



cariden

*the economics of network control*

## **Traffic Matrices for IP Networks: NetFlow, MPLS, Estimation, Regression**

*Preparing for the Future of the Internet*  
Network Information Center, México  
November 29, 2007

*Arman Maghbouleh, Cariden Technologies, Inc.*

**[www.cariden.com](http://www.cariden.com)**

(c) cariden technologies, inc., portions cisco systems



# IP Traffic Matrix Practices

2001

2003

2007

## Direct Measurement

NetFlow, RSVP, LDP, Layer 2, ...

Good when it works (half the time), but\*

\*Measurement issues

High Overhead (e.g.,  $O(N^2)$  LSP measurements, NetFlow CPU usage)

End-to-end stats not sufficient:

Missing data (e.g., LDP ingress counters not implemented)

Unreliable data (e.g., RSVP counter resets, NetFlow cache overflow)

Unavailable data (e.g., LSPs not cover traffic to BGP peers)

Inconsistent data (e.g., timescale differences with link stats)

## Estimation

Pick one of many solutions that fit link stats

(e.g., Tomogravity)

TM not accurate but good enough for planning

## Regressed Measurement

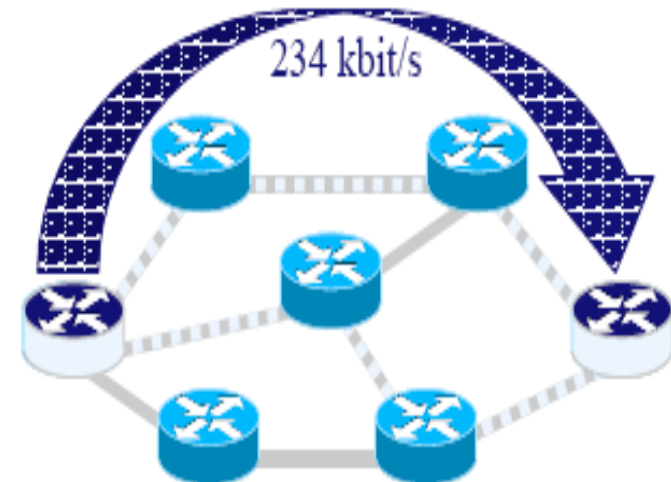
Use link stats as gold standard (reliable, available)

