UDP Performance and PCI-X Activity of the Intel 10 Gigabit Ethernet Adapter on: HP rx2600 Dual Itanium 2 SuperMicro P4DP8-2G Dual Xenon Dell Poweredge 2650 Dual Xenon

Richard Hughes-Jones

Many people helped including:
Sverre Jarp and Glen Hisdal CERN Open Lab

Sylvain Ravot, Olivier Martin and Elise Guyot DataTAG project

Les Cottrell, Connie Logg and Gary Buhrmaster SLAC

Stephen Dallison MB-NG

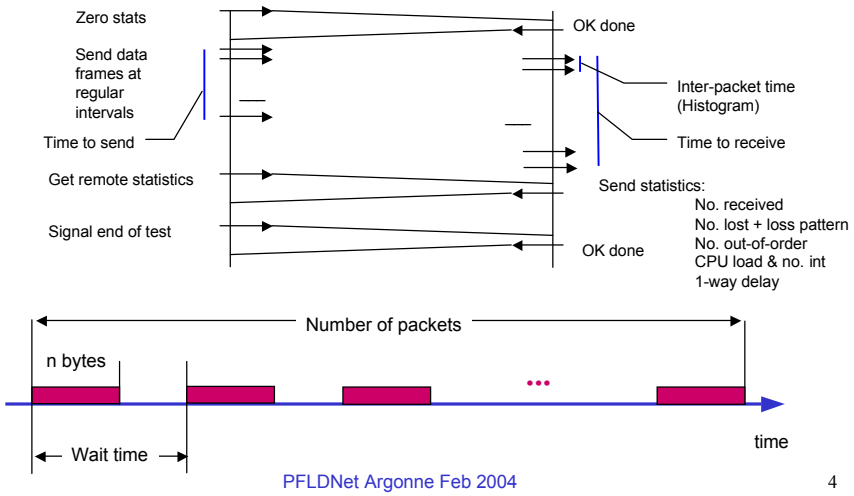
- Introduction
- 10 GigE on Itanium IA64
- 10 GigE on Xeon IA32
- 10 GigE on Dell Xeon IA32
- Tuning the PCI-X bus
- SC2003 Phoenix

Latency & Throughput Measurements

- **UDP/IP packets sent between back-to-back systems**
  - Similar processing to TCP/IP but no flow control & congestion avoidance algorithms
  - Used UDPmon test program
- **Latency**
  - Round trip times using Request-Response UDP frames
  - Latency as a function of frame size
    - Slope s given by: $s = \sum_{\text{data paths}} \left(\frac{db}{dt} \right)^{-1}$
    - Mem-mem copy(s) + pci + Gig Ethernet + pci + mem-mem copy(s)
    - Intercept indicates processing times + HW latencies
  - Histograms of 'singleton' measurements
- Tells us about:
  - Behavior of the IP stack
  - The way the HW operates
  - Interrupt coalescence
- **UDP Throughput**
  - Send a controlled stream of UDP frames spaced at regular intervals
  - Vary the frame size and the frame transmit spacing & measure:
    - The time of first and last frames received
    - The number packets received, lost, & out of order
    - Histogram inter-packet spacing received packets
    - Packet loss pattern
    - 1-way delay
    - CPU load
    - Number of interrupts
- Tells us about:
  - Behavior of the IP stack
  - The way the HW operates
  - Capacity & Available throughput of the LAN / MAN / WAN

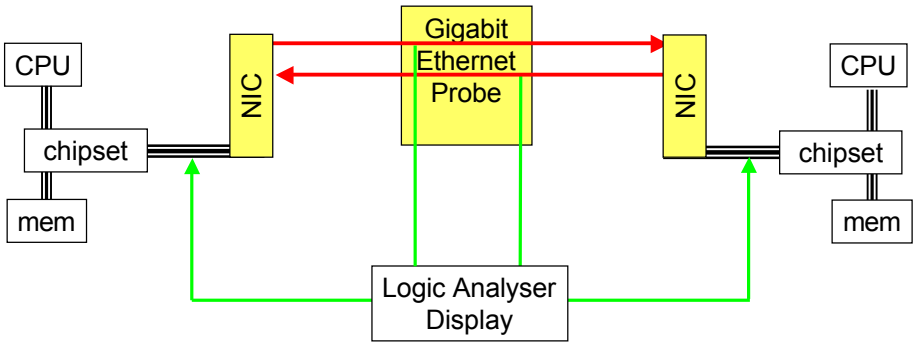
The Throughput Measurements

- **UDP Throughput**
- Send a controlled stream of UDP frames spaced at regular intervals



The PCI Bus & Gigabit Ethernet Measurements

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

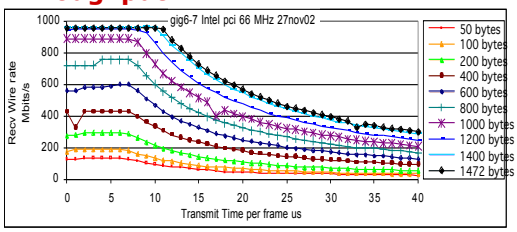


Example: The 1 Gigabit NIC Intel pro/1000

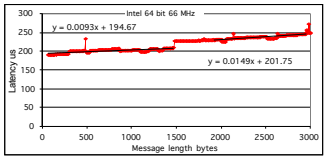
- Motherboard: Supermicro P4DP6
- Chipset: E7500 (Plumas)
- CPU: Dual Xeon 2 2GHz with 512k L2 cache
- Mem bus 400 MHz
  PCI-X 64 bit 66 MHz
- HP Linux Kernel 2.4.19 SMP
- MTU 1500 bytes

