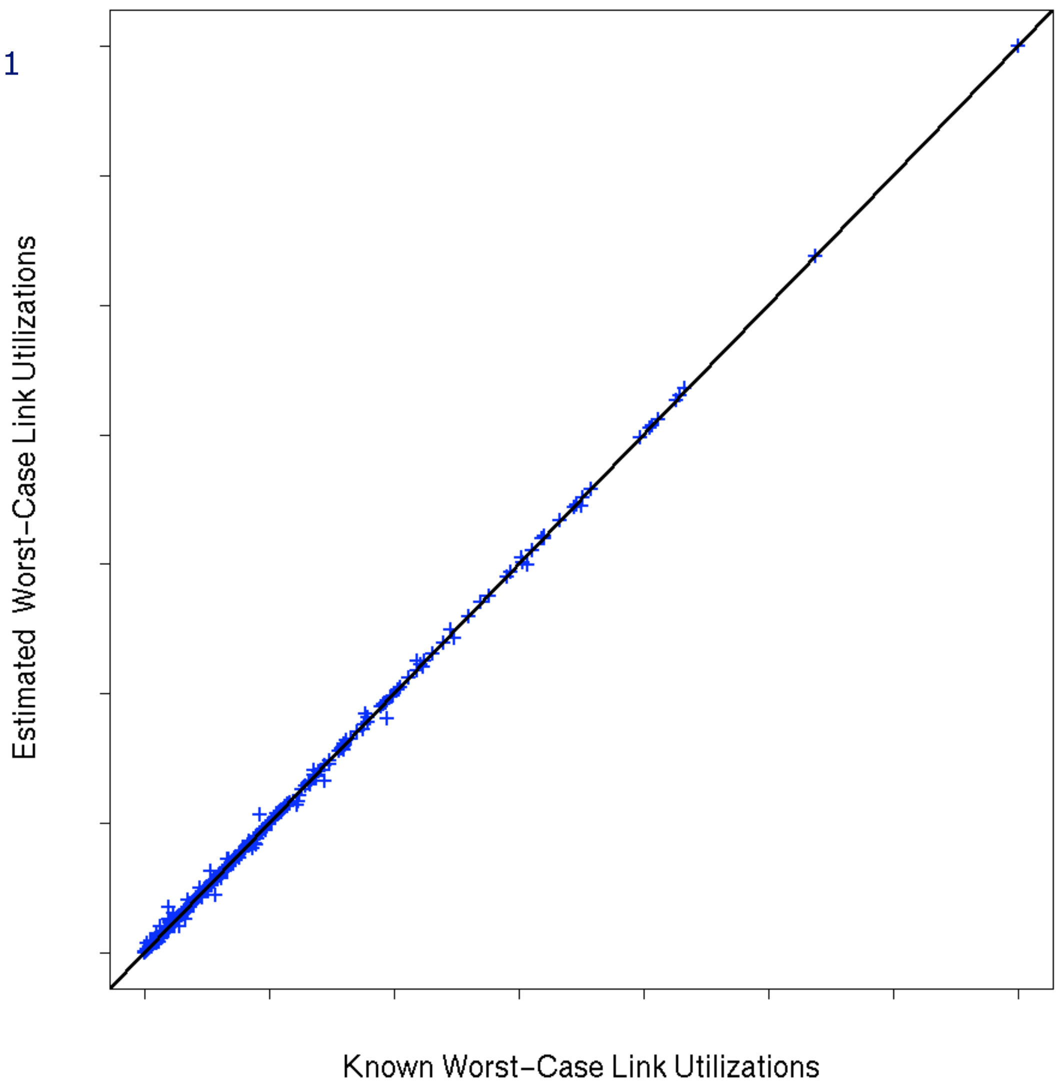
Real Network: Estimated Demands

International Tier-1
IP Backbone



Estimated Link Utilizations!

