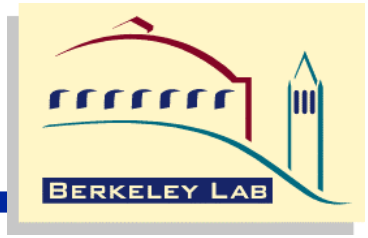


Experiences with TCP on a Routed IP over OC-12 ATM WAN

Brian L. Tierney (bltierney@lbl.gov)

Jason R. Lee (jrlee@lbl.gov)

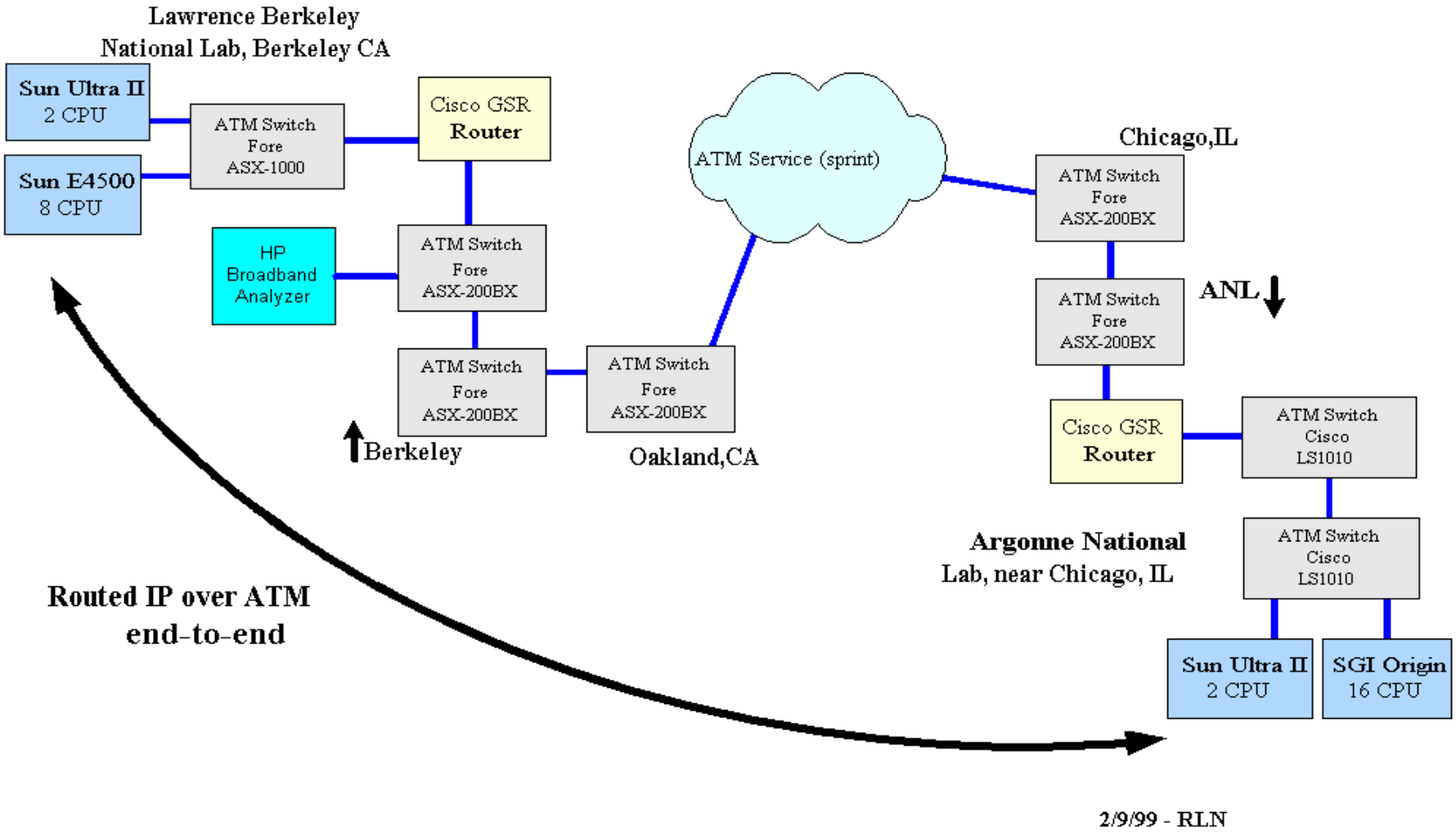
Future Technologies Group

Becca Nitzan (nitzan@es.net)

