Passive and Active Monitoring on a High-performance Network

Les Cottrell, Warren Matthews, Davide Salomoni, Connie Logg *– SLAC*

www.slac.stanford.edu/grp/scs/net/talk/pam-apr01/

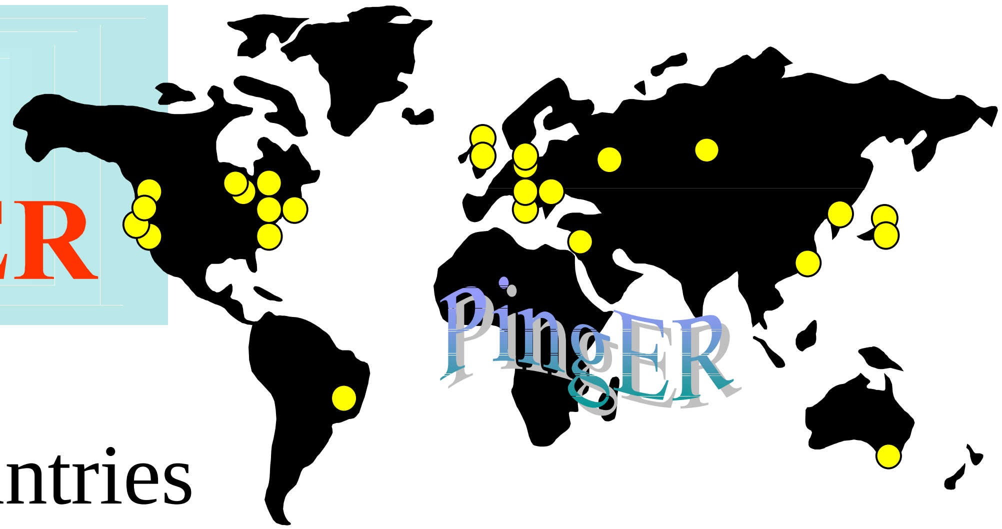
Presented at PAM-2001, Amsterdam April 23-24, 2001

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

Outline

- Results from active monitoring with PingER:
  - RTT, Loss, “jitter”
- Passive border monitoring results
- High perf throughput
  - achieving, measuring and impact
- Simulation of high perf throughput

Active WAN Monitoring/PingER

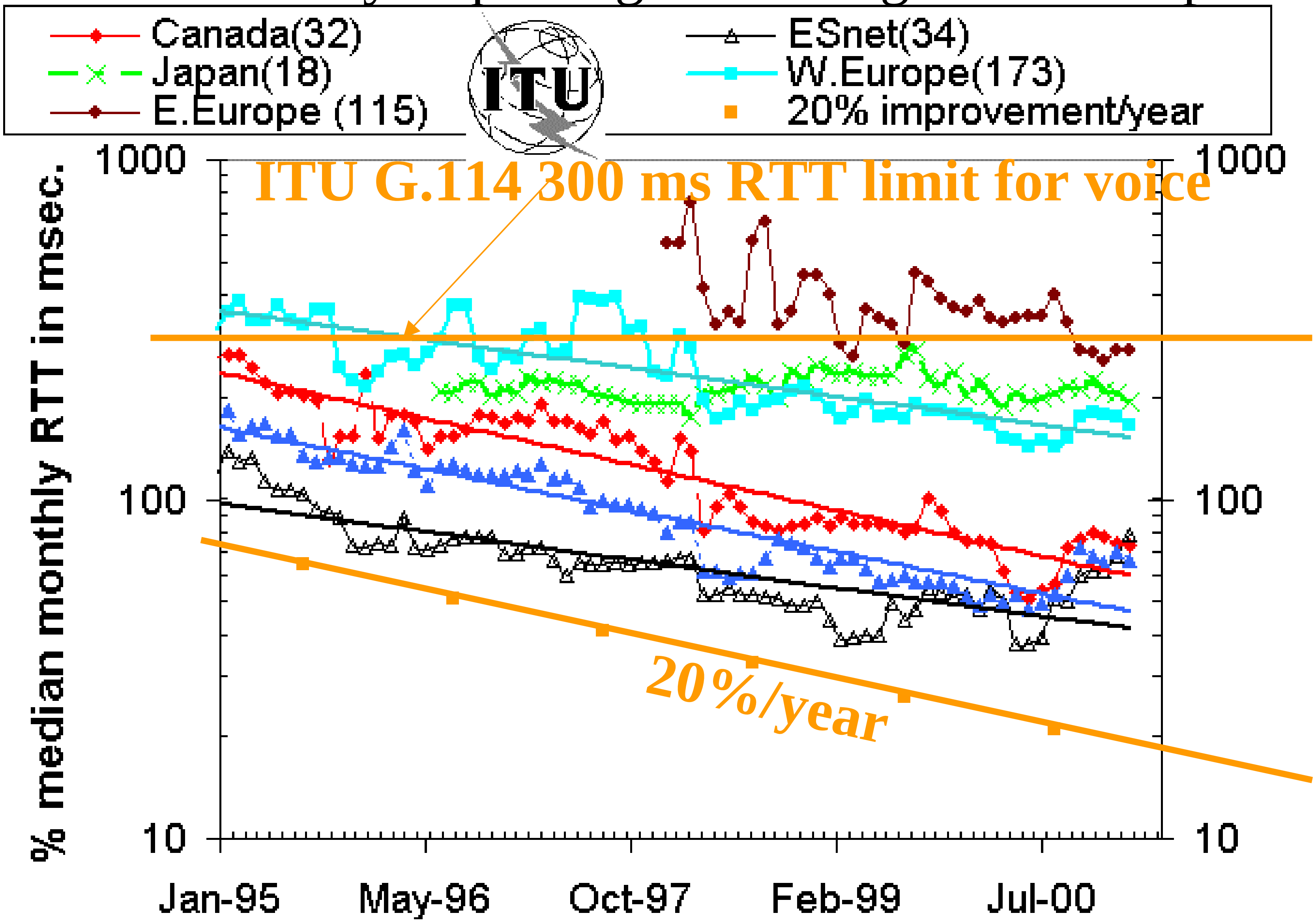


- Measurements from
  - 32 monitors in 14 countries
  - Over 600 remote hosts in over 72 countries
  - Over 3300 monitor-remote site pairs
  - Measurements go back to Jan-95
  - Reports on RTT, loss, reachability, IPDV, throughput, reordering, duplicates, looking at CLP (for bursty losses)…
- Uses ubiquitous "ping" facility of TCP/IP
- Countries monitored
  - Contain 78% of world population
  - 99% of online users of Internet

RTT from ESnet to Groups of Sites

RTT ~ distance/(0.6*c) + hops * router delay

Router delay = queuing + clocking in & out + processing



RTT Region to Region

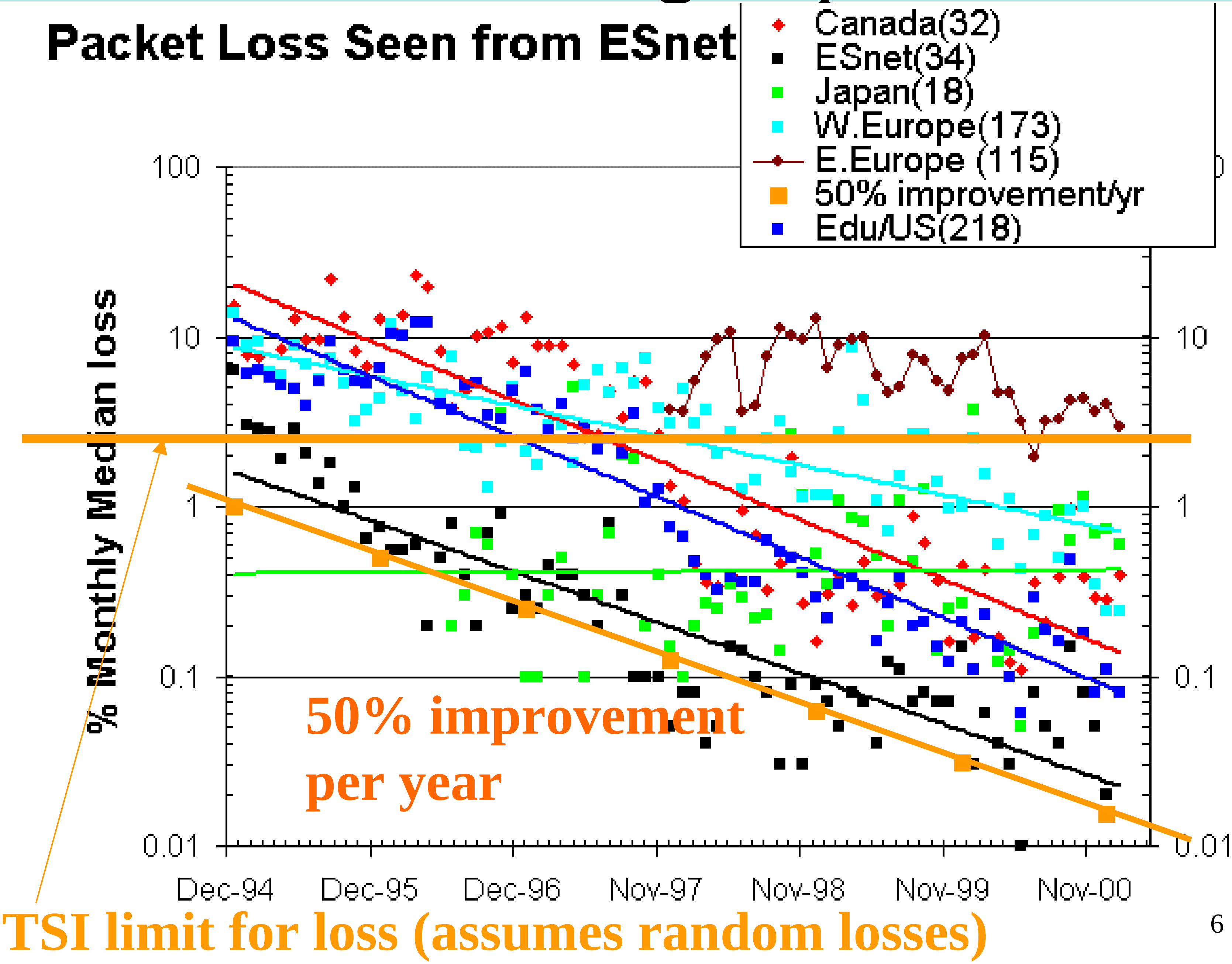
OK
White 0-64ms
Green 64-128ms
Yellow 128-256ms

NOT OK
Pink 256-512ms
Red > 512ms

WORLD	Australasia	East Europe	North America	West Europe	South America	Asia
Australasia	3.95	714.74	300.68	454.68	389.69	373.59
East Europe		359.03	235.66	87.37	278.01	319.64
North America		244.24	69.44	153.23	223.06	203.83
West Europe		385.14	163.32	42.97	260.47	290.68
South America		626.39	421.45	590.69	18.93	780.00
Asia		472.57	327.85	321.99	447.02	24.00
Africa		770.90	804.67	804.15		
Aria			772.00	416.88		
null			501.30			
Middle East			1108.97			
Central America			436.00			