Regression Framework adjusts (corrects/fills in) available

NetFlow, MPLS, measurements to match link stats

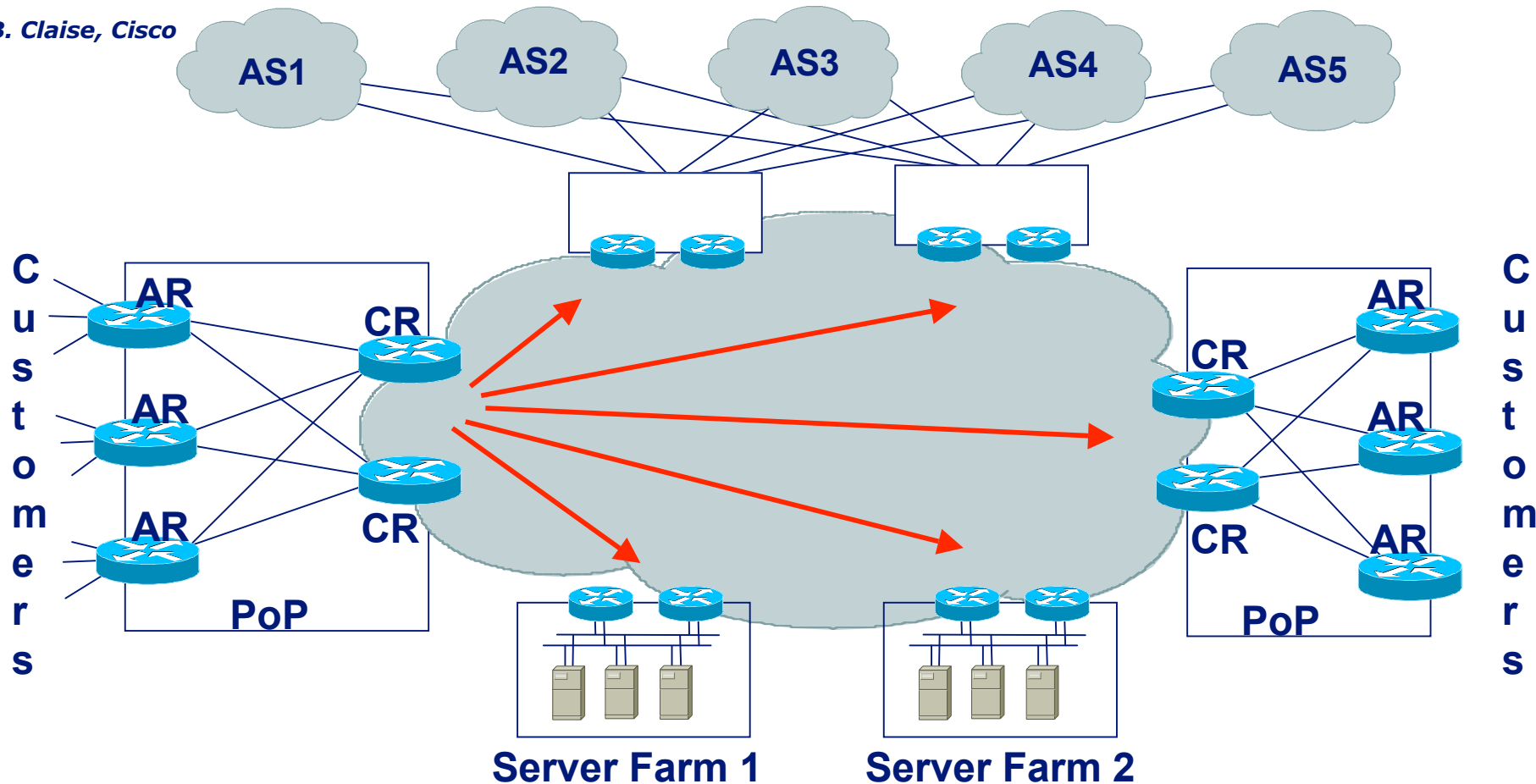
# Traffic Matrix

- Amount of data transmitted between each pair of network nodes
  - Internal vs. External
  - Per Class, per application, ...
- Crucial for
  - Resiliency planning
  - “what-if” scenarios
  - MPLS TE tunnel placement
  - IP TE



# Internal Traffic Matrix

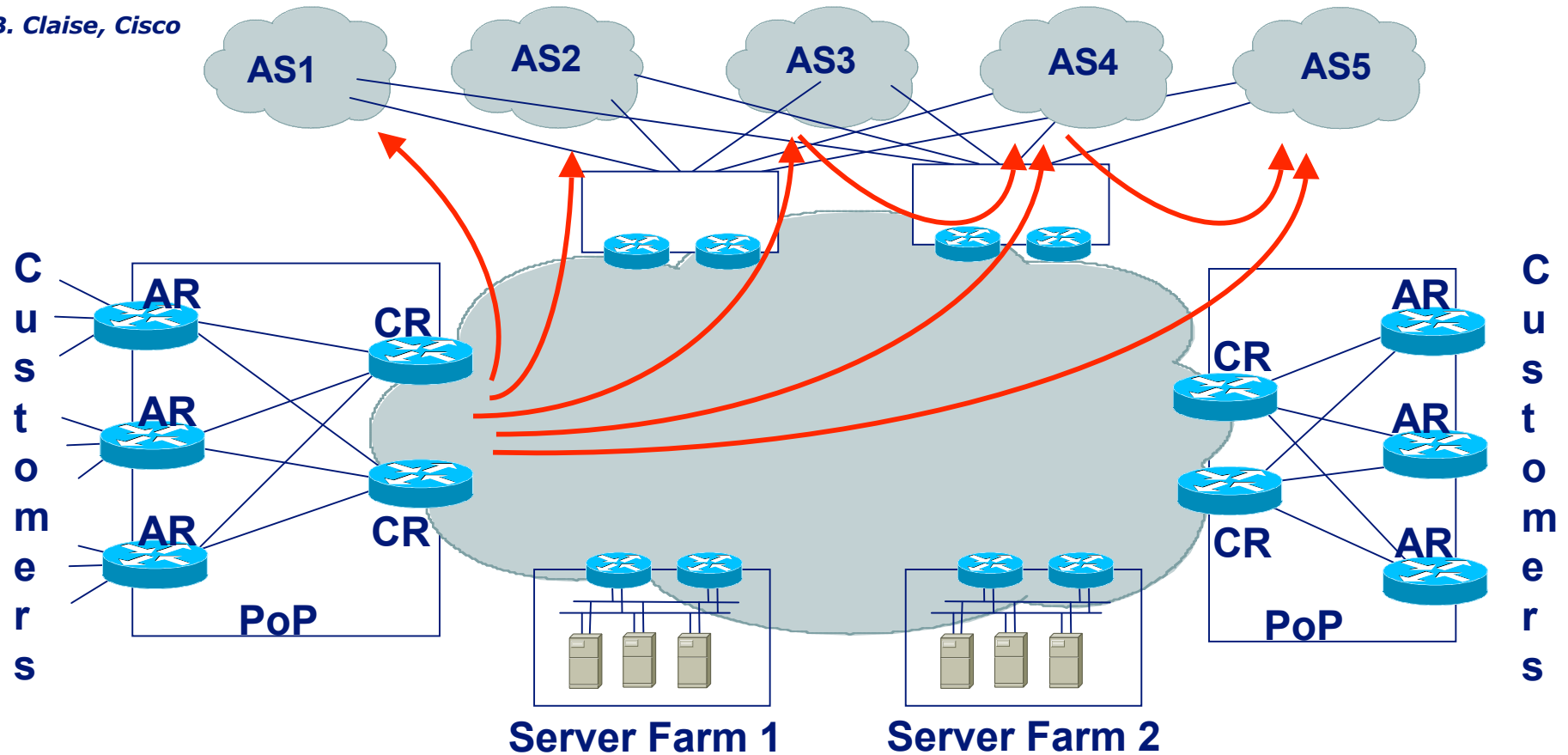
B. Claise, Cisco



- AR-to-AR or CR-to-CR
- Some PoPs, e.g. regional, may be outside MPLS mesh

# External Traffic Matrix

B. Claise, Cisco



- Router (AR or CR) to External AS (and vice versa)
- External AS to External AS (for transit providers)
- No BGP to Server Farms



