Best Practices for Determining the Traffic Matrix in IP Networks

Apricot 2005 - Kyoto, Japan

Thursday February 24, 2005
Internet Routing and Backbone Operations
Session C5-4

Thomas Telkamp, *Cariden Technologies, Inc.*

Contributors

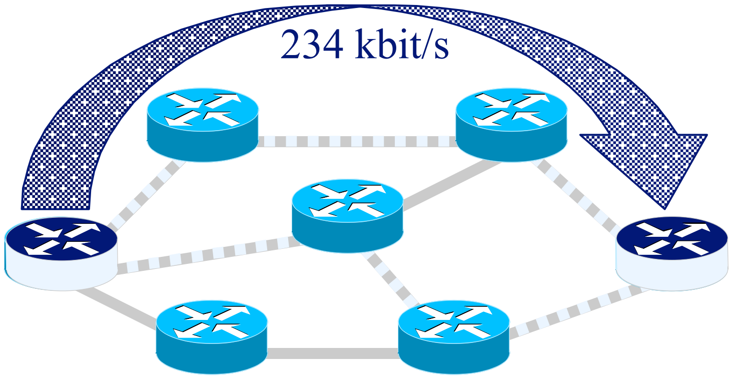
- Stefan Schnitter, *T-Systems*
  - LDP Statistics
- Benoit Claise, *Cisco Systems, Inc.*
  - Cisco NetFlow
- Mikael Johansson, *KTH*
  - Traffic Matrix Properties

Agenda

- Introduction
  - Traffic Matrix Properties
- Measurement in IP networks
  - NetFlow
  - DCU/BGP Policy Accounting
- MPLS Networks
  - RSVP based TE
  - LDP
    - Data Collection
    - LDP deployment in Deutsche Telekom
- Estimation Techniques
  - Theory
  - Example Data
- 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

