Achieving Record Speed Trans-Atlantic End-to-end TCP Throughput

***Les Cottrell** – SLAC*

Prepared for the NANOG meeting, Salt Lake City, June 2003

http://www.slac.stanford.edu/grp/scs/net/talk/nanog-jun03.html

Partially funded by DOE/MICS Field Work Proposal on Internet End-to-end Performance Monitoring (IEPM), by the SciDAC base program.

Outline

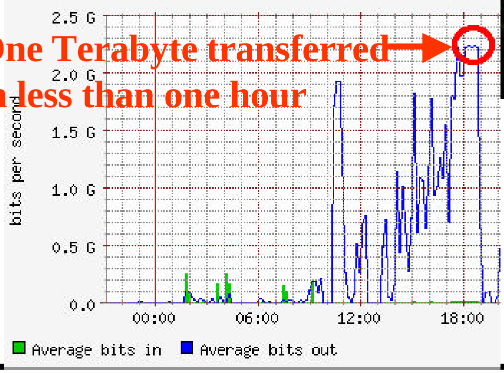
- Breaking the Internet2 Land Speed Record
  - Not be confused with:
    - ***Rocket-powered sled travels about 6,400 mph to break 1982 world land speed record***, San Francisco Chronicle May 1, 2003
- Who did it
- What was done
- How was it done?
- What was special about this anyway?
- Who needs it?
- So what’s next?
- Where do I find out more?

Who did it: Collaborators and sponsors

- **Caltech:** Harvey Newman, Steven Low, Sylvain Ravot, Cheng Jin, Xiaoling Wei, Suresh Singh, Julian Bunn
- **SLAC:** Les Cottrell, Gary Buhrmaster, Fabrizio Coccetti
- **LANL:** Wu-chun Feng, Eric Weigle, Gus Hurwitz, Adam Englehart
- **CERN:** Olivier Martin, Paolo Moroni
- ANL: Linda Winkler
- DataTAG, StarLight, TeraGrid, SURFnet, NetherLight, Deutsche Telecom, Information Society Technologies
- Cisco, Level(3), Intel
- DoE, European Commission, NSF

What was done?

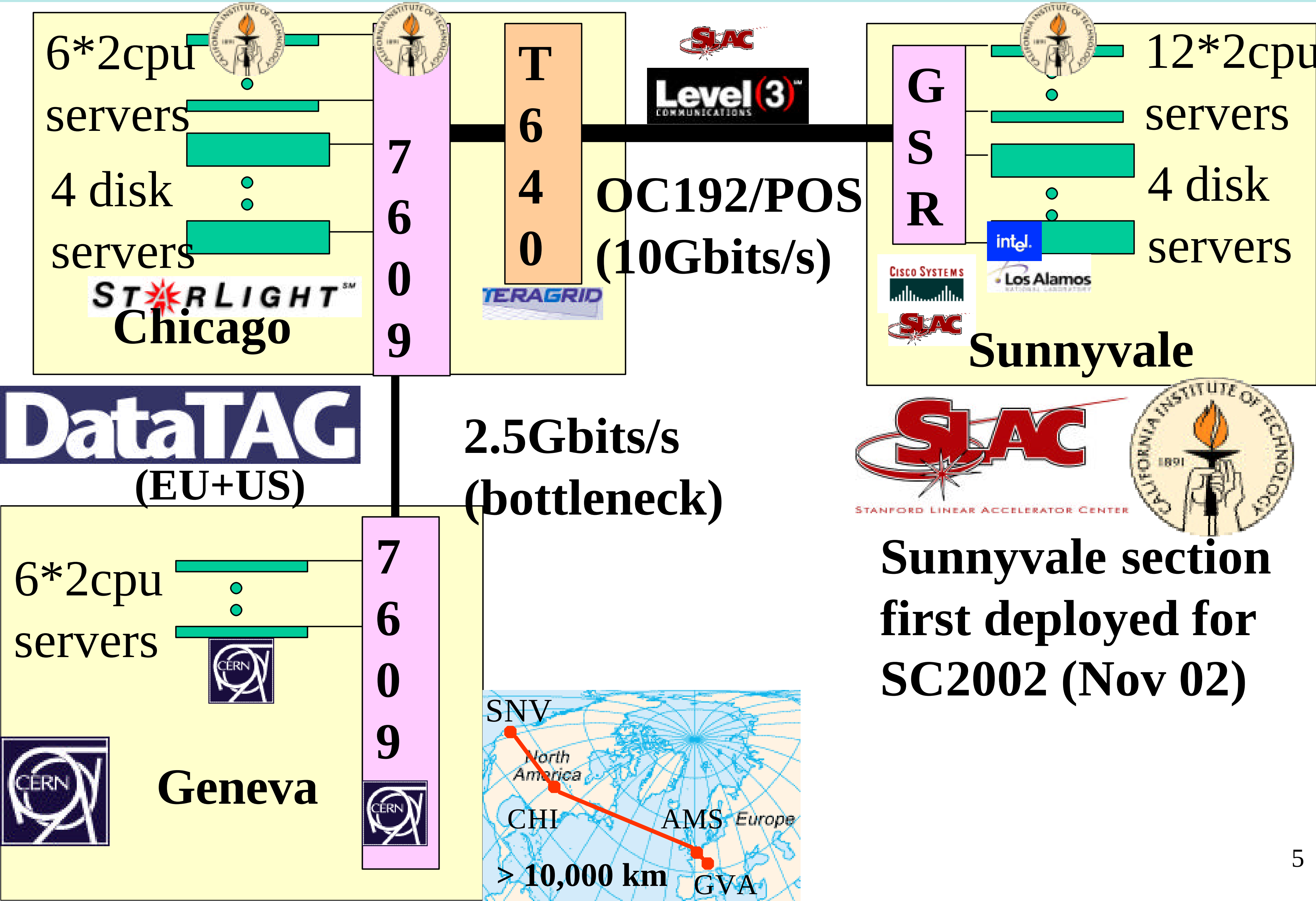
- Set a new Internet2 TCP land speed record, 10,619 Tbit-meters/sec
  - (see http://lsr.internet2.edu/)
- With 10 streams achieved 8.6Gbps across US
- **Beat the Gbps limit for a single TCP stream across the Atlantic – transferred a TByte in an hour**



