

Internet Monitoring & Tools

Les Cottrell – SLAC

Presented at the HEP Networking, Grids and Digital Divide meeting Daegu, Korea May 23-27, 2005



Partially funded by DOE/MICS Field Work Proposal on Internet End-to-end Performance Monitoring (IEPM), also supported by IUPAP



Overview

- **Why is measurement difficult yet important?**
- **LAN vs WAN**
- **SNMP**
- **Effects of measurement interval**
- **Passive**
- **Active**
- **New challenges**

Why is measurement difficult?

- Internet's evolution as a composition of independently developed and deployed protocols, technologies, and core applications
- Diversity, highly unpredictable, hard to find “invariants”
- Rapid evolution & change, no equilibrium so far
 - Findings may be out of date
- Measurement not high on vendors list of priorities
 - Resources/skill focus on more interesting and profitable issues
 - Tools lacking or inadequate
 - Implementations poor & not fully tested with new releases
- ISPs worried about providing access to core, making results public, & privacy issues
- The phone connection oriented model (Poisson distributions of session length etc.) does not work for Internet traffic (heavy tails, self similar behavior, multi-fractals etc.)

Add to that ...

- Distributed systems are very hard
 - *A distributed system is one in which I can't get my work done because a computer I've never heard of has failed.* Butler Lampson
- Network is deliberately transparent
- The bottlenecks can be in any of the following components:
 - the applications
 - the OS
 - the disks, NICs, bus, memory, etc. on sender or receiver
 - the network switches and routers, and so on
- Problems may not be logical
 - Most problems are operator errors, configurations, bugs
- When building distributed systems, we often observe unexpectedly low performance
 - the reasons for which are usually not obvious
- Just when you think you've cracked it, in steps security

Why is measurement important?

- End users & network managers need to be able to identify & track problems
- Choosing an ISP, setting a realistic service level agreement, and verifying it is being met
- Choosing routes when more than one is available
- Setting expectations:
 - Deciding which links need upgrading
 - Deciding where to place collaboration components such as a regional computing center, software development
 - How well will an application work (e.g. VoIP)
- Application steering (e.g. forecasting)
 - Grid middleware, e.g. replication manager

Passive vs. Active Monitoring

- Active injects traffic on demand
- Passive watches things as they happen
 - Network device records information
 - Packets, bytes, errors ... kept in MIBs retrieved by SNMP
 - Devices (e.g. probe) capture/watch packets as they pass
 - Router, switch, sniffer, host in promiscuous (tcpdump)
- Complementary to one another:
 - Passive:
 - does not inject extra traffic, measures real traffic
 - Polling to gather data generates traffic, also gathers large amounts of data
 - Active:
 - provides explicit control on the generation of packets for measurement scenarios
 - testing what you want, when you need it.
 - Injects extra artificial traffic
- Can do both, e.g. start active measurement and look at passively

Passive tools

- SNMP
- Hardware probes e.g. Sniffer, NetScout, can be stand-alone or remotely access from a central management station
- Software probes: snoop, tcpdump, require promiscuous access to NIC card, i.e. root/sudo access
- Flow measurement: netramet, OCxMon/CoralReef, Netflow
- Sharing measurements runs into security/privacy issues

Example: Passive site border monitoring

- Use **Cisco Netflow** in Catalyst 6509 with MSFC, on SLAC border
- Gather about 200MBytes/day of flow data
- The raw data records include source and destination addresses and ports, the protocol, packet, octet and flow counts, and start and end times of the flows
 - Much less detailed than saving headers of all packets, but good compromise
 - Top talkers history and daily (from & to), tlds, vlans, protocol and application utilization
- Use for network & security

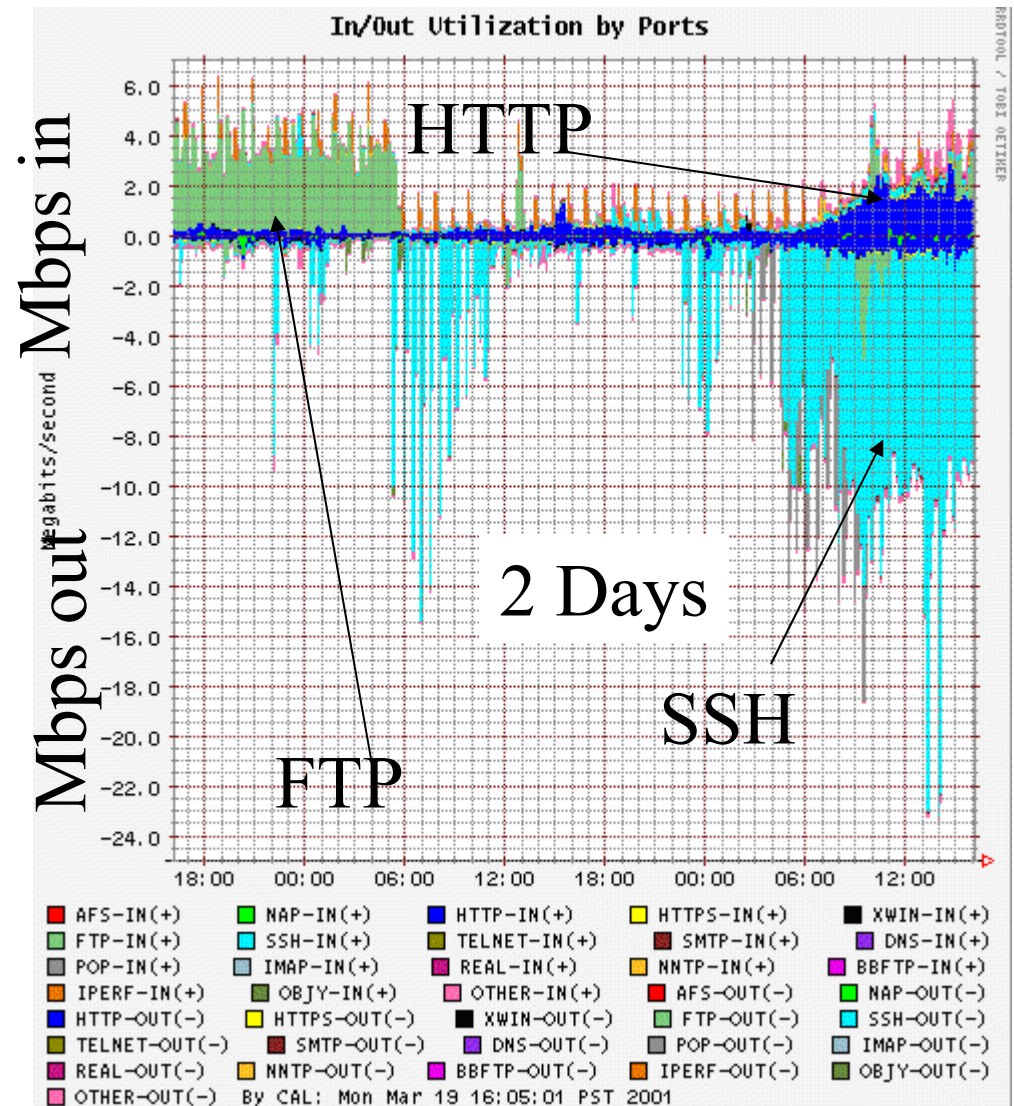
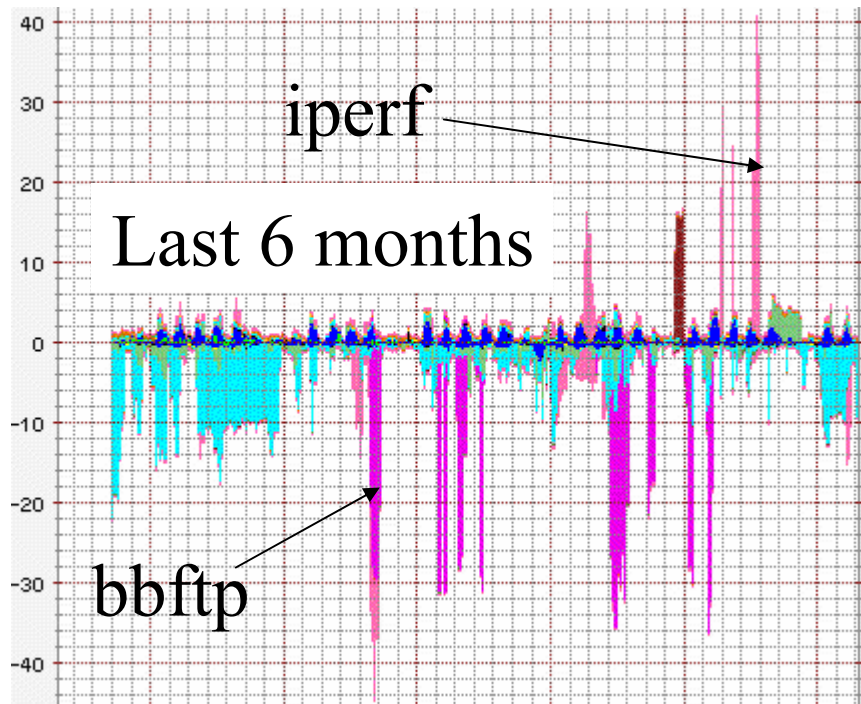
SLAC Traffic profile

SLAC offsite links:

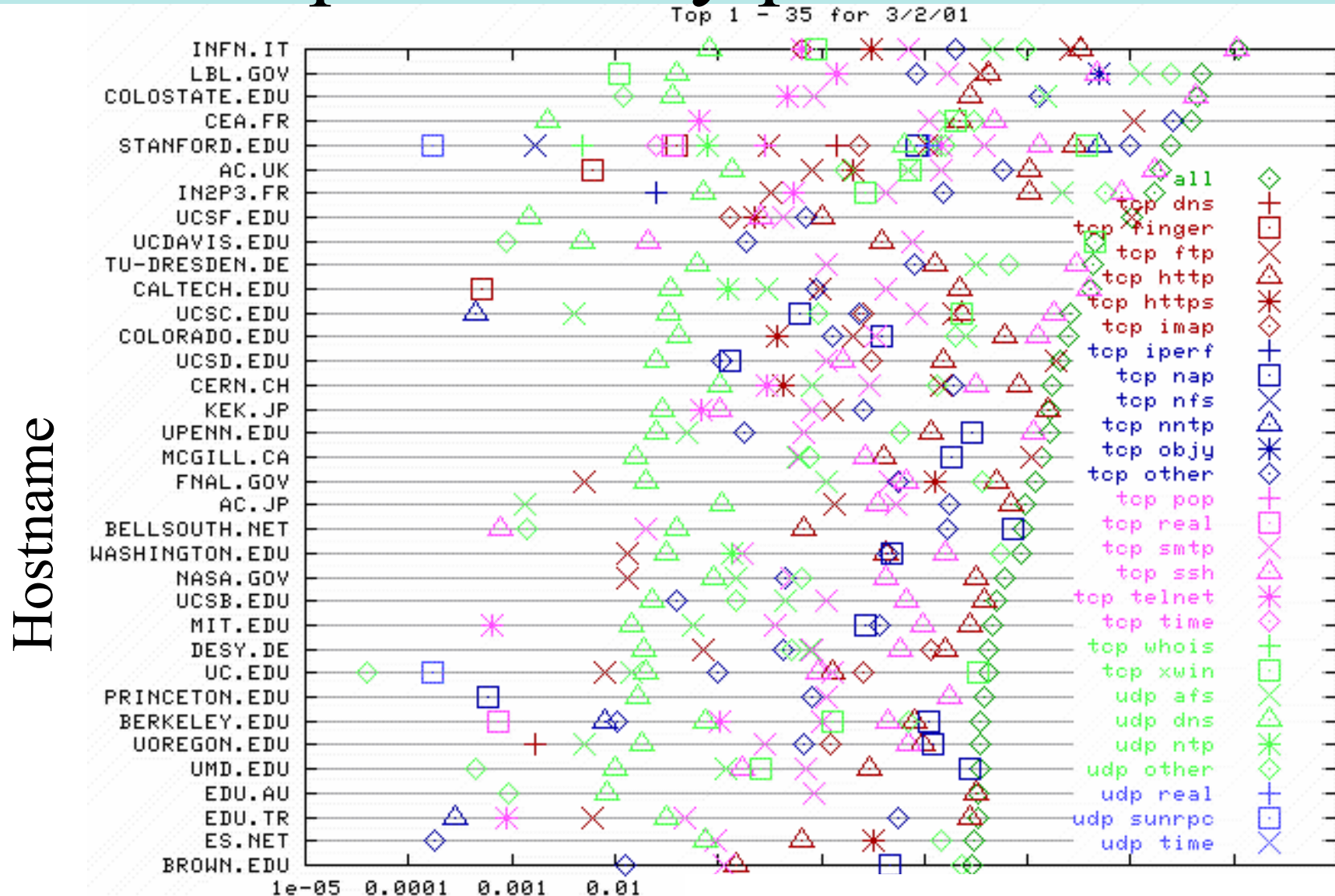
OC3 to ESnet, 1Gbps to Stanford U & thence OC12 to I2
 OC48 to NTON