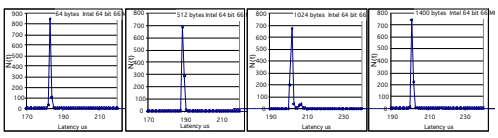
- **Intel PRO/1000 XT**

Throughput

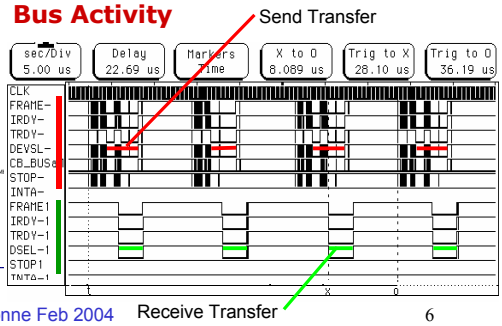


Latency



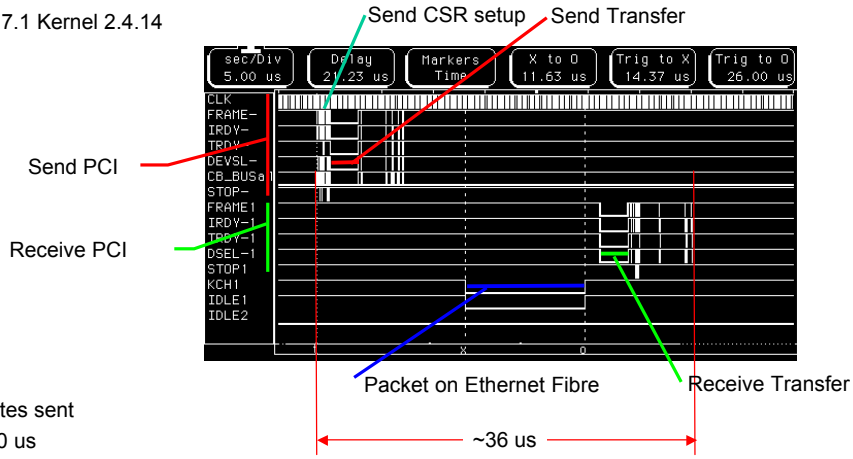


Bus Activity



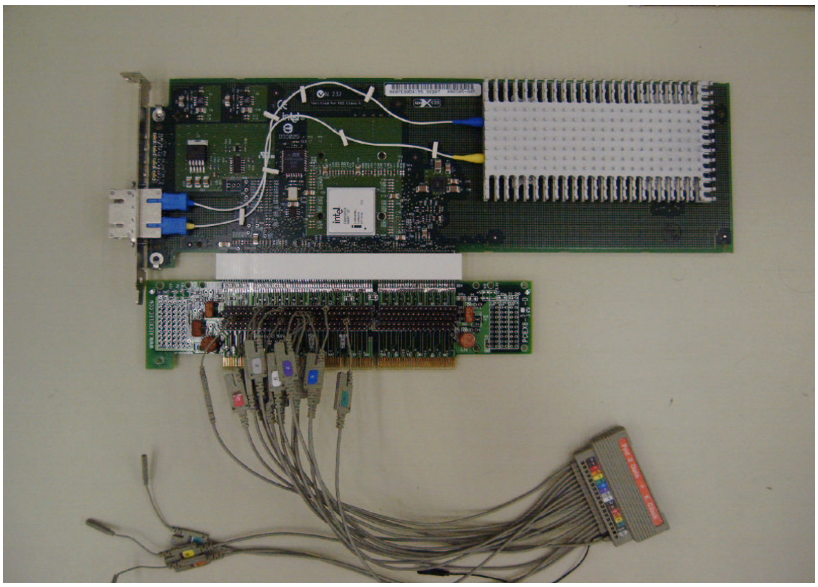
Data Flow: SuperMicro 370DLE: SysKonnect

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

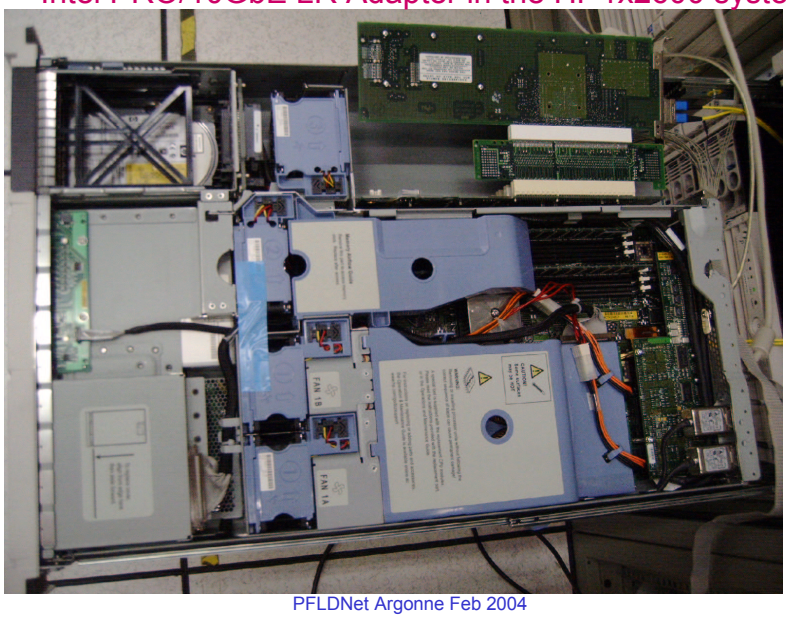


- 1400 bytes sent
- Wait 100 us
- **~8 us for send or receive**
- Stack & Application overhead ~ 10 us / node