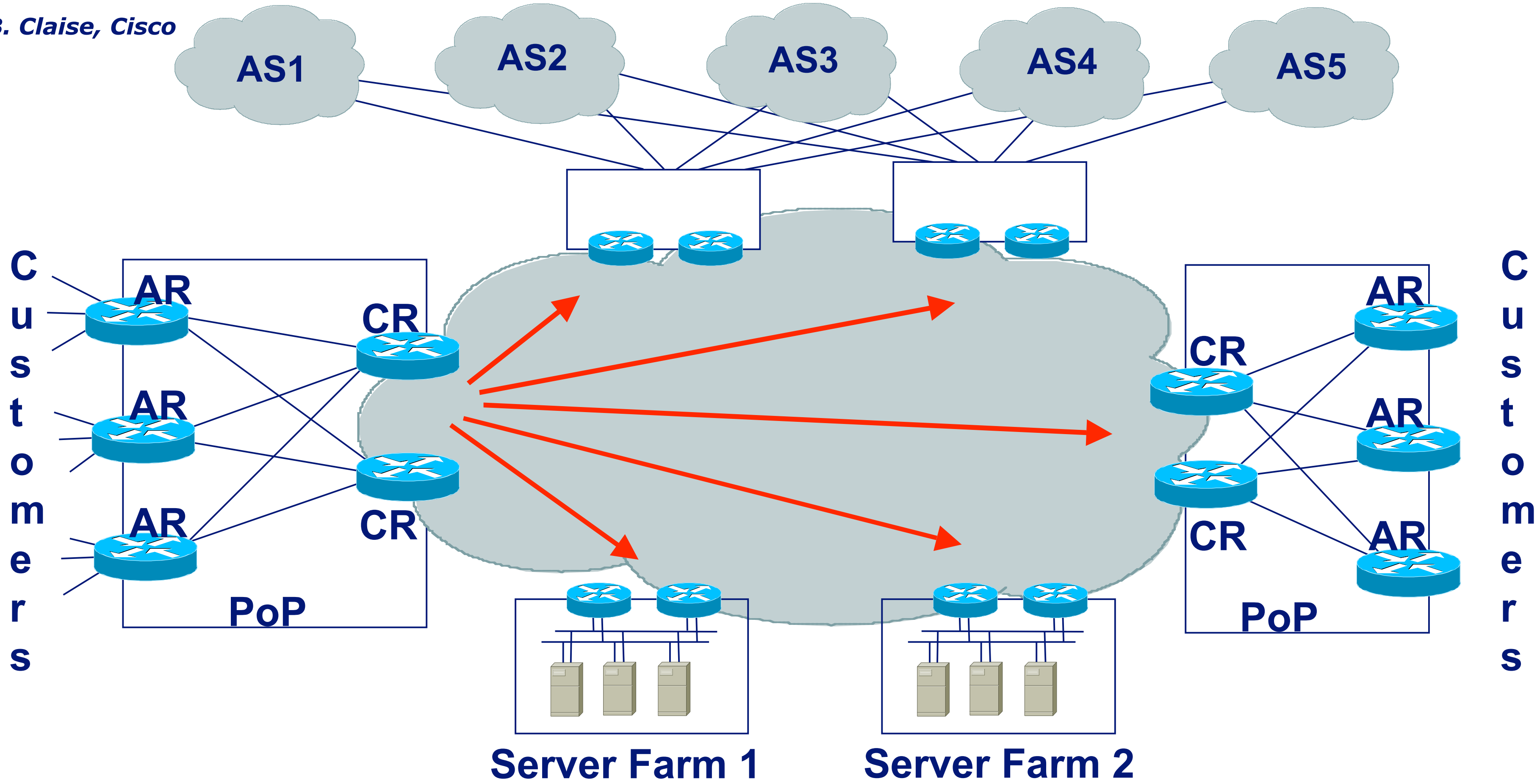


Determining the Traffic Matrix

- 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 Traffic Engineering

Internal Traffic Matrix

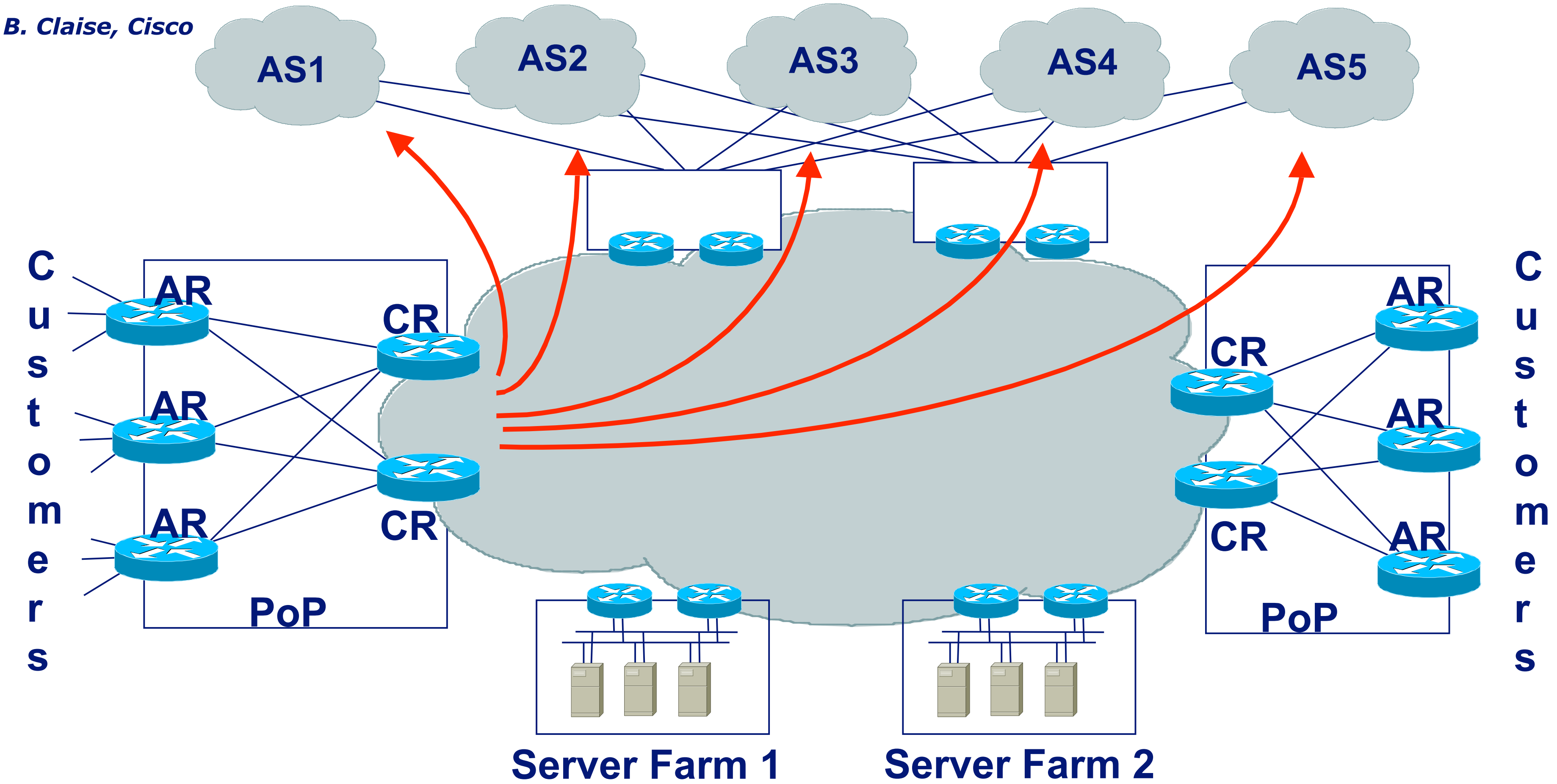
B. Claise, Cisco



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

External Traffic Matrix

B. Claise, Cisco



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

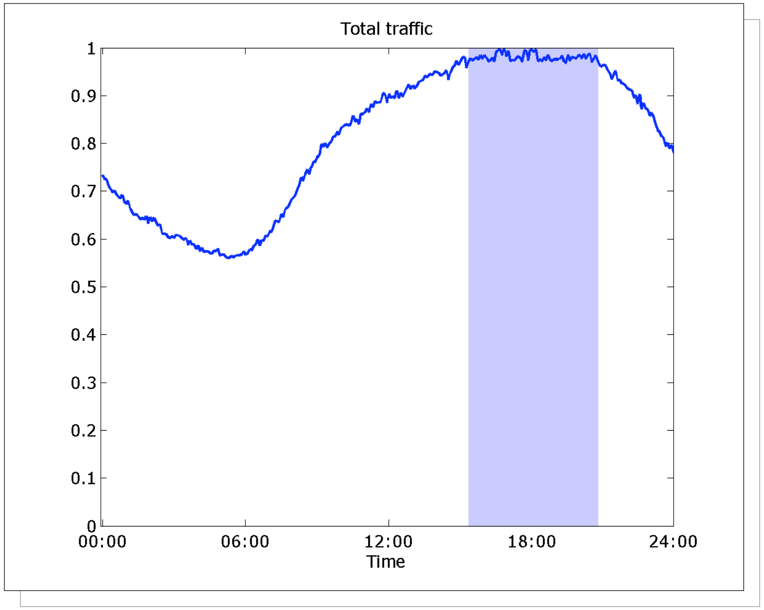
The external traffic matrix can influence the internal one