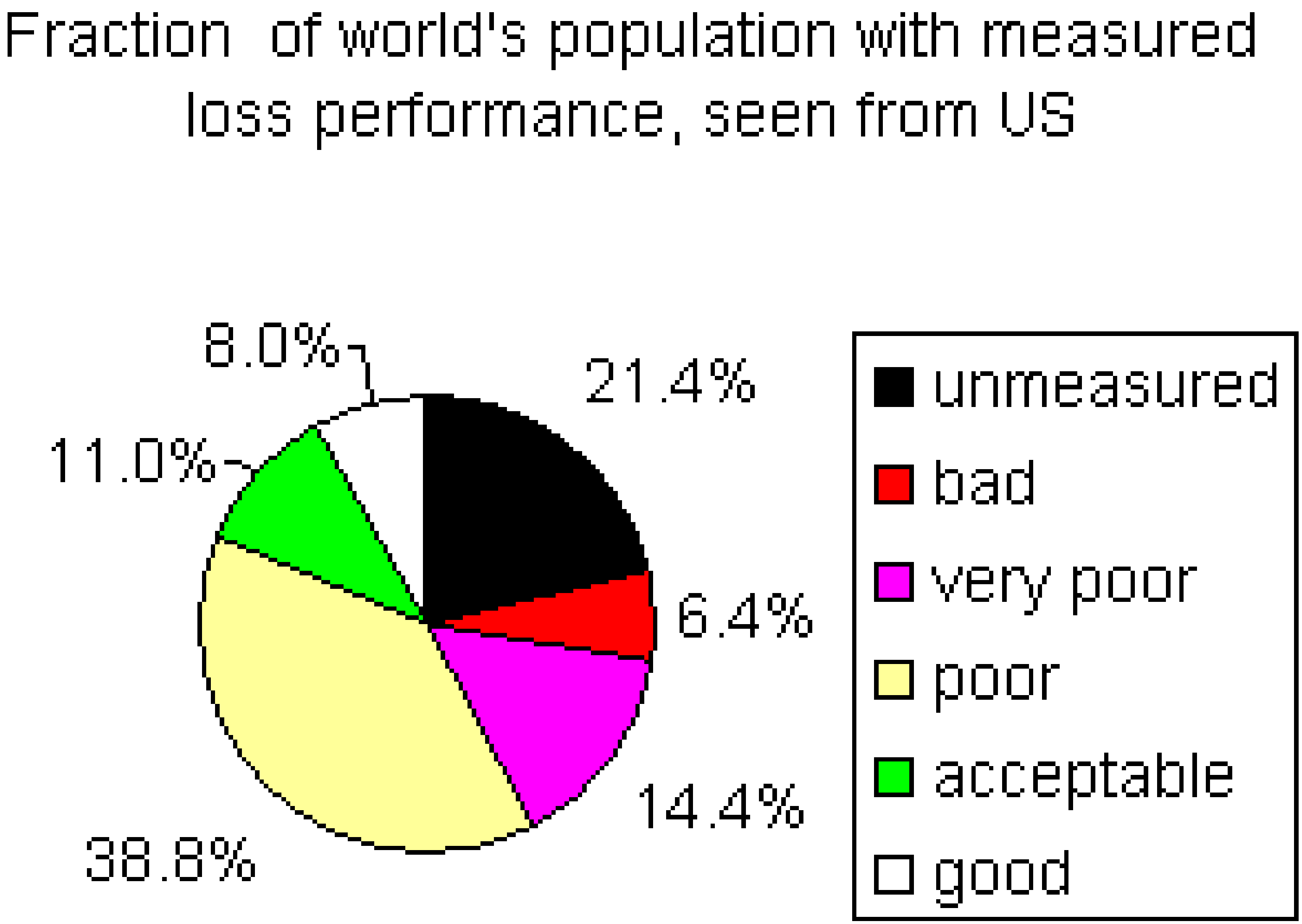
OK within regions, N. America OK with Europe, Japan

Loss seen from US to groups of Sites

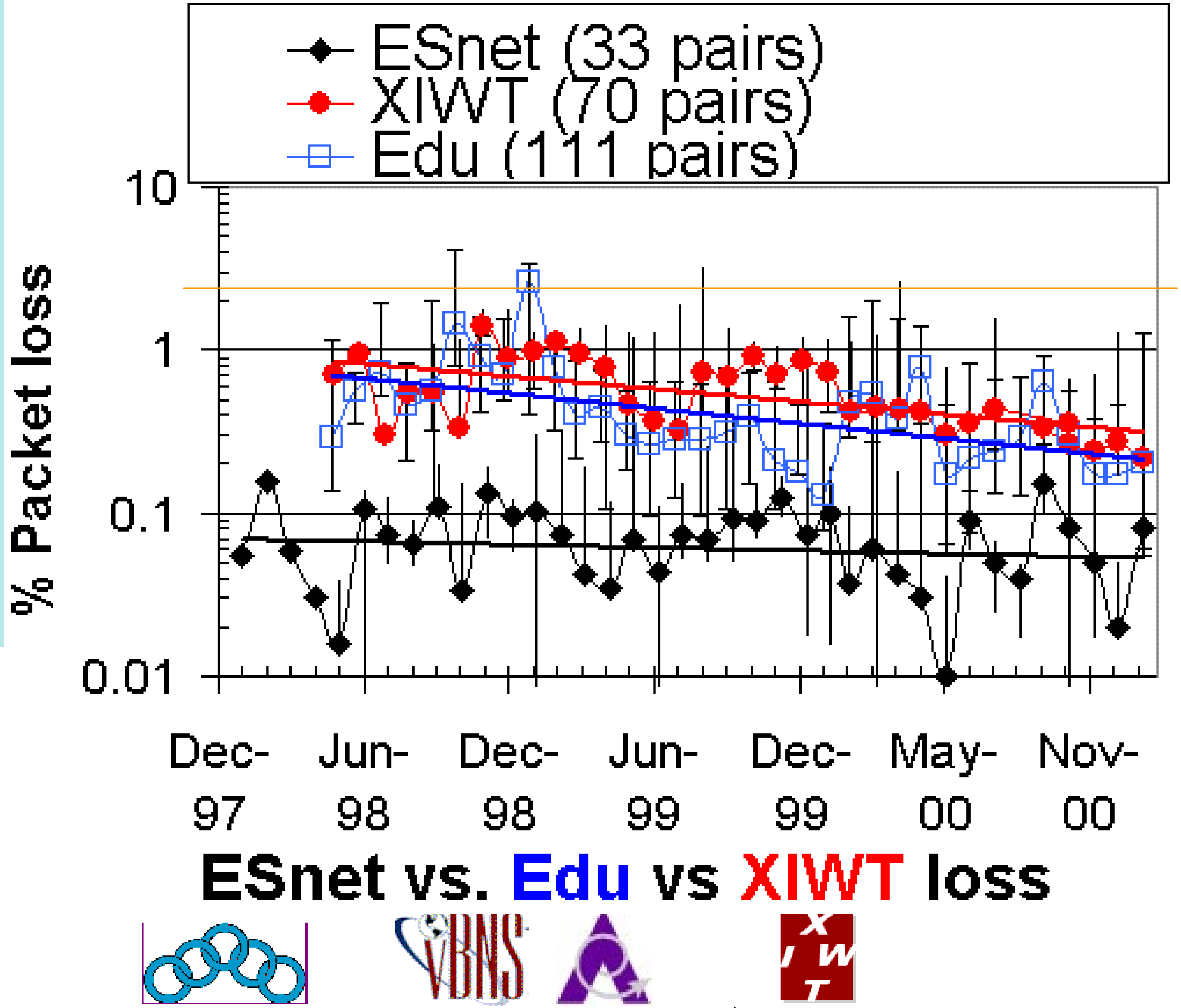


Loss to world from US

Using year 2000, fraction of world's population/country from www.nua.ie/surveys/how_many_online/



Losses within US for various Nets



In general performance is good (i.e. <= 1%)

ESnet holding steady

Edu (vBNS/Abilene) & XIWT (70% .com) improving,

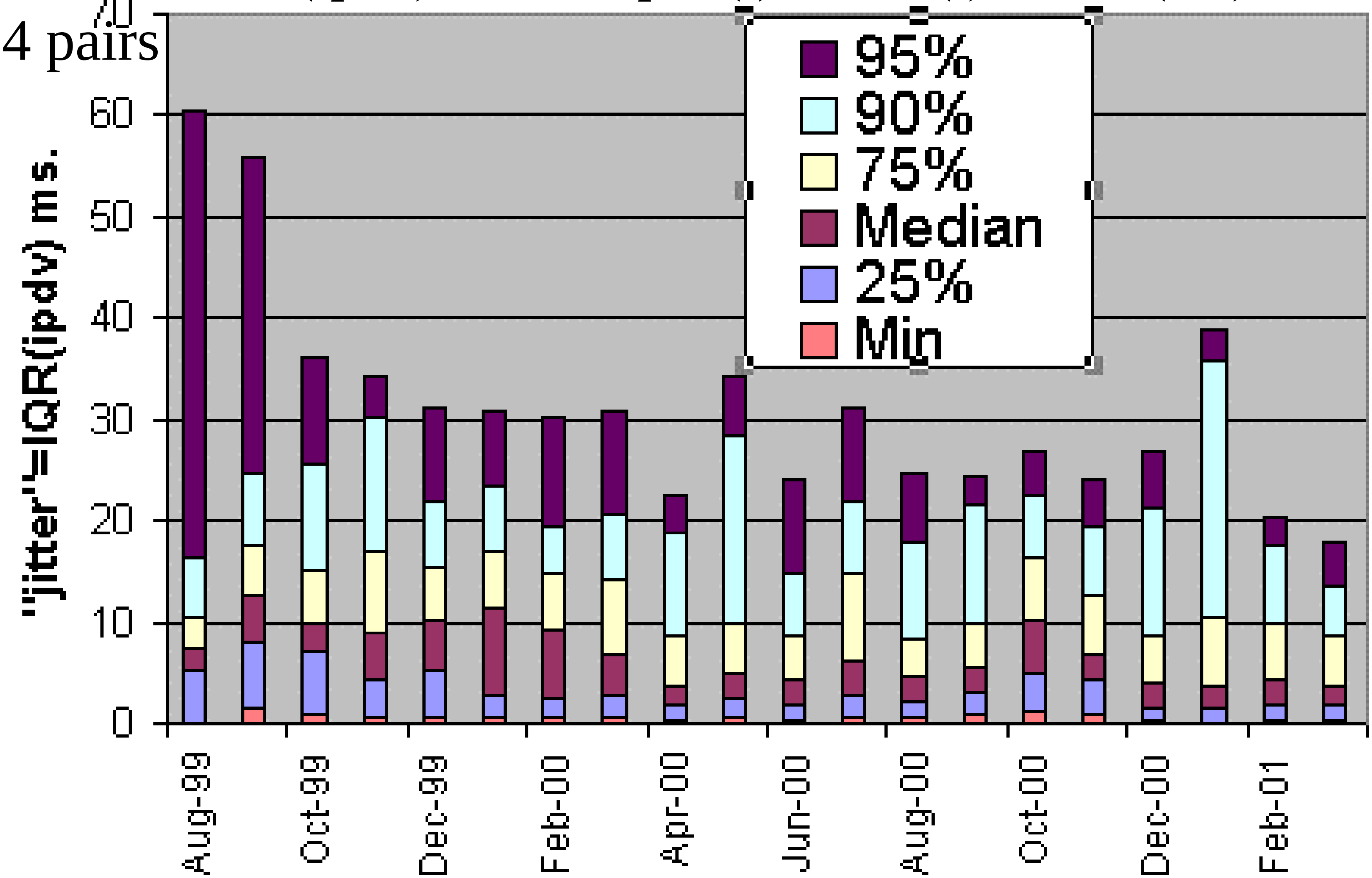
Losses between Regions

WORLD	Australasia	East Europe	North America	West Europe	South America	Asia
Australasia	0.10	6.91	0.81	1.12	0.31	1.65
East Europe		5.69	2.66	1.34	1.91	2.63
North America		6.39	0.14	0.23	0.02	0.36
West Europe		4.64	0.28	0.30	-0.20	0.43
South America		15.18	4.13	3.11	-1.37	5.02
Asia		5.93	1.68	1.30	0.00	0.32
Africa		32.73	2.10	5.58		
Aria			2.50	5.00		
null			1.12			
Middle East			3.71			
Central America			1.64			

"Jitter" from N. America to W. Europe

"Jitter" = IQR(ipdv), where *ipdv(i) =RTT(i) – RTT(i-1)*

214 pairs



ETSI: DTR/TIPHON-05001 V1.2.5 (1998-09) good speech < 75ms jitter

"Jitter" between regions

Median	US	W. Europe	E. Europe
US	2	4	15
W. Europe	4	5	22
E. Europe	18	11	29
Asia	18	14	46
S. America	20	17	38
Australasia	4	4	28
Africa	28		

90th%	US	W. Europe	E. Europe
US	9	14	25
W. Europe	13	16	39
E. Europe	134	38	130
Asia	54	292	56
S. America	81	32	53
Australasia	50	15	
Africa	171	178	265

