InfiniBand Storage System Area Networks

David Southwell, Ph.D
President & CEO – Obsidian Strategics Inc.
BB:(+1) 780.964.3283
dsouthwell@obsidianstrategics.com

Agenda

- System Area Networks and Storage
- Pertinent technology trends
- Lossy and lossless transport
- How long are your bits?
- System Area InfiniBand networks
- InfiniBand over optical networks

What is a System Area Network?

Also called a “unified fabric” - one network carries all traffic types.

(local area, management, cluster messaging, wide area and storage)

Many benefits:

- Simplified and centralised network and system management
- Lower cost (fewer adapters, cables, switches, people hours)
- Fewer “interesting” failure modes
- Green – less equipment, lower power, better airflow (cabling)
- Higher floorspace density
- Supports smaller nodes (i.e. blades)
- Better network utilisation

Effective System Area Network technologies simultaneously support a superset of all traffic type requirements, are scalable, partionable, and support QoS mechanisms to reflect priority policies across traffic types and nodes.

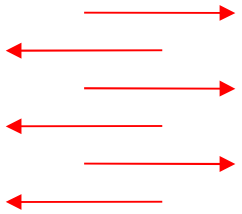
Architect for the future – technology trends

- Multi- and Many- core processors
- Virtualisation (compute, interconnect and storage)
- Solid state storage
- Blade form-factors
- Optical interconnects
- High-fidelity thin clients

Trying to move data quickly - with TCP/IP

A lossy protocol – data is deliberately dropped if it cannot be handled by a switch, router or end-point.

Recovering from data drop involves :



- Detection of drop
- Selective retransmission requests
- Drastic reductions in traffic injection rates

The higher the bit-rate, and the longer the link, the more this process kills link performance.

Developed 30+ years ago in the Mbits/s era, TCP/IP fails to deliver efficient transfers at the 10Gbits/s and higher speeds.

- It takes ~ 1GHz of CPU to process 1Gbits/s of TCP/IP
- 3000 km TCP/IP link at 10Gbits/s typically sees **< 30% link efficiency**
- This will only get worse @ 40 and 100 Gbits/s
- Offload engines just move the problem, they don't remove it

InfiniBand primer

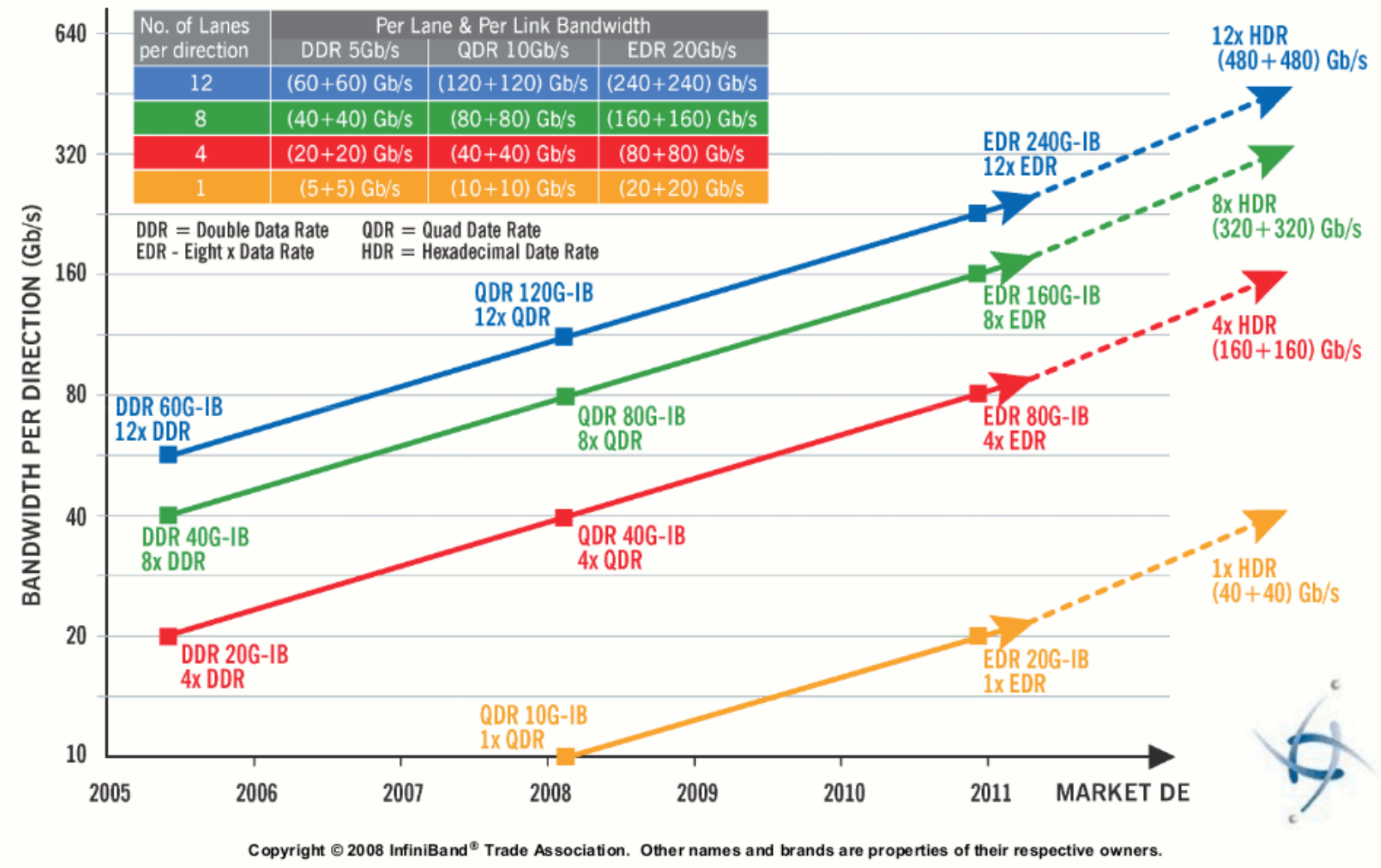
- Created by IBM, HP, Dell, Sun, Intel, Microsoft and Compaq in 2000
- The only high-performance non-proprietary cluster interconnect
- InfiniBand Trade Association (IBTA) oversees the specifications
- InfiniBand natively supported by Linux kernels since 2.6.11
- Open-source stacks provided by the OpenFabrics Alliance (OFA)

