High Bandwidth High Throughput in the MB-NG & DataTAG Projects

Richard Hughes-Jones, Stephen Dallison, Gareth Fairey
Dept. of Physics and Astronomy, University of Manchester

Robin Tasker
Daresbury Laboratory CLRC

Miguel Rio, Yee Ting Li
Dept. of Physics and Astronomy, University College London
Topography of the MB – NG Network

**Topography of the MB – NG Network**

**Manchester Domain**
- man01
- man02
- Edge Router Cisco 7609

**Boundary Router Cisco 7609**

**UKERNA Development Network**

**UCL Domain**
- lon01
- lon02
- lon03

**RAL Domain**

**Key**
- Red: Gigabit Ethernet
- Blue: 2.5 Gbit POS Access
- Green: MPLS Admin. Domains

---

e-Science All Hands Meeting 1-4 Sep 03
R. Hughes-Jones  Manchester
End Hosts
how good are they really?
End Hosts b2b & end-to-end UDP Tests

- Test with UDPmon
  Supermicro P4DP6

- Max throughput 975Mbit/s
- 20% CPU utilisation receiver packets > 1000 bytes
- 40% CPU utilisation smaller packets

- PCI: 64 bit 66 MHz
- Latency 6.1ms & well behaved
- Latency Slope $0.0761 \, \mu s/byte$
- B2B Expect: $0.0118 \, \mu s/byte$
  - PCI 0.00188
  - GigE 0.008
  - PCI 0.00188

- 6 routers

- Jitter small 2-3 $\mu$s FWHM
Signals on the PCI bus

- 1472 byte packets every 15 µs Intel Pro/1000

- PCI:64 bit 33 MHz
  - 82% usage
  - Data Transfers
    - Send setup
    - Send PCI
    - Receive PCI
    - Receive Transfers

- PCI:64 bit 66 MHz
  - 65% usage
  - Data transfers half as long
  - Data Transfers
    - Send setup
    - Send PCI
    - Receive PCI
    - Receive Transfers
Interrupt Coalescence Investigations

- Kernel parameters for Socket Buffer size rtt*BW
  - TCP mem-mem lon2-man1
  - Tx 64 Tx-abs 64
  - Rx 0 Rx-abs 128
  - 820-980 Mbit/s +- 50 Mbit/s

- Tx 64 Tx-abs 64
  - Rx 20 Rx-abs 128
  - 937-940 Mbit/s +- 1.5 Mbit/s

- Tx 64 Tx-abs 64
  - Rx 80 Rx-abs 128
  - 937-939 Mbit/s +- 1 Mbit/s
- Tx Queue located between IP stack & NIC driver
- TCP treats ‘Queue full’ as congestion!
- Results for Lon → Man
- Select txqueue.len = 2000
Network Investigations
Network Bottlenecks

- Backbones 2.5 and 10 Gbit – usually good (in Europe)
- Access links need care GEANT-NRN and Campus – SuperJANET4
- NNW – SJ4 Access: given as example of good forward planning:

10 November 2002
  1 Gbit link

24 February 2003
  26 Feb 2003
  Upgraded to 2.5 Gbit

- Trunking – use of multiple 1 Gbit Ethernet links
24 Hours HighSpeed TCP mem-mem

- TCP mem-mem lon2-man
- Tx 64 Tx-abs 64
- Rx 64 Rx-abs 128
- 941.5 Mbit/s +- 0.5 Mbit/s
TCP sharing man1-lon2

- 1 stream every 60 s:
  - man1 → lon2
  - man2 → lon2
  - man3 → lon2
- Sample every 10ms

- 1 Stream:
  - Average 940 Mbit/s
  - No Dup ACKs
  - No SACKs
  - No Sendstalls

- 2 Streams:
  - Average ~500 Mbit/s
  - Many Dup ACKs
  - Cwnd reduced

- 2 Streams:
  - Average ~300 Mbit/s
2 Streams:

- Dips in throughput due to Dup ACK
- ~4 losses /sec
- A bit regular?

Cwnd decreases:
- 1 point 33%
- Ramp starts at 62%
- Slope 70Bytes/us
TCP Protocol Stack Comparisons

- **Standard TCP**
- **HighSpeed TCP**
- **Scalable TCP**

\[ \rho = \frac{C \times RTT}{2 \times MSS}^2 \]

- Kernel on the receiver dropped packets periodically
- **MB-NG Network** rtt 6.2 ms. 
  Recovery time 1.6s
- **DataTAG Network** rtt 119 ms.
  Recovery time 590s 9.8 min
- **Throughput of the DataTAG network** was factor \(~5\) lower than that on the MB-NG network
Application Throughput
MB – NG SuperJANET4
Development Network

Gigabit Ethernet
2.5 Gbit POS Access
2.5 Gbit POS core
MPLS Admin. Domains
Gridftp Throughput HighSpeedTCP

- RAID0 Disk Tests:
  - 120 Mbytes/s Read
  - 100 Mbytes/s Write

- Int Coal 64 128
- Txqueuelen 2000
- TCP buffer 1 M byte
  (rtt*BW = 750kbytes)

- Interface throughput

- Data Rate: 520 Mbit/s

- Same for B2B tests

- So its not that simple!
Gridftp Throughput + Web100

- Throughput Mbit/s:
- See alternate 600/800 Mbit and zero
- Cwnd smooth
- No dup Ack / send stall / timeouts
http data transfers HighSpeed TCP

- Bulk data moved by web servers
- Apache web server out of the box!
- Prototype client - curl http library
- 1Mbyte TCP buffers
- 2Gbyte file
- Throughput ~720 Mbit/s
- Cwnd - some variation
- No dup Ack / send stall / timeouts
BaBar Case Study: Disk Performance

- BaBar Disk Server
  - Tyan Tiger S2466N motherboard
  - 1 64bit 66 MHz PCI bus
  - Athlon MP2000+ CPU
  - AMD-760 MPX chipset
  - 3Ware 7500-8 RAID5
  - 8 * 200Gb Maxtor IDE 7200rpm disks
- Note the VM parameter readahead max

- Disk to memory (read)
  Max throughput 1.2 Gbit/s 150 MBytes/s

- Memory to disk (write)
  Max throughput 400 Mbit/s 50 MBytes/s
  [not as fast as Raid0]
BaBar Case Study: Throughput & PCI Activity

- 3Ware forces PCI bus to 33 MHz
- BaBar Tyan to MB-NG SuperMicro
  Network mem-mem 619 Mbit/s
- Disk – disk throughput bbcp
  40-45 Mbytes/s (320 – 360 Mbit/s)
- PCI bus effectively full!

Read from RAID5 Disks

Write to RAID5 Disks
Conclusions

The MB-NG Project has achieved:

- Continuous memory to memory data transfers with an average user data rate of 940 Mbit/s for over 24 hours using the HighSpeed TCP stack.
- Sustained high throughput data transfers of 2 GByte files between RAID0 disk systems using Gridftp and bbcp.
- Transfers of 2 GByte files using the http protocol from the standard apache Web server and HighSpeed TCP that achieved data rates of ~725 Mbit/s.
- Ongoing operation and comparison of different Transport Protocols - Optical Switched Networks
- Detailed investigation of Routers, NICs & end-host performance.
- Working with e-Science groups to get high performance to the user.

- Sustained data flows at Gigabit rates are achievable
- Use Server quality PCs not Supermarket PCs + care with interfaces
- Be kind to the Wizards!
More Information Some URLs

- MB-NG project web site: http://www.mb-ng.net/
- DataTAG project web site: http://www.datatag.org/
- UDPmon / TCPmon kit + writeup:
  http://www.hep.man.ac.uk/~rich/net
- Motherboard and NIC Tests:
  www.hep.man.ac.uk/~rich/net/nic/GigEth_tests_Boston.ppt
- TCP tuning information may be found at:
  http://www.ncne.nlanr.net/documentation/faq/performance.html
  & http://www.psc.edu/networking/perf_tune.html
Backup Slides
EU Review Demo Consisted of:

- Raid0 Disk
- GridFTP
- Data over TCP Streams
- GridFTP
- Raid0 Disk

Node Monitoring
Site Monitoring
Dante Monitoring
Throughput on the day!

TCP ACKs

Data
~400 Mbit/s

EDG Wp7 Network Throughput Monitor

Connect To: Status: Connected to 192.91.244.12

Time Series Plot for MBits in

Time Series Plot for MBits out
Some Measurements of Throughput CERN-SARA

◆ Using the GÉANT Backup Link
  - 1 GByte file transfers
  - Blue Data
  - Red TCP ACKs

◆ Standard TCP
  - Average Throughput 167 Mbit/s
  - Users see 5 - 50 Mbit/s!

◆ High-Speed TCP
  - Average Throughput 345 Mbit/s

◆ Scalable TCP
  - Average Throughput 340 Mbit/s
What the Users Really find:

◆ CERN – RAL using production GÉANT
  ▪ CMS Tests 8 streams
  ▪ 50 Mbit/s @ 15 MB buffer
  ▪ Firewall 100 Mbit/s

◆ NNW – SJ4 Access
  ▪ 1 Gbit link