ETHERNET V. INFINIBAND

Reliable Transport

- InfiniBand uses hardware based retransmission
- InfiniBand uses both link level and end-to-end CRC’s
- Ethernet is a best efforts delivery, allowed to drop packets and relies on the TCP/IP protocol for reliability which is typically implemented in SW for retransmission
- The effort to implement TCP/IP in hardware has been proven much more challenging than what people imagined. TCPoffloads cards have not been very successful and have not been shown to lower latency
- TCP/IP is the major performance bottleneck for bandwidths of 10G and above
- InfiniBand delivers reliability at the hardware level providing higher throughput, less latency and rarely causes jitter. This enables the use without TCP/IP

Flow Control

- InfiniBand uses credit based flow control for each link which means that InfiniBand switch chips can be built with much smaller on-chip buffers than Ethernet
- Ethernet switches rely on explicit packet drops for flow control, which requires larger buffers because the cost of retransmission is very high.
- This technical difference enables the building larger and lower cost switch chips for InfiniBand vs. Ethernet
- This has resulted in larger, higher density, lower contention, lower cost InfiniBand switches with lower cost per port than their 10Ge equivalents
  - maximum 40G InfiniBand, zero contention port density is 648 ports
  - maximum 10G Ethernet, zero contention port density is 384 ports

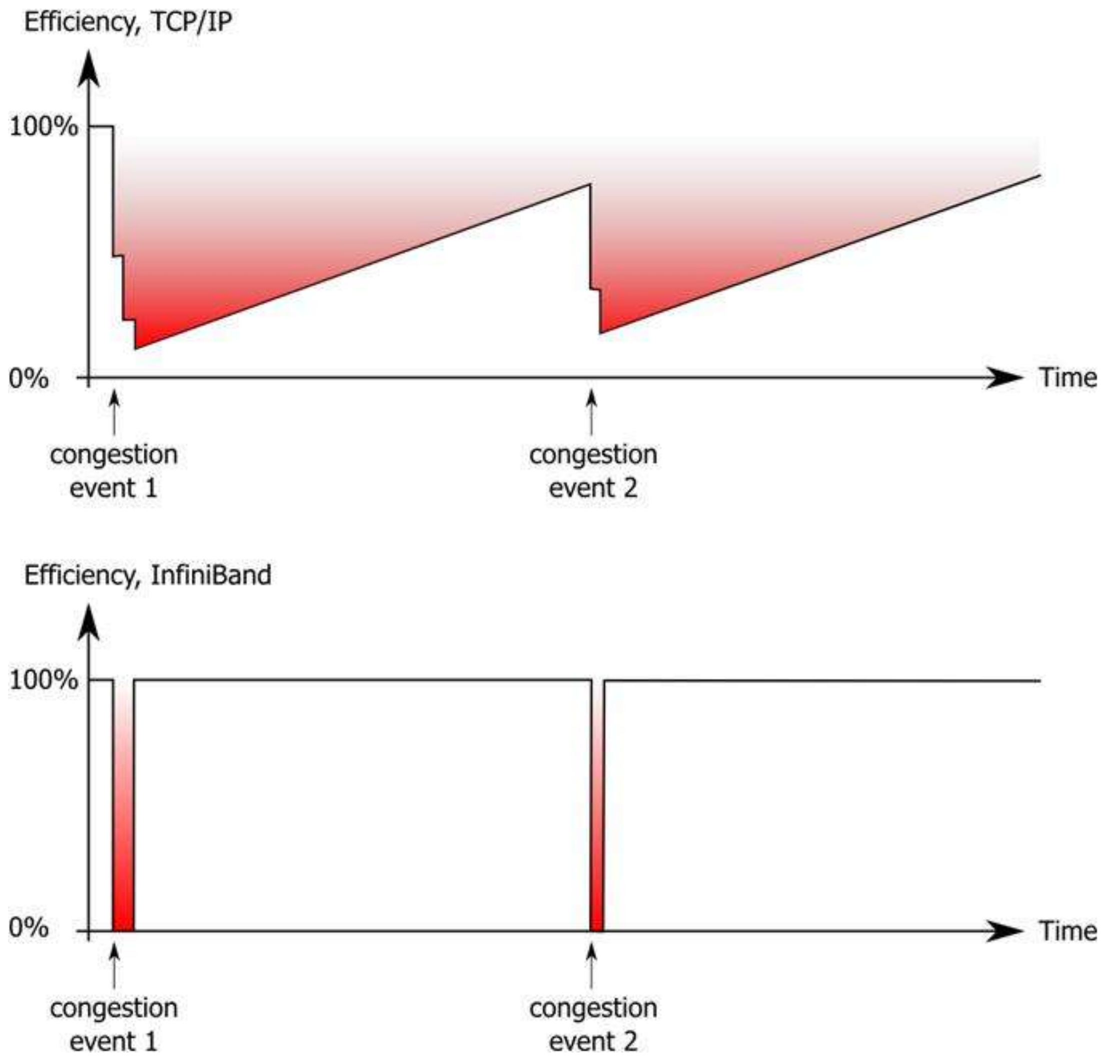
- InfiniBand has late packet invalidation which enables cut-through switching for low latency non-blocking performance spanning the entire fabric
- Virtually all Ethernet switches are designed for L3/L4 switching, which requires packet rewrite, and which requires a store-forward architecture
- Where Ethernet supports cut-through, reliable deployment is limited to local clusters and small subnets due to the need to prevent propagation of invalid packets.