Median	ESnet	Edu	Com
ESnet	2	6	
Edu	2	1	
Com			5

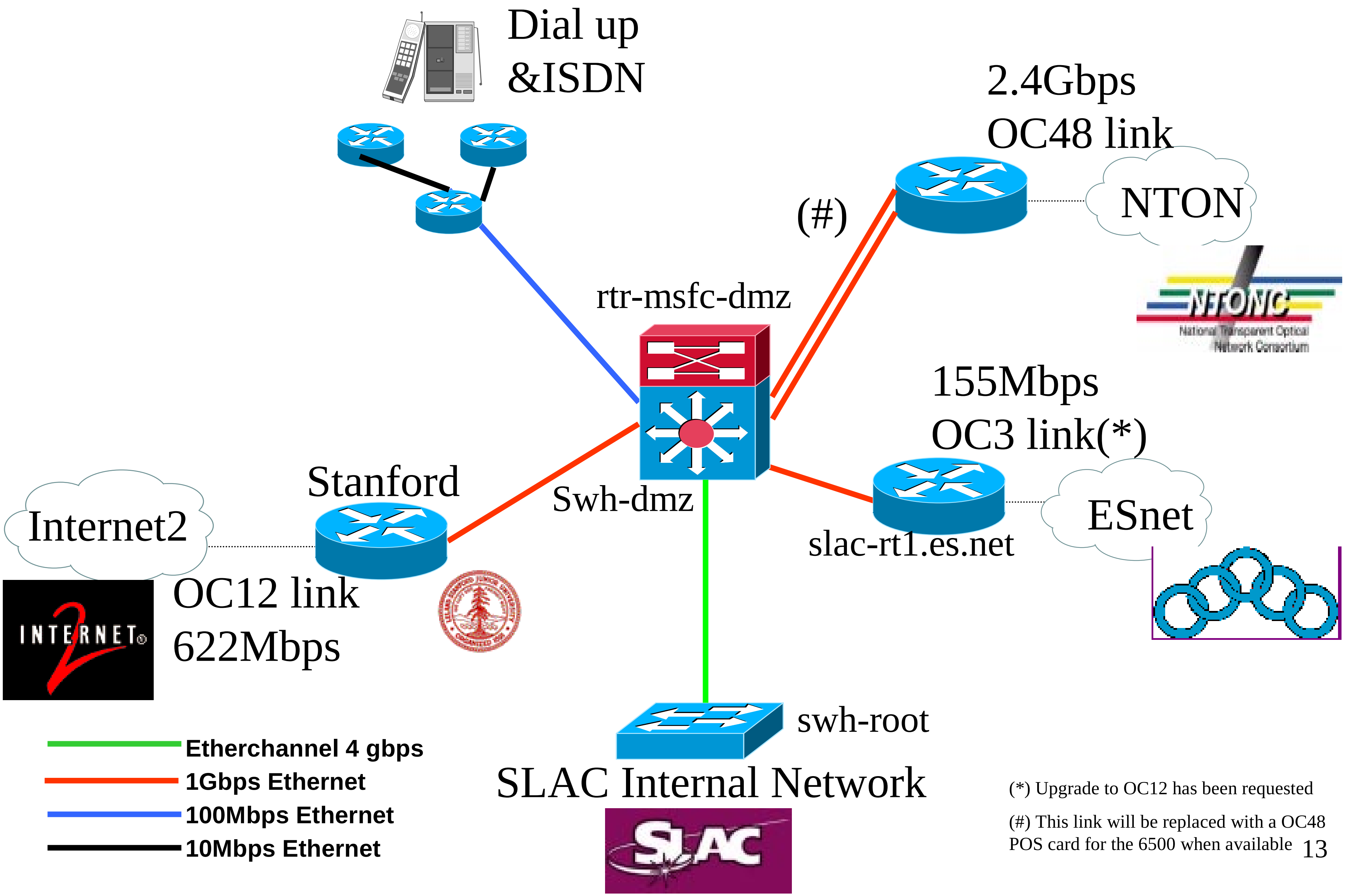
ETSI: DTR/TIPHON-05001 V1.2.5 (1998-09)

75ms=Good 125ms=Med 225ms=Poor

Passive site border monitoring

- Use SNMP to get utilization etc.
- Used to use OC3Mon with CoralReef for flows etc. but now have GigE interfaces
- Use **Cisco Netflow** in Catalyst 6509 with MSFC, only on border at the moment
- Gather about 200MBytes/day of flow data
- Date recorded in binary every 10 minutes into RRD
- 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 OC3Mon, but good compromise
  - Top talkers history and daily (from & to), tlds, vlans, protocol and application utilization, flow times, time series, distributions
- Use for network & security

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

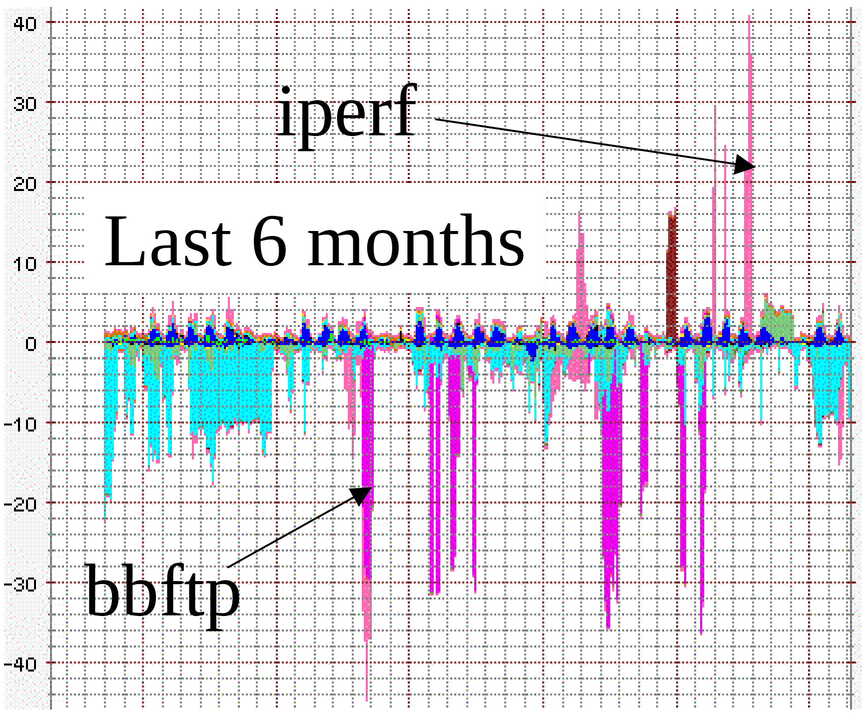
SLAC Traffic profile

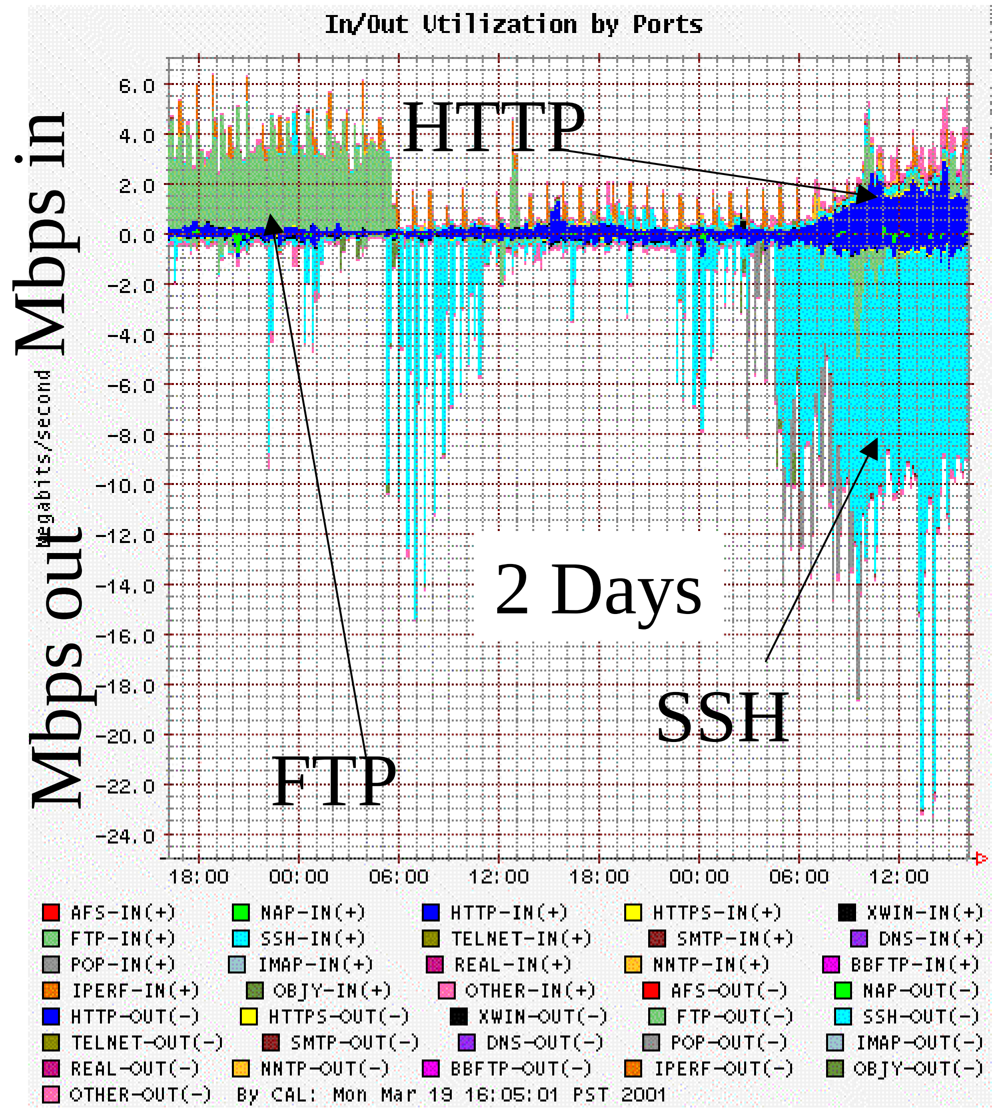
SLAC offsite links:

