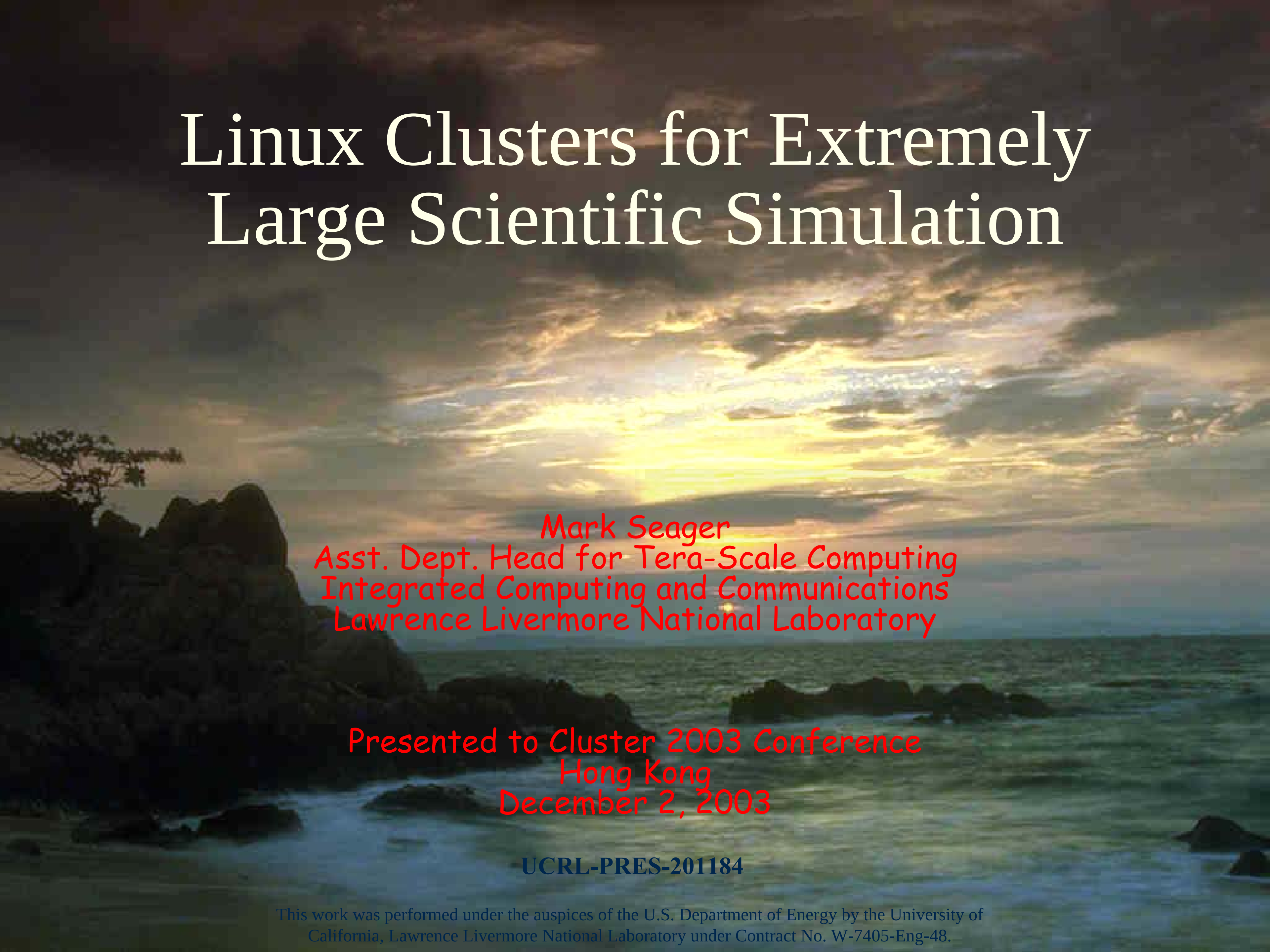


Linux Clusters for Extremely Large Scientific Simulation

Mark Seager
Asst. Dept. Head for Tera-Scale Computing
Integrated Computing and Communications
Lawrence Livermore National Laboratory

Presented to Cluster 2003 Conference
Hong Kong
December 2, 2003

UCRL-PRES-201184

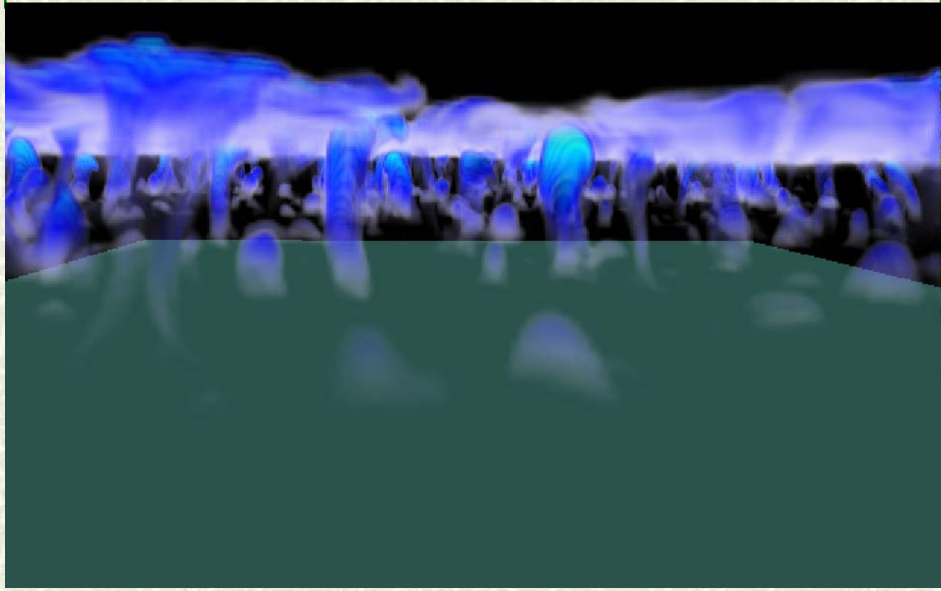
This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

Overview

- Computing at LLNL today and tomorrow
- LLNL platform and Linux strategy
- MCR+ALC multi-cluster Linux cluster architecture
- Next generation simulation environment
- Science runs having huge impact on programs at LLNL

The day when the scientist, no matter how devoted, can make significant progress alone and without material help is past…

E.O Lawrence
On accepting the Nobel Prize



David Stevens – Cumulus Convection

LLNL has huge installed base of clusters for scientific simulation and modeling

These machines represent 32.7 teraFLOP/s peak aggregate computing capability and bring computational science of scale to the ASCI program and to unclassified researchers from every program at LLNL.

Lawrence Livermore National Laboratory
MCR Cluster - October 2002

MCR – 21st Top500 Number 3 with 7.634 out of 11.2 teraFLOP/s

Worlds fastest Linux cluster with 1,152 Dual Xeon 2.4GHz nodes with Quadrics ELAN3 interconnect

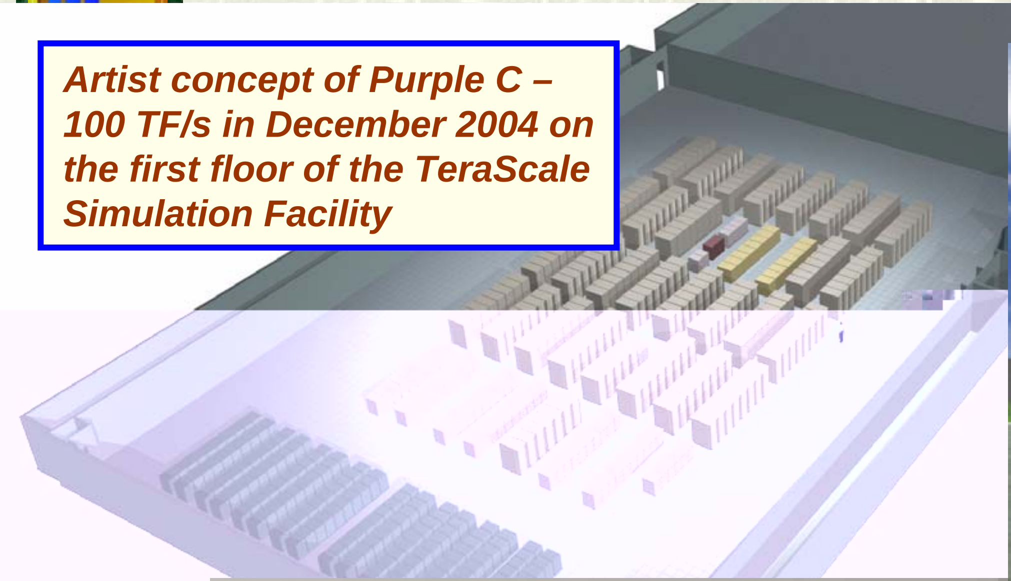
ASCI White – 18th Top500 Number 1 with 7.304 out of 12.3 teraFLOP/s

Worlds largest AIX pSeries machine with 512 16-way Power3 375 MHz NighHawk-2 with dual plane Colony interconnect

