THE UNIVERSITY of MANCHESTER

UDPmon and TCPstream
Tools to understand Network Performance

Richard Hughes-Jones
The University of Manchester

www.hep.man.ac.uk/~rich/net/

UDPmon: The Measurements (1)

THE UNIVERSITY of MANCHESTER

- **Latency**
- Round trip times measured using Request-Response UDP frames
- Latency as a function of frame size
  - Slope gives sum of individual data transfer rates end-to-end
    - Mem copy + pci + Gig Ethernet + pci + mem copy
  - Intercept indicates processing times + HW latencies
- Histograms of ‘singleton’ measurements
- Tells us about:
  - Behavior of the IP stack
  - The way the HW operates
  - Performance of the LAN / MAN / WAN

UDPmon: The Measurements (2)

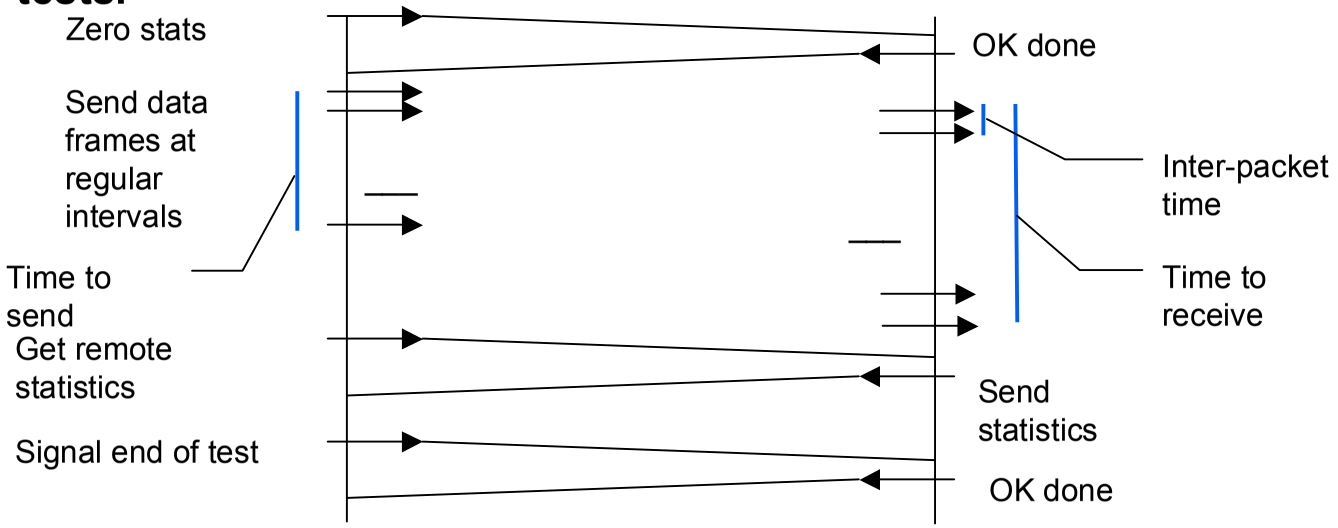
THE UNIVERSITY of MANCHESTER

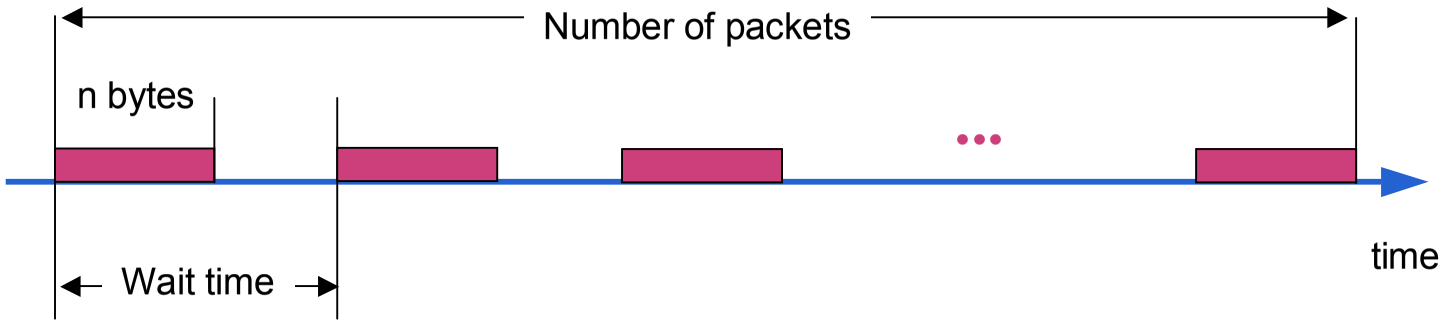
- **UDP Throughput**
- Send a controlled stream of UDP frames spaced at regular intervals
- Vary the frame size and the frame transmit spacing
- At the receiver record
  - The time of first and last frames received
  - The number packets received, the number lost, number out of order
  - The received inter-packet spacing is histogramed
  - The time each packet is received → provides packet loss pattern
  - CPU load, Number of interrupts
- Use the Pentium CPU cycle counter for times and delay
  - Few lines of user code
- Tells us about:
  - Behavior of the IP stack
  - The way the HW operates
  - Capacity and Available throughput of the LAN / MAN / WAN

UDPmon: The Works

THE UNIVERSITY of MANCHESTER

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

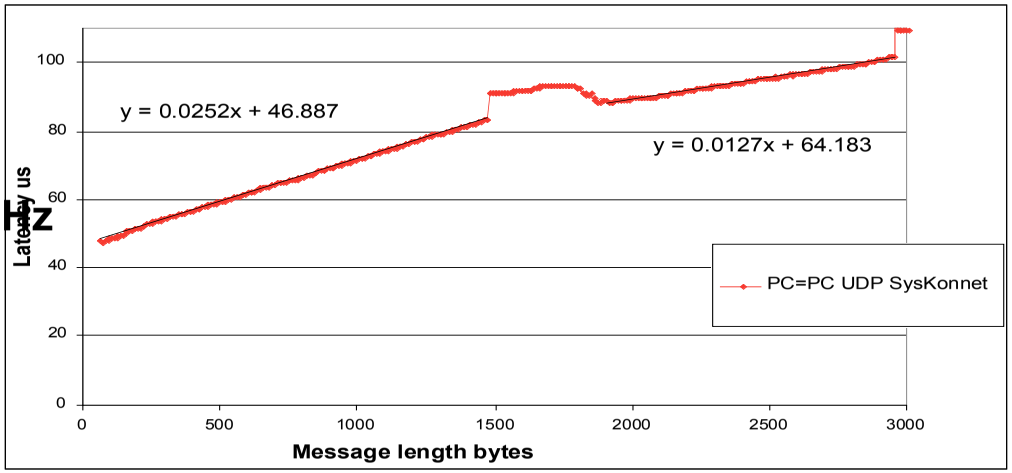


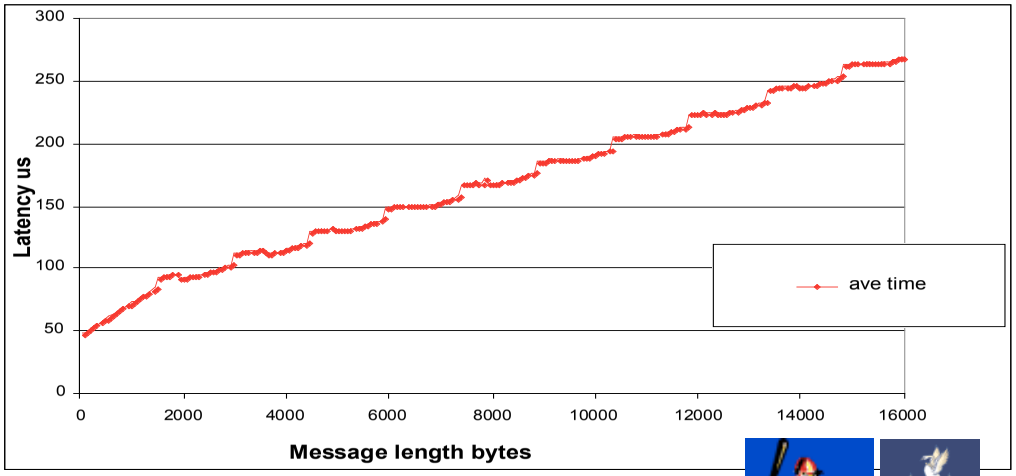


Latency: Tests B2B PCs with Gigabit Ethernet

THE UNIVERSITY of MANCHESTER

- **NIC SysKonnect SK-9843**
- Motherboard: SuperMicro 370DLE
- Chipset: ServerWorks III LE
- CPU: PIII 800 MHz **PCI:64 bit 33 MHz**
- RedHat 7.1 Kernel 2.4.14
- Latency low & well behaved
- Latency Slope **0.0252 us/byte**
- Expect:
  - PCI 0.00758
  - GigE 0.008
  - PCI 0.00758
  - **0.0236 us/byte**