International Tier-1
IP Backbone



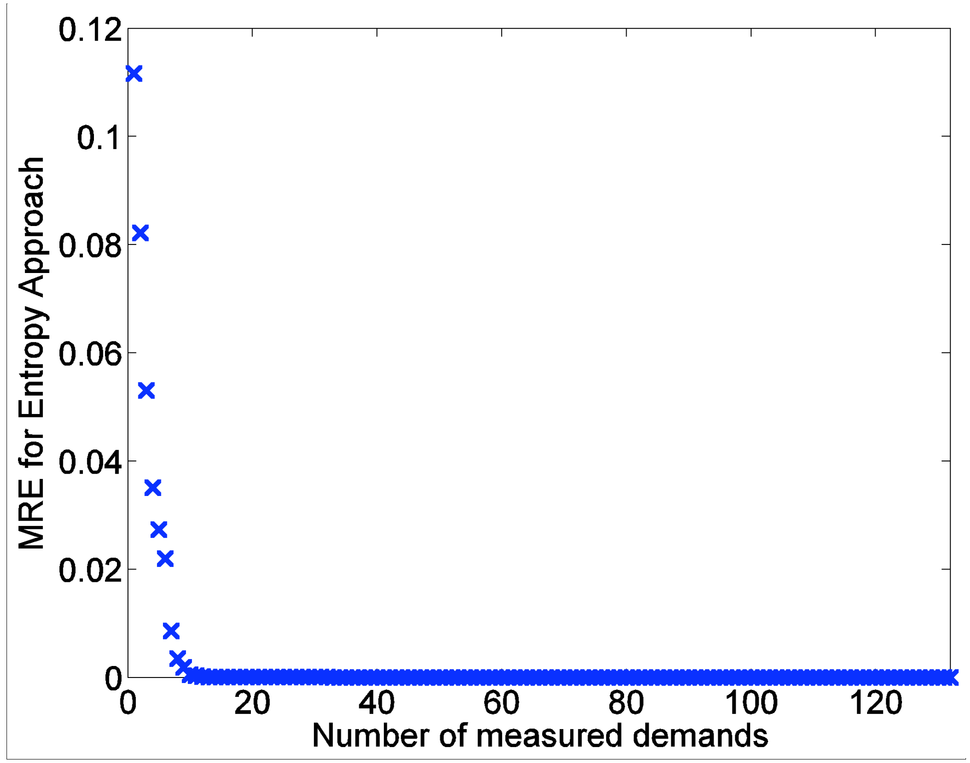
Demand Estimation Results

- Individual demands
  - Inaccurate estimates…
- 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

Estimation with Measurements

- Estimation techniques can be used in combination with demand meadsurements
  - E.g. NetFlow or partial MPLS mesh

- This example: Greedy search to find demands which decreases MRE (Mean Relative Error) most.
  - A small number of measured demands account for a large drop in MRE



Data from [1]

Estimation Summary

- Algorithms have been published
  - Commercial tools are available
  - Implement yourself?
- Can be used in multiple scenarios:
  - Fully estimate Traffic Matrix
  - Estimate Peering traffic when Core Traffic Matrix is know
  - Estimate unknown demands in a network with partial MPLS mesh (LDP or RSVP)
  - Combine with NetFlow
    - Measure large demands, estimate small ones
- Also see AT&T work
  - E.g. Nanog29: *How to Compute Accurate Traffic Matrices for Your Network in Seconds* [2]