Traffic Matrix Properties

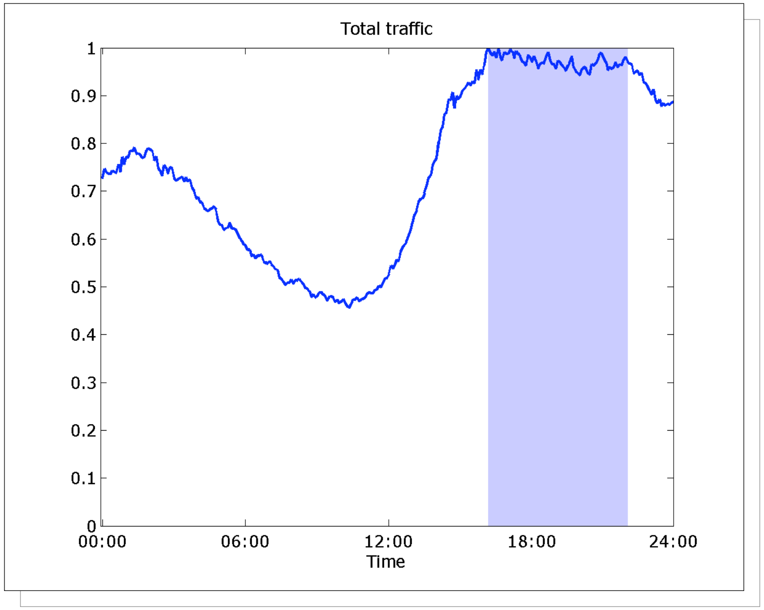
- Example Data from Tier-1 IP Backbone
  - Measured Traffic Matrix (MPLS TE based)
  - European and American subnetworks
  - 24h data
  - See [1]
- Properties
  - Temporal Distribution
    - How does the traffic vary over time
  - Spatial Distribution
    - How is traffic distributed in the network?

Total traffic and busy periods

European subnetwork



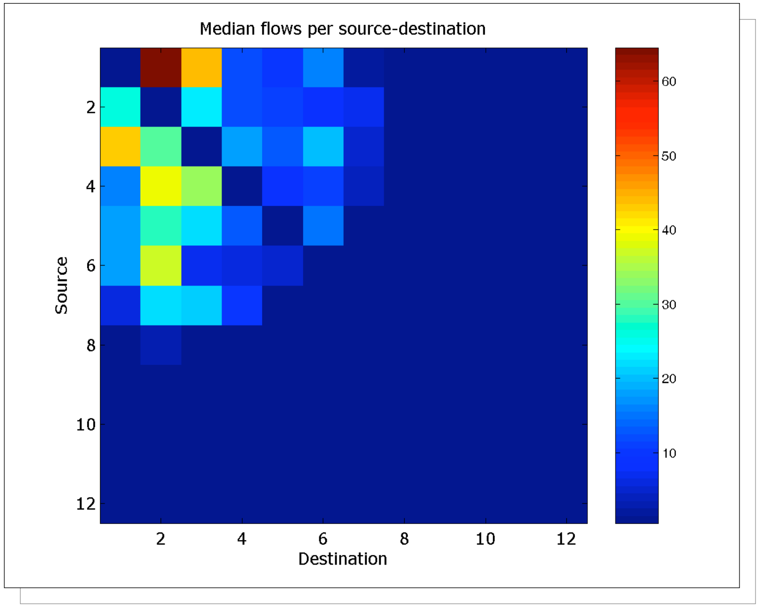
American subnetwork



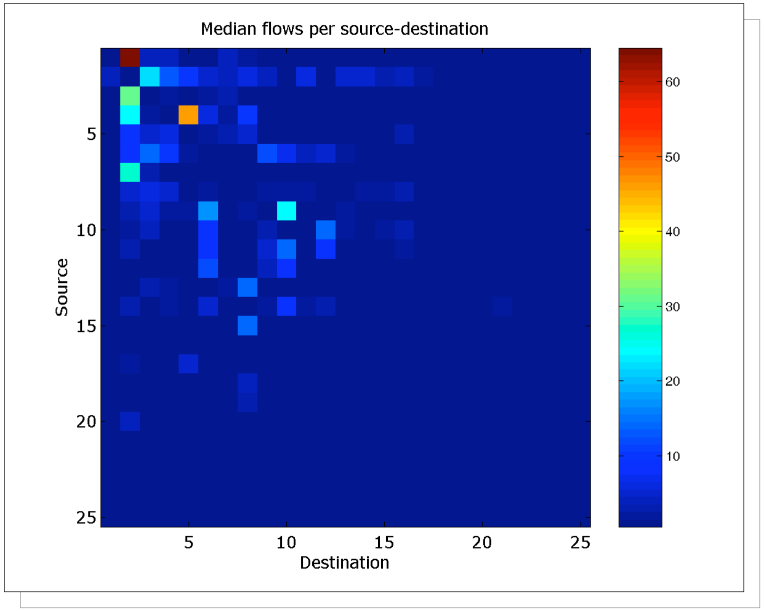
Total traffic very stable over 3-hour busy period