ASCI Linux Cluster – 21st Number 6 with 6.586 out of 9.2 teraFLOP/s

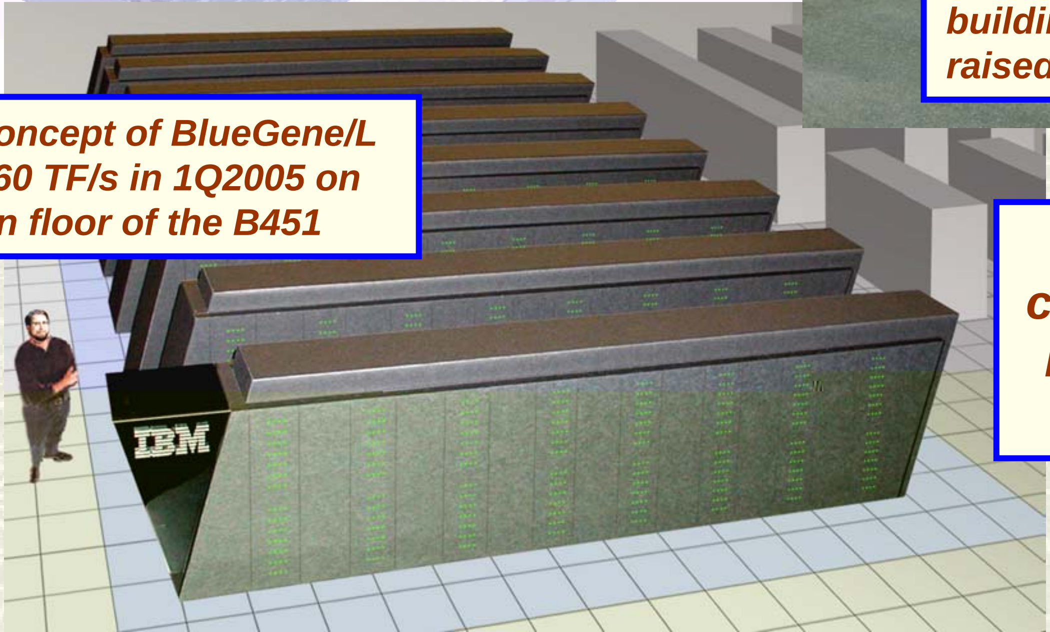
World’s second fastest Linux cluster with 960 Dual Xeon 2.4GHz nodes with Quadrics ELAN3 interconnect

The leading edge of scientific simulation is approaching 0.5 petaFLOP/s…



Artist concept of Purple C – 100 TF/s in December 2004 on the first floor of the TeraScale Simulation Facility

Artist rendition of TeraScale Simulation Facility. This 253,000 ft^2 building has 48,000 ft^2 of computer raised floor and 21MW of power



Artist concept of BlueGene/L – 180-360 TF/s in 1Q2005 on the main floor of the B451

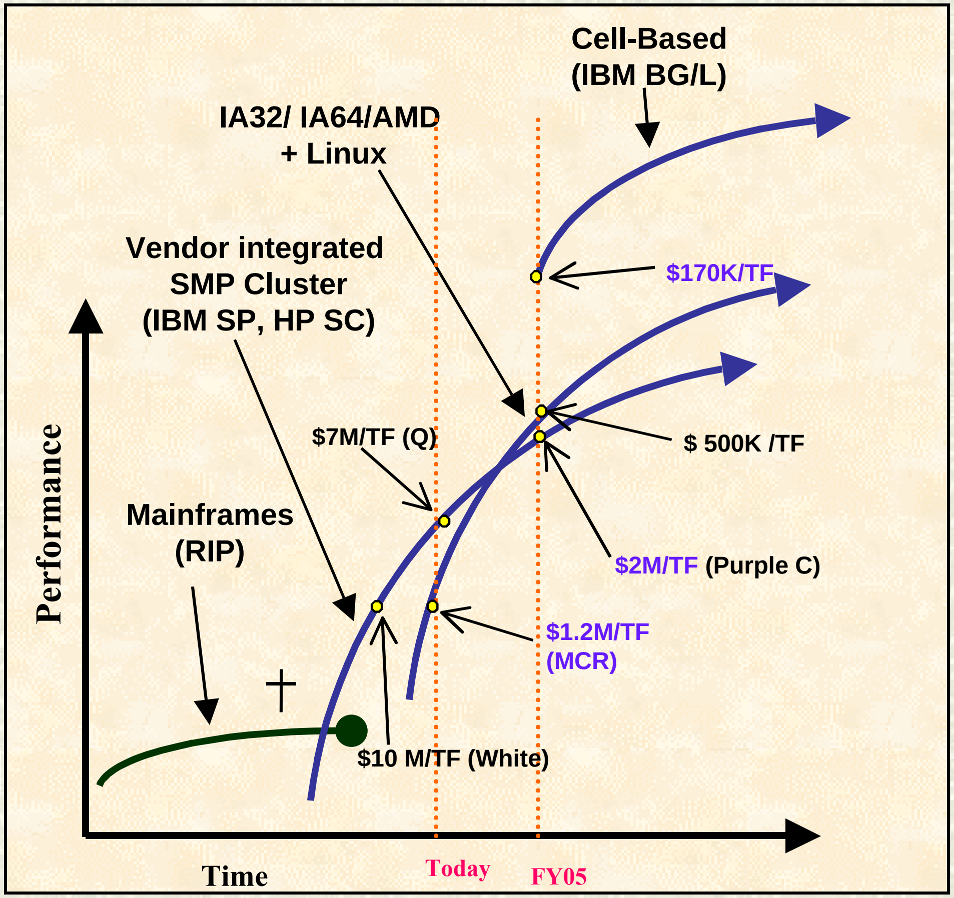
DOE investments in computational capability at LLNL represented on this slide are almost $400M

Our platform strategy is to straddle multiple technology curves to appropriately balance risk and benefit

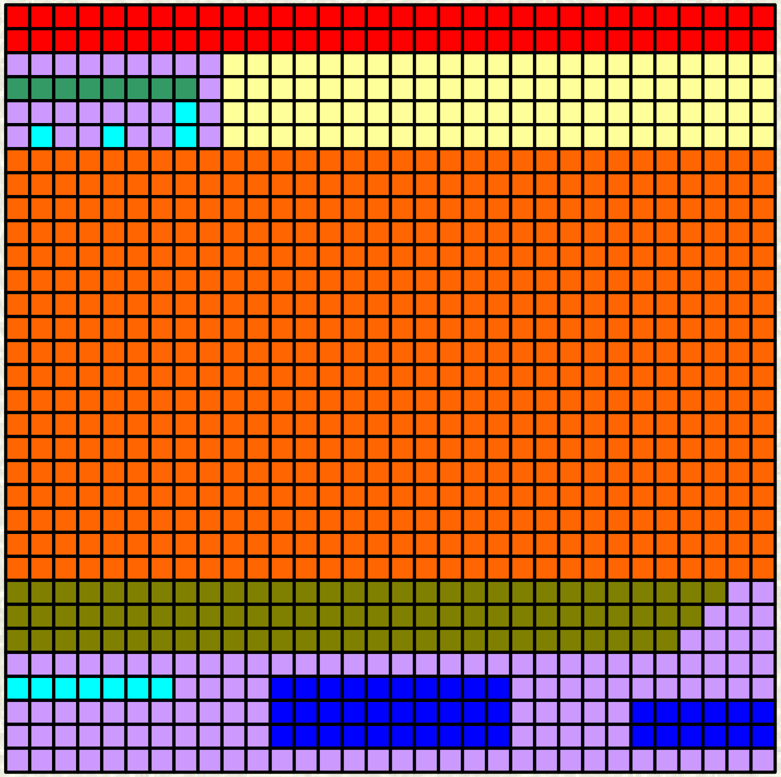
Three complementary curves…

1. Delivers to today's stockpile's demanding needs
   - Production environment
   - For "must have" deliverables now
2. Delivers transition for next generation
   - "Near production", riskier environment
   - Capacity/capability systems in a strategic mix
3. Delivers affordable path to petaFLOP/s
   - Research environment, leading transition to petaflop systems?

Any given technology curve is ultimately limited by Moore's Law

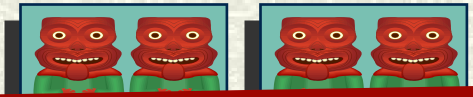


Livermore Model makes incredible demands on computing infrastructure

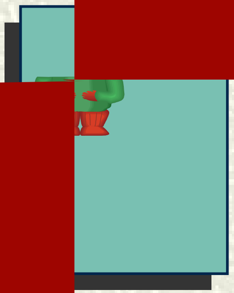


- Interactive and batch usage
  - Debugging and visualization
- Mix of job sizes
  - Capability runs at a significant portion of platform
  - Capacity runs at a small fraction of platform
- Mix of runtimes
  - Short run-time for setup
  - Many months of science run
- Many users
  - Require progress on jobs