Traffic Matrix Estimation Case-Study

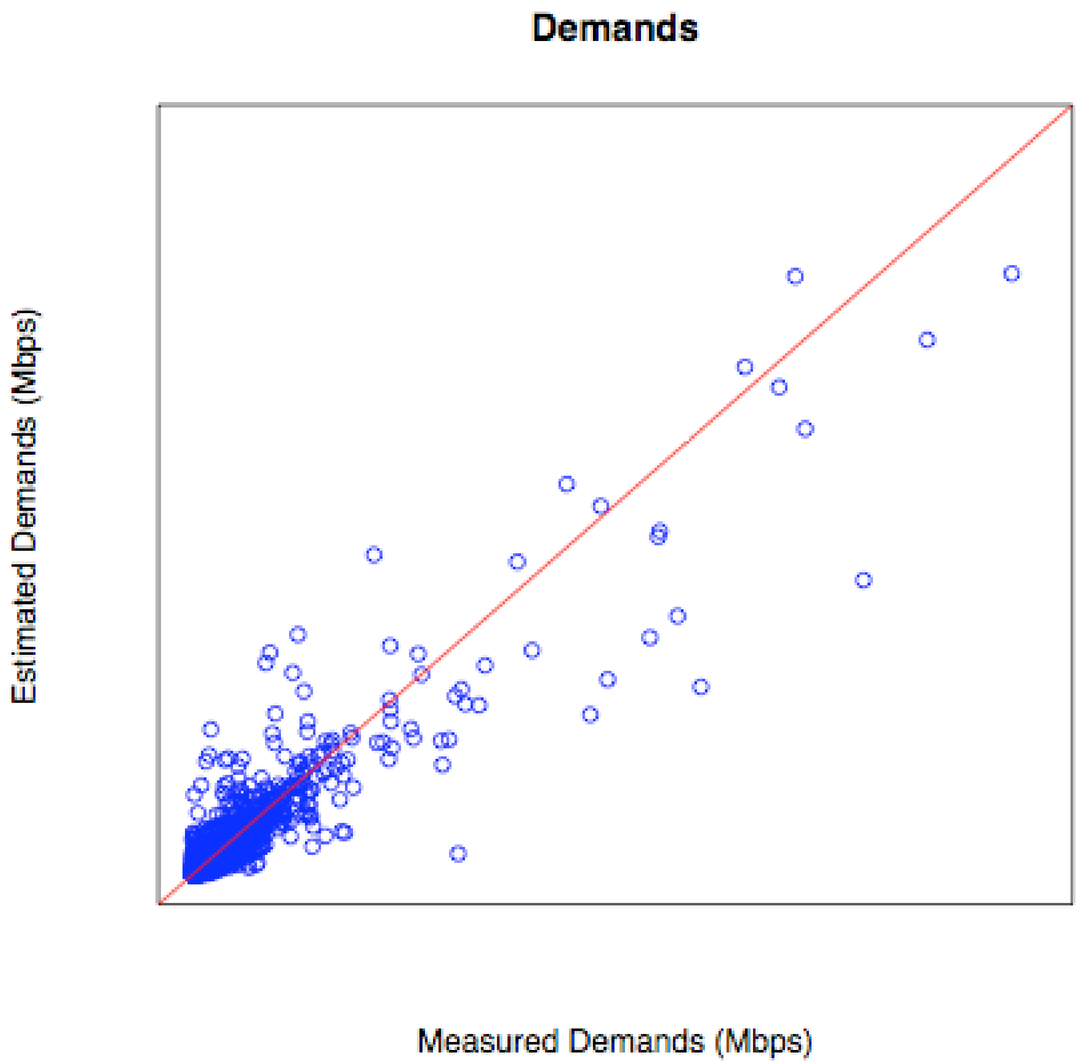
TM Estimation Case-Study

- Large ISP network
  - 77 Routers
  - 166 Circuits
- Known Traffic Matrix
  - Direct MPLS measurement
- Case-study will evaluate:
  - How does estimated TM compare to known TM?
  - How well do tools that require a TM work when given the estimated TM?
- TM estimation using Cariden MATE Software
  - Demand Deduction tool