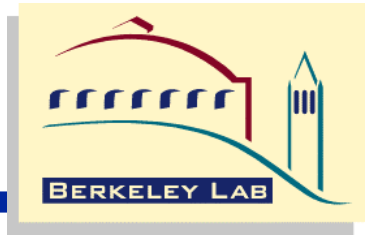
ESNet

Lawrence Berkeley National Laboratory

Network Configuration



Protocol Overhead

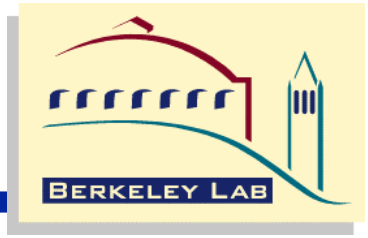


- **What is the maximum possible throughput?**

  —**Available Bandwidth after Protocol Overhead**

Protocol	OC-12 (MTU=9180)
Line rate	622.080 Mbps
To ATM	600.768 Mbps
To AAL	544.092 Mbps
To IP	541.966 Mbps
To appl. via TCP	539.605 Mbps

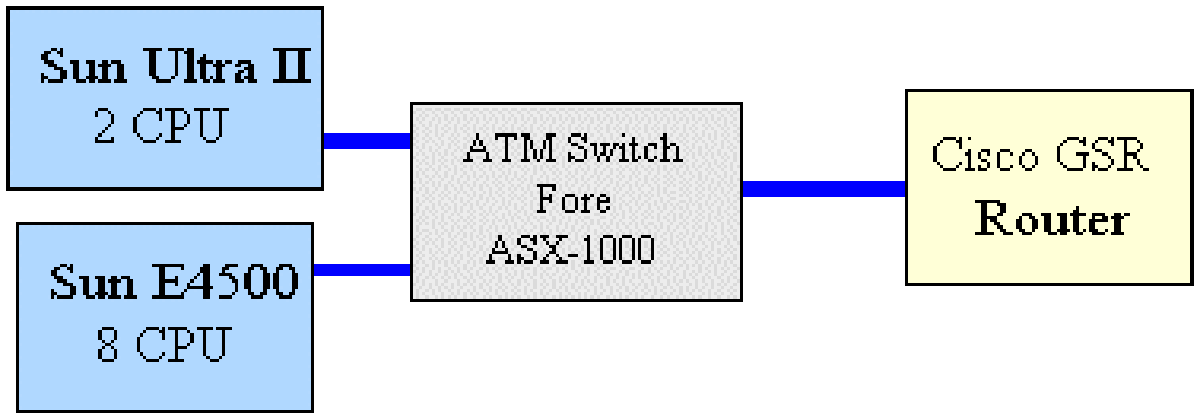
TCP Testing



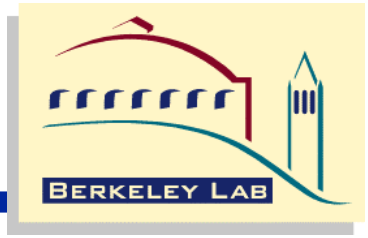
- Tested TCP throughput using ttcp with 4 MB send and receive buffers on a “private” network: no other traffic during these tests
  - —Results: Throughput = 150 to 300 Mbps
- Installed the SACK patch (RFC 2018) from Sun
  - —Results: Throughput = 300 to 480 Mbps
- Variance due to cell loss: GSR routers reported 3-4 packet losses during typical 3-5 minute test
- Replaced port in Oakland ATM switch to try to correct cell loss problem
  - —Results: 340 to 480 Mbps: less packet loss now, resulting in smaller range of throughput

Loopback Testing

- **Setup a loopback through the GSR router at LBNL (local test from Sun A to GSR, IP level routed back to Sun B)**
  - **Results: 513 Mbps TCP throughput (same speed as host to host without the switch or router)**
  - **This gave us a performance baseline**
  - **GSR is not a bottleneck**