http://www.openfabrics.org

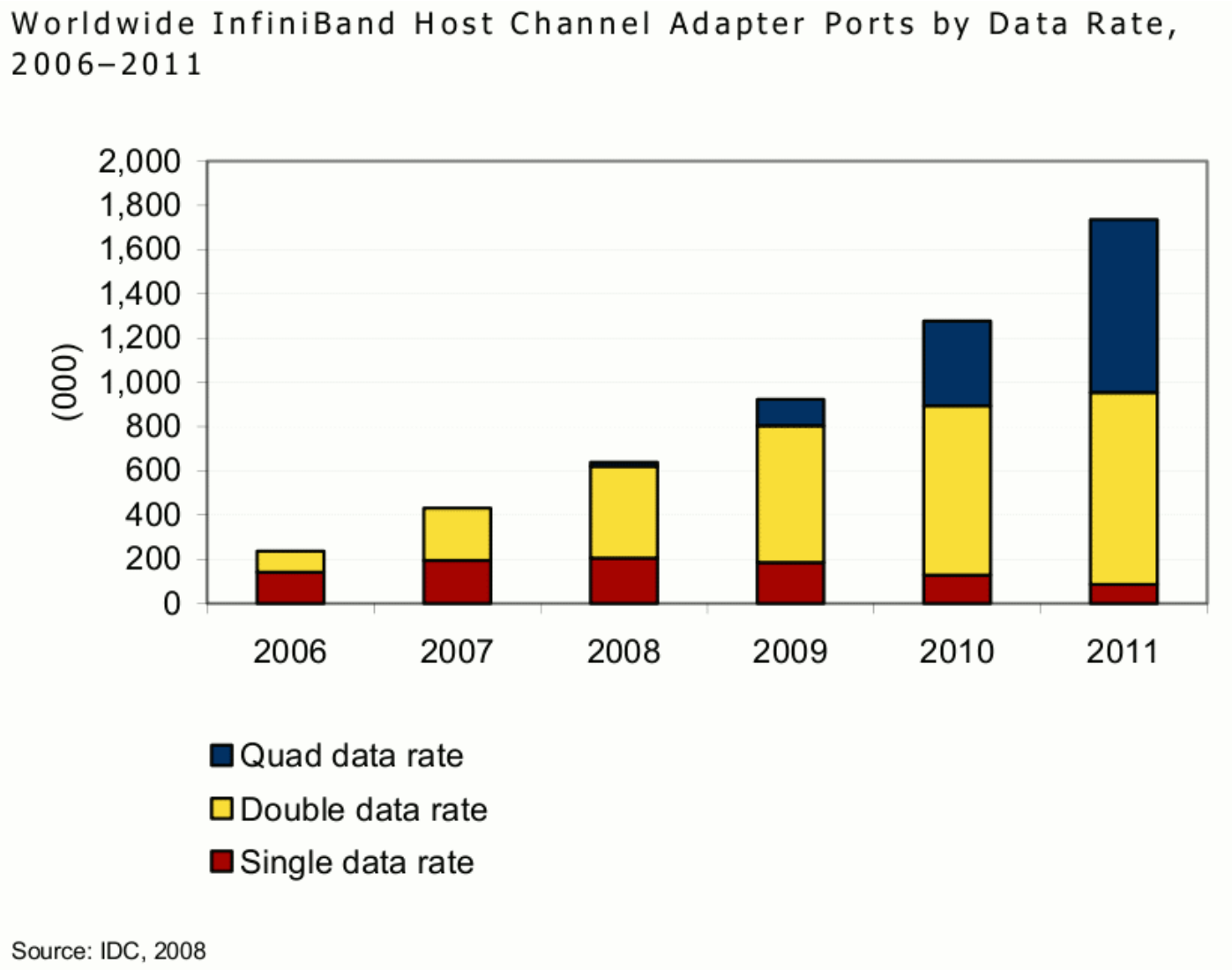
http://www.infinibandta.org

- A switched-fabric interconnect architecture
- Scales in performance (96Gbits/s today, road-mapped to 384GBits/s)
- Low-latency switches, adapters (sub-µs user-space to user-space)
- Supports Remote DMA (zero copy memory-to-memory over fabric)
- Scales to thousands of nodes, supporting useful topologies (fat tree)
- 3 of the 5 fastest machines in the top500 list use InfiniBand fabrics
- Cost effective even for small server clusters