Profile

bulk-data xfer dominates



Top talkers by protocol

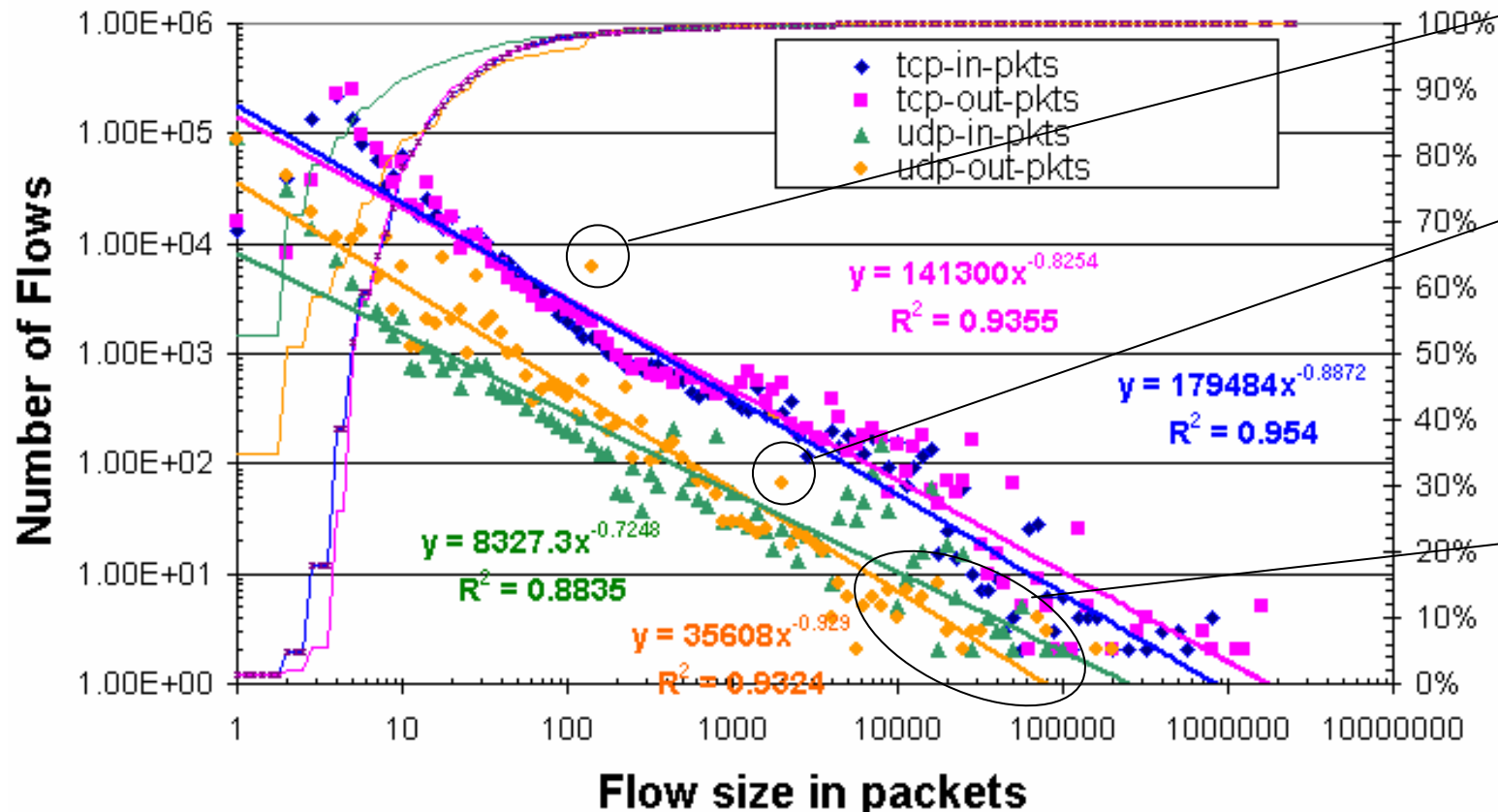


Volume dominated by single
Application - bbcp

1 100 10000
MBytes/day (log scale) 10

Flow sizes

Flow size distribution at SLAC border April 9, 2001



Heavy tailed, in ~ out, UDP flows shorter than TCP, packet~bytes
 75% TCP-in < 5kBytes, 75% TCP-out < 1.5kBytes (<10pkts)
 UDP 80% < 600Bytes (75% < 3 pkts), ~10 * more TCP than UDP
 Top UDP = AFS (>55%), Real(~25%), SNMP(~1.4%)

SNMP

Real
A/V

AFS
file
server

Flow lengths

- 60% of TCP flows less than 1 second
- Would expect TCP streams longer lived
 - But 60% of UDP flows over 10 seconds, maybe due to heavy use of AFS

Some Active Measurement Tools

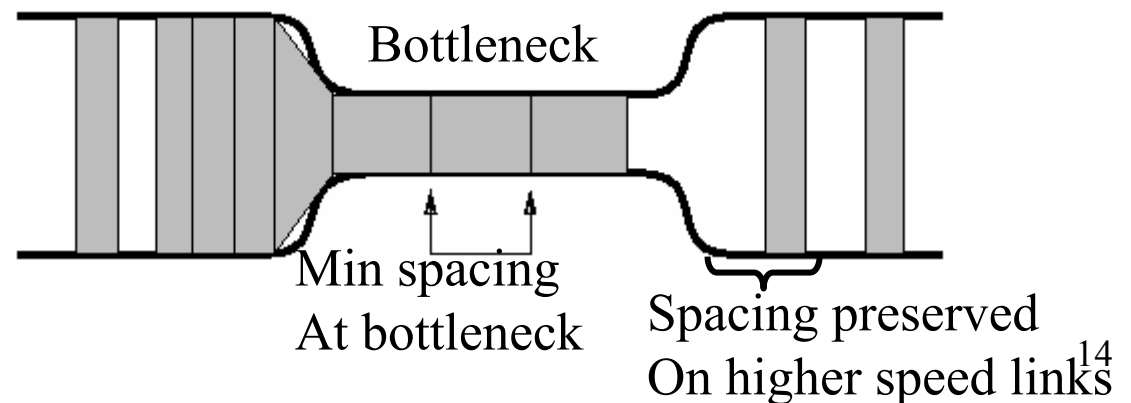
- Ping connectivity, RTT & loss
 - flavors of ping, fping, Linux vs Solaris ping
 - but blocking & rate limiting
- Alternative synack, but can look like DoS attack
- Sting: measures one way loss
- Traceroute
 - Reverse traceroute servers
 - Traceroute archives
- Combining ping & traceroute,
 - traceping, pingroute
- Pathchar, pchar, pipechar, bprobe, abing etc.
- Iperf, netperf, ttcp, FTP ...

Path characterization

- Pathchar/pchar
 - sends multiple packets of varying sizes to each router along route
 - plot min RTT vs packet size to get bandwidth
 - calculate differences to get individual hop characteristics
 - measures for each hop: BW, queuing, delay/hop
 - can take a long time
 - may be able to ID location of bottleneck

- Abing/pathload/pathchirp

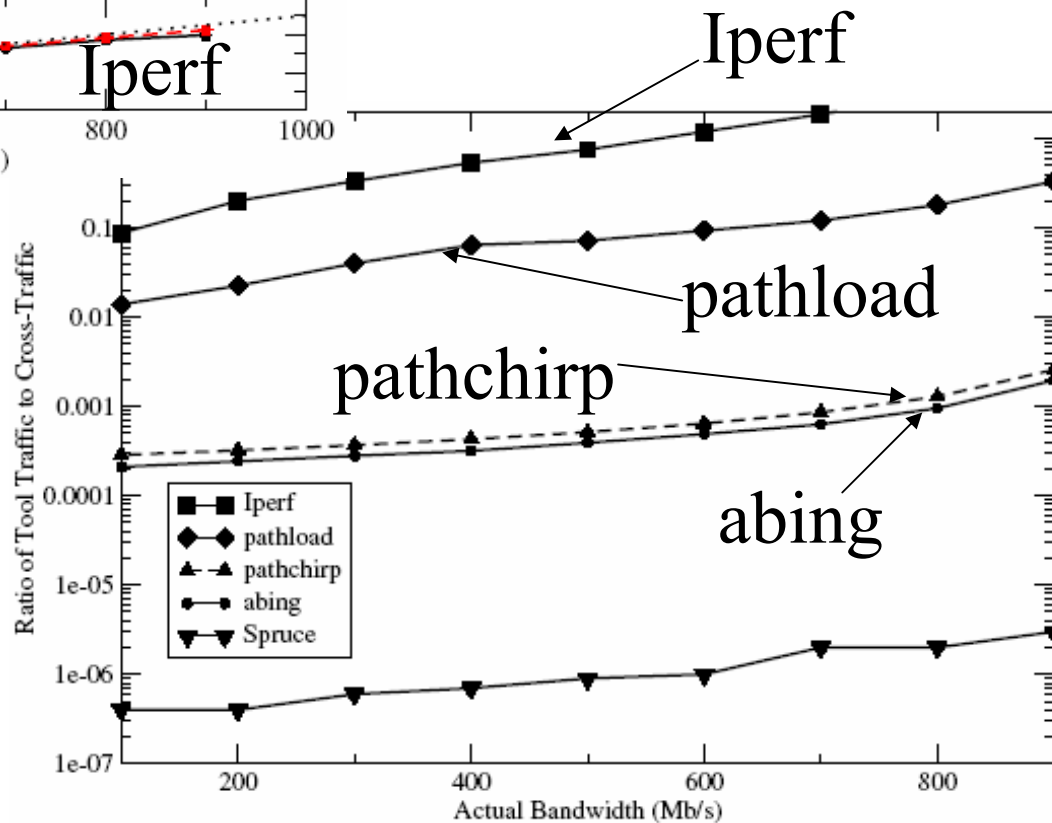
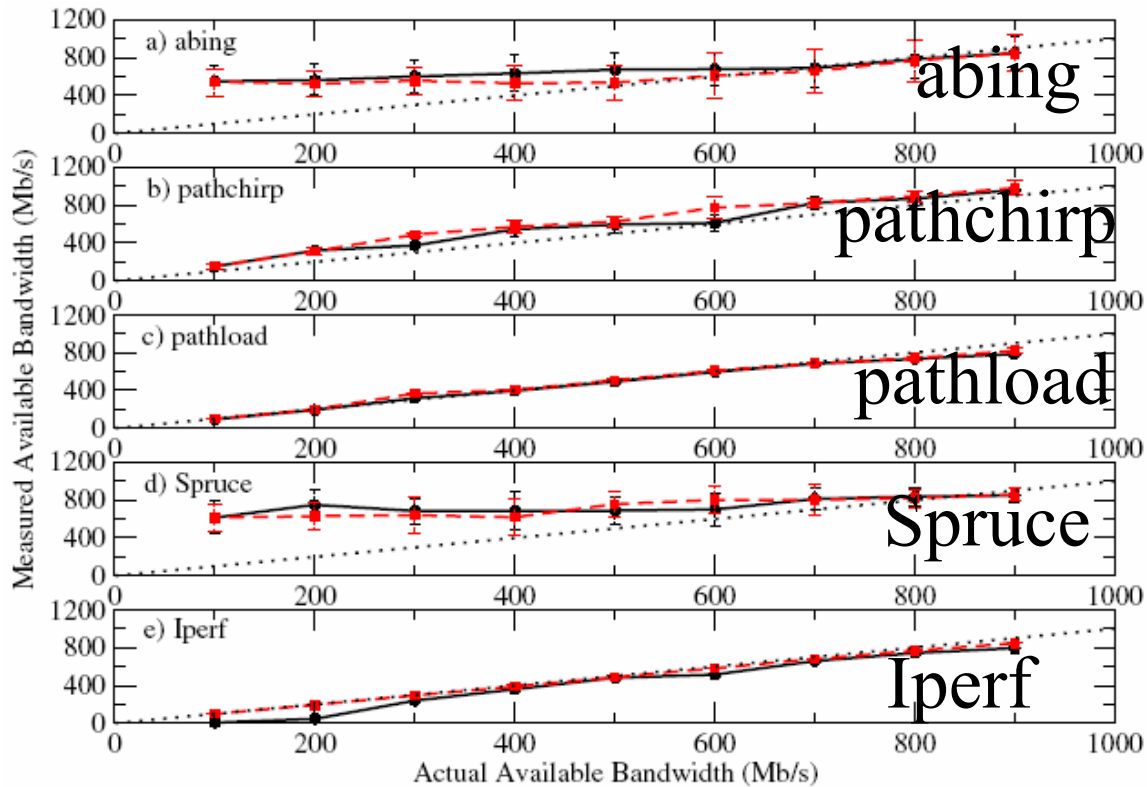
- Sends packets with known separation. measure separation at other e:
- Much faster
- Finds bottleneck bw