# Measuring the Traffic Matrix

## Flows

- NetFlow
  - Not trivial but relatively common
  - Aggregate flows to network end points using BGP feed
- BGP Policy Accounting & Destination Class Usage
  - Limited to 16 or 64 buckets

## MPLS LSPs

- LDP
  - Used for VPNs
  - $O(N^2)$  meas.
- RSVP
  - Used for MPLS TE
  - $O(N^2)$  meas. if full mesh

# NetFlow

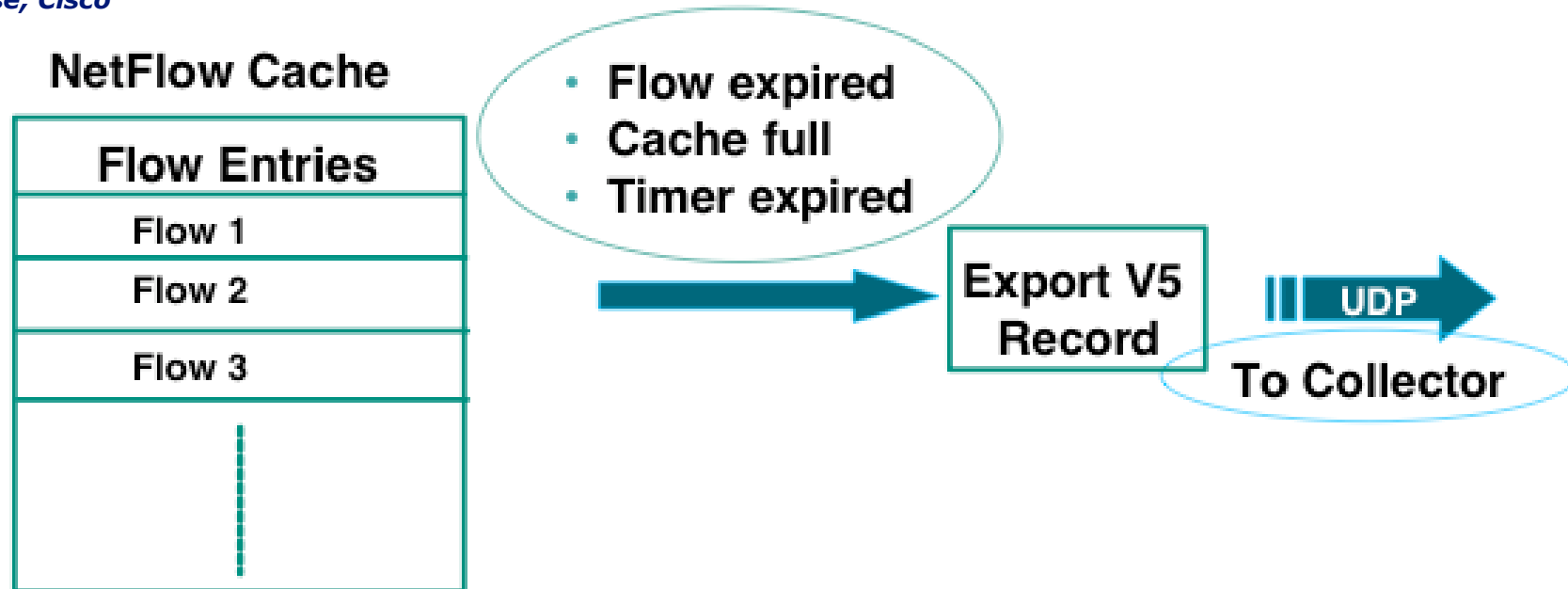


# NetFlow Background

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

# NetFlow Export

B. Claise, Cisco



- 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 Versions

- **Version 5**
  - most commonly available
- **Version 7**
  - on the switches
- **Version 8**
  - the Router Based Aggregation
- **Version 9** (circa 2003)
  - BGP NextHop TOS Aggregation
- **Supported by multiple vendors**
  - Cisco
  - Juniper
  - others



# Traffic Matrices from NetFlow v5/v8

- Enable NetFlow on *edge-of-model* interfaces
  - AR-CR interface for internal TM,
  - Peer-AR interface for external TM
- Export data to central collector(s)
- Correlate flows with *edge-of-model*
  - BGP passive peer on collector(*usual*)
  - Infer from peer-AS field
  - Static (e.g. list of address space for no-BGP server farm)
- Aggregate flow counts



# BGP Passive Peer on the Collector

- Export v5 with IP address or v8 with prefix aggregation (instead of peer-as or destination-as for source and destination)
- Correlate IP to iBGP NextHop
  - Can also look up: source/destination AS, AS Path, BGP communities, ... for BGP TE, peering analysis, etc.
- Existing Software
  - (<http://www.switch.ch/network/projects/completed/TF-NGN/floma/software.html>)
  - Arbor PeakFlow (marketed for security/DDoS detection)
  - Network Signature's BENTO
  - Ixia IxTraffic
  - Adlex (now Compuware)
  - ASFLOW (open source)
  - ...