InfiniBand Bandwidth Roadmap



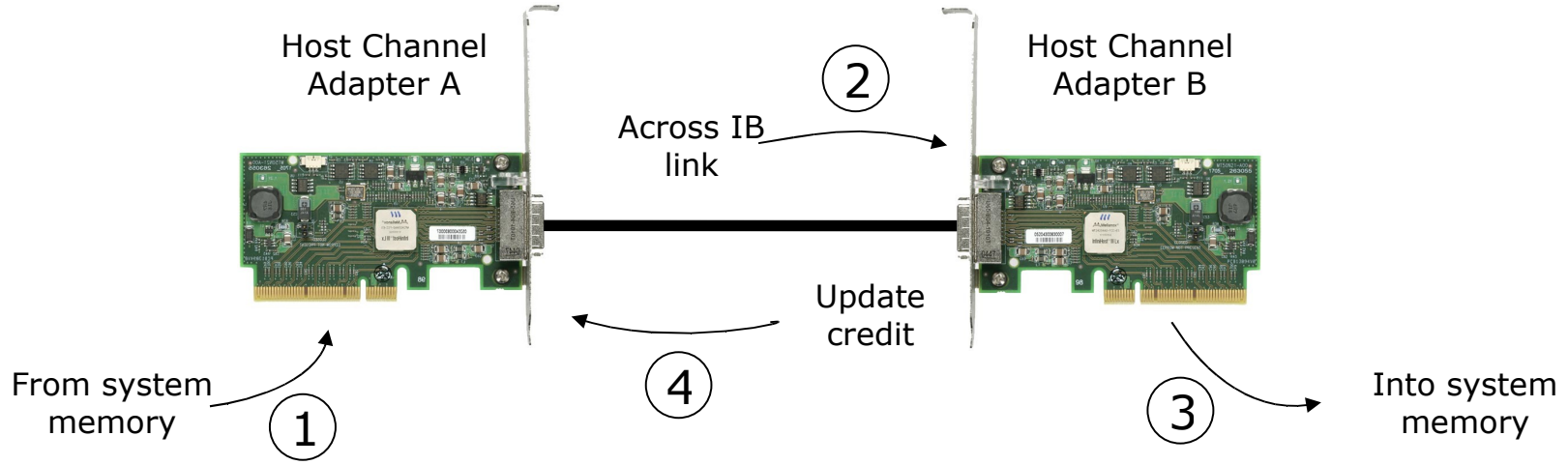
InfiniBand Penetration Projections



Moving data quickly - with InfiniBand

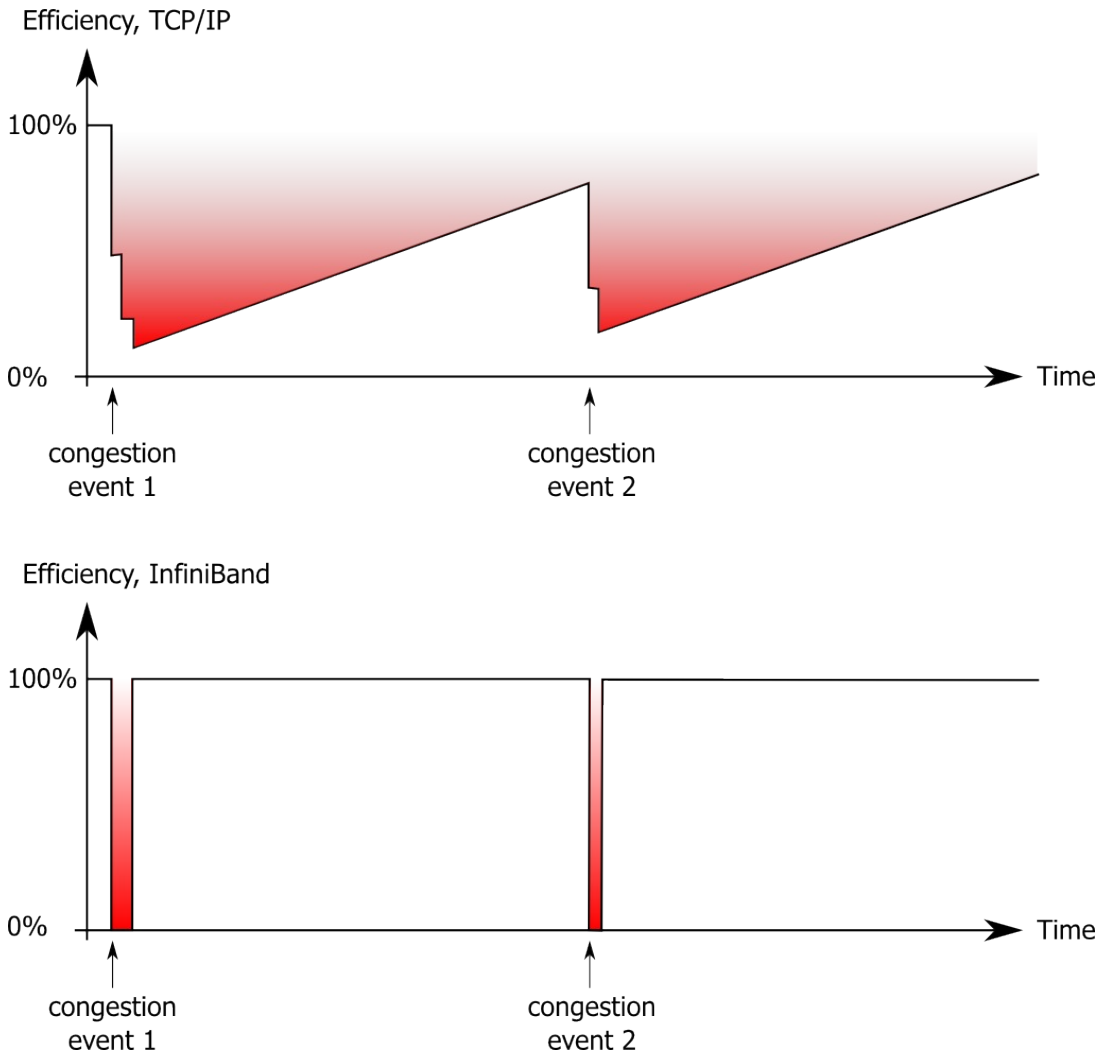
InfiniBand employs a lossless buffer-credit flow control protocol:

- Each link end-point advertises receive buffer capacity to the other
- Data is sent without warning if it fits into known receive buffer capacity
- As buffer space is freed up, buffer-credit packets update available capacity



It is intrinsically more efficient to prevent congestion by delaying a transmission, than to blindly transmit anyway and incur the penalties associated with clean up after data is dropped.