Network throughput

- Iperf/thrulay
 - Client generates & sends UDP or TCP packets
 - Server receives receives packets
 - Can select port, maximum window size, port , duration, parallel streams, Mbytes to send etc.
 - Client/server communicate packets seen etc.
 - Reports on throughput
 - Requires sever to be installed at remote site, i.e. friendly administrators or logon account and password
- Applications
 - GridFTP, bbcp, bbftp (single, multi-stream file to file)

Intrusiveness VS Accuracy



Active Measurement Projects

- PingER (ping)
- AMP (ping)
- One way delay:
 - Surveyor (now defunct), RIPE (mainly Europe), owamp
- IEPM-BW (bandwidth, throughput ...)
- NIMI (mainly a design infrastructure)
- NWS (mainly for forecasting)
- Skitter
- All projects measure routes
- For a detailed comparison see:
 - www.slac.stanford.edu/comp/net/wan-mon/iepm-cf.html
 - www.slac.stanford.edu/grp/scs/net/proposals/infra-mon.html

Some Challenges

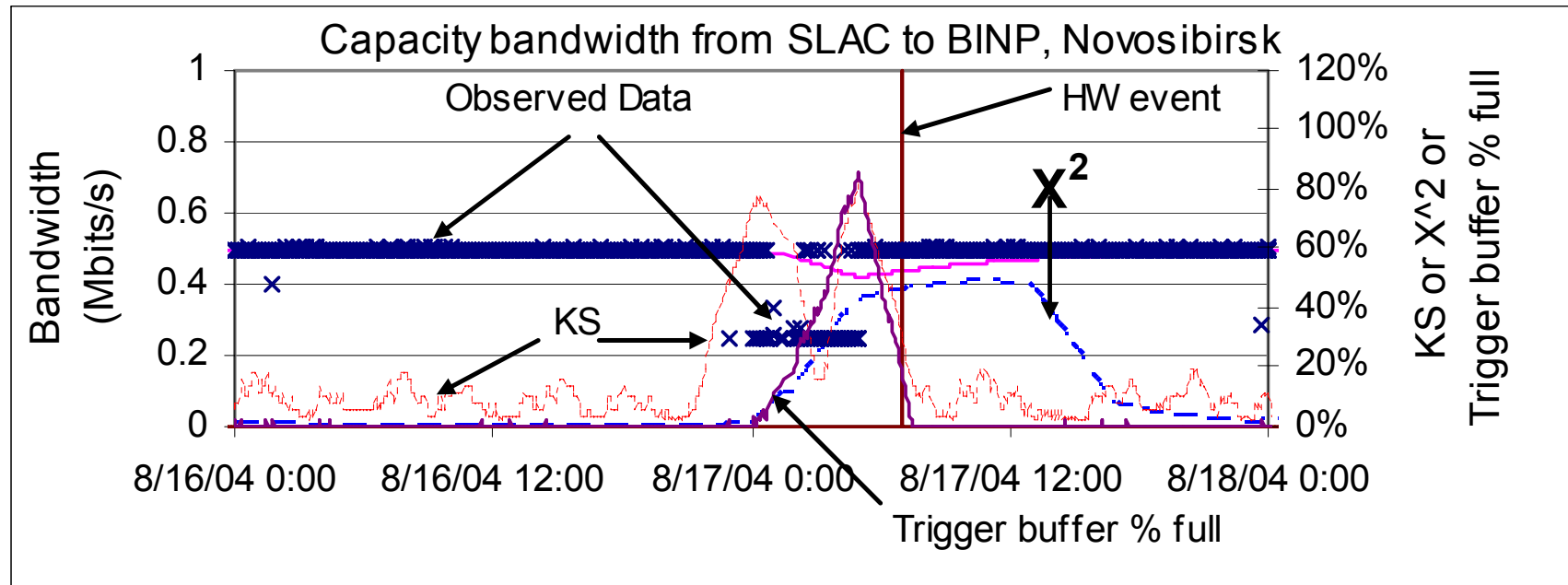
- High performance links
- Dedicated circuits
- Visualizing topologies, e.g. traceroutes
- Reviewing thousands of graphs to spot anomalies
 - Automated anomalous event detection
 - Gathering more information & alerting
- Guiding middleware
 - Need long term forecasts,
 - Web services
 - E.g. scheduling wavelengths, or QoS services

Hi-perf Challenges

- Packet loss hard to measure by ping
 - For 10% accuracy on BER $1/10^8 \sim 1$ day at 1/sec
 - Ping loss \neq TCP loss
- Iperf/GridFTP throughput at 10Gbits/s
 - To measure stable (congestion avoidance) state for 90% of test takes ~ 60 secs ~ 75 GBytes
 - Requires scheduling implies authentication etc.
- Using packet pair dispersion can use only few tens or hundreds of packets, however:
 - Timing granularity in host is hard (sub μ sec)
 - NICs may buffer (e.g. coalesce interrupts. or TCP offload) so need info from NIC or before
- Security: blocked ports, firewalls, keys vs. one time passwords, varying policies ... etc.

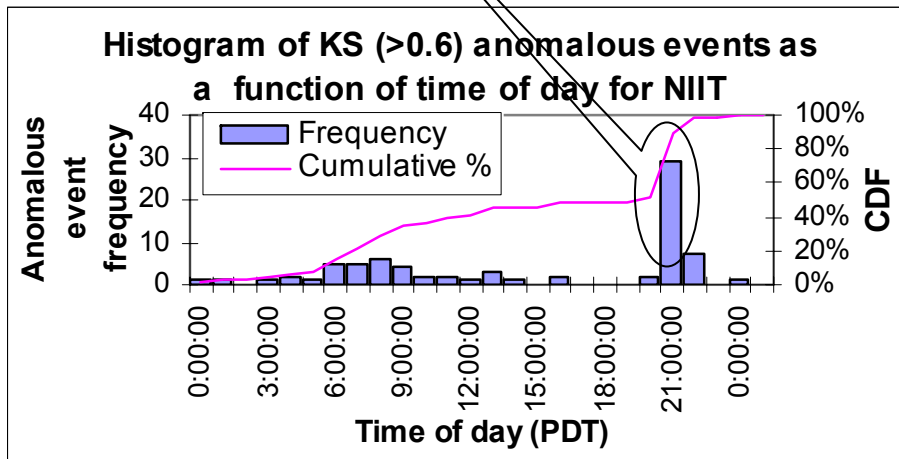
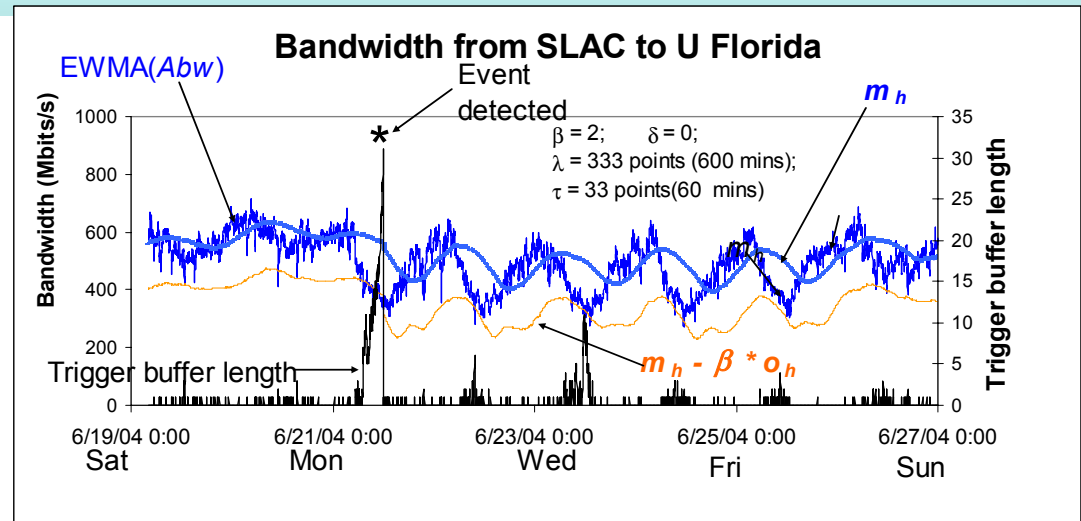
Anomalous Event Detection

- Relatively easy to spot steps in performance if the time series is normally pretty flat
 - Plateau algorithm, nicely intuitive
 - Kolmogorov Smirnov