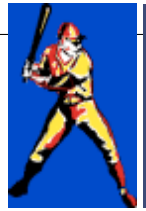




www.hep.man.ac.uk/~rich/net/nic/GigEth_tests_Boston.ppt

CAIDA Bandwidth Estimation Meeting SDSC June 2002

R. Hughes-Jones Manchester

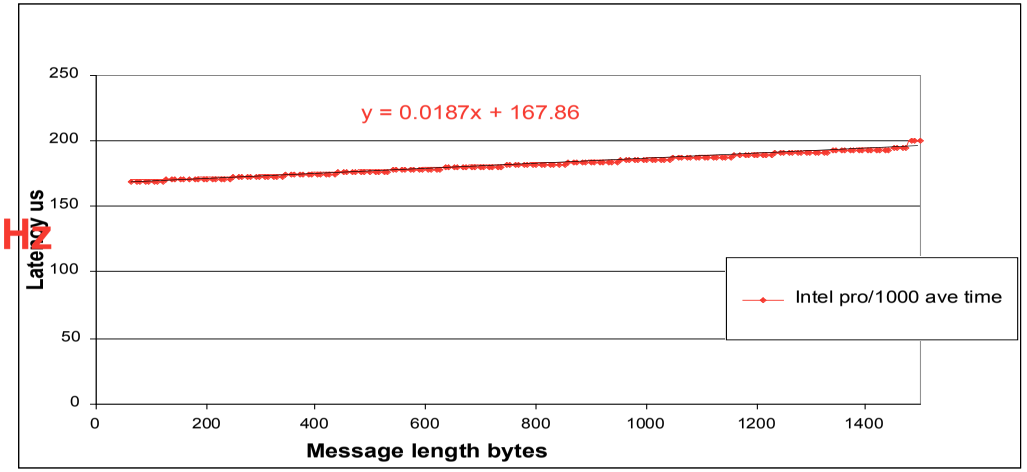


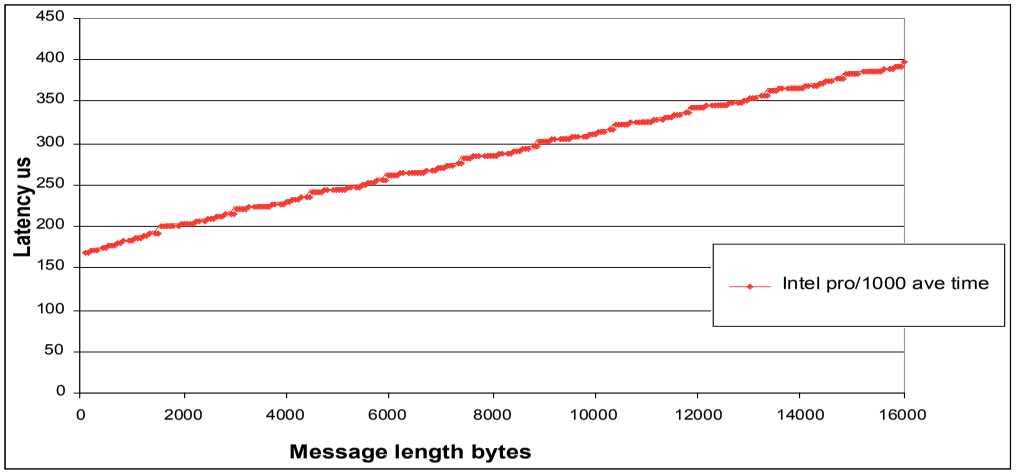
BRUNEL UNIVERSITY WEST LONDON

Latency: Tests B2B PCs with Gigabit Ethernet

THE UNIVERSITY of MANCHESTER

- **NIC Intel Pro/1000**
- Motherboard: SuperMicro 370DLE
- Chipset: ServerWorks III LE
- CPU: PIII 800 MHz **PCI:64 bit 66 MHz**
- RedHat 7.1 Kernel 2.4.14
- Latency high but well behaved
- Indicates Interrupt coalescence
- Slope **0.0187 us/byte**
- Expect:
  - PCI 0.00188
  - GigE0.008
  - PCI 0.00188
  - **0.0118 us/byte**

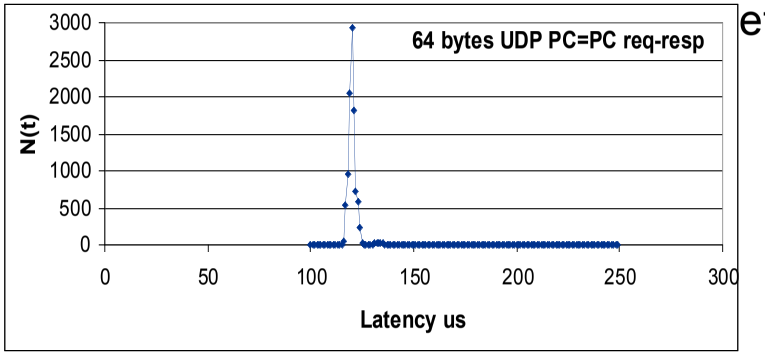




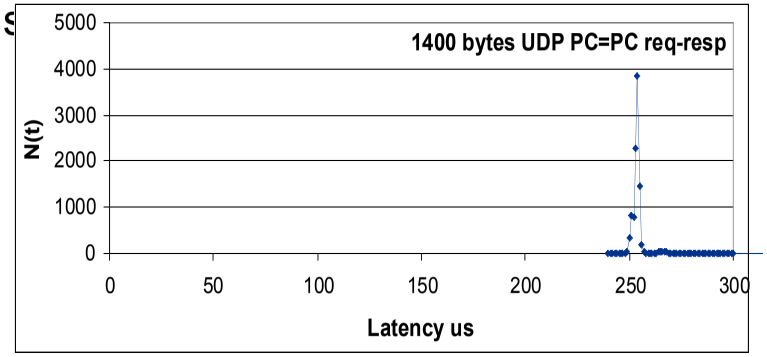
Latency: Histograms

THE UNIVERSITY of MANCHESTER

- Motherboard: SuperMicro 370DLE Chipset: ServerWorks III LE
- CPU: PIII 800 MHz **PCI:32 bit 33 MHz 100Mbit NICs** RedHat 5.2 Kernel 2.2.14
- LAN 2 PC B2B