Provide identical pro [redacted]
across scalable platf

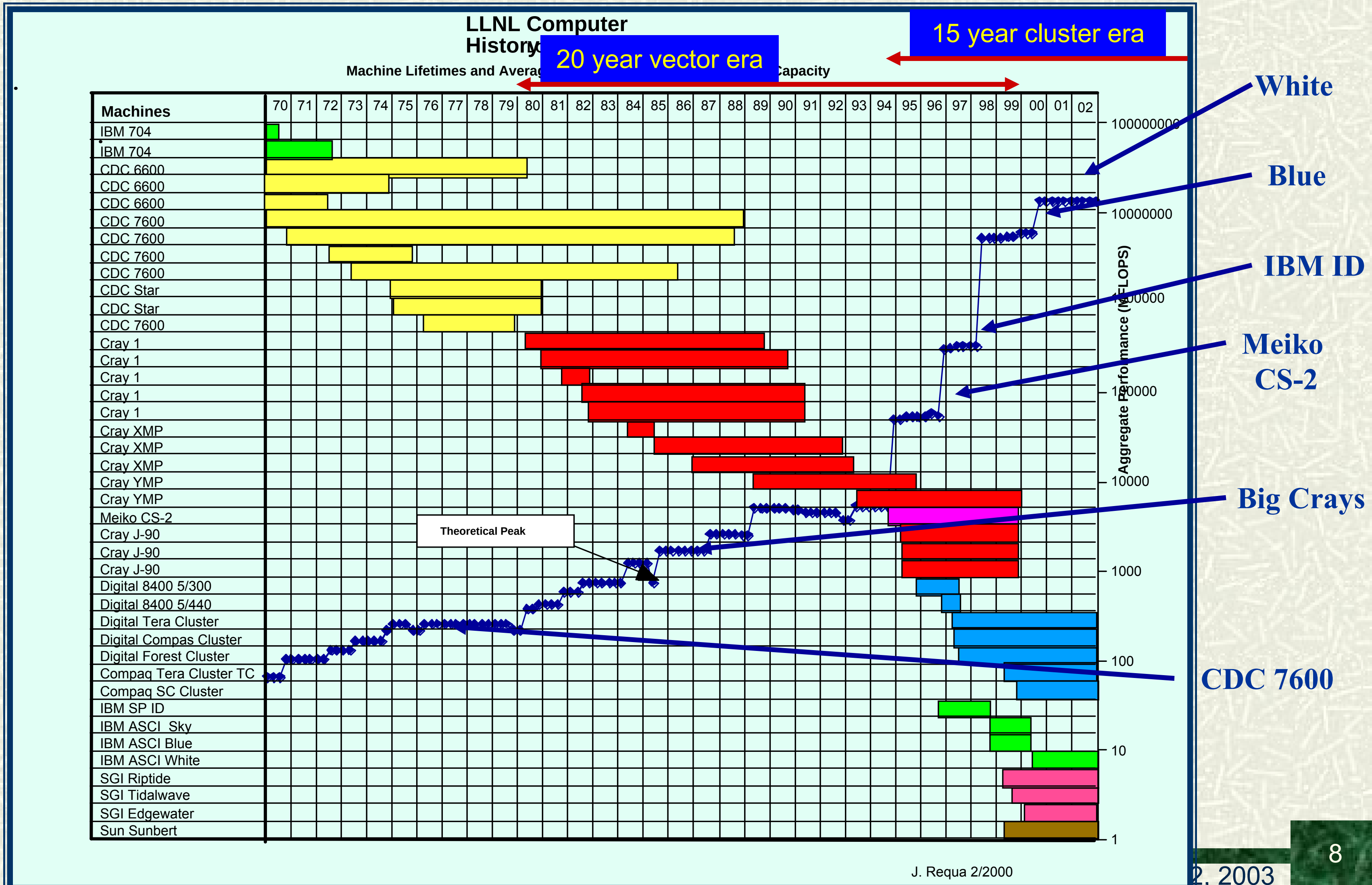


[redacted]

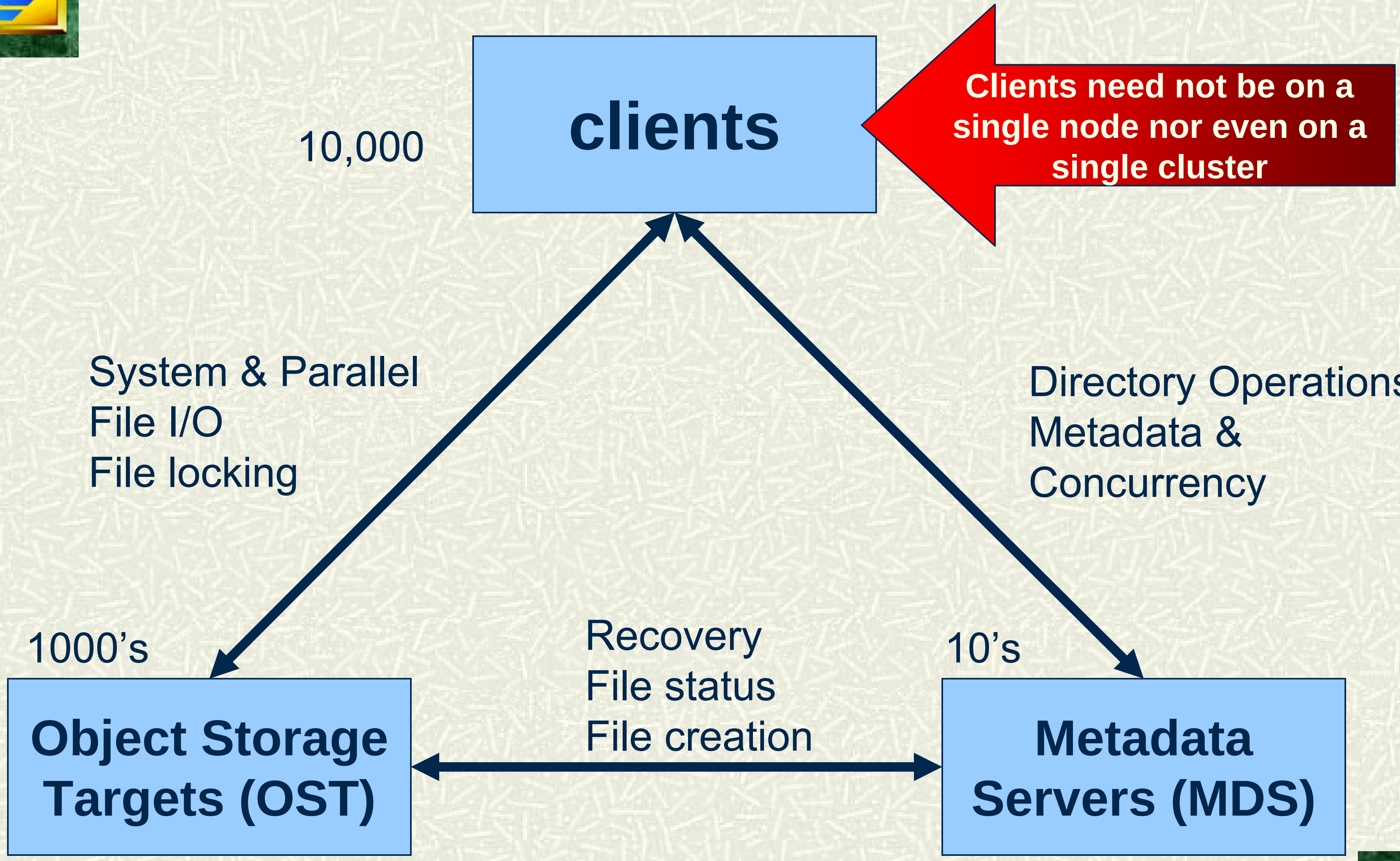


What is the essence of the killer micro revolution?

By the end of Purple in 2009, the Livermore model (clustered SMPs with incrementally improving functionality) will provide more than three-quarters the programming model longevity of Vector era.



Lustre global file system architecture designed to scale to PetaBytes of storage and TB/s bandwidth across multiple systems



clients

10,000

Clients need not be on a single node nor even on a single cluster

System & Parallel
File I/O
File locking

Directory Operations,
Metadata &
Concurrency

1000's

Object Storage Targets (OST)

Recovery
File status
File creation

10's

Metadata Servers (MDS)