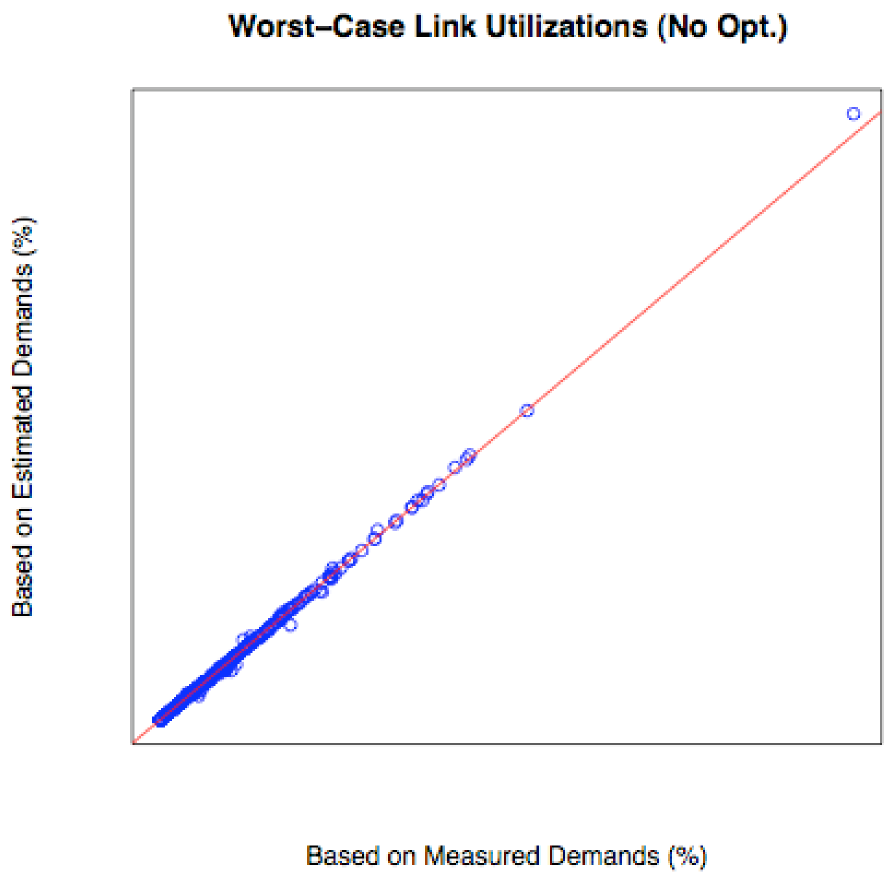
Procedure

- Start with current network and known TM
  - save as "PlanA" (with TM "Known")
- IGP Simulation for non-failure
- Save Link Utilizations and Node In/Out traffic
- Estimate Traffic Matrix
  - New TM: "Estimated"
  - Save as "PlanB"
- Do an IGP *Metric Optimization* on both networks
  - Using known TM in planA
  - Using estimated TM in PlanB
- Simulate IGP routing on both optimized networks
  - using *known* Traffic matrix for both
- Compare Results!

