



OPENFABRICS
ALLIANCE

SAND2006-7153C



Enabling Large-Scale Scientific Computing with InfiniBand: Experiences with the World's Largest InfiniBand Cluster

Matt Leininger, Ph.D.

**Sandia National Laboratories
Scalable Computing R&D
Livermore, CA**

15 November 2006

DOE/ASC has Evaluated Several Generations of InfiniBand



2001-2002: Nitro I & II: IB blade reference designs (SNL) 2.2 GHz Xeon processors, small clusters, funded early MPI/IB work, and Cadillac (LANL) 128 node cluster



2003: Catalyst: 128 nodes 4X PCI-X IB (SNL), Blue Steel: 256 dual nodes 4X PCI-X (LANL), 96 nodes 4X PCI-X Viz Red RoSE (SNL)

2004: Catalyst: Added 85 nodes 4X PCIe IB, 288 port IB switch(SNL), ~300 nodes 4X PCIe Viz Red Rose (SNL)

2005: Thunderbird and Talon: 4,480 and 128 dual 3.6 Ghz nodes, 4X PCIe IB (SNL)
Lustre/IB production @ SNL Red RoSE



SNL, LANL, LLNL have ~12,000 nodes of IB today and more to come



DOE Goals for InfiniBand



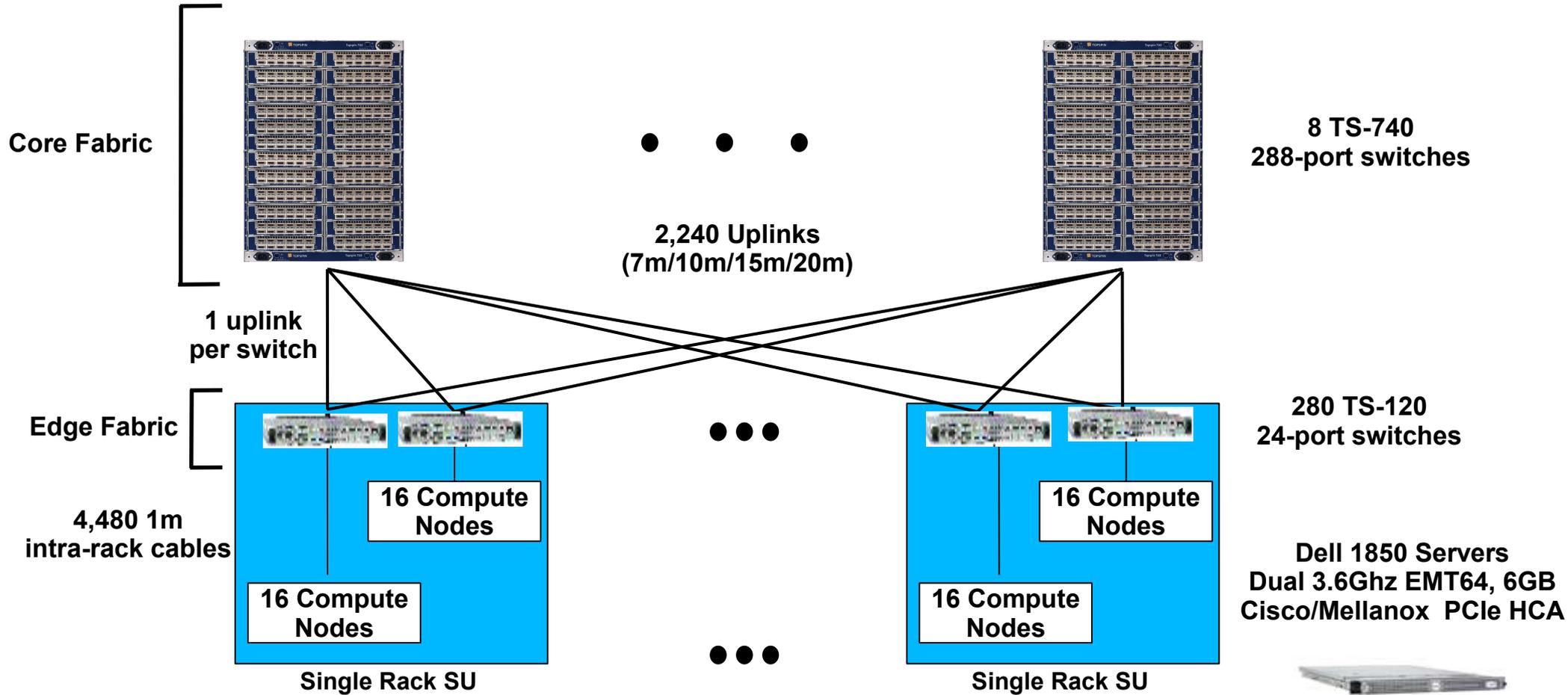
- To accelerate the development of an Linux IB software stack for HPC
 - High performance (high bandwidth, low latency, low CPU overhead)
 - Scalability
 - Robustness
 - Portability
 - Reliability
 - Manageability
 - Single open source SW stack, diagnostic and management tools supported across multiple (i.e. all) system vendors
 - Integrate IB SW stack into mainline Linux kernel at kernel.org
 - Get stack into Linux distributions (RedHat, SuSE, etc.)