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

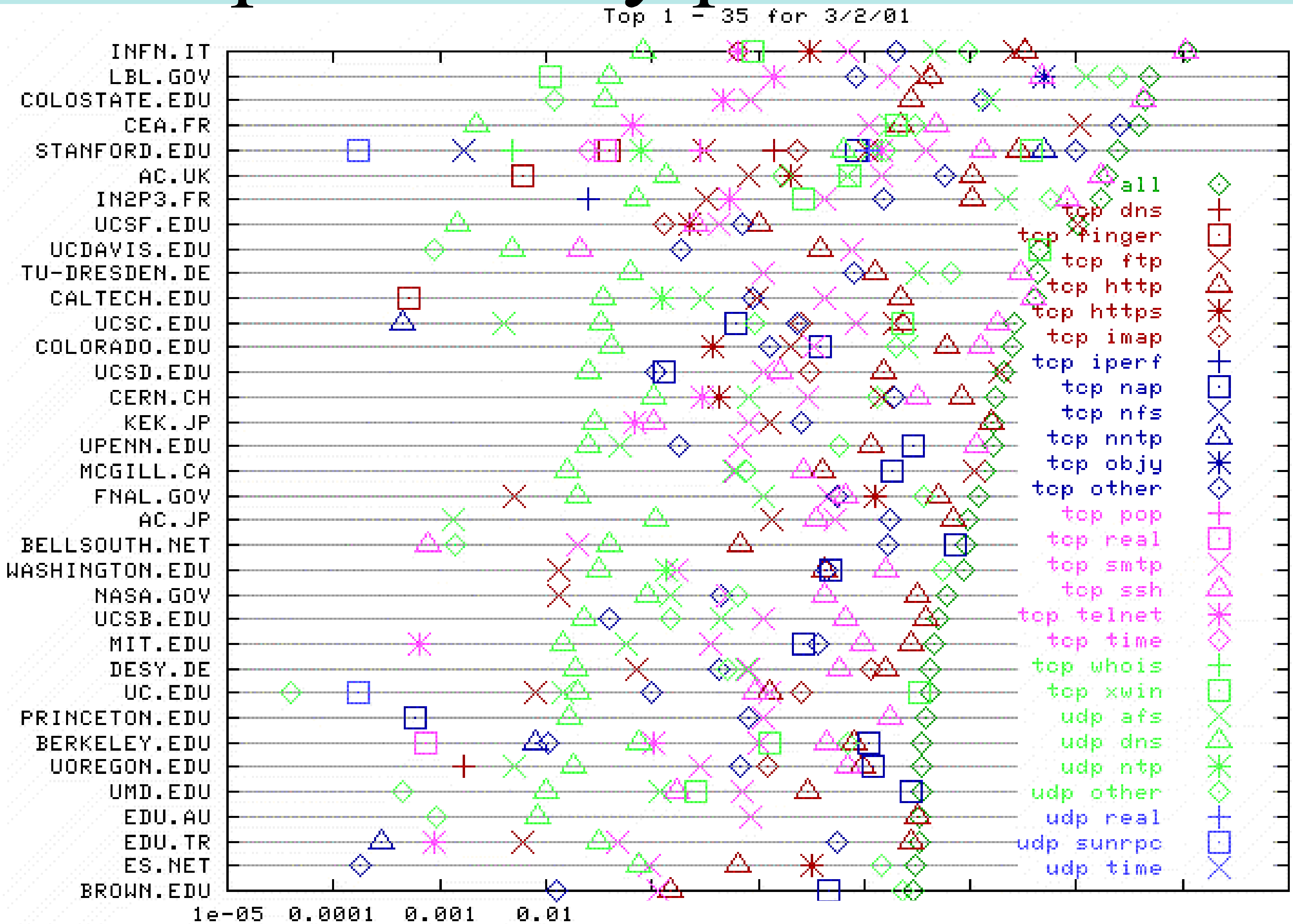
Profile

- bulk-data xfer dominates





Top talkers by protocol



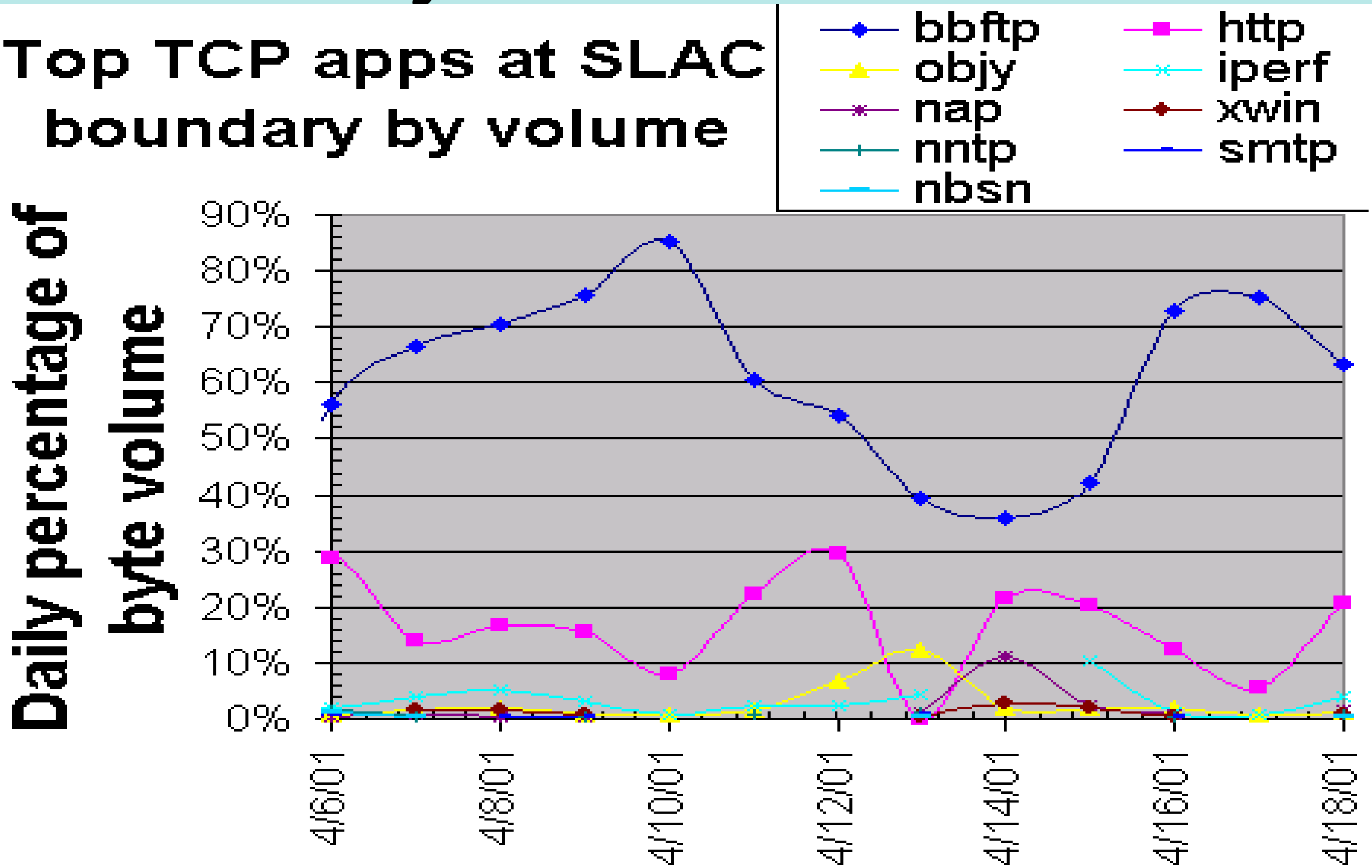
Hostname

1 100 10000

MBytes/day (log scale)