Estimated Demands

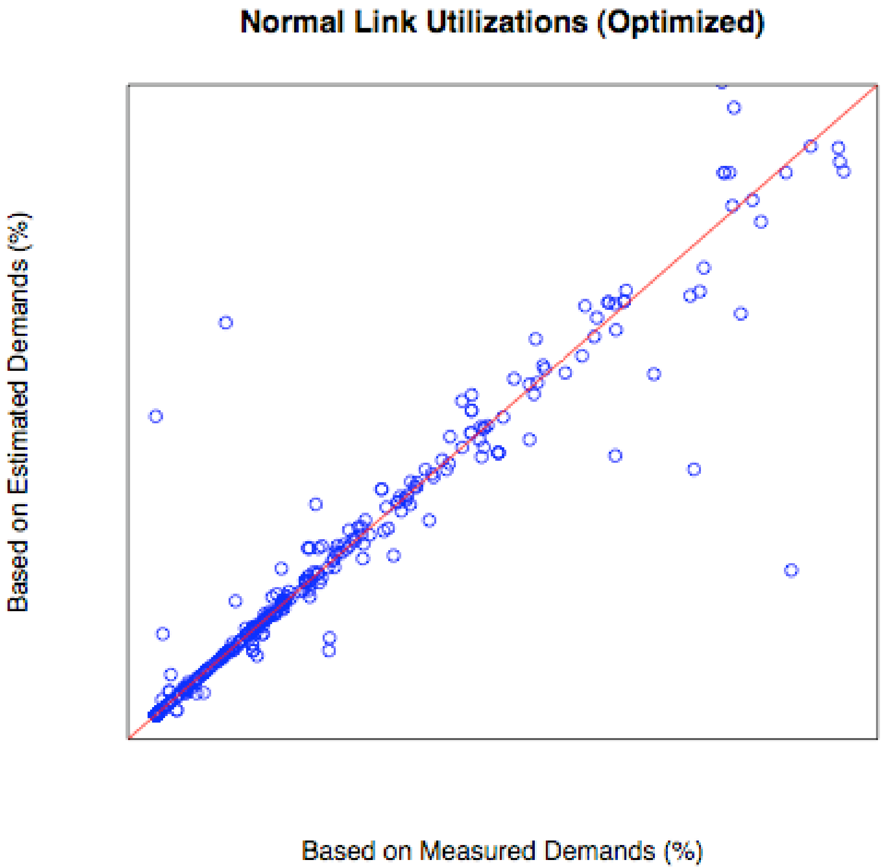


Worst-Case Link Util. (No. Opt)



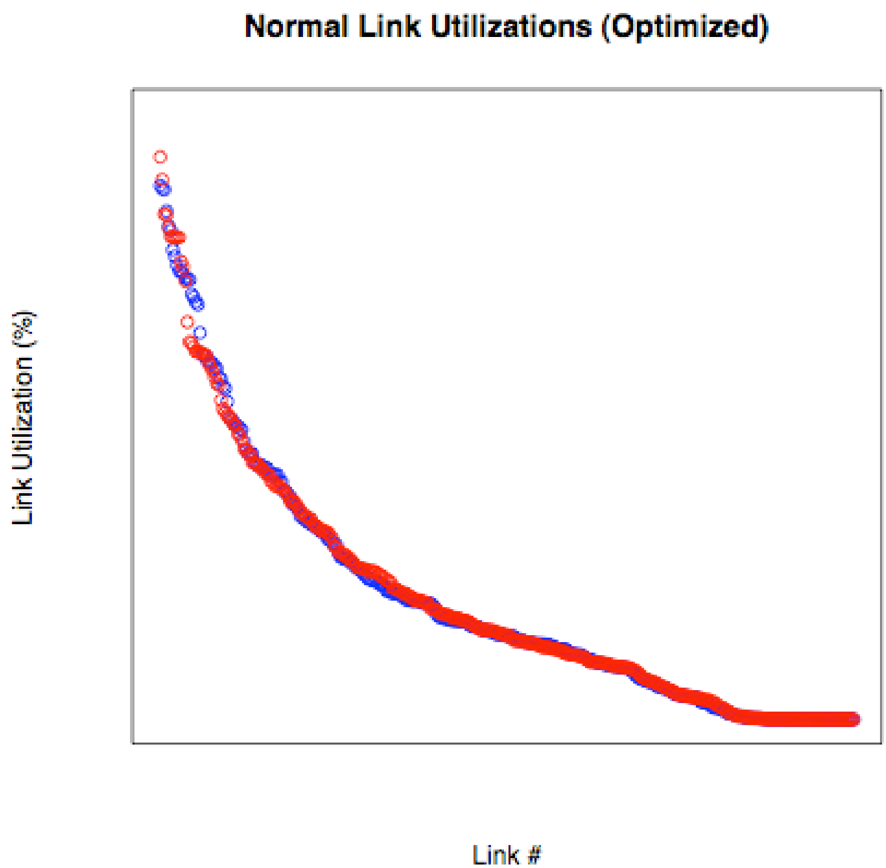
- No Metric Optimization
- PlanA Traffic Matrix:
  - Known
- PlanB Traffic Matrix:
  - Estimated
- IGP Simulation
  - Circuit + SRLG failures
- Compare Worst-Case Link Utilizations (in %)

Normal Link Utilizations (Opt.)