Spatial demand distributions

European subnetwork



American subnetwork



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

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/BGP Policy Accounting
  - Routers collect aggregated destination statistics
- MPLS
  - RSVP
    - Measurement of Tunnel/LSP counters
  - LDP
    - Measurement of LDP counters
- Estimation
  - Estimate Traffic Matrix based on Link Utilizations

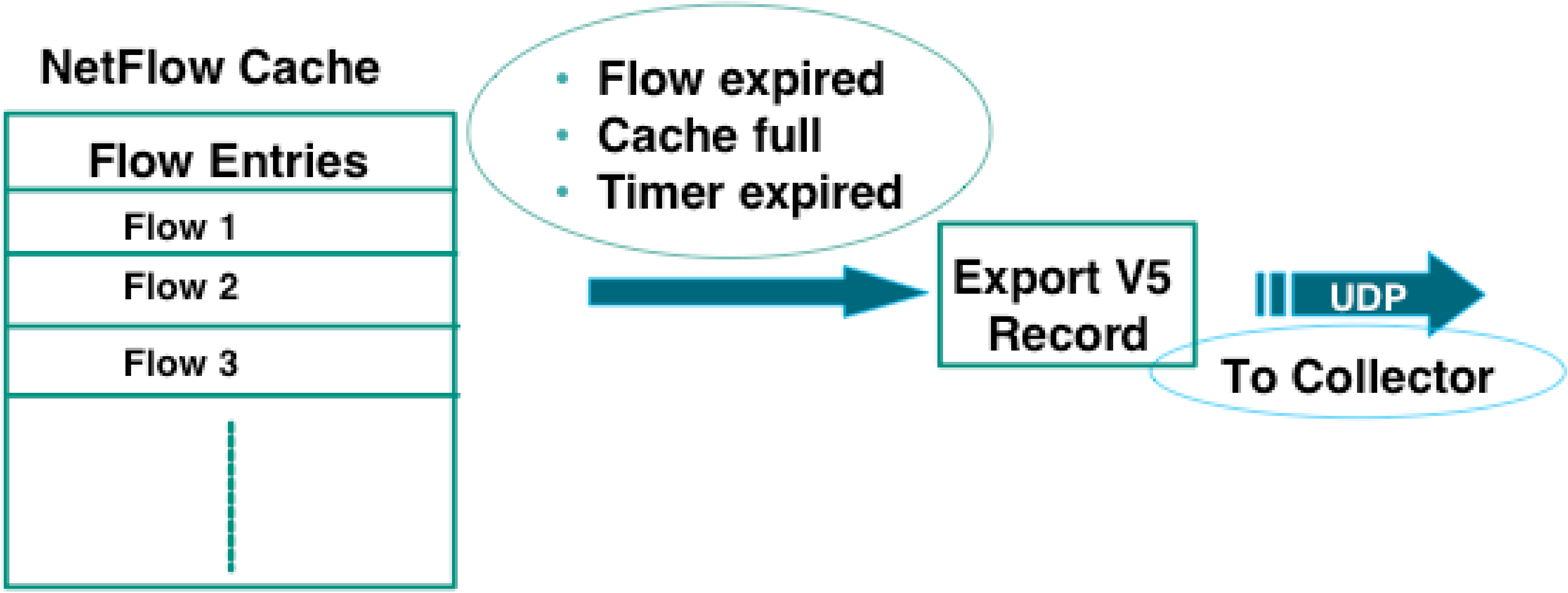
NetFlow: Versions

- 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

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

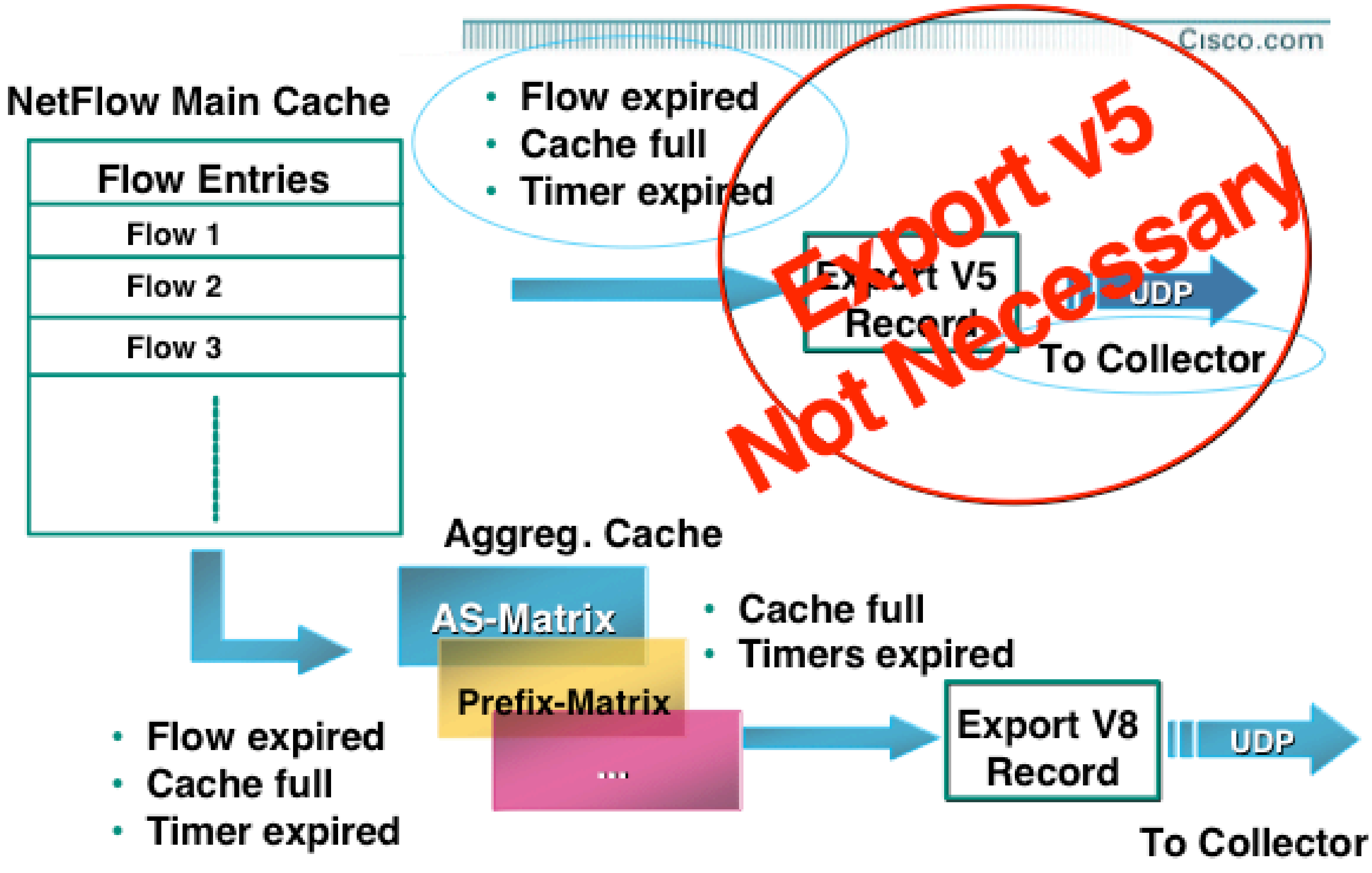
NetFlow Version 8

- 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 Aggregation (Version 9)

- 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

NetFlow Summary

- Building a Traffic Matrix from NetFlow data is not trivial