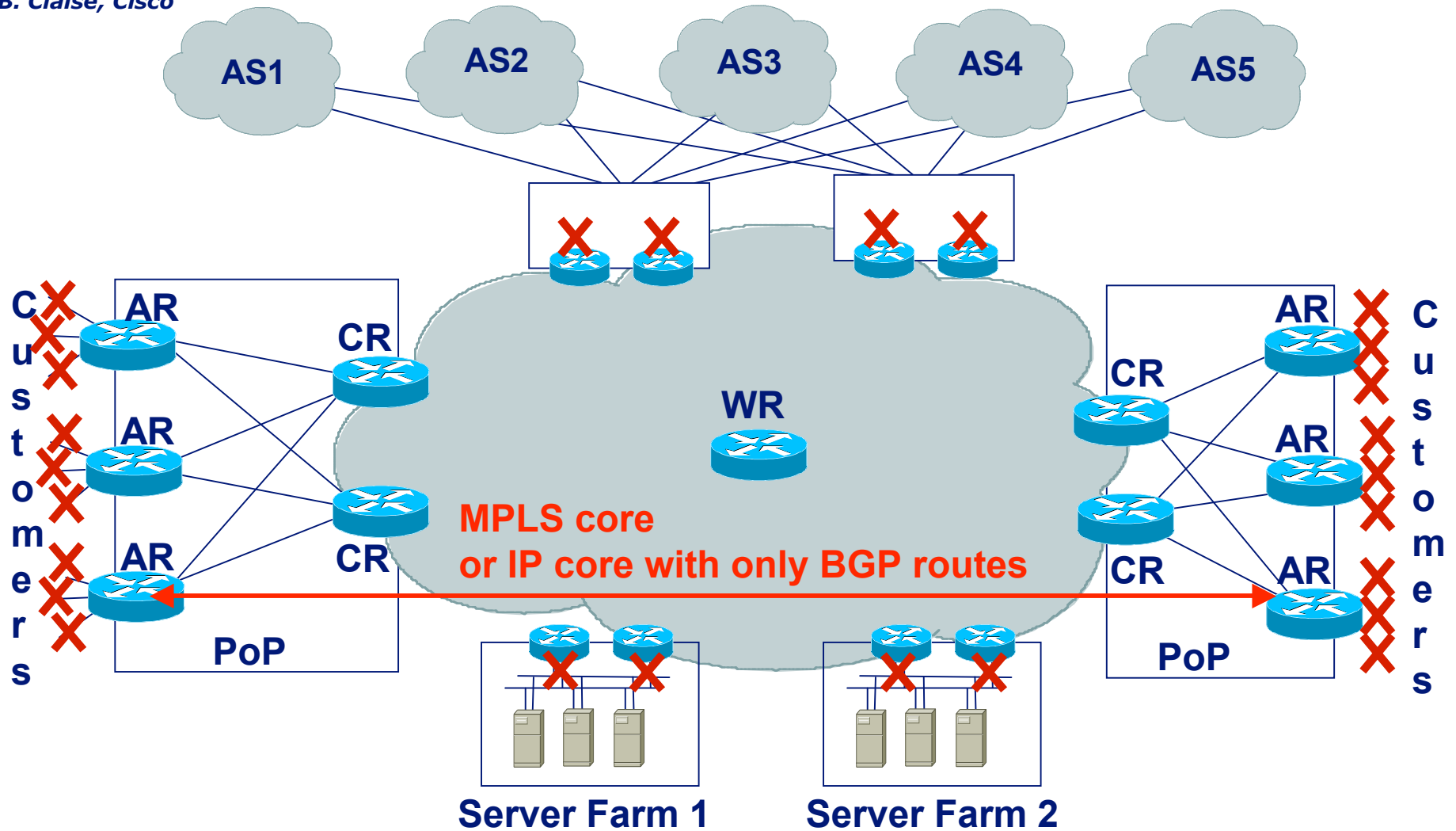


# BGP NextHop TOS Aggregation (v9)

- Aggregation scheme in NetFlow v9  
Router does Flow-NextHop correlation  
Exports traffic matrix (very convenient!)
- Configure on Ingress Interfaces
- Only for BGP routes
  - Non-BGP routes will have next-hop 0.0.0.0
- Only for IP packets
  - IP to IP, or IP to MPLS

# BGP NextHop TOS Aggregation


B. Claise, Cisco



# In Practice: NetFlow stats may not match link stats

- NetFlow stats undercount but not consistently:-(

Router implementation matters!  
Sampling is one cause but not always.



Interface	Traffic via SNMP(Mbps)	NetFlow/ SNMP
6	81	0.56
1	338	0.57
17	145	0.58
9	333	0.6
1	2210	0.61
33	4150	0.61
8	147	0.61
3	2290	0.62
32	1380	0.62
11	516	0.62
7	500	0.62
12	602	0.66
31	673	0.68