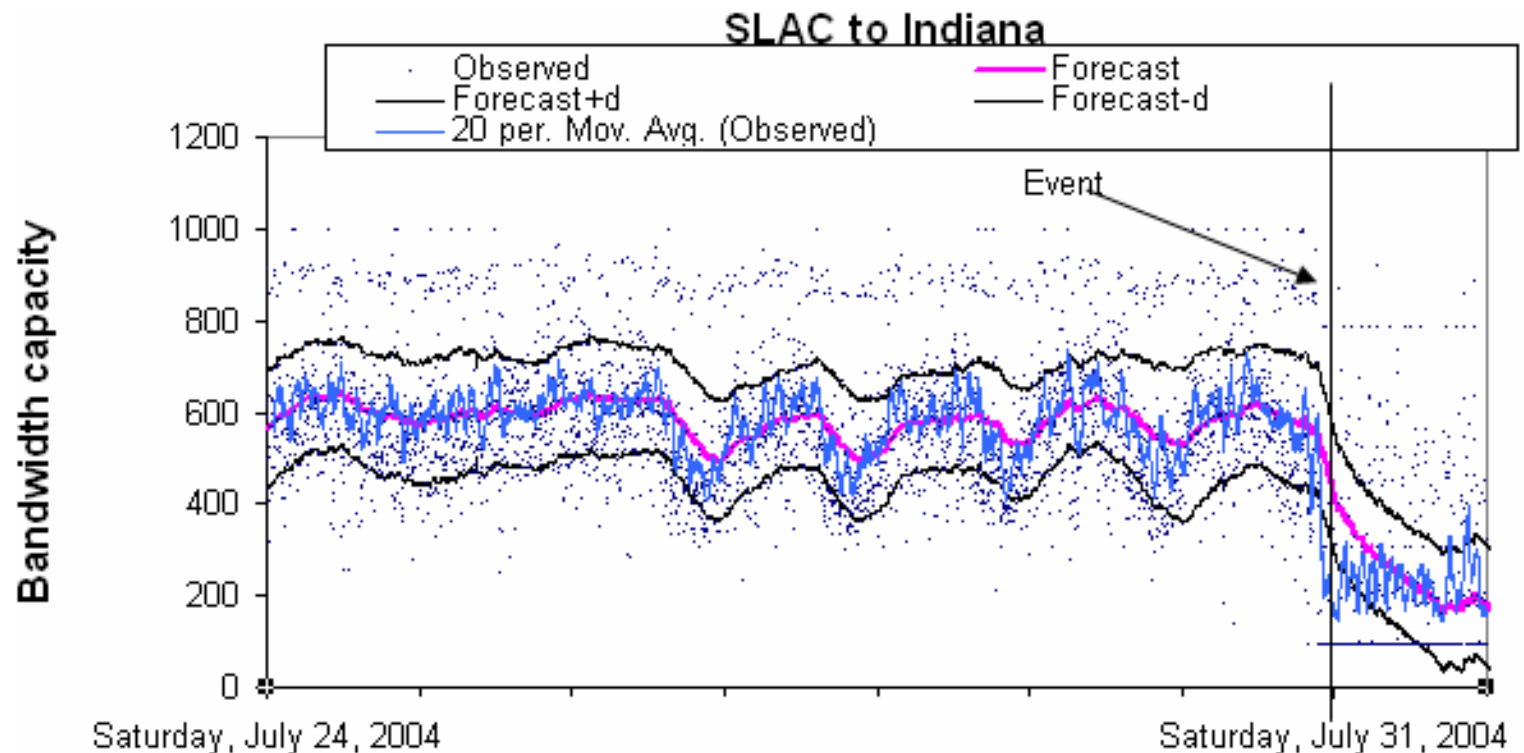
Seasonal Variations

- Unfortunately some (10-29%) paths show large diurnal changes
- These can cause false positives



Holt Winters

- So use Holt-Winters triple exponential weighted moving averages
 - Short term smoothing
 - Long term linear trends
 - Seasonal smoothing
- **Much better agreement**, removes diurnal & week start false positives
- **Also gives long term forecasts** – can use for scheduling etc.



Visualizing traceroutes

- One compact page per day
- One row per host, one column per hour
- One character per traceroute to indicate pathology or change (usually period(.) = no change)
- Identify unique routes with a number
 - Be able to inspect the route associated with a route number
 - Provide for analysis of long term route evolutions

[Yesterday's Summary](#) | [Reverse Traceroute Summary](#) | [Directory of Historical Traceroutes](#)

Checking a box for a node(s) and an hour(s) and pressing SUBMIT will provide topology m

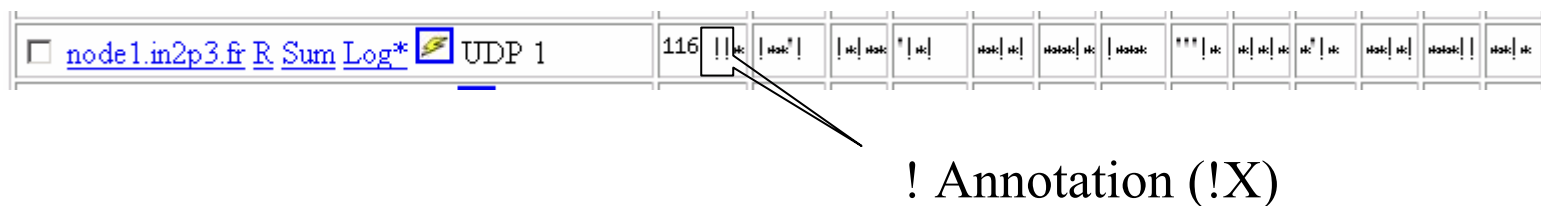
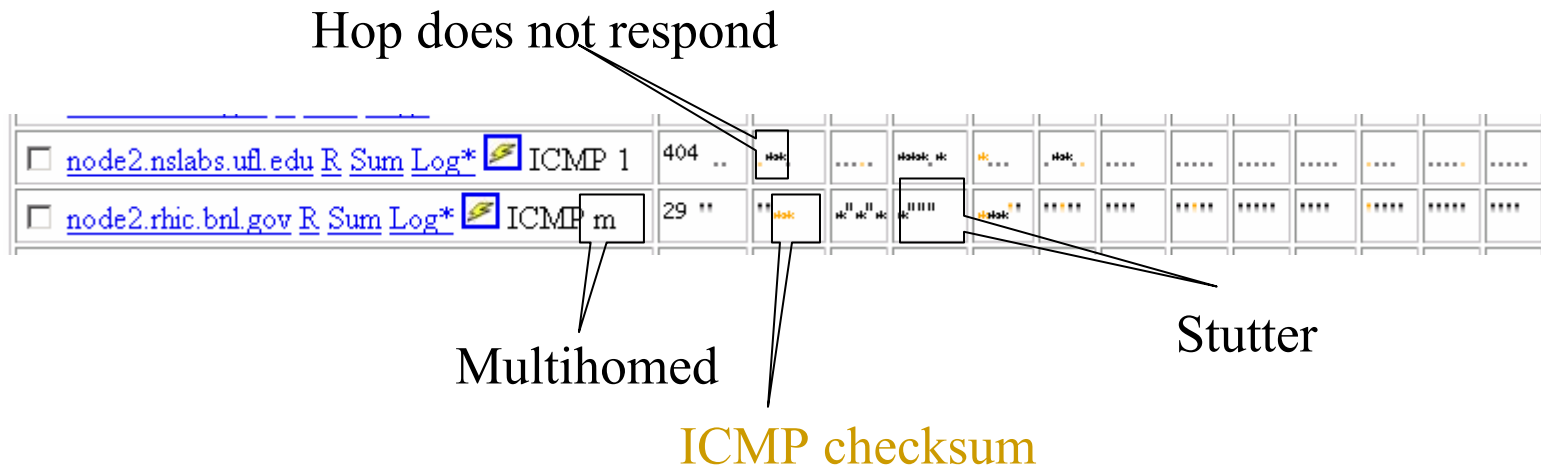
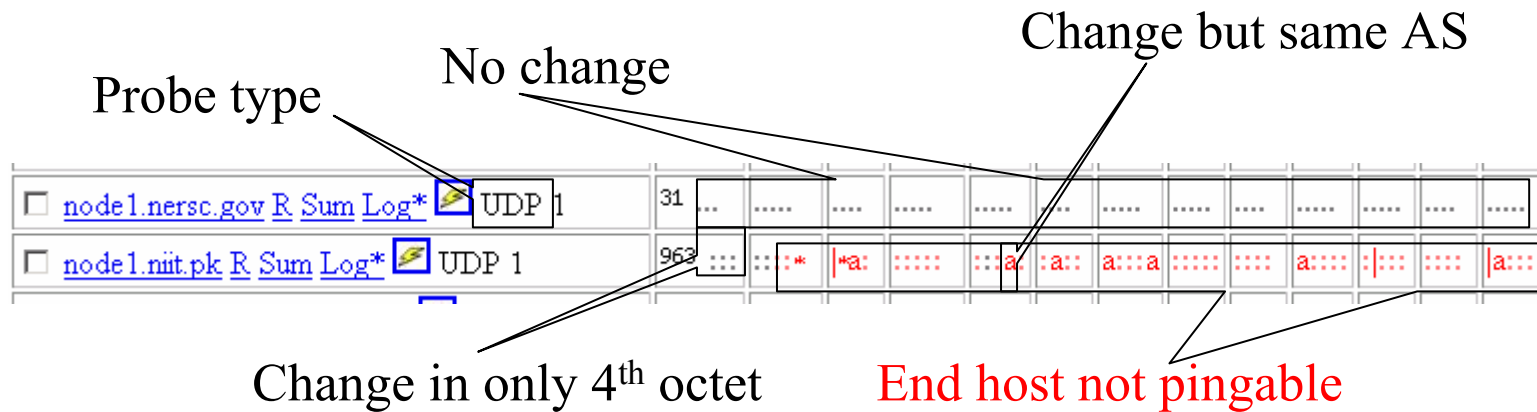
NODE \ Hour (Pacific Time)=>	<input type="checkbox"/> 00	<input type="checkbox"/> 01	<input type="checkbox"/> 02	<input type="checkbox"/> 03	<input type="checkbox"/> 04
<input type="checkbox"/> node1.cacr.caltech.edu* R Sum Log*	202
<input type="checkbox"/> node1.cesnet.cz* R Sum Log*	35 ...	68	35
<input type="checkbox"/> node1.clrc.ac.uk* R Sum Log*	91 ...	112	91
<input type="checkbox"/> node1.dl.ac.uk* R Sum Log*	97 ...	155	97
<input type="checkbox"/> node1.ece.rice.edu* R Sum Log*	241
<input type="checkbox"/> node1.fnal.gov* R Sum Log*	8 ...	48	8
<input type="checkbox"/> node1.in2p3.fr* R Sum Log*	29 ...	131	30

Route # at start of day, gives idea of route stability

Multiple route changes (due to GEANT), later restored to original route


Period (.) means no change 23

Pathology Encodings



Navigation

traceroute to CCSVSN04.IN2P3.FR (134.158.104.199), 30 hops max, 38 byte packets
 1 rtr-gsr-test (134.79.243.1) 0.102 ms
 ...
 13 in2p3-lyon.cssi.renater.fr (193.51.181.6) 154.063 ms !X

[node1.in2p3.fr](#) [R](#) [Sum](#) [Log*](#)  UDP 1 116 ||*|

```

#date          time numhops epoch      rtno  route
07/08/2004 00:10:46      13 1089270646  116  (134.79.243.1), (134.
07/08/2004 00:25:41      14 1089271541  115  (134.79.243.1), (134.
07/08/2004 00:40:25      15 1089272425  114  (134.79.243.1), (134.
07/08/2004 00:55:24      13 1089273324  116  (134.79.243.1), (134.
  