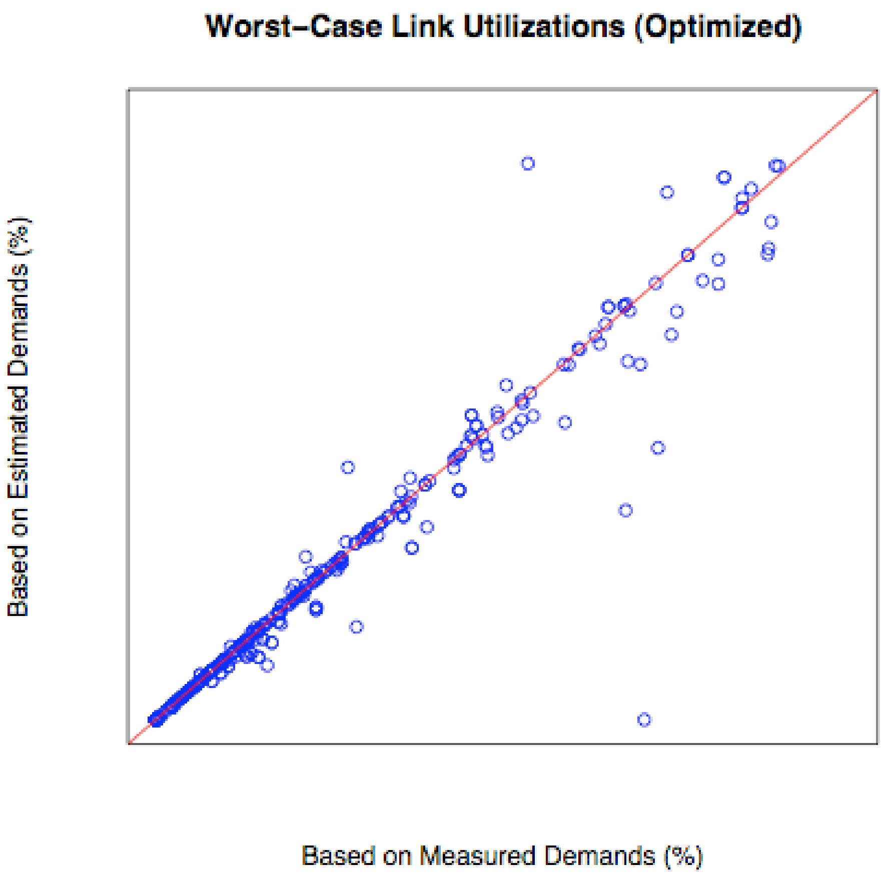
- IGP Metric Optimization
  - PlanA Traffic Matrix:
    - Known
  - PlanB bandwidth level:
    - Estimated
- IGP Simulation
  - PlanA Traffic Matrix:
    - Known
  - PlanB bandwidth level:
    - Original
- Compare Base Link Utilizations (in %)
  - non-failure

Normal Link Utilizations (Opt.)