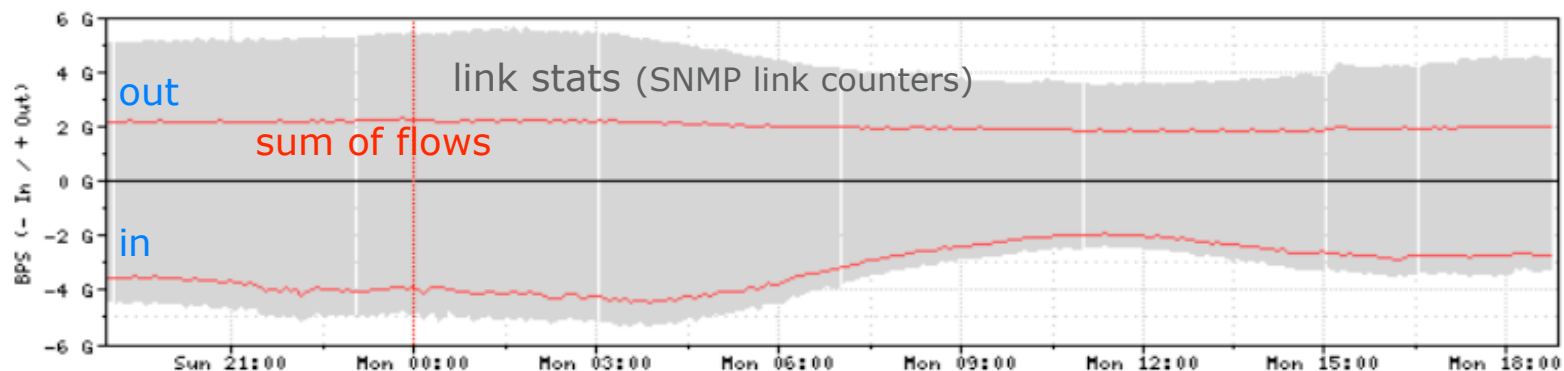
Top-of-the-line CR

Interface	Traffic via SNMP(Mbps)	NetFlow/ SNMP
45	1760	0.77
49	1730	0.78
58	1730	0.79
79	7750	0.8
74	7570	0.82
30	1350	0.85
34	7260	0.85
31	8840	0.86
61	7330	0.86
71	6310	0.86
39	1730	0.87
94	12710	0.87
98	12590	0.87
5	1760	0.88
27	11130	0.88
35	11130	0.88
36	5380	0.88
37	68	0.88
38	5320	0.89
39	71	0.89
8	1380	0.92
42	1370	0.93
57	1720	0.93
48	1720	0.94
44	1360	0.97
26	1730	0.98
29	0.396	13.31

Data from NetFlow tool in an operational ISP network

## NetFlow in Practice (2)

- Stats can clip at crucial times
  - NetFlow cache overflows at high traffic
  - CPU stops counting NetFlow when busy
- NetFlow and SNMP timescale mismatch
  - 10- or 15-minute typical (flows expire) vs. 2- or 5-minute SNMP link stats
- Poor implementations (e.g., bad outbound accounting)



# MPLS

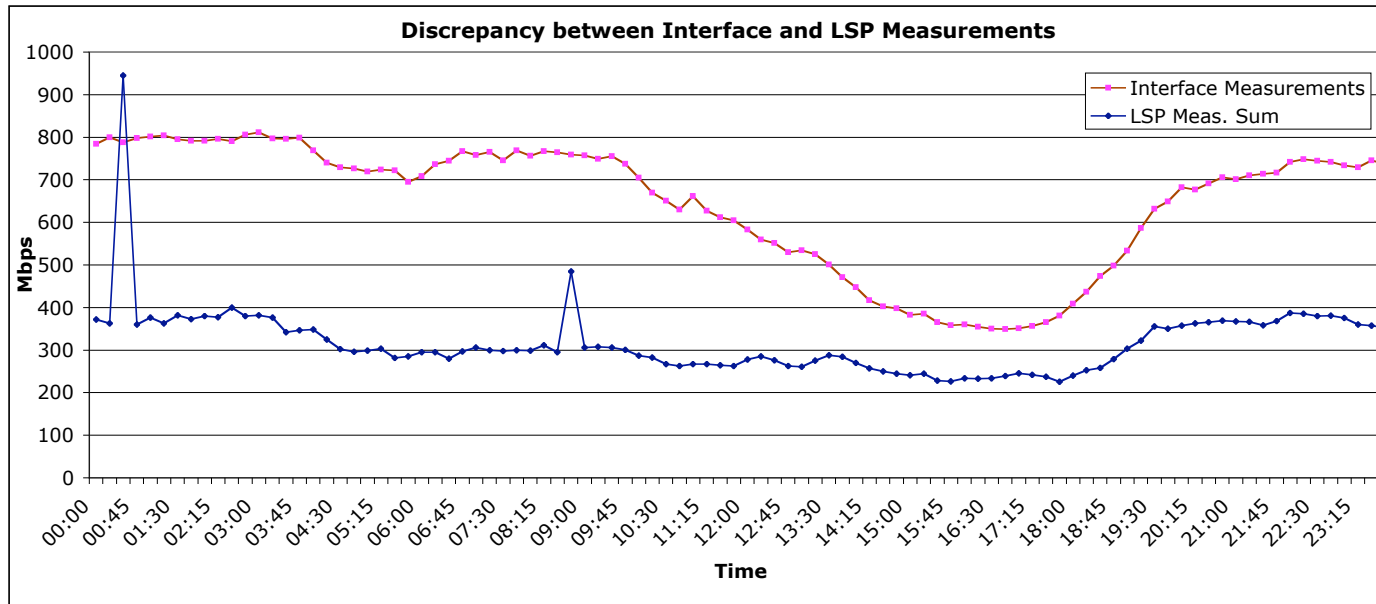
- MPLS LSPs (should be able to) provide internal traffic matrix directly
  - LDP: MPLS-LSR-MIB (or equivalent)
    - Mapping FEC to exit point of LDP cloud
    - Counters for packets that enter FEC (ingress)
    - Counters for packets switched per FEC (transit)
  - RSVP counters
- Does not provides external traffic matrix