The latency impact of store forward is quite significant:
for a 500Byte packet at 1 Gbps it is 5.7 usec, for a 1500 Byte packet at 1 Gbps it is 16.3 usec

Store and forward adds this overhead for every hop!

- InfiniBand has end to end Congestion management as part of the existing standard (IBA 1.2)
  - Congestion detection at receiver sends notification messages to sender to reduce rate
  - Policies determine recovery rates
- Ethernet relies on TCP
- Issues with current TCP congestion management algorithms on high bandwidth, long haul circuits limit single session throughput due to the window sizing halving for each congestion event

TCP/IP recovery for congestion events



Courtesy of Obsidian Research

- InfiniBand includes a complete management stack which provides high levels of configurability and observability.
- Ethernet relies on TCP/IP to correct errors
- InfiniBand enables an end-to-end solution to be deployed and run reliably without the overhead of TCP-IP
- Ethernet relies on add-ons such as trunking and spanning tree to add resiliency into Layer 2 networks.
- Ethernet Spanning tree is active:standby and takes seconds to switch. This switching causes multicast flooding in most switches.
- InfiniBand preselects failover paths and switches almost instantly

The challenges of reliable networks

- Reliable delivery has a downside. The packets have to be saved somewhere before they can be delivered
- InfiniBand transfers this buffering from the TCP socket on the server to the network
- All Credit-based networks suffer from congestion
- When receivers are not fast enough, packets build up in the network. This backpressure slows the sender but causes congestion on shared links potentially impacting other nodes or applications
  - The Ethernet Pause feature has similar impact and is how FCoE achieves reliable delivery
- InfiniBand uses independent Tx/Rx buffers for each Virtual Lane to mitigate this impact
- Requires careful QoS design to minimise the onset of congestion and utilize the VL's

The challenges of mesh networks

- Mesh networks present multiple active paths to link two nodes
- Spanning tree solves this by pruning the mesh and reducing it to one active path
- Applications, particularly those using RDMA require in-order delivery. This can only be achieved by having a fixed path between two nodes
- This constrains bandwidth usage and requires more sophisticated path selection algorithms with load balancing
- Reconfiguration events can result in looping packets consuming all bandwidth
- Topology design and path selections has to address potential loops

- RDMA is part of the standard with InfiniBand. It is a mandatory requirement and has been extensively interoperability tested. This allows multivendor configurations to be safely deployed
- RDMA on Ethernet mostly requires a matched pair of NIC's.
- iWARP added the (latency) cost of TCP to overcome Ethernets reliability problems
- RDMAoEthernet is an emerging standard and relies on FCoE reliable delivery to avoid the need for a TCP layer. This then requires a Convergence Enhanced Ethernet NIC to be fully deployed end-to-end for RDMA
- InfiniBand RDMA, written by OFED, is standard in Linux 2.6.14 and later. Torvold will not permit another RDMA implementation in the stock kernel. Ethernet manufacturers are slowly adding OFED support to their cards