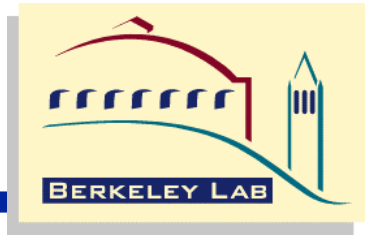


HP ATM Tester



- Added add an HP Broadband Analyzer at LBNL, and setup a loopback in the GSR at ANL
  - —HP Analyzer does “GCRA (Generic Cell Rate Algorithm) compliance testing” (traffic shaping)
- Discovered a bug in the GSR policing code
  - —GSR has PCR (Peak Cell Rate), SCR (sustained) and MBS (Max burst size) "equivalents" that are setable. SCR was being ignored.
  - —The GSR was honoring the PCR and not the SCR. This was tested by issuing pings FROM the GSR to the HP
- After fixing this bug: achieved 572 Mbps (ATM rate) (max ATM over OC-12 = 600 Mbps)

Summary of TCP Performance



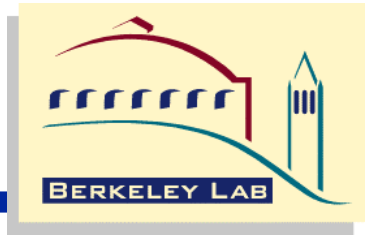
- Early informal testing

Test	Throughput Range (Mbits/sec)
TCP: Local LBL loop through GSR	400 to 513 Mbps
TCP: LBNL to ANL (no SACK)	150 to 300 Mbps
TCP: LBNL to ANL (with SACK)	300 to 480 Mbps

- Current Results (10 GB transfer, shared link)

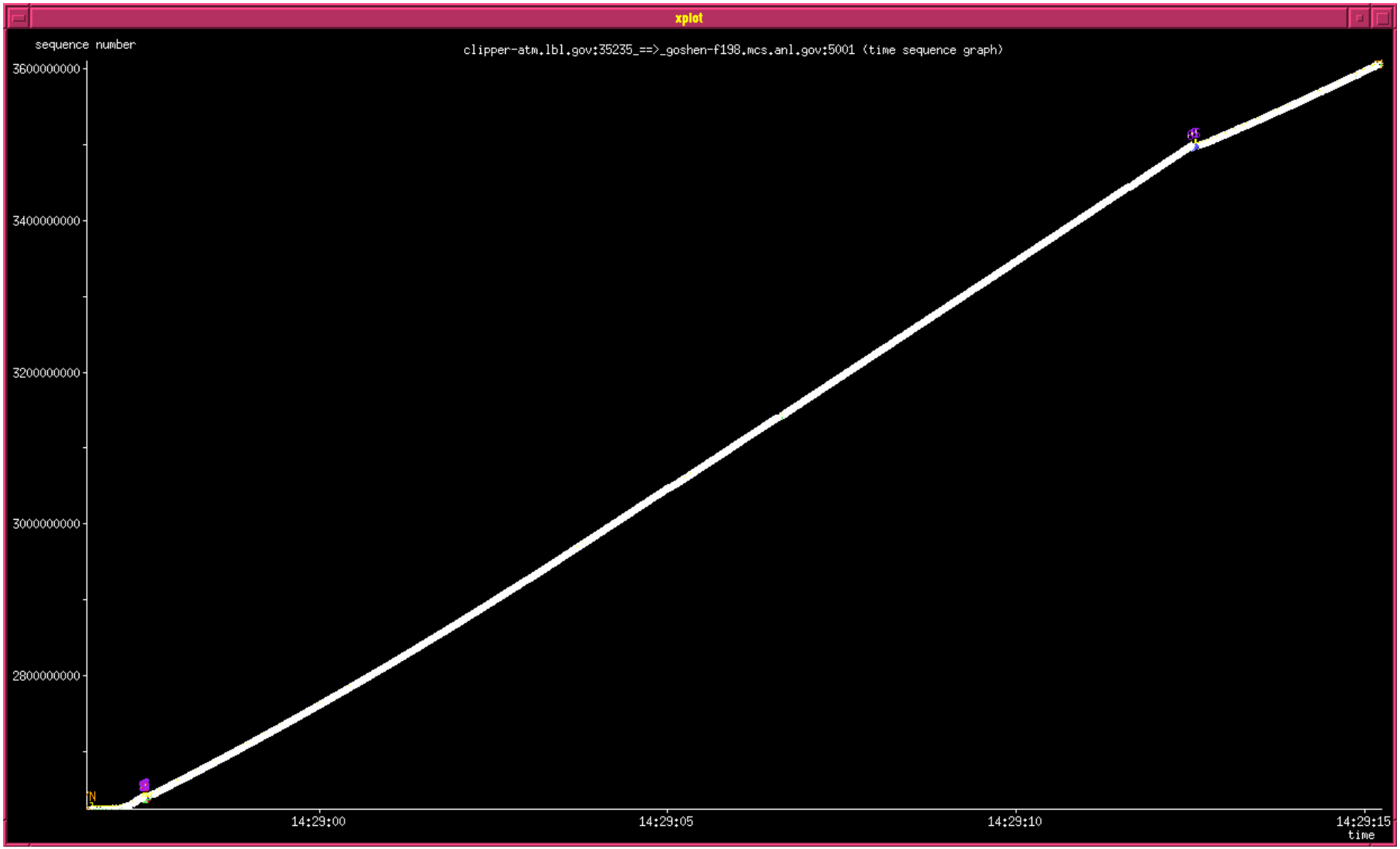
Test	Min	Max	Average	Std
ANL to LBNL	278	393	346	36.69
LBNL to ANL	285	387	352	23.59