Volume dominated by single Application - bbftp

Not your normal Internet site

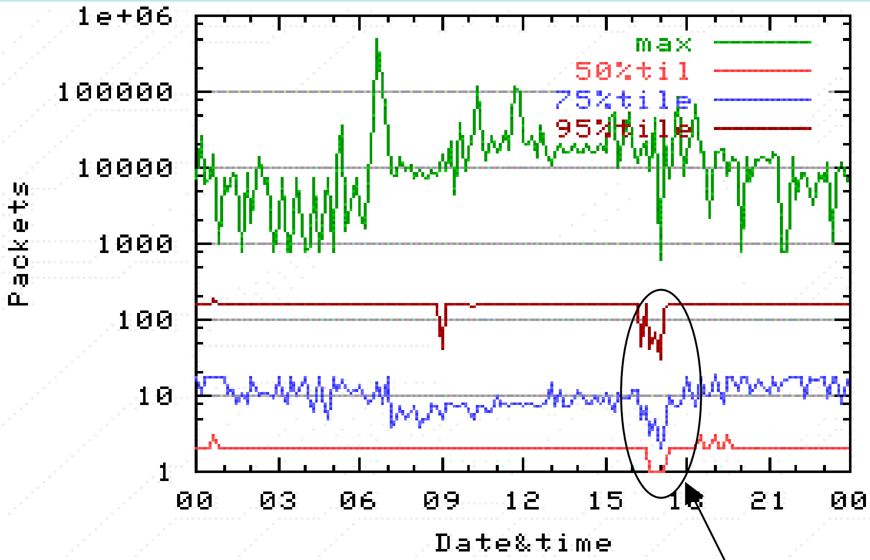


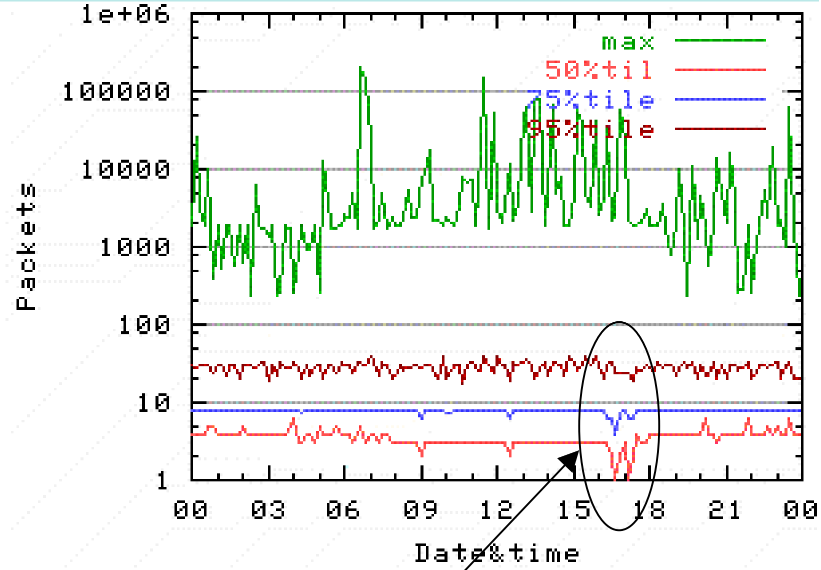
50-300Gbytes/day

Ames IXP: approximately 60-65% was HTTP, about 13% was NNTP

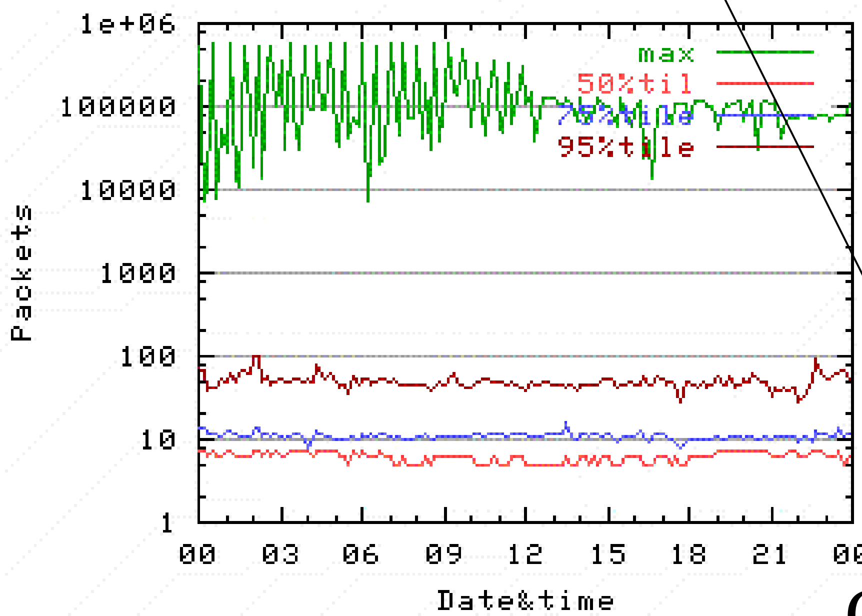
Uwisc: 34% HTTP, 24% FTP, 13% Napster

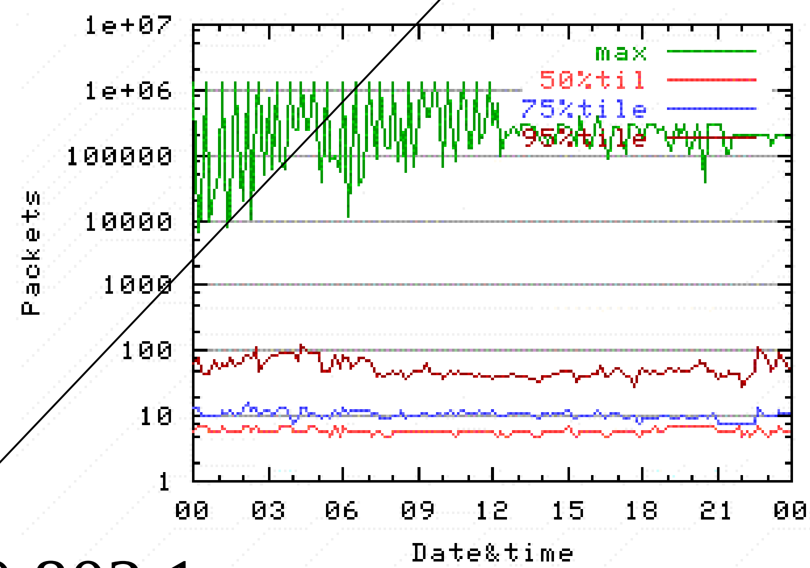
Time series





UDP





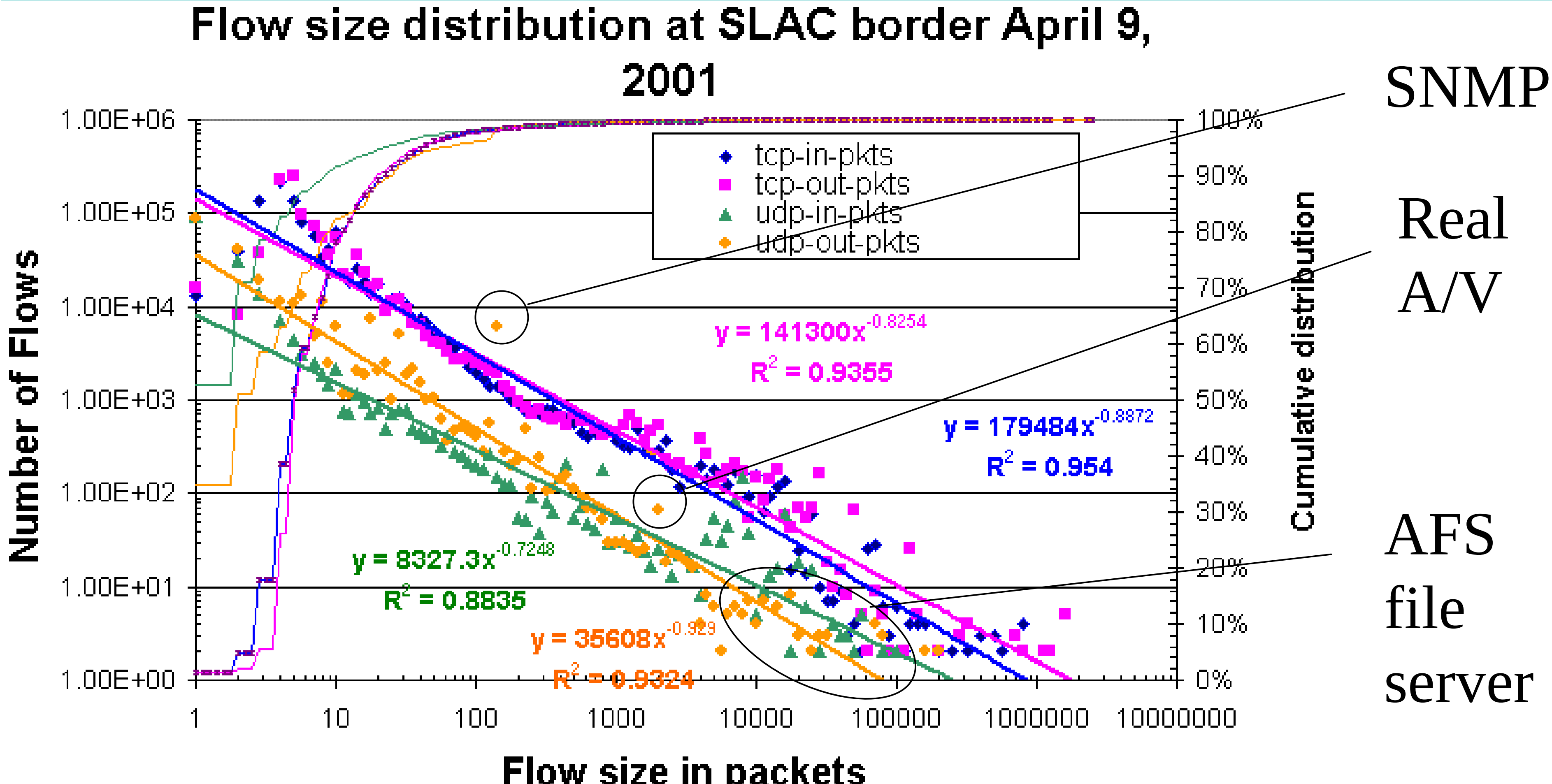
TCP

Outgoing

Cat 4000 802.1q vs. ISL

Incoming

Flow sizes



Confirms Nevil Brownlee's data measured at SDSC:

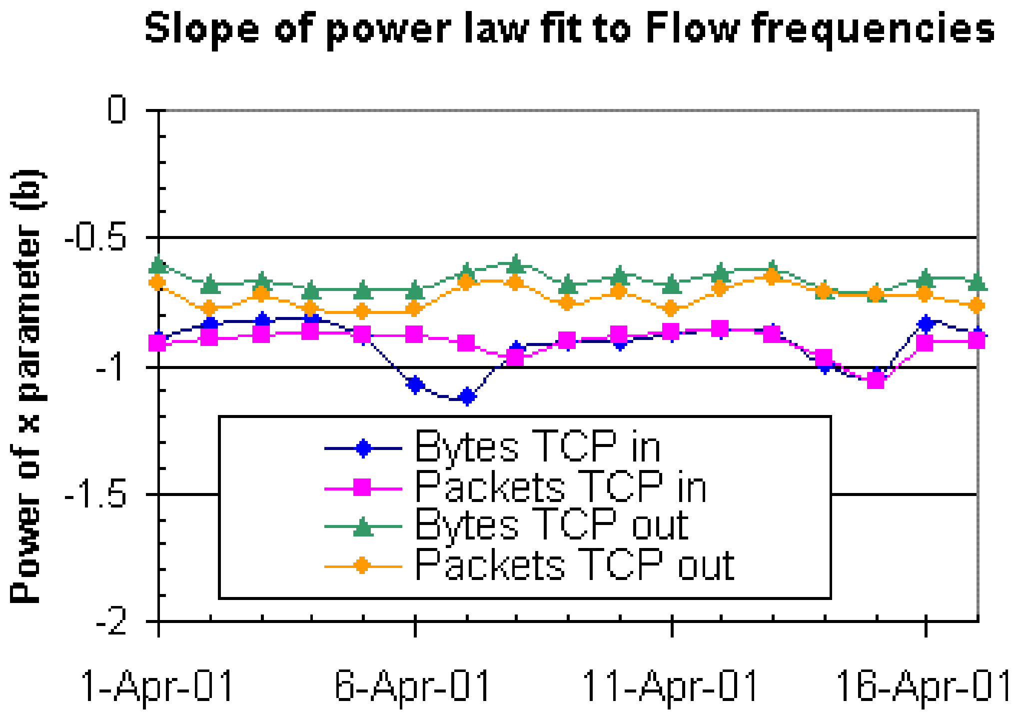
Heavy tailed, in ~ out, UDP flows shorter than TCP, packet~bytes

75% TCP-in < 5kBytes, 75% TCP-out < 1.5kBytes (<10pkts)

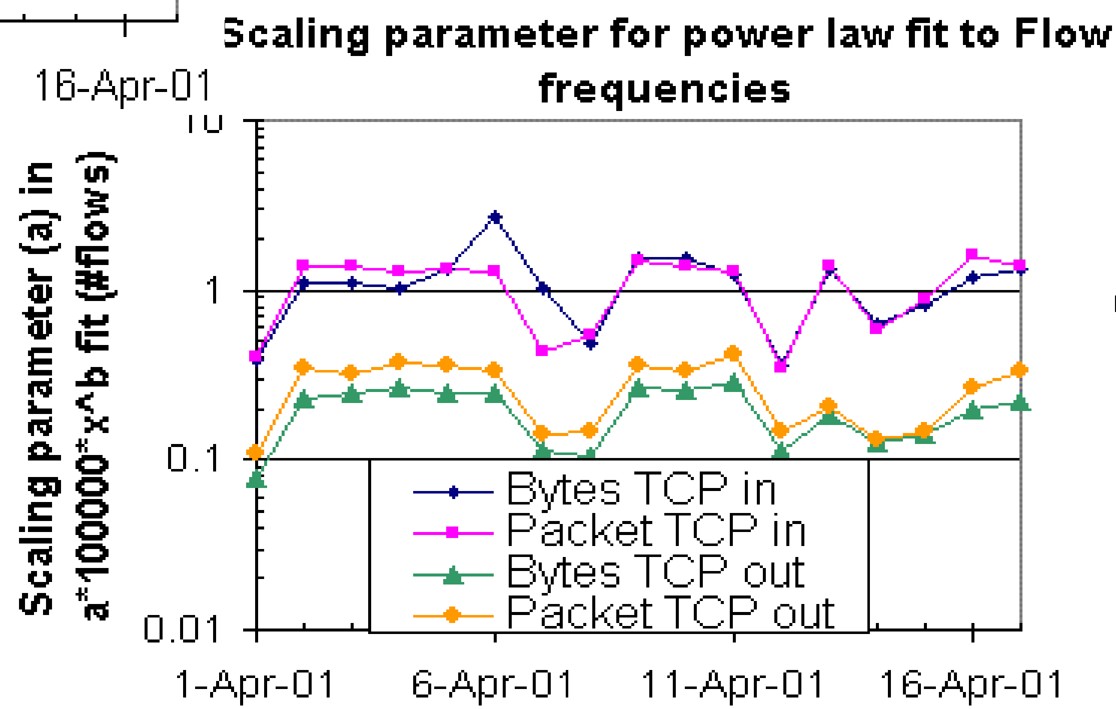
UDP 80% < 600Bytes (75% < 3 pkts), ~10 * more TCP than UDP

Top UDP = AFS (>55%), Real(~25%), SNMP(~1.4%)

Power law fit parameters by time

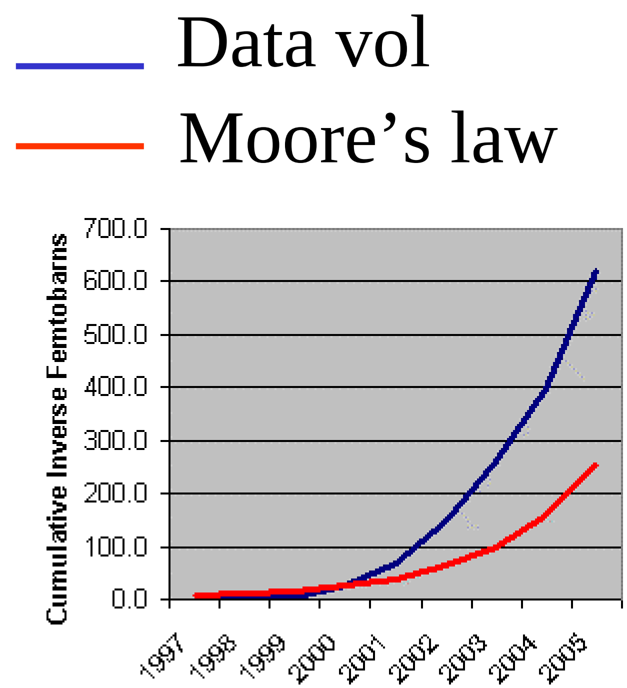


Just 2 parameters provide a reasonable description of the flow size distributions



App: High Speed Bulk Throughput

- **Driven by:**
  - Data intensive science, e.g. data grids
  - HENP data rates, e.g. BaBar 300TB/year, collection doubling yearly, i.e. PBytes in couple of years
  - Data rate from experiment ~ 20MBytes/s ~ 200GBytes/d
  - Multiple regional computer centers (e.g. Lyon-FR, RAL-UK, INFN-IT, LBNL-CA, LLNL-CA, Caltech-CA) need copies of data
  - Boeing 747 high throughput, BUT poor latency (~ 2 weeks) & very people intensive

