ASCI Terascale Simulation Requirements and Deployments

David A. Nowak

ASCI Program Leader

Mark Seager

ASCI Terascale Systems Principal Investigator

Lawrence Livermore National Laboratory

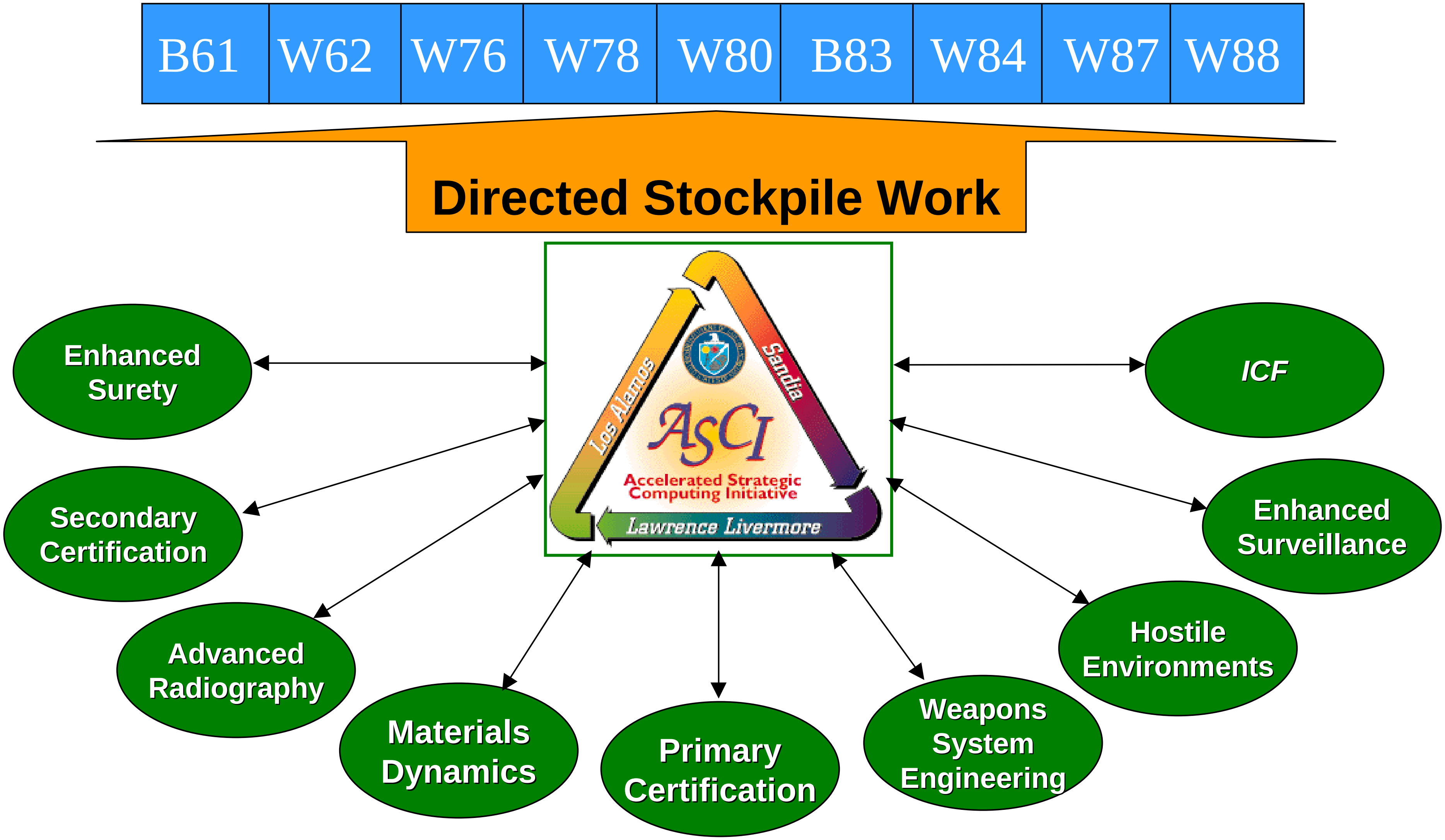
University of California

*Work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Overview

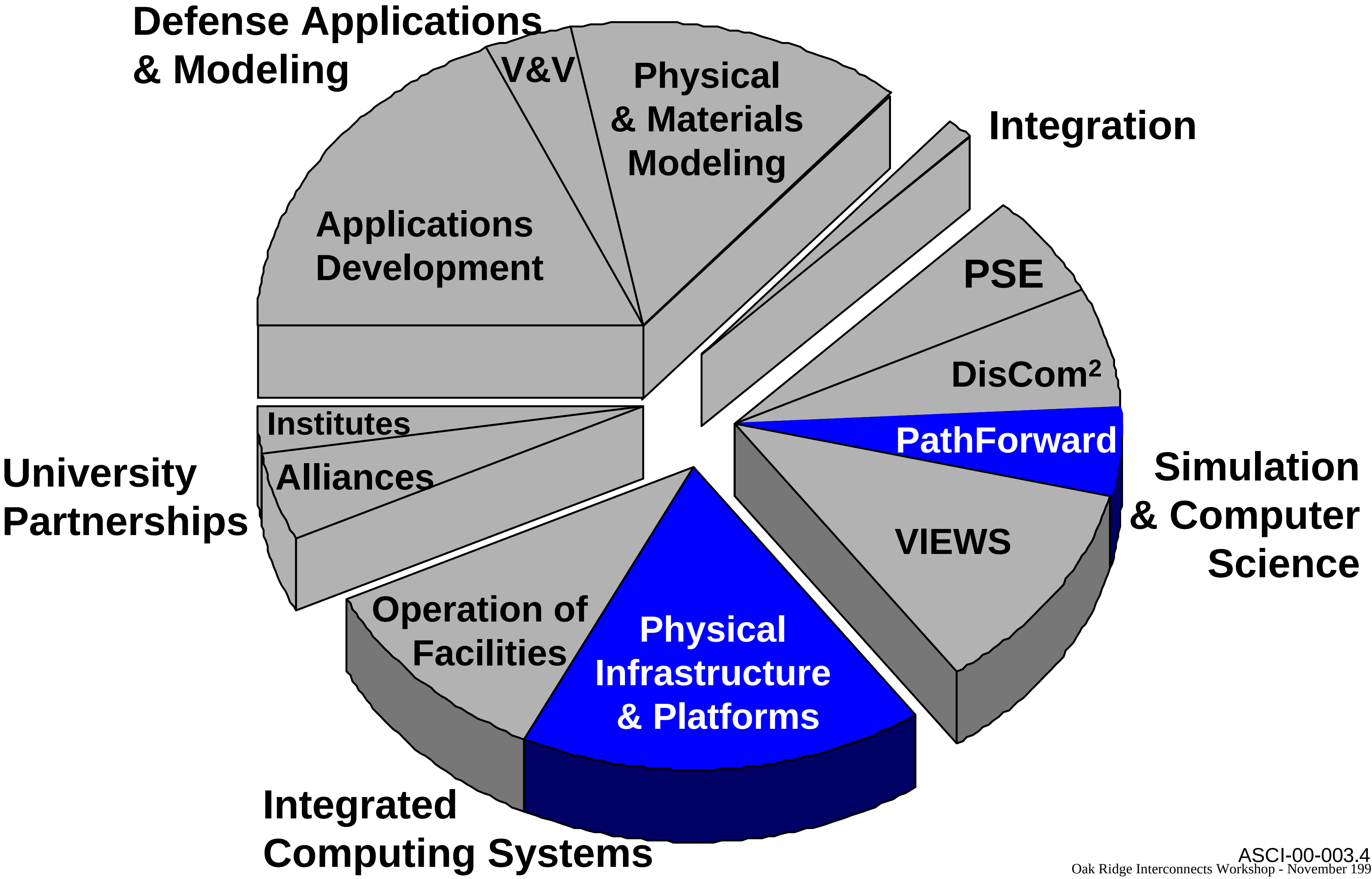
- ASCI program background
- Applications requirements
- Balanced terascale computing environment
- Red Partnership and CPLANT
- Blue-Mountain partnership
  - Sustained Stewardship TeraFLOP/s (SST)
- Blue-Pacific partnership
  - Sustained Stewardship TeraFLOP/s (SST)
- White partnership
- Interconnect issues for future machines