Why the Large Variance in Throughput?



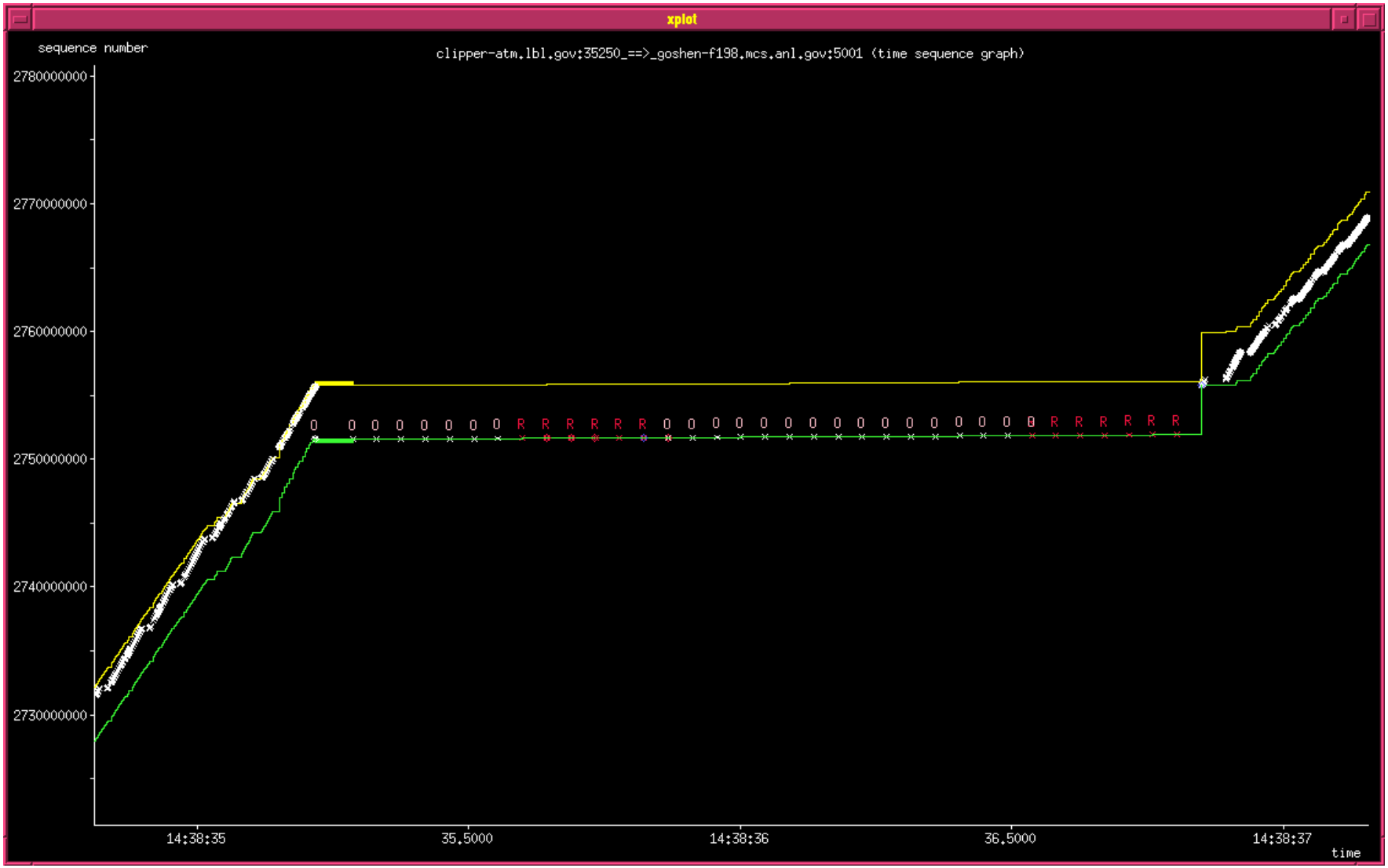
- Most TCP traces show 1-3 “glitches” during a 10 GB transfer (see traces on following slides)
  - —GSR router at ANL reports CRC errors on input
  - —very hard to determine source of these errors