10 Gigabit Ethernet NIC with the PCI-X probe card.



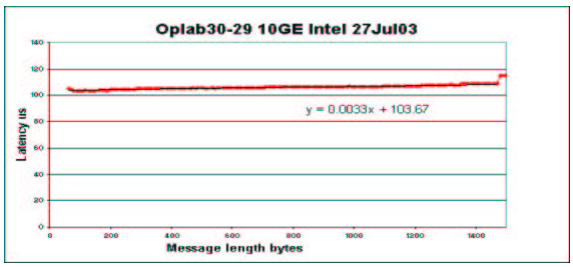
Intel PRO/10GbE LR Adapter in the HP rx2600 system

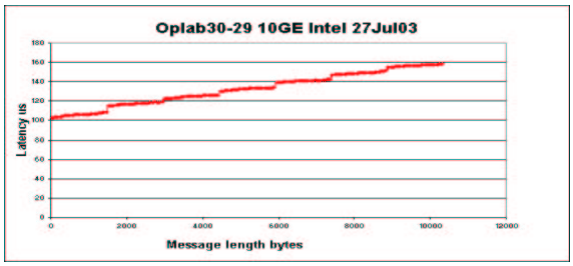


10 GigE on Itanium IA64: UDP Latency

- Motherboard: HP rx2600 IA 64
- Chipset: HPzx1
- CPU: Dual Itanium 2 1GHz with 512k L2 cache
- Mem bus dual 622 MHz 4.3 GByte/s
- PCI-X 133 MHz
- HP Linux Kernel 2.5.72 SMP
- **Intel PRO/10GbE LR Server Adapter**
- NIC driver with
  - RxIntDelay=0
  - XsumRX=1 XsumTX=1
  - RxDescriptors=2048 TxDescriptors=2048
- **MTU 1500 bytes**

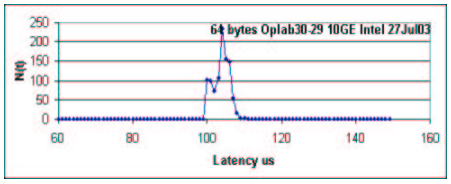
- Latency 100 µs & very well behaved
- Latency Slope **0.0033 µs/byte**
- B2B Expect: **0.00268 µs/byte**
  - PCI 0.00188 **
  - 10GigE 0.0008
  - PCI 0.00188

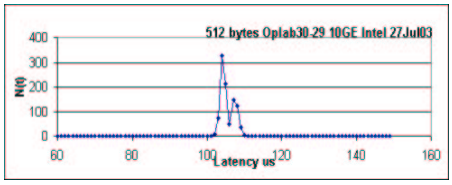


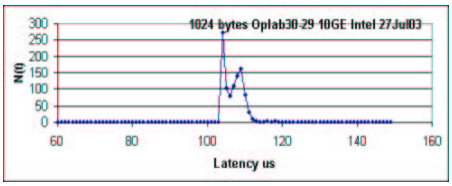


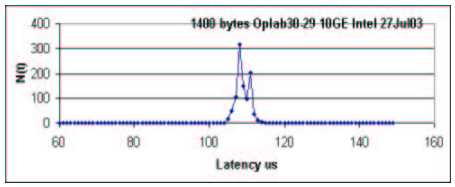
10 GigE on Itanium IA64: Latency Histograms

- Double peak structure with the peaks separated by 3-4 μs
- Peaks are ~1-2 μs wide
- Similar to that observed with 1 Gbit Ethernet NICs on IA32 architectures









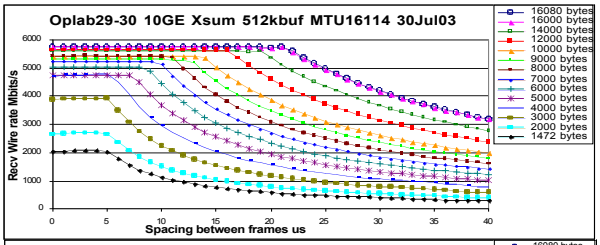
10 GigE on Itanium IA64: UDP Throughput

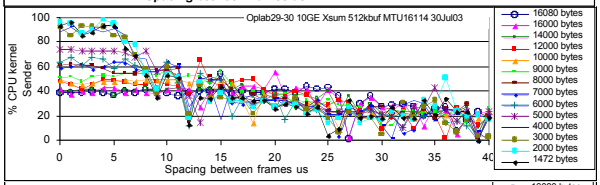
- **HP Linux Kernel 2.5.72 SMP**
- MTU 16114 bytes
- Max throughput **5.749 Gbit/s**

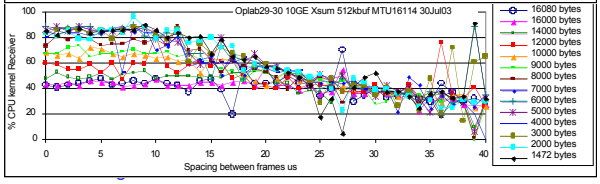
- Int on every packet
- No packet loss in 10M packets

- Sending host, 1 CPU is idle
- For 14000-16080 byte packets, one **CPU is 40% in kernel mode**
- As the packet size decreases load rises to ~90% for packets of 4000 bytes or less.