- Promise of 10GbaseT has always been held out to lower Ethernet prices.
- Technical challenges of running 10G over RG45 has been immense
  - Requires 6W to drive
  - Needs Cat-6A or Cat-7 cable so will rarely run over existing infrastructure
  - Uses a block level error detection scheme. Requires full block to be loaded. Adds 2µS to every hop
  - Few vendors support 10GbaseT for these reasons
- SFP+ is most popular 10G option.
  - Small form factor gives same packing density as RG45
  - 1W and 100nS latency
  - Available in both Cu and Optical (LC format)
  - Comparable to QFSP used by InfiniBand (and 40G Ethernet)

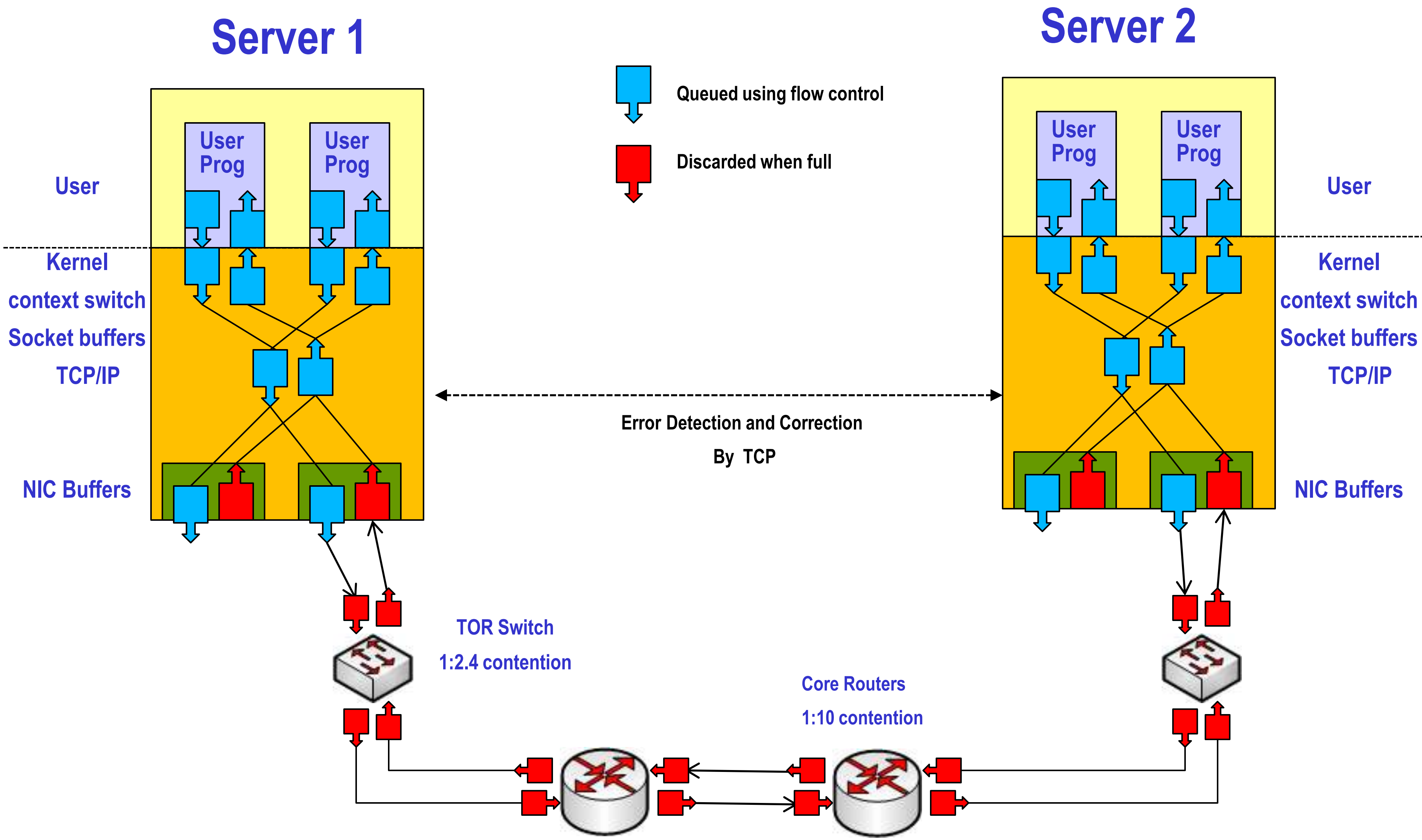
Long Distance

- InfiniBand is commonly viewed as fit for local clusters only. This is incorrect and was caused by the fat and short copper cables.
- InfiniBand and Ethernet share the same cabling at 40G an above. At 10G cables are similar (SFP+ v. QFSP) and have similar physical constraints
- InfiniBand Fibre cables are available up to 4km
- The higher scalability of InfiniBand subnets (48K ports) means that remote sites can be safely bridged without incurring the penalties of routing delays
- Long distance InfiniBand switches provide the necessary packet buffering to support distances of thousands of miles - e.g. US DoD coast-to-coast InfiniBand fabric