QsNet Elan3 modified fat-tree interconnect for a 1152 node, 11.06 TF/s cluster

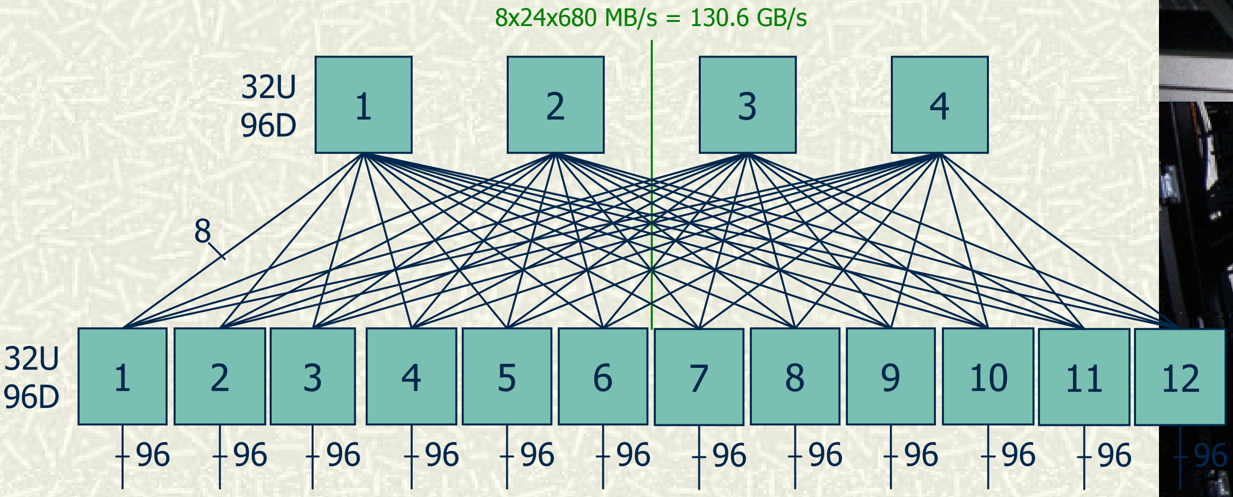
F = 1152x9.6 GF/s = 11.059 TF/s

Node B:F = 680 MB/s / 9.6 GF/s = 0.071

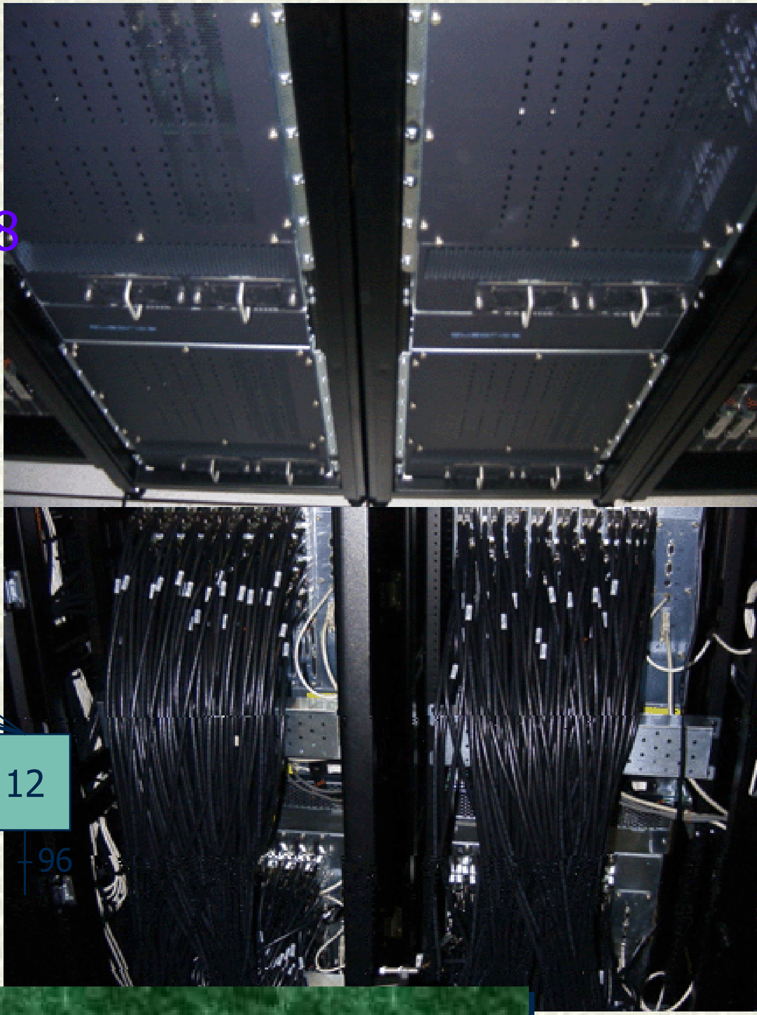
System B:F = 130.6 GB/s / 11.059 TF/s = 0.0118

Node : System = 6.0

Requires 16 QsNet Elan3 128-way switches



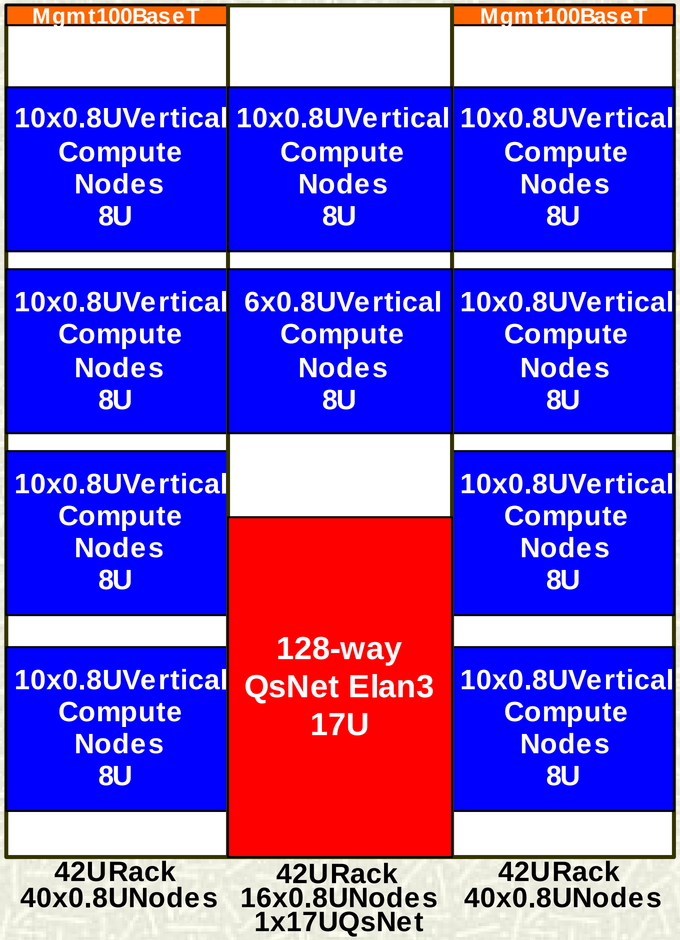
1152 Elan3 ports, 680 MB/s (bi-directional) each



This scales up to 14x96 or 1,344 ports

96x0.8U Compute node scalable unit with first level QsNet Elan3

96 compute nodes and 1 first level switch



Mgmt100BaseT

10x0.8UVertical Compute Nodes 8U

10x0.8UVertical Compute Nodes 8U

10x0.8UVertical Compute Nodes 8U

10x0.8UVertical Compute Nodes 8U

42URack
40x0.8UNodes

10x0.8UVertical Compute Nodes 8U

6x0.8UVertical Compute Nodes 8U

128-way QsNet Elan3 17U

42URack
16x0.8UNodes
1x17UQsNet

Mgmt100BaseT

10x0.8UVertical Compute Nodes 8U

10x0.8UVertical Compute Nodes 8U

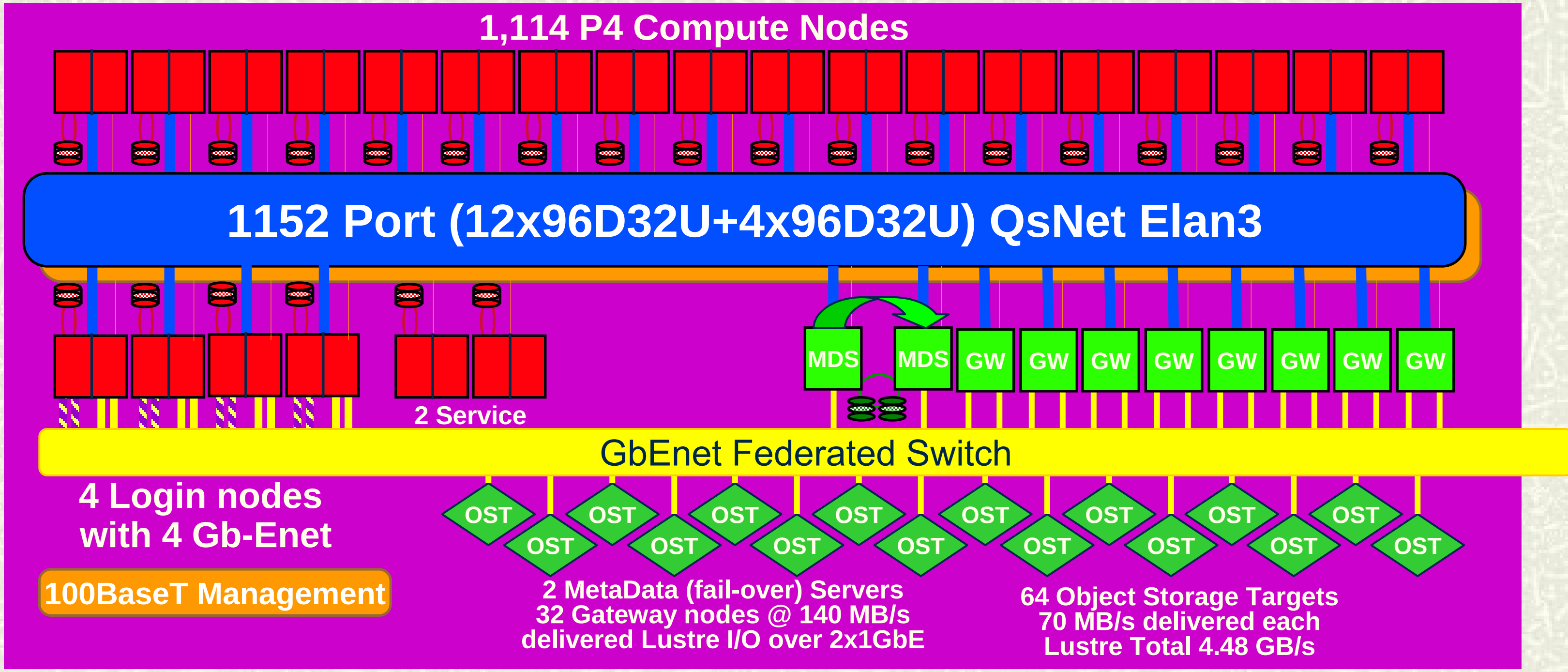
10x0.8UVertical Compute Nodes 8U

10x0.8UVertical Compute Nodes 8U

42URack
40x0.8UNodes

MCR System Architecture for 1,152 nodes, 11.1 TF/s peak[†]

[†]Cluster wide file system leverages DOE/NNSA ASCI PathForward Open Source Lustre development