# LDP in Practice

- Only transit statistics, no ingress statistic  
(on many versions of Cisco's IOS)
- Missing values  
(expected when making tens of thousands of measurements)
- Can take many minutes  
(important for tactical, quick response, TE)
- Not address external TM (of course)

# RSVP Possible Issues



Data from operational network:  
150 LSPs in one link

- Undercount link stats
- Not track well
- Volatile

- Also

- Problematic counters:

- reset on path reroute on many Junos implementations
    - missing all together on many Alcatel Lucent SR platforms

- Issues with  $O(N^2)$ : missing values, time, ...



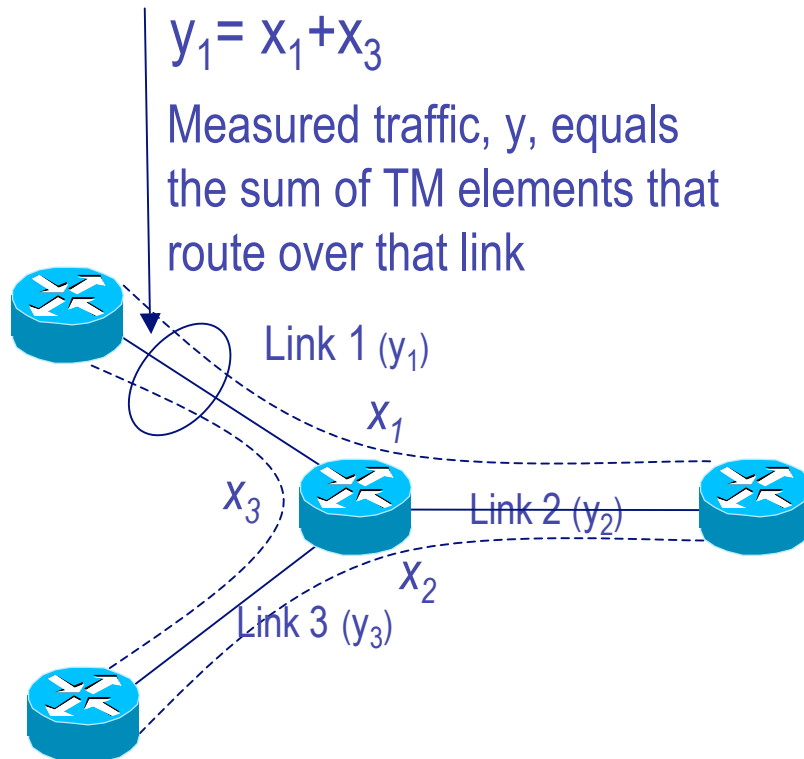
# LSP Stats Summary

- LSP stats good enough when:
  - Only need internal traffic matrix
  - Have full mesh of LSPs
  - Not getting bitten by various platform issues
  - Long-term analysis (not quick enough for tactical Ops)
- Otherwise, if use LSP stats, need to watch out for
  - missing
  - unreliable
  - unavailable
  - inconsistent
  - slow-to-gather data

# Estimation based on Link Stats (e.g. Tomogravity)

# Estimation Background

## Tomography, Tomogravity\*, etc.



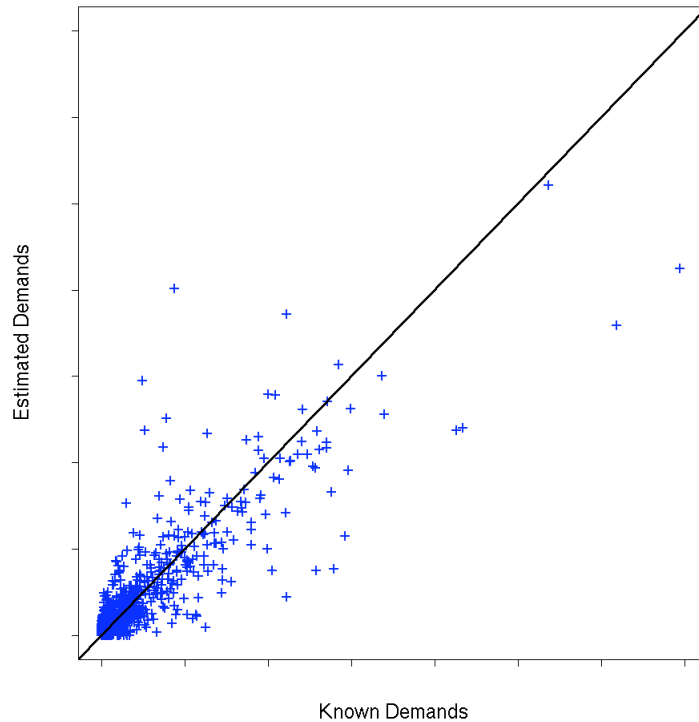
Interface Stats	Routing Matrix A	Traffic Matrix (as a vector)
$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$	$= \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$	$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \epsilon$