```







Date/Time	Hop 1	Hop 2	Hop 3	Hop 4
07/08_00:10	SLAC 0.102 ms	SLAC 0.210 ms	(192.68.191.146) 0.286 ms slac-rt4.es.net	(134.55.209.6) 0.610 n snv-pos
07/08_00:25	SLAC 0.100 ms	SLAC 0.239 ms	(192.68.191.146) 0.273 ms slac-rt4.es.net	(134.55.209.6) 0.633 n snv-pos
07/08_00:40	SLAC 0.107 ms	SLAC 0.273 ms	(192.68.191.146) 0.309 ms slac-rt4.es.net	(134.55.209.6) 0.676 n snv-pos
07/08_00:55	SLAC 0.261 ms	SLAC 0.236 ms	(192.68.191.146) 0.315 ms slac-rt4.es.net	(134.55.209.6) 0.669 n snv-pos

```

#rt#  firstseen  lastseen  route
0  1086844945  1089705757  ...,192.68.191.83,137.164.23.41,137.164.22.37, ...,131.215.xxx.xxx
1  1087467754  1089702792  ...,192.68.191.83,171.64.1.132,137, ...,131.215.xxx.xxx
2  1087472550  1087473162  ...,192.68.191.83,137.164.23.41,137.164.22.37, ...,131.215.xxx.xxx
3  1087529551  1087954977  ...,192.68.191.83,137.164.23.41,137.164.22.37, ...,131.215.xxx.xxx
4  1087875771  1087955566  ...,192.68.191.83,137.164.23.41,137.164.22.37, ..., (n/a),131.215.xxx.xxx
5  1087957378  1087957378  ...,192.68.191.83,137.164.23.41,137.164.22.37, ...,131.215.xxx.xxx
6  1088221368  1088221368  ...,192.68.191.146,134.55.209.1,134.55.209.6, ...,131.215.xxx.xxx
7  1089217384  1089615761  ...,192.68.191.83,137.164.23.41,(n/a), ...,131.215.xxx.xxx
8  1089294790  1089432163  ...,192.68.191.83,137.164.23.41,137.164.22.37,(n/a), ...,131.215.xxx.xxx
  
```

History Channel

[Today's Summary](#) | [Previous day's Summary](#) | [Reverse Traceroute Summary](#) | [Directory of Historical Traceroutes](#) | [Help](#)

	Parent Directory	14-Apr-2004 09:42
	2002 09/	30-Apr-2004 16:18
	2002 10/	30-Apr-2004 16:23
	2002 11/	30-Apr-2004 16:28
	2002 12/	30-Apr-2004 16:32
	2003 01/	30-Apr-2004 16:38

Character encoding of routes

- A '!' indicates that the traceroute was exactly the same as the previous one.
- A '!' indicates that the traceroute was exactly the same as the previous one, but that the datapoint is from the bw-tests regular run and not the more frequent times an hour runs.
- A '!' indicates that the traceroute was exactly the same as the previous one, but an ! annotation was found in the traceroute.
- A '|' indicates that the last hop was not reachable (i.e. the traceroute terminated after 30 hops, possibly the end host is behind a firewall).
- A red '|' indicates that the unreachable last hop, was also not pingable (probably host was unreachable).

AS' information

[Today's Summary](#) | [Previous day's Summary](#) | [Directory of Historical Traceroutes](#) | [Help](#)

SUBMIT Topology request | **SUBMIT Traceroute/ASN request** | RESET FIELDS | trace

NODE \ Hour (Pacific Time)=>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	00	01	02	03	04	05	06	07
<input checked="" type="checkbox"/> node1.binp.nsk.su R Sum Log UDP 1	0
<input type="checkbox"/> node1.cacr.caltech.edu R Sum Log UDP 1	0

traceroute to rainbow.inp.nsk.su (193.124.167.29), 30 hops max, 38 byte packets AS5402: BINP
1 rtr-gsr-test (134.79.243.1) 0.134 ms AS3671: SU-SLAC
2 rtr-dmz1-ger (134.79.135.15) 0.242 ms AS3671: SU-SLAC
3 slac-rt4.es.net (192.68.191.146) 0.339 ms SLAC-1: Stanford
4 snv-pos-slac.es.net (134.55.209.1) 0.933 ms AS293: Energy
5 chicr1-oc192-snvcr1.es.net (134.55.209.54) 48.989 ms AS293: Energy
6 aoacr1-oc192-chicr1.es.net (134.55.209.58) 69.059 ms AS293: Energy
7 aoapr1-ge0-aoacr1.es.net (134.55.209.110) 69.592 ms AS293: Energy
8 198.124.216.126 (198.124.216.126) 256.832 ms AS291: ESnet-CIDR-A
9 keksw2-ns.kek.jp (130.87.4.35) 266.092 ms AS2505: KEK



Changes in network topology (BGP) can result in dramatic changes in performance

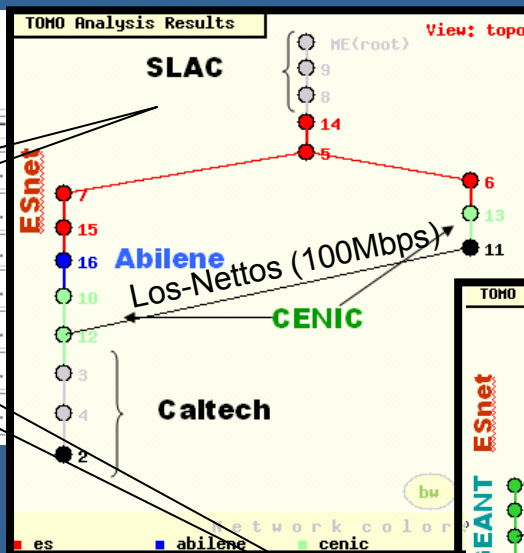
Traceroute Analysis for 10/09/2003

[Yesterday's Summary](#) | [Reverse Traceroute Summary](#) | [Directory of Historical Traceroutes](#)

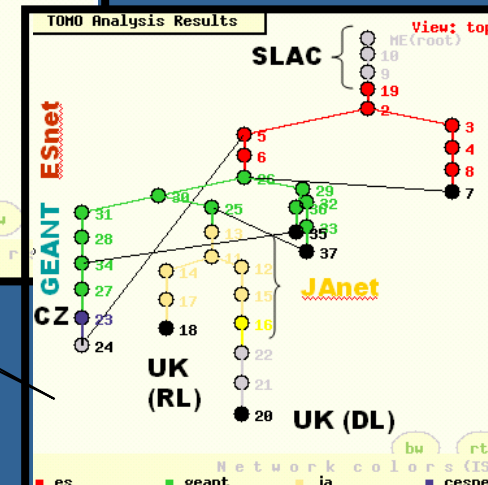
SUBMIT Topology request | RESET FIELDS

Hour (PST) →

Remote host	NODE \ Hour =>	00	01	02	14	15	16	17
<input checked="" type="checkbox"/> node1.cacr.caltech.edu* R Sum Log*	105	110	105
<input type="checkbox"/> node1.cesnet.cz* R Sum Log*	35	...	96 95	93
<input type="checkbox"/> node1.clrc.ac.uk* R Sum Log*	67	...	71 67	67
<input type="checkbox"/> node1.dl.ac.uk* R Sum Log*	97	...	102 97	97
<input type="checkbox"/> node1.ece.rice.edu* R Sum Log*	104
<input type="checkbox"/> node1.fnal.gov* R Sum Log*	8
<input type="checkbox"/> node1.in2p3.fr* R Sum Log*	29	...	72 100	82



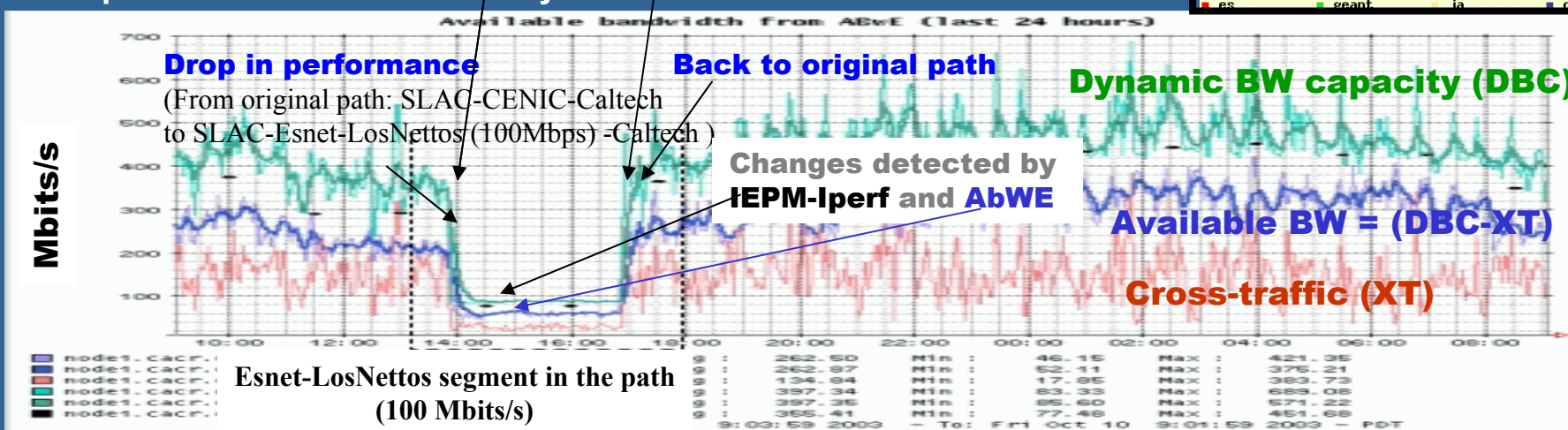
Samples of traceroute trees generated from the table



Snapshot of traceroute summary table

Notes:

1. Caltech misrouted via Los-Nettos 100Mbps commercial net 14:00-17:00
2. ESnet/GEANT working on routes from 2:00 to 14:00
3. A previous occurrence went un-noticed for 2 months
4. Next step is to auto detect and notify



ABwE measurement one/minute for 24 hours Thurs Oct 9 9:00am to Fri Oct 10 9:01am

Dedicated Optical Circuits

- Could be whole new playing field, today's tools no longer applicable:
 - No jitter (so packet pair dispersion no use)
 - Instrumented TCP stacks a la Web100 may not be relevant
 - Layer 1 switches make traceroute less useful
 - Losses so low, ping not viable to measure
 - High speeds make some current techniques fail or more difficult (timing, amounts of data etc.)