One Terabyte transferred in less than one hour

When	**From**	**To**	**Bottle-neck**	**MTU**	**Streams**	**TCP**	**Thru-put**
Nov '02 (SC02)	Amsterdam	Sunny-vale	1 Gbps	9000B	1	Standard	923 Mbps
Nov '02 (SC02)	Balti-more	Sunny-vale	10 Gbps	1500	10	FAST	8.6 Gbps
Feb '03	Sunny-vale	Geneva	2.5 Gbps	9000B	1	Standard	2.38 Gbps

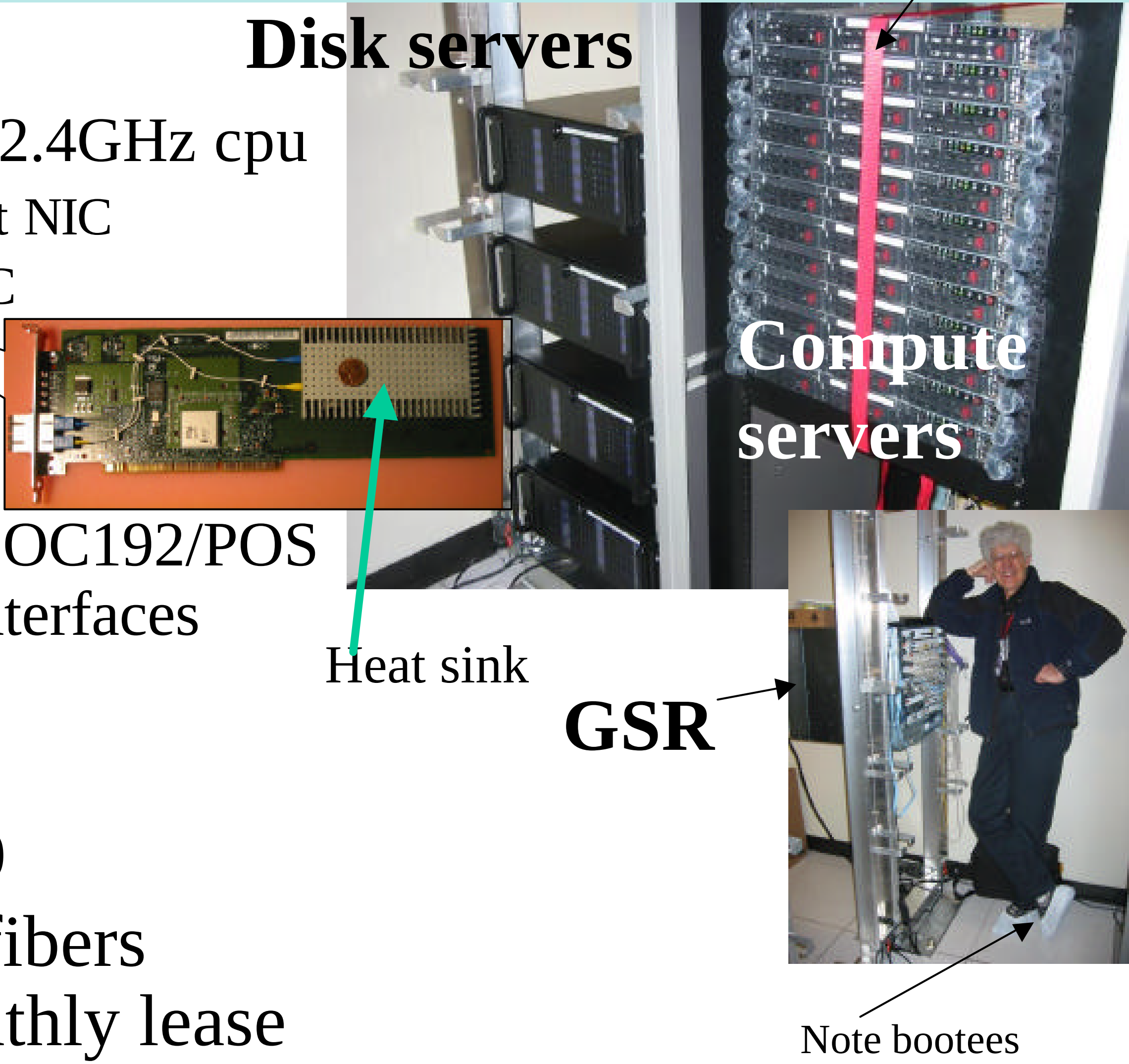
How was it done: Typical testbed



Sunnyvale section first deployed for SC2002 (Nov 02)

Typical Components

- CPU
  - Pentium 4 (Xeon) with 2.4GHz cpu
    - For GE used Syskonnect NIC
    - For 10GE used Intel NIC
  - Linux 2.4.19 or 20
- Routers