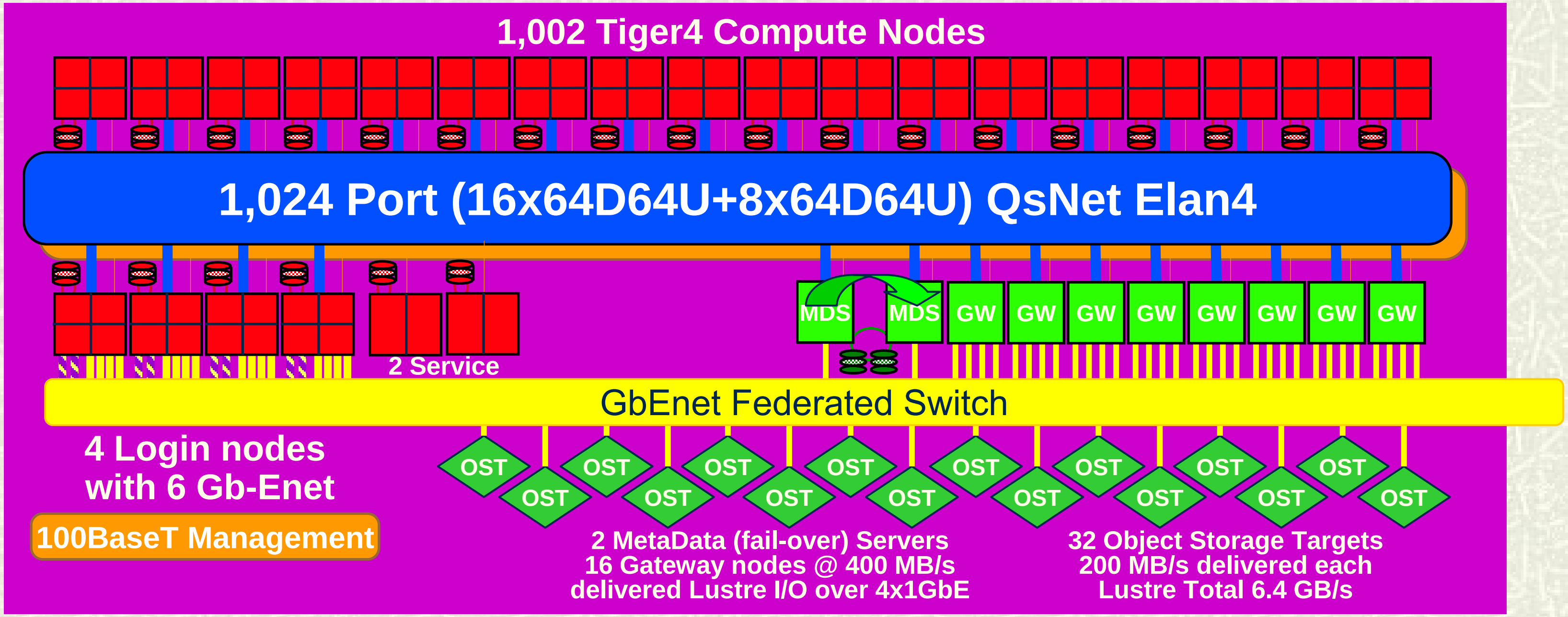

1,114 P4 Compute Nodes
1152 Port (12x96D32U+4x96D32U) QsNet Elan3
MDS
MDS
GW
GW
GW
GW
GW
GW
GW
GW
2 Service
GbEnet Federated Switch
4 Login nodes with 4 Gb-Enet
100BaseT Management
OST
2 MetaData (fail-over) Servers
32 Gateway nodes @ 140 MB/s delivered Lustre I/O over 2x1GbE
64 Object Storage Targets
70 MB/s delivered each
Lustre Total 4.48 GB/s

System Parameters

- Dual 2.4 GHz Pentium 4 Prestonia nodes with 4.0 GB PC2100 DDR SDRAM
- Aggregate 11.1 TF/s peak, 4.608 GiB memory
- <5 μs, 320 (400) MB/s MPI latency and Bandwidth over QsNet
- Lustre cluster wide file system with 4.48 GB/s delivered bandwidth
- Support 120 MB/s transfers to Archive over dual Jumbo Frame Gb-Enet and QSW links from each Login node
- 115.2 TB in local disk in 120 GB/node ATA100 disk
- 13 B:F = 110 TB global parallel file system in multiple RAID5
- Complete remote management of consolidated consoles and LinuxBIOS boot

October 30, 2002 - Achieved 7.634 TF/s on MPLinpack (69% of peak)

Thunder System Architecture for 1,024 nodes, 23 TF/s peak



1,002 Tiger4 Compute Nodes

1,024 Port (16x64D64U+8x64D64U) QsNet Elan4

4 Login nodes
with 6 Gb-Enet

100BaseT Management

2 MetaData (fail-over) Servers
16 Gateway nodes @ 400 MB/s
delivered Lustre I/O over 4x1GbE

32 Object Storage Targets
200 MB/s delivered each
Lustre Total 6.4 GB/s

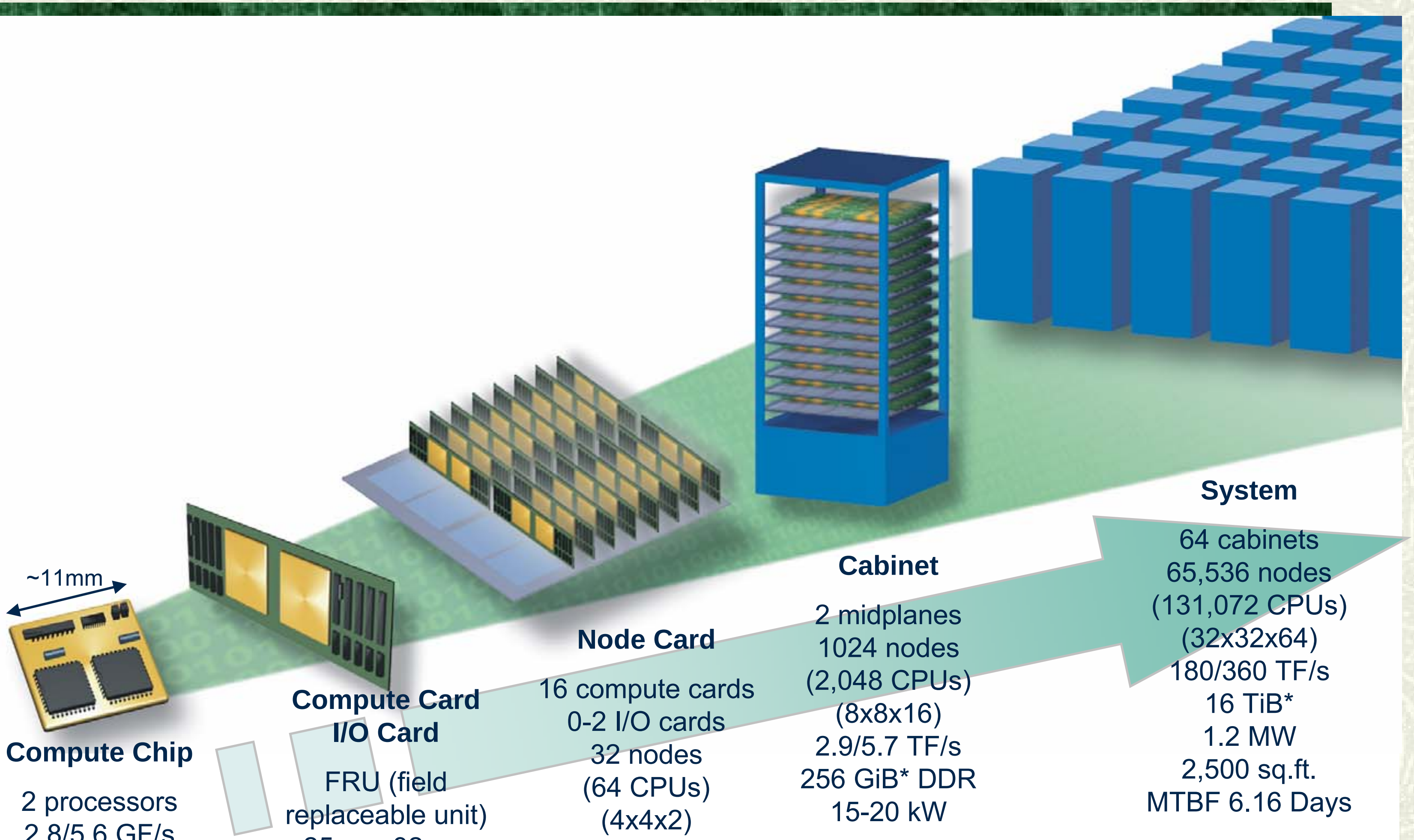
System Parameters

- Quad 1.4 GHz Itanium2 Madison Tiger4 nodes with 8.0 GB DDR266 SDRAM
- <3 µs, 900 MB/s MPI latency and Bandwidth over QsNet Elan4
- Support 400 MB/s transfers to Archive over quad Jumbo Frame Gb-Enet and QSW links from each Login node
- 75 TB in local disk in 73 GB/node UltraSCSI320 disk
- 50 MB/s POSIX serial I/O to any file system
- 8.7 B:F = 192 TB global parallel file system in multiple RAID5
- Lustre file system with 6.4 GB/s delivered parallel I/O performance
  - •MPI I/O based performance with a large sweet spot
  - •32 < MPI tasks < 4,096
- Software RHEL 3.0, CHAOS, SLURM/DPCS, MPICH2, TotalView, Intel and GNU Fortran, C and C++ compilers

Thunder build will be complete in Jan 2004
Contracts with

- **California Digital Corp for nodes and integration**
- **Quadrics for Elan4**
- **Data Direct Networks for global file system**
- **Cluster File System for Lustre support**

BlueGene/L is scaled up with a few unique components and IBM’s system on a chip technology developed for the embedded marketplace



Compute Chip

2 processors
2.8/5.6 GF/s
4 MiB* eDRAM

(compare this with a 1988 Cray YMP/8 at 2.7 GF/s)

Compute Card
I/O Card

FRU (field replaceable unit)
25mmx32mm
2 nodes (4 CPUs)
(2x1x1)
2.8/5.6 GF/s
256/512 MiB* DDR
15 W