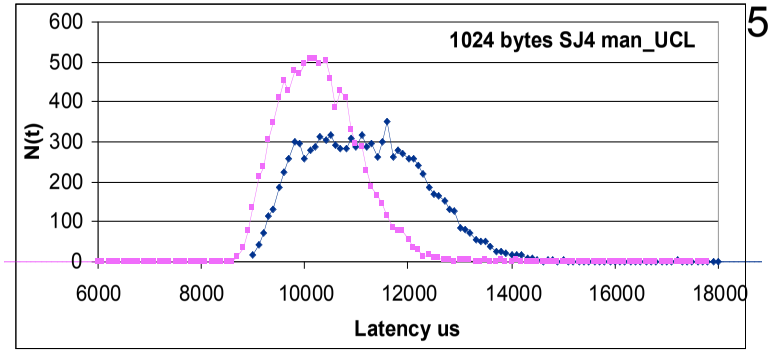


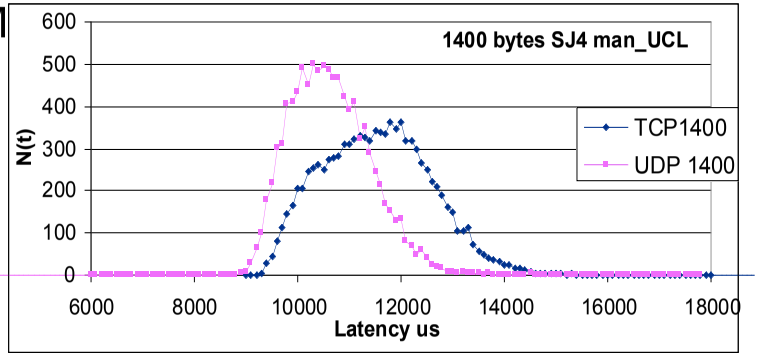
et s



- Man – UCL LAN – MAN – SuperJANET4 – MAN – LAN

5M



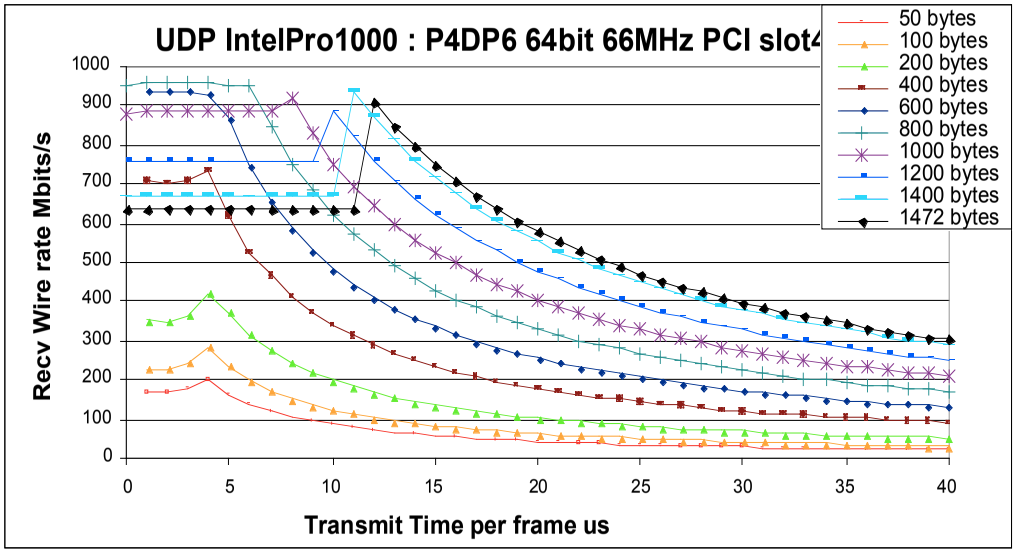


UDP Throughput: Intel Pro/1000 on B2B P4DP6

THE UNIVERSITY of MANCHESTER

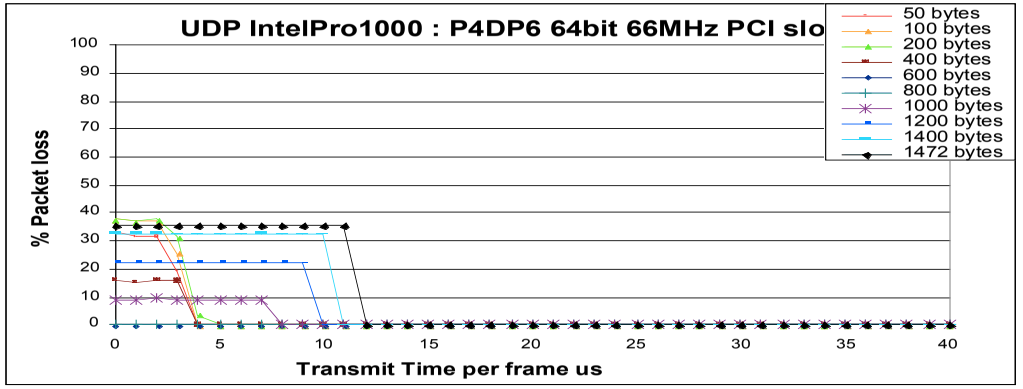
- Motherboard: SuperMicro P4DP6 Chipset: Intel E7500 (Plumas)
- CPU: Dual Xeon Prestonia (2cpu/die) 2.2 GHz **Slot 4: PCI, 64 bit, 66 MHz**
- RedHat 7.2 Kernel 2.4.14

- Max throughput **950Mbit/s**
- Some throughput drop for packets >1000 bytes



- Loss NIC dependent
- Loss not due to user ➔ Kernel moves

- Traced to discards in the receiving IP layer ???

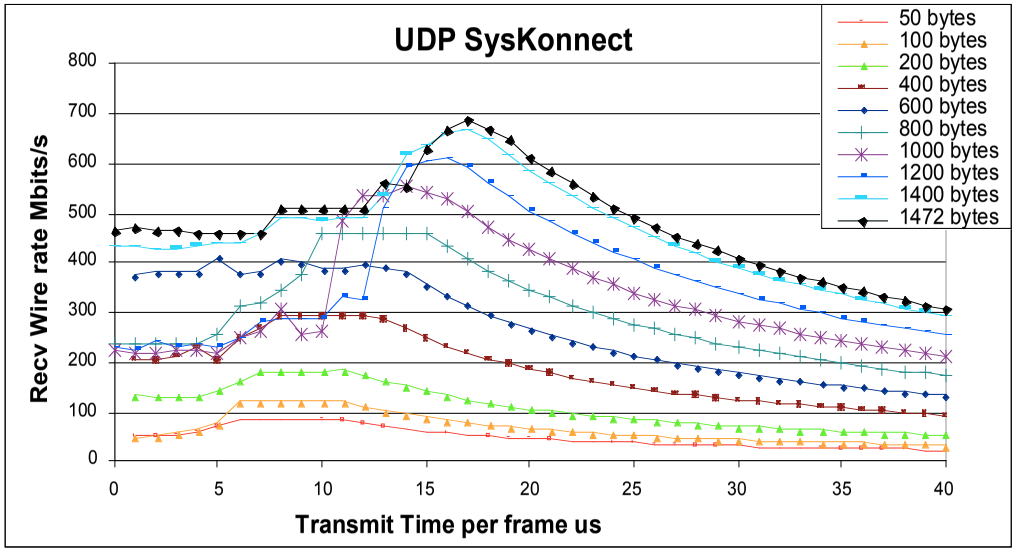


UDP Throughput: SysKonnect SK-9843 B2B

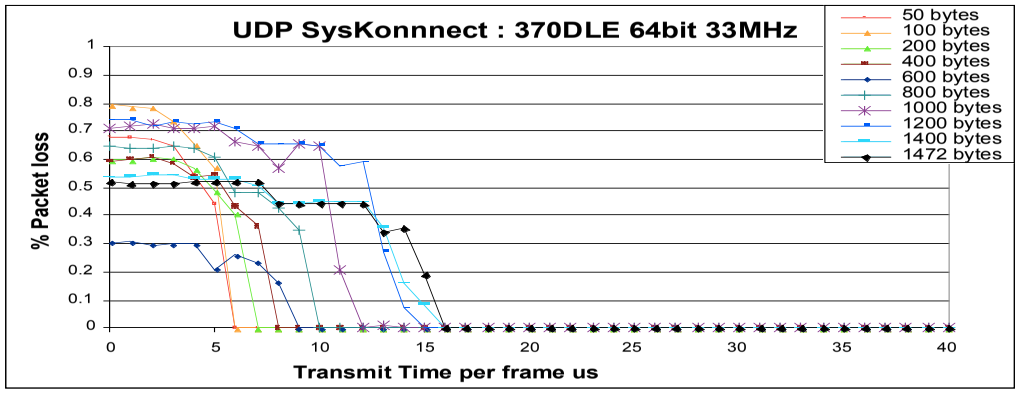
THE UNIVERSITY of MANCHESTER

- Motherboard: SuperMicro 370DLE Chipset: ServerWorks III LE Chipset
- CPU: PIII 800 MHz **PCI:64 bit 33 MHz**
- RedHat 7.1 Kernel 2.4.14