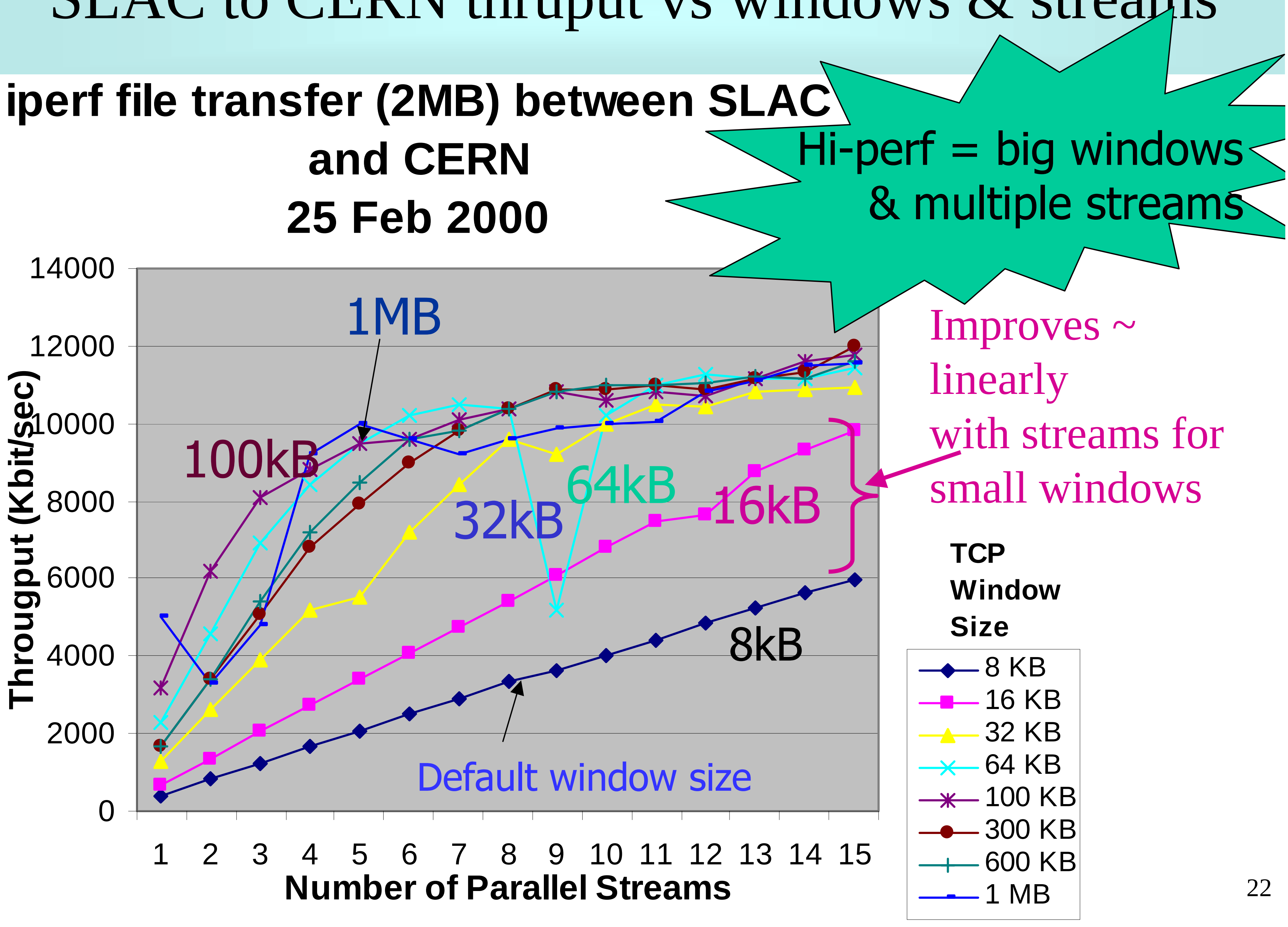


- So need high-speed networks and ability to utilize
  - High speed today = few hundred GBytes/day

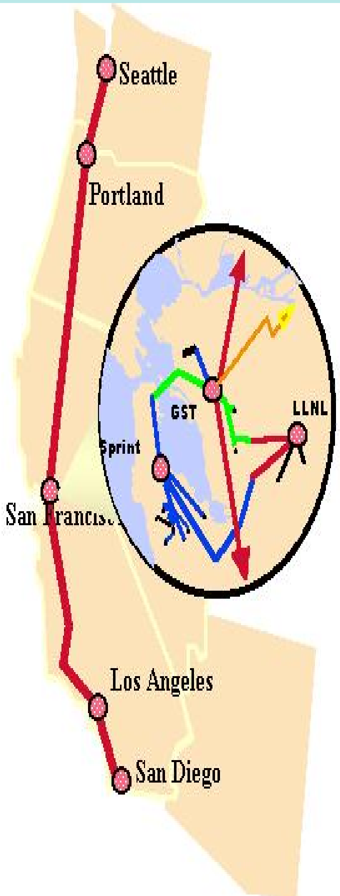
Measuring TCP throughput

- Selected about a dozen major collaborator sites in CA, CO, IL, FR, UK over last 9 months
  - Of interest to SLAC
  - Can get logon accounts
- Use **`iperf`**
  - Choose window size and # parallel streams
  - Run for 10 seconds together with ping (loaded)
  - Stop iperf, run ping (unloaded) for 10 seconds
  - Change window or number of streams & repeat
- Record streams, window, throughput (Mbits/s), loaded & unloaded ping responses

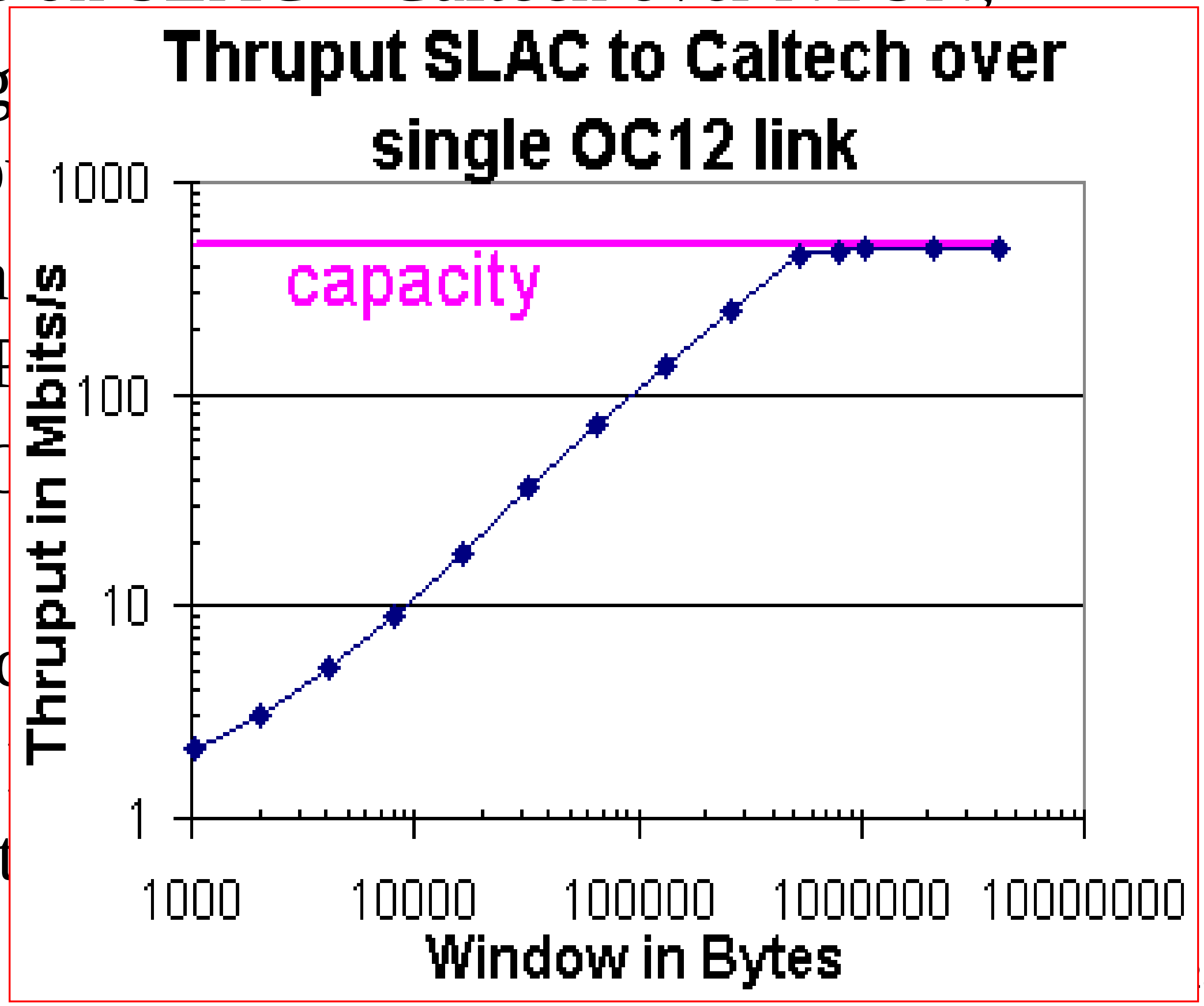
SLAC to CERN thruput vs windows & streams



Progress towards goal: 100 Mbytes/s Site-to-Site



- Focus on SLAC – Caltech over NTON;
- Using
  & do
- Repla
  with F
- SLAC
  and 2
- Caltec
- ~500
  recent



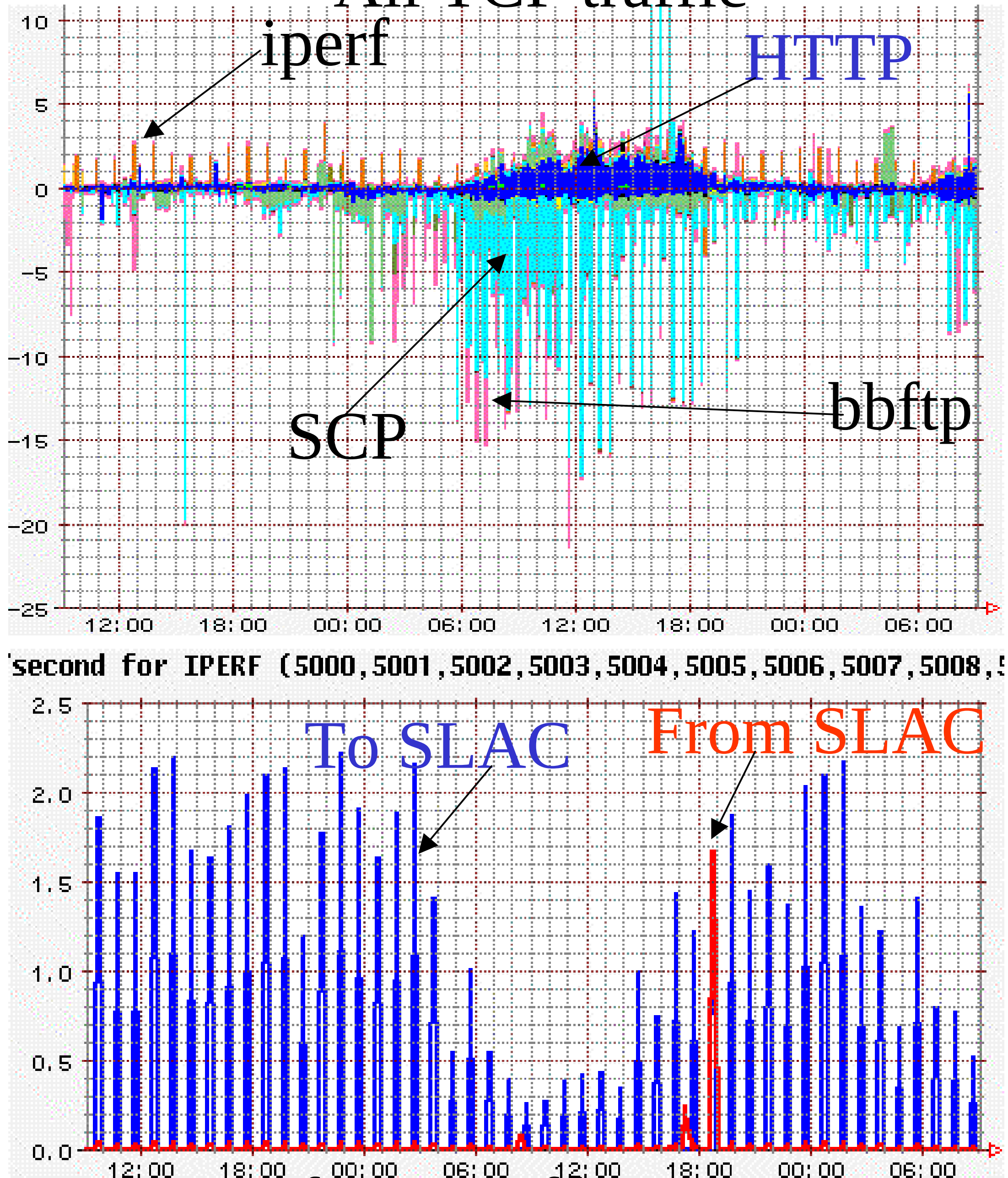
National Transparent Optical Network Consortium

SC2000 WAN Challenge

- SC2000, Dallas to SLAC RTT ~ 48msec
  - SLAC/FNAL booth: Dell PowerEdge PIII 2 * 550MHz with 64bit PCI + Dell 850MHz both running Linux, each with GigE, connected to Cat 6009 with 2GigE bonded to Extreme SC2000 floor switch
  - NTON: OC48 to GSR to Cat 5500 Gig E to Sun E4500 4*460MHz and Sun E4500 6*336MHz
- Internet 2: 300 Mbits/s
- NTON 960Mbits/s Dallas to SLAC mem-to-mem
- Details:
  - www-iepm.slac.stanford.edu/monitoring/bulk/sc2k.html