Typical Trace

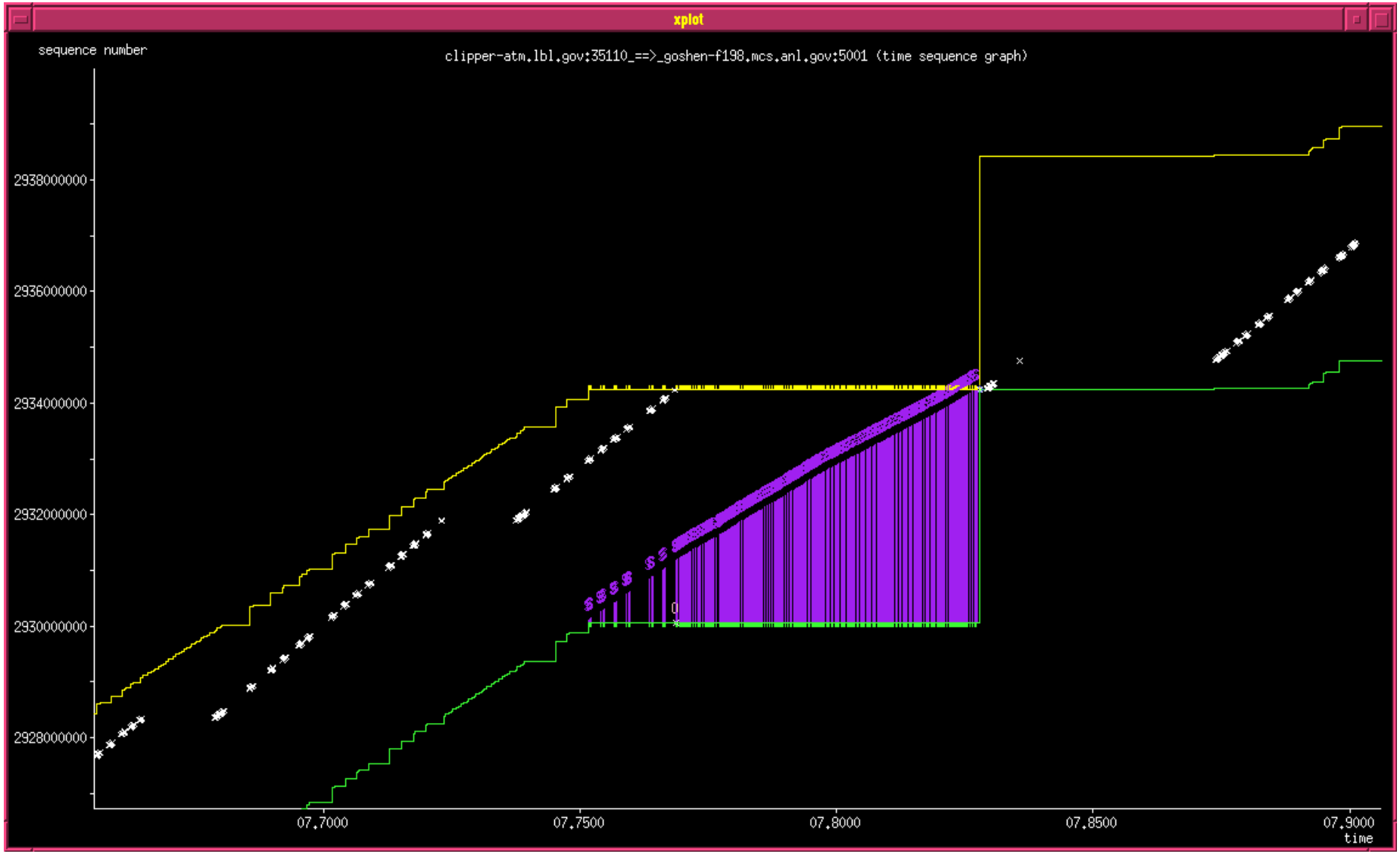


Without SACK

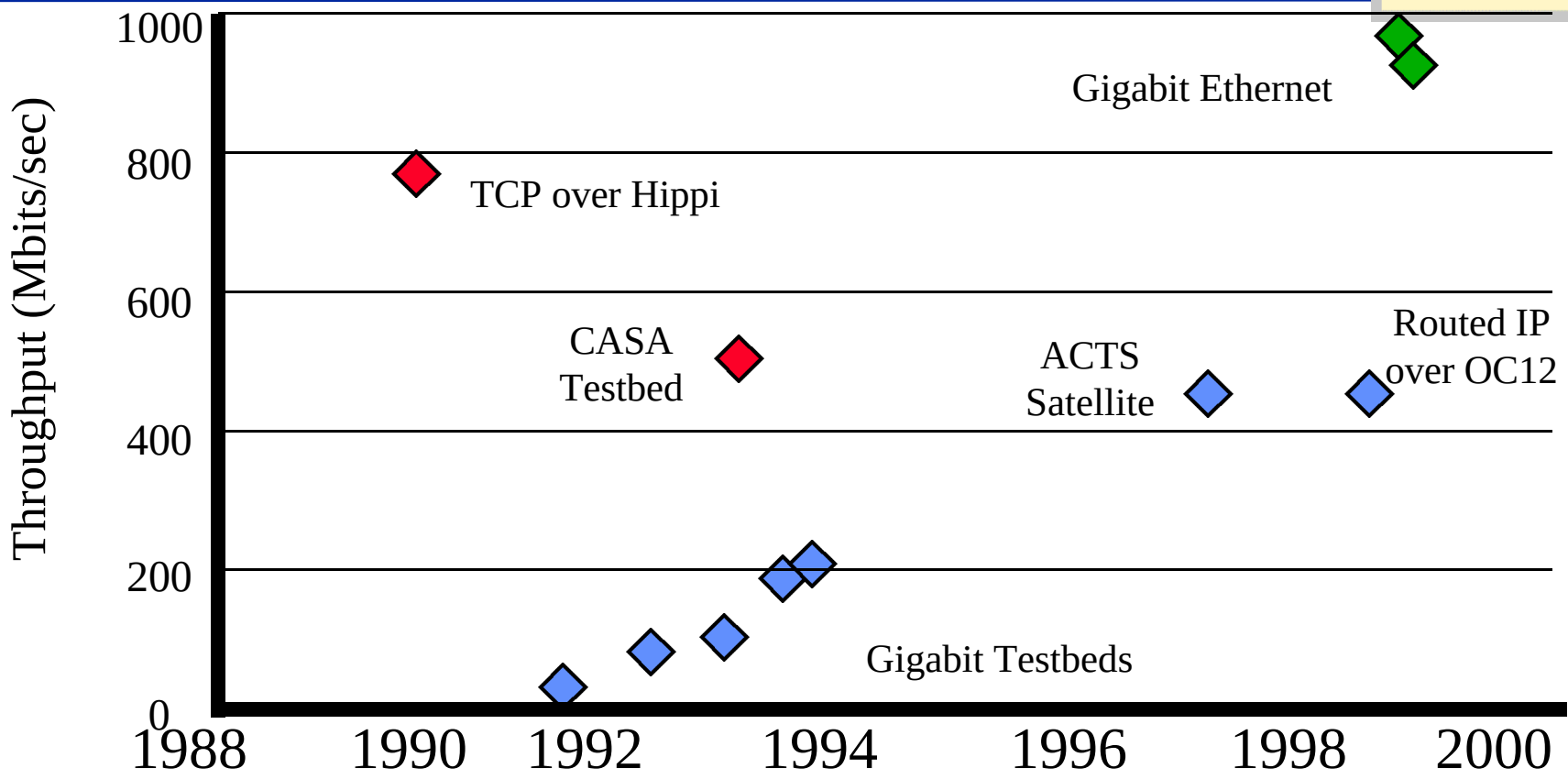
Without SACK: close up



SACK Close Up



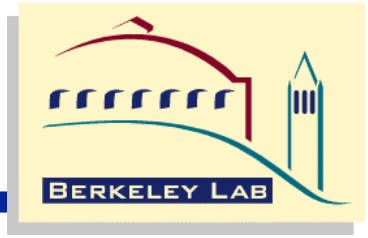
History of TCP over "Gigabit" Networks



TCP Performance Issues:

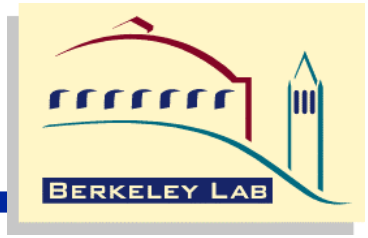
- **window Scale option**
- **zero copy TCP**
- **Host memory**
- **HW checkum**
- **Cell pacing**
- **Switch buffer size**
- **Early Packet Discard**
- **SACK**
- **RED**

TCP History: Sample Results



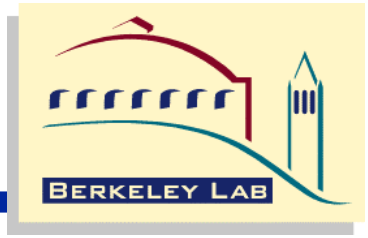
- Cray to Cray, Hippi LAN: 780 Mbps; PSC, 1990
- Cray to Intel, Hippi over Sonet, 500 Mbps, CASA, 1993
- Gigabit Testbeds, OC-3 and OC-12, 1993-94
  - Magic: 40 to 130 Mbps
  - Bagnet: 40-90 Mbps
  - Aurora: 215 Mbps
  - VistaNet and Nectar: 200 Mbps
- ACTS: 480 Mbps over OC-12, 1998
- LBNL to SLAC (MAN test): 480 Mbps
- LBNL to ANL through IP routers: 480 Mbps
- Sun/Alteon Gigabit Ethernet tests: 990 Mbps

Issues



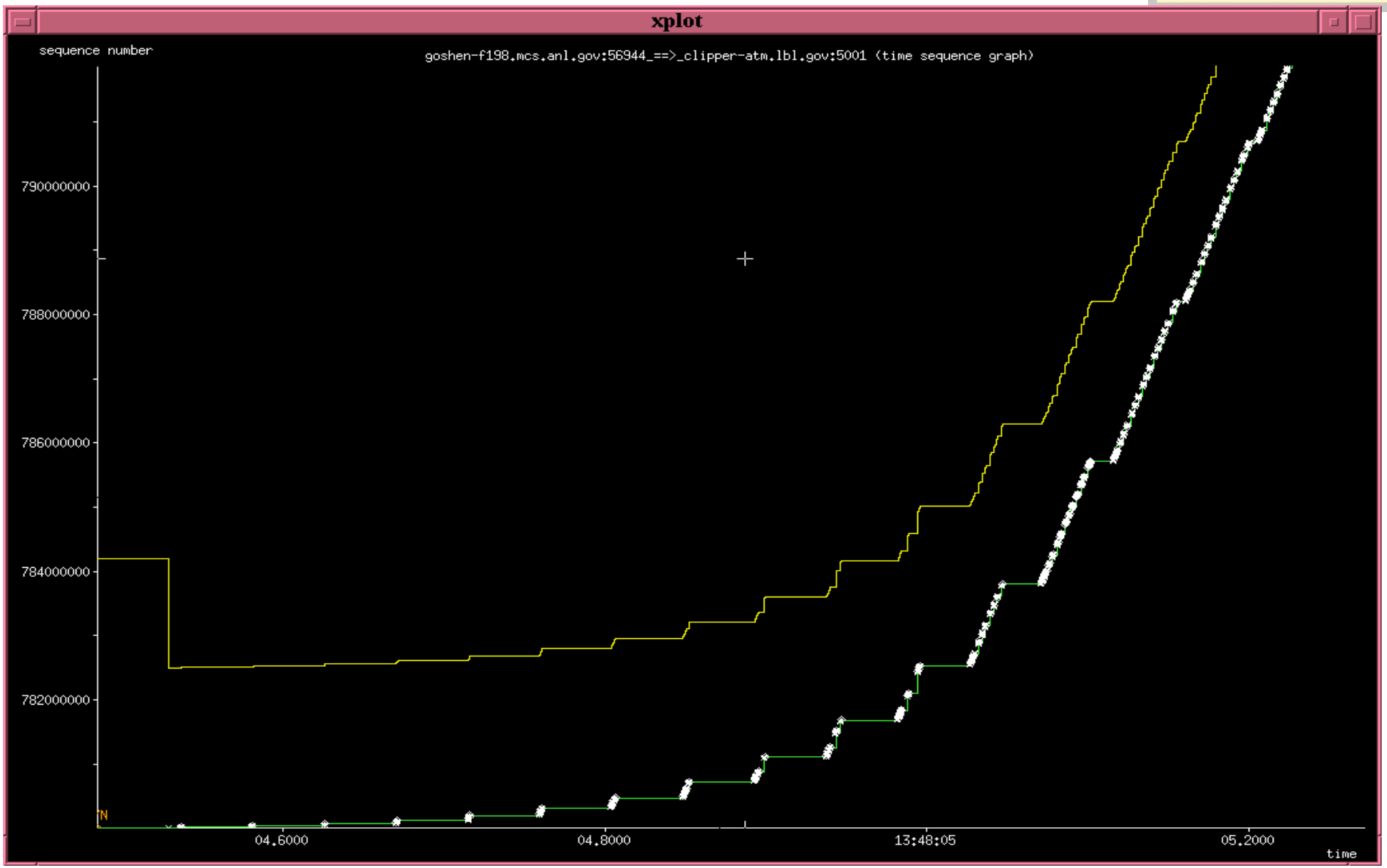
- **Still very hard to find problems**
  - **ATM switches still do not accurately report cell loss (it was a LOT of work to track down the bad ATM switch card)**
  - **Can not see into ISP ATM cloud**
  - **Often not getting what you are paying for**

Issues

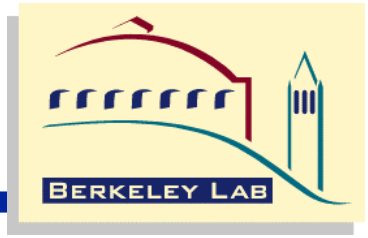


- What else might be done to improve the TCP throughput?
  - —TCP is very sensitive to packet/cell loss
  - —Takes 12 RTT's to ramp up to full window size
- New TCP enhancements might help:
  - —TCP Vegas: more sophisticated bandwidth estimation scheme
  - —Increasing TCP's Initial Window based on previous connection (slow-start restart)

TCP Slow Start



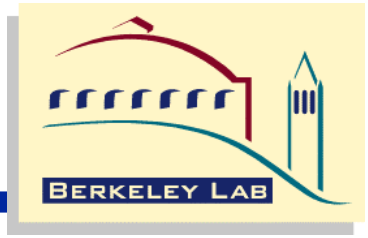
Effects of TCP Slow-Start



- **Network Characteristics**
  - —**RTT = 45 ms, bw = 450 Mbps, MTU = 9180** Bytes**, round trips stalled in slow start = 11**
- **Would "slow-start restart" help?**
  - —**Only for relatively small files**

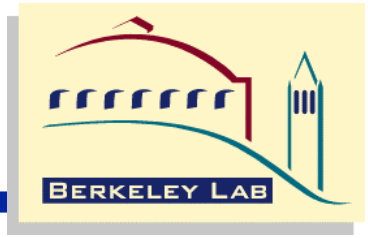
File Size	Minimum Transfer Time	"wasted" time	Speed Up
10 GB	178 sec	.54 sec	.30%
1 GB	18.2 sec	.54 sec	2.9%
100 MB	2.31 sec	.54 sec	23.4%

What Next?



- TCP over OC-48
  - What are the issues?
  - Memory/Bus bandwidth is again an issue
  - Discussion

Conclusions



- High throughput TCP is possible in long distance OC-12 environment, but TCP is quite sensitive to packet/cell loss
- SACK option helps quite a bit when there is cell loss
- GSR Router does not appear to be a bottleneck
- Very difficult to locate the source of cell loss

- Useful URLs:
  - http://www.es.net/
  - http://www-didc.lbl.gov/
  - http://www.psc.edu/networking/all_sack.html (links to several SACK implementations)