  - Need to correlate Source/Destination information with routers or PoPs
  - Commercial Products
- 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

NetFlow Summary

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

DCU/BGP Policy Accounting

- DCU: Destination Class Usage
  - Juniper
- BGP Policy Accounting
  - Cisco
- Accounting traffic according to the route it traverses
  - For example based on BGP communities
- Supports up to 16 (DCU) or 64 (BGP PA) 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

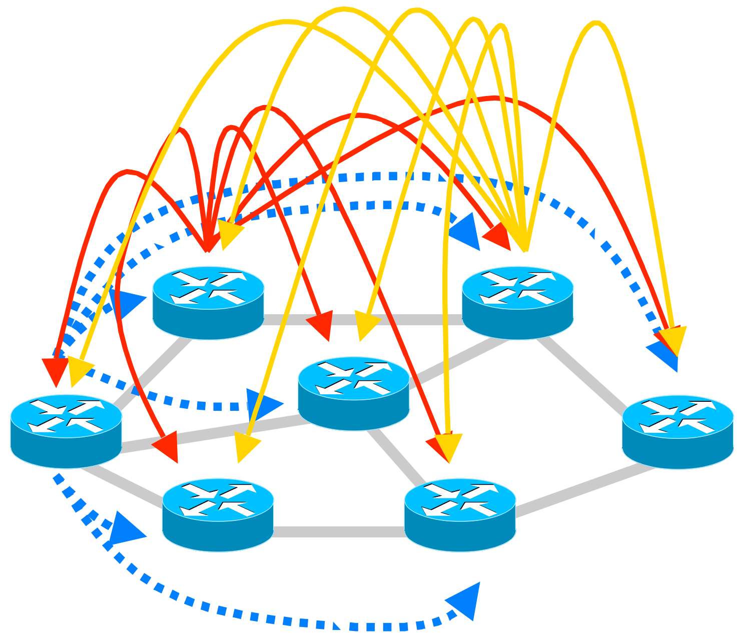
MPLS Based Methods

- Two methods to determine traffic matrices:
  - Using RSVP-TE tunnels
  - Using LDP statistics
    - As described in [4]

- Some comments on Deutsche Telekom's practical implementation

RSVP-TE Based Method

- 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;