- Given  $Y$  and  $A$  solve for  $X$  (minimize  $\epsilon$ )
- Many solutions to above
  - Pick some *likely*  $X$   
(e.g., most conformant with gravity model)

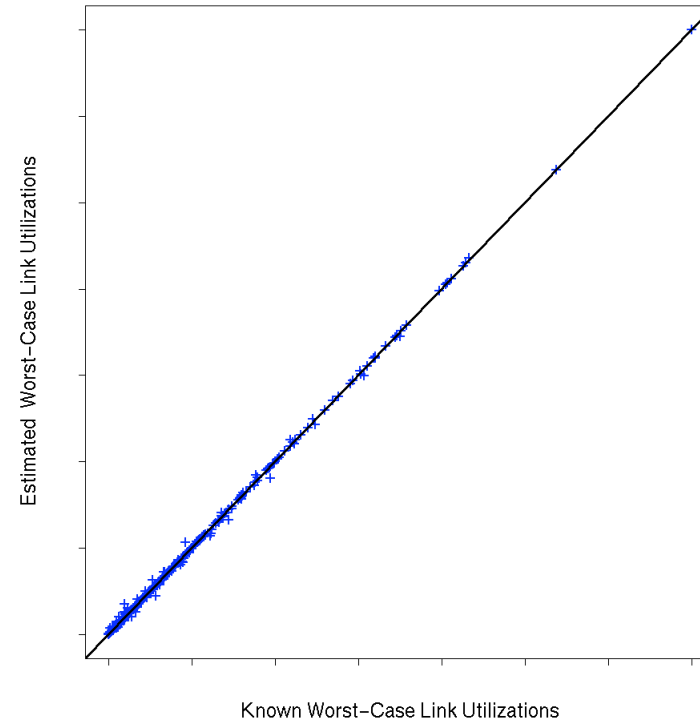
\* Zhang et al. (2004)

# Estimation Results

International Tier-1  
IP Backbone

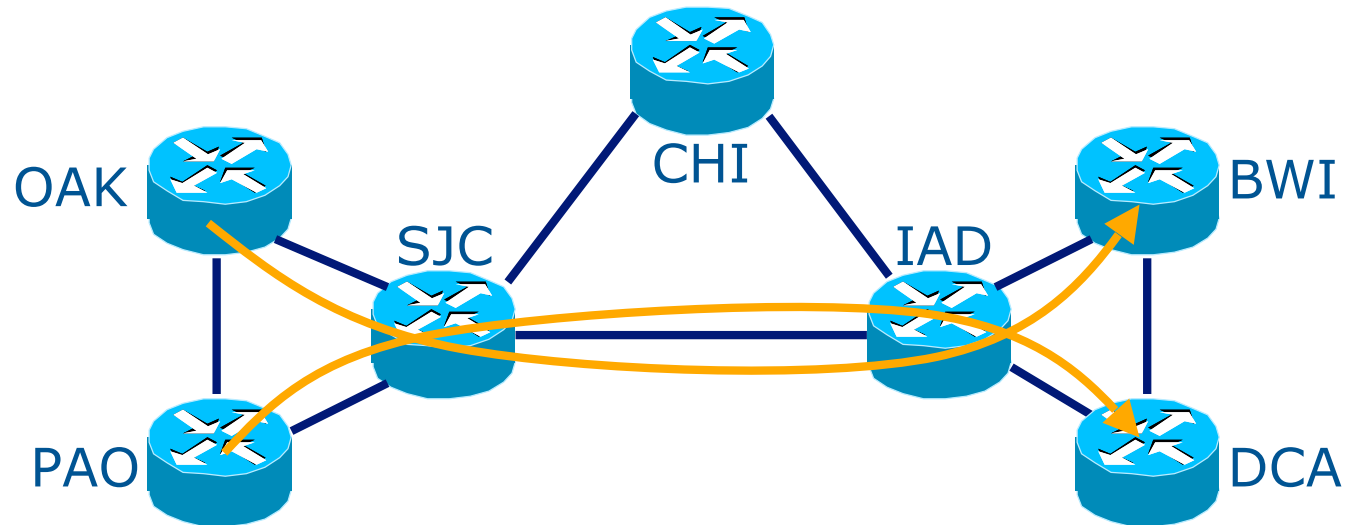


- Individual estimates are not accurate



- Results of using (the inaccurate) estimates in, failure analysis, for example, are accurate!

# 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)

# Regressed Measurements



# Regressed Measurements Overview

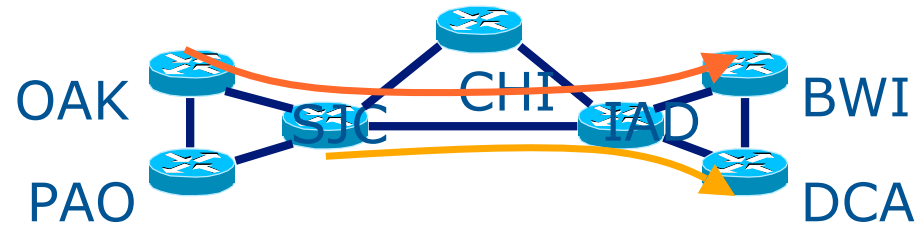
- Use interface stats as gold standard
  - Traffic management policies, almost always, based on interface stats (e.g.,
    - ops alarm if 5-min average utilization goes >90%
    - traffic engineering considered if any link util approach 80%
    - cap planning guideline is to not have link util above 90% under any single failure)
- Mold NetFlow, LSP stats, ... to match interface stats