- Max throughput **690Mbit/s**
- No packet loss



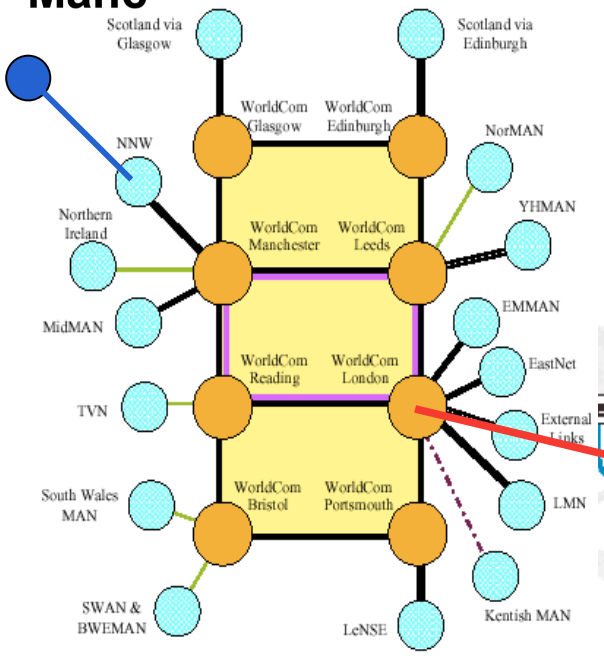
- Packet loss during drop



European Topology: NRNs & Geant

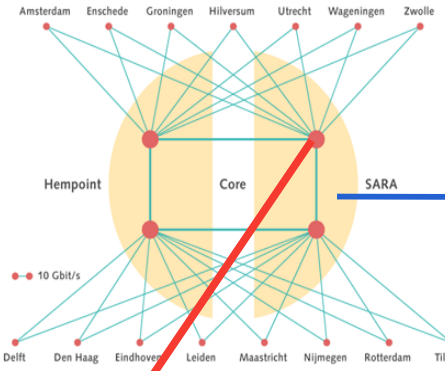
THE UNIVERSITY of MANCHESTER

Manc

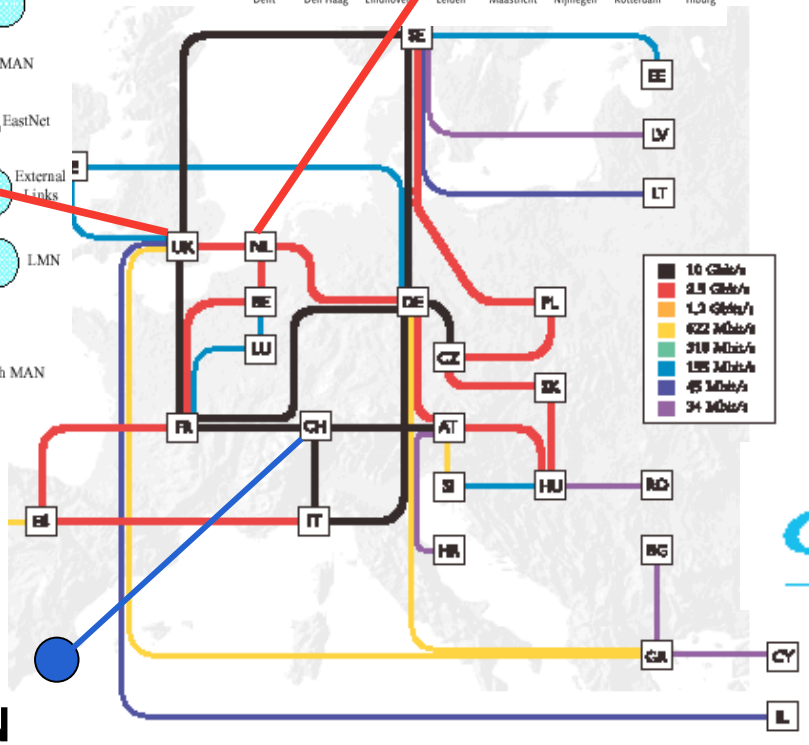


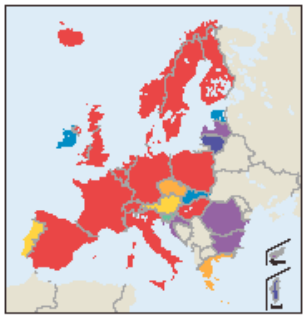
SuperJANET4

SURFnet



UvA





CERN

GEANT

CAIDA Bandwidth Estimation Meeting SDSC June 2002
R. Hughes-Jones Manchester

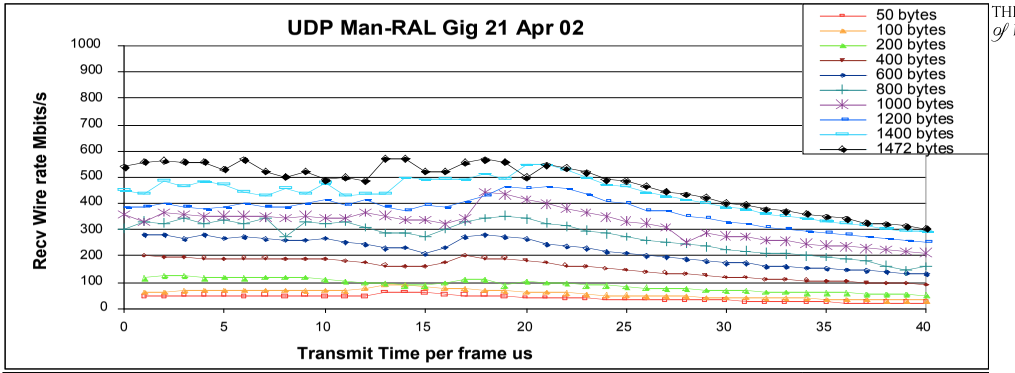
Gigabit Throughput on the Production WAN

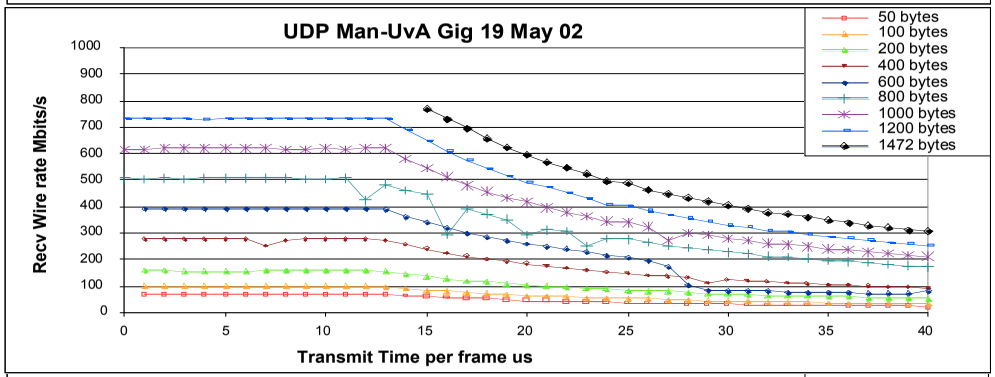
THE UNIVERSITY of MANCHESTER

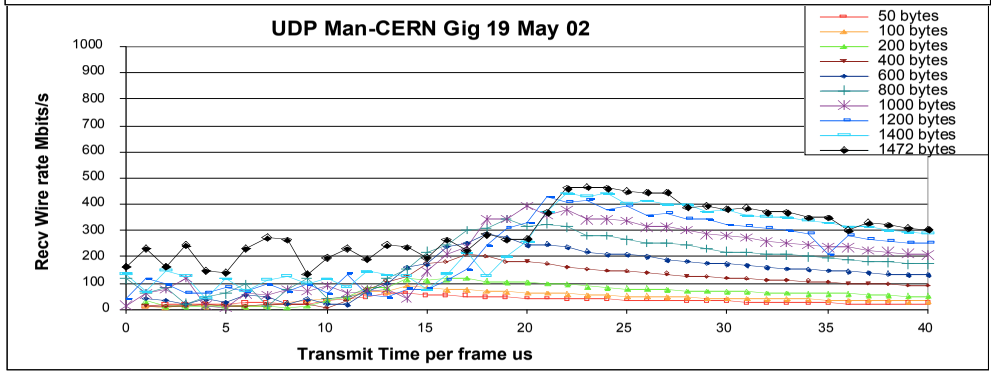
- **Manc - RAL 570 Mbit/s**
- 91% of the 622 Mbit access link between SuperJANET4 and RAL
- 1472 bytes propagation ~21 s