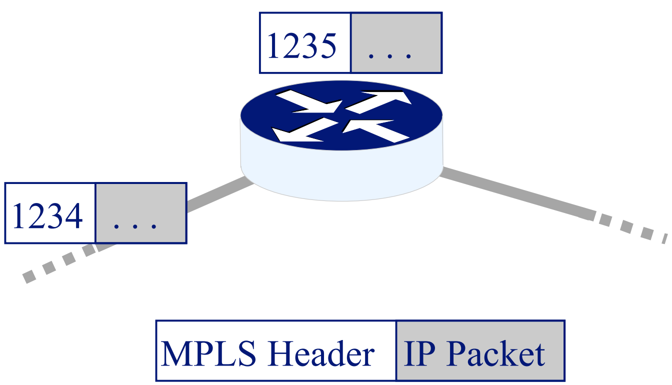


RSVP-TE: Pro’s and Con’s

- 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
…	…	…		…

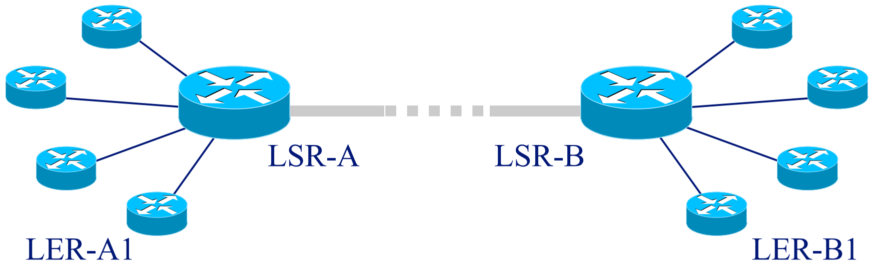
Traffic matrices with LDP statistics

- 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)* . (see [4])

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.

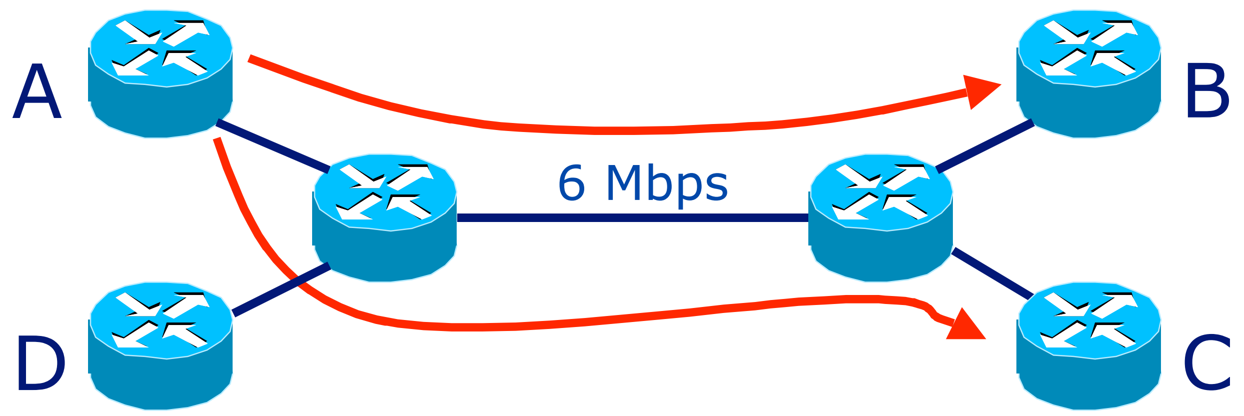
Conclusions for LDP method

- The LDP method can be implemented in a multi-vendor network
    - E.g. Deutsche Telekom's global MPLS Backbone
        - continuous calculation of traffic matrices (15min averages)
        - Commodity PC
- It does not require the definition of explicitly routed LSPs
- See ref. [4]