A successful Stockpile Stewardship Program requires a successful ASCI

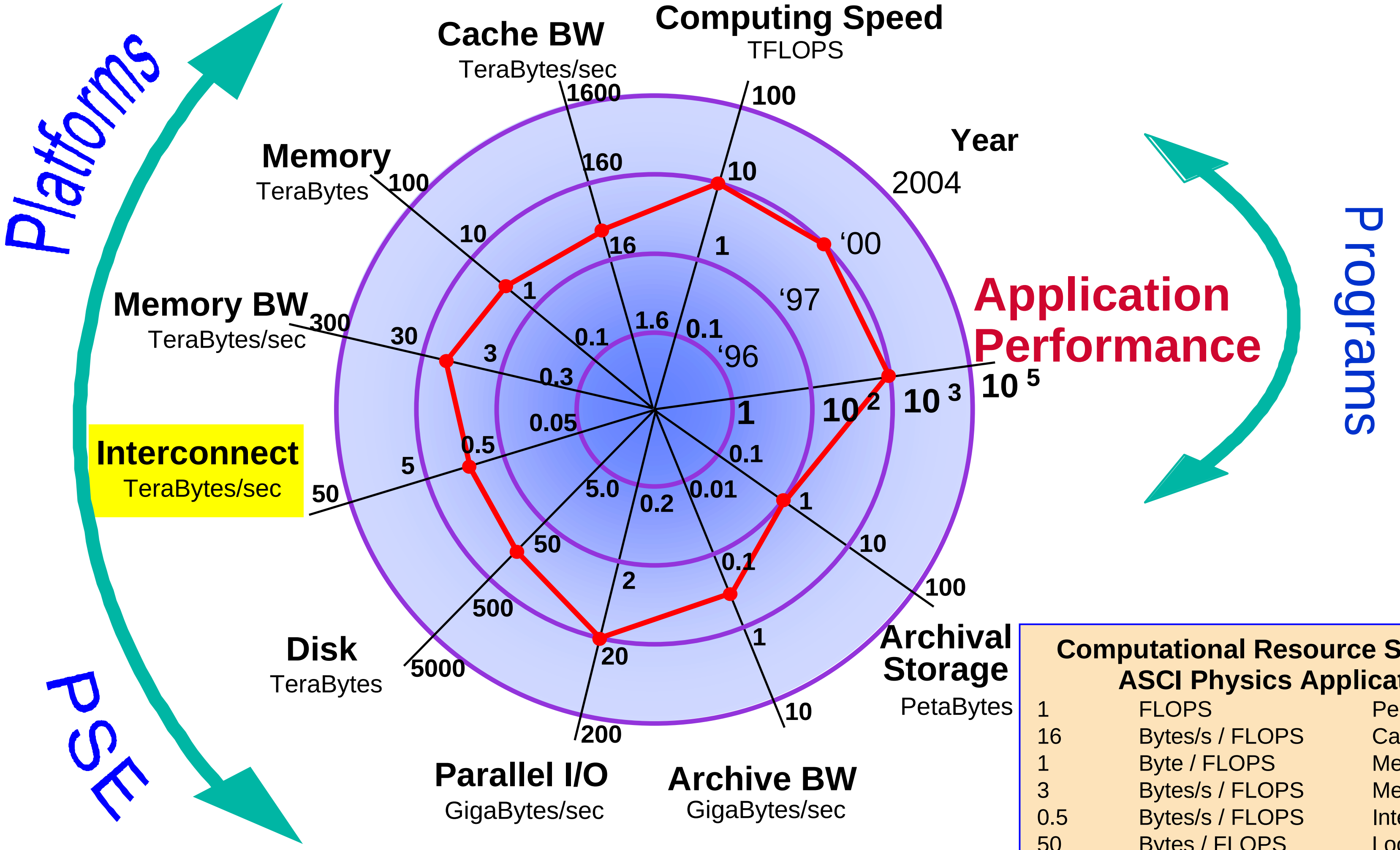


B61	W62	W76	W78	W80	B83	W84	W87	W88

ASCI's major technical elements meet Stockpile Stewardship requirements


Defense Applications & Modeling
V&V
Physical & Materials Modeling
Applications Development
Integration
PSE
DisCom²
PathForward
VIEWS
Simulation & Computer Science
Institutes
Alliances
University Partnerships
Operation of Facilities
Physical Infrastructure & Platforms
Integrated Computing Systems