- **Manc-UvA (SARA) 750 Mbit/s**
- SJANET4 + Geant + SURFnet

- **Manc – CERN 460 Mbit/s**
- CERN PC had a 32 bit PCI bus



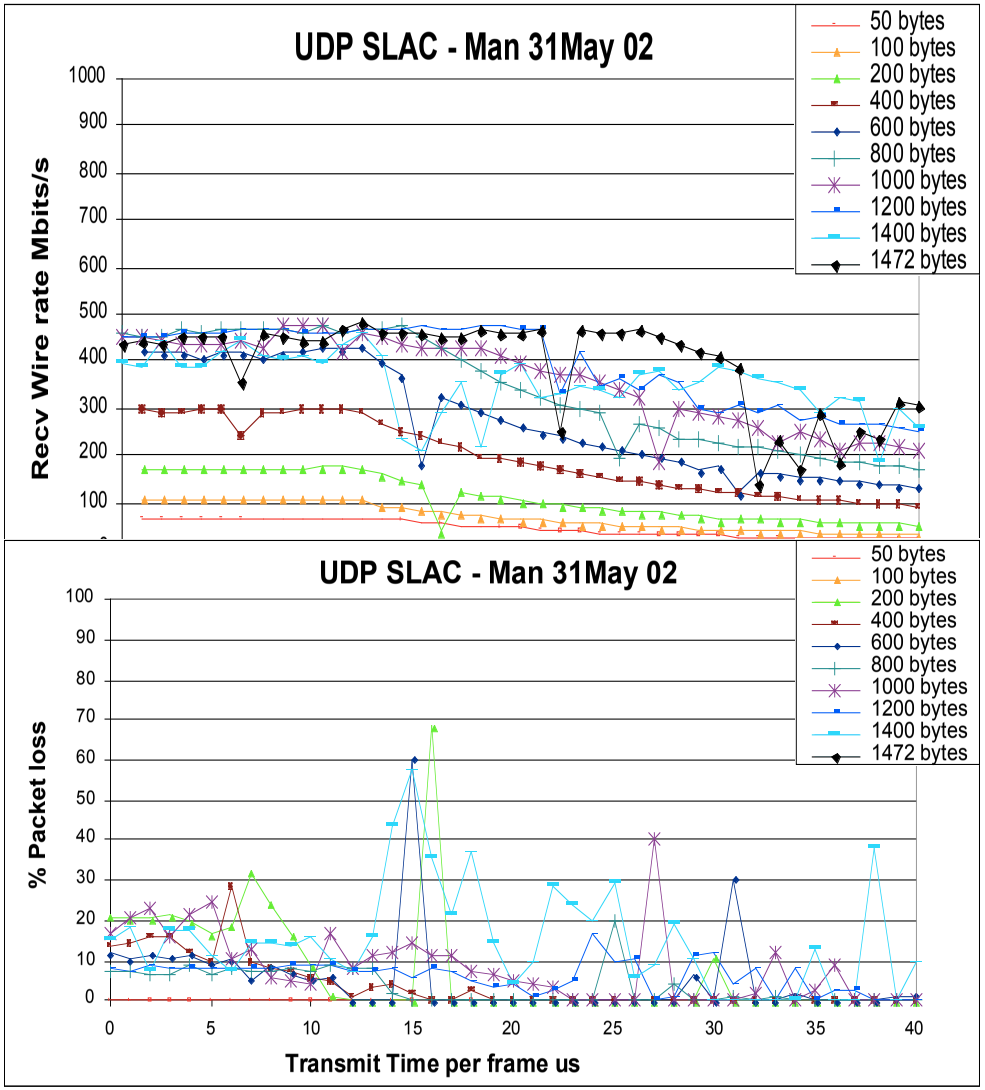




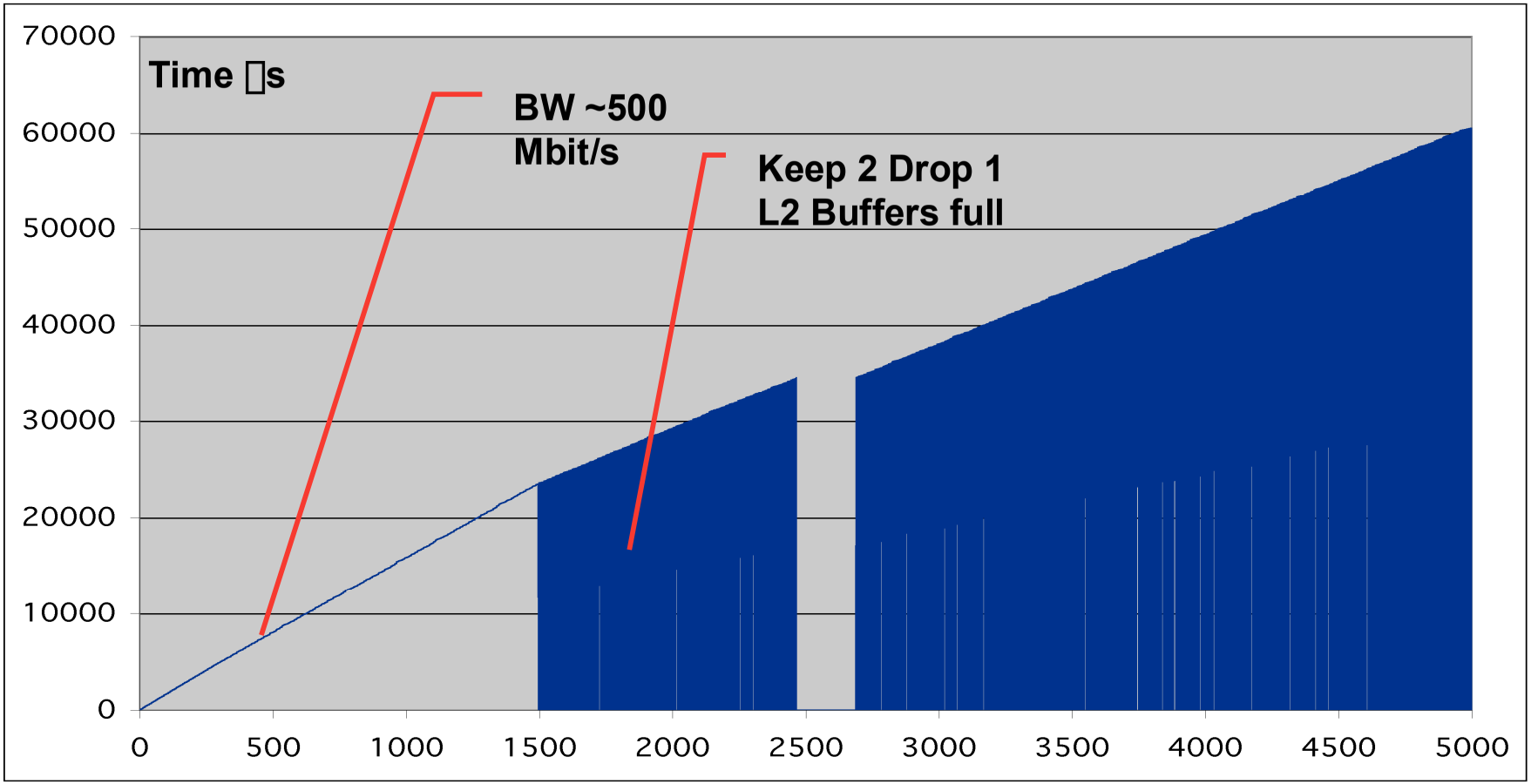
UDP Throughput: SLAC - Man

THE UNIVERSITY of MANCHESTER

- **SLAC – Manc 470 Mbit/s**
- 75% of the 622 Mbit access link
- SuperJANET4 peers with ESnet at 622Mbit in NY