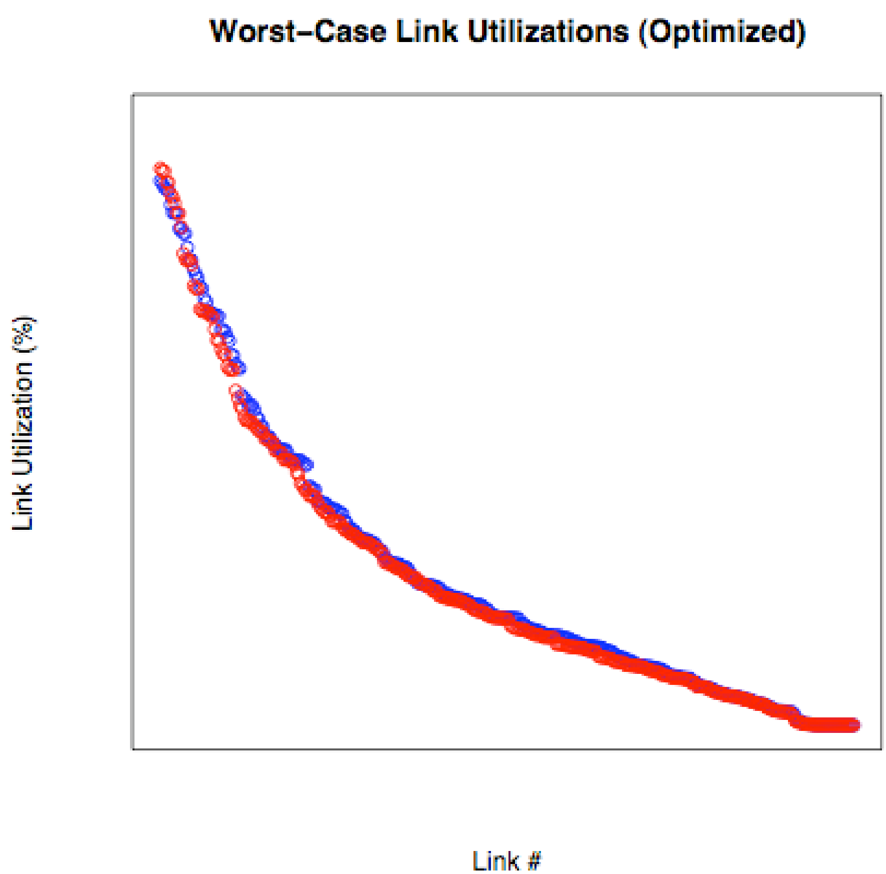
- Scenario: same as previous slide
- Compare *Sorted* Link Utilizations
  - non-failure
- Colors:
  - based on measured demands: BLUE
  - based on estimated demands: RED

Worst-Case Link Utilizations (Opt)



- Scenario: same
- Compare Worst-Case Link Utilizations (in %)
  - Circuits + SRLG failures

Worst-Case Link Utilizations (Opt)



- Scenario: same
- Compare *Sorted* Worst-Case Link Utilizations (in %)
  - Circuits + SRLG failures
- Colors:
  - based on measured demands: BLUE
  - based on estimated demands: RED

TM Estimation Case-Study

- Works very well on this ISP topology/traffic!
  - Also on AT&T, and all other networks we tried
- Even more accurate if used in combination with demand measurements
  - E.g. from NetFlow, DCU or MPLS

Summary & Conclusions

Overview

- “Traditional” NetFlow (Version 5)
  - Requires a lot of resources for collection and processing
  - Not trivial to convert to Traffic Matrix
- BGP NextHop Aggregation NetFlow provides almost direct measurement of the Traffic Matrix
  - Verion 9 export format
  - Only supported by Cisco in newer IOS versions
- Juniper DCU is too limited (only 16 classes) to build a full Traffic Matrix
  - But could be used as adjunct to TM Estimation

Overview

- MPLS networks provide easy access to the Traffic Matrix
  - Directly measure in RSVP TE networks
  - Derive from switching counters in LDP network
- Very convenient if you already have an MPLS network, but no reason to deploy MPLS just for the TM

- Estimation techniques can provide reliable Traffic Matrix data
  - Very useful in combination with partially know Traffic Matrix (e.g. NetFlow, DCU or MPLS)

Contact

Thomas Telkamp

Cariden Technologies, Inc.

telkamp@cariden.com

Stefan Schnitter

T-Systems

Stefan.Schnitter@t-systems.com

References

1. A. Gunnar, M. Johansson, and T. Telkamp, “Traffic Matrix Estimation on a Large IP Backbone - A Comparison on Real Data”, *Internet Measurement Conference 2004.* Taormina, Italy, October 2004.
2. 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.
3. AT&T Tomogravity page:
   http://www.research.att.com/projects/tomo-gravity/
4. S. Schnitter, T-Systems; M. Horneffer, T-Com. “Traffic Matrices for MPLS Networks with LDP Traffic Statistics.” Proc. Networks 2004, VDE-Verlag 2004.
5. Y. Vardi. “Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data.” J.of the American Statistical Association, pages 365–377, 1996.