Advantages of InfiniBand over Ethernet

- Cut through design with hardware generated link and end2end CRC's and late packet invalidation
  - Avoids packet buffering required by Ethernet
  - 5uS compared with 20uS Ethernet latency
- Implicit layer 2 trunking, bundles of 1,4,12 physical links into a single “logical” channel. Handled transparently by the hardware
  - Ethernet trunking is vendor option, implemented in NIC driver rather than hardware.
  - Ethernet confused by competing standards e.g. Cisco Etherchannel
  - Ethernet does not stripe an individual packet whilst InfiniBand does
- Standardized RDMA to lower CPU overhead
  - Ethernet currently has vendor specific RDMA, requires matched pairs of cards and device driver support. Effort to standardize ongoing in Ethernet community
- Legacy Ethernet protocol constrains large switch implementation – max possible today:
  - InfiniBand 648 port zero contention 40Gbps, 3052 ports at 20Gbps, no contention.
  - Cisco Nexus 7000 32 port 10G blade with a total of 8 per chassis (256 ports) but limited to 80G fabric i.e. 8:1 contention. Still not shipping.

Ethernet Packet Drops and discards



Impact on latency and jitter

Whilst latency savings are only small for 10ge v. InfiniBand, their is a big advantage with less jitter for InfiniBand compared with Ethernet