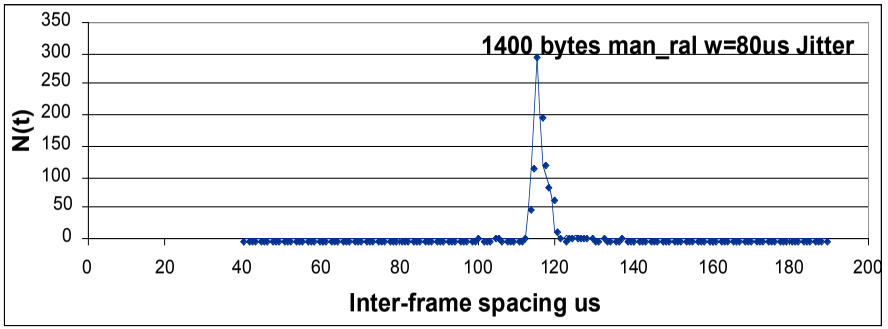


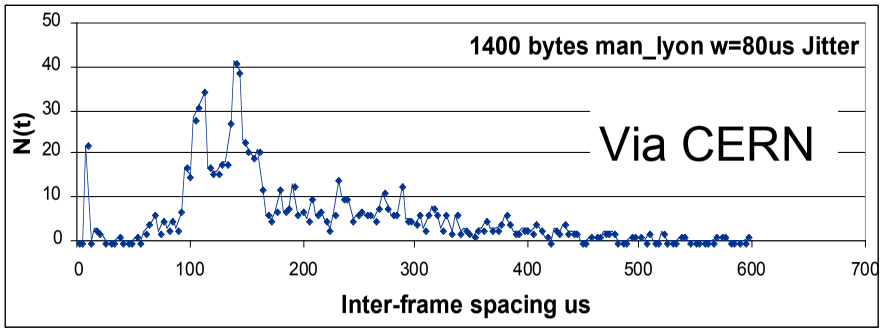
Packet Loss: Amsterdam – Chicago Lambda

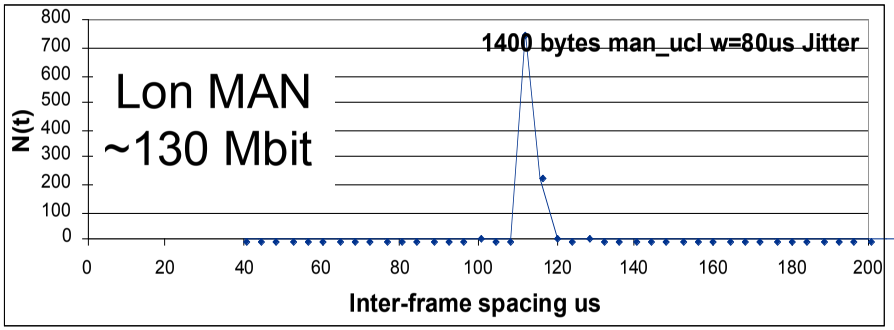


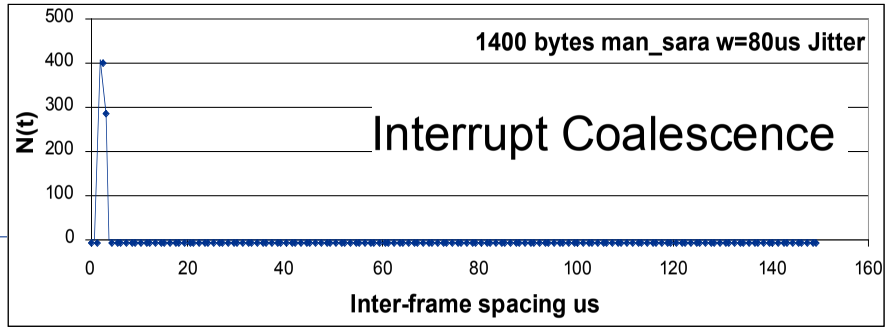
UDP Frame Jitter (Nov 01)

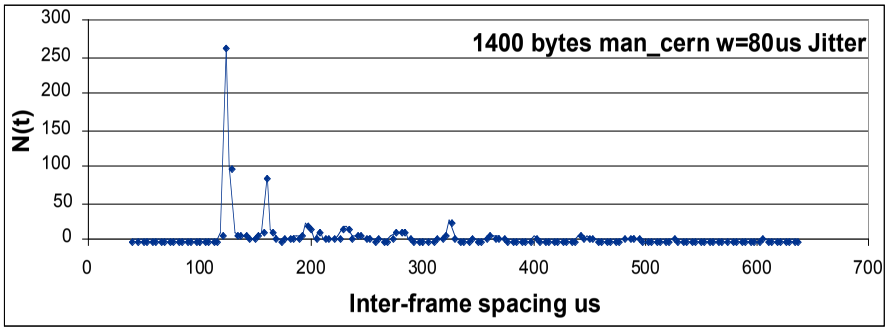
THE UNIVERSITY of MANCHESTER

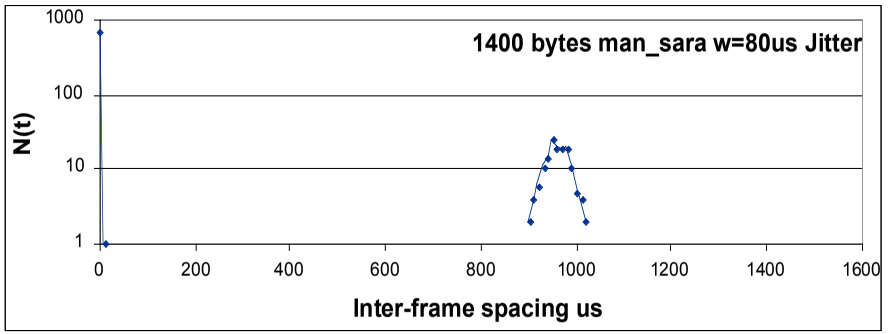












TCPstream

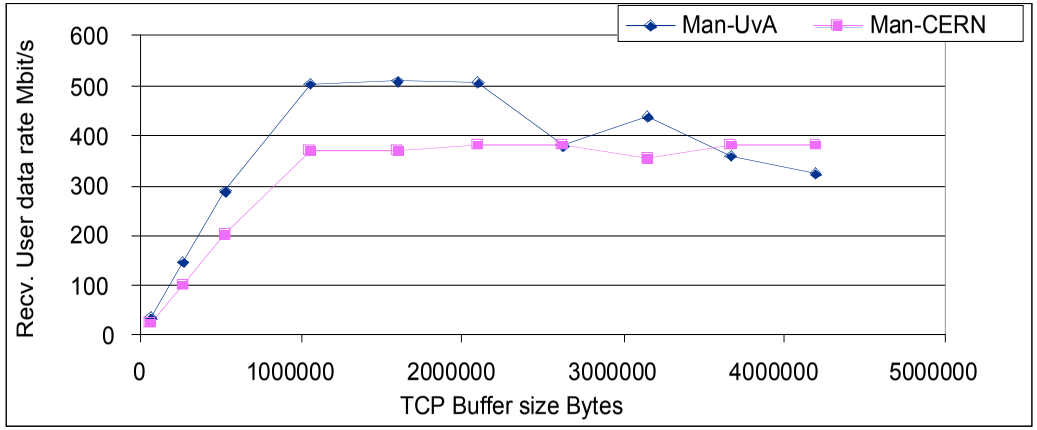
THE UNIVERSITY of MANCHESTER

- **TCP Throughput vs : msg size, transmit spacing, n-streams, window size**
- 1 control TCP link n test streams
- Memory – memory
- Only 1 process – no threads – no context switching
- Streams treated equally not round Robin but next available stream
- At the receiver record
  - The time of first and last message received
  - The amount of data
  - (CPU load, Number of interrupts)
- Use the Pentium CPU cycle counter for times and delay
  - Few lines of user code
- Tells us about:
  - Behavior of the IP stack
  - Available TCP throughput of the LAN / MAN / WAN

Gigabit TCP on the Production WAN (01)

THE UNIVERSITY of MANCHESTER

- **Throughput vs TCP buffer size**

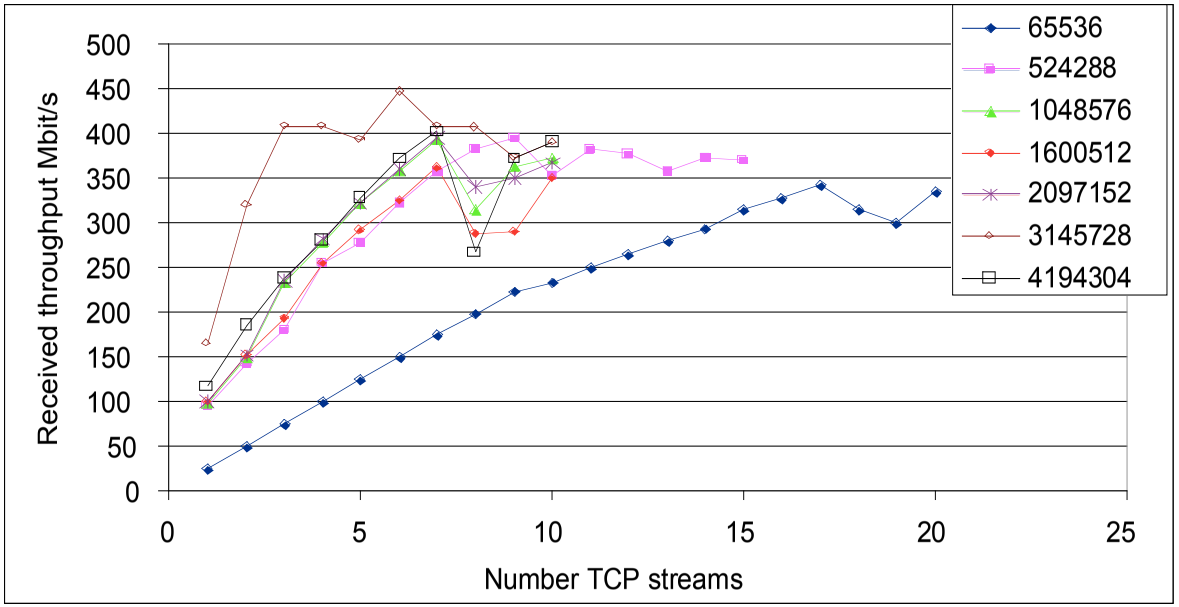


- TCP window sizes in Mbytes calculated from RTT*bandwidth

Link	RTT ms	Expected BW for 1 Gbit/s	BW Measured by UDP 750 Mbit/s	460 Mbit/s
Man – Ams	14.5	1.81	1.36	
Man – CERN	21.4	2.68		1.23