Node Card

16 compute cards
0-2 I/O cards
32 nodes
(64 CPUs)
(4x4x2)
90/180 GF/s
8 GiB* DDR

Cabinet

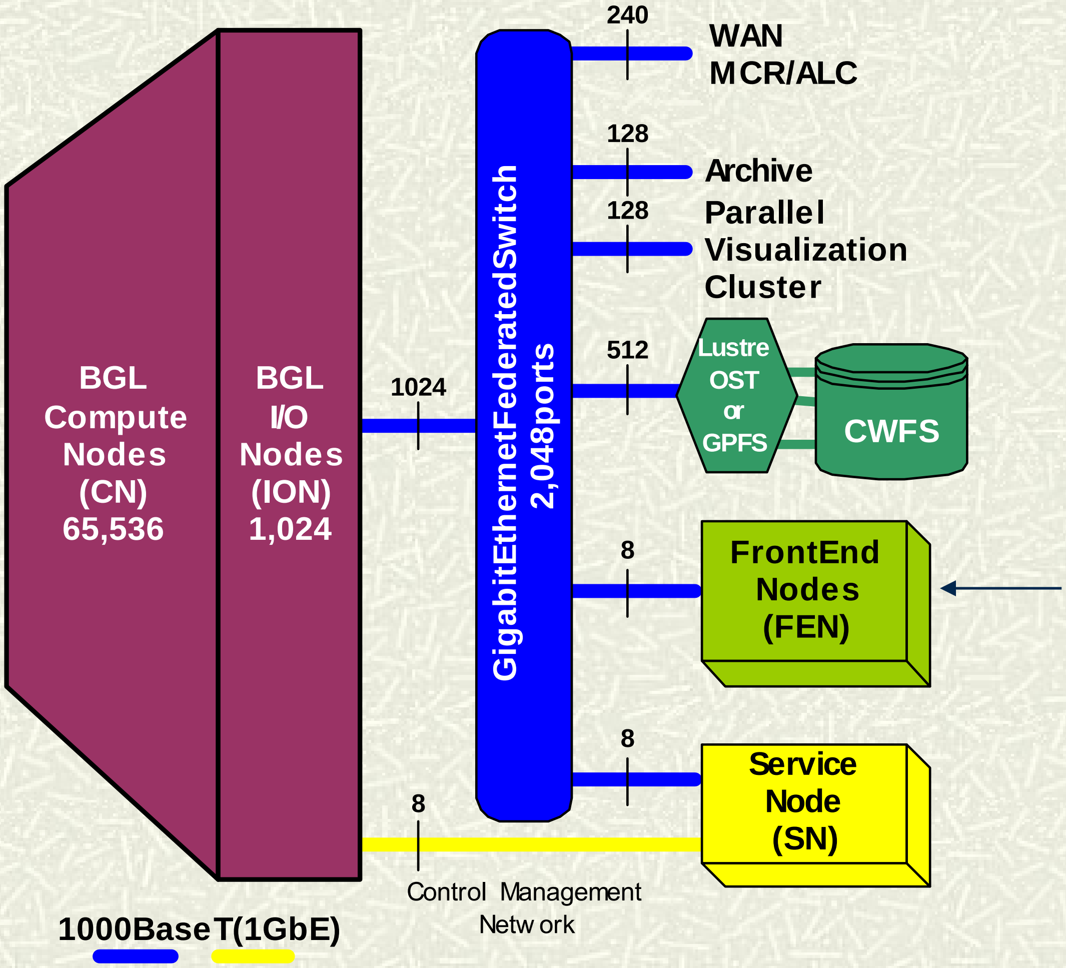
2 midplanes
1024 nodes
(2,048 CPUs)
(8x8x16)
2.9/5.7 TF/s
256 GiB* DDR
15-20 kW

System

64 cabinets
65,536 nodes
(131,072 CPUs)
(32x32x64)
180/360 TF/s
16 TiB*
1.2 MW
2,500 sq.ft.
MTBF 6.16 Days

* http://physics.nist.gov/cuu/Units/binary.html

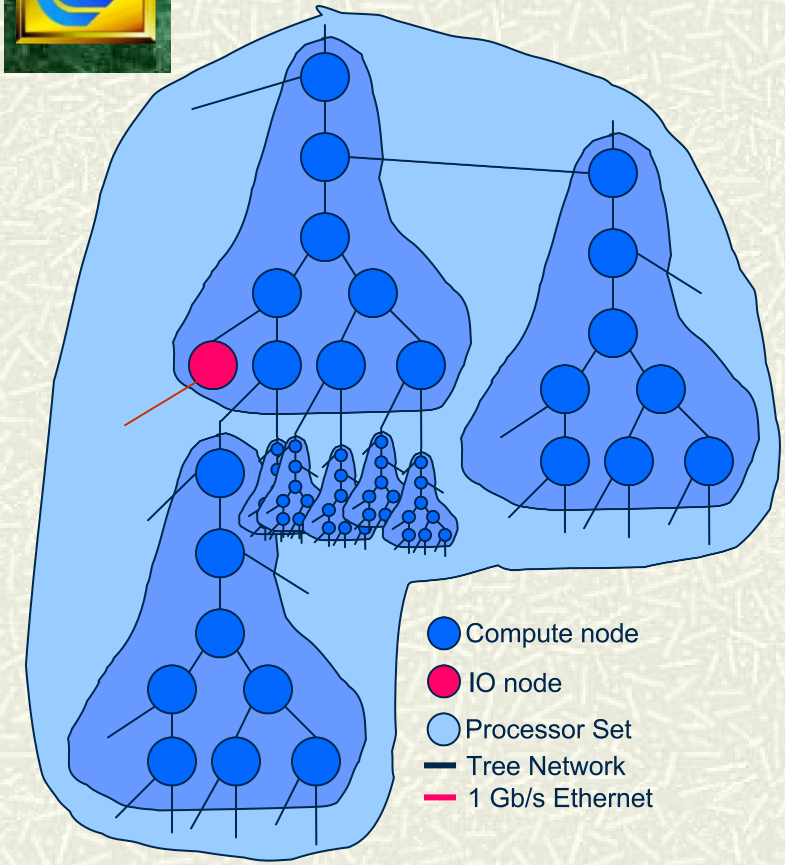
BlueGene/L IO architecture presents opportunity for highly integrated simulation environment



user logs in here

Integration strategy is to utilize BlueGene/L as fulcrum for OCF (Open Computing Facility) Enterprise Wide File System

File I/O communication path between I/O node and its 64 compute nodes is the tree network



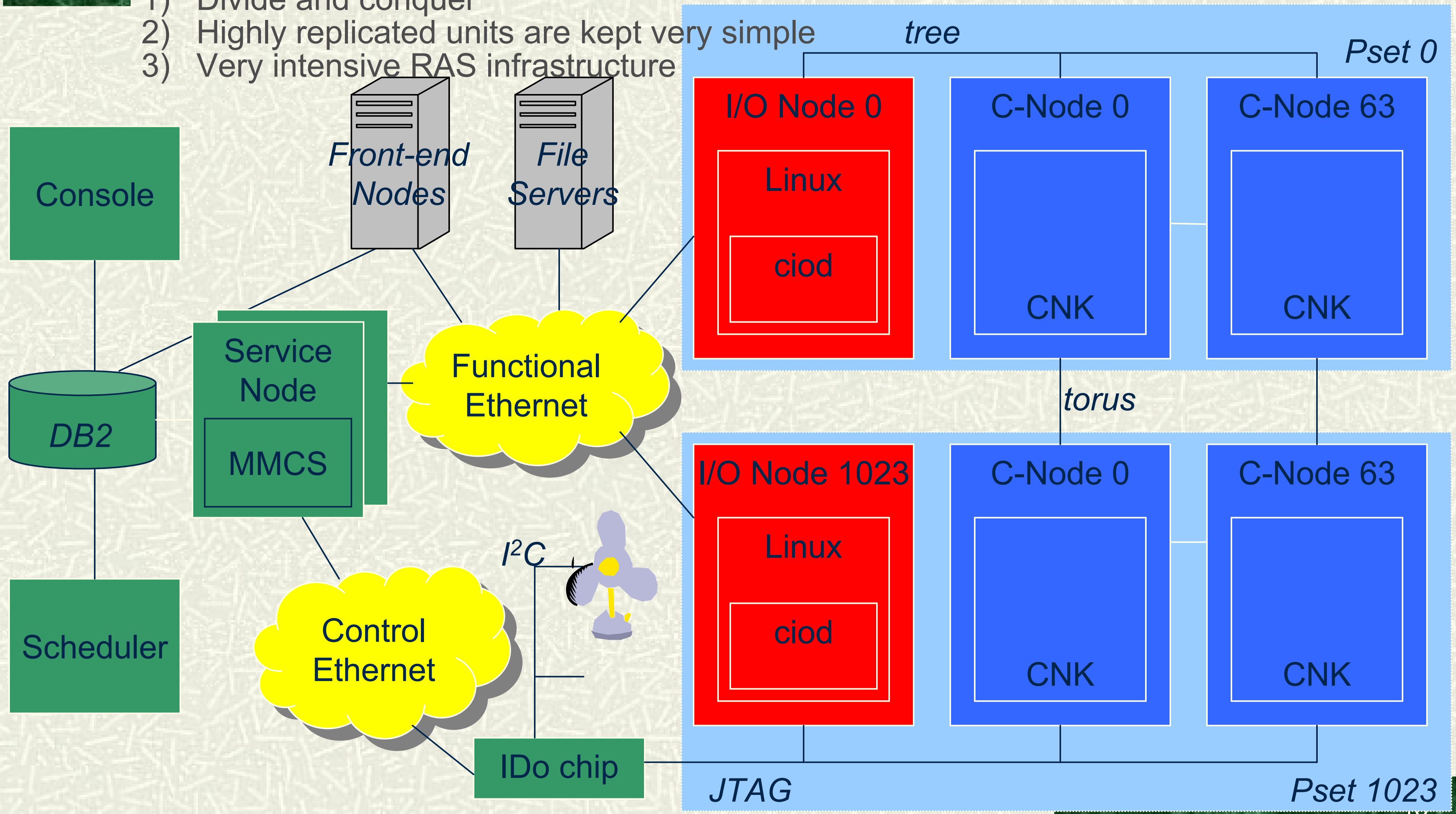
- IO sub-tree consists of 8 compute nodes
- BGL has 8 IO sub-trees per I/O node