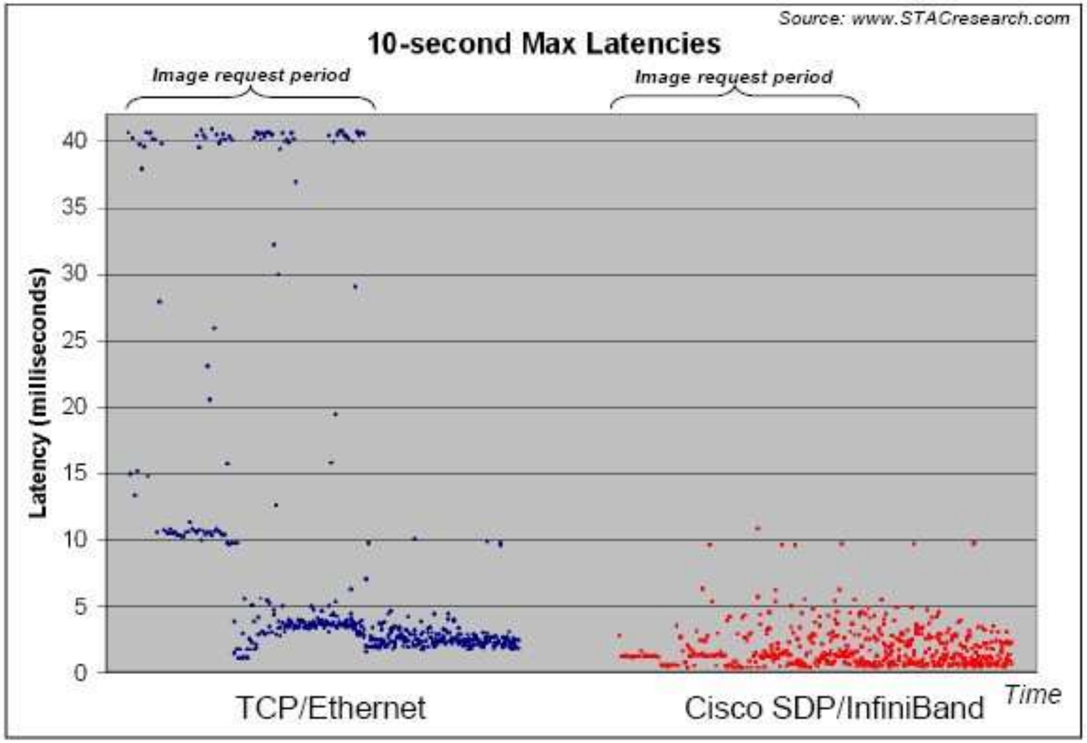
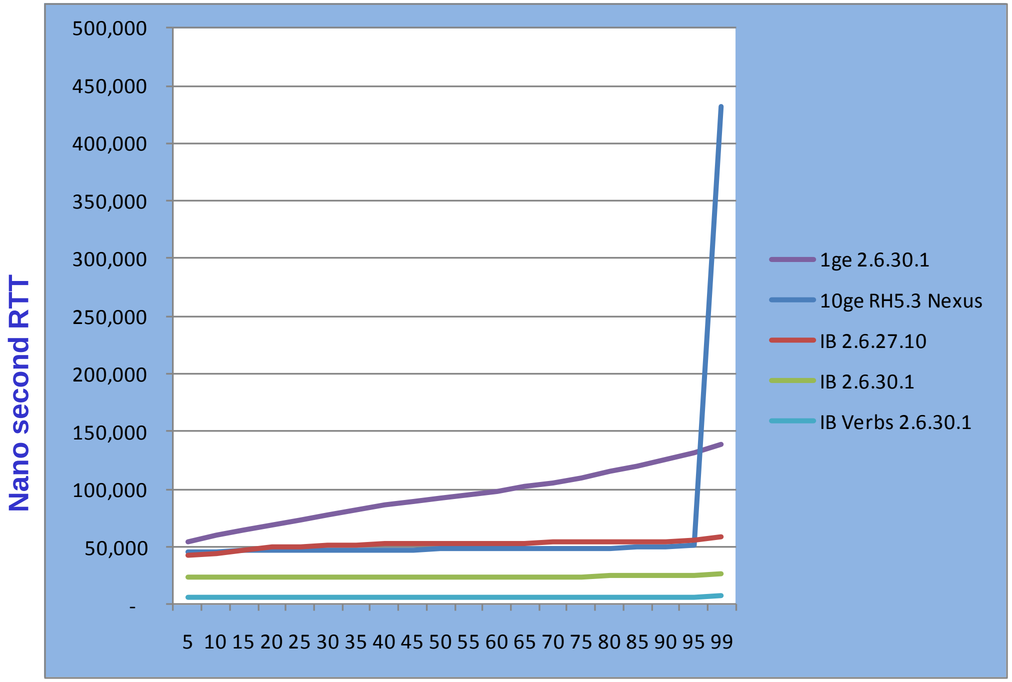


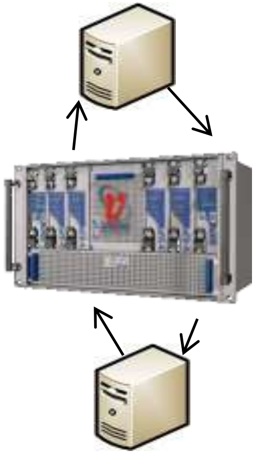
Diagram courtesy of STAC research

Multicast Latency RTT in microseconds

Informatix Solutions

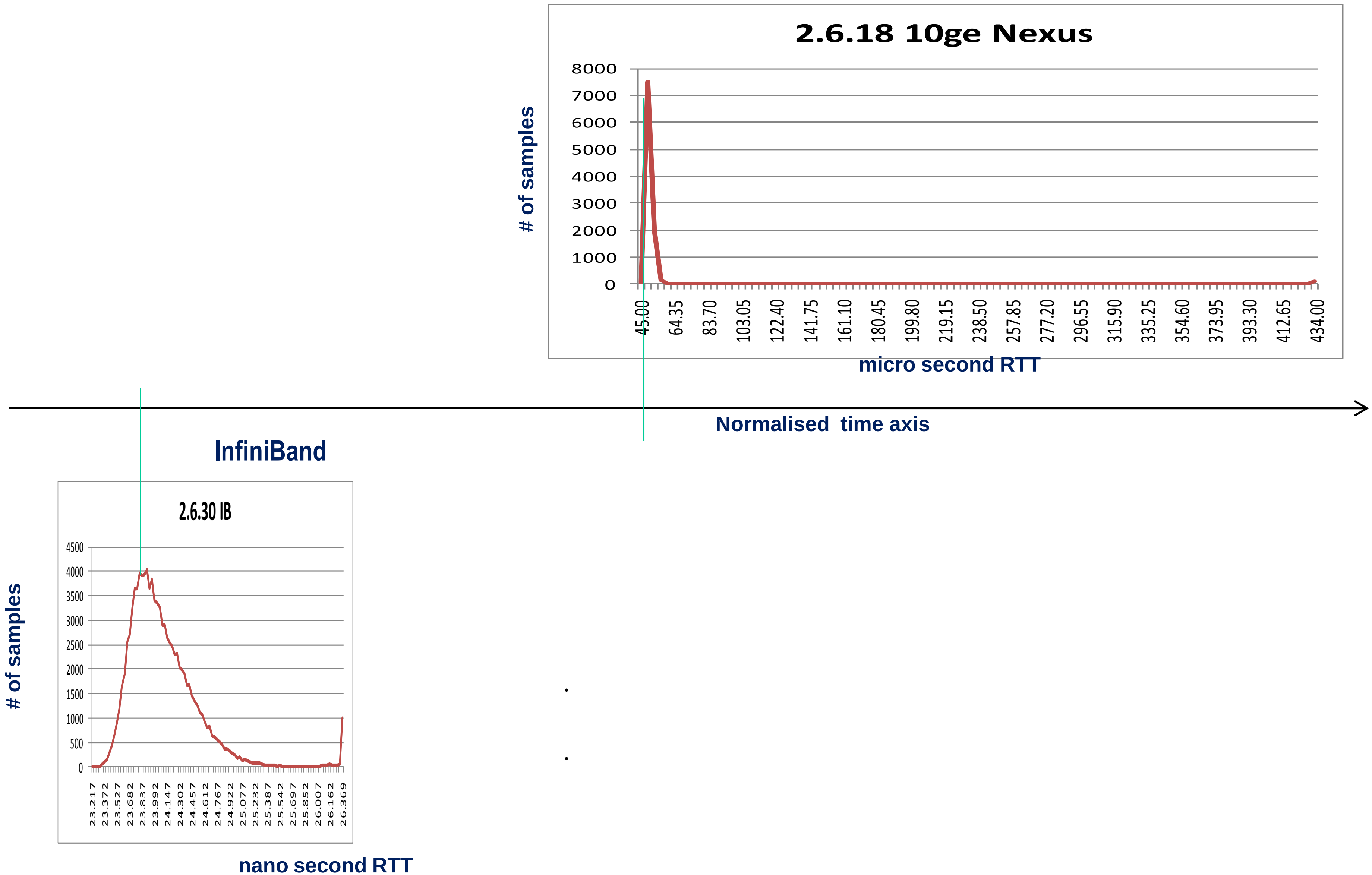
Interface	BW	Min (µS)	Avg (µS)	Median (µS)	Max (µS)	STD (µS)
Ethernet	1G	41.64	92.14	92.38	151.69	23.65
Ethernet	10G	45.00	65.48	48.00	3485.0	172.2
IPoIB (bonded)	80G	32.11	51.28	52.64	199.20	4.30
IPoIB	40G	23.22	24.86	23.99	1582	27.29
IB VERB	40G	5.79	6.21	6.14	46.90	0.51