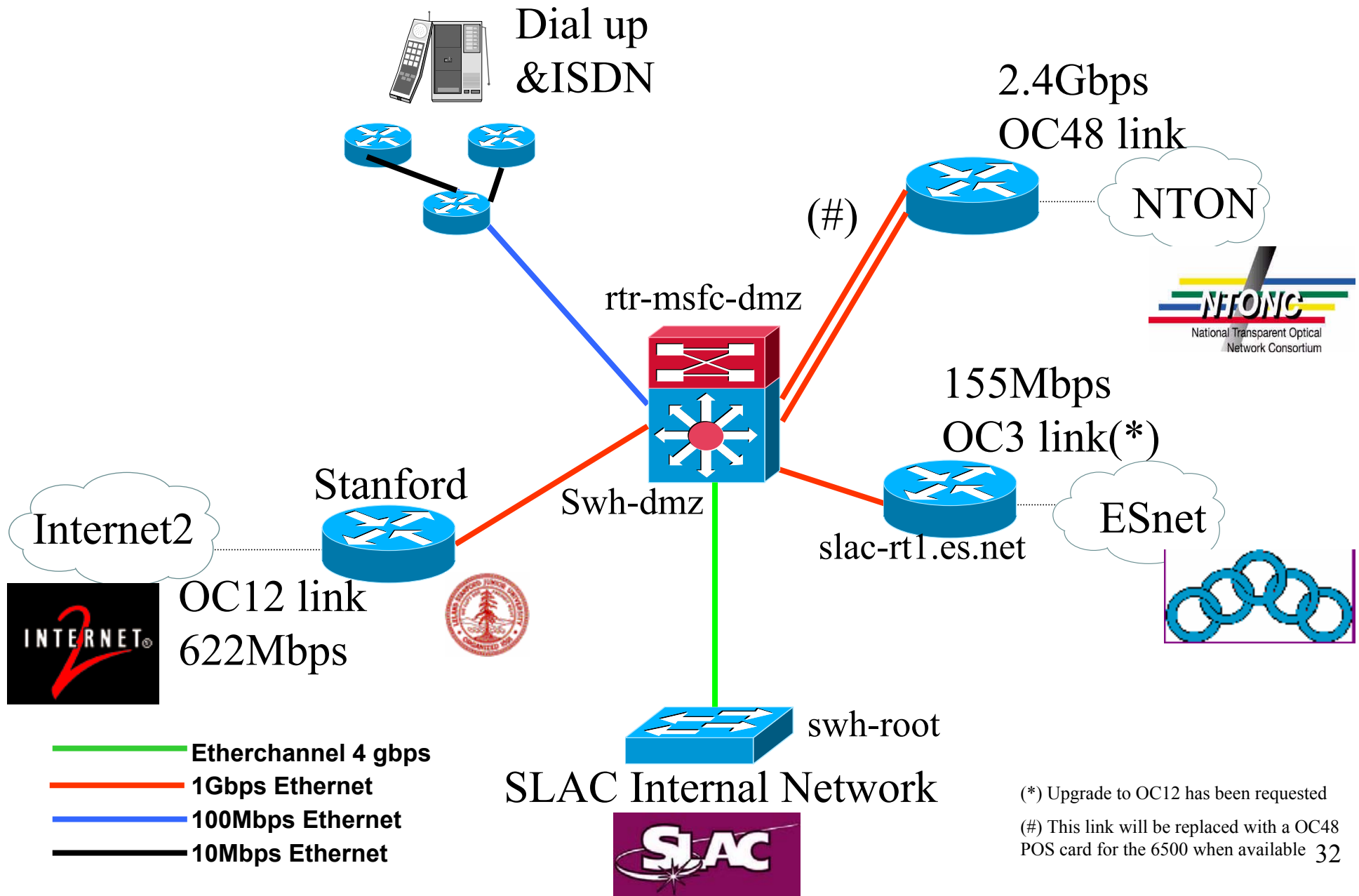
Future work

- Apply E2E anomaly detection to multiple metrics (RTT, available bandwidth, achievable throughput), multi-routes
- Apply forecasting & anomaly detection to passive data
 - If patterns stay stable for a long time (weeks)
 - Put together time series from multiple separate flows
 - Interpolate and use with Holt-Winters
- Detect not just size of bottleneck but location
 - Then can apply QoS just to poor link rather than whole path

More Information

- Tutorial on monitoring
 - www.slac.stanford.edu/comp/net/wan-mon/tutorial.html
- RFC 2151 on Internet tools
 - www.freesoft.org/CIE/RFC/Orig/rfc2151.txt
- Network monitoring tools
 - www.slac.stanford.edu/xorg/nmtf/nmtf-tools.html
- Ping
 - <http://www.ping127001.com/pingpage.htm>
- IEPM/PingER home site
 - www-iepm.slac.stanford.edu/
- IEEE Communications, May 2000, Vol 38, No 5, pp 130-136

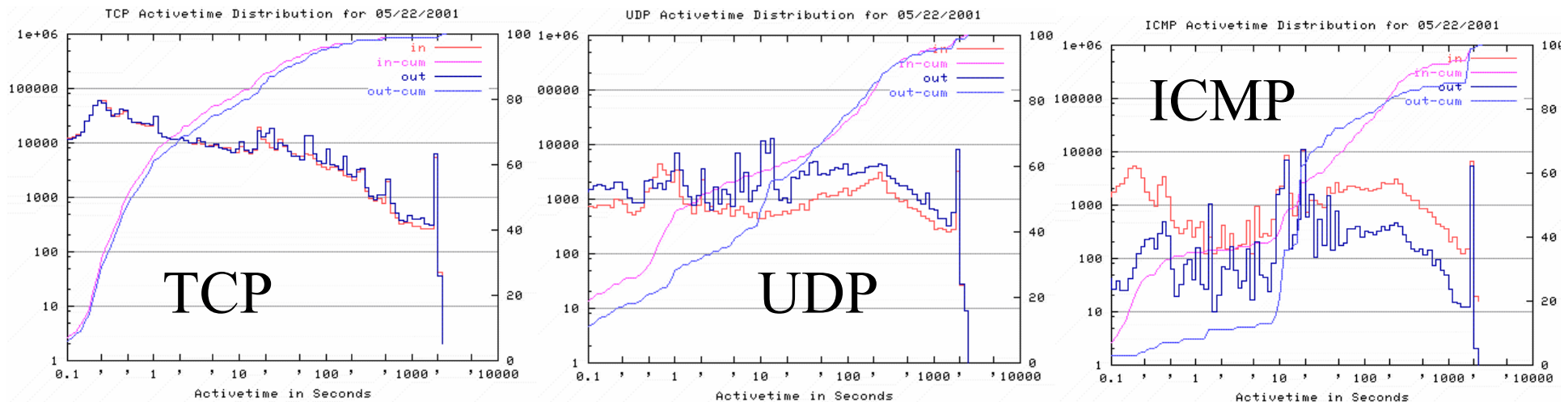
Simplified SLAC DMZ Network, 2001



(*) Upgrade to OC12 has been requested
 (#) This link will be replaced with a OC48 POS card for the 6500 when available 32

Flow lengths

- Distribution of netflow lengths for SLAC border
 - Log-log plots, linear trendline = power law
 - Netflow ties off flows after 30 minutes
 - TCP, UDP & ICMP “flows” are \sim log-log linear for longer (hundreds to 1500 seconds) flows (heavy-tails)
 - There are some peaks in TCP distributions, timeouts?
 - Web server CGI script timeouts (300s), TCP connection establishment (default 75s), TIME_WAIT (default 240s), tcp_fin_wait (default 675s)



Traceroute technical details

Rough traceroute algorithm

```
ttl=1; #To 1st router
```

```
port=33434; #Starting UDP port
```

```
while we haven't got UDP port unreachable {
```

```
    send UDP packet to host:port with ttl
```

```
    get response
```

```
        if time exceeded note roundtrip time
```

```
    else if UDP port unreachable
```

```
        quit
```

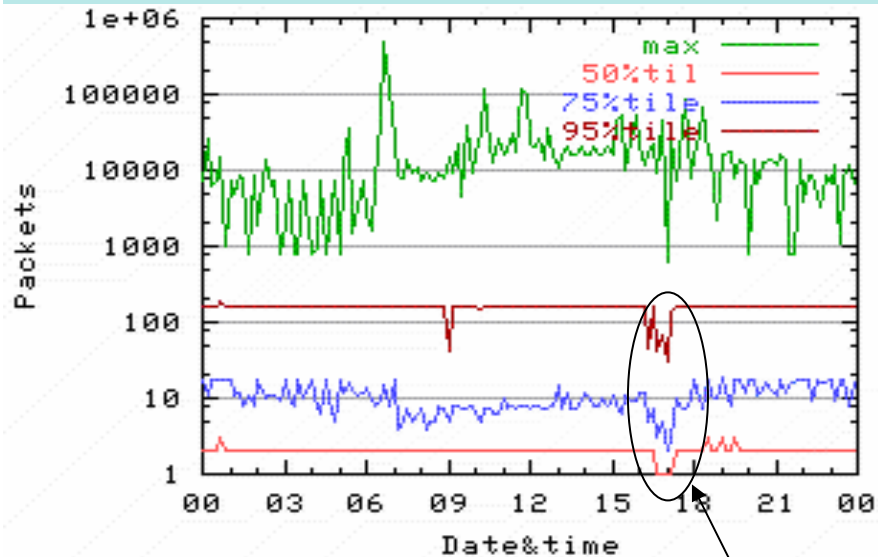
```
    print output
```

```
    ttl++; port++
```

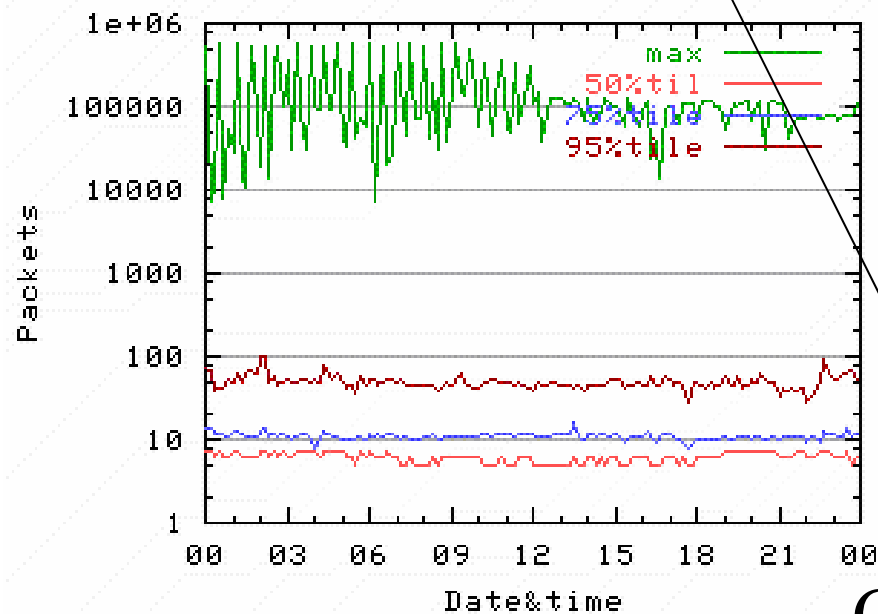
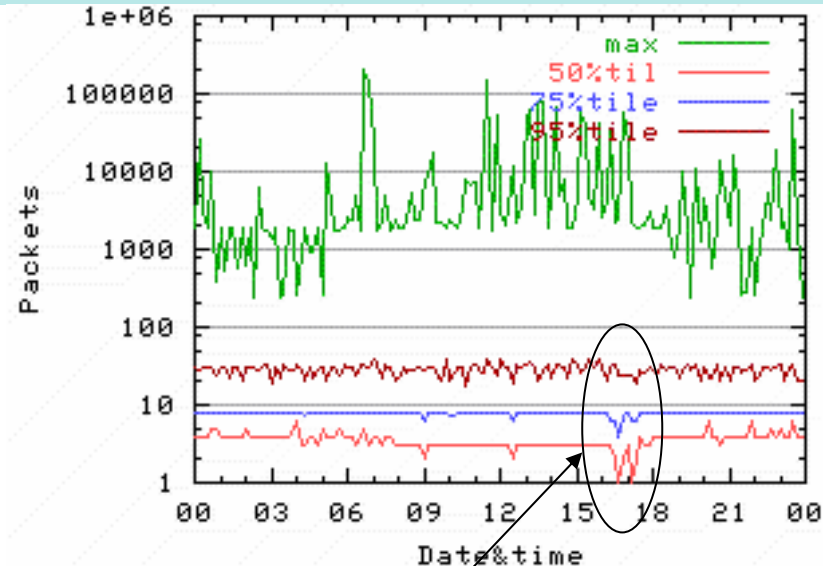
```
}
```

- Can appear as a port scan
 - SLAC gets about one complaint every 2 weeks.

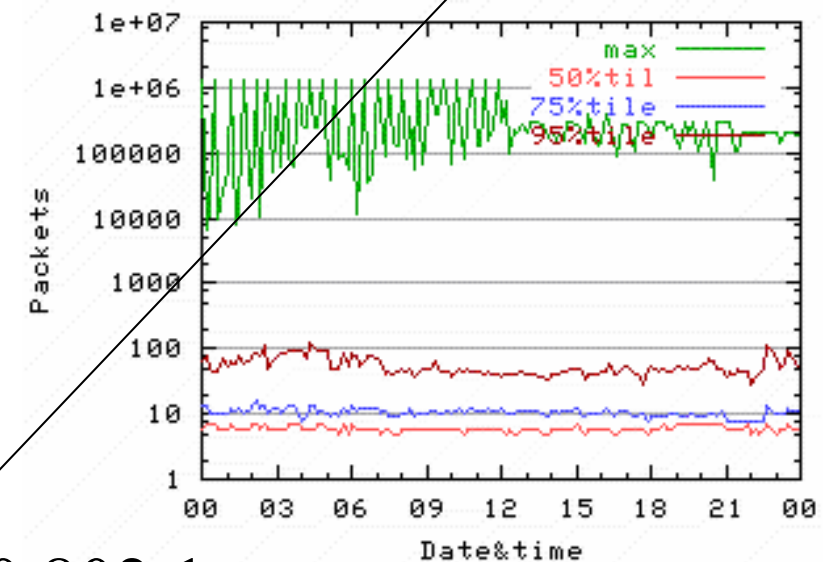
Time series



UDP



TCP



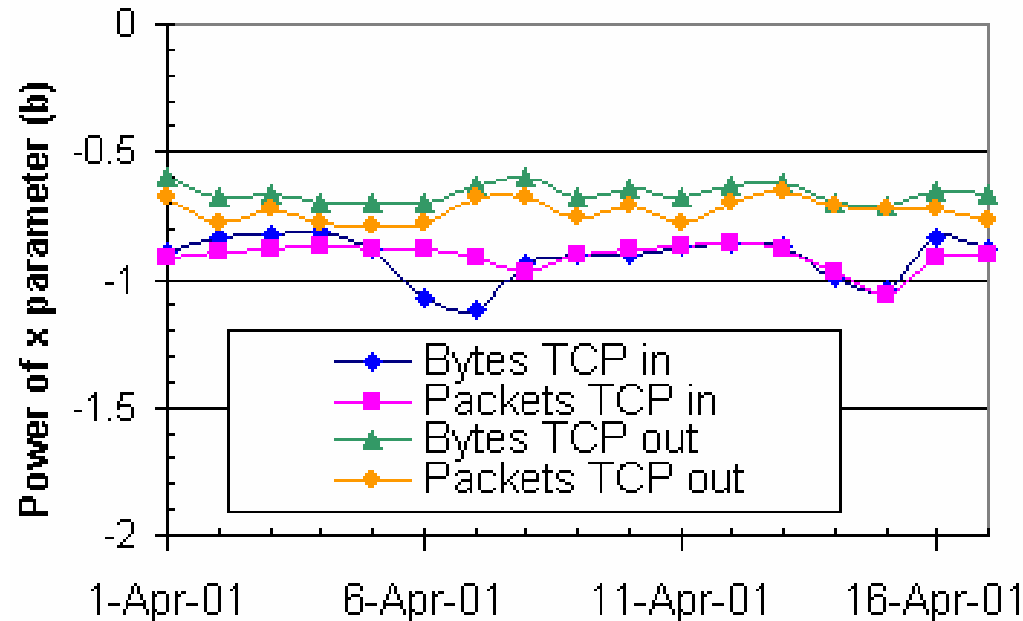
Outgoing

Cat 4000 802.1q
vs. ISL

Incoming

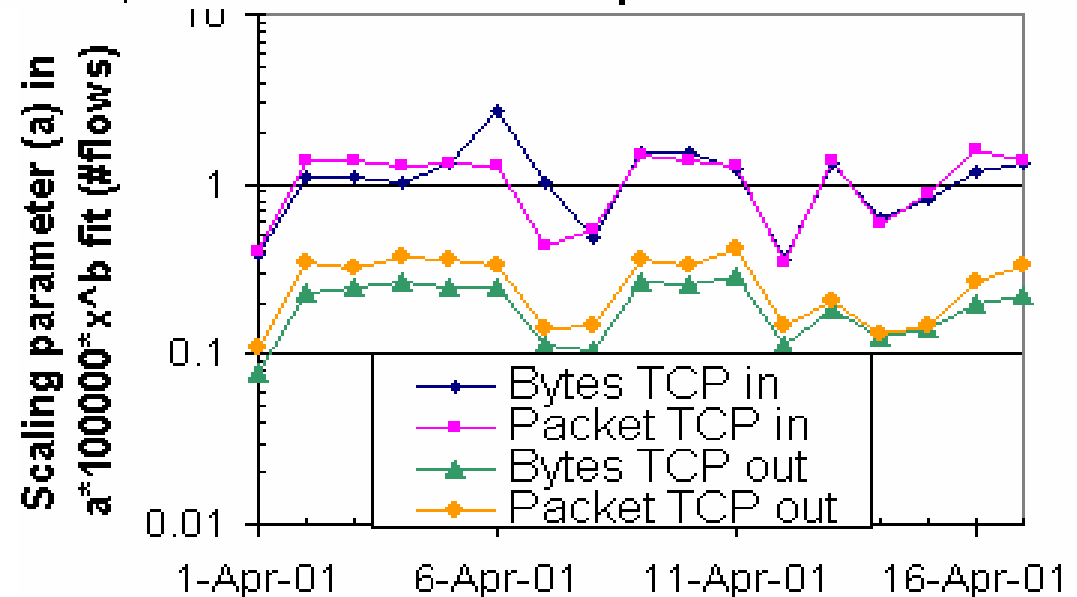
Power law fit parameters by time

Slope of power law fit to Flow frequencies



Just 2 parameters provide a reasonable description of the flow size distributions

Scaling parameter for power law fit to Flow frequencies

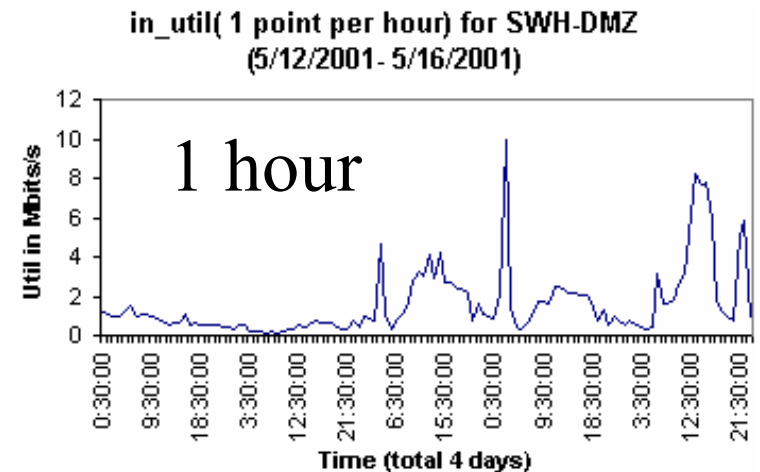
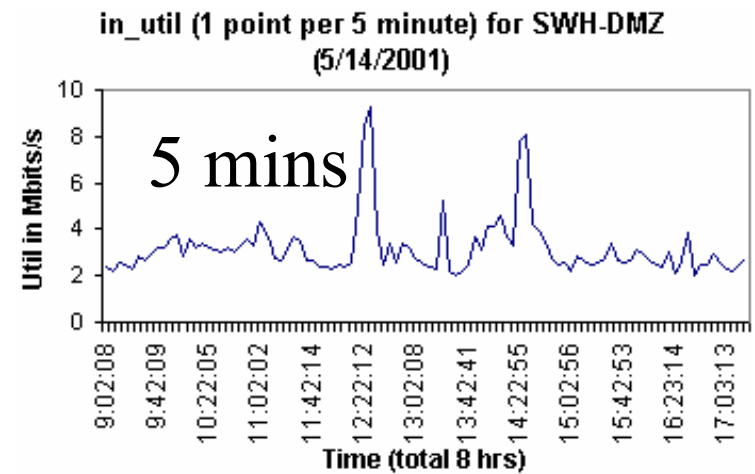
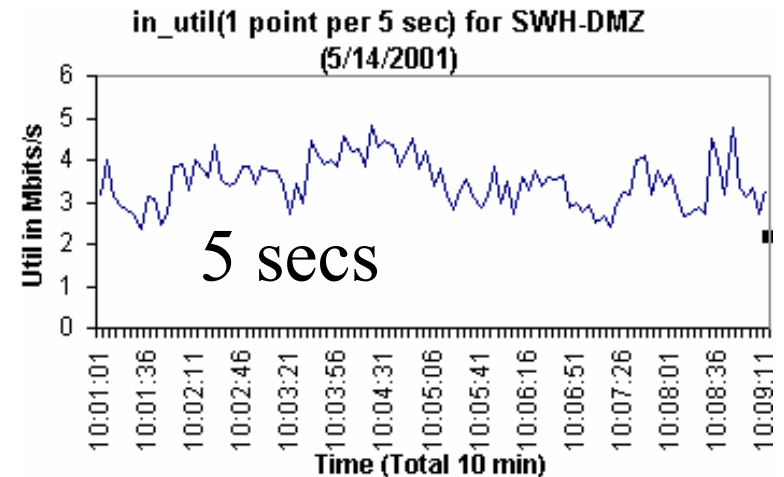


Averaging/Sampling intervals

- Typical measurements of utilization are made for 5 minute intervals or longer in order not to create much impact.
- Interactive human interactions require second or sub-second response
- So it is interesting to see the difference between measurement made with different time frames.

Utilization with different averaging times

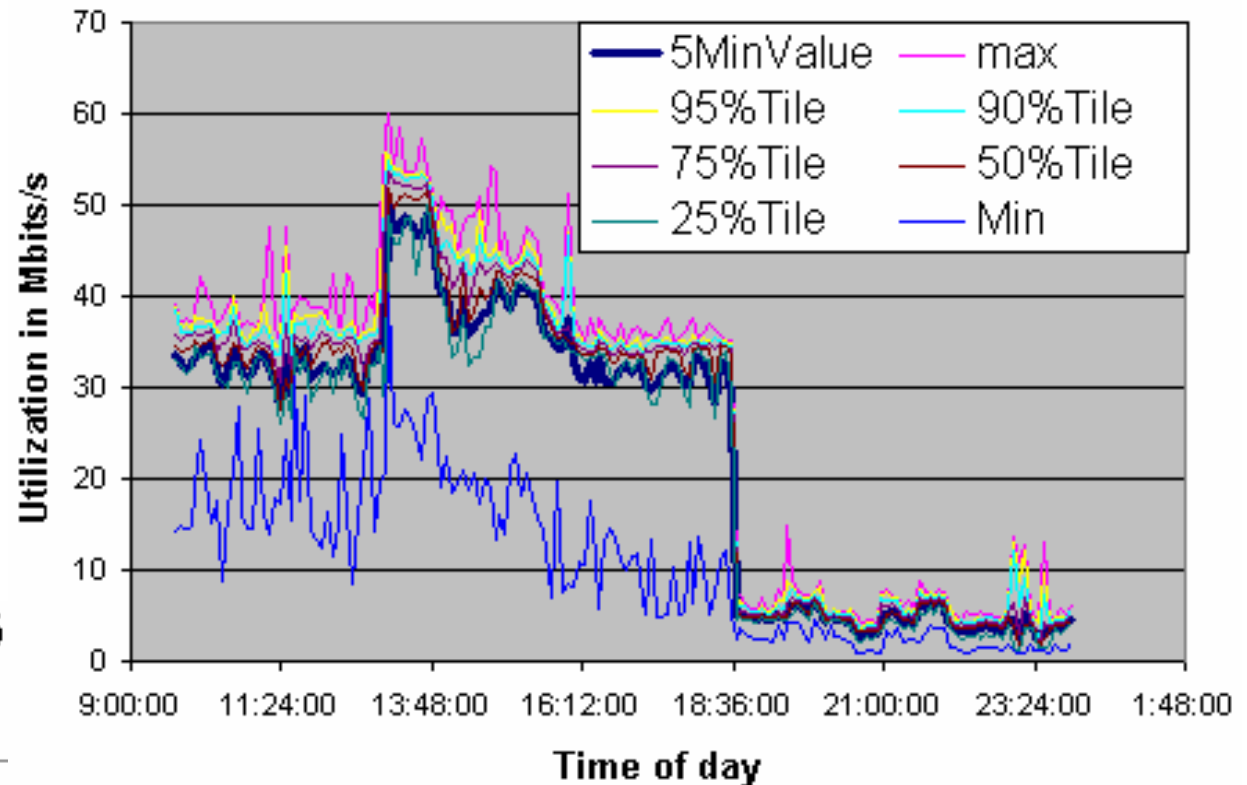
- Same data, measured Mbits/s every 5 secs
- Average over different time intervals
- Does not get a lot smoother
- May indicate multi-fractal behavior



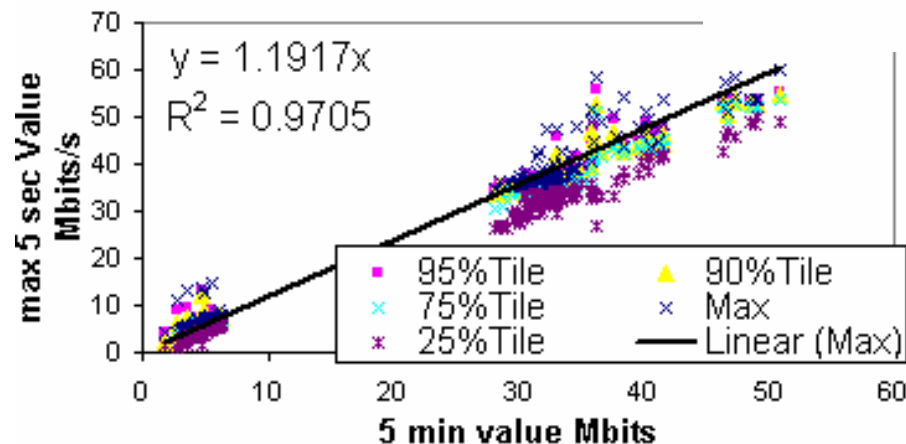
Lot of heavy FTP activity

- The difference depends on traffic type
- Only 20% difference in max & average

Trendlines for 5min & 5 sec out utilizations at SLAC border for May 14, 2001

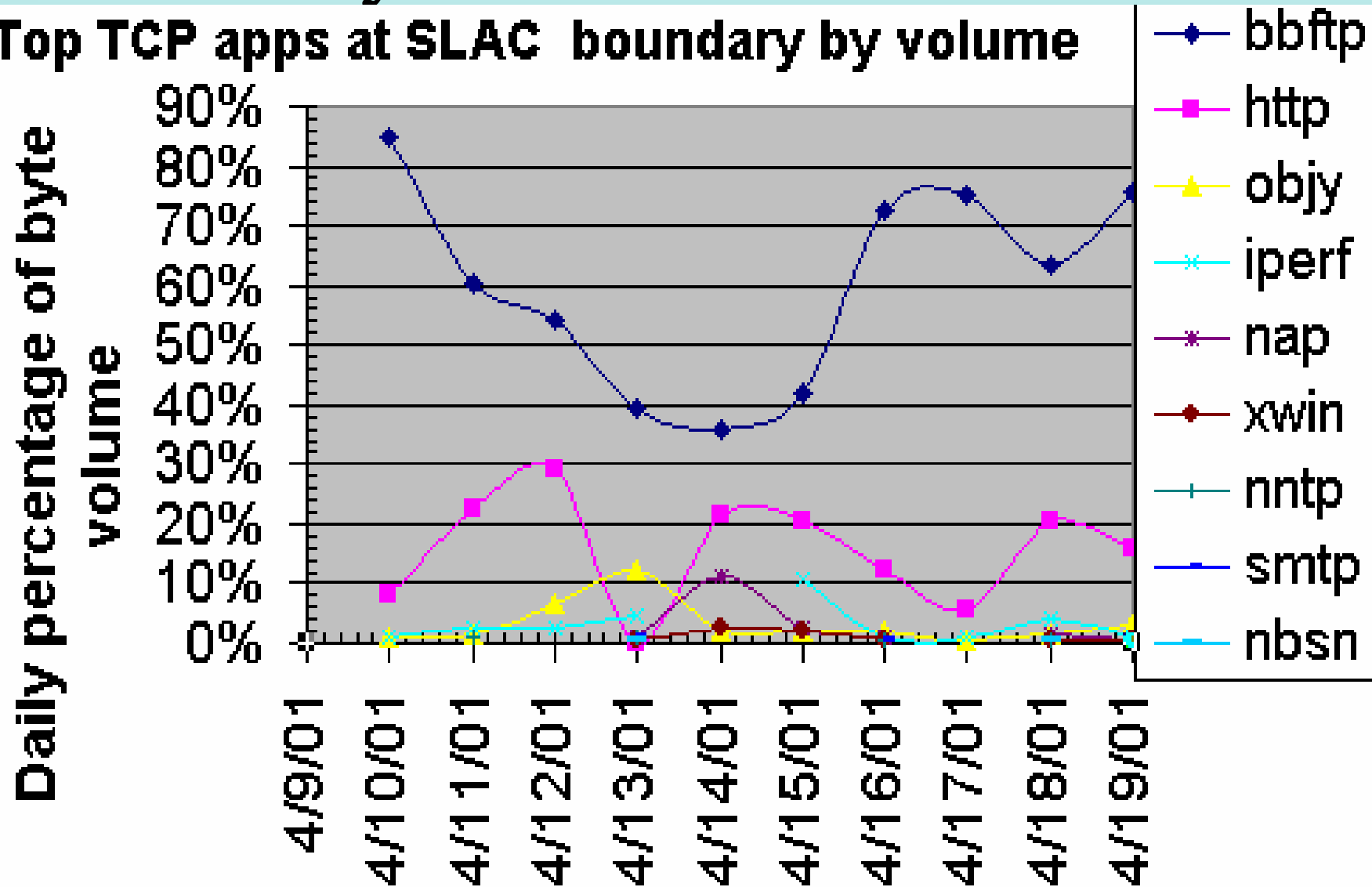


In Utilization for 5 min & 5 intervals 5-22-2001



Not your normal Internet site

Top TCP apps at SLAC boundary by volume

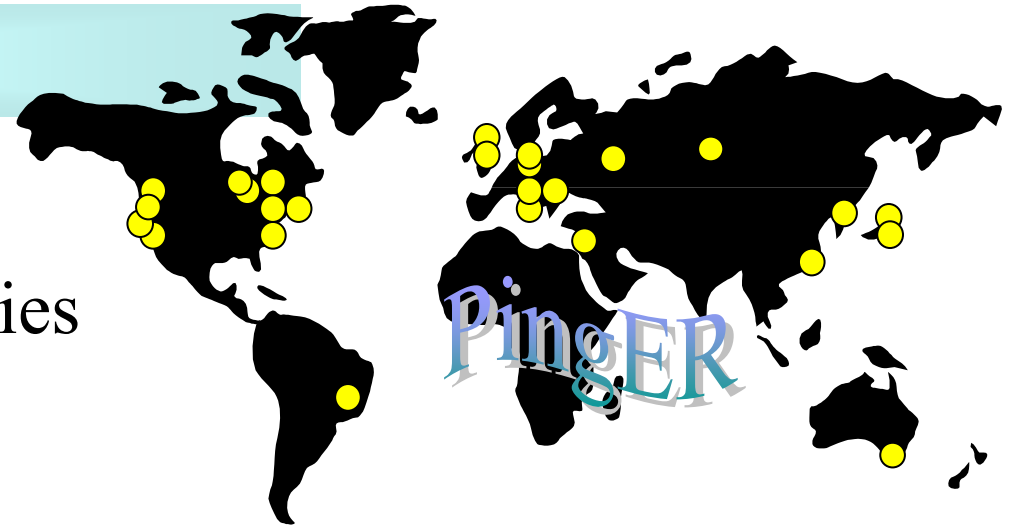


Ames IXP: approximately 60-65% was HTTP, about 13% was NNTP
Uwisc: 34% HTTP, 24% FTP, 13% Napster

PingER cont.

- Monitor timestamps and sends ping to remote site at regular intervals (typically about every 30 minutes)
- Remote site echoes the ping back
- Monitor notes current and send time and gets RTT
- Discussing installing monitor site in Pakistan
 - provide real experience of using techniques
 - get real measurements to set expectations, identify problem areas, make recommendations
 - provide access to data for developing new analysis techniques, for statisticians etc.

PingER

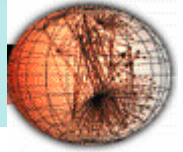


- Measurements from
 - 38 monitors in 14 countries
 - Over 600 remote hosts
 - Over 120 countries
 - Over 3300 monitor-remote site pairs
 - Measurements go back to Jan-95
 - Reports on RTT, loss, reachability, jitter, reorders, duplicates ...
- Uses ubiquitous “ping” facility of TCP/IP
- Countries monitored
 - Contain over 80% of world population
 - 99% of online users of Internet

Surveyor & RIPE, NIMI

- Surveyor & RIPE use dedicated PCs with GPS clocks for synchronization
 - Measure 1 way delays and losses
 - Surveyor mainly for Internet 2
 - RIPE mainly for European ISPs
- NIMI (National Internet Measurement Infrastructure) more of an infrastructure for measurements and some tools (I.e. currently does not have public available data, regularly updated)
 - Mainly full mesh measurements on demand

Skitter



- Makes ping & route measurements to tens of thousands of sites around the world. Site selection varies based on web site hits.
 - Provide loss & RTTs
 - Skitter & PingER are main 2 sites to monitor developing world.

“Where is” a host – cont.

- Find the Autonomous System (AS) administering
 - Use reverse traceroute server with AS identification, e.g.:
 - www.slac.stanford.edu/cgi-bin/nph-traceroute.pl
 - ...
 - 14 lhr.comsats.net.pk (210.56.16.10) [AS7590 - COMSATS] 711 ms (ttl=242)
 - Get contacts for ISPs (if know ISP or AS):
 - <http://puck.nether.net/netops/nocs.cgi>
 - Gives ISP name, web page, phone number, email, hours etc.
 - Review list of AS's ordered by Upstream AS Adjacency
 - www.telstra.net/ops/bgp/bgp-as-upsstm.txt
 - Tells what AS is upstream of an ISP
 - Look at real-time information about the global routing system from the perspectives of several different locations around the Internet
 - Use route views at www.antc.uoregon.edu/route-views/
- Triangulate RTT measurements to unknown host from multiple places

Who do you tell

- Local network support people
- Internet Service Provider (ISP) usually done by local networker
 - Use puck.nether.net/netops/nocs.cgi to find ISP
 - Use www.telstra.net/ops/bgp/bgp-as-upsstm.txt to find upstream ISPs
- Give them the ping and traceroute results

Achieving throughput

- User can't achieve throughput available (Wizard gap)
- Big step just to know what is achievable

TCP Achievable Bandwidth