  - Cisco GSR 12406 with OC192/POS & 1 and 10GE server interfaces (loaned, list > $1M)
  - Cisco 760x
  - Juniper T640 (Chicago)
- Level(3) OC192/POS fibers (loaned SNV-CHI monthly lease cost ~ $220K)

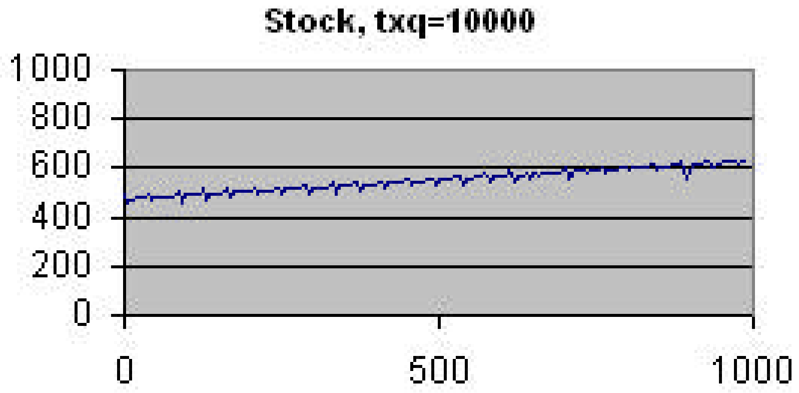


Challenges

- PCI bus limitations (66MHz * 64 bit = 4.2Gbits/s at best)
- At 2.5Gbits/s and 180msec RTT requires 120MByte window
- Some tools (e.g. bbcp) will not allow a large enough window – (bbcp limited to 2MBytes)
- Slow start problem at 1Gbits/s takes about 5-6 secs for 180msec link,
  - i.e. if want 90% of measurement in stable (non slow start), need to measure for 60 secs
  - need to ship >700MBytes at 1Gbits/s

- After a loss it can take over an hour for stock TCP (Reno) to recover to maximum throughput at 1Gbits/s
  - i.e. loss rate of 1 in ~ 2 Gpkts (3Tbits), or BER of 1 in $3.6*10^{12}$

Sunnyvale-Geneva, 1500Byte MTU, stock TCP

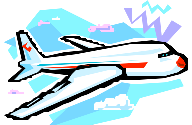


What was special? 1/2

- End-to-end application-to-application, single and multi-streams (not just internal backbone aggregate speeds)
- TCP has not run out of stream yet, scales from modem speeds into multi-Gbits/s region
  - TCP well understood, mature, many good features: reliability etc.
  - Friendly on shared networks
- New TCP stacks only need to be deployed at sender
  - Often just a few data sources, many destinations
  - No modifications to backbone routers etc
  - No need for jumbo frames
- Used Commercial Off The Shelf (COTS) hardware and software

What was Special 2/2

- Raise the bar on expectations for applications and users
  - Some applications can use Internet backbone speeds
  - Provide planning information
- The network is looking less like a bottleneck and more like a catalyst/enabler
  - Reduce need to colocate data and cpu
  - No longer ship literally truck or plane loads of data around the world



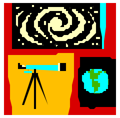
FedEx®

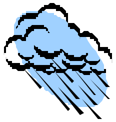
  - Worldwide collaborations of people working with large amounts of data become increasingly possible

Who needs it?

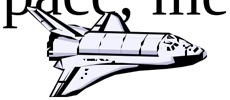
- HENP – current driver
  - Multi-hundreds Mbits/s and Multi TByte files/day transferred across Atlantic today
    - SLAC BaBar experiment already has almost a PByte stored

  - Tbits/s and ExaBytes (10^{18}) stored in a decade
- Data intensive science:
  - Astrophysics, Global weather, Bioinformatics, Fusion, seismology…

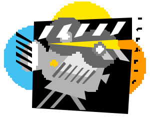




- Industries such as aerospace, medicine, security …

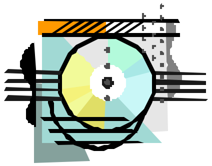


- Future:
  - Media distribution
    - Gbits/s=2 full length DVD movies/minute



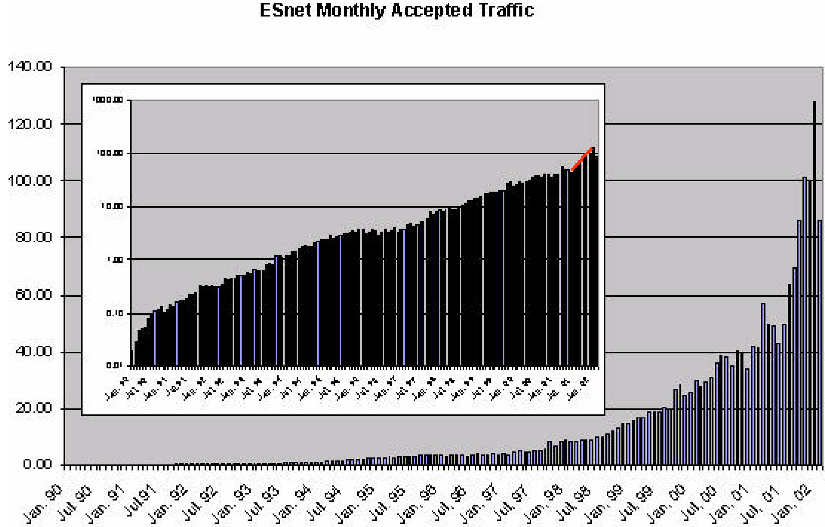
    - 2.36Gbits/s is equivalent to
      - Transferring a full CD in 2.3 seconds (i.e. 1565 CDs/hour)
      - Transferring 200 full length DVD movies in one hour (i.e. 1 DVD in 18 seconds)