Impact of cross-traffic on Iperf between SLAC & W. Europe

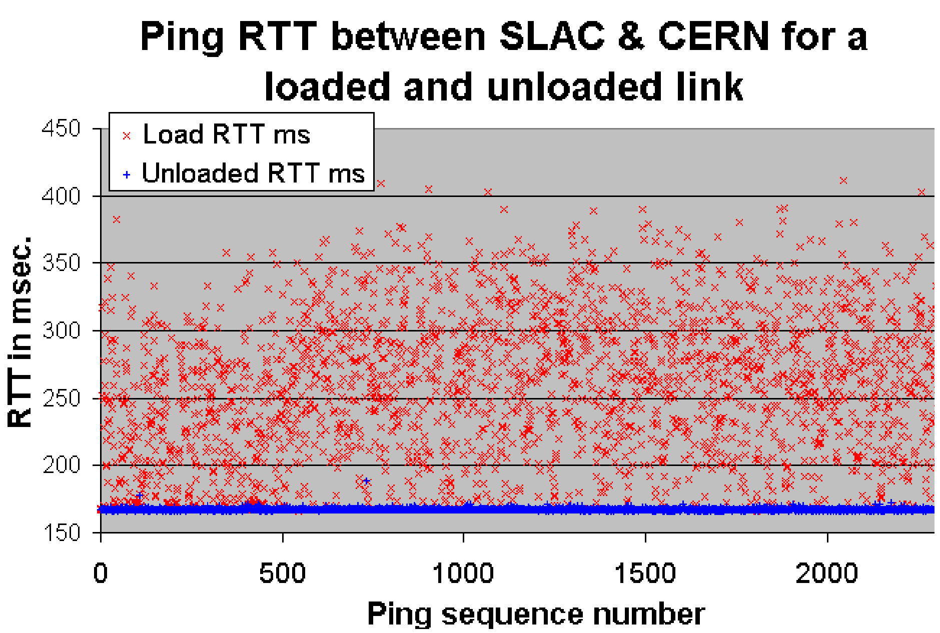
All TCP traffic

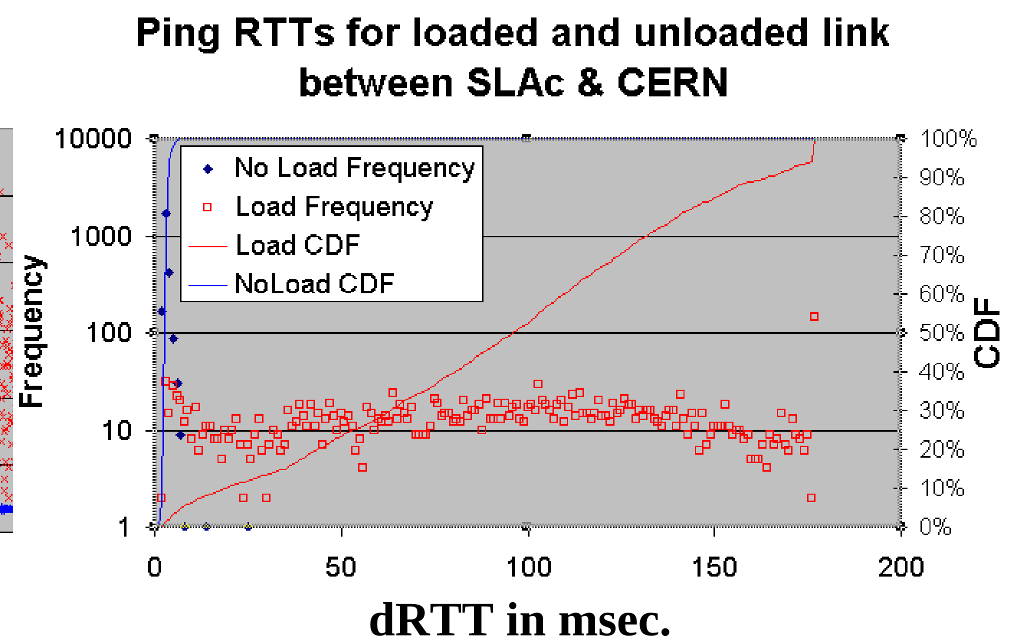


Iperf port traffic

Impact on Others

- Make ping measurements with & without iperf loading
  - Loss loaded(unloaded)
  - RTT





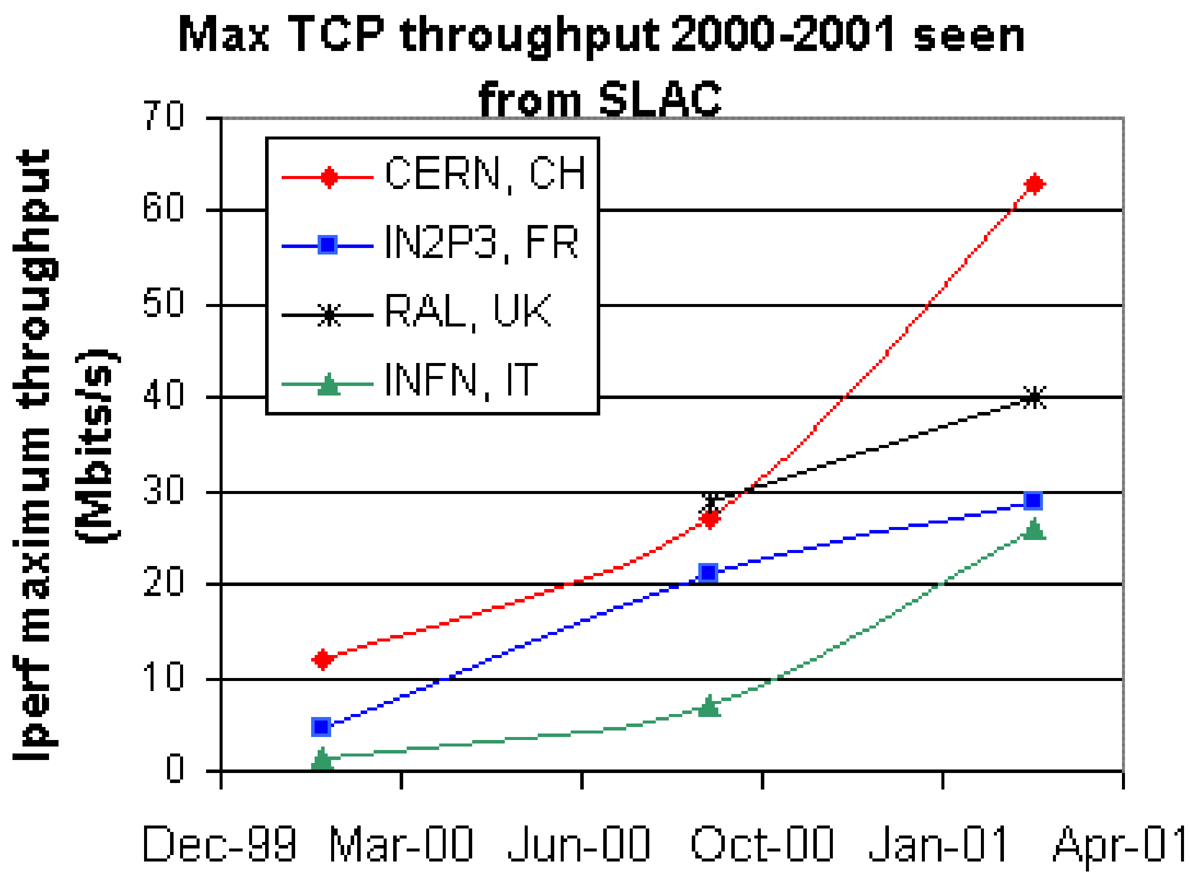
Improvements for major International BaBar sites

Links are being improved: ESnet, PHYnet, GARR, Janet, TEN-155

Improvements to come:

IN2P3 => 155Mbps

RAL => 622Mbps



Throughput improvements of 2 to 16 times in a year

Iperf throughput conclusions 1/2

- Can saturate bottleneck links
- For a given **`iperf`** measurement, streams share throughput equally.
- For small window sizes throughput increases linearly with number of streams
- Predicted optimum window sizes can be large (> Mbyte)
- Need > 1 stream to get optimum performance
- Can get close to max thruput with small (<=32Mbyte) with sufficient (5-10) streams
- Improvements of 5 to 60 in thruput by using multiple streams & larger windows
- Loss not sensitive to throughput

Iperf thruput conclusions 2/2

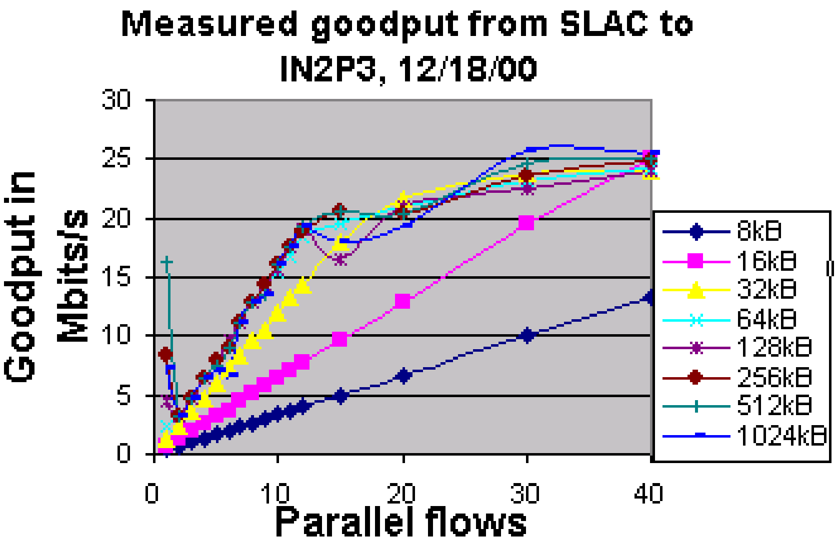
- For fixed ***streams*window*** product, streams are more effective than window size:

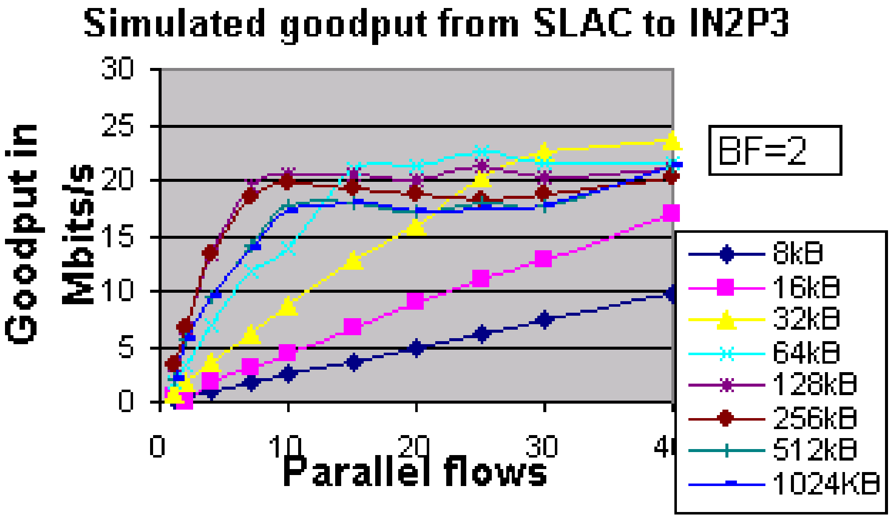
Site	Window	Streams	Throughput
CERN	256kB	2	9.45Mbits/
CERN	64kB	8	26.8Mbits/s
Caltech	256kB	2	1.7Mbits/s
Caltech	64kB	8	4.6Mbits/s

- There is an optimum number of streams above which performance flattens out
- See www-iepm.slac.stanford.edu/monitoring/bulk/