OpenFabrics was formed around these goals

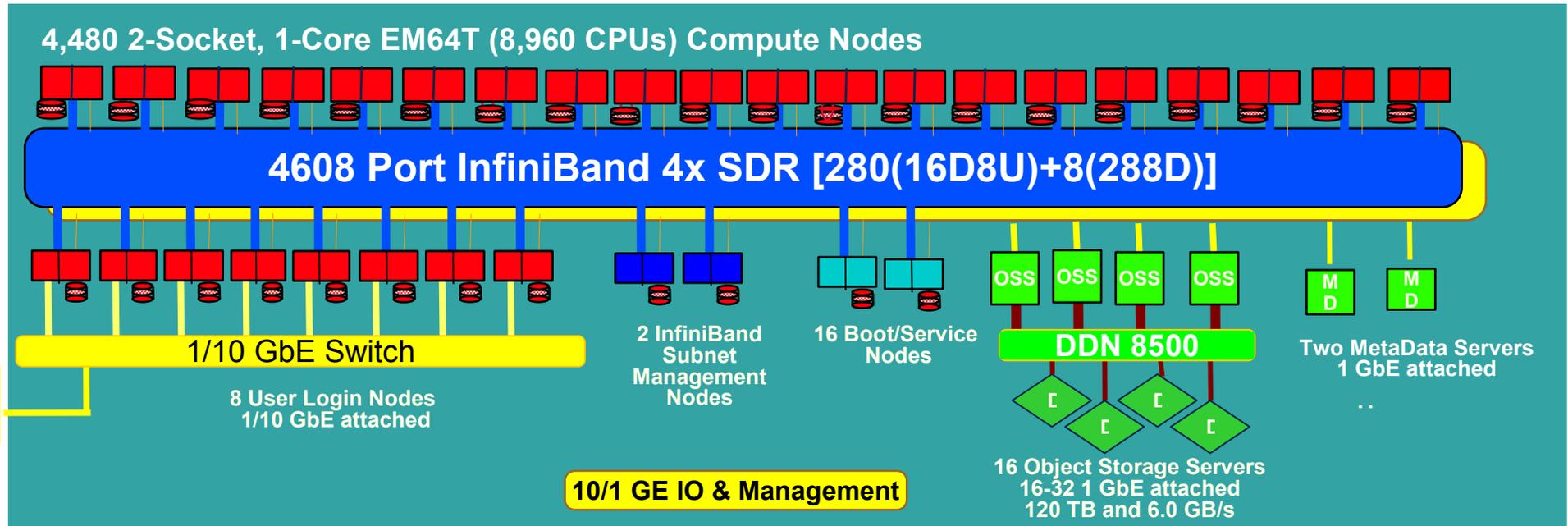
DOE ASC PathForward program has been funding OpenFabrics development since early 2005

Sandia Thunderbird Cluster

8,960 Processor, 65TF/s



Sandia Thunderbird Architecture



System Parameters

- 14.4 GF/s dual socket 3.6 GHz single core Intel SMP nodes DDR-2 400 SDRAM
- 50% blocking (2:1 oversubscription of InfiniBand fabric)
- ~300 InfiniBand switches to manage
- ~9,000 InfiniBand ports
- ~33,600 meters (or 21 miles) of 4X InfiniBand copper cables
- ~10,000 meters (or 6 miles) of copper Ethernet cables
- 26,880 1 GB DDR-2 400 SDRAM modules
- 1.8 MW of power, 400 tons of cooling

#5 in Top500
38.2 Tflops on 3721 nodes
71% efficiency

Nodes	Stack	Runtime	Memory	Result	Efficiency	Date
3721	MVAPICH,VAPI	7.35 hrs	68%	38.27 TF	71.42%	2005
4347	OpenMPI,OFED	6.72 hrs	65%	52.57 TF	83.98%	2006
4347	OpenMPI,OFED	8.37 hrs	70%	52.71 TF	84.20%	2006
4347	OpenMPI,OFED	9.44 hrs	73%	53.00 TF	84.66%	2006

The efficiencies at large scale were possible because of

- OpenFabrics (OFED 1.0)
- OpenMPI 1.1.2
- Memfree HCA firmware
- Stunt mode Linux (no RAID, no HD, no IPMI, no PFS, no random daemons)