Example terascale computing environment in CY00 with ASCI White at LLNL



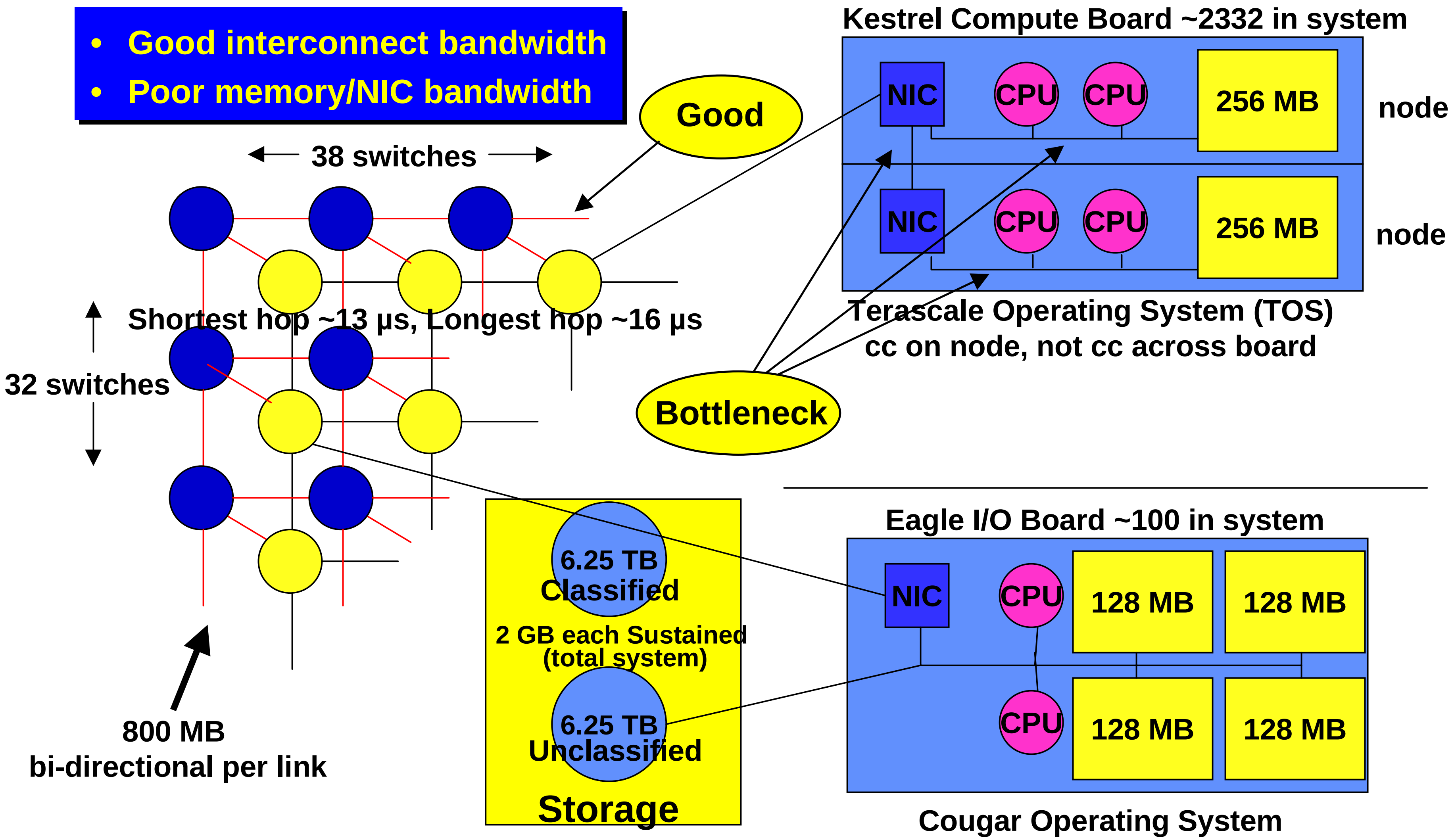
ASCI is achieving programmatic objectives, but the computing environment will not be in balance at LLNL for the ASCI White platform.

Computational Resource Scaling for ASCI Physics Applications

1	FLOPS	Peak Compute
16	Bytes/s / FLOPS	Cache BW
1	Byte / FLOPS	Memory
3	Bytes/s / FLOPS	Memory BW
0.5	Bytes/s / FLOPS	Interconnect BW
50	Bytes / FLOPS	Local Disk
0.002	Byte/s / FLOPS	Parallel I/O BW
0.0001	Byte/s / FLOPS	Archive BW
1000	Bytes / FLOPS	Archive

SNL/Intel ASCI Red

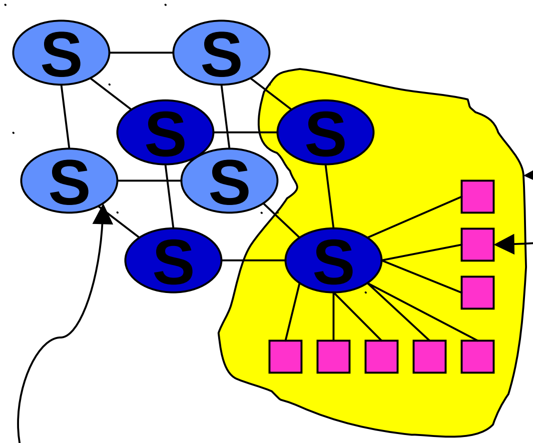
- Good interconnect bandwidth
- Poor memory/NIC bandwidth



Aggregate link bandwidth = 1.865 TB/s

SNL/Compaq ASCI C-Plant

C-Plant is located at Sandia National Laboratory
Currently a Hypercube, C-Plant will be reconfigured as a mesh in 2000
C-Plant has 50 scalable units
LINUX Operating System
C-Plant is a “Beowulf” configuration



Scalable Unit:

- •16 “boxes”
- •2   16 port Myricom switches
  - •160 MB each direction
  - •320 MB total

“Box:”

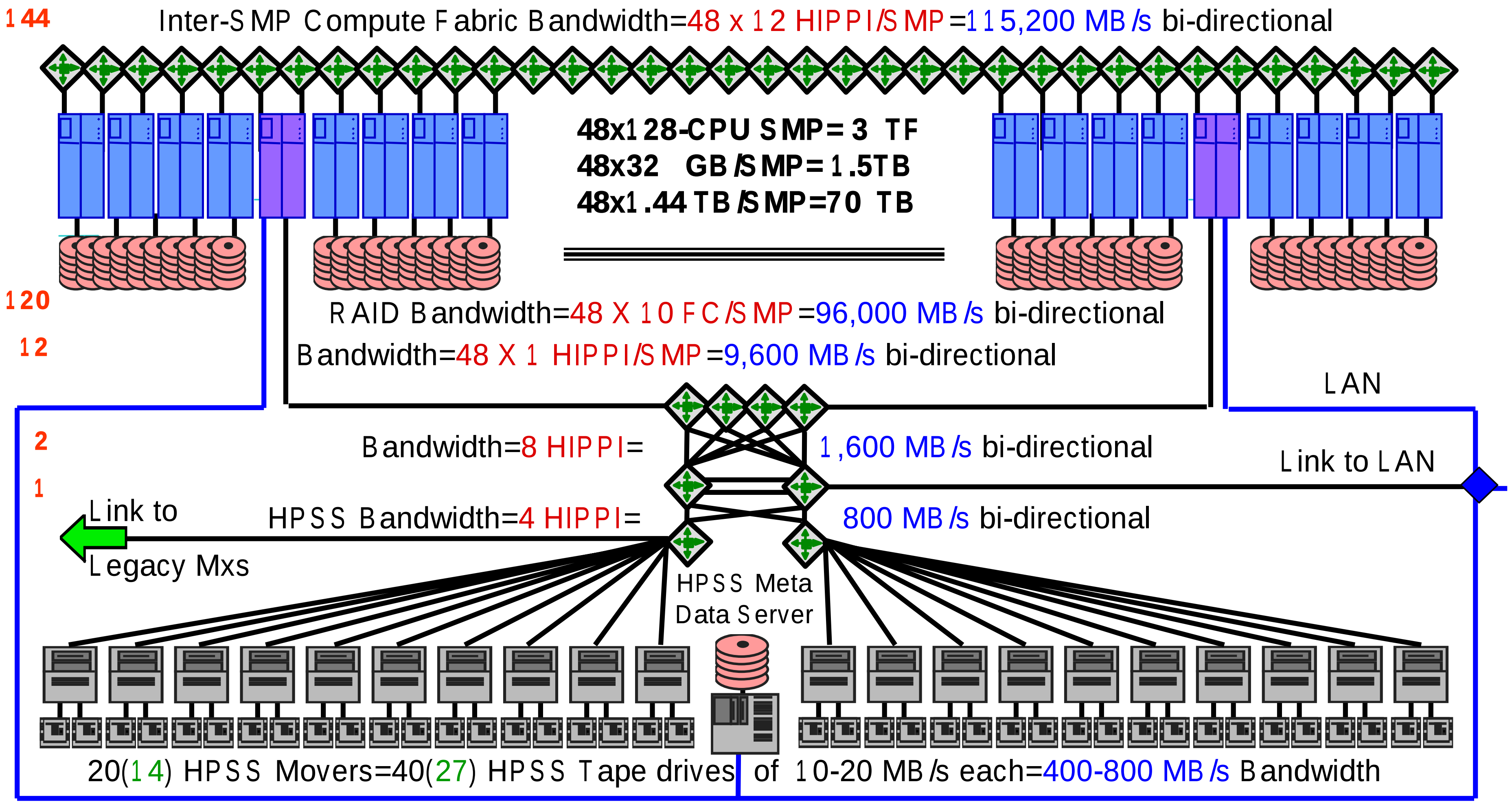
- •500 MHz Compaq eV56 processor
- •192 MB SDRAM
- •55 MB NIC
- •Serial port
- •Ethernet port
- •______ Disk space