Efficiency – graphical comparison

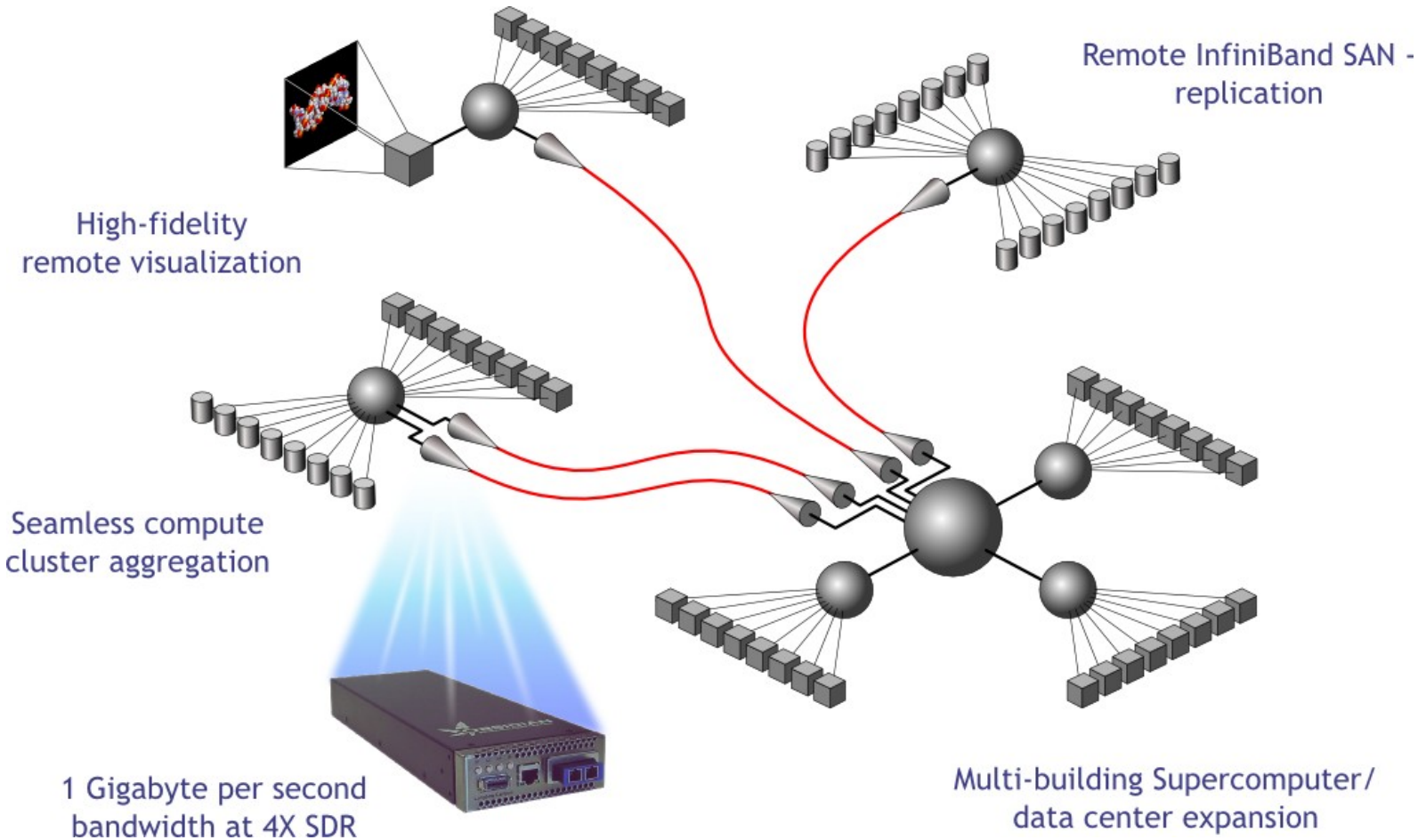


Bits aren't what they used to be...

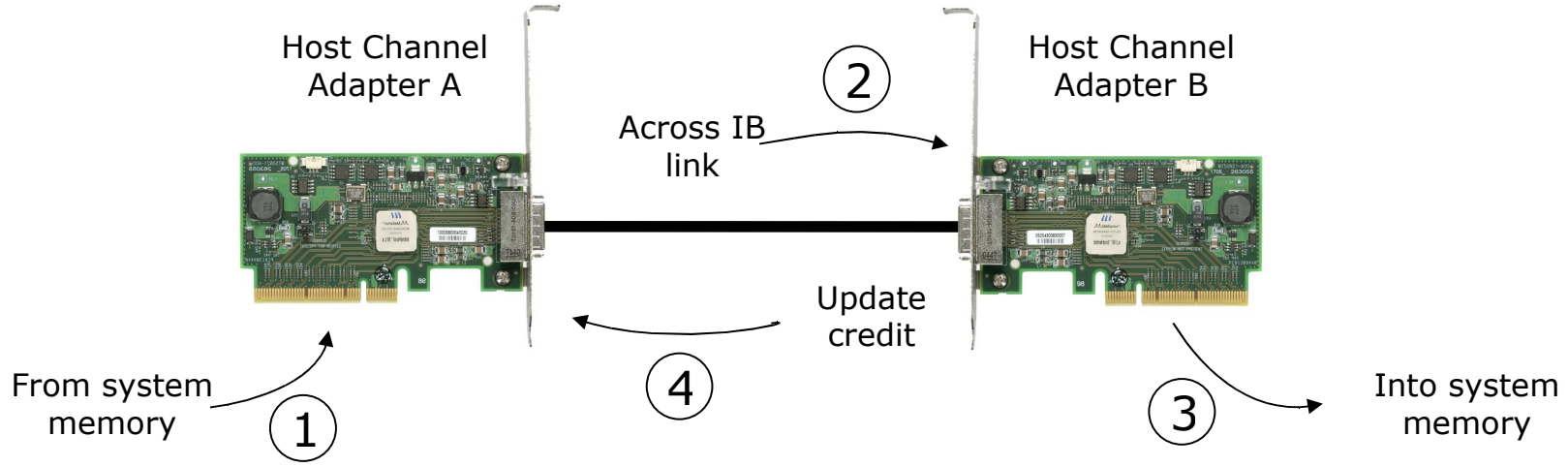
Year	Technology	Bit Length	Data on 10km link	
1973	3 MBit Eth	99 m	12	Bytes
1978	10 Mbit Eth	30 m	40	Bytes
1993	100MBit Eth	3 m	400	Bytes
1996	1 Gbit Eth	20 cm	6	KBytes
2001	4X SDR IB	2 cm	60	Kbytes
2005	4X DDR IB	1 cm	120	KBytes
2008	4X QDR IB	5 mm	240	Kbytes
2008	12X QDR IB	2 mm	600	Kbytes
201?	12X HDR IB	420 µm	2.8	MBytes

- Efficient storage transports must be lossless!
- Flow control will become progressively more critical as we move faster

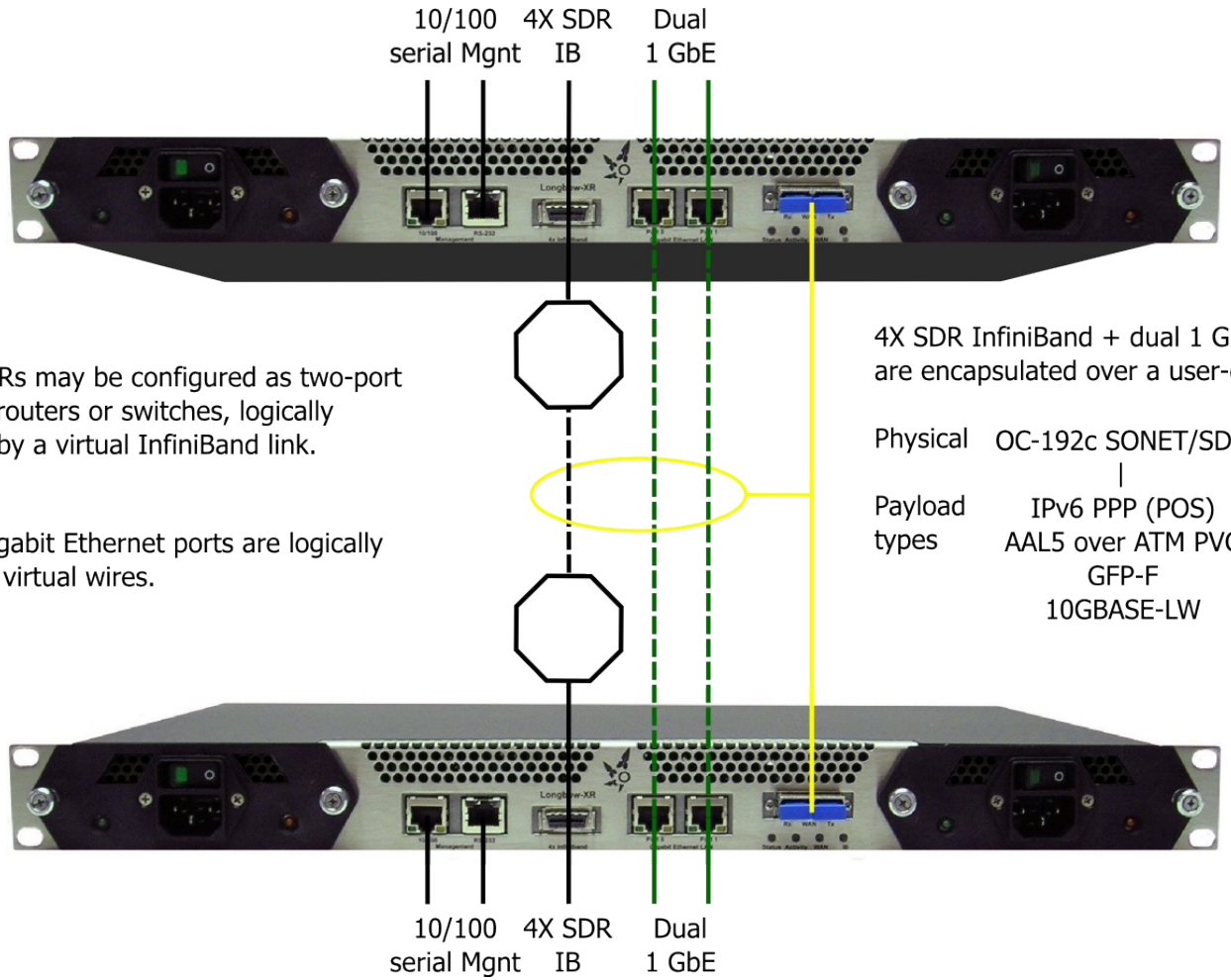
InfiniBand System Area Networks



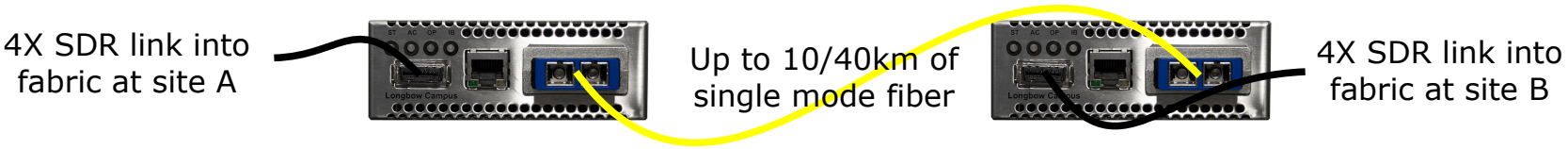
Flashback – Buffer credit starvation



Longbow X Series – Wide Area InfiniBand



Longbow C Series – lightpath InfiniBand



Longbow Cs work over direct lightpaths – dark fiber or WDM systems.

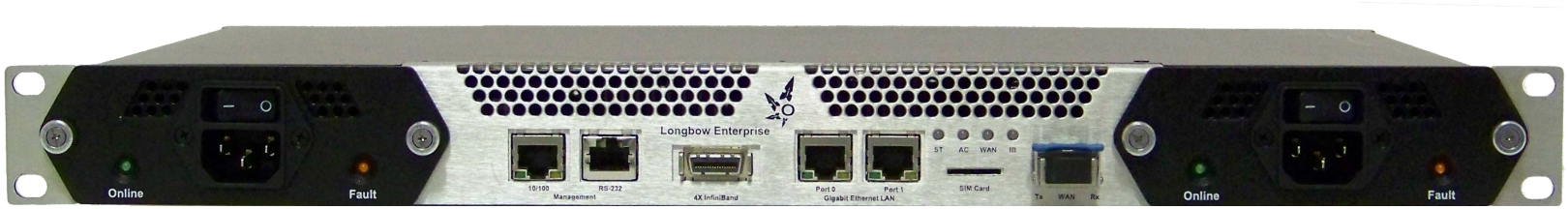
Port-to-port latency is ~800 nanoseconds.

Not internally redundant; typically deployed in trunked configurations – higher aggregate performance, leveraging InfiniBand protocol failover.

Rack-mounts up to four devices per shelf – 1U for 4GBytes/s of 10km/40km InfiniBand...

Longbow E Series

- Supports Enterprise Data Centre applications over global terrestrial networks with integrated line-rate AES-192 crypto
- Scheduled for evaluations starting in Q4’08
- Demonstrated in NASA's booth at SC|07 in advanced prototype form



Full Line Rate native InfiniBand:

- Range extension over 10GbEthernet or (DF / WDM) lightpaths
- Inter-subnet Routing
- Standards-based Authentication and Encryption
- Firewall filtering, packet inspection (future)

Longbow deployment scenarios

Used in pairs, Longbow devices natively connect remote InfiniBand equipment - within and between data centres, clusters and supercomputers - across standard optical networks.

Longbow X Series – Global range

Longbow C Series – 10/40km

Transparent, high-bandwidth, low-latency, secure, robust, standards-based

- Data Centre replication
- Data Centre/ supercomputer expansion
- Data Centre interlinks – Grid / Cloud Storage Computing
- Latency-sensitive messaging (automated trading)
- Global data streaming (military, surveillance, science)
- Remote visualization
- Next-Generation Storage Area Networks
- Cluster Clustering

Optical transports for long haul InfiniBand

InfiniBand bandwidth scales like no other protocol – high capacity optical channels can be filled with a single logical flow – no need to look for aggregated workloads.

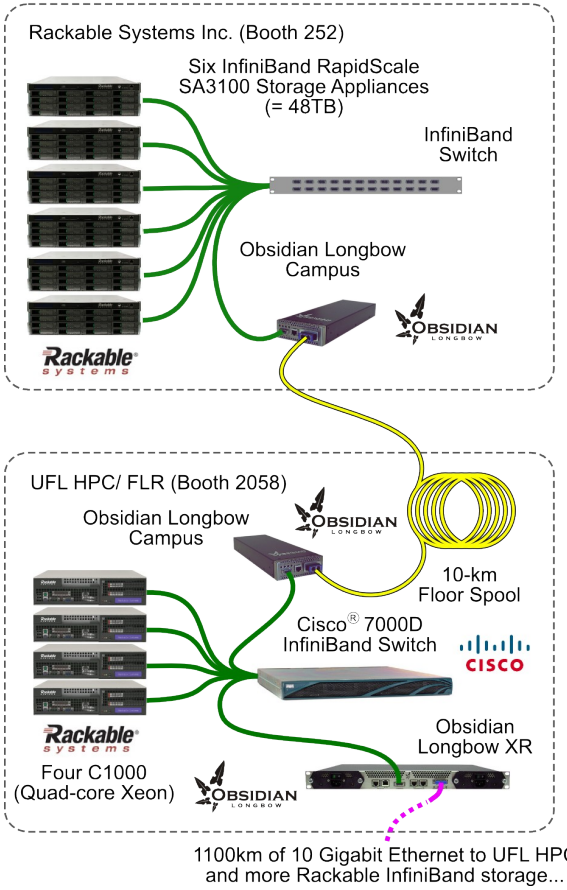
The Longbow roadmap is able to use faster optical modulation as well as WDM techniques – maximizing leverage from optical infrastructure at all stages of roll-out.

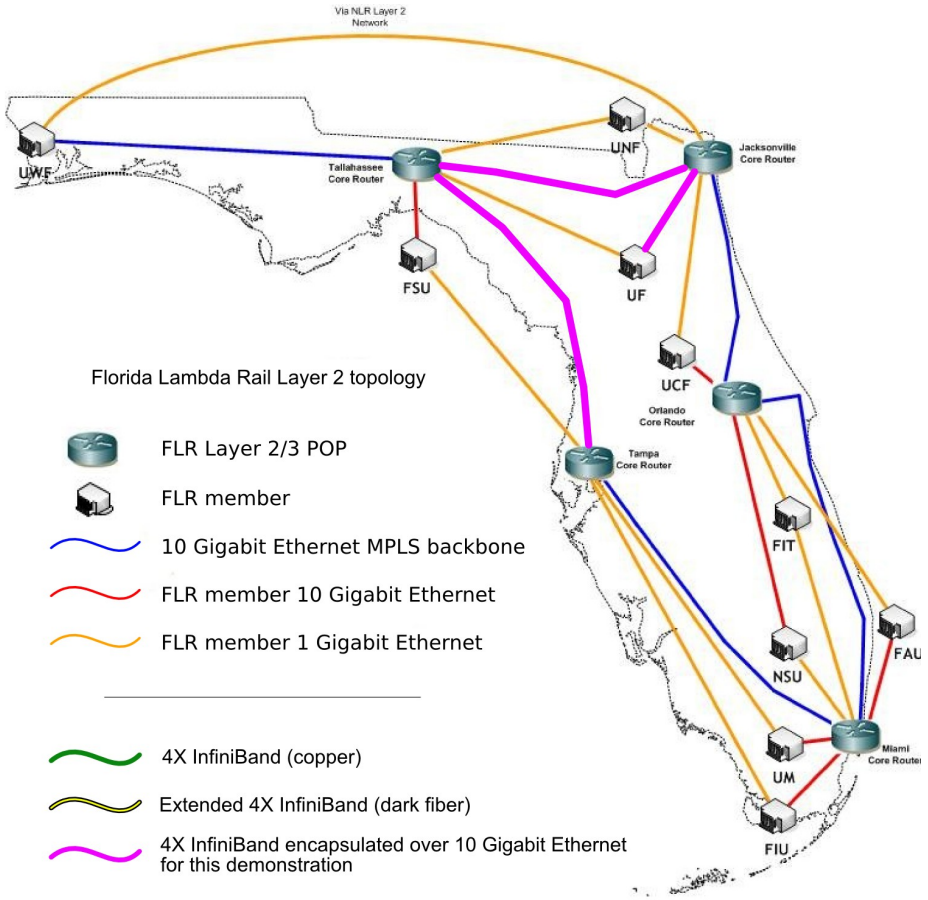
InfiniBand Configuration	# InfiniBand channels	InfiniBand channel payload bitrate	Optical Configuration	# Optical wavelengths	Optical wavelength bitrate	**Bandwidth (full duplex)**	Status
4X SDR	4	2 GBits/s	10G	1	10 GBits/s	**1 Gbytes/s**	Production
4X QDR	4	8 GBits/s	3λ WDM	3	11.1 GBits/s	**4 Gbytes/s**	Development
4X QDR	4	8 GBits/s	40G	1	40 GBits/s	**4 Gbytes/s**	Future
12X QDR	12	8 GBits/s	10λ WDM	10	10 GBits/s	**12 Gbytes/s**	Development
12X QDR	12	8 GBits/s	3λ WDM	3	40 GBits/s	**12 Gbytes/s**	Future
12X QDR	12	8 GBits/s	100G	1	100 GBits/s	**12 Gbytes/s**	Future

NASA deployment

Wide-Area IB SAN @ SC|06

Dr. Charles Taylor and Dr. Craig Prescott have been using Longbows for Campus and Wide Area InfiniBand native storage, with Rackable Systems (Terrascale).

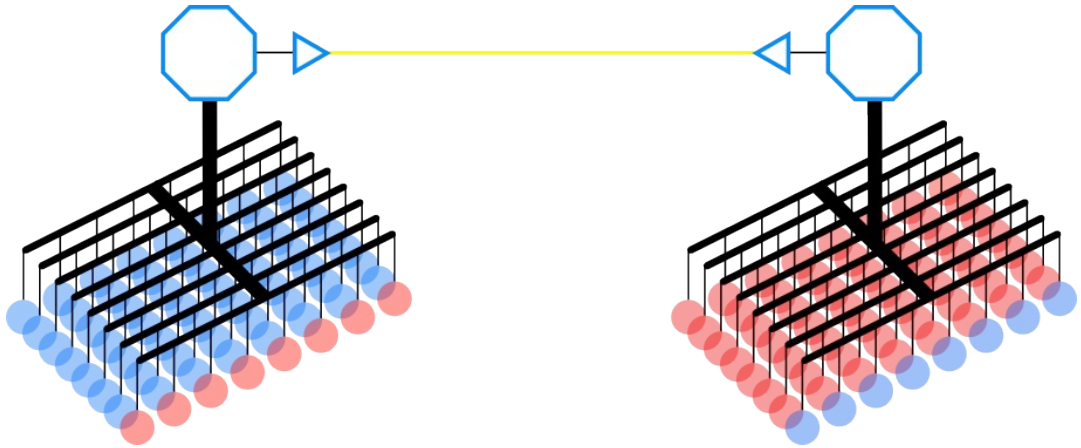




Cluster Clustering – InfiniBand Grid

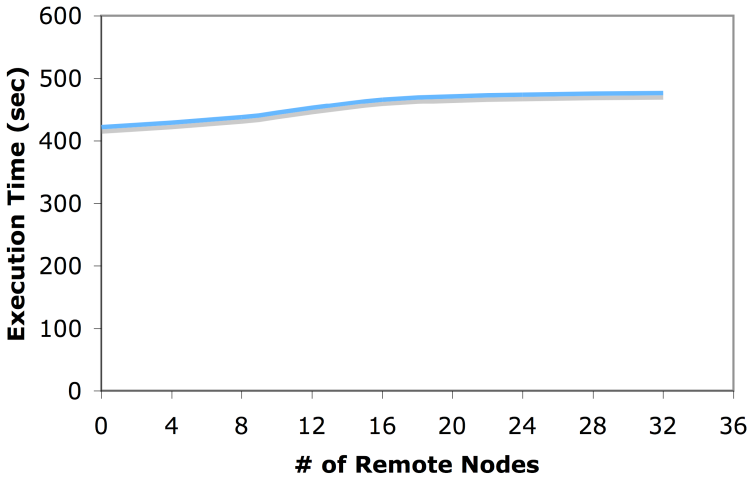
Six satellite campus sites, with InfiniBand-based clusters in each – how effective would it be to unify them into a single **campus area grid**?

Dr. Dan Stanzione (Arizona State) measured performance for Real World application codes distributed across two sites 2.5 km apart but connected by a single pair of Longbow Campus devices using dedicated dark fibre...

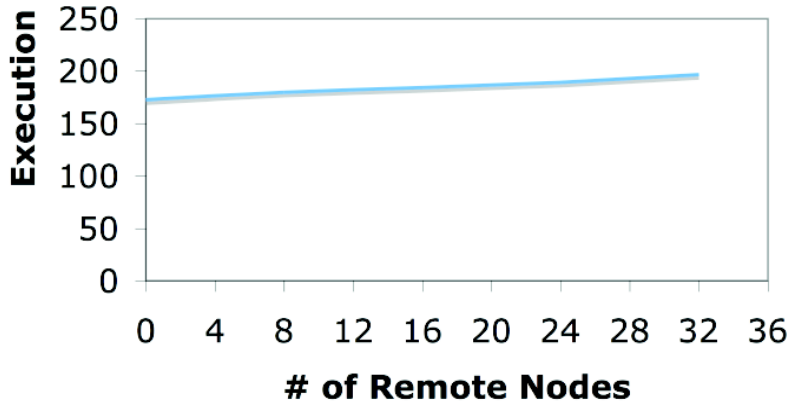


64 processors were used in each test; graphs were created by varying the number of processors separated from the others by the Longbow connection (0 ... 32) and measuring the resulting run times.

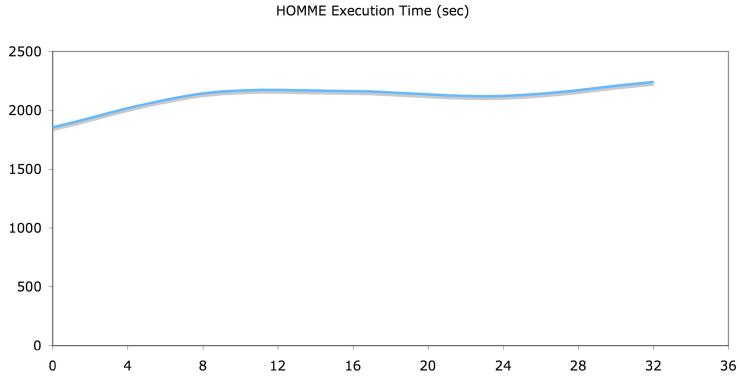
ASU results – MILC, WRF & HOMME



MIMD Lattice Computation (MILC)

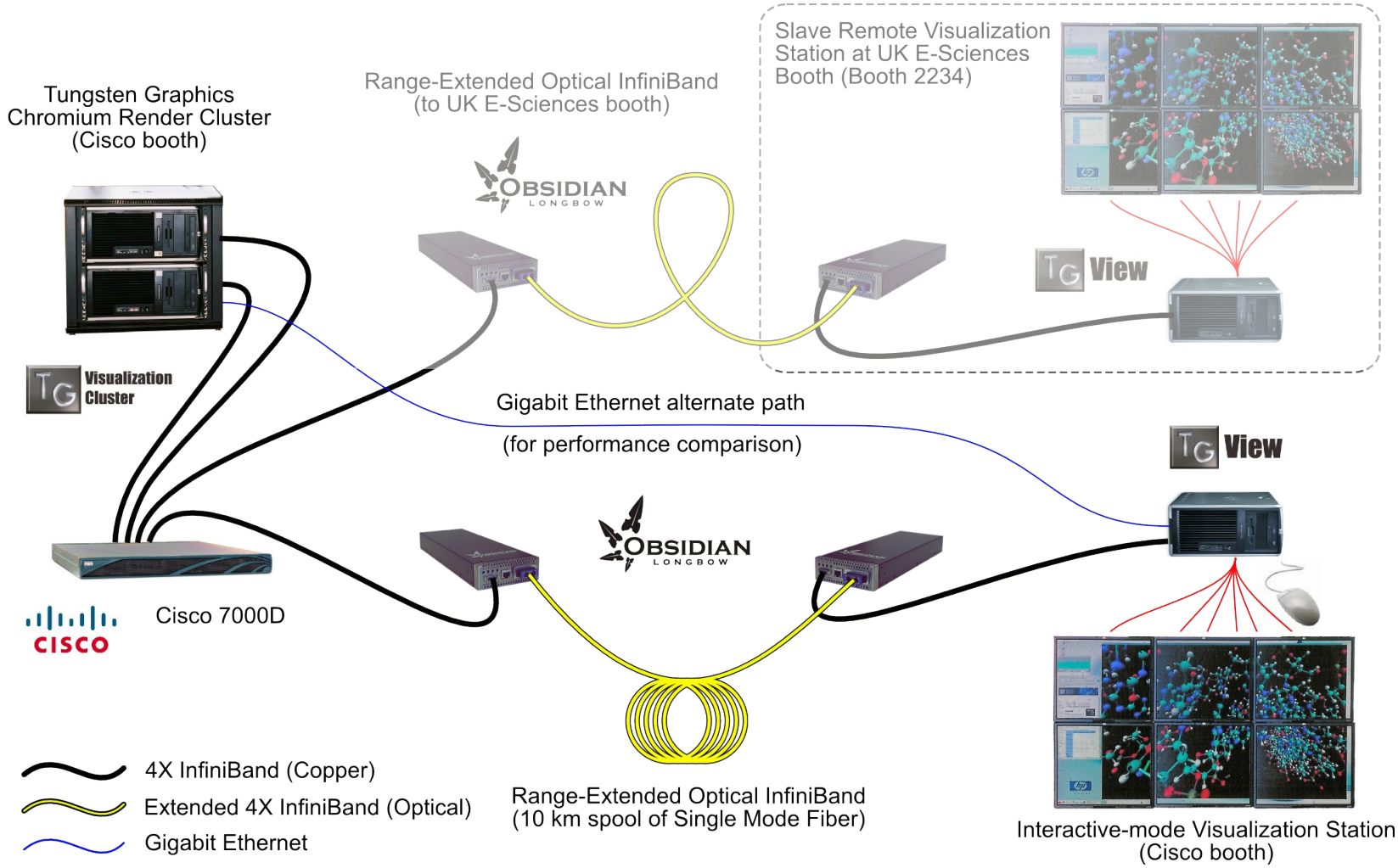


Weather Research and Forecasting (WRF) Code



High Order Methods Modelling Environment (HOMME)

High performance remote visualization



High performance remote visualization

Conclusions

- We are entering a network-constrained computing era
- Storage stresses bandwidth, latency, scalability and redundancy

Architect for the future – there will be a fork-lift upgrade which ever way you proceed beyond "standard" 10GbE…

InfiniBand is a strong candidate for future System Area Networks within high-performance computing applications and more generally in distributed data centre environments.