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

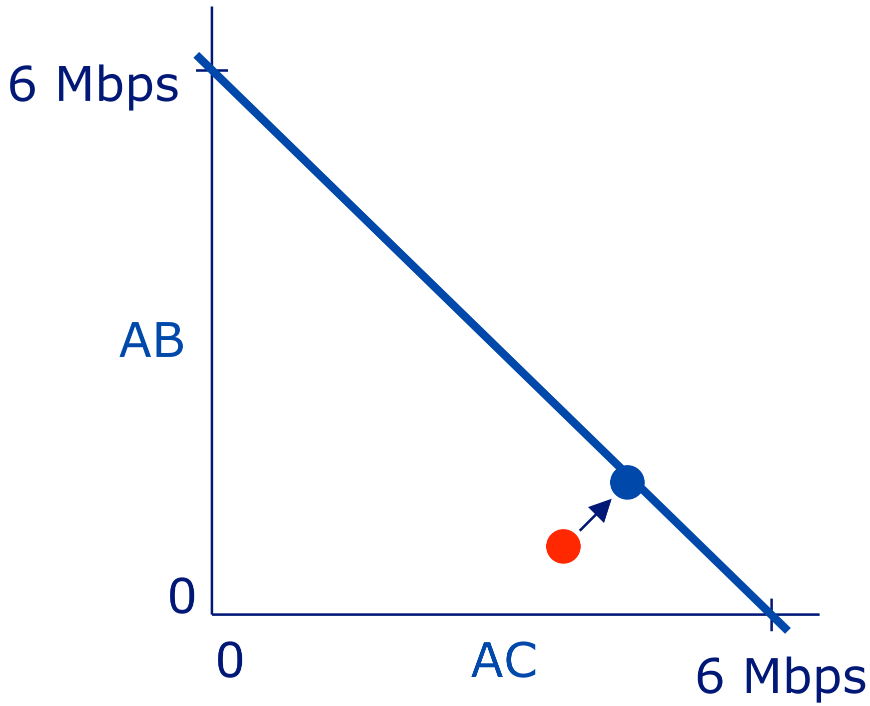


y: link utilizations
A: routing matrix
x: point-to-point demands

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

Example

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



Additional information

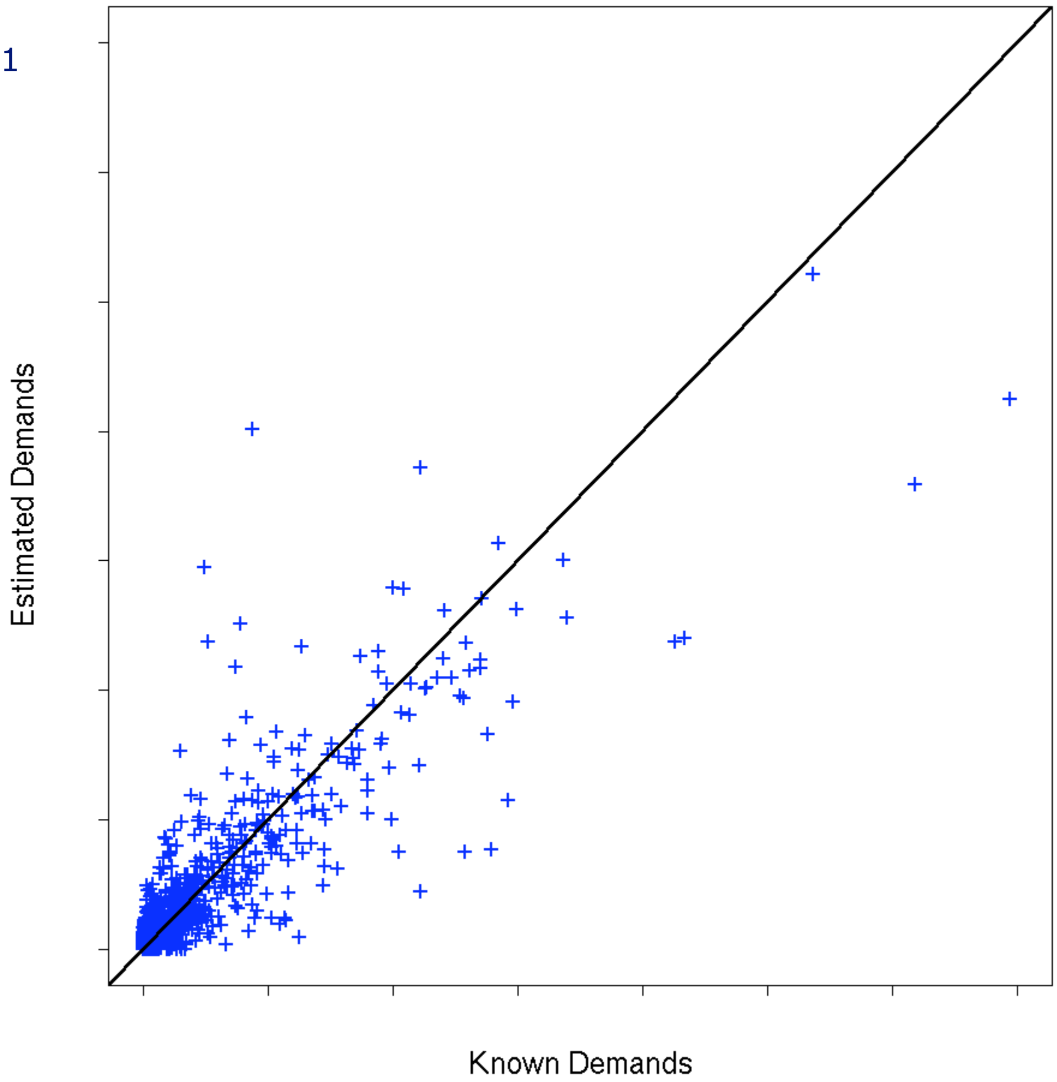
E.g. Gravity Model (every source sends the same percentage as all other sources of it's total traffic to a certain destination)