Gigabit TCP on the Production WAN (10)

THE UNIVERSITY of MANCHESTER



- **Throughput vs n-streams**
- Default buffer size slope = ~25 Mbit/s/stream up to 9 streams then 15 Mbit/s/stream
- With larger buffers rate of increase per stream is larger
- Plateaus at about 7 streams giving a total throughput of ~400 Mbit/s

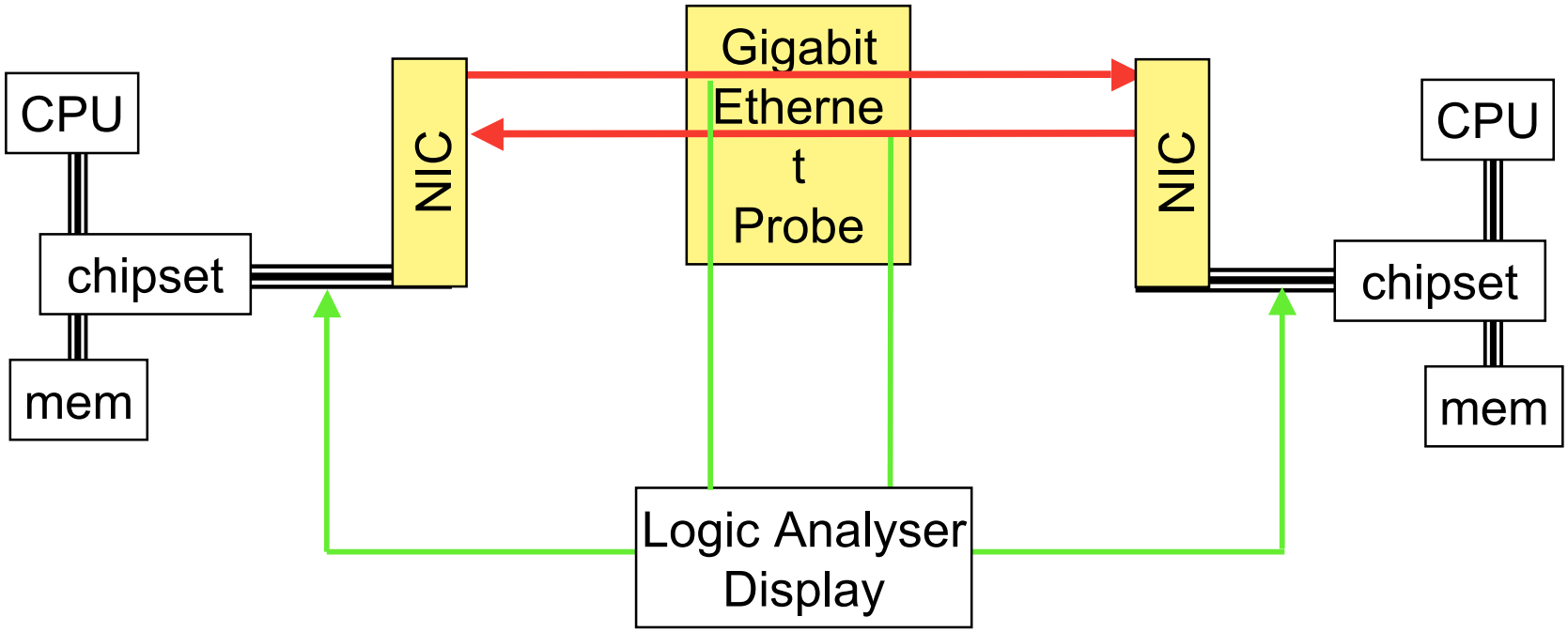
More Information Some URLs

- UDPmon / TCPstream 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
- Motherboard and NIC Tests:
  www.hep.man.ac.uk/~rich/net/nic/GigEth_tests_Boston.ppt
- PPNCG Home page with Stop Press:
  http://ppncg.rl.ac.uk/
- PPNCG Page for monitoring Grid, HEP & Astronomy Sites
  http://icfamon.dl.ac.uk/ppncg/astronomy.html
- IEPM PingER home site:
  http://www-iepm.slac.stanford.edu/
- IEPM-BW site:
  http://www-iepm.slac.stanford.edu/bw

The Measurements (4)

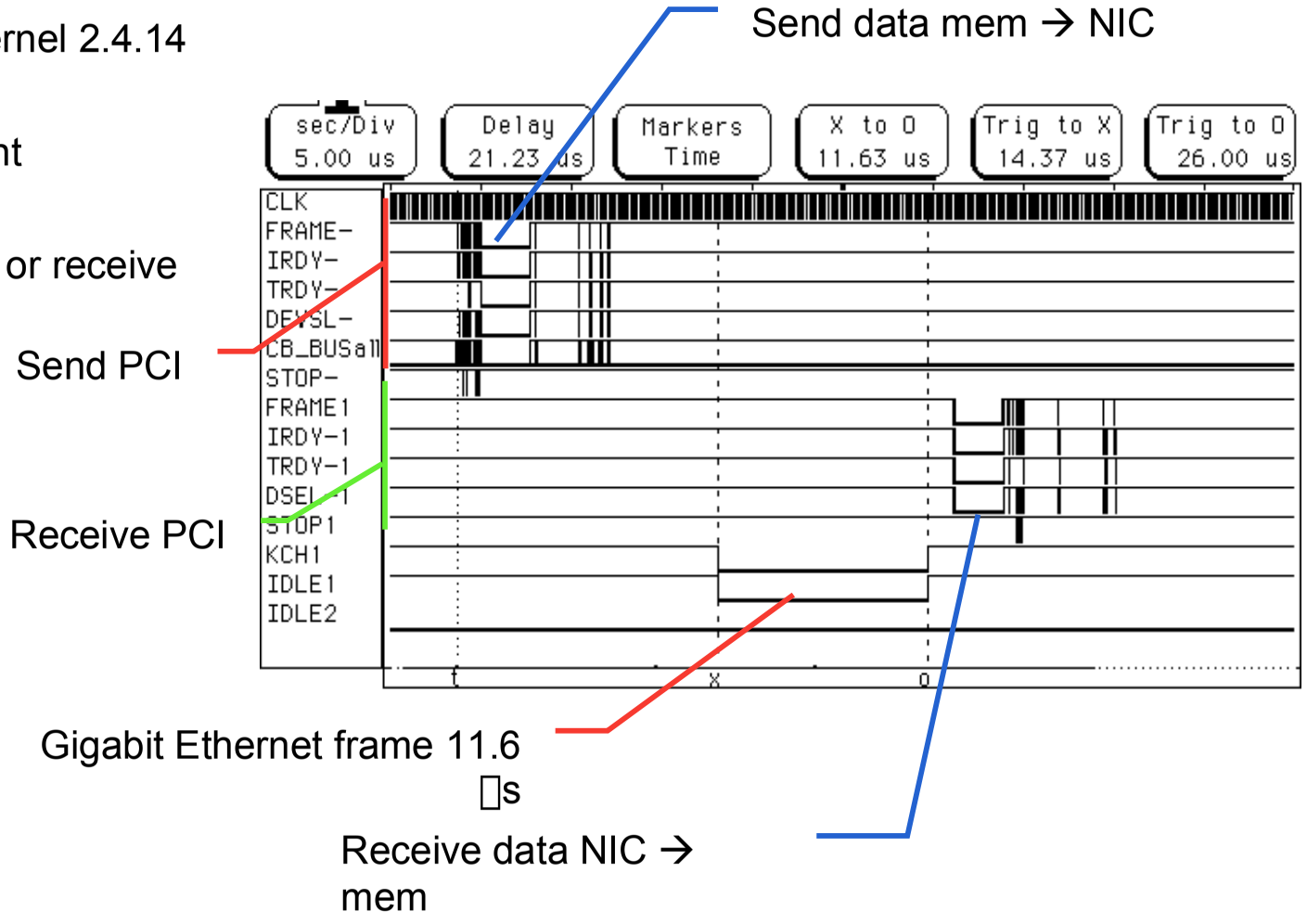
THE UNIVERSITY of MANCHESTER

- **PCI Activity**
- Logic Analyzer with
  - PCI Probe cards in sending PC
  - Gigabit Ethernet Fiber Probe Card
  - PCI Probe cards in receiving PC



PCI: SysKonnect SK-9843

- Motherboard: SuperMicro 370DLE Chipset: ServerWorks III LE Chipset
- CPU: PIII 800 MHz **PCI:64 bit 66 MHz**
- RedHat 7.1 Kernel 2.4.14
- SK300
- 1400 bytes sent
- Wait 100 us
- ~8 us for send or receive

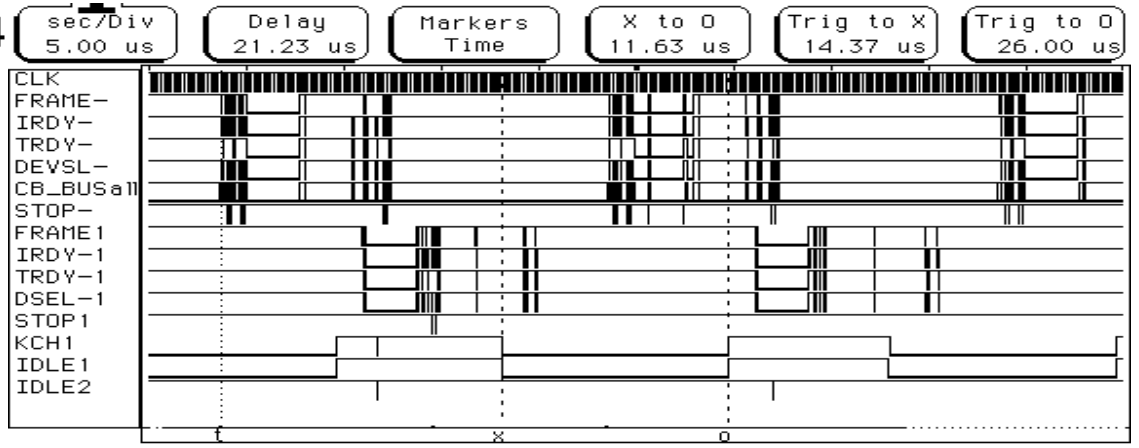


PCI: SysKonnect SK-9843

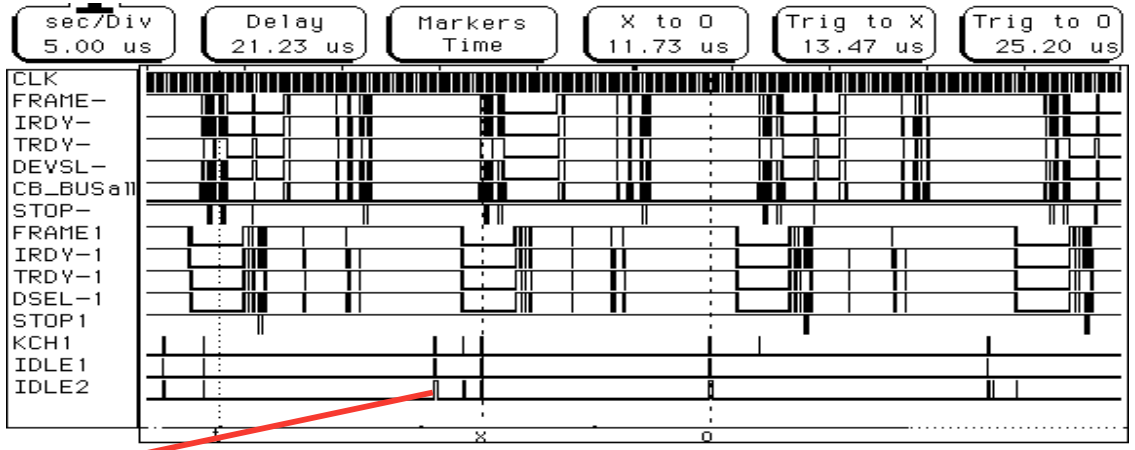
THE UNIVERSITY of MANCHESTER

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

- SK301
- 1400 bytes sent
- Wait 20 us



- Sk303
- 1400 bytes sent
- Wait 10 us
- **Frames are back-to-back**
- **Can drive at line speed**
- **Cannot go any faster !**



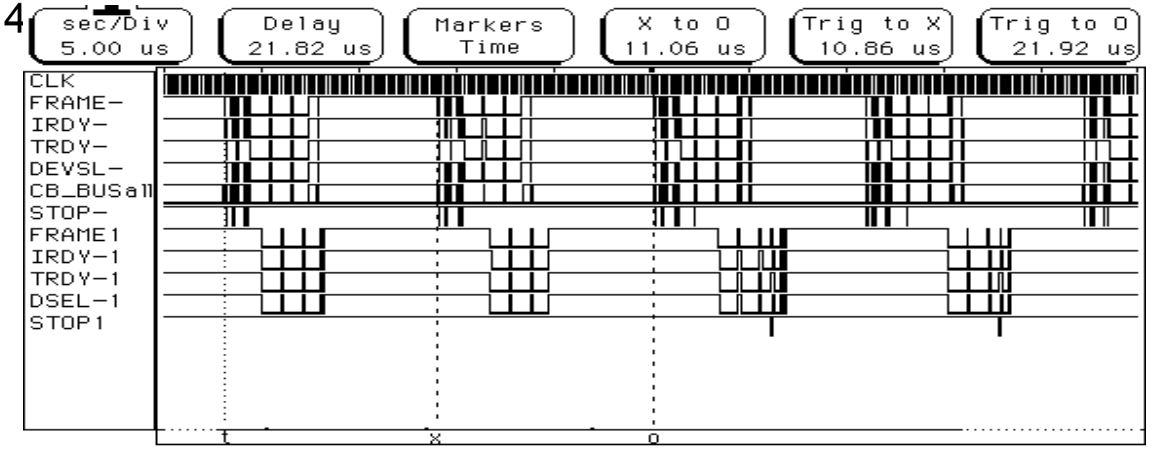
Gig Eth frames back to back

PCI: Intel Pro/1000

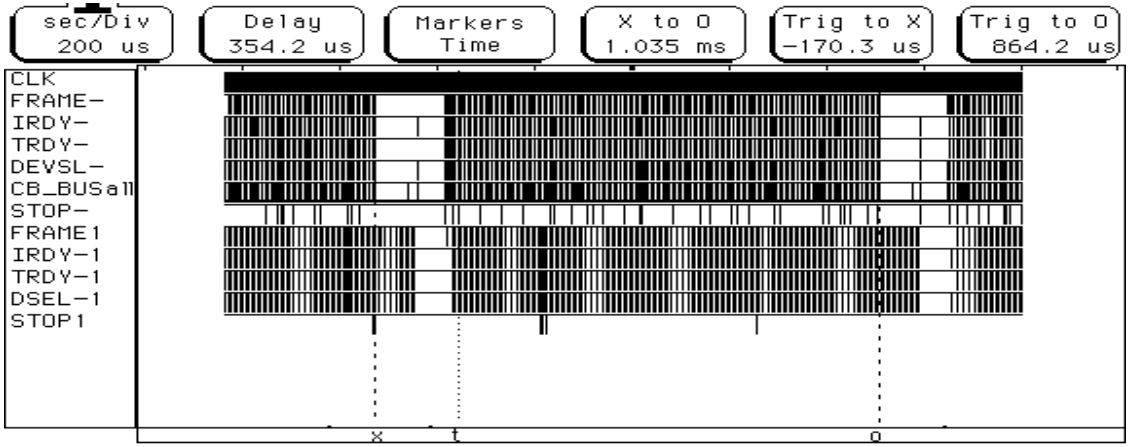
THE UNIVERSITY of MANCHESTER

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

- IT66M212
- 1400 bytes sent
- Wait 11 us
- ~4.7us on send PCI bus
- PCI bus ~45% occupancy
- ~ 3.25 us on PCI for data recv



- IT66M212
- 1400 bytes sent
- Wait 11 us
- Packets lost
- Action of pause packet?



PCI: Intel Pro/1000 on P4DP6

THE UNIVERSITY of MANCHESTER

- Motherboard: SuperMicro P4DP6 Chipset: Intel E7500 (Plumas)
- CPU: Dual Xeon Prestonia (2cpu/die) 2.2 GHz **Slot 3-5: PCI, 64 bit, 66 MHz**
- RedHat 7.2 Kernel 2.4.14

- ITP4010
- 1400 bytes sent
- Wait 8 us
- ~5.14us on send PCI bus
- PCI bus ~68% occupancy
- ~ 2 us on PCI for data recv