10ge was Cisco Nexus 5010

Testing carried out at Intel fasterLAB



2.6.18 10ge Nexus
# of samples
0
1000
2000
3000
4000
5000
6000
7000
8000
45.00
64.35
83.70
103.05
122.40
141.75
161.10
180.45
199.80
219.15
238.50
257.85
277.20
296.55
315.90
335.25
354.60
373.95
393.30
412.65
434.00
micro second RTT
Normalised time axis
InfiniBand
2.6.30 IB
# of samples
0
500
1000
1500
2000
2500
3000
3500
4000
4500
23.217
23.372
23.527
23.682
23.837
23.992
24.147
24.302
24.457
24.612
24.767
24.922
25.077
25.232
25.387
25.542
25.697
25.852
26.007
26.162
26.369
nano second RTT

40/100G Ethernet v. 40G InfiniBand

Informatix Solutions

Ethernet

- IEEE 802.3ba standard definition it is not expected to be agreed before mid-2010
- Currently only carrier class switches are available. Need DC class switches before deployable to the server
  - Switch port cost is around $40K for 40G Ethernet
  - Ethernet 10G Cisco Nexus port currently around $930
  - 10G Ethernet dual port NIC ~$1000
- A 40G longline is currently 6x the cost of a 10G longline
- A 100G longline is currently 20x the cost of a 10G longline

InfiniBand

- InfiniBand 40G standard is agreed, and can be configured for 120G by using 12x
- 40G InfiniBand products for both switches and servers have been shipping in volume since 2009
  - InfiniBand 40G switch port already < $300 (36-port)
  - InfiniBand 40G HCA dual port ~$850

Swot Analysis from client Design Study

Five sites using existing long haul circuits.

Costs covered purchase of all network equipment and purchase of HCA's for InfiniBand option, using vendor list prices.

Ethernet option only provided 1Gb/s server attach.

By the time we got to deploy the InfiniBand products had been upgraded to 40G within these budgetary estimates.

Solution	Budgetary Cost	Strengths	Weakness
Cisco Catalyst 6500	€1,437K	Widest installed Proven technology Risk adverse	Poor Bandwidth usage Complexity of configuration Costly given provided functionality Same as everybody else – no latency advantage
Nortel ERS8600	€ 919K	Well proven High B/W usage through Active:Active L2 links Simpler L2 management than Cisco Better POP scalability through multipoint support (SW upgrade in 2009) Risk Neutral Lowest cost solution	Different to Cisco so small learning curve
InfiniBand	€1,330K	Lowest latency solution High B/W usage through Active:Active First to deploy pan-Market low latency solution in Europe Includes 20gb/s server attach providing additional application performance benefits	Learning curve of new technology First installation in Financial Services for Europe, for this distributed IB fabric Could be considered bleeding edge solution and therefore highest risk

Comparison with Ethernet - summary

Ethernet

- Best effort delivery. Any device may drop packets
- Relies on TCP/IP to correct any errors
- Subject to microbursts
- Store and forward. (cut-through usually limited to local cluster)
- Standardization around compatible RDMA NICs only now starting – need same NICs are both ends
- Trunking is an add-on, multiple standards an extensions
- Spanning Tree creates idle links
- Now adding congestion management for FCoE but standards still devloping
- Carries legacy from it's origins as a CSMA/CD media
  - Ethernet switches not as scalable as InfiniBand

Provisioned port cost for 10Ge approx. 40% higher than cost of 40G InfiniBand

InfiniBand

- Guaranteed delivery. Credit based flow control
- Hardware based re-transmission
- Dropped packets prevented by congestion management
- Cut through design with late packet invalidation
- RDMA baked into standard and proven by interoperability testing
- Trunking is built into the architecture
- All links are used
- Must use QoS when sharing with different applications
- Supports storage today
- Green field design which applied lessons learnt from previous generation interconnects.
- Legacy protocol support with IPoIB, SRP, vNICs and vHBAs.

Related info

- Hedge by deploying the Mellanox VPI range of HCA's. These dual port (CX4) cards can be configured to run InfiniBand or 10G Ethernet. They support OFED on both, and RDMAoE. HCA has drivers for Linux and Windows.

- See also:
  - Serialization costs
  - Multicast
  - Ethernet to InfiniBand gateways