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

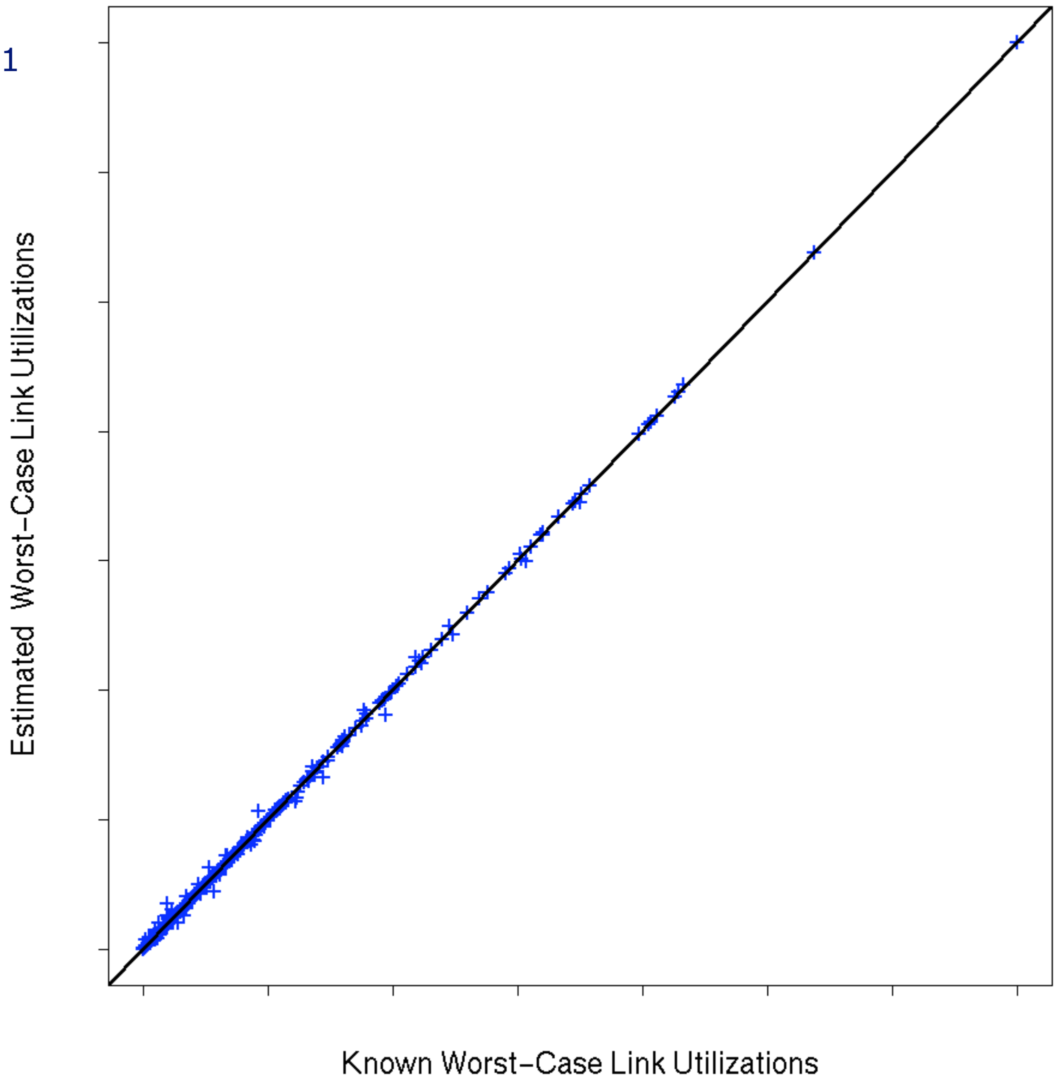
Real Network: Estimated Demands

International Tier-1
IP Backbone



Estimated Link Utilizations!

International Tier-1
IP Backbone



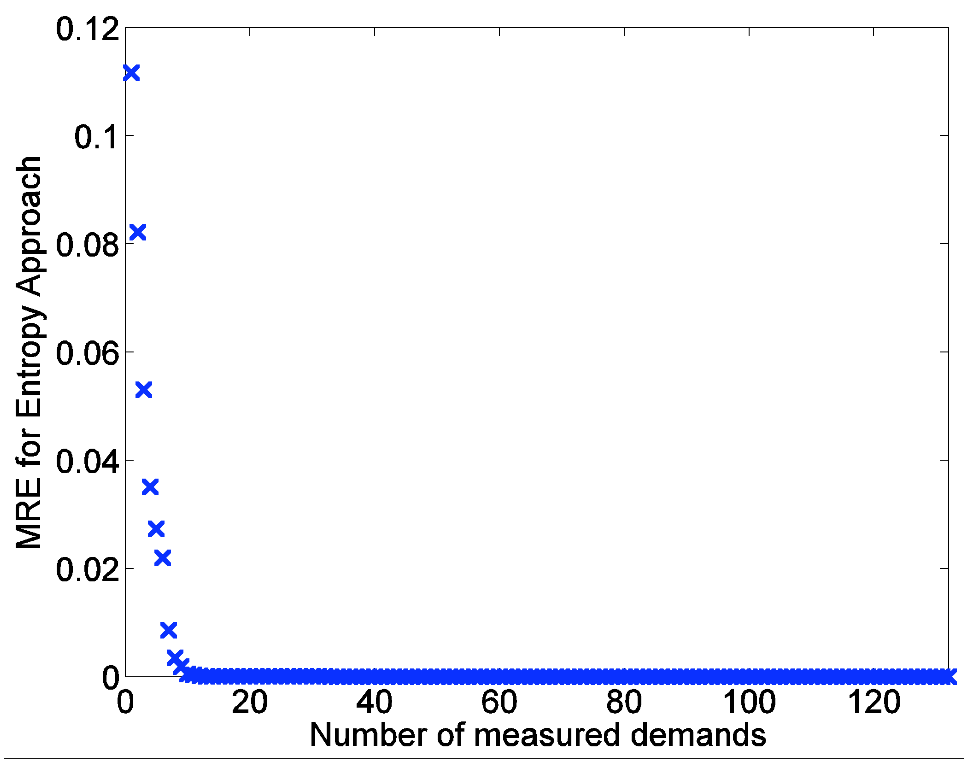
Demand Estimation Results

- 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/DCU/BGP Policy Accounting
    - 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]

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
  - BGP only
  - Only supported by Cisco in newer IOS versions
- DCU/BGP Policy Accounting 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 justification 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

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.