

Delivering HPC Performance at Scale

October 2011

Joseph Yaworski
QLogic
Director – HPC Product Marketing
Office: 610-233-4854
Joseph.Yaworski@QLogic.com

- **QLogic Overview**
- **TrueScale[®] Performance Design**
- **History Behind InfiniBand**
- **Examples of Performance at Scale**



QLogic: A Global Company

- **Headquarters**
 - Aliso Viejo, California
- **Products**
 - Networking for HPC & Storage
 - # 1 or # 2 in the target markets we serve
- **Employees**
 - Over 1000
- **Financial Position**
 - 7 straight years of market share growth
 - FY11 Revenue = \$597.2M
 - No debt, strong cash position
- **Member of the S&P 500 traded on NASDAQ**
 - Symbol = QLGC



Focused on End-to-End High Performance Computing Solutions



ASIC Design

- Scalable high bandwidth
- Low latency under load
- Power Optimization



Switch & HCA Development

- Modular & scalable to 864 ports
- Signal integrity
- Advanced feature set
- Fabric optimization routing routines



System Architecture

- Designed for HPC
- MPI performance tuned interface - PSM
- Message rates 30 M/s



Fabric Management

- Advanced installation and verification tools
- Real time fabric display/viewer
- Fabric virtualization
- Fabric QoS
- Integration with industry leading job schedulers



Application Integration

- Integrated with multiple file systems
- Performance optimized with over 70 applications
- NetTrack Development Center

InfiniBand History Lesson

Bit of History

Early 2000 Timeframe

Original InfiniBand Focus

Applications

I/O Focused ULPs

Verbs Provider / Driver

Traditional Offload HCA

InfiniBand Wire Transports

- **Before InfiniBand**
 - Competing Standards – **NextGenIO & FutureIO**
- **Early InfiniBand Focus**
 - Designed for the enterprise data center market and an **IO paradigm**
 - **Backbone network** as a replacement for Ethernet and Fibre Channel
 - Incorporate best data center features of all interconnects and protocols
 - **Performance Req.: Millions of IOP's**
- **Servers**
 - Single Core / Dual Socket
 - Limited processor speed
 - Slower PCI, PCI-X buses

Bit of History

Mid 2000 Timeframe

InfiniBand HPC Focus

Applications

MPI Libraries
(Verb-based)

I/O Focused
ULPs

Verbs Provider / Driver

Traditional
Offload HCA

InfiniBand Wire Transports

InfiniBand Finds It's Niche

- High Performance Computing Clusters market
- Low-Latency / High Bandwidth advantages
- **Primarily messaging paradigm – MPI**
- Cluster sizes: 1000s of nodes
- **Performance Req: 10M msg/sec**

Verbs - Retrofitted for HPC

- Based on RDMA and QP programming model
- Connection oriented approach with heavy-weight state
- Poor match to MPI semantics
- RDMA model requires significant memory pinning for send and receive

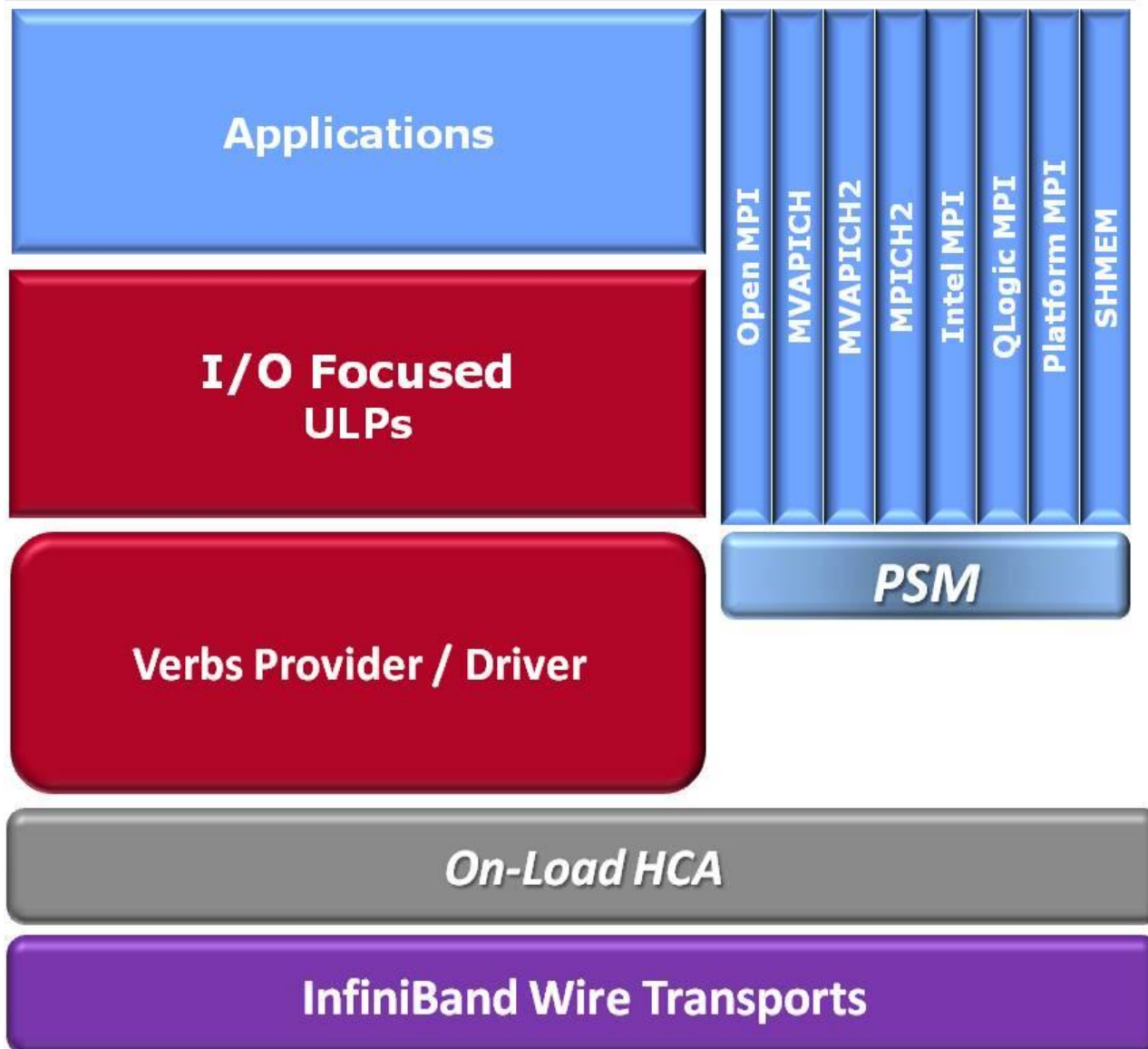
Servers

- Multiple Cores per CPU
- Multi-socket servers are the norm
- Processors faster with more internal bandwidth
- PCI-express

Bit of History

Mid 2000 Timeframe

Optimized InfiniBand HPC Implementation

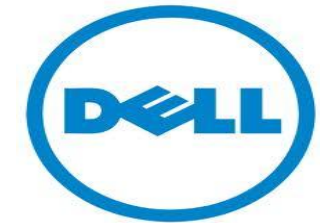


Performance Scaled Messaging

- **PSM is specifically designed for MPI**
 - Light weight - 1/10th the user space code of Verbs
- **Connectionless with minimal on-adapter state**
 - No Chance of Cache Misses as the Fabric Scales
- **High MPI message rate –**
 - Short message efficiency
- **Amenable to receiving out-of-order packets**

Designed to Scale with Today's Servers

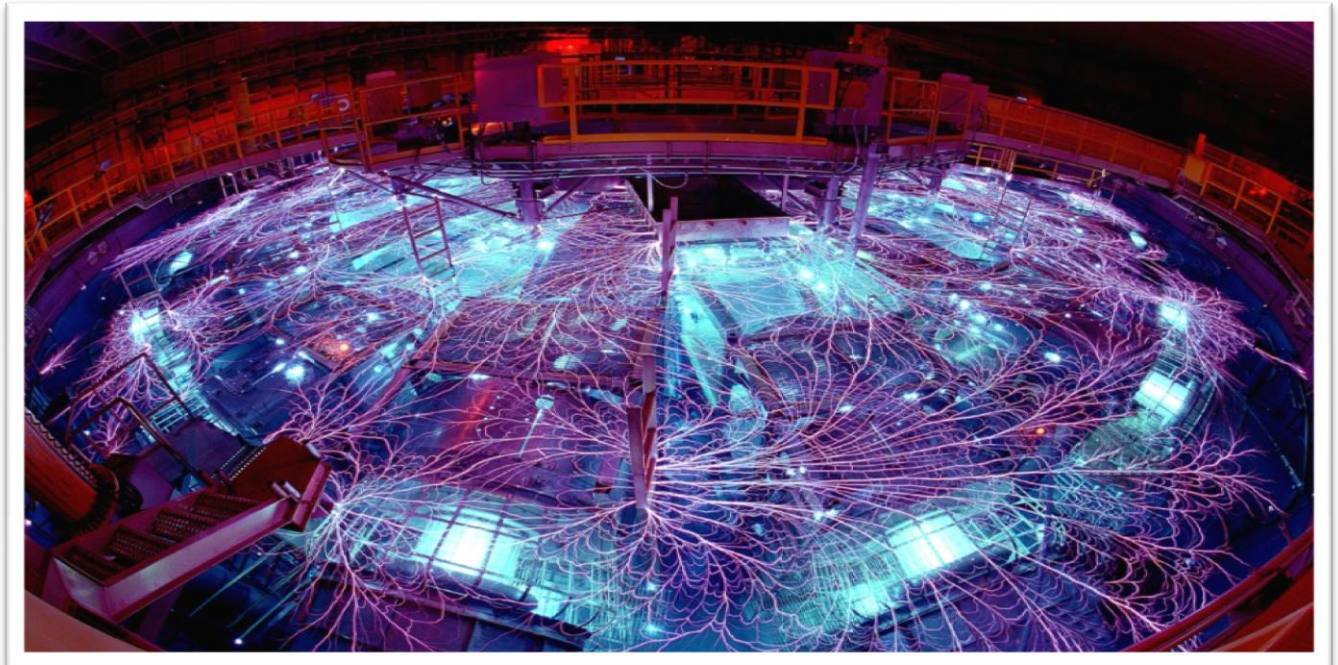
- **Dense Multiple Core CPU's**
- Multi-socket servers are the norm
- Processors faster with more internal bandwidth
- PCI-express



Exploiting high-performance computing to solve global energy, climate change and security challenges

Enabling breakthrough scientific discoveries using leading edge-technologies and partnerships

Chose Dell and QLogic TrueScale



Scalable Linux Clusters: Enabling Scientific Discoveries

November 17, 2010

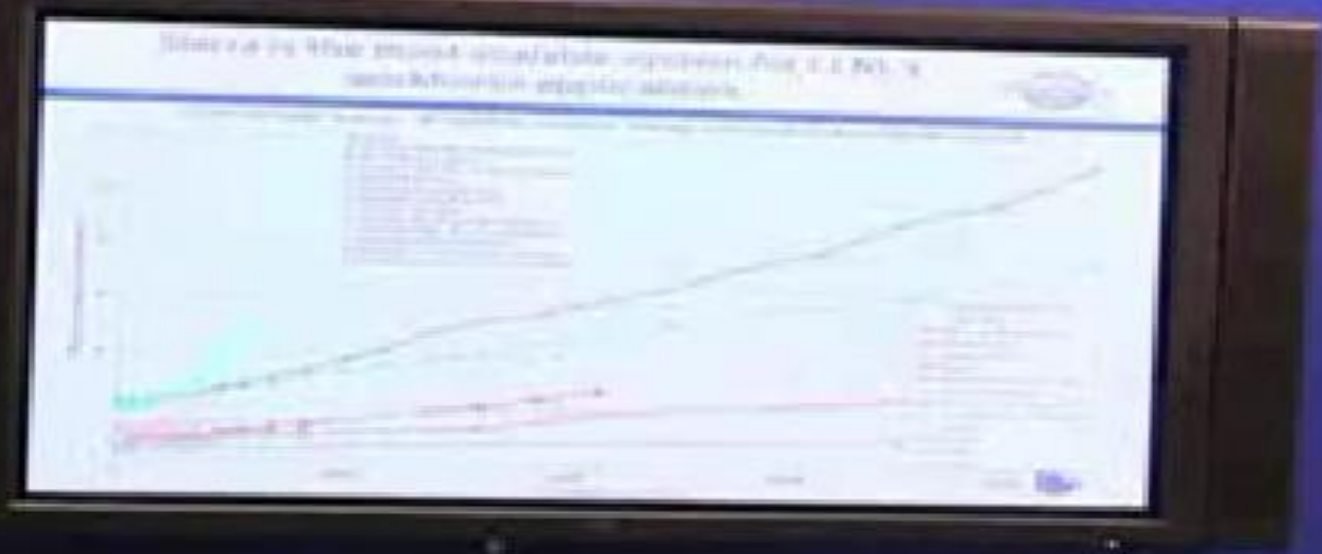


Matt Leininger

Deputy for Advanced Technology Projects

**S&T Principal Directorate - Computation Directorate
Lawrence Livermore National Laboratory**

**This work performed under the auspices of the U.S. Department of Energy by Lawrence
Livermore National Laboratory under Contract DE-AC52-07NA27344
LLNL-PRES-XXXXXX**



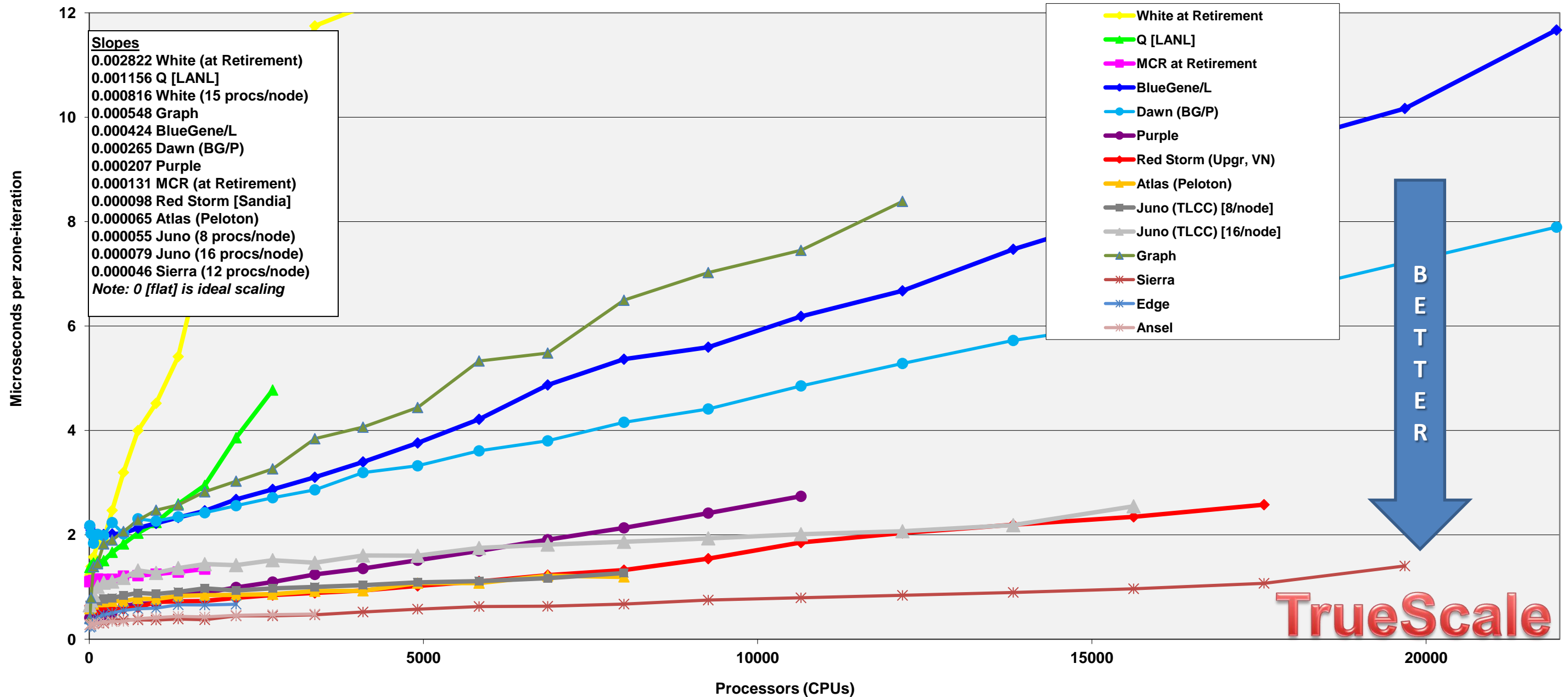
QLogic SC10
Storage Presentation Schedule

The poster features a red and blue color scheme with a central graphic. Below the title, there is a list of items, likely a schedule or product specifications, though the text is too small to read clearly.



Sierra is the most scalable system LLNL has ever deployed

Historical Weak Scaling - 3D Radiation problem's average zone-iteration grind time per machine



Early performance and scalability features still under evaluation



- Scalability of the IB fabric is best of class
- Typical latency are 1-2 us
- Message injection rate is of fabric is one element of scalability (~27-30M msg/sec for 4byte)
- MPI collectives benefit from all the above
- Advanced routing and congestion control features are under evaluation
- QLogic PSM layer released open source and in OpenFabrics

<http://www.qlogic.com/Products/Pages/HPCLearnMore.aspx>

Tri-Labs Linux Compute Cluster 2

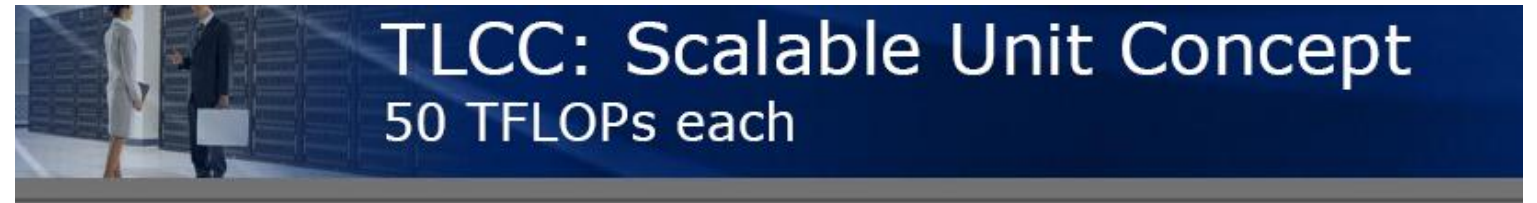
TLCC2 – Next Generation Deployment to TLCC

QLogic InfiniBand chosen for the DoE TLCC2 deployment

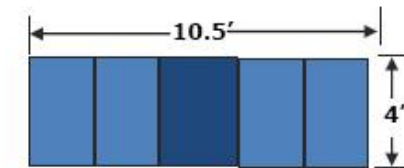
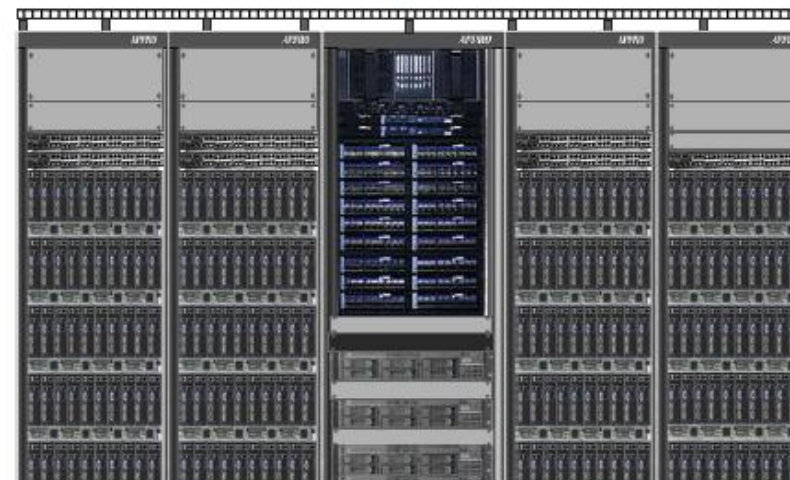
- Intel Xeon 'Sandy Bridge' processors, QLogic QDR InfiniBand
- 6-Pflops / 20K nodes when fully deployed
- Bids heavily influenced by LLNL findings

DOE Labs

- LLNL – Lawrence Livermore National Labs
- LANL – Los Alamos National Labs
- SNL – Sandia National Labs



Two SU Configuration:



- Five (5) 48U Racks, each with one PDU
- 308 x Compute Nodes (1 x QDR Appro Blades)
- Twelve (12) Gateway nodes (2 x QDR Blades)
- Seven (7) 48-port Ethernet Switches
- One (1) 324 port IB Switch (fully populated)
- Two (2) RPS (boot/management) Node
- Two (2) LSM (login) Node

TrueScale Benchmark Win
Against
Next Generation InfiniBand Offerings

Shared Success with Acer



National Applied Research Laboratories

National Center for High-Performance Computing

NCHC provides the highest levels of computing performance and lowest power consumption to support Taiwan's research and academic communities efficiently.

The Results

- **Computing capability: +170 Tflops** (>512 compute nodes, over 25,000 cores)
- **I/O Capacity: 3 MB/second/core** (DDN SFA array with Lustre)
- **Interconnect fabric: Dedicated QLogic MPI and I/O fabrics**
- **Power consumption: < 1000 kW**



The Ultimate in Performance



Shared Success with Dell

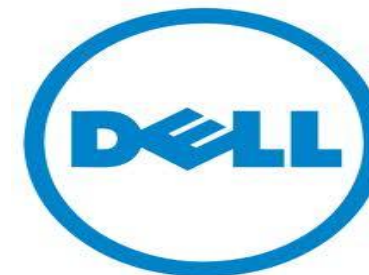
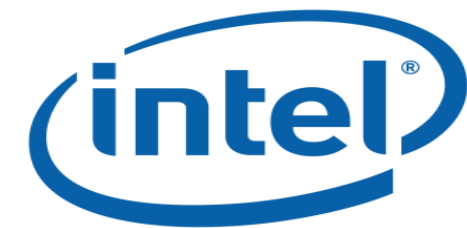
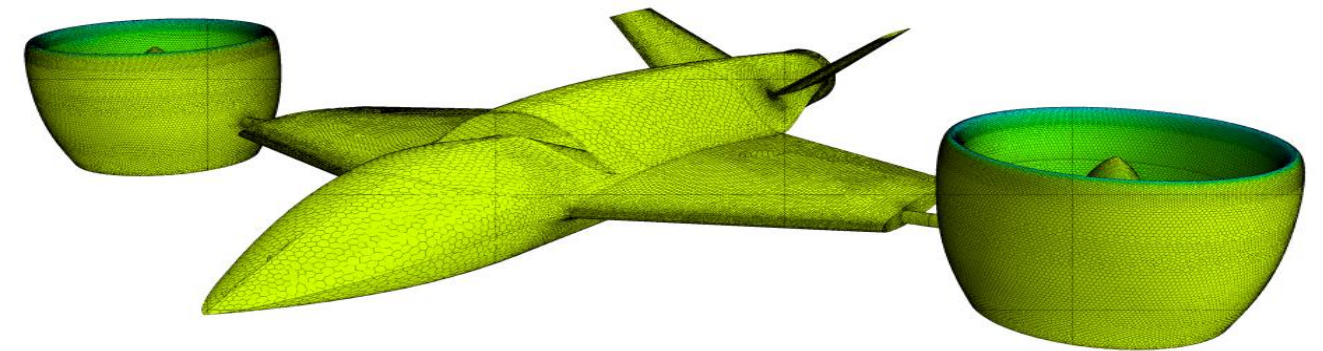


“Test flew” the AD-150 at the QLogic NetTrack Developer Center using

- Dell PowerEdge® HPC Cluster
- CD-adapco STAR-CCM+
- QLogic TrueScale InfiniBand

The results? **98% FASTER** time to solution

- Able to run more and larger models
- Better design validation
- Reduced costs through better simulation and less physical prototyping



Did You Know...

QLogic InfiniBand accelerates today's breakthrough discoveries to harness tomorrow's energy



Vestas



QLogic TrueScale InfiniBand Accelerates HPC Innovations for these Premier Automotive Brands...



Take High Performance Computing Sky-High with QLogic TrueScale InfiniBand



BAE SYSTEMS

QLogic TrueScale InfiniBand.

Designed for HPC. And Used By These
Premier EMEA Weather Centres...



Shared Success with HP



ARAMCO chose HP with QLogic TrueScale InfiniBand

Recently installed 512 node cluster purpose-built for their HPC workloads

10 times faster than their previous system

- 6+ TFLOPS
- Would rank in the top 100 of the Top500 report

End-to-end QLogic TrueScale InfiniBand solution ensures

- Unsurpassed messaging rates
- Highest effective application bandwidth
- Absolute lowest latencies



Key Recent Customer Wins



DOE Tri-Labs - TLCC2





QLOGIC[®]
The Ultimate in Performance

The logo features a stylized white symbol above the brand name 'QLOGIC' in a bold, sans-serif font. Below the brand name is the tagline 'The Ultimate in Performance' in a smaller, white, sans-serif font. The background is a vibrant blue with abstract, glowing lines and patterns that suggest a high-tech or data-driven environment.