# LSP Example for Regression

- Each LSP measurement adds a row to the  $Y=AX$

- RSVP measurement for OAK->BWI

$$Y_{\text{RSVP-OAK->BWI}} = X_{\text{OAK->BWI}}$$



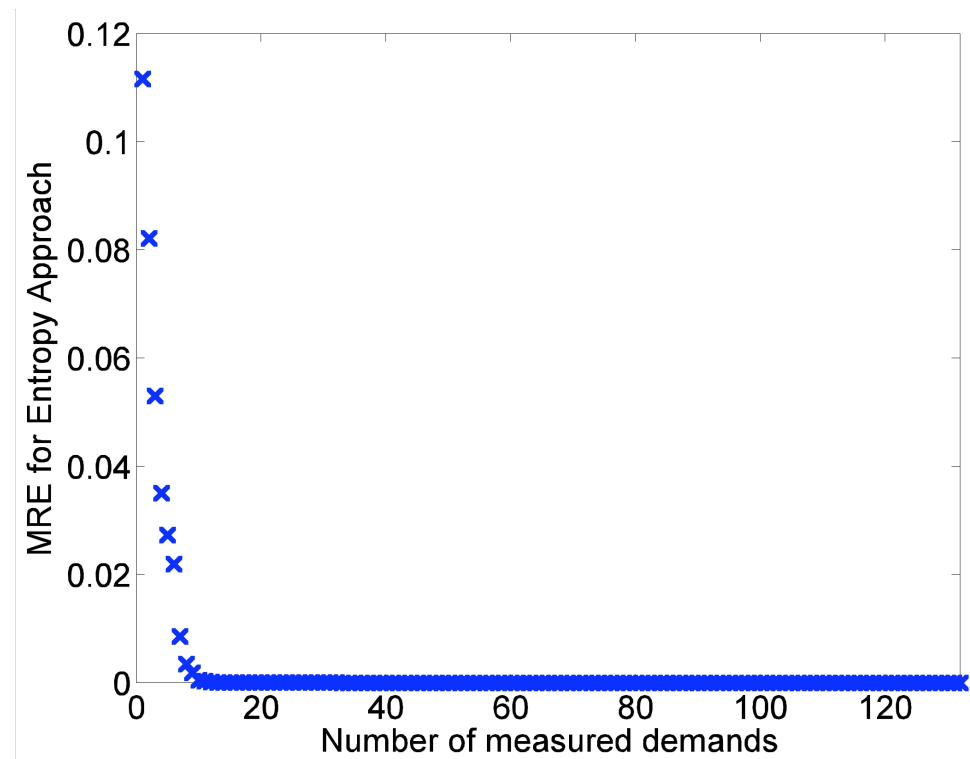
- Transit LDP measurement for SJC->BWI:

$$Y_{\text{Transit-SJC->DCA}} = X_{\text{OAK->DCA}} + X_{\text{PAO->DCA}}$$

- Solve for  $X$  such that there is strict conformance with link stat  $Y$  values with other measurements matched as best possible.

# Role of Netflow, LSP Stats,...

- Provide guidance on TM ratios, approximate scale  
(more stable than actual values)
- Can improve TM estimates with just a few measurements





# Regressed Measurements Sample

- Topology discovery done in real-time
- LDP measurements rolling every 30 minutes
- 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)



# Regressed Measurements Summary

- 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

# Overall Summary

- Direct Measurement works well sometimes
  - Netflow OK on some equipment
  - LSP counters OK on some equipment and if only care for internal traffic matrix
  - Watch out for scaling, speed and measurement mismatch with link stats
- Estimation on link stats works sometimes
  - Has great speed (order of time to measure link stats)
  - Validity for given topology must be verified
- Regression is most flexible
  - Provides a spectrum of solutions between measurement and estimation
- Best practice is to start simple, verify, add complexity only if required



# Best Practice: Start Simple, Verify

- Collect data over a few weeks
  - Link stats plus LSP and NetFlow stats (as available)
  - Make sure data set contains some failures:-)
- LSP or NetFlow stats good enough? (if so stop)
  - Compare sum of LSP, NetFlow against link counters
  - Compare failure utilization prediction against reality
- Link-based estimation good enough? (if so stop)
  - Again, test prediction against reality after failure
- Use Regressed Measurements on available data
  - Test, stop if predictions good enough
  - Otherwise add stats incrementally (e.g., additional NetFlow coverage)
  - Repeat this step until predictions are good

# References

- **Telkamp 2007**
  - Best Practices for Determining the Traffic Matrix in IP Networks V 3.0, NANOG 39, February 2007, Toronto.
- **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/>
- **Schnitter and Horneffer 2004**
  - S. Schnitter, T-Systems; M. Horneffer, T-Com. "Traffic Matrices for MPLS Networks with LDP Traffic Statistics." Proc. Networks 2004, VDE-Verlag 2004.



# MPLS aware NetFlow (reference)

- 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



# NetFlow Version 8 (reference)

- 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: Asymmetric BGP traffic

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