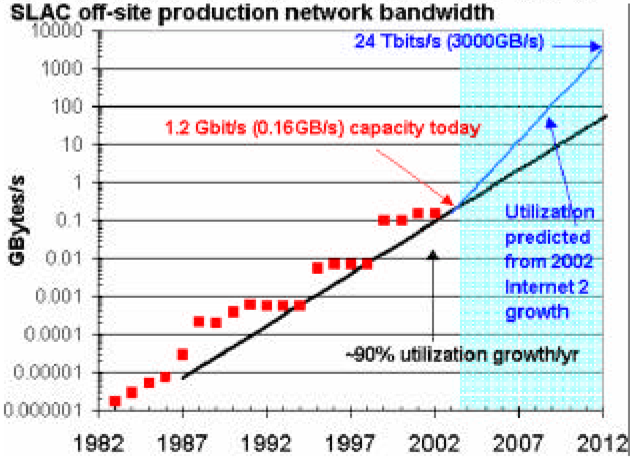
    - Will sharing movies be like sharing music today?

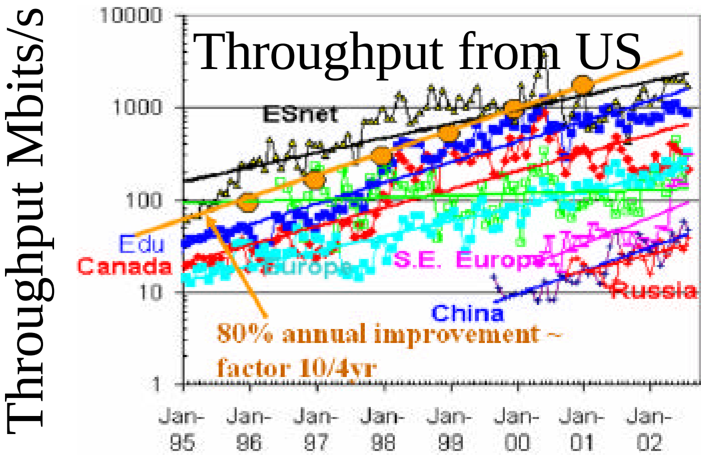


When will it have an impact

- ESnet traffic doubling/year since 1990
- SLAC capacity increasing by 90%/year since 1982
  - SLAC Internet traffic increased by factor 2.5 in last year
- International throughput increase by factor 10 in 4 years
- So traffic increases by factor 10 in 3.5 to 4 years, so in:
  - 3.5 to 5 years 622 Mbps => 10Gbps
  - 3-4 years 155 Mbps => 1Gbps
  - 3.5-5 years 45Mbps => 622Mbps
- 2010-2012:
  - 100s Gbits for high speed production net end connections
  - 10Gbps will be mundane for R&E and business
  - Home: doubling ~ every 2 years, 100Mbits/s by end of decade?







Impact

- Caught technical press attention
  - On TechTV and ABC Radio
  - Reported in places such as CNN, the BBC, Times of India, Wired, Nature
  - Reported in English, Spanish, Portuguese, French, Dutch, Japanese

abc RADIO NETWORKS
america listens to abc

techtv

BBCi

CNN

nature

WIRED

NEWSFACTOR™ NETWORK
NEWSFACTOR.com

InfoWorld

Slashdot

THE TIMES OF INDIA

What’s next?

- Break 2.5Gbits/s limit
- Disk-to-disk throughput & useful applications
  - Need faster cpus (extra 60% MHz/Mbits/s over TCP for disk to disk), understand how to use multi-processors
- Evaluate new stacks with real-world links, and other equipment
  - Other NICs
  - Response to congestion, pathologies
  - Fairnesss
  - Deploy for some major (e.g. HENP/Grid) customer applications
- Understand how to make 10GE NICs work well with 1500B MTUs
- **Move from “hero” demonstrations to commonplace**

More Information

- Internet2 Land Speed Record Publicity
  - www-iepm.slac.stanford.edu/lsr/
  - www-iepm.slac.stanford.edu/lsr2/
- 10GE tests
  - www-iepm.slac.stanford.edu/monitoring/bulk/10ge/
  - sravot.home.cern.ch/sravot/Networking/10GbE/10GbE_test.html