- Tree link is 2b+2b @ 1.4GHz = 350+350MB/s
  - 350 MB/s broadcast from IO node to 64 compute nodes
  - 5.5+5.5 MB/s effective simultaneous point to point bandwidth/compute node
- I/O link to file system is 1 Gb/s Ethernet
  - Achievable bandwidth below about 90+90 MB/s
- Aggregate tree BW through a node is 2.1 GB/s
  - Achievable with multiple global combine operations
- Arithmetic operations implemented in tree
  - Integer/FP max/min
  - Integer add/subtract, bitwise logical ops
- Latency of tree less than 2.5 μs to top, additional 2.5 μs to broadcast to all
- Global sum over 64K in less than 2.5 μs (to top of tree)
- Partitioned with Torus boundaries
- Flexible local routing table

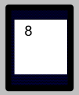
Blue Gene/L is really a 1,024 node Linux cluster with 64 floating point accelerators

Three key scalable software design ideas:

1) Divide and conquer
2) Highly replicated units are kept very simple
3) Very intensive RAS infrastructure



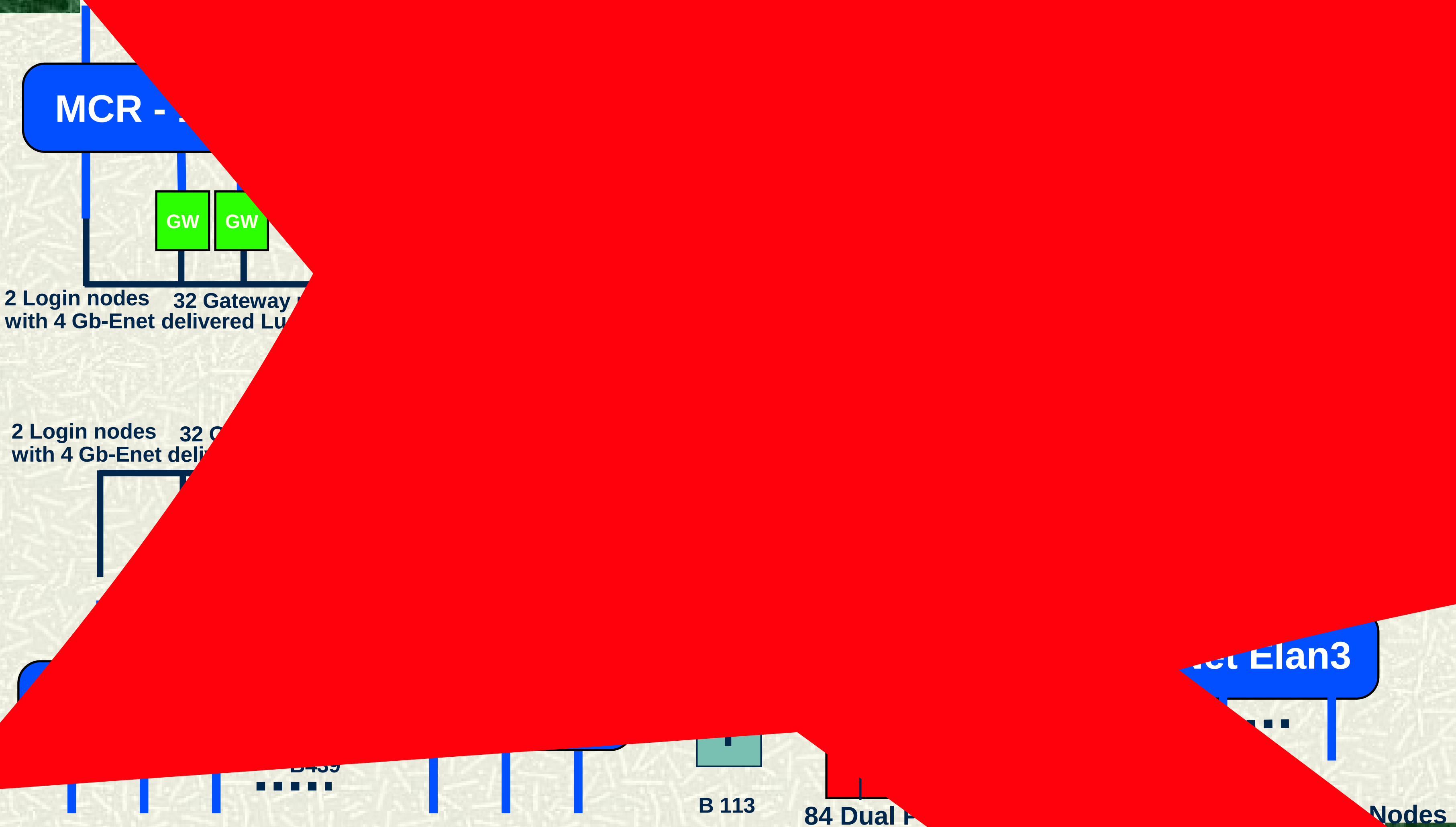
Federated 1 Gb Ethernet switching infrastructure requires distributed network between multiple building



Assumptions:

1) Data flows thru this federated switch are primarily to/from Lustre.
2) BGL in B453, Thunder in B451 and MCR, ALC in B439

[redacted]



MCR -

GW

GW

2 Login nodes
with 4 Gb-Enet delivered

32 Gateway

2 Login nodes
with 4 Gb-Enet

32

Elan3

B439

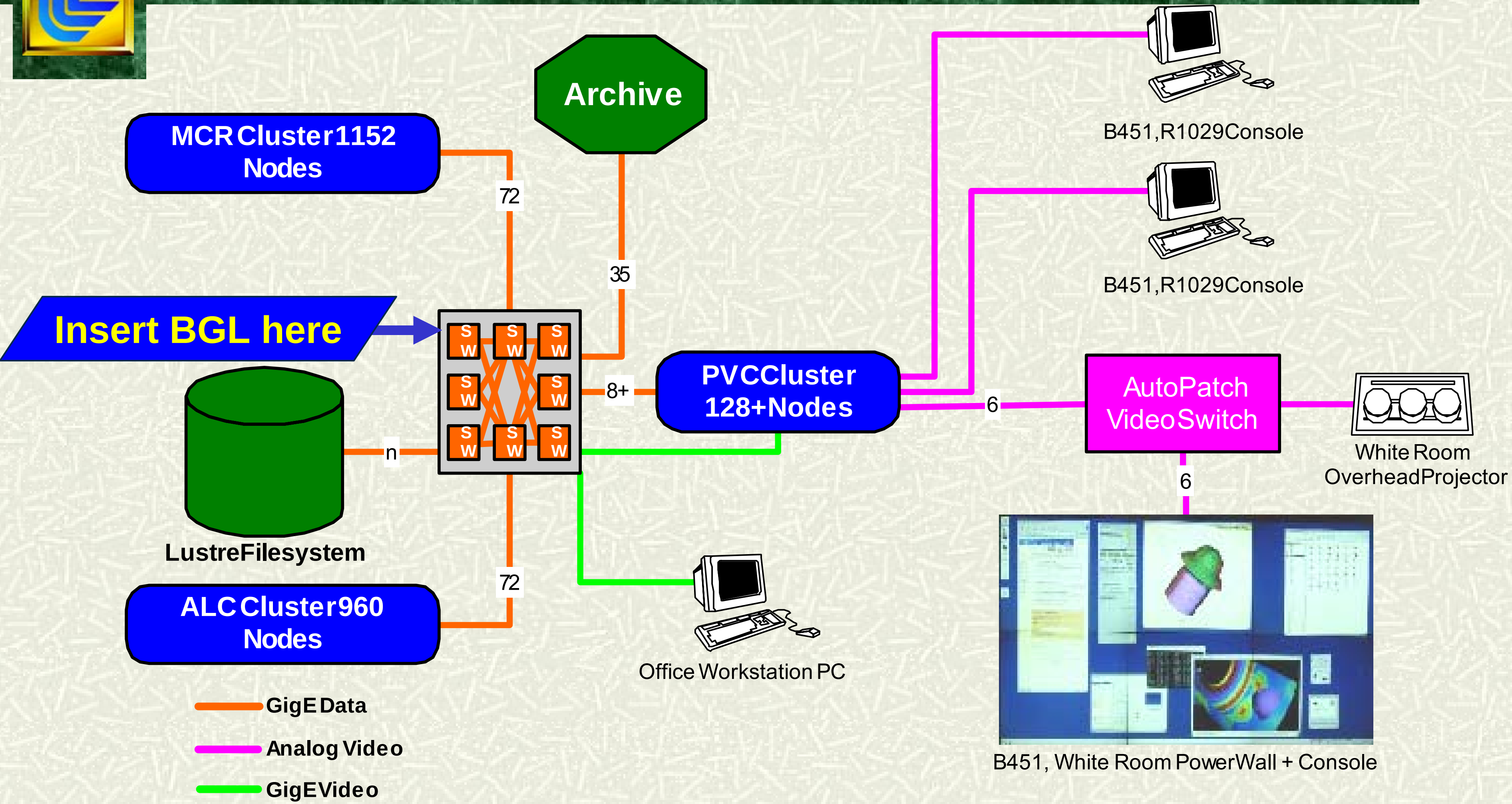
B 113

84 Dual

Nodes

924 Dual P4 Compute Nodes

Next generation of parallel visualization technology will cleanly integrate into BGL simulation environment