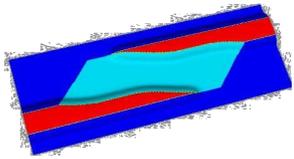


Thunderbird Infiniband Software

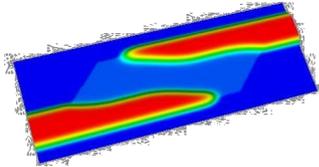


- Sandia Thunderbird Production Computing (4,480 nodes; 8,960 processors)
 - **Past year**
 - RHEL4
 - Running Cisco/Mellanox VAPI proprietary software stack
 - CiscoSM and Cisco diagnostics
 - MVAPICH1
 - **Currently upgrading production environment**
 - OFED v1.0/1.1
 - OpenMPI v1.1.2 or v1.2
 - RHEL4U4
 - OpenFabrics management and diagnostic tools

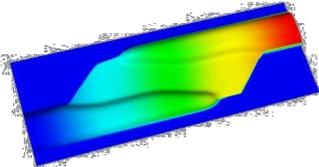
Thunderbird is being used in an increasing number of simulation and optimization application problems



Nominal depth

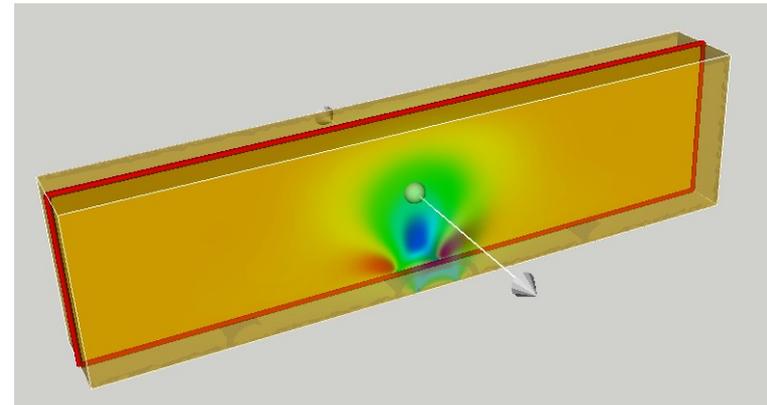


Etched depth

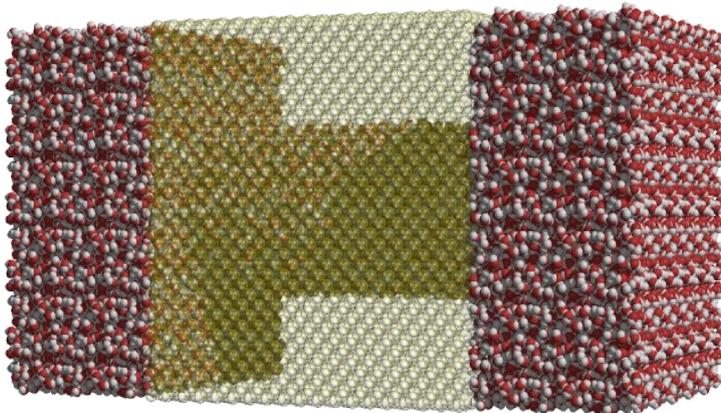


Travel time

Etch-compensated minimum-dispersion microfluidic displacer obtained through shape optimization. Sundance was used to simulate the device manufacturing and performance (K.Long, A. Skulan, S. Margolis, and P. Boggs, 2005)



Flow velocity above an electrode gap in a microfluidic channel. Sundance was used to develop a simulator coupling electrical, thermal, and fluid effects. (K. Long and B. van Bloemen Waanders, 2006)



Electric potential in a nanoscale pore in a biological membrane. Sundance was used as the continuum-level component in a coupled atomistic-continuum simulator (Debuschere and Adalsteinsson, 2006)



OPENFABRICS
ALLIANCE

Petascale InfiniBand Cluster Requirements



- Sandia Cplant, LLNL MCR, and LANL Pink Clusters (1500-2000 processors)
 - Commodity HW, high speed interconnect, Linux, and other open source SW
 - Showed that it was possible to bring scientific computing to the masses
- Sandia Thunderbird
 - Pushing scientific simulations
 - Scalability, InfiniBand, OpenFabrics, OpenMPI, and Linux to 4000 nodes
- Future - Petascale InfiniBand Linux clusters
 - 4000 nodes with multi-core CPUs feasible in next 2-3 years
- Scalability to this level will require:
 - < 1us pt2pt latency and increased message injection rate (15-20M/s) for small messages
 - Hardware and OF software support for congestion control architecture
 - Fully adaptive routing (addition to IB spec.)
 - Cheap reliable fiber for 4X/12X DDR and QDR (match the cost of copper)
 - High performance (near line rate - SDR, DDR, QDR) native IB-IB routing
 - Reliable multicast (up to a minimum of 128 peers)
 - More requirements presented at Sonoma Workshop and joint OFA-IBTA workshop (<http://openfabrics.org/conference.html>)

Achieving these goals will be a collaborative effort between OFA, IBTA, and HPC community



OPENFABRICS
ALLIANCE



For more information

Matt Leiningner mleini@sandia.gov