Hypercube

This is a research project — a long way from being a production system.

LANL/SGI/Cray ASCI Blue Mountain

3.072 TeraOPS Peak



144 Inter-SMP Compute Fabric Bandwidth=48 x 12 HIPPI/SMP=115,200 MB/s bi-directional

48x128-CPU SMP=3 TF
48x32 GB/SMP=1.5TB
48x1.44 TB/SMP=70 TB

120 RAID Bandwidth=48 X 10 FC/SMP=96,000 MB/s bi-directional

12 Bandwidth=48 X 1 HIPPI/SMP=9,600 MB/s bi-directional

LAN

2 Bandwidth=8 HIPPI=1,600 MB/s bi-directional

Link to LAN

1 Link to Legacy Mxs

HPSS Bandwidth=4 HIPPI=800 MB/s bi-directional

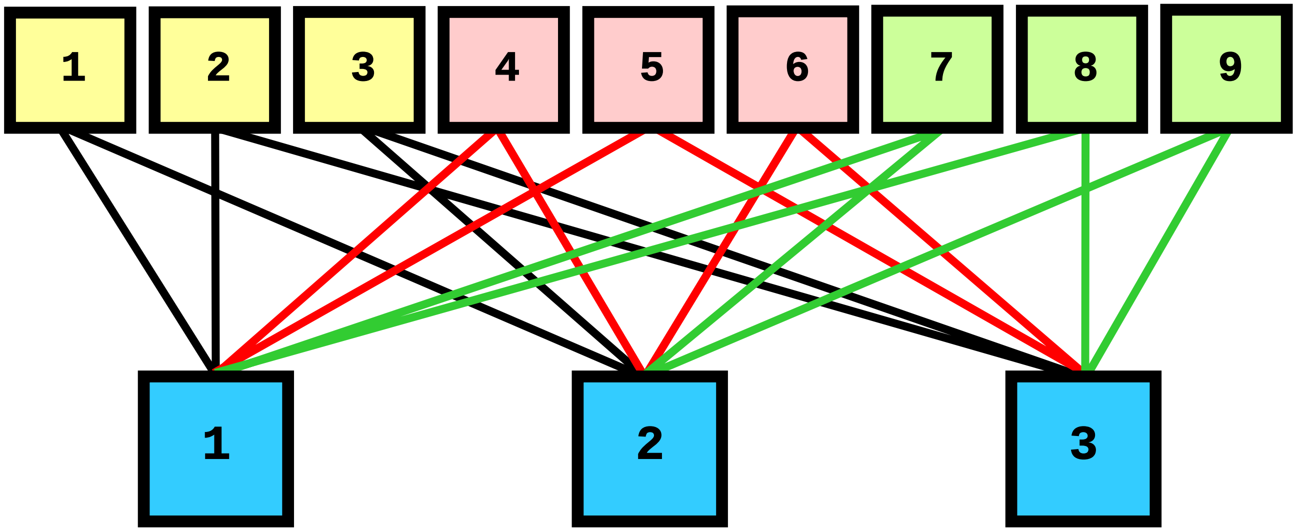
HPSS Meta Data Server

20(14) HPSS Movers=40(27) HPSS Tape drives of 10-20 MB/s each=400-800 MB/s Bandwidth

Aggregate link bandwidth = 0.115 TB/s

Blue Mountain
Planned GSN Compute Fabric

9 Separate 32x32 X-Bar Switch Networks



Expected Improvements

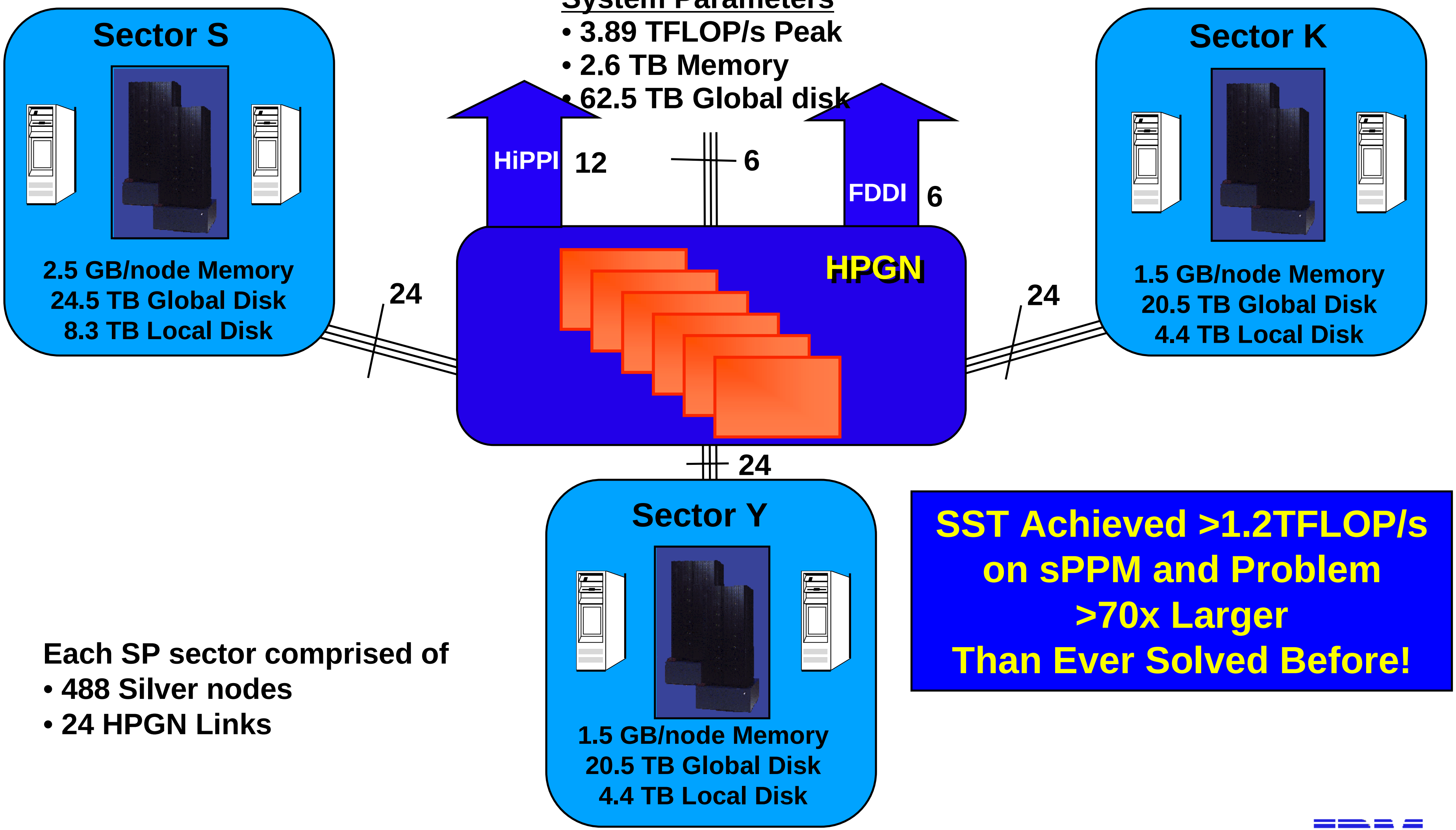
Throughput	115,200 MB/s	=>	460,800 MB/s,	4x
Link Bandwidth	200 MB/s	=>	1,600 MB/s,	8x
Round Trip Latency	110 µs	=>	~ 10 µs	,11x

3 Groups of 16 Computers each

Aggregate link bandwidth = 0.461 TB/s

LLNL/IBM Blue-Pacific

3.889 TeraOP/s Peak



System Parameters

- 3.89 TFLOP/s Peak
- 2.6 TB Memory
- 62.5 TB Global disk

Each SP sector comprised of

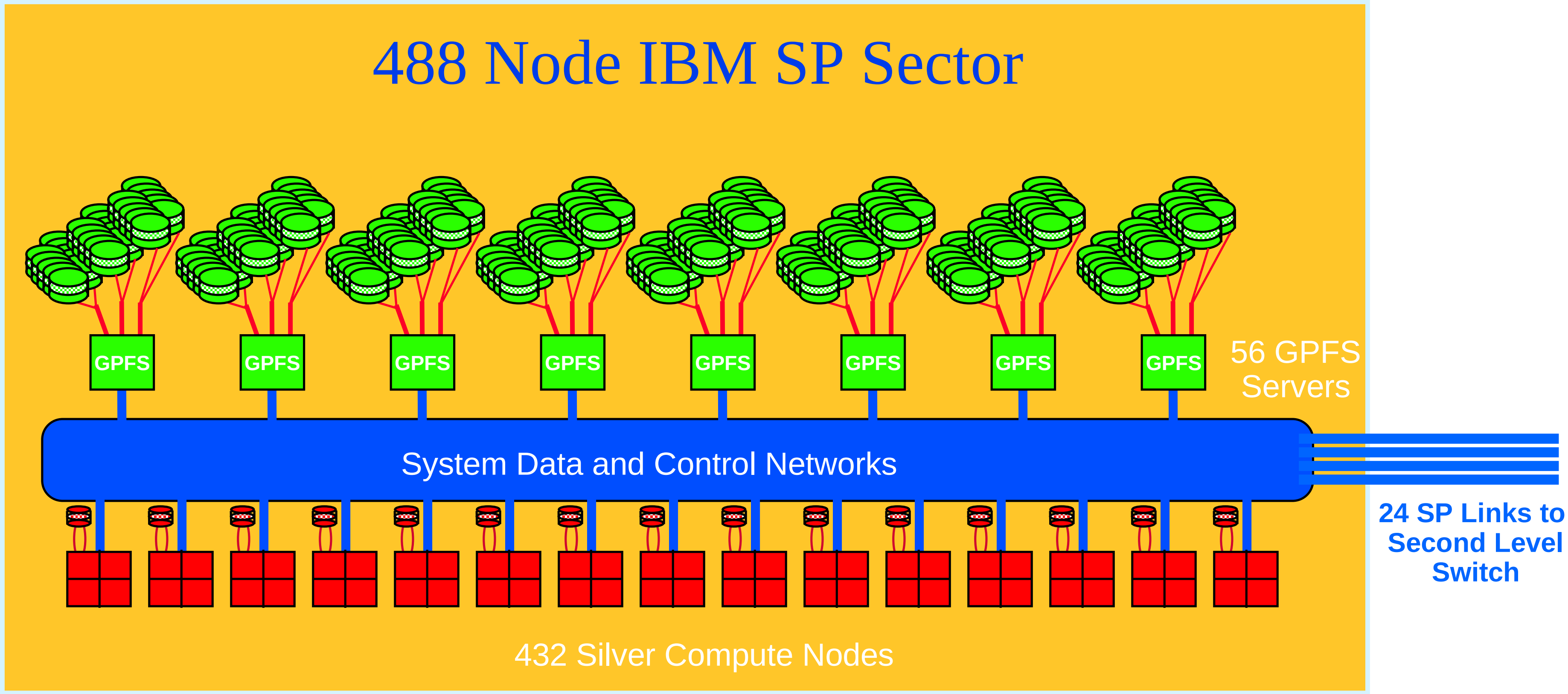
- 488 Silver nodes
- 24 HPGN Links

SST Achieved >1.2TFLOP/s on sPPM and Problem >70x Larger Than Ever Solved Before!

Aggregate link bandwidth = 0.439 TB/s

IBM

I/O Hardware Architecture of SST


488 Node IBM SP Sector
GPFS
56 GPFS Servers
System Data and Control Networks
24 SP Links to Second Level Switch
432 Silver Compute Nodes

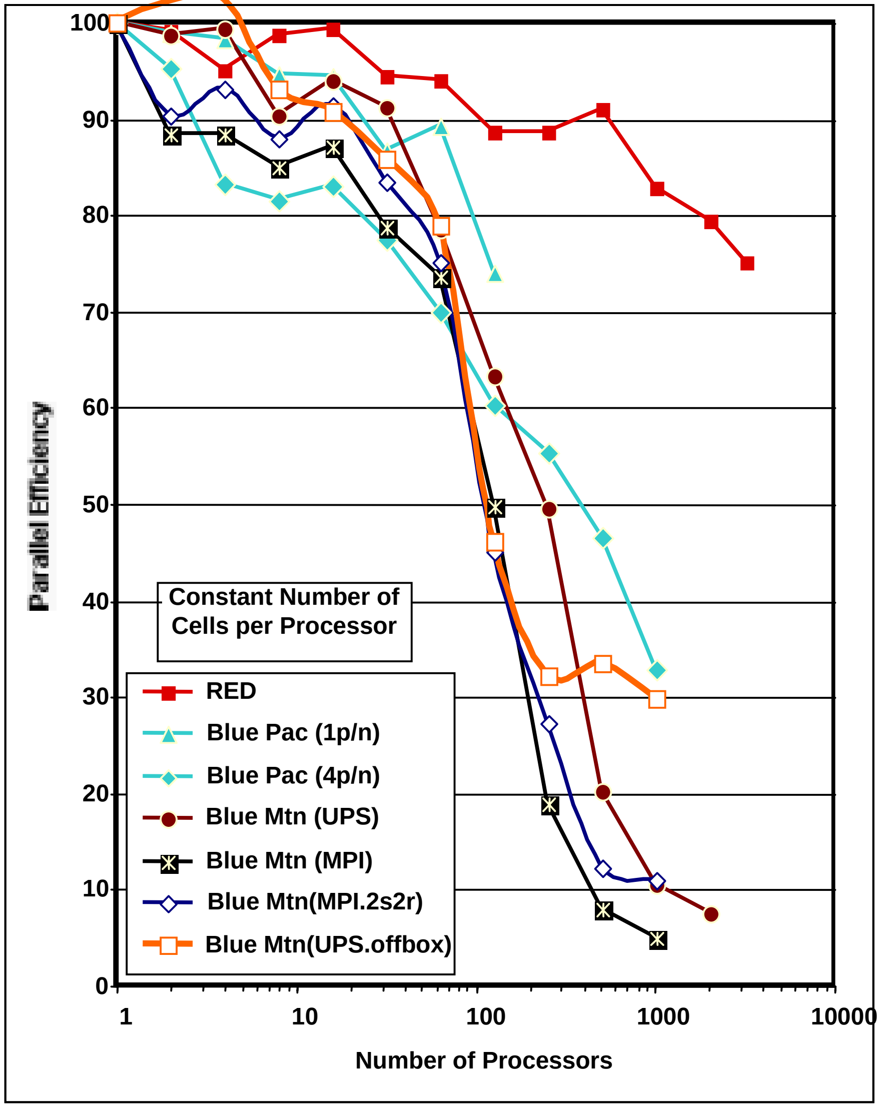
Each SST Sector

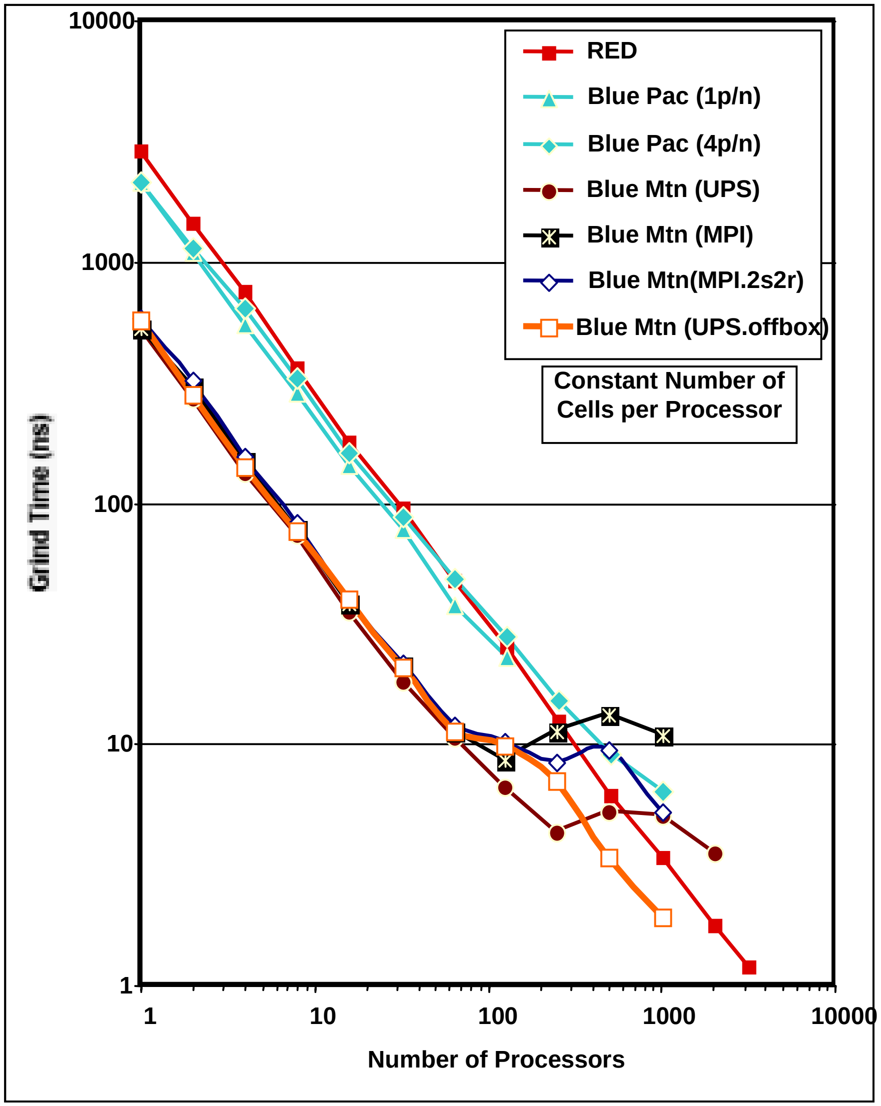
- Has local and global I/O file system
- 2.2 GB/s delivered global I/O performance
- 3.66 GB/s delivered local I/O performance
- Separate SP first level switches
- Independent command and control
- **Link bandwidth = 300 Mb/s Bi-directional**

Full system mode

- Application launch over full 1,464 Silver nodes
- 1,048 MPI/us tasks, 2,048 MPI/IP tasks
- High speed, low latency communication between all nodes
- Single STDIO interface

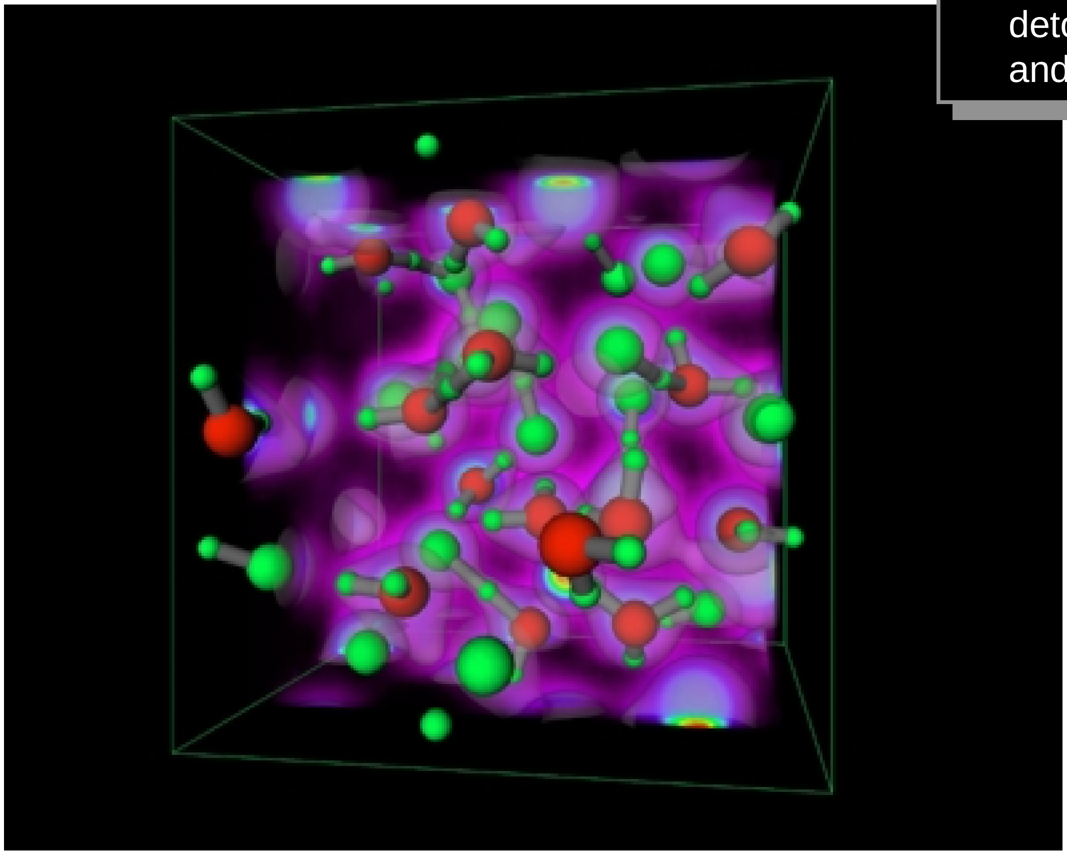
Partisn (S_N-Method) Scaling





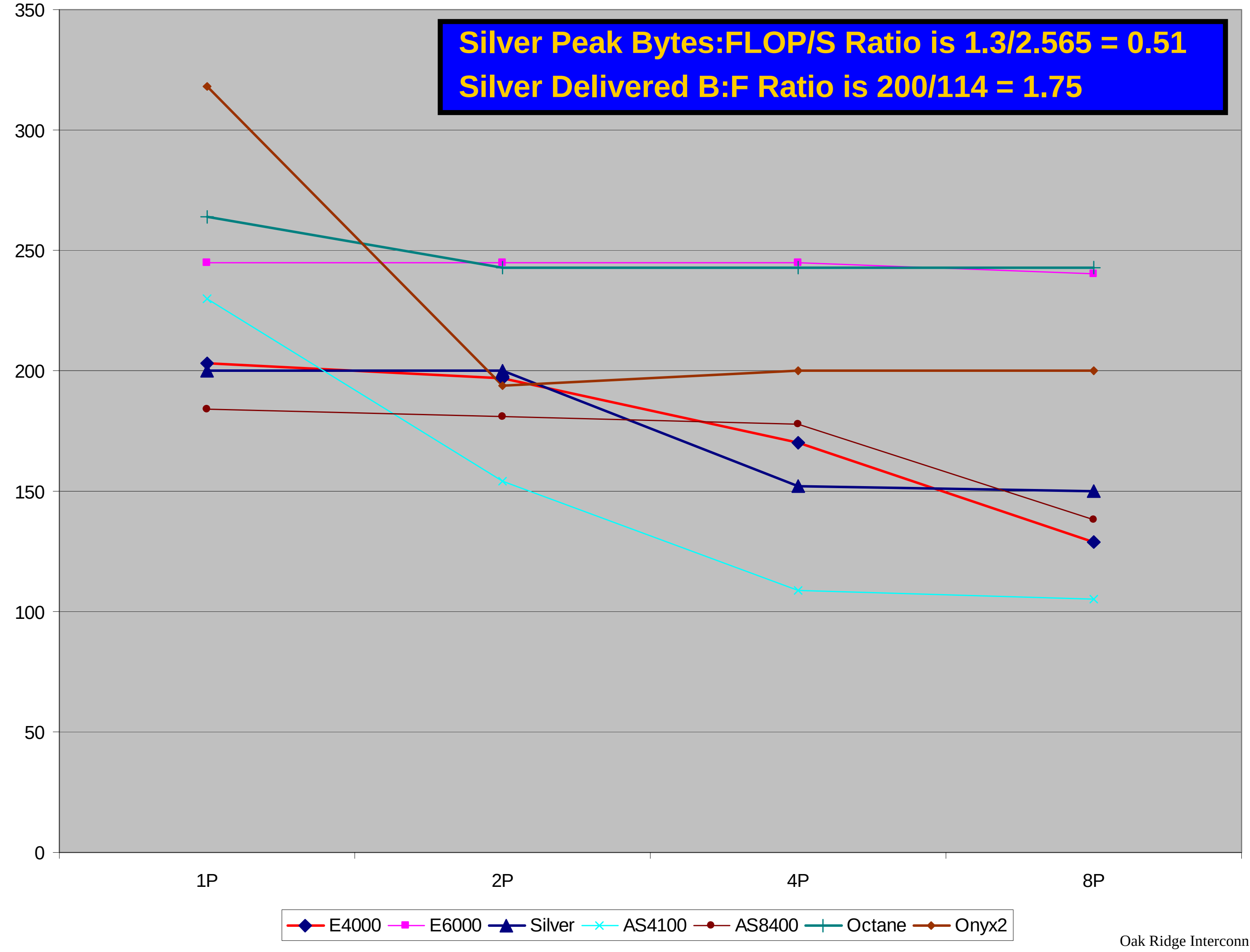
The JEEP calculation adds to our understanding the performance of insensitive high explosives

- This calculation involved 600 atoms (largest number ever at such a high resolution) with 1,920 electrons, using about 3,840 processors
- This simulation provides crucial insight into the detonation properties of IHE at high pressures and temperatures.



- **Relevant experimental data (e.g., shock wave data) on hydrogen fluoride (HF) are almost nonexistent because of its corrosive nature.**
- **Quantum-level simulations, like this one, of HF- H_2O mixtures can substitute for such experiments.**

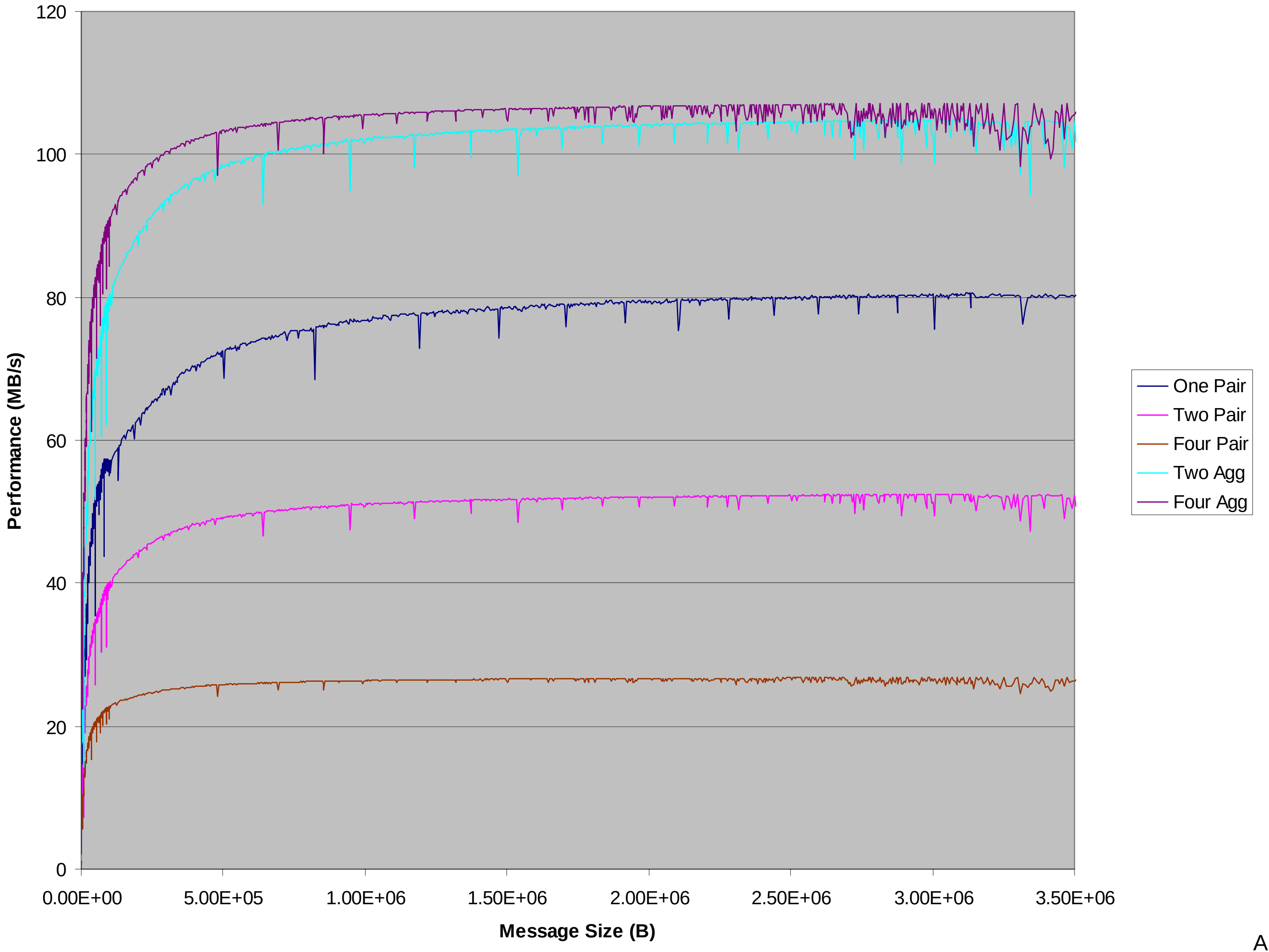
Silver Node delivered memory bandwidth is around 150-200 MB/s/process



Silver Peak Bytes:FLOP/S Ratio is 1.3/2.565 = 0.51

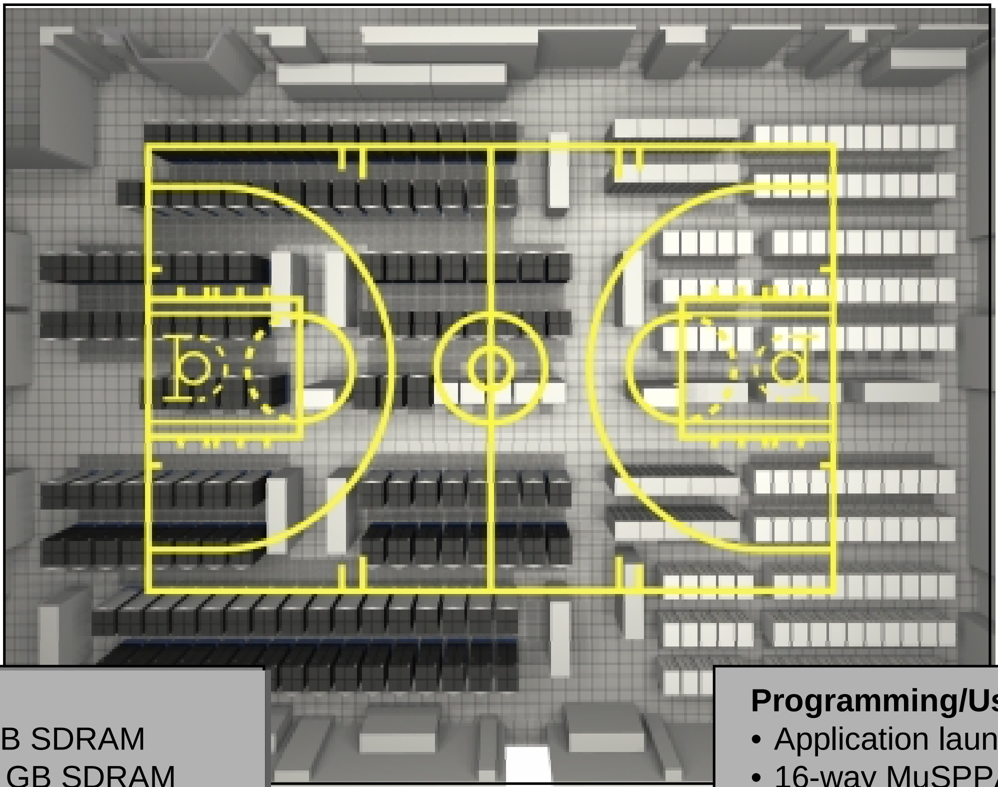
Silver Delivered B:F Ratio is 200/114 = 1.75

MPI_SEND/US delivers low latency and aggregate high bandwidth, but counter intuitive behavior per MPI task



LLNL/IBM White

10.2 TeraOPS Peak



MuSST (PERF) System

- 8 PDEBUG nodes w/16 GB SDRAM
- ~484 PBATCH nodes w/8 GB SDRAM
- 12.8 GB/s delivered global I/O performance
- 5.12 GB/s delivered local I/O performance
- 16 GigaBit Ethernet External Network
- Up to 8 HIPPI-800

Programming/Usage Model

- Application launch over ~492 NH-2 nodes
- 16-way MuSPPA, Shared Memory, 32b MPI
- 4,096 MPI/US tasks
- Likely usage is 4 MPI tasks/node with 4 threads/MPI task
- Single STDIO interface

Aggregate link bandwidth = 2.048 TB/s

Five times better than the SST; Peak is three times better

Ratio of Bytes:FLOPS is improving

Interconnect issues for future machines — Why Optical? —

- Need to increase Bytes:FLOPS ratio
  - Memory bandwidth (cache line) utilization will be dramatically lower for codes that utilize arbitrarily connected meshes and adaptive refinement indirect addressing.
  - Interconnect bandwidth must be increased and latency must be reduced to allow a broader range of applications and packages to scale well
- To get very large configurations (30 70 100 TeraOPS) larger SMPs will be deployed
  - For fixed B:F interconnect ratio this means that more bandwidth coming out of an SMP
  - Multiple pipes/planes will be used Optical reduces cable count
- Machjne footprint is growing 24,000 square feet may require optical
- Network interface paradigm
  - Virtual memory direct memory access
  - Low-latency remote get/put
- Reliability Availability and Serviceability (RAS)

ASCI Terascale Simulation Requirements and Deployments

David A. Nowak

ASCI Program Leader

Mark Seager

ASCI Terascale Systems Principal Investigator

Lawrence Livermore National Laboratory

University of California

*Work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.