- Receiving host both CPUs busy
- **16114 bytes 40% kernel mode**
- Small packets 80 % kernel mode
- TCP gensink data rate was 745 MBytes/s = 5.96 Gbit/s







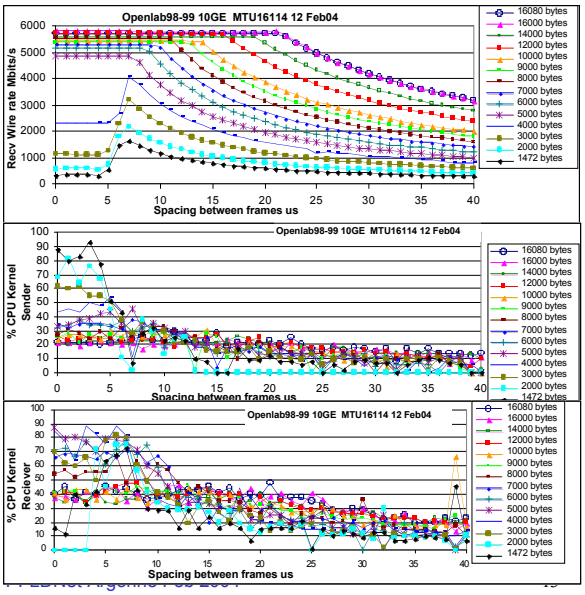
10 GigE on Itanium IA64: UDP Throughput [04]

- **HP Linux Kernel 2.6.1 #17 SMP**
- MTU 16114 bytes
- Max throughput **5.81 Gbit/s**

- Int on every packet
- Some packet loss pkts < 4000 bytes

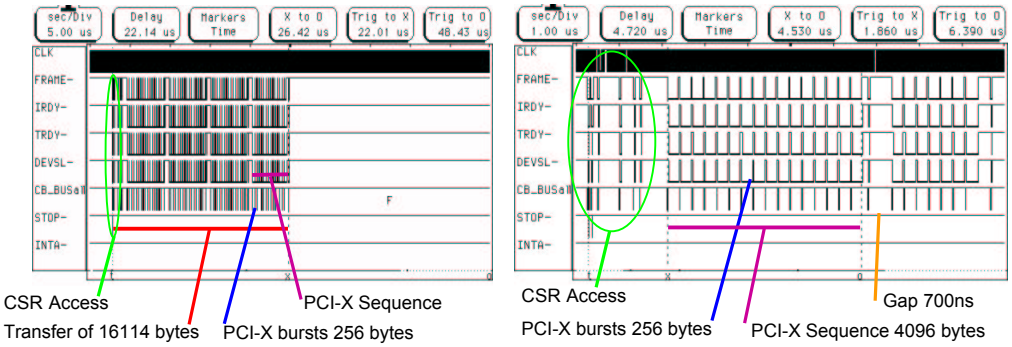
- Sending host, 1 CPU is idle – but swap over
- For 14000-16080 byte packets, one **CPU is 20-30% in kernel mode**
- As the packet size decreases load rises to ~90% for packets of 4000 bytes or less.

- Receiving host 1 CPU is idle – but swap over
- **16114 bytes 40% kernel mode**
- Small packets 70 % kernel mode



10 GigE on Itanium IA64: PCI-X bus Activity

- **16080 byte packets every 200 µs** **Intel PRO/10GbE LR Server Adapter MTU 16114**
- **setpci -s 02:1.0 e6.b=2e (22 26 2a)** **mmrbc 4096 bytes (512 1024 2048)**

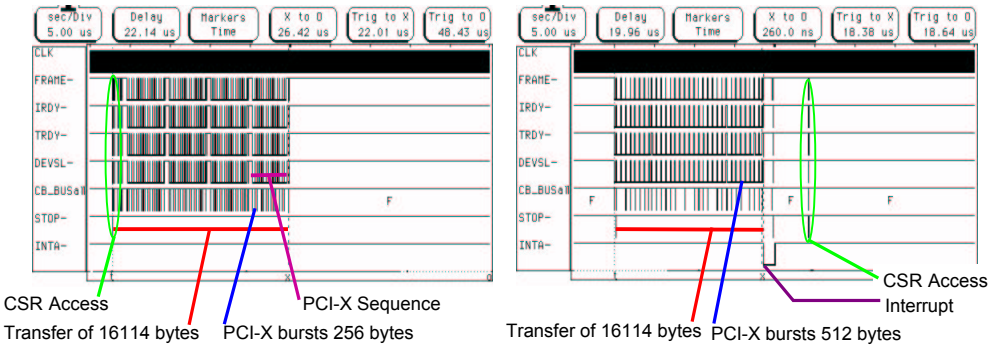


- **PCI-X Signals transmit - memory to NIC**
- **Interrupt and processing: 48.4 µs after start**
- **Data transfer takes ~22 µs**
- **Data transfer rate over PCI-X: 5.86 Gbit/s**

- **Made up of 4 PCI-X sequences of ~4.55 µs then a gap of 700 ns**
- **Sequence contains 16 PCI bursts 256 bytes**
- **Sequence length 4096 bytes (mmrbc)**