Archive

MCR Cluster 1152 Nodes

72

35

Insert BGL here

SW SW SW
SW SW
SW SW SW

8+

n

PVC Cluster 128+ Nodes

Lustre Filesystem

72

ALC Cluster 960 Nodes

Office Workstation PC

B451, R1029 Console

B451, R1029 Console

6

AutoPatch Video Switch

White Room Overhead Projector

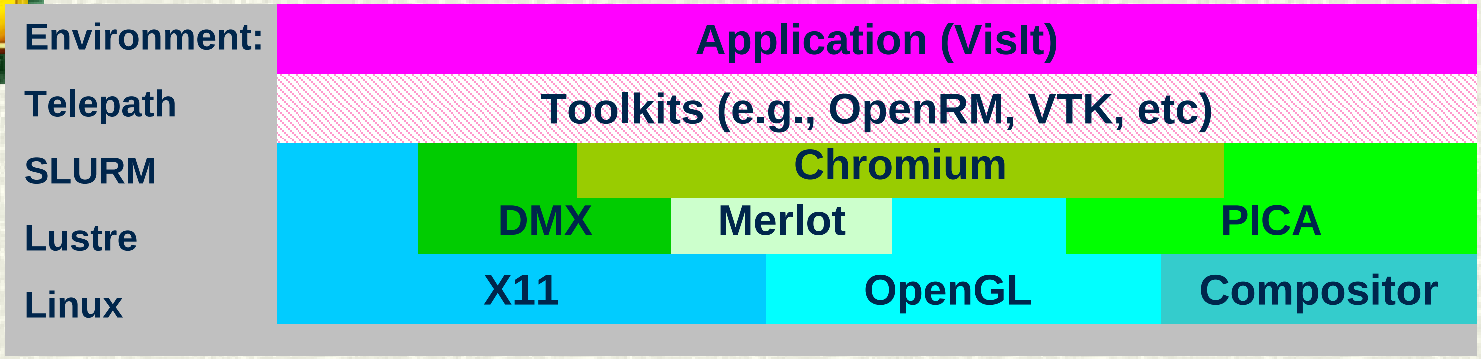
6

B451, White Room PowerWall + Console

- GigE Data
- Analog Video
- GigE Video

All components of parallel visualization solution are built on commodity technology and are scalable

LLNL visualization based on ASCI tri-Laboratory distributed, parallel rendering software stack



Environment:
Telepath
SLURM
Lustre
Linux

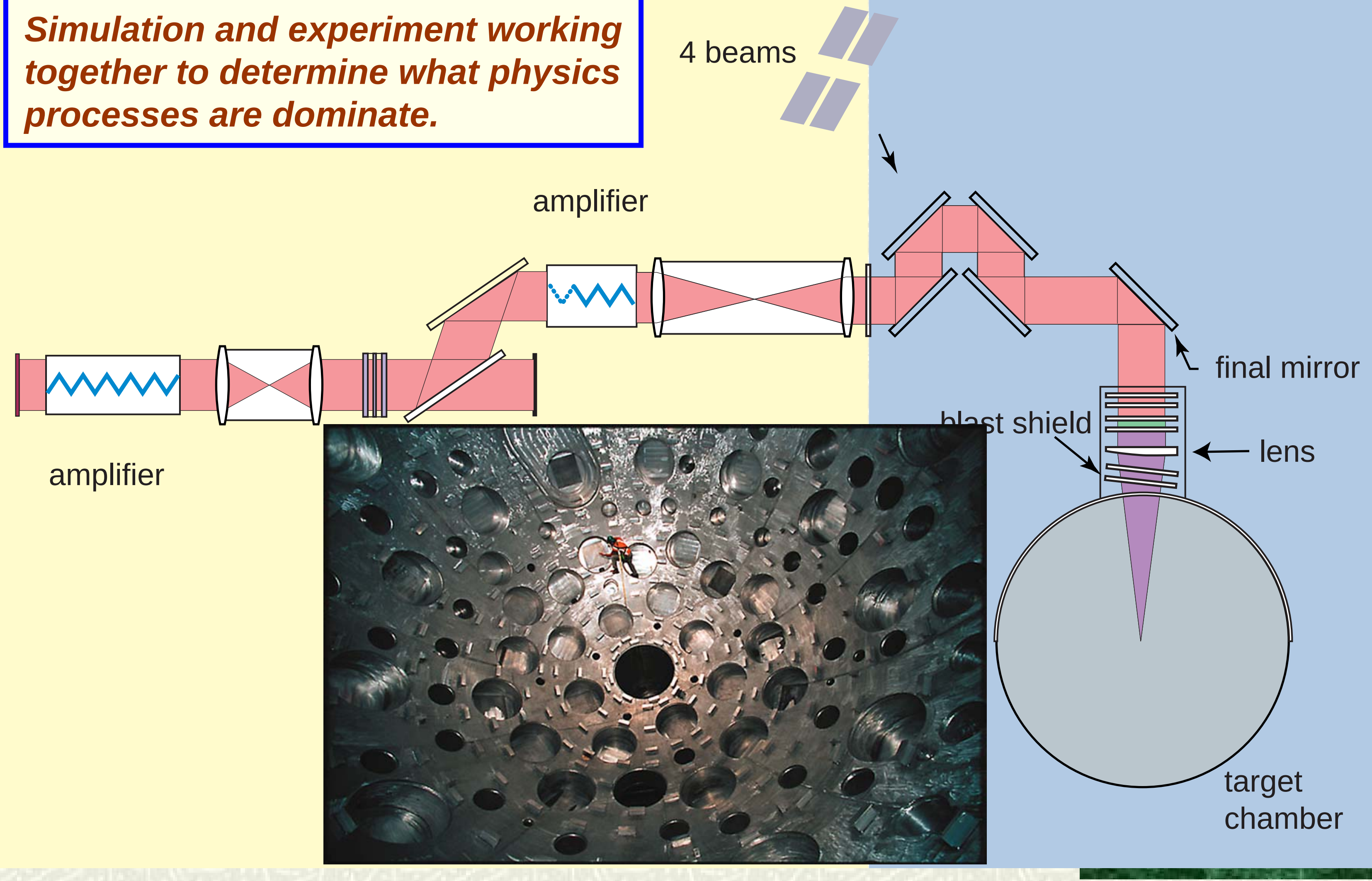
Application (VisIt)
Toolkits (e.g., OpenRM, VTK, etc)
Chromium
DMX
Merlot
PICA
X11
OpenGL
Compositor

- Multi-year Open Source development effort, delivering a distributed graphics API stack. The backbone of a portable, cluster based scalable rendering system:
  - Tiled X11 services
  - OpenGL
  - Visualization scheduling/display control
  - Applications
- Based on core local services
  - Unix/X11/OpenGL/Scheduling
  - Other HW: compositors, distance, etc
- Notable in the release are the Chromium distributed parallel OpenGL API and the DMX distributed multi-headed X11 server.

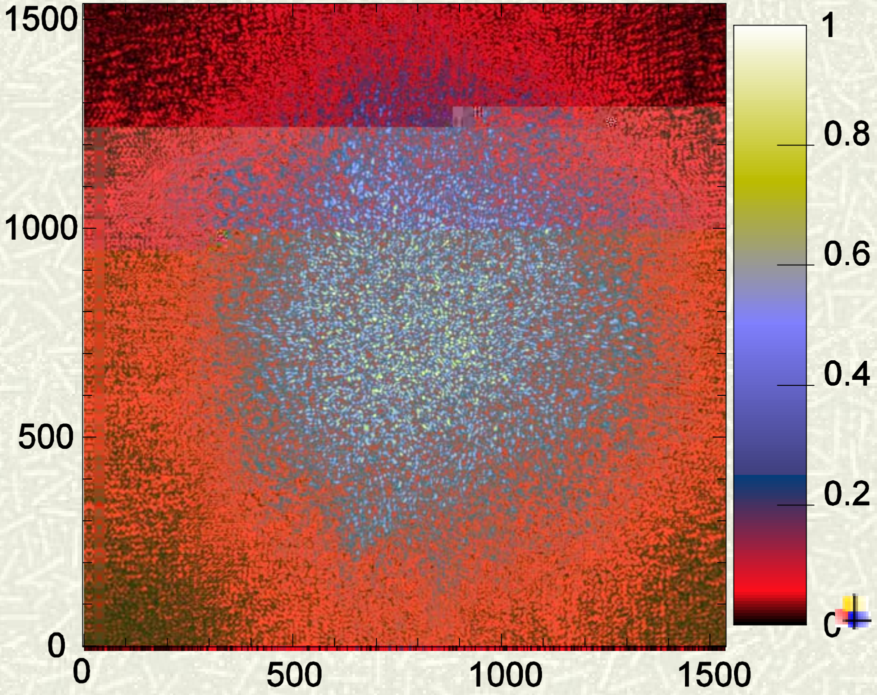
- Infrastructure
  - Telepath – Scheduling & display routing
  - VTK – Distributed visualization algorithms
  - Chromium – Parallel OpenGL rendering
  - DMX – 2D distributed display scaling
  - PICA – Parallel compositing model
  - Merlot – Image delivery infrastructure
- Applications
  - VisIt, ParaView – