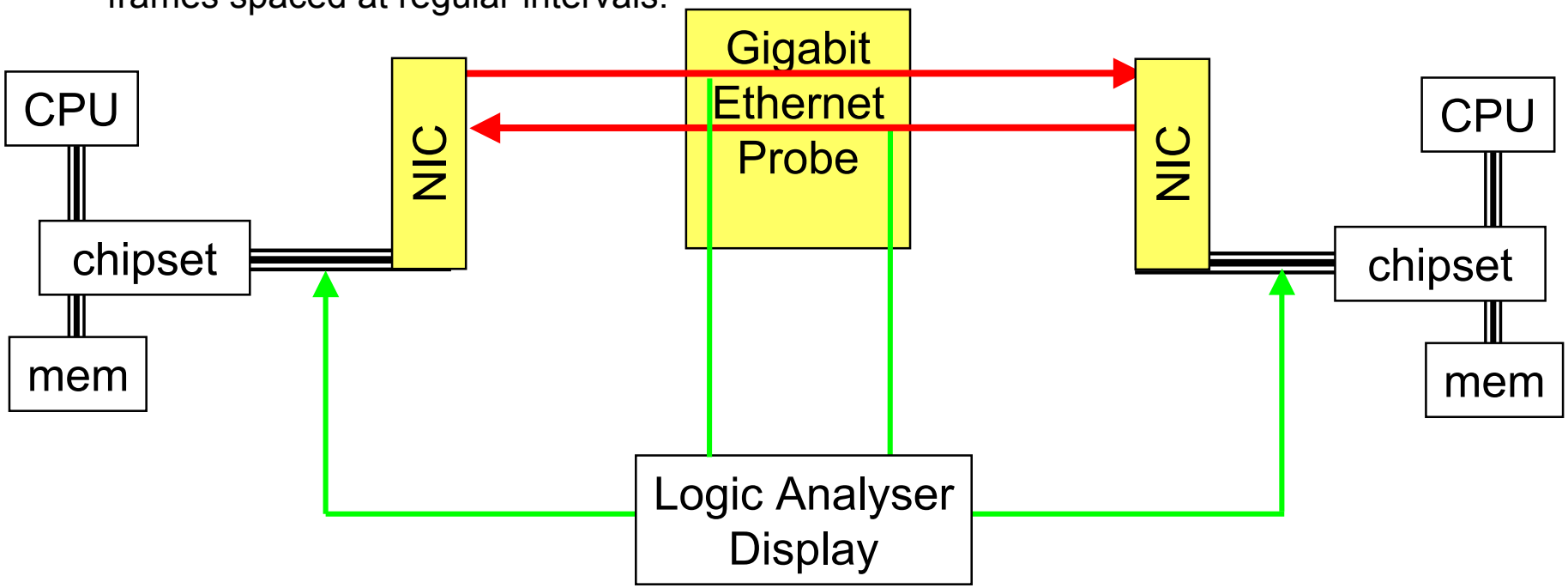
High Data Rate Transmission for vlbiGRID using the Academic Network

Related projects include:
MB-NG, DataGrid, DataTAG, UKLIGHT

Richard Hughes-Jones, Ralph Spencer, Steve Parsley
The University of Manchester & JIVE

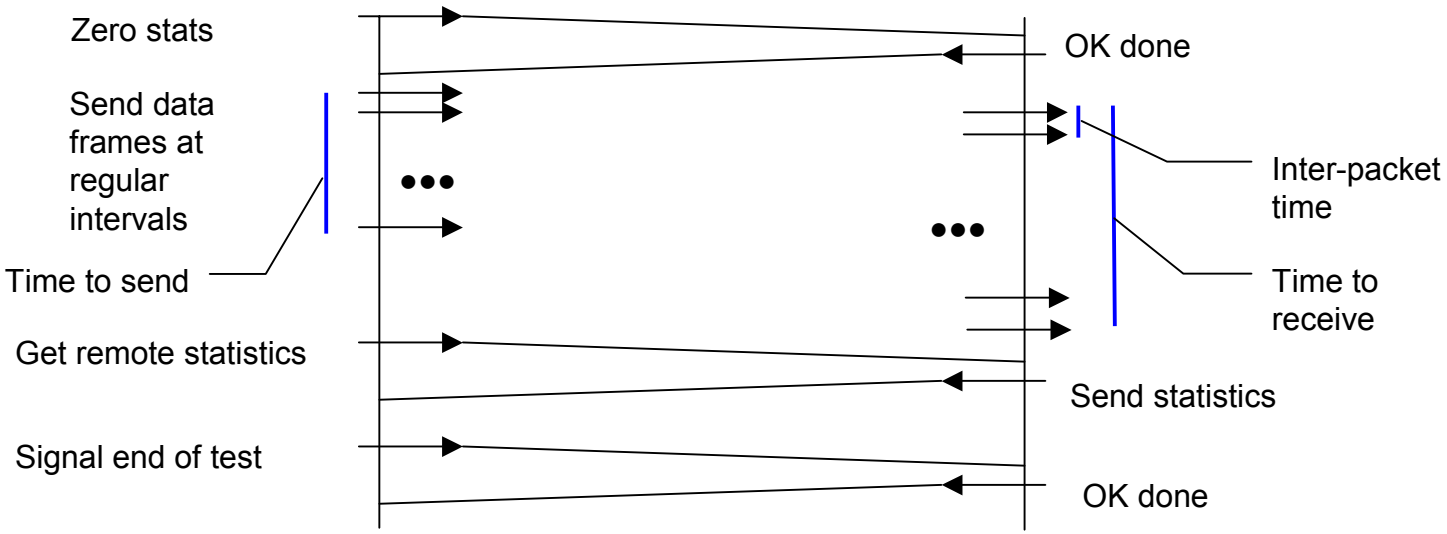

Evaluating Motherboards & Gigabit NICs

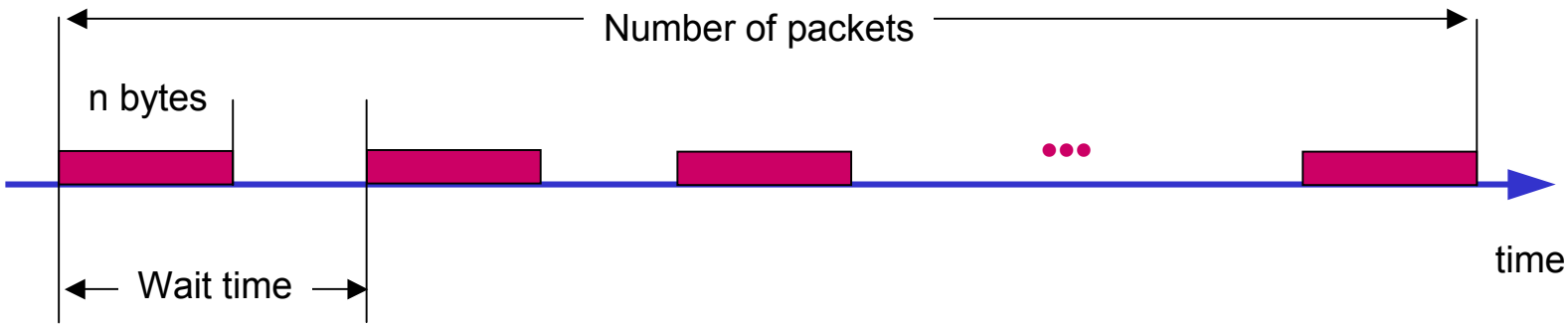
- **Latency**
- Round trip times with UDP frames
  - Slope gives sum of individual data transfers end-to-end
  - Histograms
- **UDP Throughput – Capacity & Available**
  - Send a controlled stream of UDP frames spaced at regular intervals.
- **PCI Activity**
- Logic Analyzer with
  - PCI Probe cards in sending PC
  - Gigabit Ethernet Fiber Probe Card
  - PCI Probe cards in receiving PC



UDPmon: The Works

- **How it works:**
  - **The 'Zero_stats' request also provides an interlock against concurrent tests.**



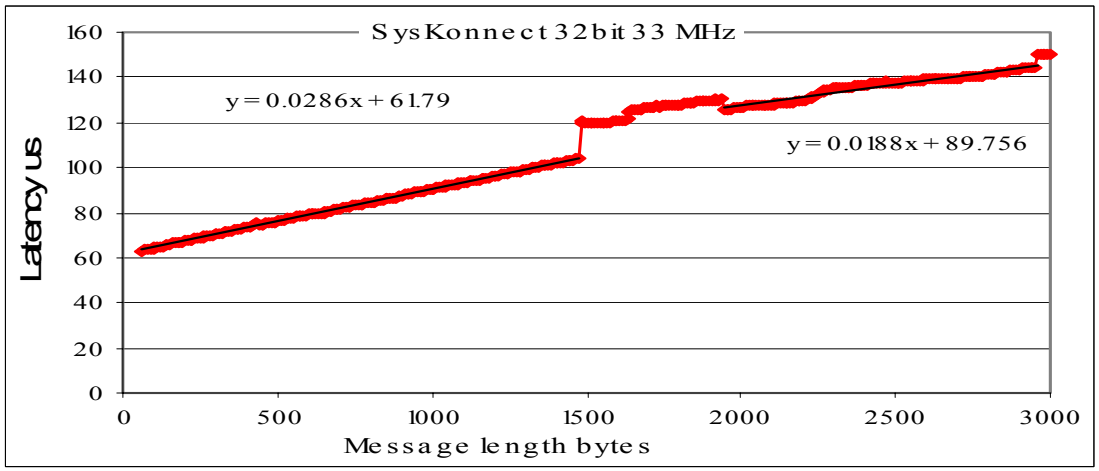


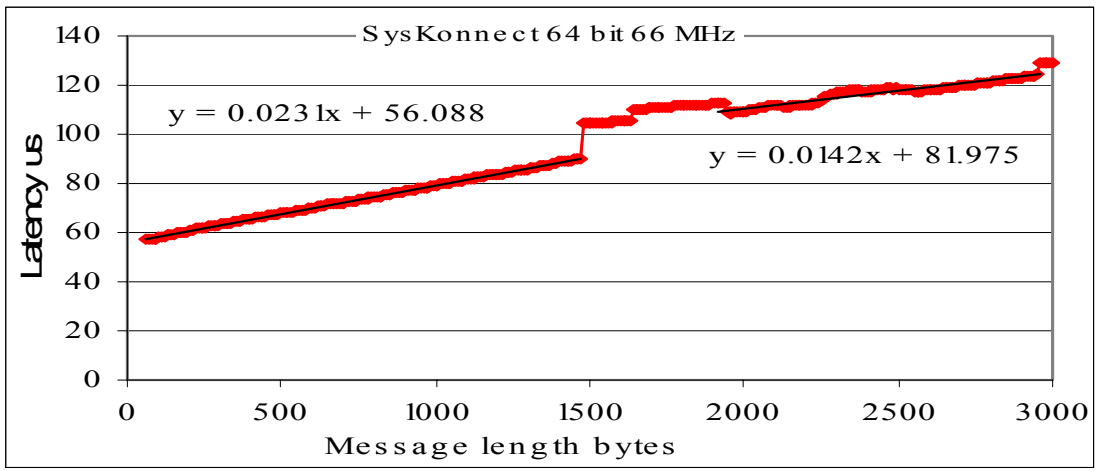
SuperMicro 370DLE: Latency: SysKonnect

- Motherboard: SuperMicro 370DLE Chipset: ServerWorks III LE Chipset
- CPU: PIII 800 MHz
- RedHat 7.1 Kernel 2.4.14

- **PCI:32 bit 33 MHz**
- Latency small 62 µs & well behaved
- Latency Slope **0.0286 µs/byte**
- Expect: **0.0232 µs/byte**
  - PCI 0.00758
  - GigE 0.008
  - PCI 0.00758

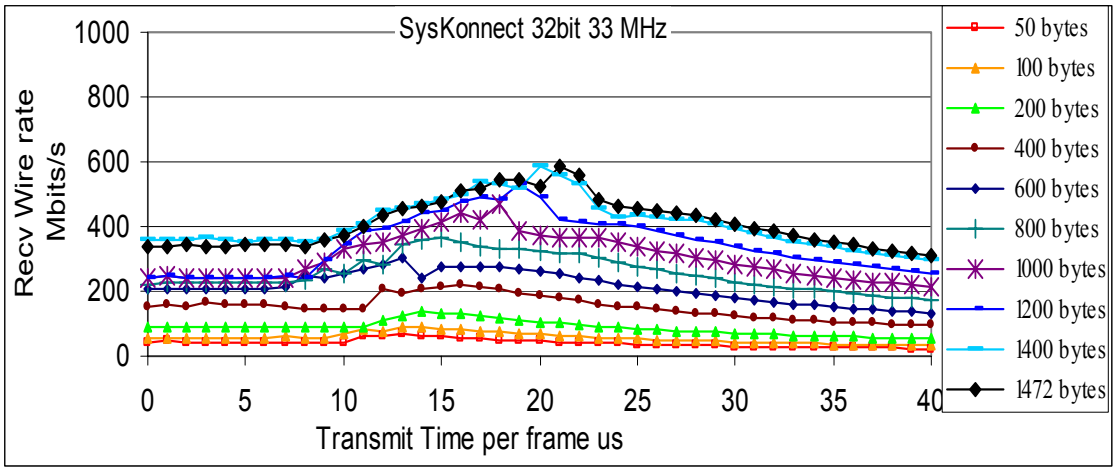
- **PCI:64 bit 66 MHz**
- Latency small 56 µs & well behaved
- Latency Slope **0.0231 µs/byte**
- Expect: **0.0118 µs/byte**
  - PCI 0.00188
  - GigE 0.008
  - PCI 0.00188
- Possible extra data moves ?



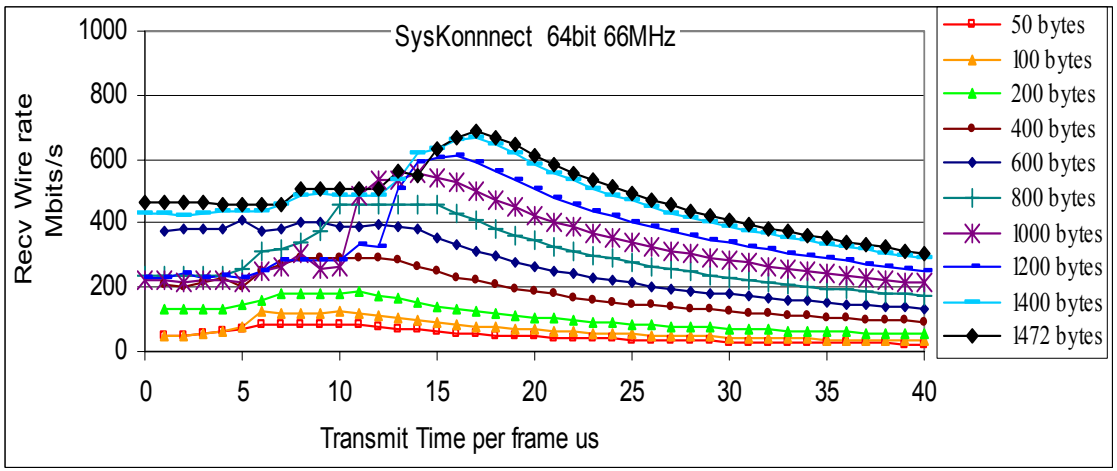


SuperMicro 370DLE: Throughput: SysKonnect

- Motherboard: SuperMicro 370DLE Chipset: ServerWorks III LE Chipset
- CPU: PIII 800 MHz
- RedHat 7.1 Kernel 2.4.14
- **PCI:32 bit 33 MHz**
- Max throughput **584Mbit/s**
- No packet loss >18 us spacing



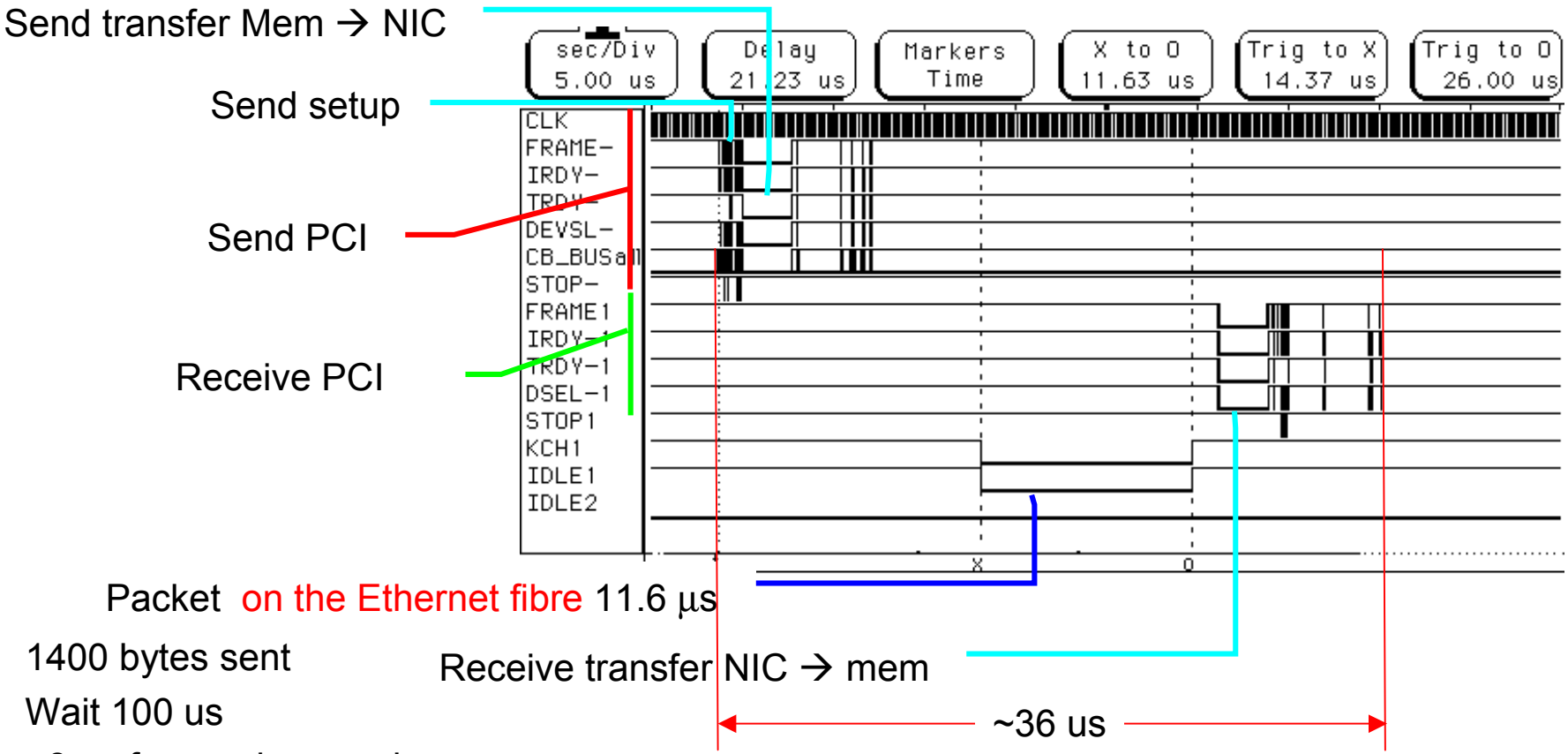
- **PCI:64 bit 66 MHz**
- Max throughput **720 Mbit/s**
- No packet loss >17 us spacing



- Packet loss during BW drop

SuperMicro 370DLE: PCI: SysKonnect

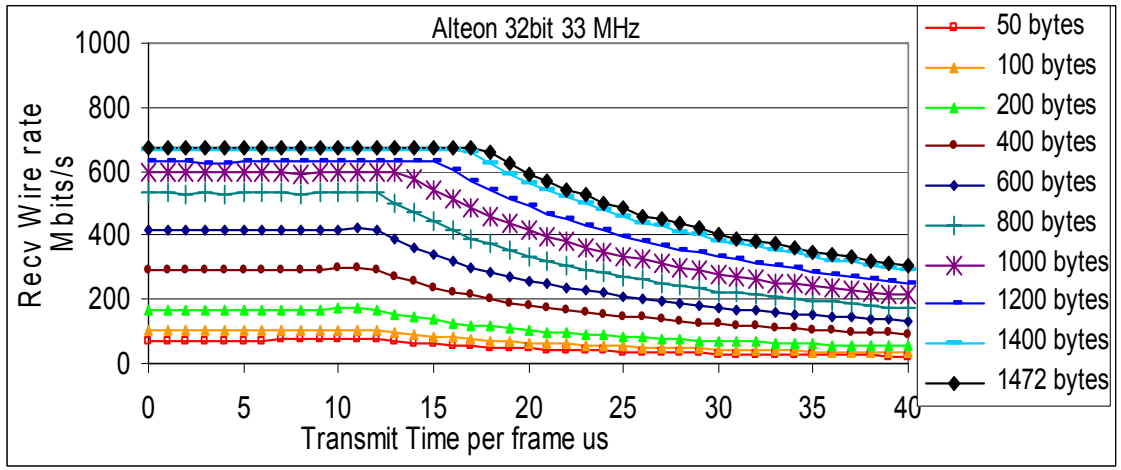
- Motherboard: SuperMicro 370DLE Chipset: ServerWorks III LE Chipset
- CPU: PIII 800 MHz **PCI:64 bit 66 MHz**
- RedHat 7.1 Kernel 2.4.14



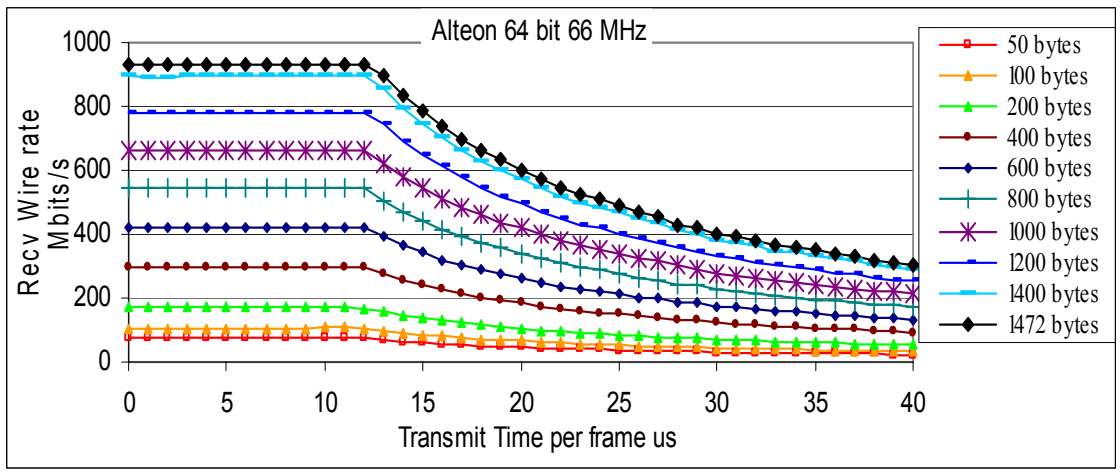
- 1400 bytes sent
- Wait 100 us
- ~8 us for send or receive
- Stack & Application overhead ~ 10 us / node

SuperMicro 370DLE: Throughput: Alteon

- Motherboard: SuperMicro 370DLE Chipset: ServerWorks III LE Chipset
- CPU: PIII 800 MHz
- RedHat 7.1 Kernel 2.4.14
- **PCI:64 bit 33 MHz**
- Max throughput **674Mbit/s**
- Packet loss < 10 us spacing



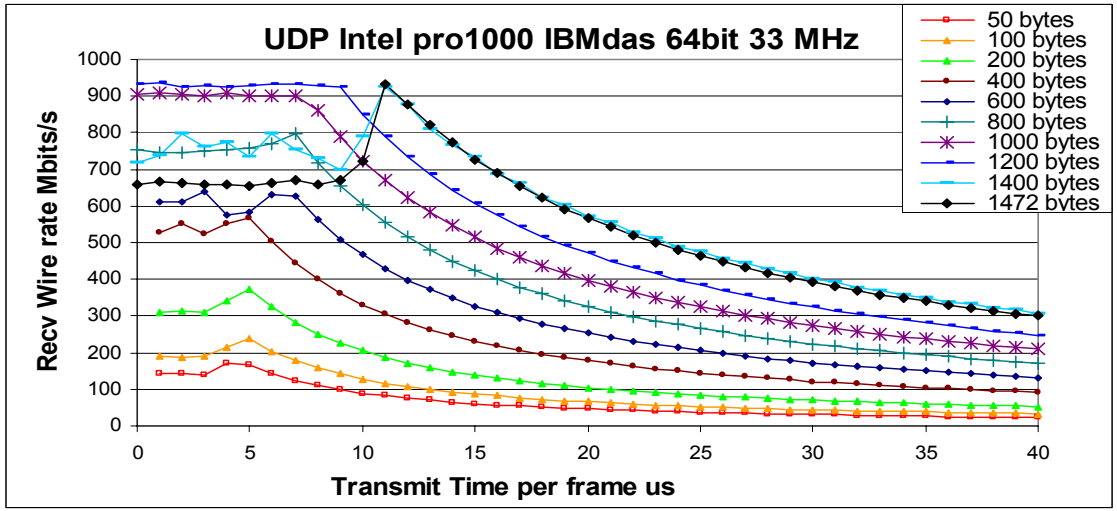
- **PCI:64 bit 66 MHz**
- Max throughput **930 Mbit/s**
- Packet loss < 10 us spacing



- Packet loss during BW drop

IBM das: Throughput: Intel Pro/1000

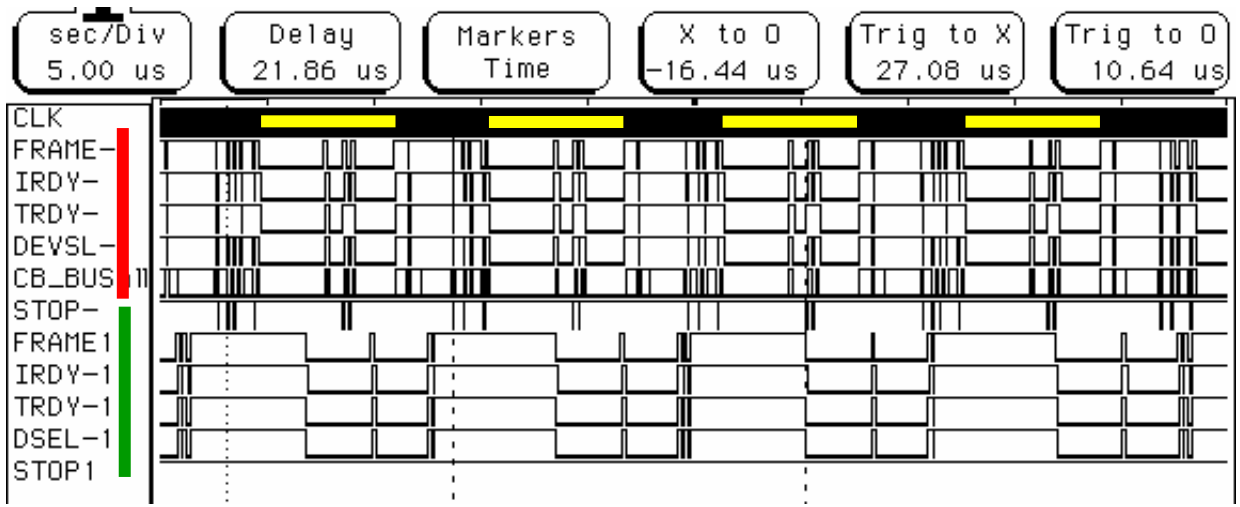
- Motherboard: IBM das Chipset:: ServerWorks CNB20LE
- CPU: Dual IBM 340 1GHz **PCI:64 bit 33 MHz**
- RedHat 7.1 Kernel 2.4.14



- Max throughput 930Mbit/s
- No packet loss > 12 us
- Clean behaviour

- 1400 bytes sent
- 11 us spacing
- Signals clean
- ~9.3us on send PCI bus
- **PCI bus ~82% occupancy**

- ~ 5.9 us on PCI for data recv.

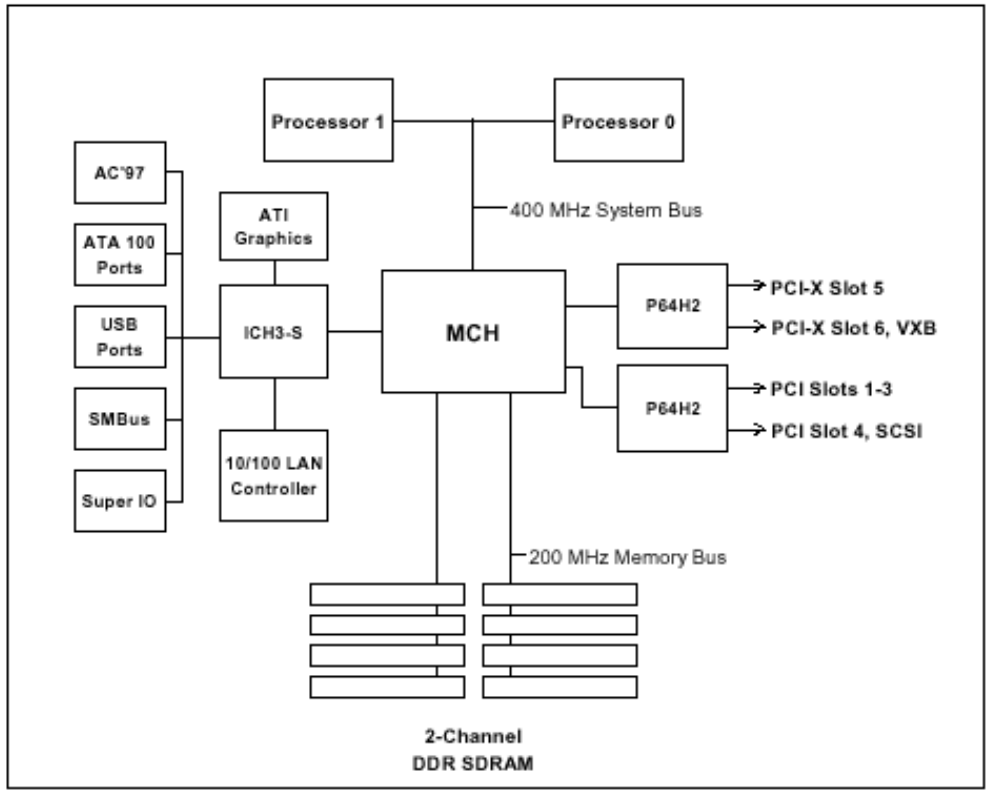


The SuperMicro P4DP6 Motherboard

- **Dual** Xeon Prestonia (2cpu/die)
- 400 MHx Front side bus

- Intel® E7500 Chipset
- **6 PCI-X slots**
- **4 independent PCI buses**
- Can select:
  - 64 bit 66 MHz PCI
  - 100 MHz PCI-X
  - 133 MHz PCI-X

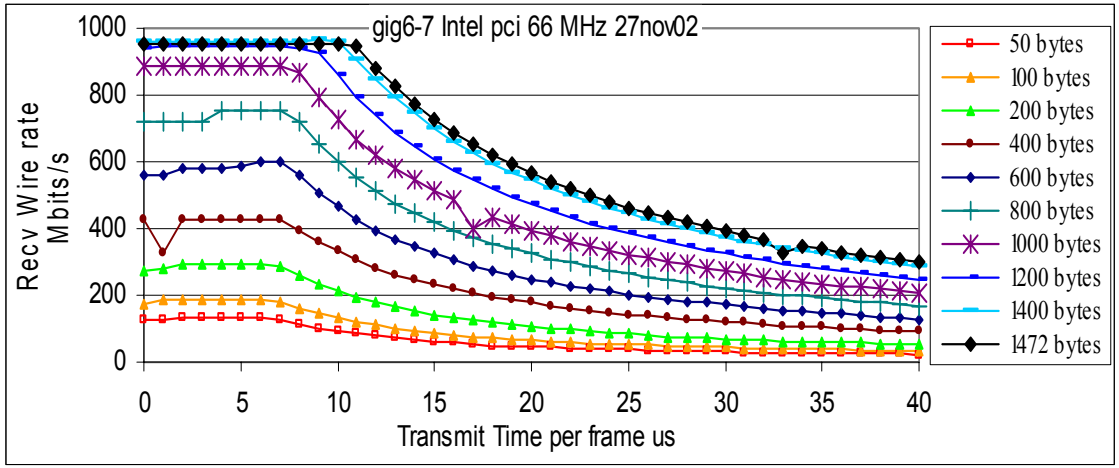
- 2 100 Mbit Ethernet
- Adaptec AIC-7899W dual channel SCSI
- UDMA/100 bus master/EIDE channels
  - data transfer rates of 100 MB/sec burst
- **P4DP8-2G dual Gigabit Ethernet**



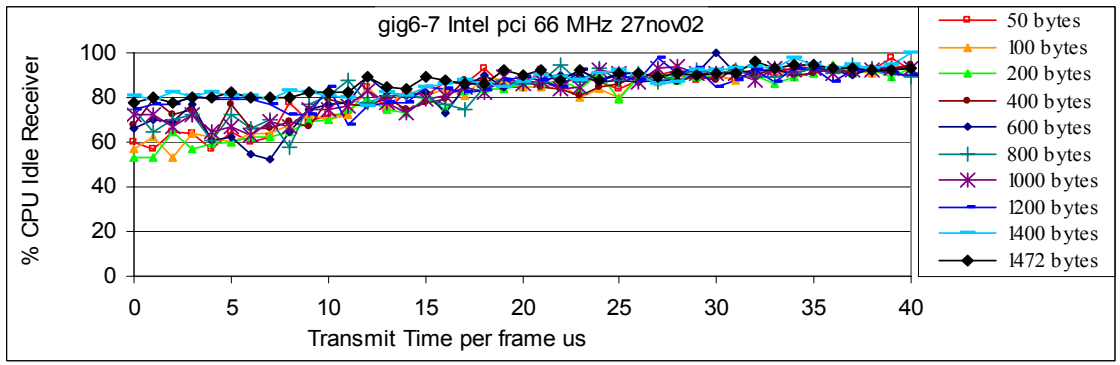
SuperMicro P4DP6: Throughput Intel Pro/1000

- Motherboard: SuperMicro P4DP6 Chipset: Intel E7500 (Plumas)
- CPU: Dual Xeon Prestonia 2.2 GHz **PCI, 64 bit, 66 MHz**
- RedHat 7.2 Kernel 2.4.19

- Max throughput **950Mbit/s**
- No packet loss



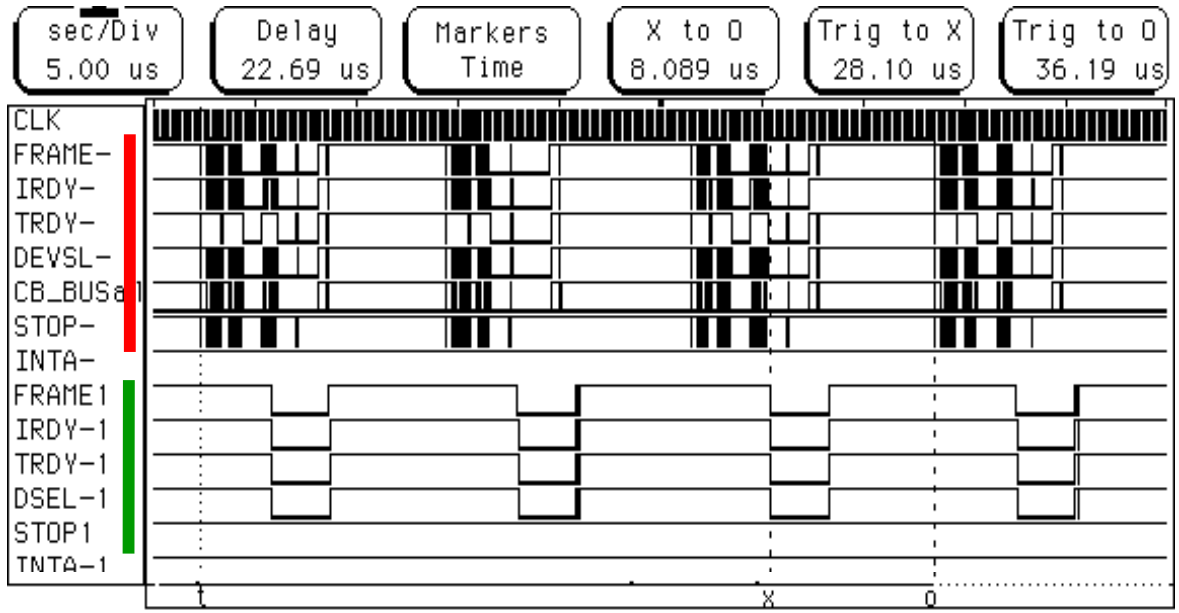
- CPU utilisation on the receiving PC was ~ 25 % for packets > than 1000 bytes
- 30- 40 % for smaller packets



SuperMicro P4DP6: PCI Intel Pro/1000

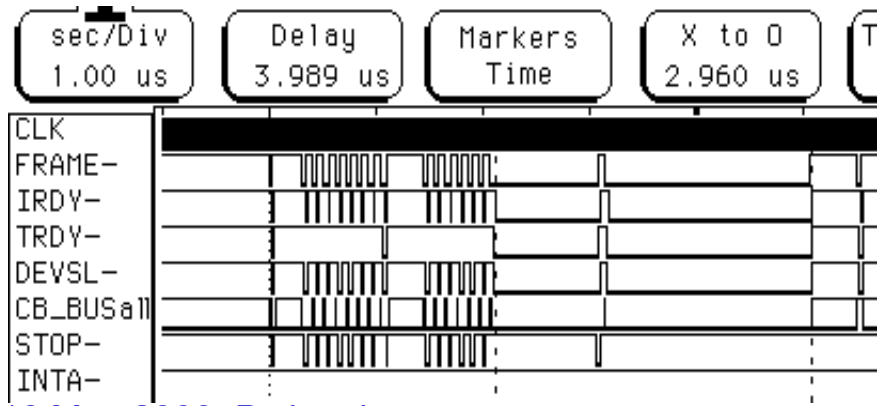
- Motherboard: SuperMicro P4DP6 Chipset: Intel E7500 (Plumas)
- CPU: Dual Xeon Prestonia 2.2 GHz **PCI, 64 bit, 66 MHz**
- RedHat 7.2 Kernel 2.4.19

- 1400 bytes sent
- Wait 12 us
- ~5.14us on send PCI bus
- PCI bus ~68% occupancy
- ~ 3 us on PCI for data recv



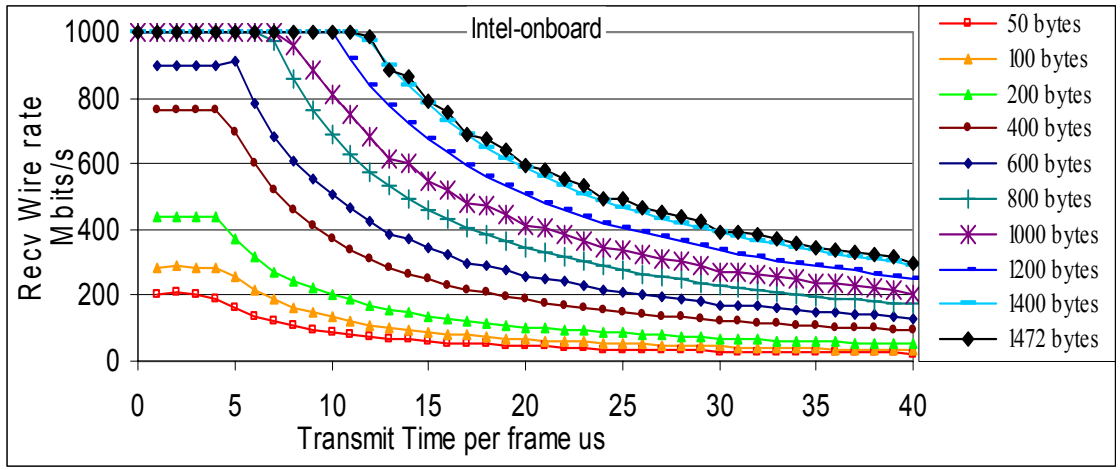
- CSR access inserts PCI STOPs
- NIC takes ~ 1 us/CSR
- CPU faster than the NIC !

- Similar effect with the SysKonnect NIC

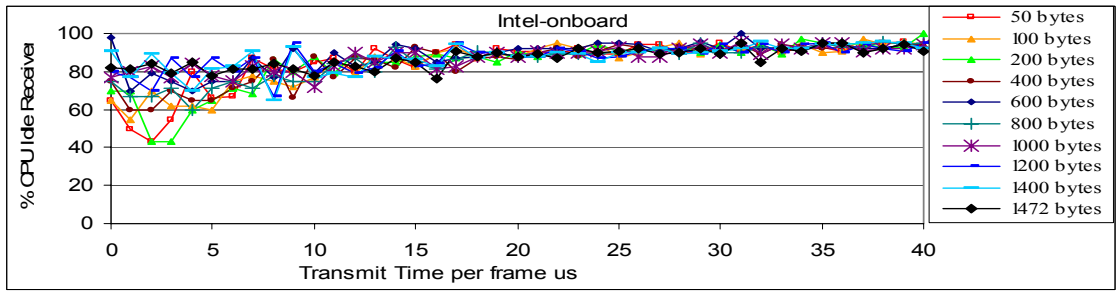


SuperMicro P4DP8-G2: Throughput Intel onboard

- Motherboard: SuperMicro P4DP8-G2 Chipset: Intel E7500 (Plumas)
- CPU: Dual Xeon Prestonia 2.4 GHz **PCI-X:64 bit**
- RedHat 7.3 Kernel 2.4.19



- Max throughput 995Mbit/s
- No packet loss



- 20% CPU utilisation receiver packets > 1000 bytes
- 30% CPU utilisation smaller packets

Summary on the Motherboards & NICs

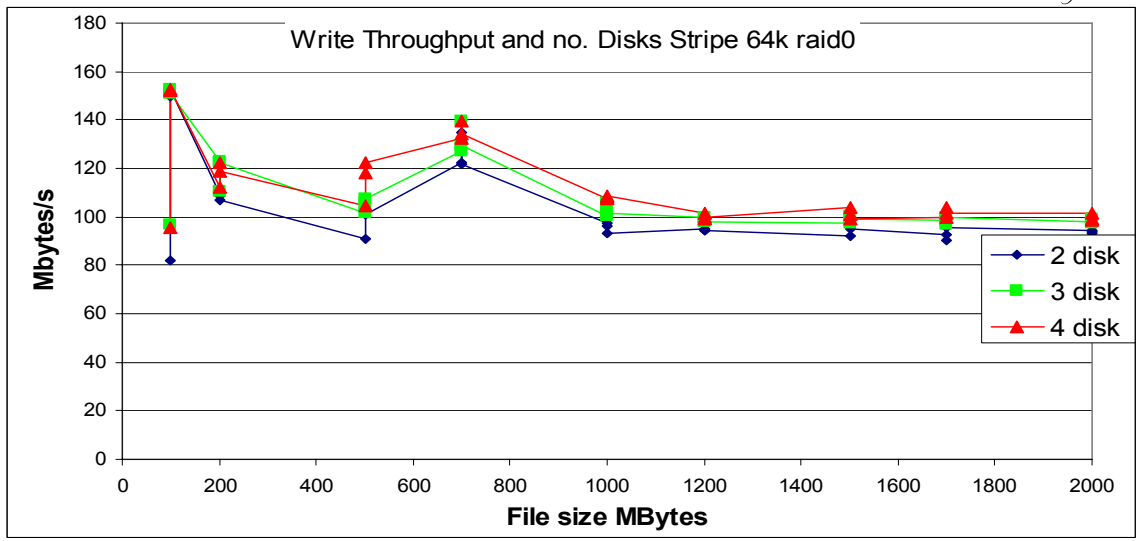
- All NICs & motherboards were stable – 1000s GBytes of transfers
- Alteon could handle 930 Mbit/s on 64bit/66MHz
- **SysKonnect** gave 720-876 Mbit/s improving to **876-990 Mbit/s on later boards**
- **Intel** gave 910 – 950 Mbit/s and **950-995 Mbit/s on later boards**

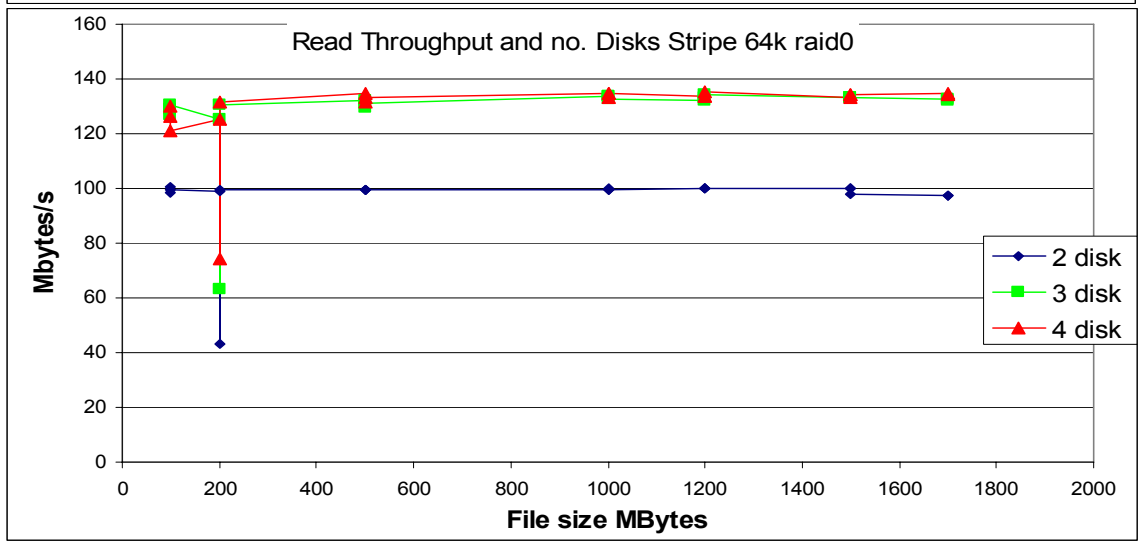
- PCI and GigEthernet signals show 800 MHz CPU can drive large packets at line speed.
- More CPU power is required for receiving – loss due to IP discards
  - **Rule of thumb at least 1 GHz CPU power free for 1 Gbit**

- Times for DMA transfers scale with PCI bus speed but CSR access is constant
  - New PCI-X and on-board controllers are better
- Buses: 64 bit 66MHz PCI or faster PCI-X are required for performance
  - **32 bit 33 MHz PCI bus is REALLY busy !!**
  - **64bit 33 MHz are > 80% used**

Raid0 Performance

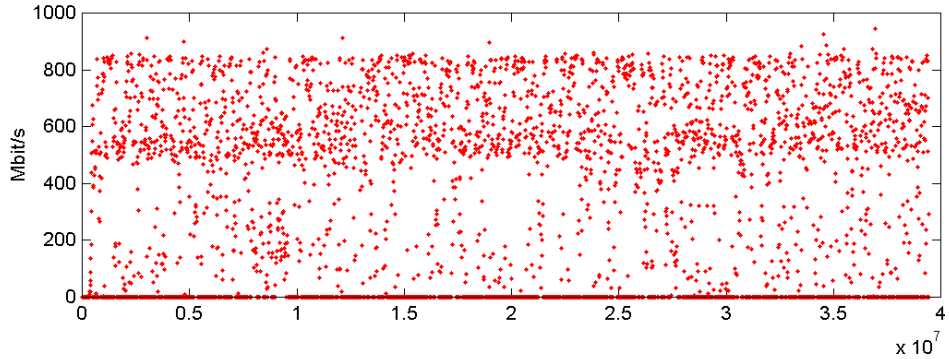
- Maxdor 3.5 Series DiamondMax PLus 9 120 Gb ATA/133
- 3ware 7500-8 controller
- Write
  Slight increase with number of disks
- Read
- 3 Disks OK
- Write 100 MBytes/s
- Read 130 MBytes/s

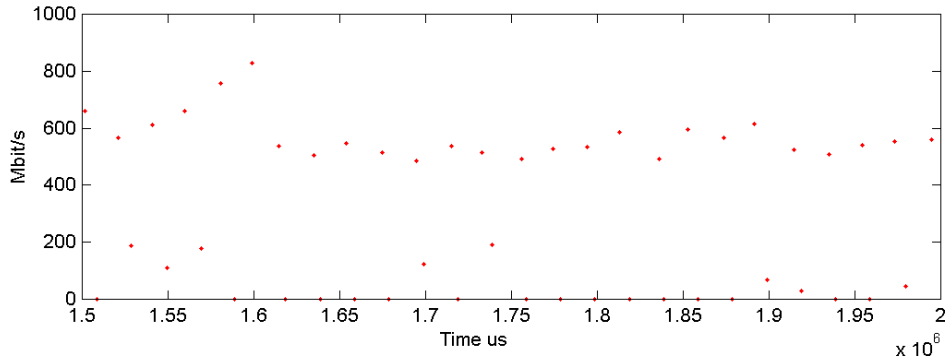


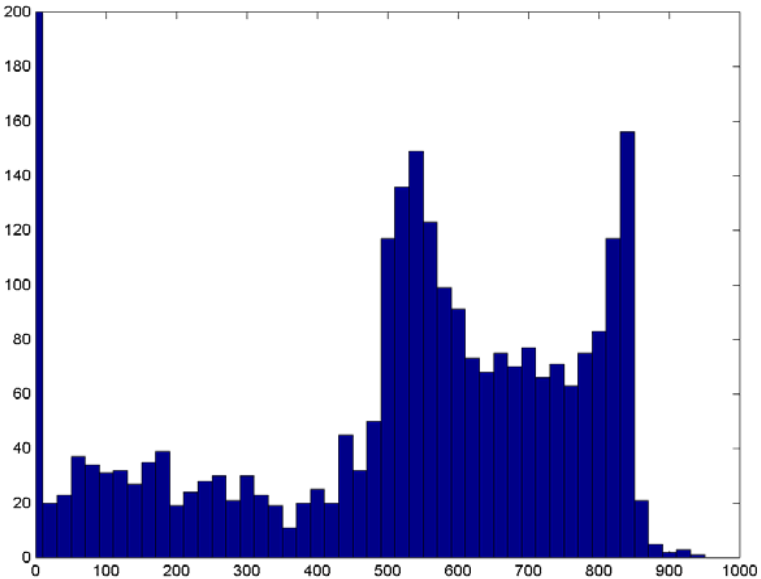


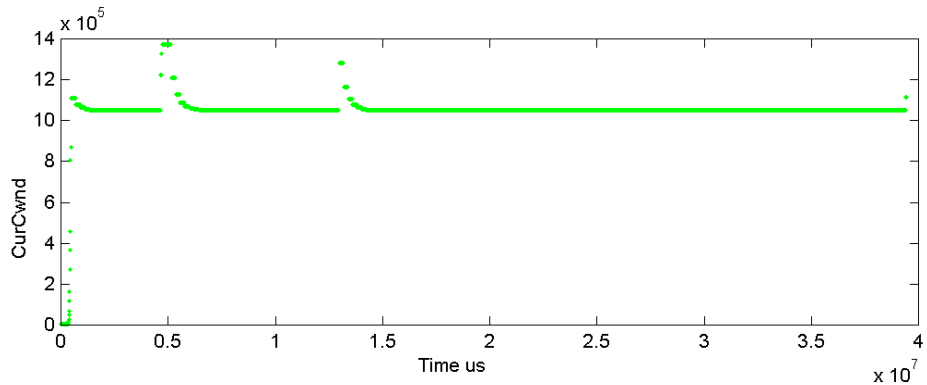
Gridftp Throughput + Web100

- Long term Wire rate: ~520Mbit/s:
- See alternate 600/800 Mbit and zero
- Cwnd smooth
- No dup Ack / send stall / timeouts



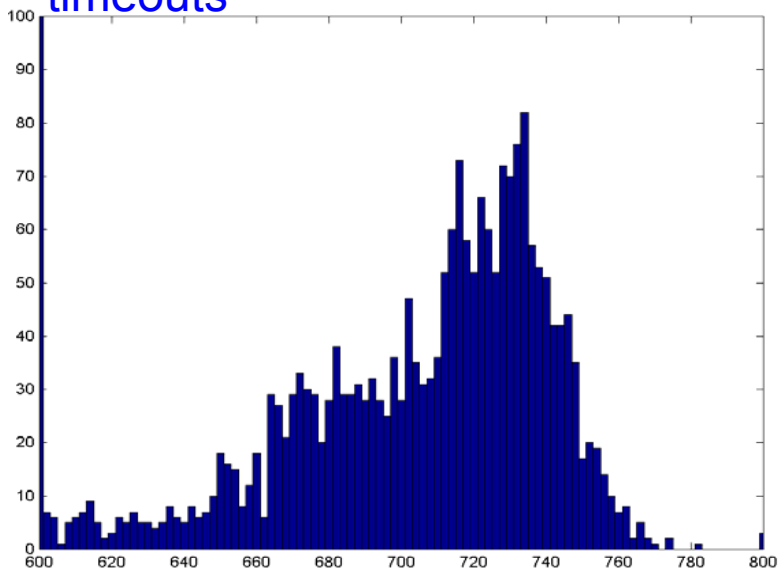


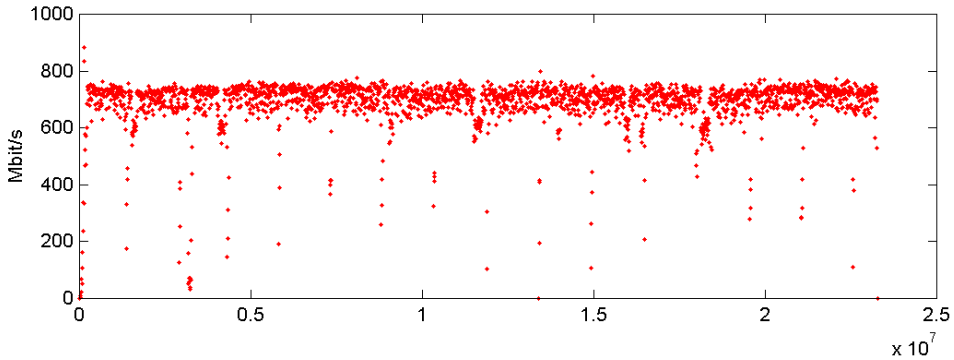


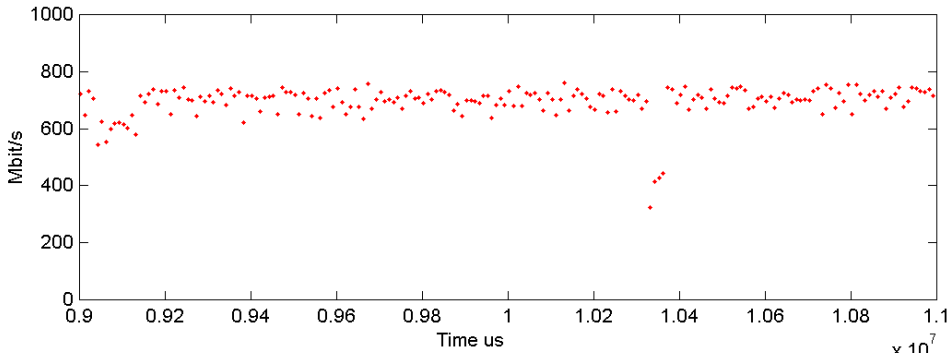


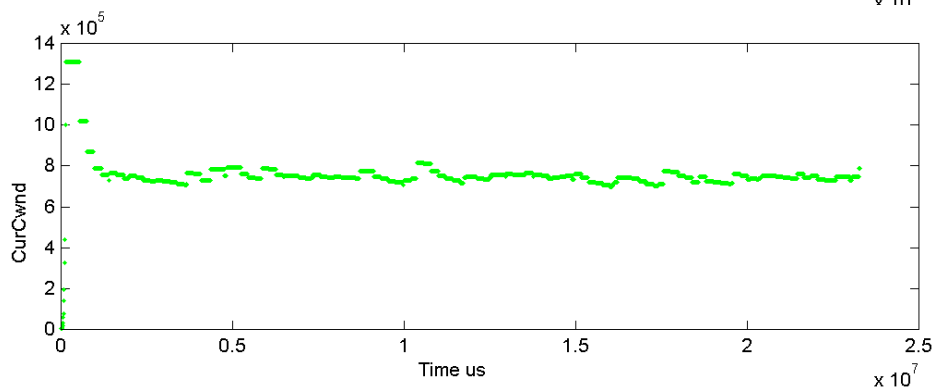
http data transfers HighSpeed TCP

- Apachie web server out of the box!
- prototype client - curl http library
- 1Mbyte TCP buffers
- Same HW & 2Gbyte file
- Throughput ~725 Mbits/s
- Cwnd - some variation
- No dup Ack / send stall / timeouts



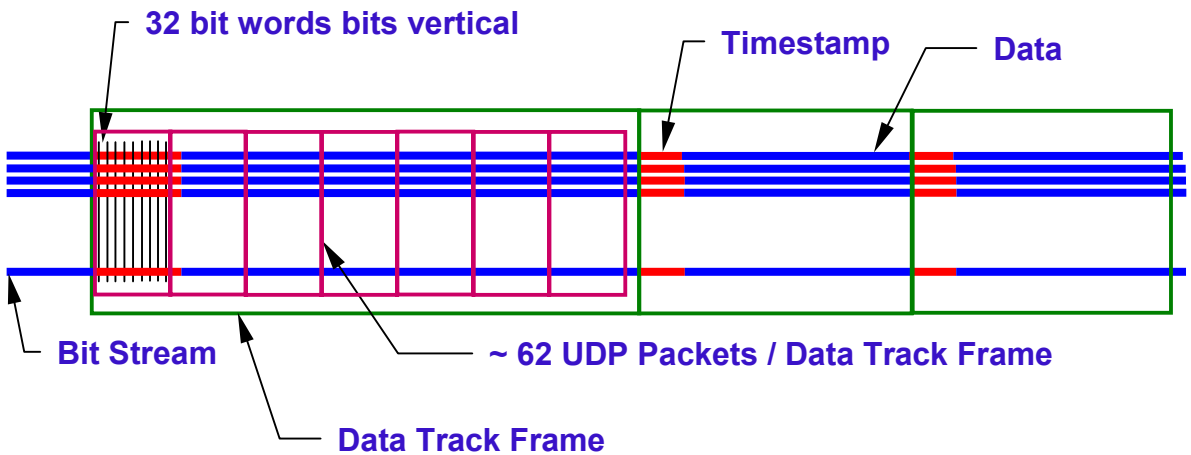






Radio Astronomy VLBI Demos at iGrid2002 and ER2002

- Web based demonstration sending VLBI data
  - A controlled stream of UDP/IP packets
  - 500 Mbit/s on the production network Manchester → Amsterdam

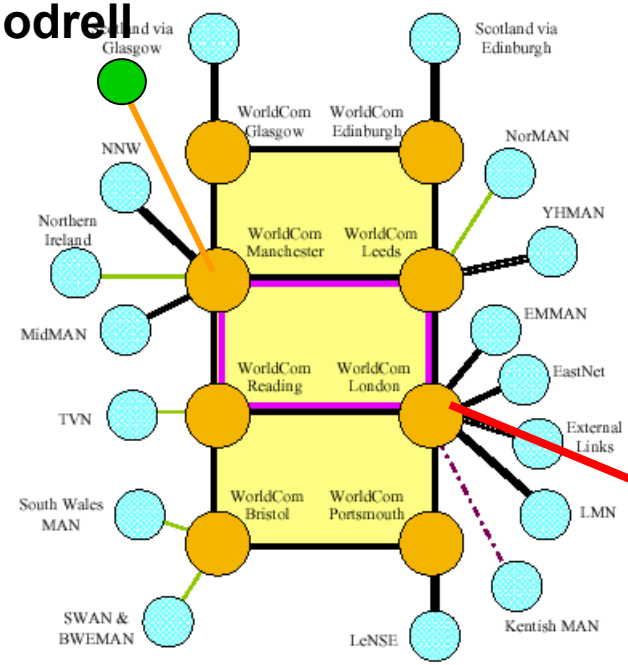


European Topology: NRNs, Geant, Sites

JIVE
JOINT INSTITUTE FOR VLBI IN EUROPE

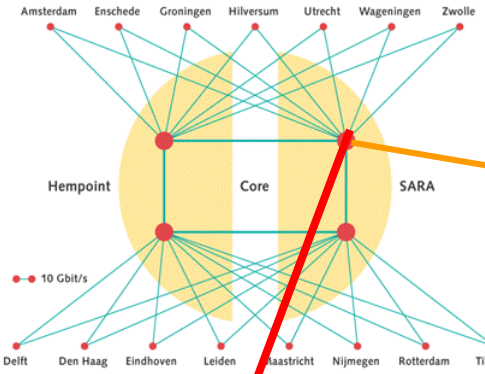
THE UNIVERSITY of MANCHESTER

Manchester
Jodrell



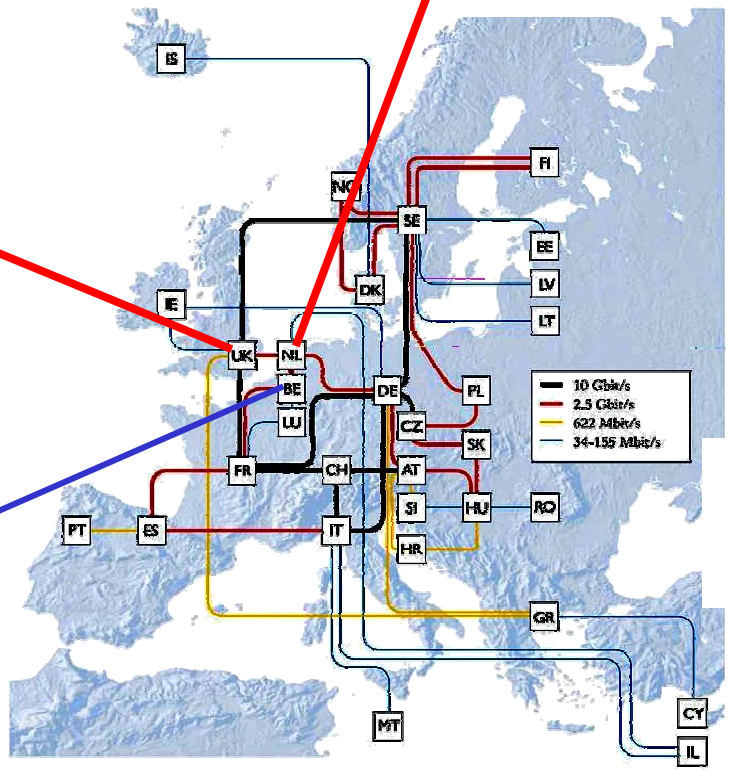
SuperJANET4

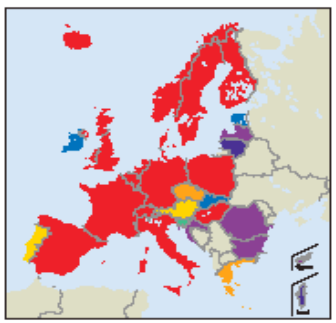
SURFnet



Amsterdam

Dwingeloo

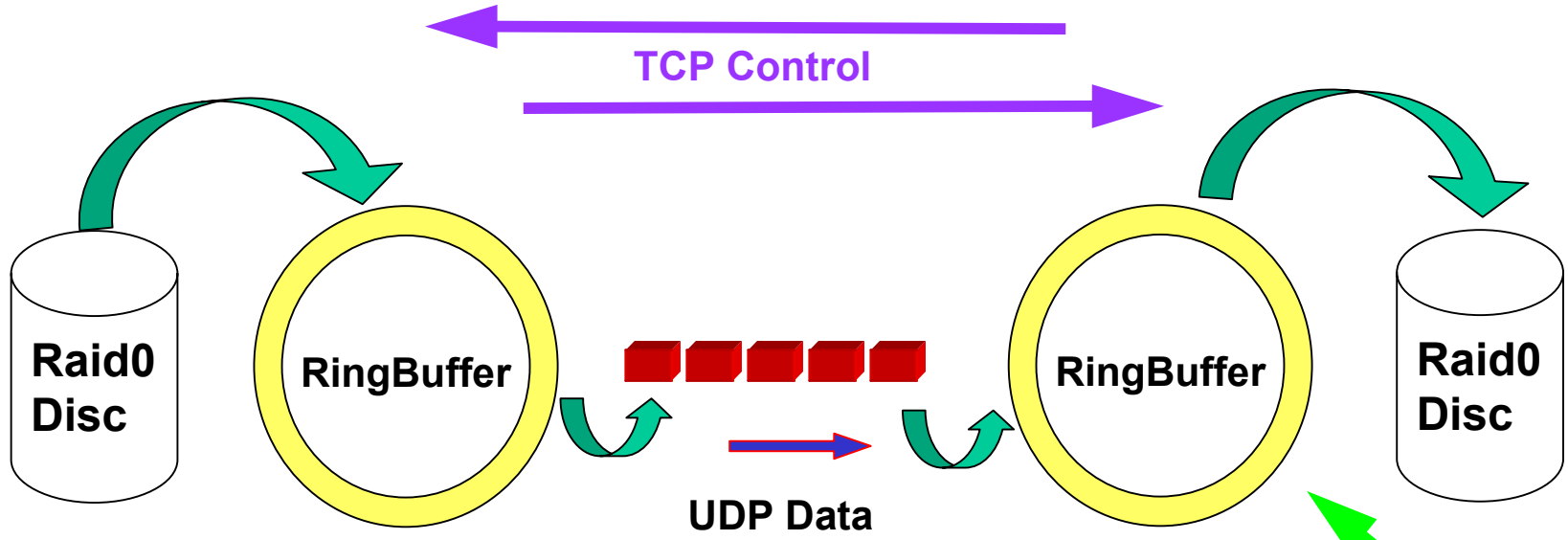




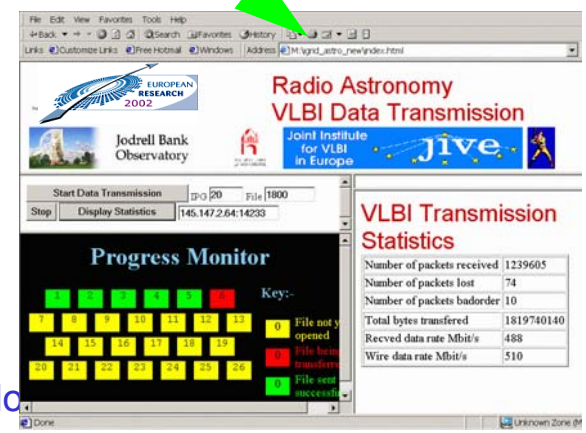
Brussels
ER 2002

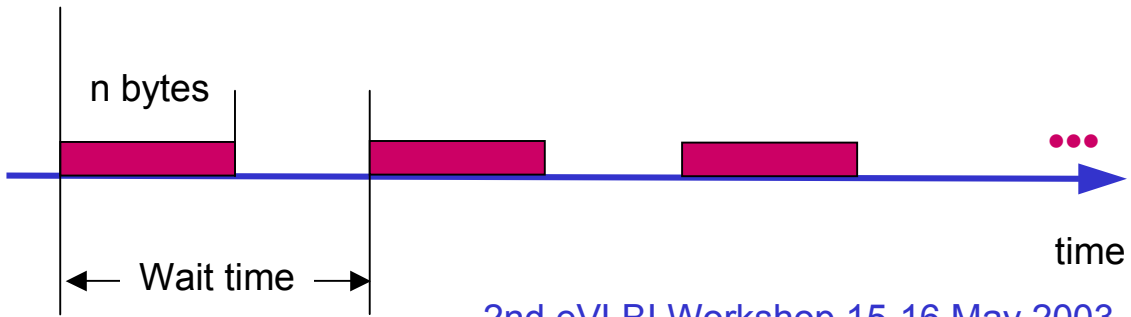
GÉANT

The Works:

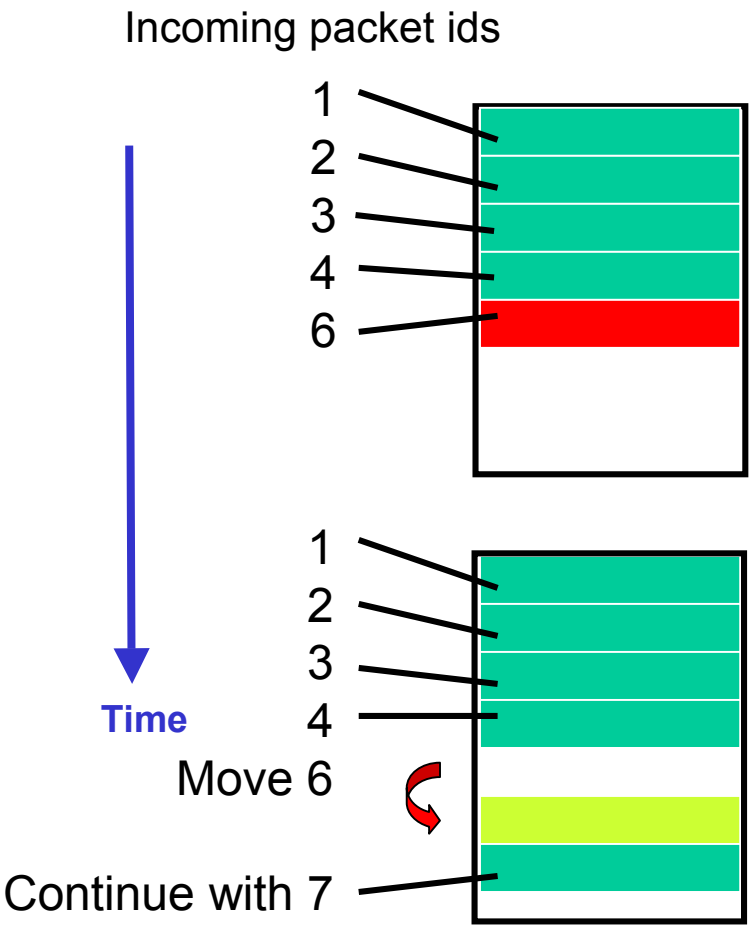


Web Interface



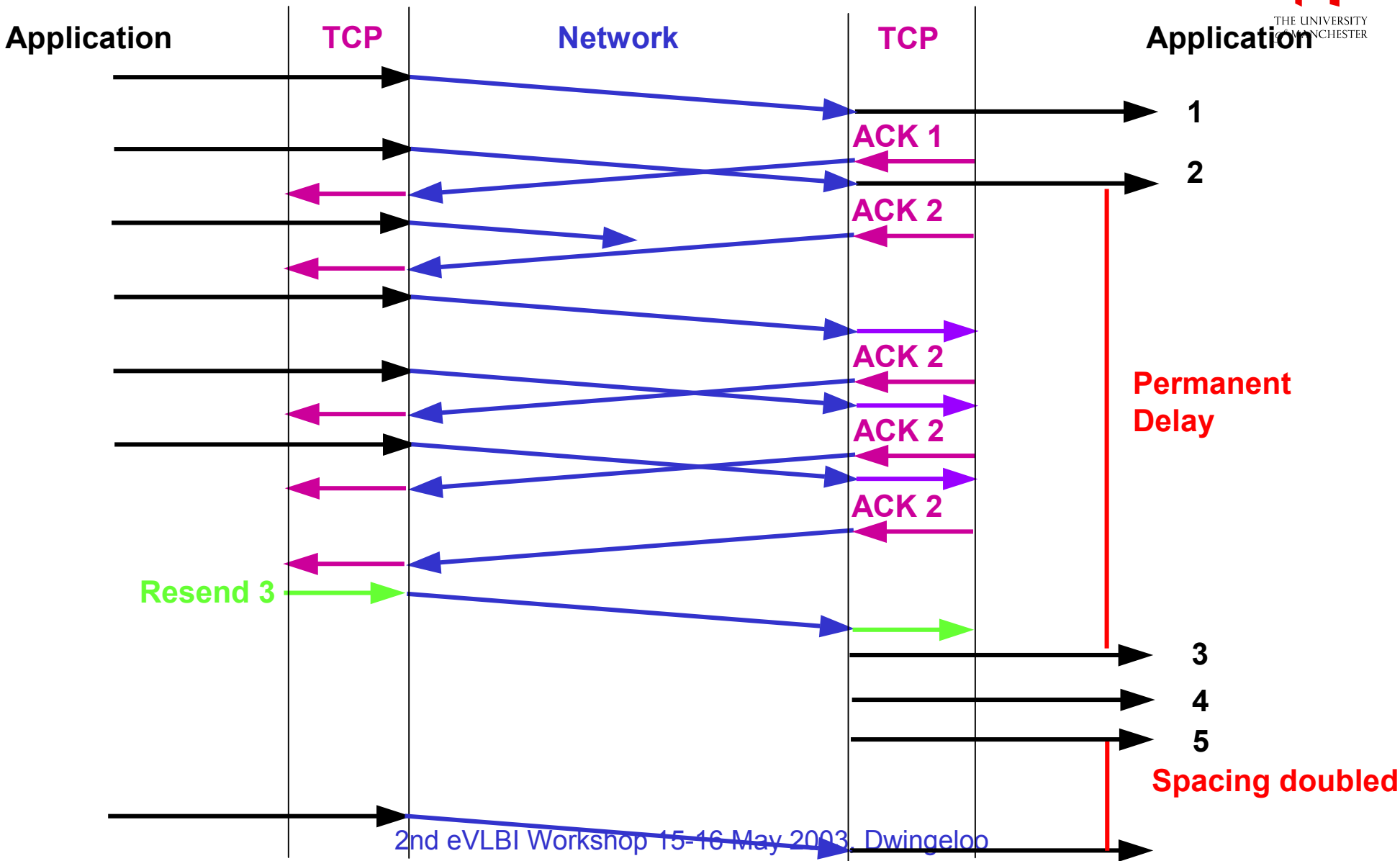


Receiving Packets



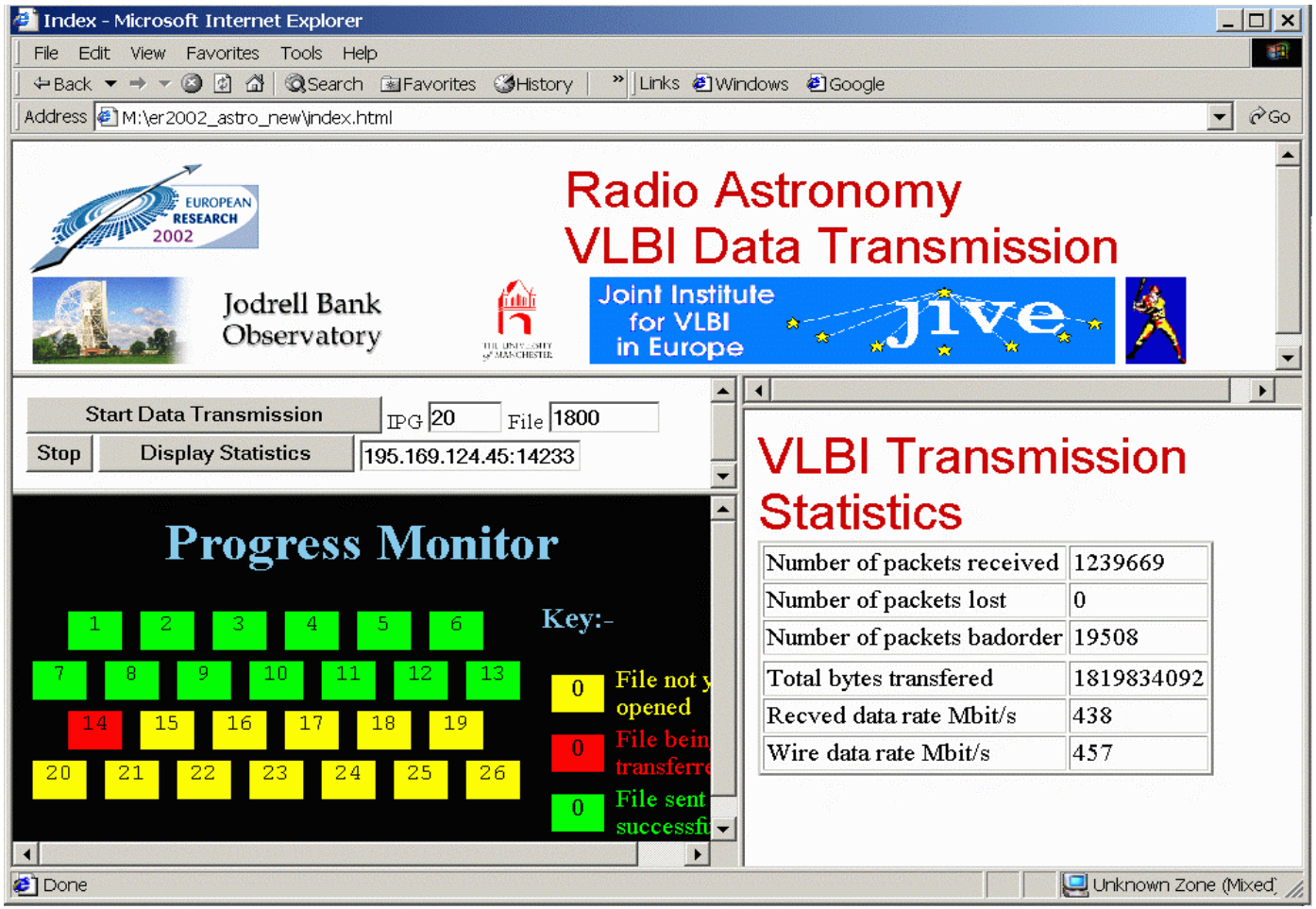
- **Packets have**
  Header:
  packet id + timestamp
  Data
  - Put header directly into control area
  - Put data directly into ring buffer
  - No extra copy
- **Assume UDP packet arrive in order none lost**
  - Deposit data into next consecutive slot
- **Inspect Header**
  - Move data onward to correct location
  - Record:
    - The received inter-packet spacing
    - 1-way delay

TCP in Action



The VLBI Demonstration

- BE & LBE Qos on GÉANT Backbone
- Web site www.hep.man.ac.uk/~rich/VLBI_web

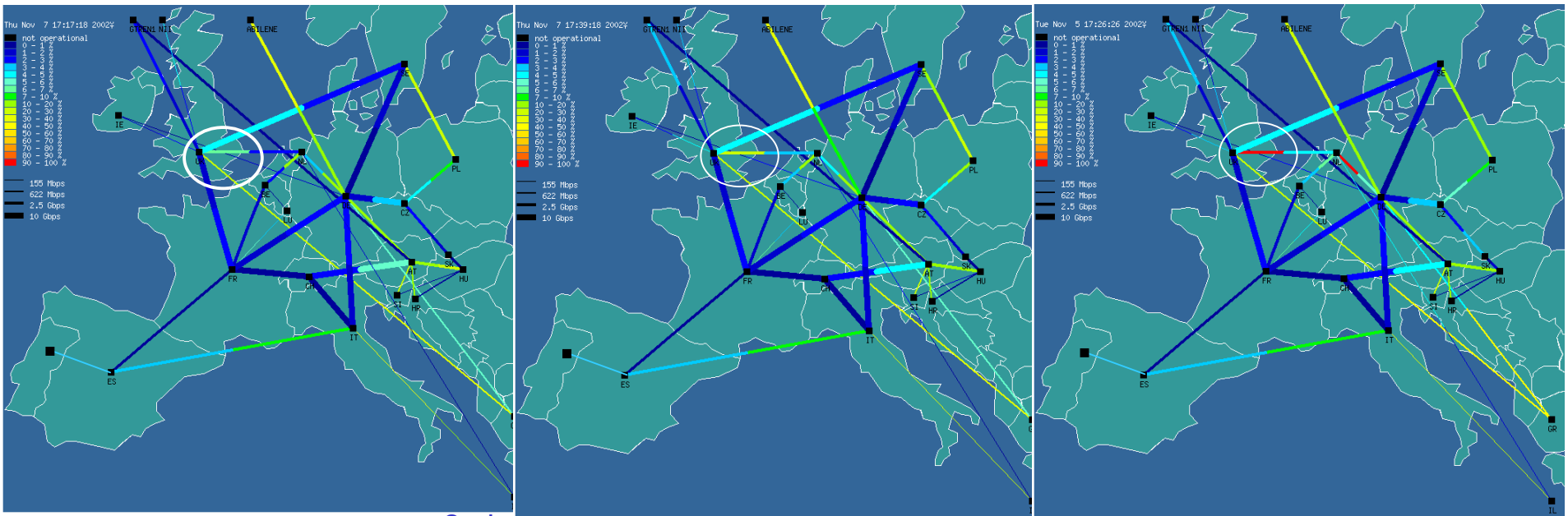


The GÉANT Weathermap

- **Normal Traffic**

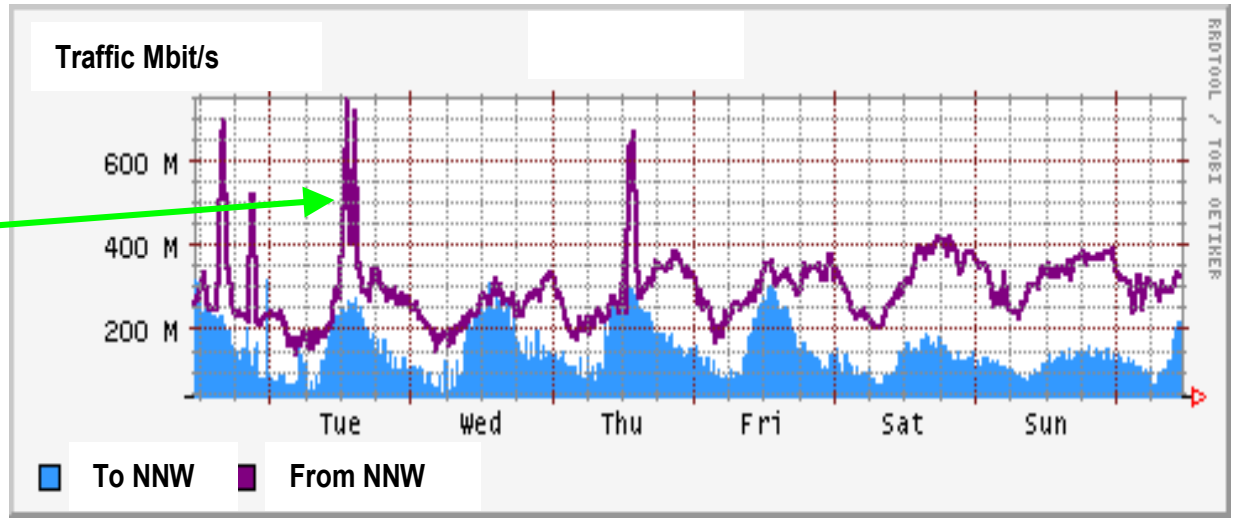
- **Normal Traffic**
  +
- **Radio Astronomy Data**
  - 500 Mbit/s

- **Normal Traffic**
  +
- **Radio Astronomy Data**
  +
- **Less Than Best Effort**
  - 2.0 Gbit/s



Manchester Campus Access links: iGrid2002

- **Manchester – SuperJANET Access link**
  - **1 Gbit/s**
- **500 Mbit UDP/IP**
- **70% Usage**
- **24-26 September 2002**

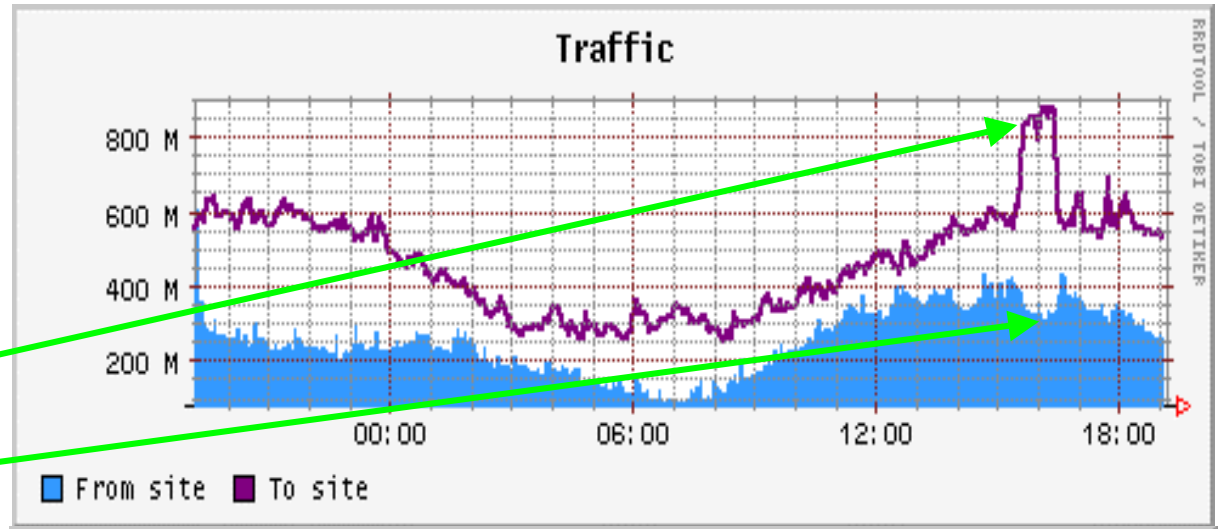


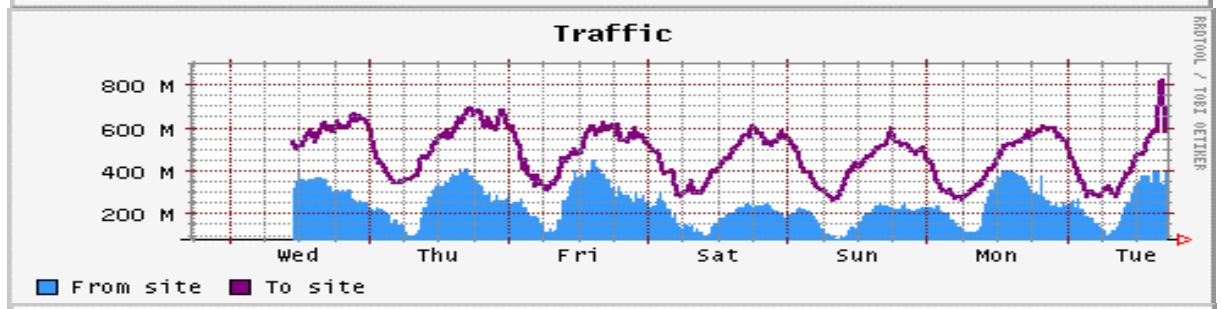
Manchester Campus Access links: ER2002

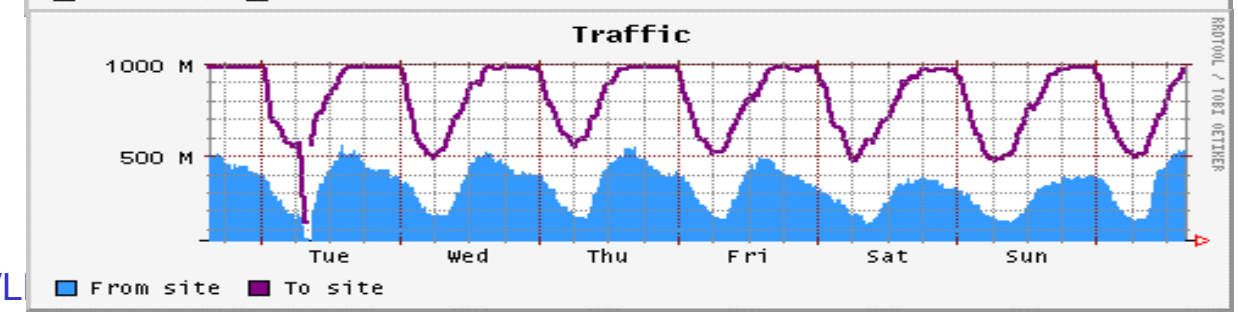
- By 10 November 2002:
- Tests Prior to the Demo
- Manchester – SuperJANET Access link
  - 1 Gbit/s
- 500 Mbit UDP/IP
- Incoming depressed

- UKERNA offered us a private 1 Gbit direct connection

- In February 2003 !

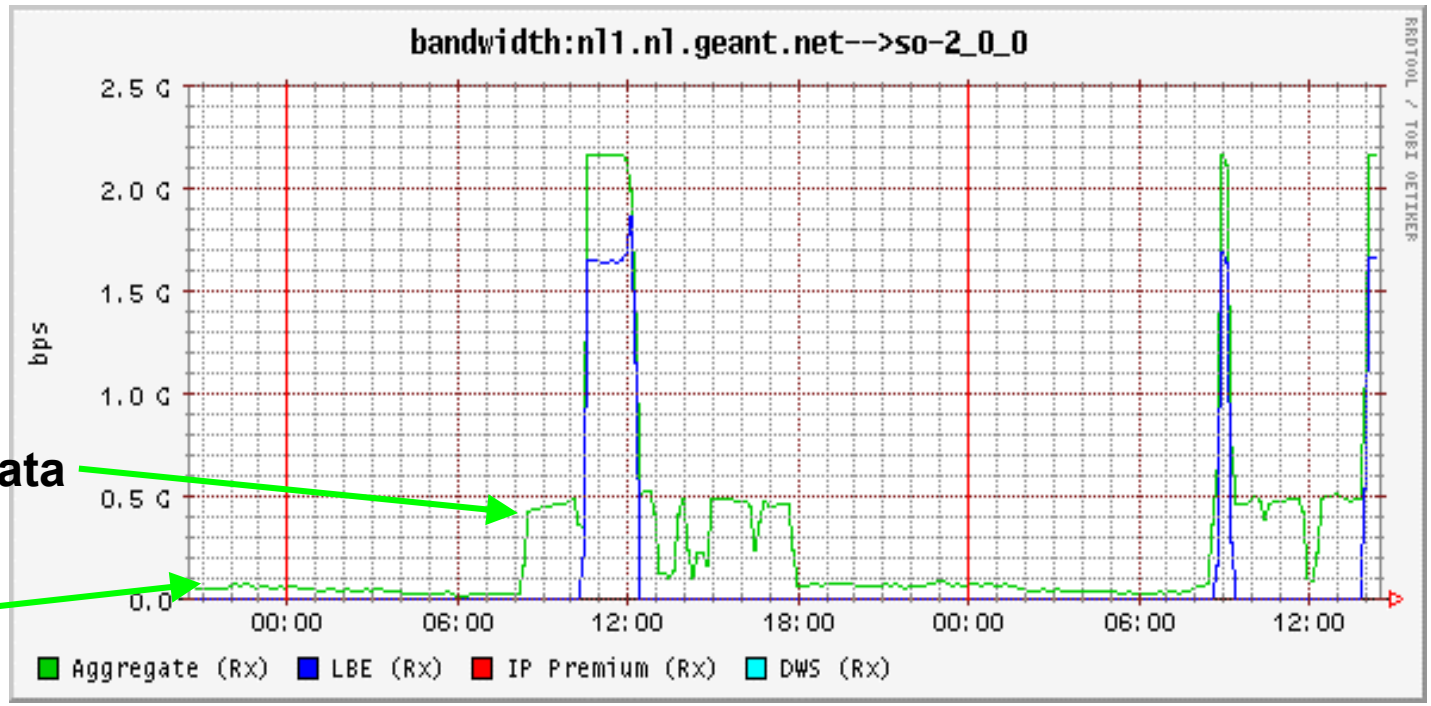






GÉANT Backbone during ER2002

- Dante Backbone
- UK-NL link
  - Incoming at NL
- 500 Mbit/s VLBI data
- Normal Traffic



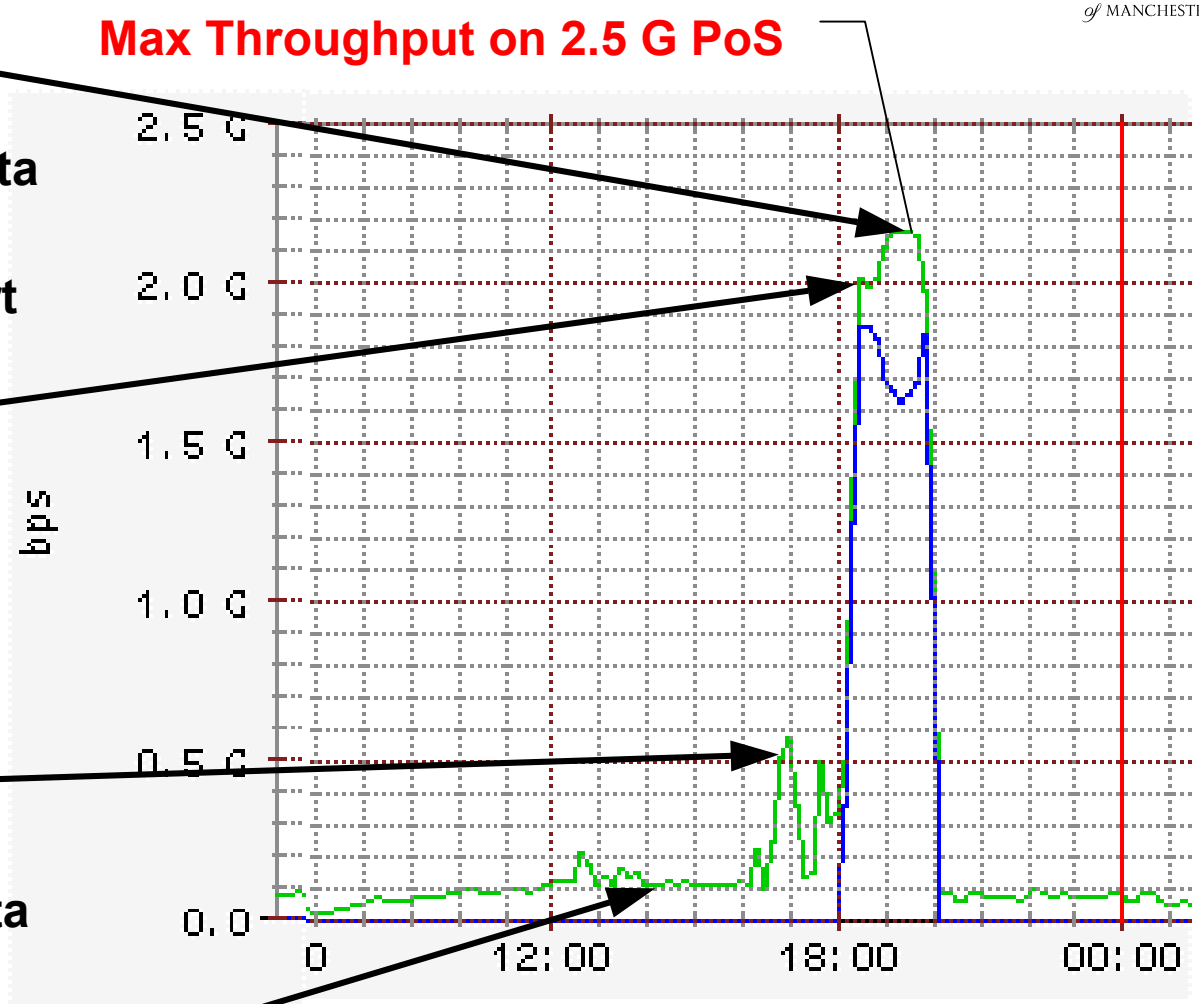
Traffic QoS Classes on GÉANT Backbone

- **Normal Traffic**
  +
- **Radio Astronomy Data**
  +
- **Less Than Best Effort**
  - 2.0 Gbit/s

- **Normal Traffic**
  +
- **Less Than Best Effort**
  - 2.0 Gbit/s

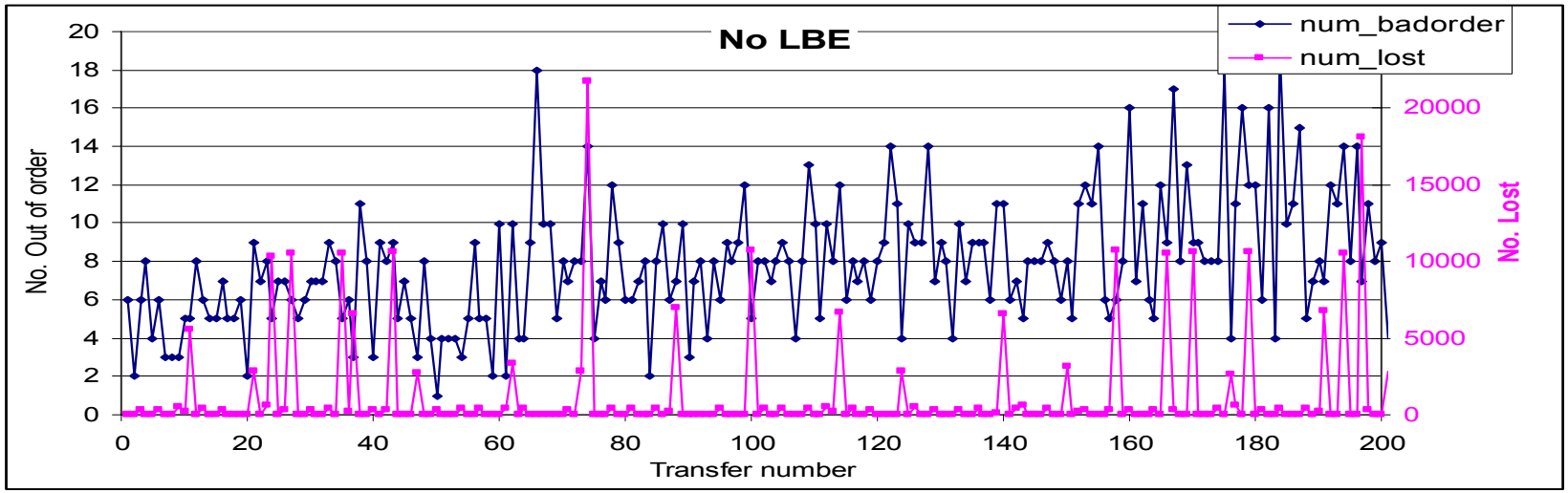
- **Normal Traffic**
  +
- **Radio Astronomy Data**
  - 500 Mbit/s

- **Normal Traffic**

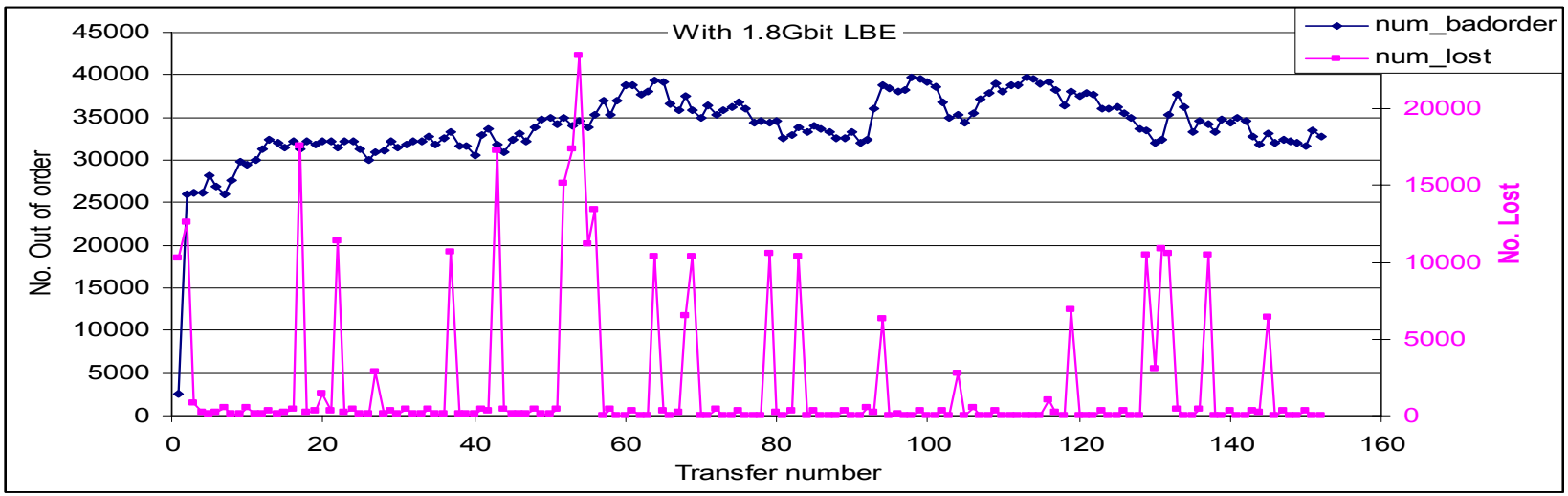


Some Measurements made during ER2002

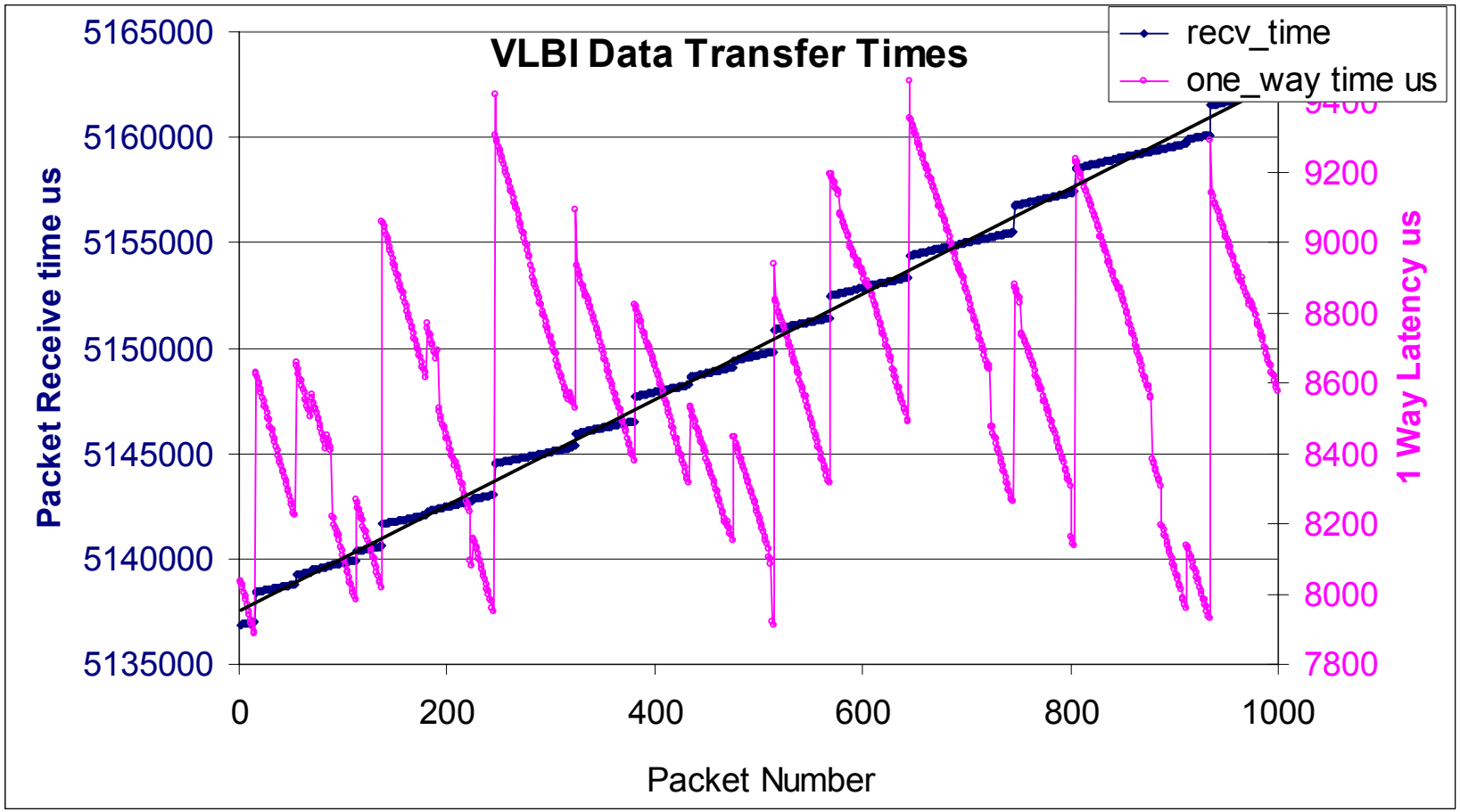
- **1.24 M packets**



0.4 %

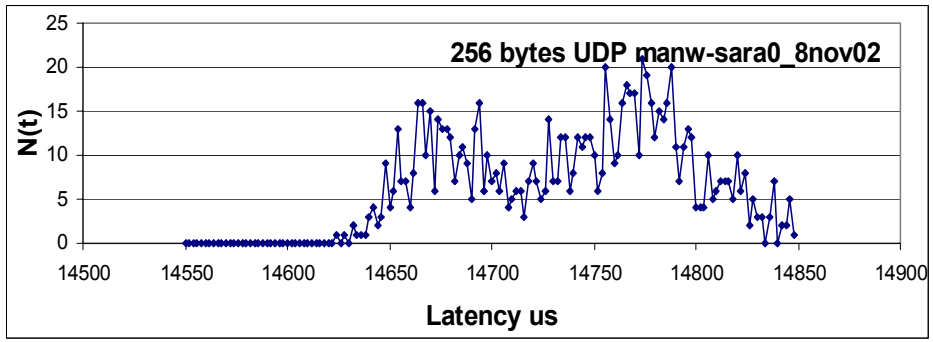


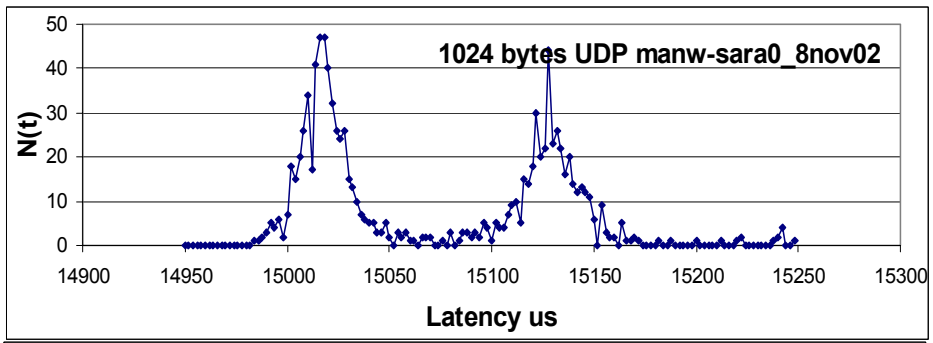
1-way delay during iGrid2002

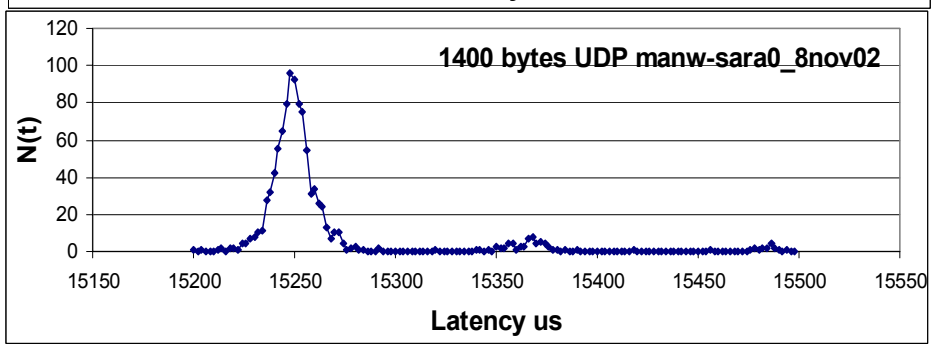


Latency Histograms

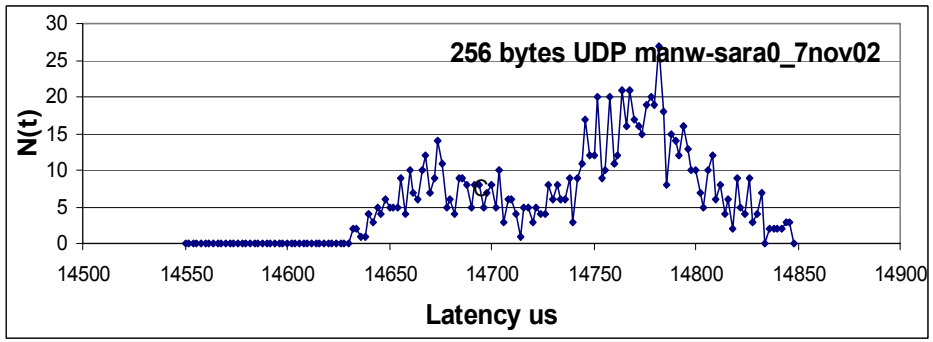
No LBE

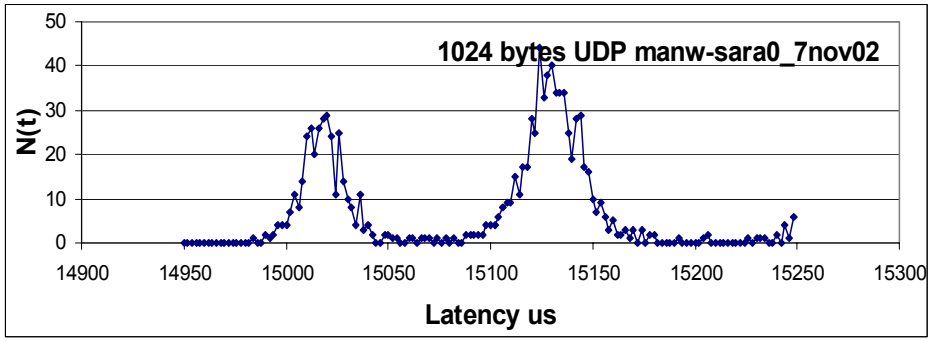


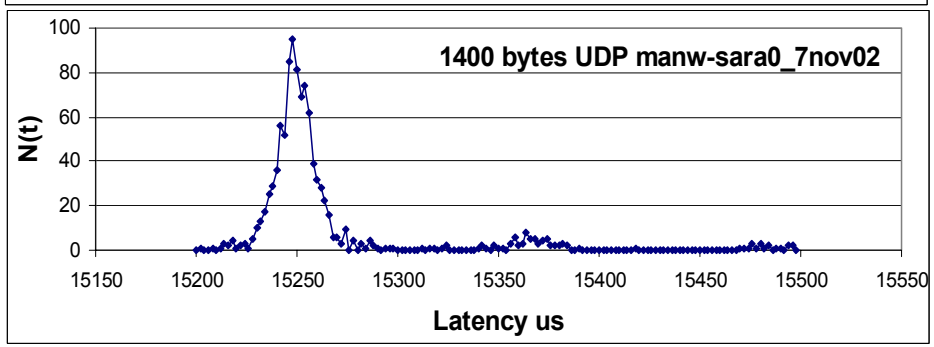




2.0 Gbit LBE UK→ DE

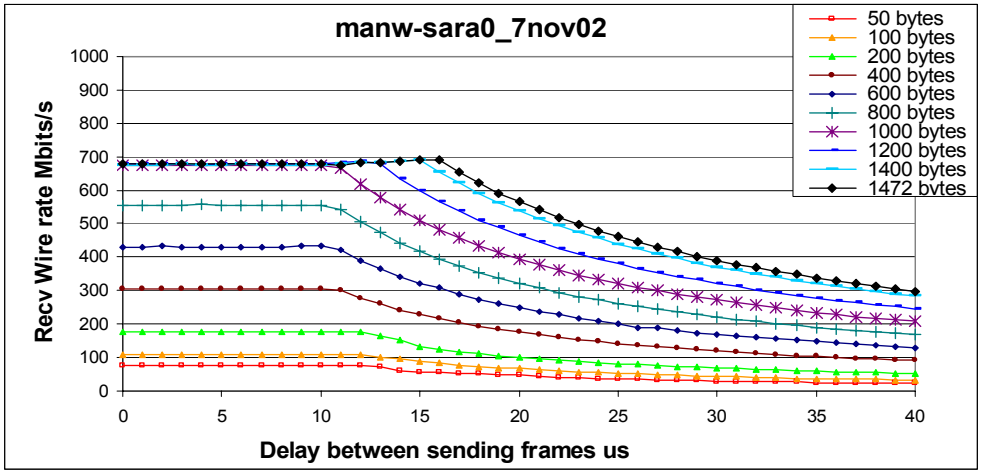


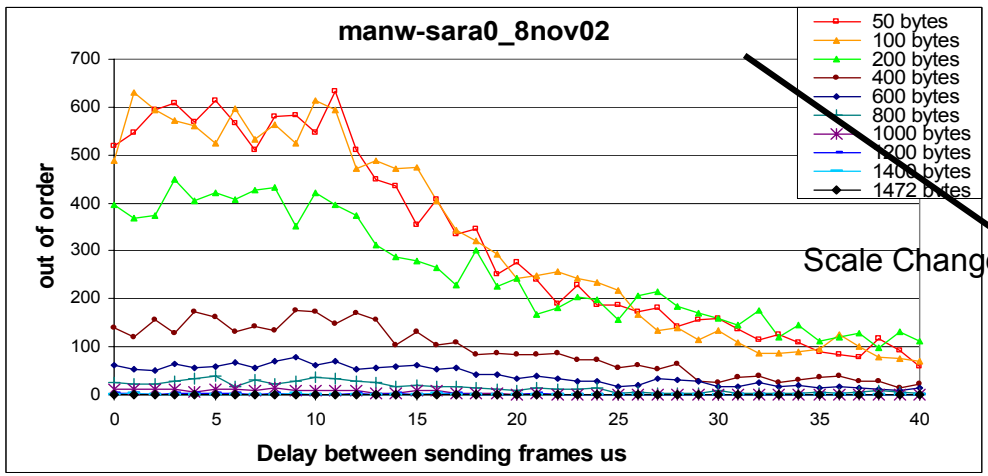




Throughput

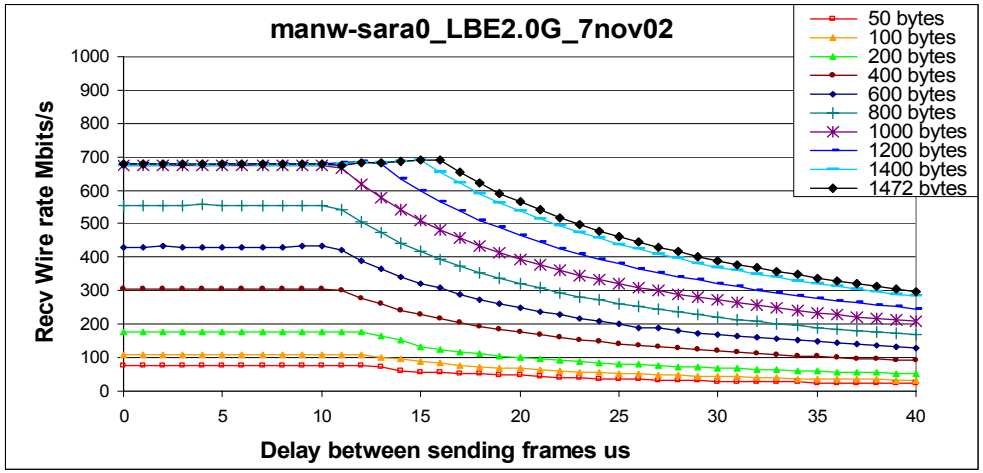
No LBE

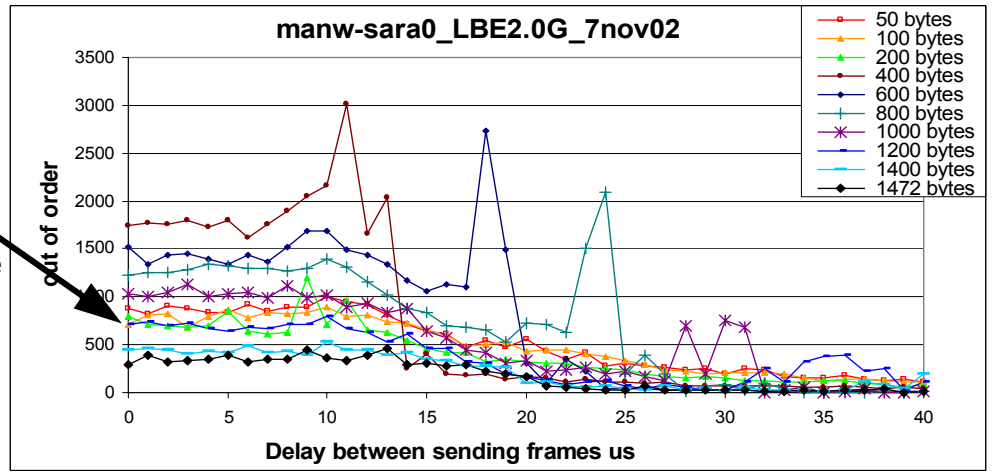




Scale Change

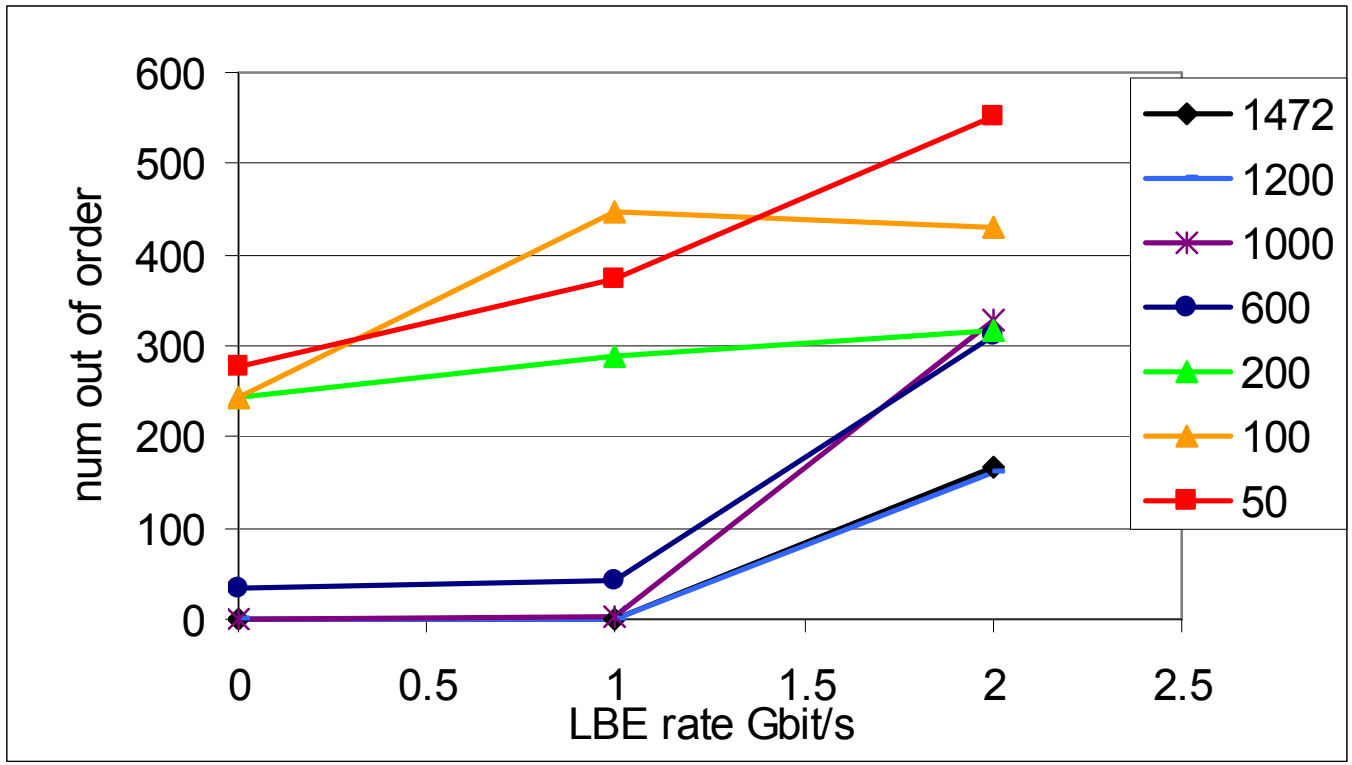
2.0 Gbit LBE UK→ DE





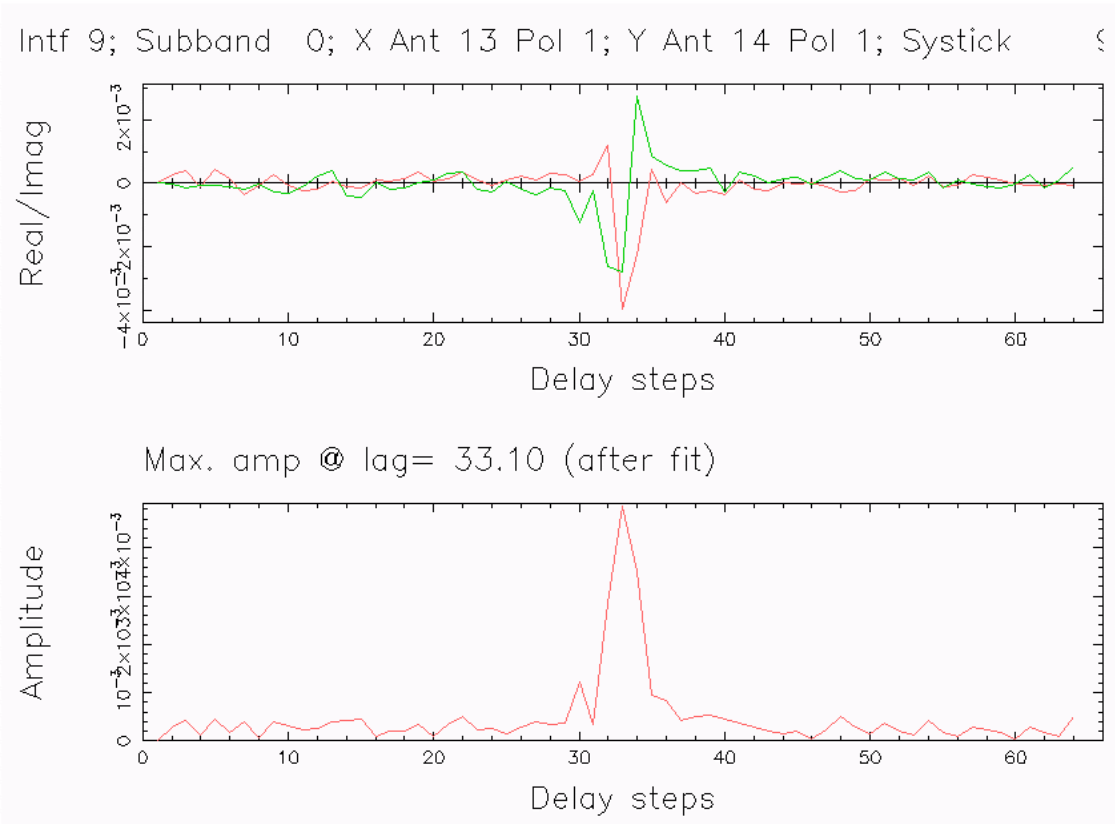
Packet re-ordering

- **20 µs Transmit packet spacing 10,000 sent**
- **Different behaviour for small packets**
- **Still some re-ordering at 40 µs spacing**
  - **~100 for 100 bytes**
  - **20 for >1000 bytes**



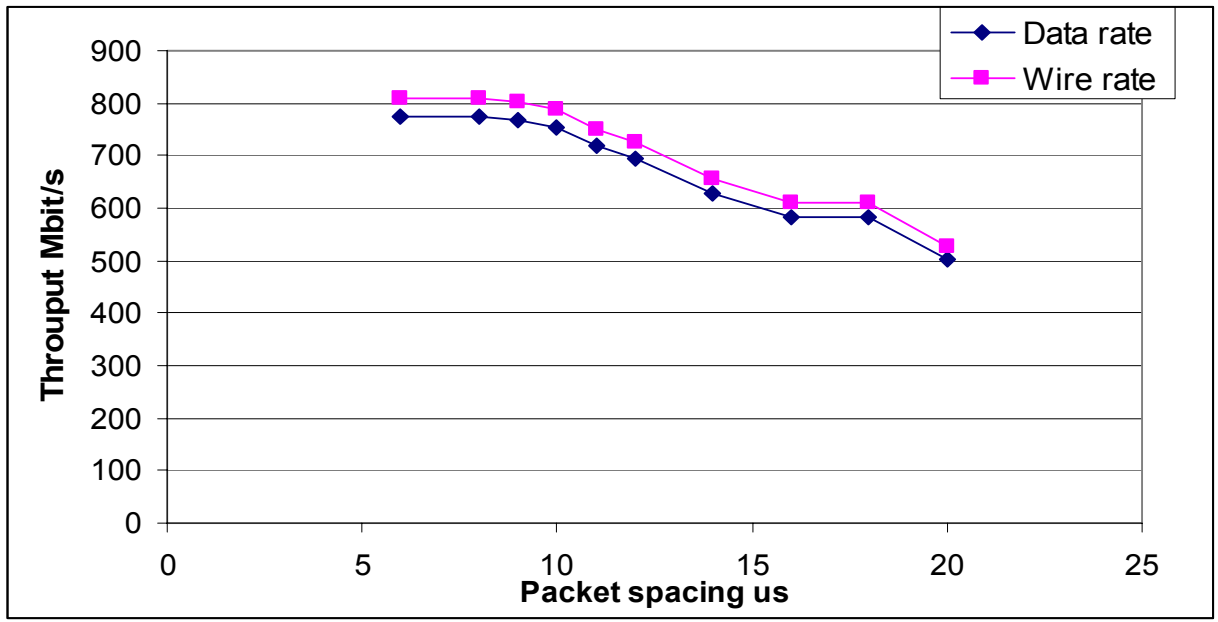
Lambda to Dwingeloo

- **Became Operational during iGrid2002**
- **Data from Manchester were sent by ftp to Dwingeloo (Westerbork data)**
- **Received on PCEVN with Gigabit Ethernet**
- **Correlated with JB Data from tape**

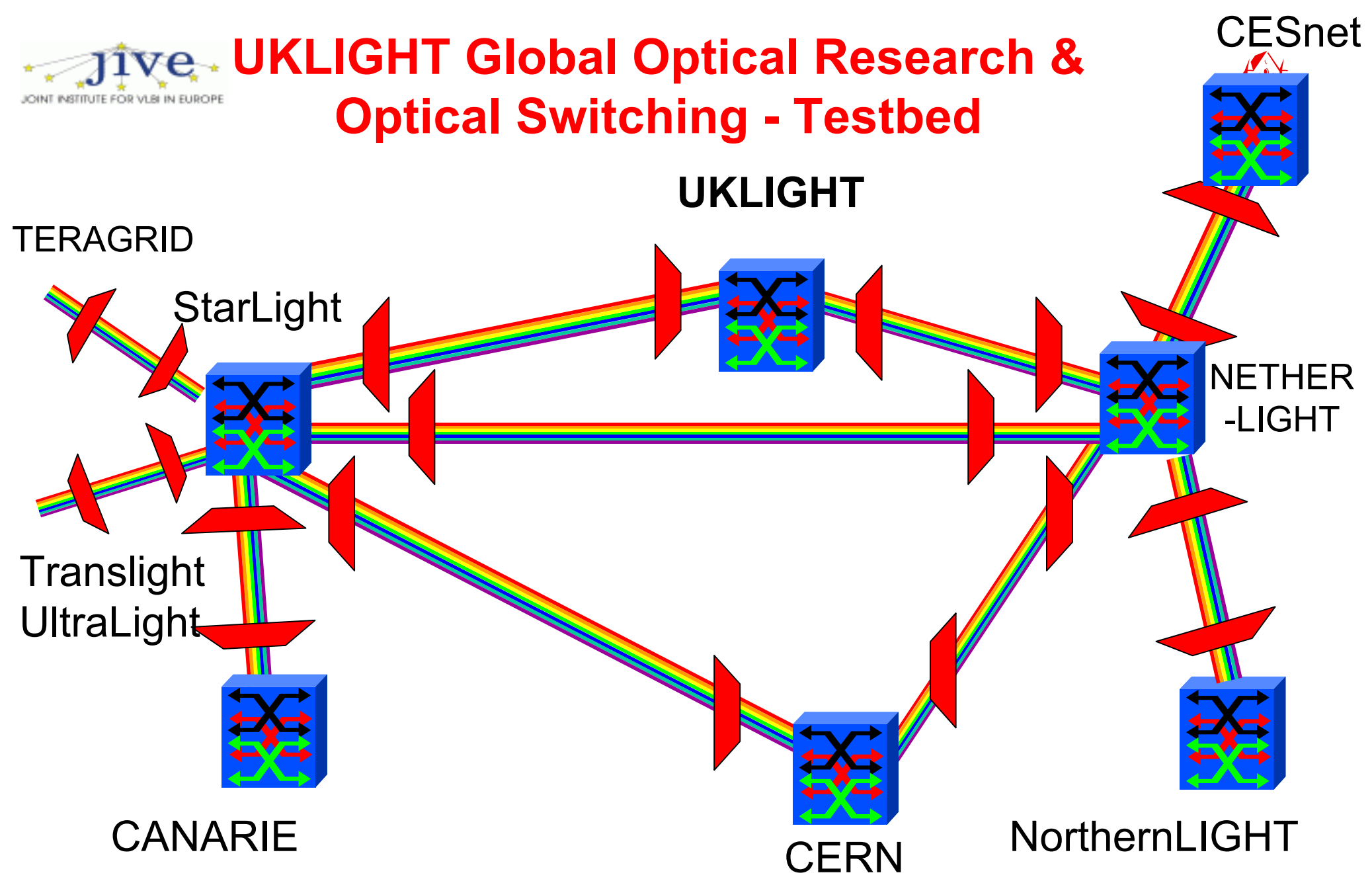


Tests on the MB-NG Development Network

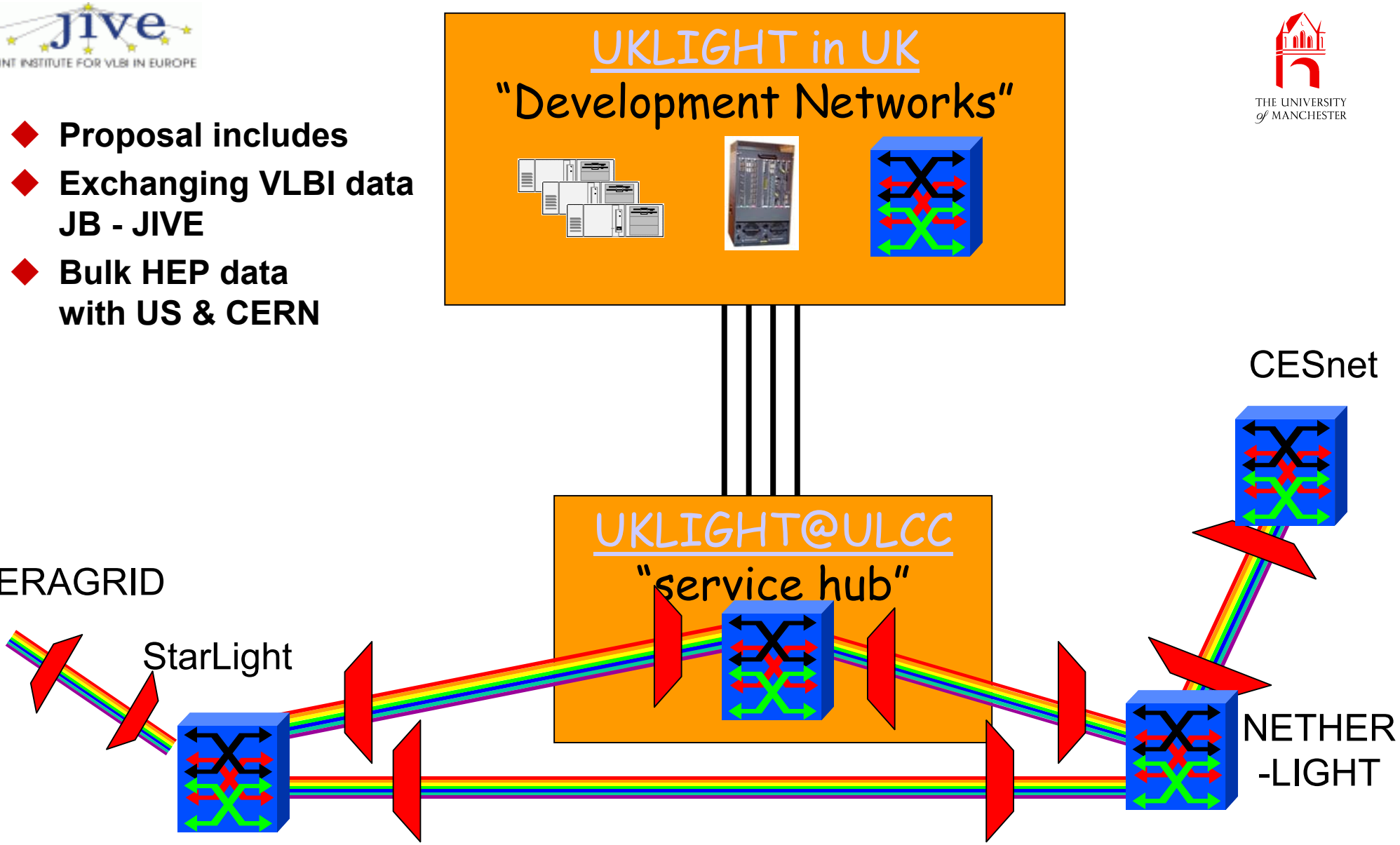
- **Sending from Manchester to London**
- **Reading form 3ware Raid0**
- **Full 1500 byte Ethernet packet transmitted every 12.48 µs on Gigabit Ethernet**
- **Rise to 8 µs indicates interaction between reading data from disk and buffering it in the IP stack**



UKLIGHT Global Optical Research & Optical Switching - Testbed



- **Proposal includes**
- **Exchanging VLBI data JB - JIVE**
- **Bulk HEP data with US & CERN**



Summary, Conclusions & Thanks

- The production network can transfer VLBI data at over 500 Mbit/s
- Simple UDP/IP is fine
- Packet loss acceptable provided link not saturated
- End Hosts must have sufficient power –
  - “Server Quality” Motherboards
  - High speed disk sub-systems
  - Well designed NIC

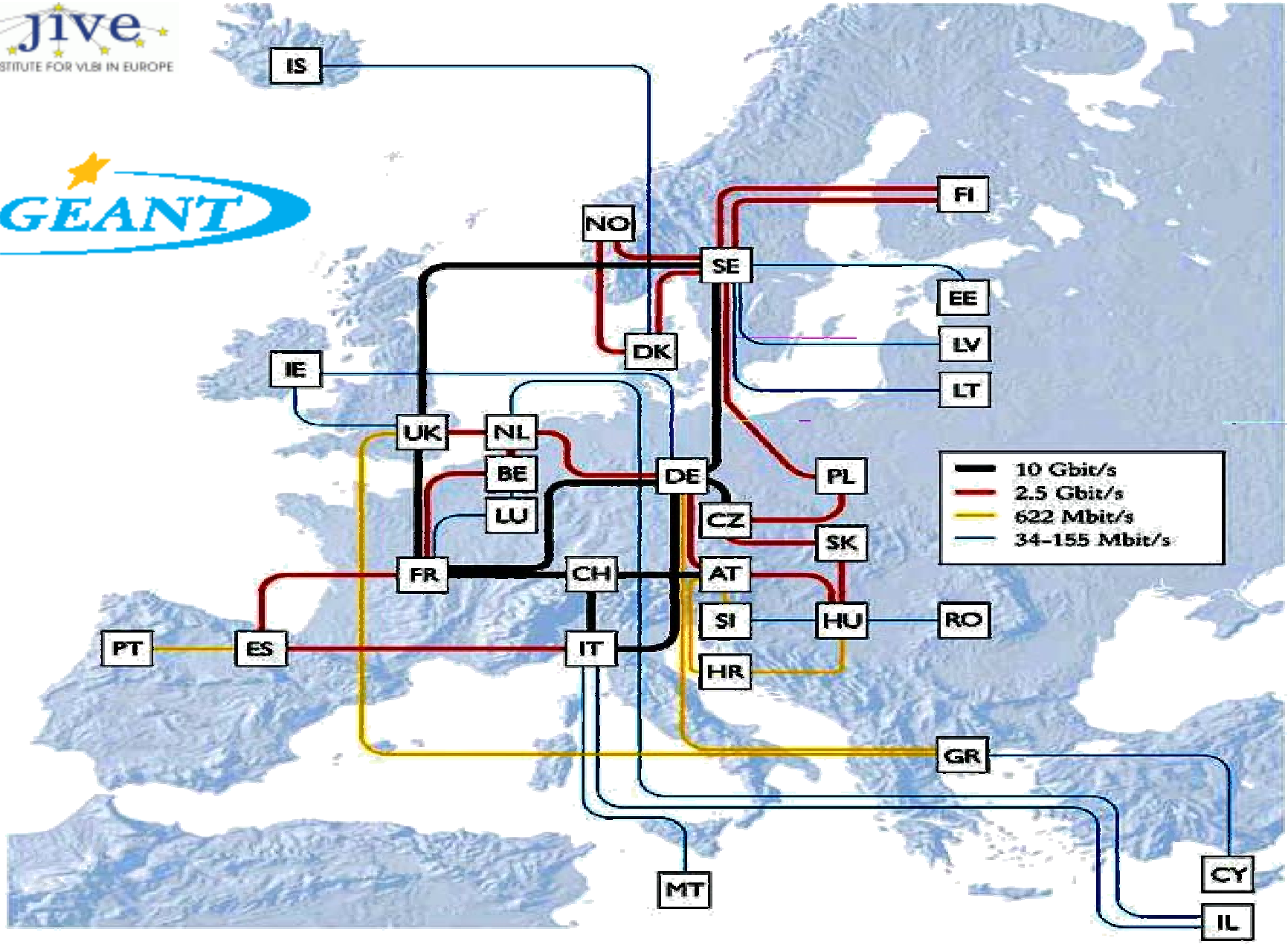
- Many thanks to all who helped

Acknowledgements

- Boston Ltd supply of the SuperMicro servers & donation of SCSI disks
  - Dev Taygi, Alex Gili-Ross
- Brunell University
  - Peter van Santen
- Dante
  - Marian Garcia, Nicolas Simar, Dale Robertson, Roberto Sabatino,
- JIVE
  - Steve Parsley, Sergei Pogrebenko
- Metsähovi Radio Observatory
  - Ari Mujunen
- SURFnet & Universiteit van Amsterdam
  - Pieter de Boer, Erik.Radius, Cees de Laat, Antony Antony, Mieke van de Berg
- UKERNA
  - Robin Arrack, Bob Day, Jeremy Sharp
- University of Manchester
  - Richard Hughes-Jones, Ralph Spencer, Paul Burgess,
  - Robin Hughes-Jones Design of web interface (Undergraduate, Oxford University)

More Information & Some URLs

- VLBI Demos at iGrid2002 and ER2002
  www.hep.man.ac.uk/~rich/VLBI_web
  http://www.hep.man.ac.uk/u/rich/VLBI_web/igrid2002_astro/iGrid2002_v8.pdf
- UDPmon / TCPmon kit + writeup
  http://www.hep.man.ac.uk/~rich/net
- ATLAS Investigation of the Performance of 100Mbit and Gigabit Ethernet Components Using Raw Ethernet Frames
  http:// www.hep.man.ac.uk/~rich/atlas/atlas_net_note_draft5.pdf
- DataGrid WP7 Networking:
  http://www.gridpp.ac.uk/wp7/index.html
- DataTAG
  http://datatag.web.cern.ch/datatag
- Motherboard and NIC Tests:
  www.hep.man.ac.uk/~rich/net/nic/GigEth_tests_Boston.ppt
  http://datatag.web.cern.ch/datatag/pfldnet2003/papers/hughes-jones.pdf
- PFLDNet Workshop
  http://datatag.web.cern.ch/datatag/pfldnet2003/index.html


IS
NO
SE
FI
EE
LV
LT
DK
IE
UK
NL
BE
LU
DE
PL
CZ
SK
FR
CH
AT
SI
HU
RO
PT
ES
IT
HR
GR
MT
CY
IL
10 Gbit/s
2.5 Gbit/s
622 Mbit/s
34-155 Mbit/s

Throughput Measured for 1472 byte Packets

NIC Motherboard	Alteon AceNIC	SysKonnect SK-9843	IntelPro1000
SuperMicro 370DLE; Chipset: ServerWorks III LE PCI 32bit 33 MHz RedHat 7.1 Kernel 2.4.14	674 Mbit/s	584 Mbit/s 0-0 µs	
SuperMicro 370DLE Chipset: ServerWorks III LE PCI 64bit 64 MHz RedHat 7.1 Kernel 2.4.14	930 Mbit/s	720 Mbit/s 0-0 µs	910 Mbit/s 400-120 µs
IBM das Chipset: CNB20LE; PCI 64bit 32 MHz RedHat 7.1 Kernel 2.4.14		790 Mbit/s 0-0 µs	930 Mbit/s 400-120 µs
SuperMicro P4DP6 Chipset: Intel E7500; PCI 64bit 64 MHz RedHat 7.2 Kernel 2.4.19-SMP		876 Mbit/s 0-0 µs	950 Mbit/s 70-70 µs
SuperMicro P4DP8-G2 Chipset: Intel E7500; PCI 64bit 64 MHz RedHat 7.2 Kernel 2.4.19-SMP		990 Mbit/s 0-0 µs	995 Mbit/s 70-70 µs