10 GigE on Itanium IA64: PCI-X bus Activity

- 16080 byte packets every 200 µs Intel PRO/10GbE LR Server Adapter MTU 16114
- setpci -s 02:1.0 e6.b=2e (22 26 2a) mmrbc 4096 bytes (512 1024 2048)



- PCI-X Signals transmit - memory to NIC
- Interrupt and processing: 48.4 µs after start
- Data transfer takes ~22 µs
- Data transfer rate over PCI-X: 5.86 Gbit/s

- PCI-X Signals receive – NIC to memory
- Interrupt every packet
- Data transfer takes ~18.4 µs
- Data transfer rate over PCI-X : 7.014 Gbit/s
- Note: receive is faster cf the 1 GE NICs

10 GigE on Xeon IA32: UDP Latency

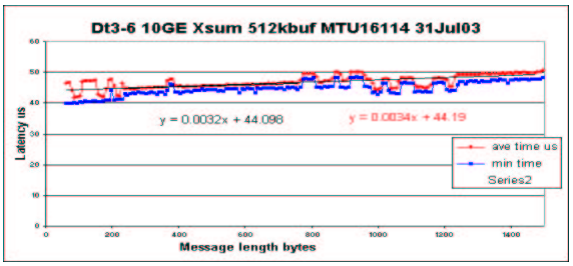
- Motherboard: Supermicro P4DP8-G2
- Chipset: Intel E7500 (Plumas)
- CPU: Dual Xeon 2.2GHz with 512k L2 cache
- Mem bus 400 MHz

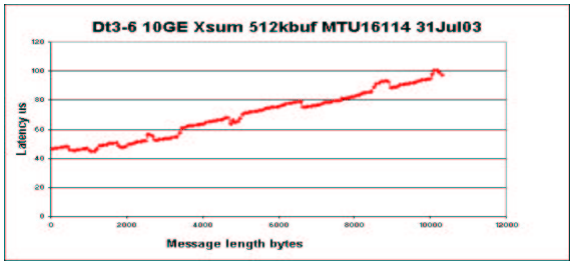
- PCI-X 133 MHz
- RedHat Kernel 2.4.21 SMP
- Intel(R) PRO/10GbE Network Driver v1.0.45

- **Intel PRO/10GbE LR Server Adapter**

- NIC driver with
  - RxIntDelay=0
  - XsumRX=1 XsumTX=1
  - RxDescriptors=2048 TxDescriptors=2048
- MTU 1500 bytes

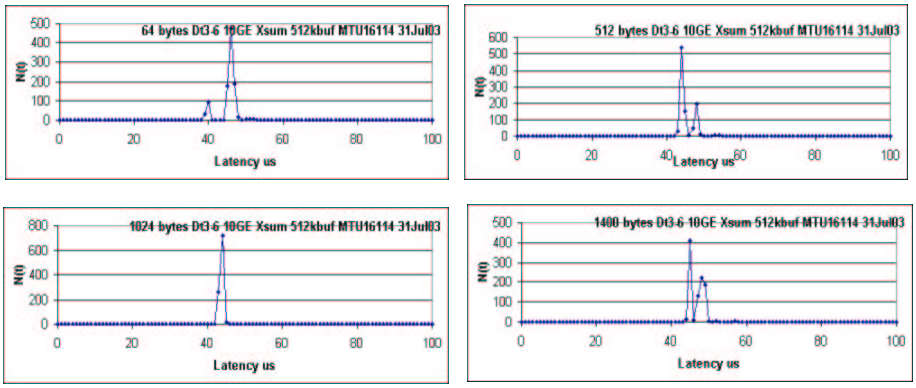
- Latency 144 µs & reasonably behaved
- Latency Slope **0.0032 µs/byte**
- B2B Expect: **0.00268 µs/byte**





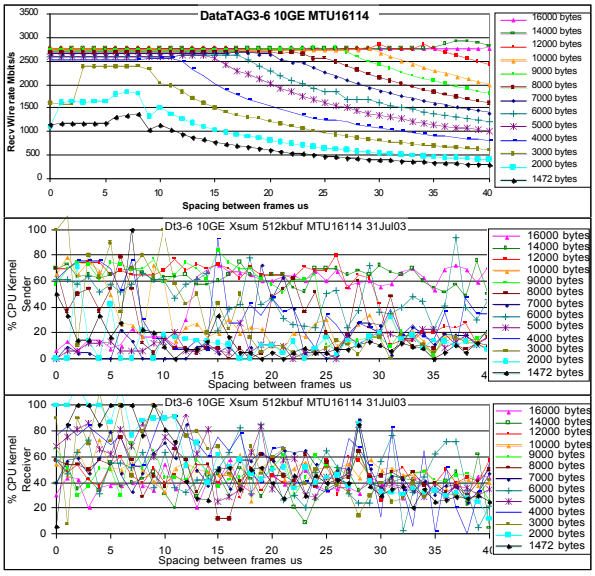
10 GigE on Xeon IA32: Latency Histograms

- Double peak structure with the peaks separated by 3-4 μs
- Peaks are ~1-2 μs wide
- Simliar to that observed with 1 Gbit Ethernet NICs on IA32 architectures



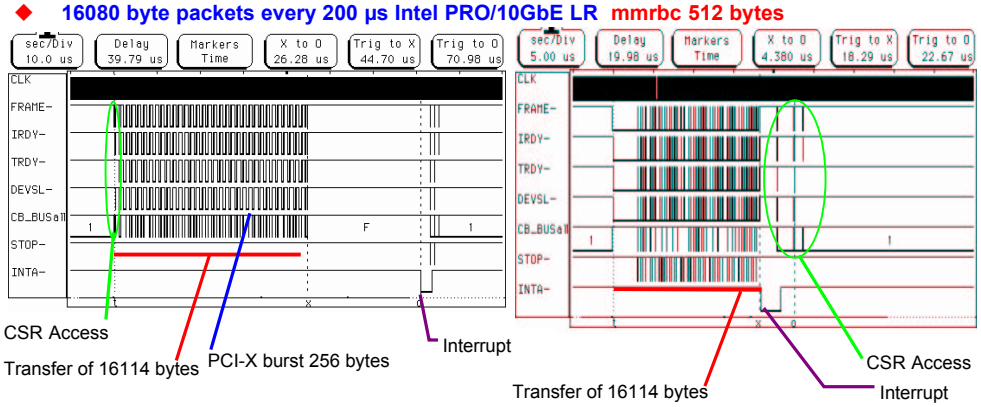
10 GigE on Xeon IA32: Throughput

- MTU 16114 bytes
- Max throughput **2.75 Gbit/s** mmrbc 512
- Max throughput 3.97 Gbit/s mmrbc 4096 bytes
- Int on every packet
- No packet loss in 10M packets
- Sending host,
- For closely spaced packets, the other CPU is ~60-70 % in kernel mode
- Receiving host
- Small packets 80 % in kernel mode
- >9000 bytes ~50% in kernel mode



10 GigE on Xeon IA32: PCI-X bus Activity

- **16080 byte packets every 200 µs Intel PRO/10GbE LR mmrbc 512 bytes**



- **PCI-X Signals transmit - memory to NIC**
- **Interrupt and processing: 70 µs after start**
- **Data transfer takes ~44.7 µs**
- **Data transfer rate over PCI-X: 2.88Gbit/s**

- **PCI-X Signals receive – NIC to memory**
- **Interrupt every packet**
- **Data transfer takes ~18.29 µs**
- **Data transfer rate over PCI-X : 7.014Gbit/s same as Itanium**

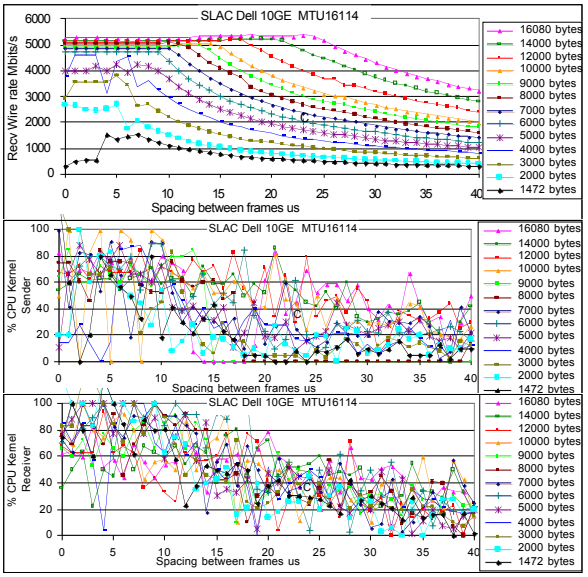
10 GigE on Dell Xeon: Throughput

- MTU 16114 bytes
- Max throughput **5.4 Gbit/s**

- Int on every packet
- Some packet loss pkts < 4000 bytes

- Sending host,
- For closely spaced packets, one CPU is ~70% in kernel mode
- CPU usage swaps

- Receiving host
- 1 CPU is idle but CPU usage swaps
- For closely spaced packets ~80 % in kernel mode



10 GigE on Dell Xeon : UDP Latency

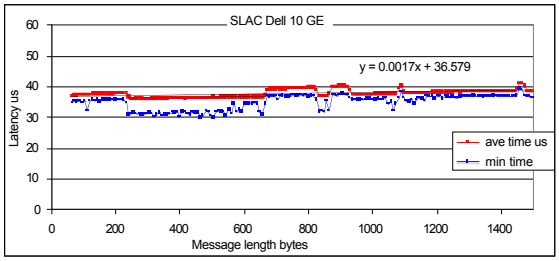
- Motherboard: Dell Poweredge 2650
- Chipset: Intel E7500 (Plumas)
- CPU: Dual Xeon 3.06 GHz with 512k L2 cache
- Mem bus 533 MHz

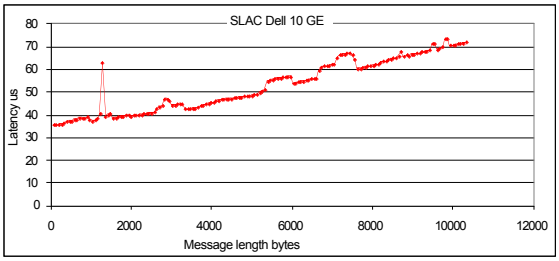
- PCI-X 133 MHz
- RedHat Kernel 2.4.20 altAIMD
- Intel(R) PRO/10GbE Network Driver v1.0.45

- **Intel PRO/10GbE LR Server Adapter**

- NIC driver with
  - RxIntDelay=0
  - XsumRX=1 XsumTX=1
  - RxDescriptors=2048 TxDescriptors=2048
- MTU 16114 bytes

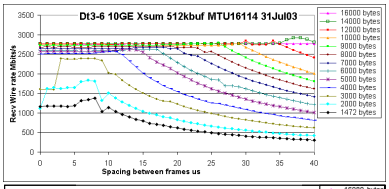
- Latency 36 µs with some steps
- Latency Slope **0.0017 µs/byte**
- B2B Expect: **0.00268 µs/byte**



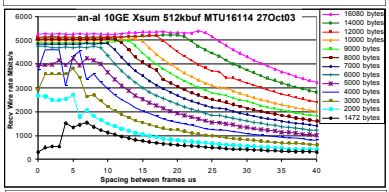


10 GigEthernet: Throughput

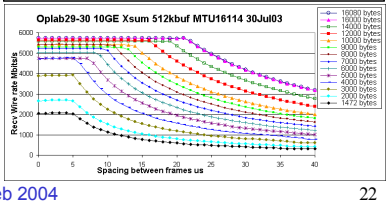
- **1500 byte MTU gives ~ 2 Gbit/s**
- **Used 16144 byte MTU max user length 16080**
- **DataTAG Supermicro PCs**
- **Dual 2.2 GHz Xeon CPU FSB 400 MHz**
- **PCI-X mmrbc 512 bytes**
- **wire rate throughput of 2.9 Gbit/s**



- **SLAC Dell PCs giving a**
- **Dual 3.0 GHz Xeon CPU FSB 533 MHz**
- **PCI-X mmrbc 4096 bytes**
- **wire rate of 5.4 Gbit/s**

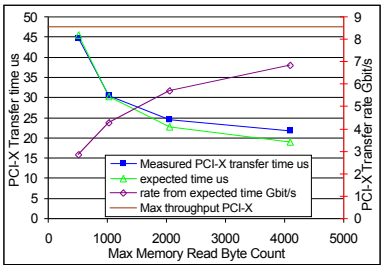


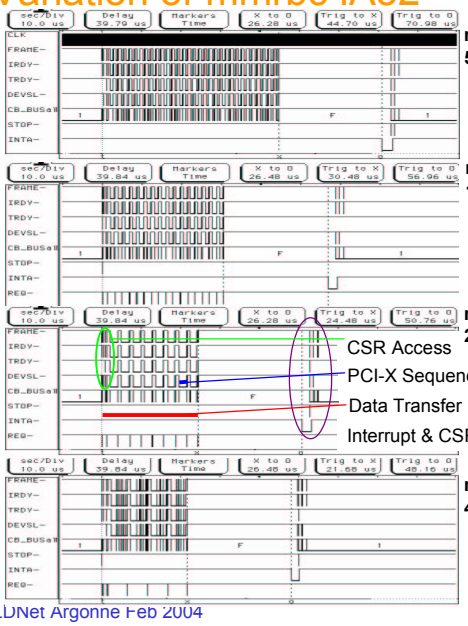
- **CERN OpenLab HP Itanium PCs**
- **Dual 1.0 GHz 64 bit Itanium CPU FSB 400 MHz**
- **PCI-X mmrbc 4096 bytes**
- **wire rate of 5.7 Gbit/s**



Tuning PCI-X: Variation of mmrbc IA32

- 16080 byte packets every 200 µs
- Intel PRO/10GbE LR Adapter
- PCI-X bus occupancy vs mmrbc
- Plot:
  - Measured times
  - Times based on PCI-X times from the logic analyser
  - Expected throughput





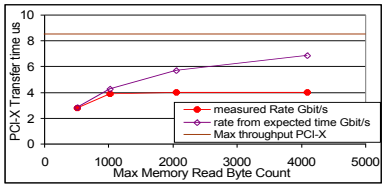
Tuning PCI-X: Throughput vs mmrbc

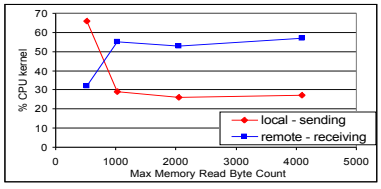
- DataTag IA32
- 2.2 GHz Xeon
- 400 MHz FSB
- 2.7 – 4.0 Gbit.s

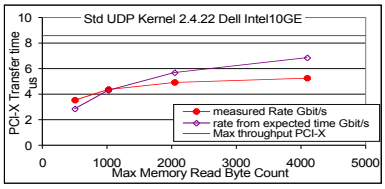
- SLAC Dell
- 3.0 GHz Xeon
- 533 MHz FSB
- 3.5 - 5.2 Gbit/s

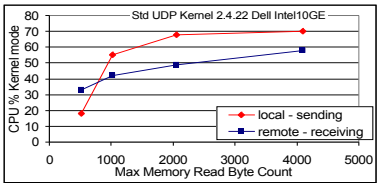
- OpenLab IA64
- 1.0 GHz Itanium
- 622 MHz FSB
- 3.2 - 5.7 Gbit/s

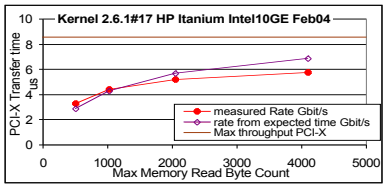
- CPU load is 1 CPU not average

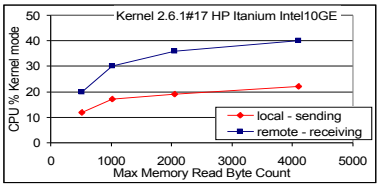








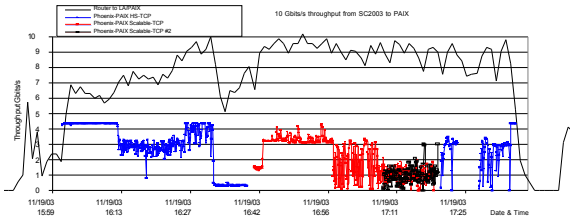




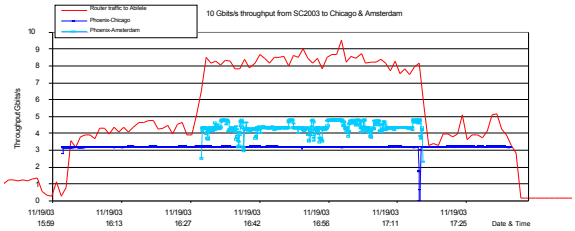
10 GigEthernet at SC2003 BW Challenge

- Three Server systems with 10 GigEthernet NICs
- Used the DataTAG altAIMD stack 9000 byte MTU
- Send mem-mem iperf TCP streams From SLAC/FNAL booth in Phoenix to:

  - Pal Alto PAIX
  - rtt 17 ms , window 30 MB
  - Shared with Caltech booth
  - 4.37 Gbit hstcp I=5%
  - Then 2.87 Gbit I=16%
  - Fall corresponds to 10 Gbit on link
  - 3.3Gbit Scalable I=8%
  - Tested 2 flows sum 1.9Gbit I=39%



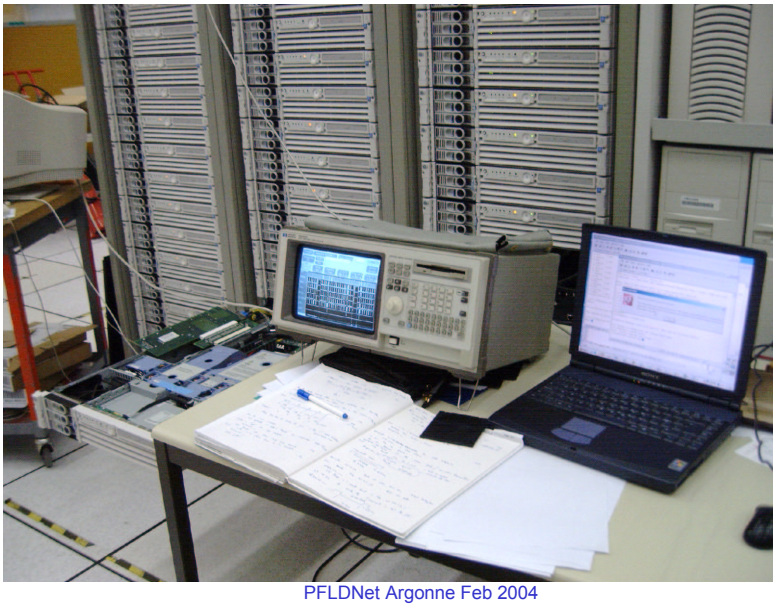
  - Chicago Starlight
  - rtt 65 ms , window 60 MB
  - Phoenix CPU 2.2 GHz
  - 3.1 Gbit hstcp I=1.6%
  - Amsterdam SARA
  - rtt 175 ms , window 200 MB
  - Phoenix CPU 2.2 GHz
  - 4.35 Gbit hstcp I=6.9%
  - Very Stable
  - Both used Abilene to Chicago



Summary & Conclusions

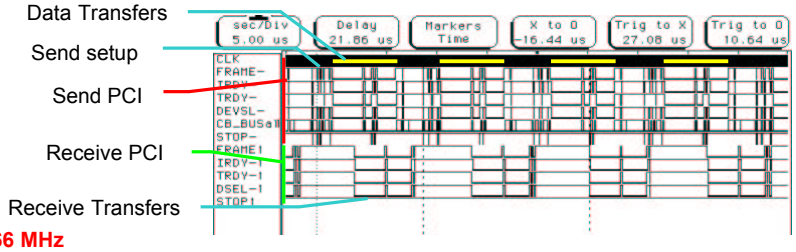
- Intel PRO/10GbE LR Adapter and driver gave stable throughput and worked well
- Need large MTU (9000 or 16114) – 1500 bytes gives ~2 Gbit/s
- PCI-X tuning mmrbc = 4096 bytes increase by 55% (3.2 to 5.7 Gbit/s)
- PCI-X sequences clear on transmit gaps ~ 950 ns
- Transfers: transmission (22 µs) takes longer than receiving (18 µs)
- Tx rate 5.85 Gbit/s Rx rate 7.0 Gbit/s (Itanium) (PCI-X max 8.5Gbit/s)
- CPU load considerable 60% Xenon 40% Itanium
- BW of Memory system important – crosses 3 times!
- Sensitive to OS/ Driver updates
- More study needed

Test setup with the CERN Open Lab Itanium systems



1 GigE on IA32: PCI bus Activity 33 & 66 MHz

- **1472 byte packets every 15 µs Intel Pro/1000**
- **PCI:64 bit 33 MHz**
- **82% usage**



- **PCI:64 bit 66 MHz**
- **65% usage**
- **Data transfers half as long**