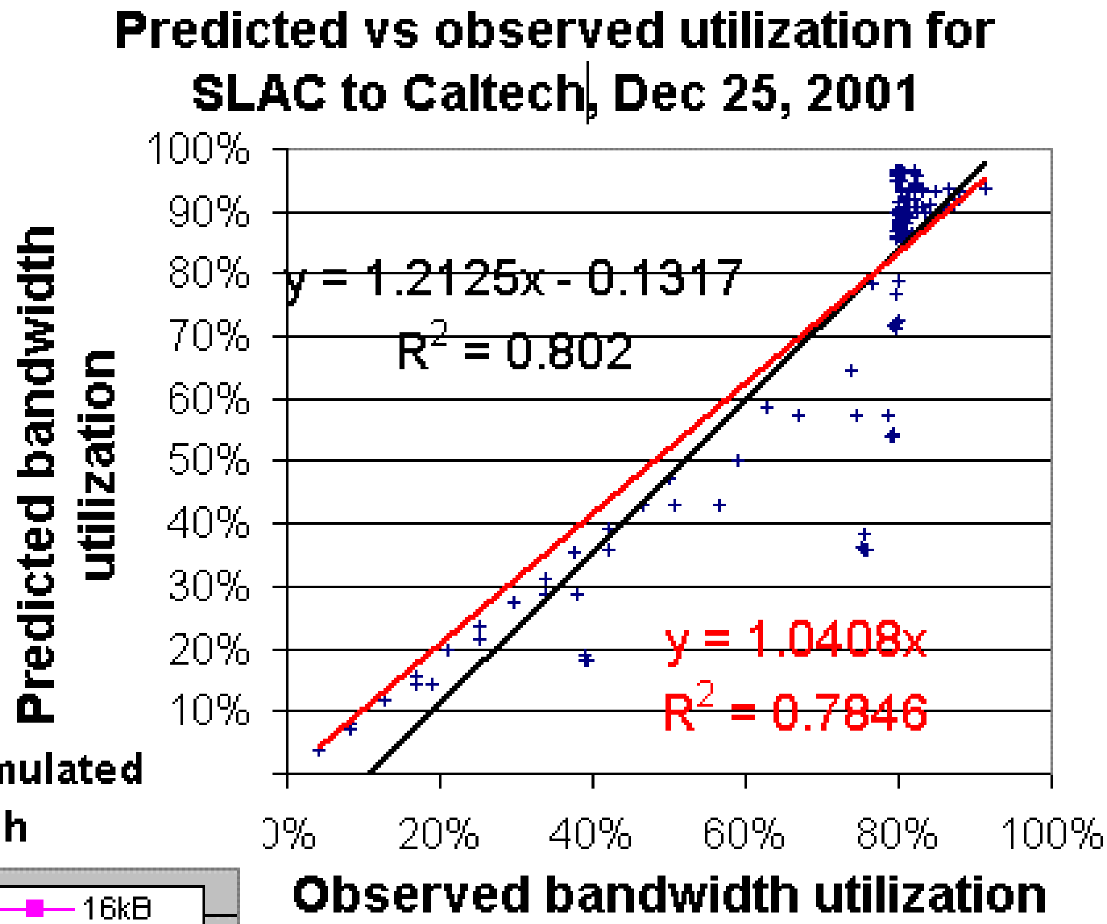
Network Simulator (ns-2)

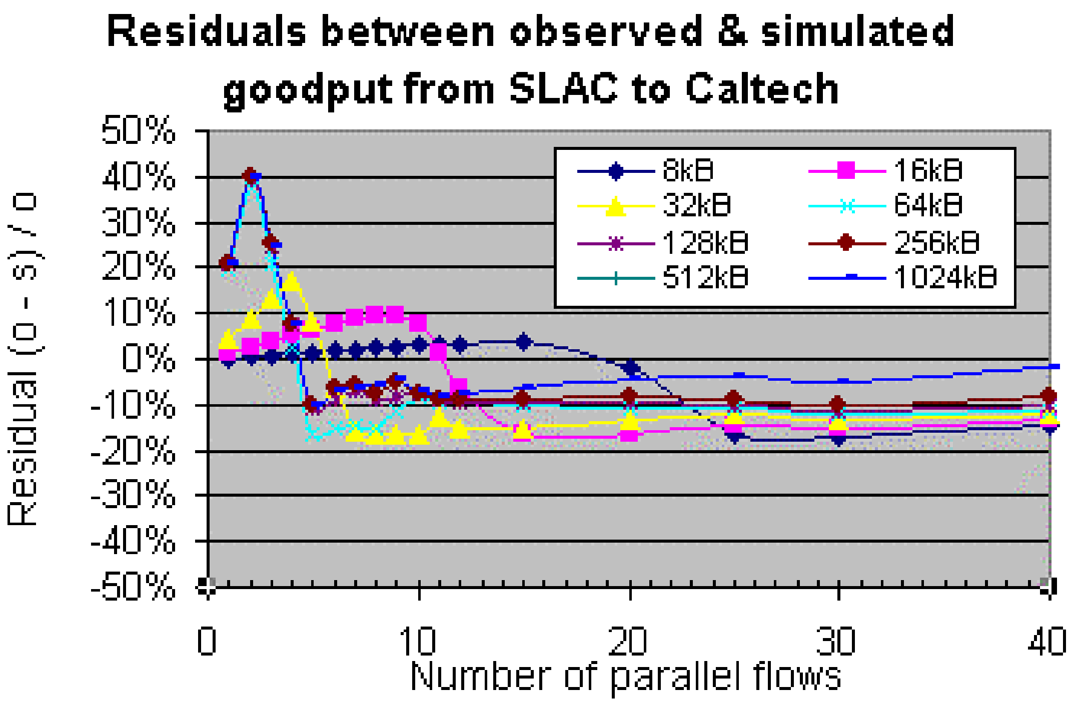
- From UCB, simulates network
  - Choice of stack (Reno, Tahoe, Vegas, SACK…)
  - RTT, bandwidth, flows, windows, queue lengths …
- Compare with measured results
  - Agrees well
  - Confirms observations (e.g. linear growth in throughput for small window sizes as increase number of flows)



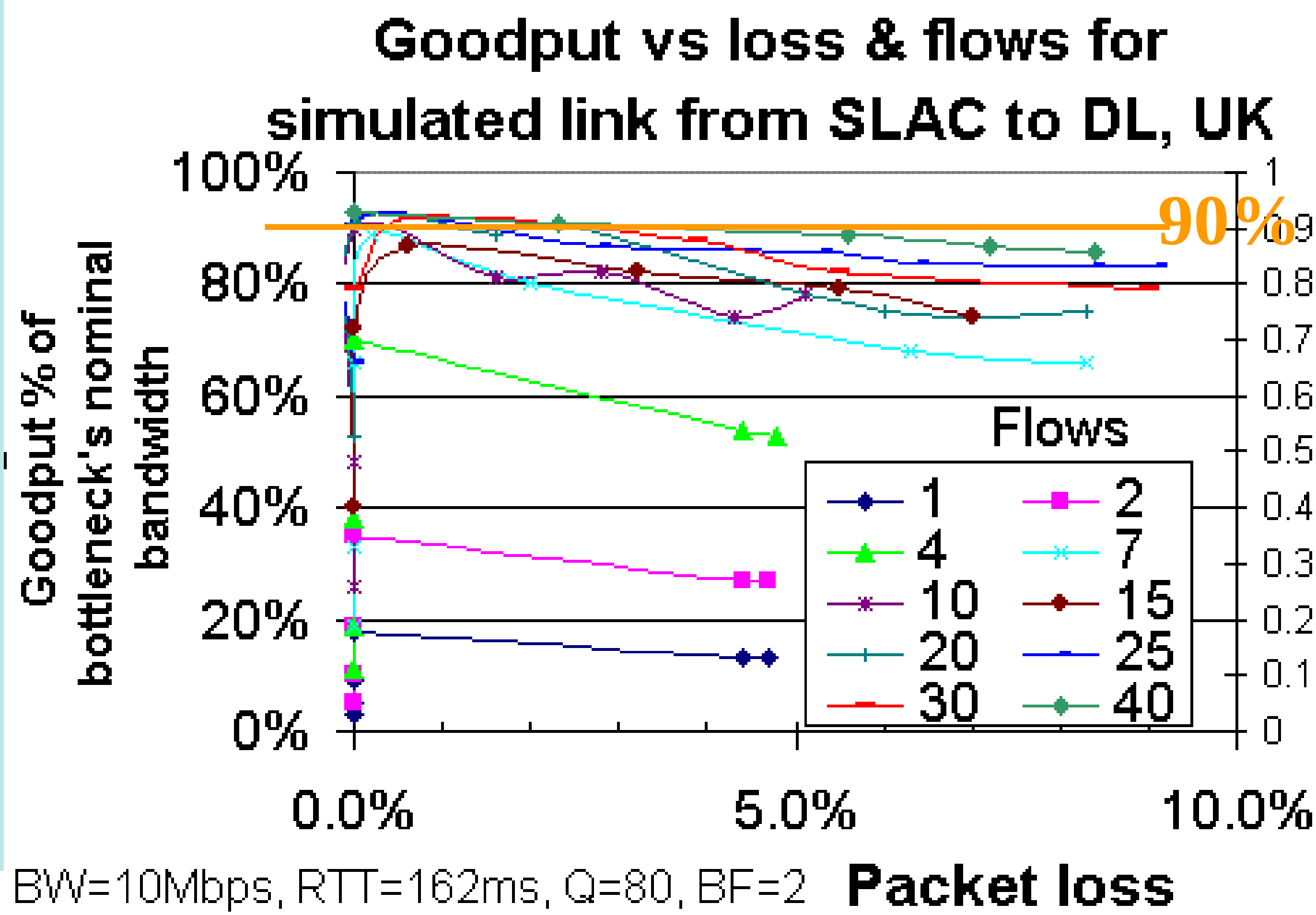


Agreement of ns2 with observed





Ns-2 thruput & loss predict



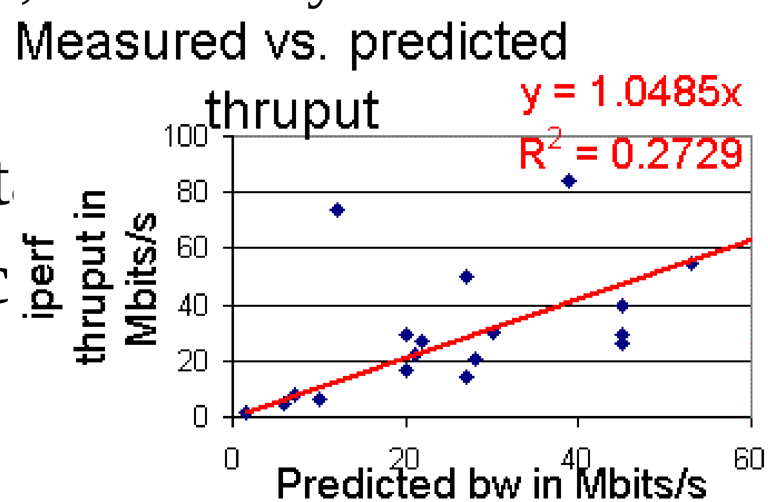
•Indicates on unloaded link can get 70% of available bandwidth without causing noticeable packet loss

•Can get over 80-90% of available bandwidth

•Can overdrive: no extra throughput BUT extra loss

Simulator benefits

- No traffic on network (nb throughput can use 90%)
- Can do what if experiments
- No need to install iperf servers or have accounts
- No need to configure host to allow large windows
- BUT
  - Need to estimate simulator parameters, e.g.
    - RTT use ping or synack
    - Bandwidth, use pchar, pipechar etc., moderately accurate
- AND ***its not the real thing***
  - Need to validate vs. observed dat
  - Need to simulate cross-traffic etc



WAN thruput conclusions

- High FTP performance across WAN links is possible
  - Even with 20-30Mbps bottleneck can do > 100Gbytes/day
- OS must support **big windows** selectable by application
- Need **multiple parallel streams**
- **Loss is important** in particular interval between losses
- Compression looks promising, but needs cpu power
- Can get close to max thruput with small (<=32Mbyte) with sufficient (5-10) streams
- **Improvements of 5 to 60 in thruput** by using multiple streams & larger windows
- Impacts others users, need **Less than Best Effort** QoS service

More Information

- This talk:
  - www.slac.stanford.edu/grp/scs/net/talk/slac-wan-perf-apr01.htm
- IEPM/PingER home site
  - www-iepm.slac.stanford.edu/
- Transfer tools:
  - http://hepwww.rl.ac.uk/Adye/talks/010402-ftp/html/sld015.htm
- TCP Tuning:
  - www.ncne.nlanr.net/training/presentations/tcp-tutorial.ppt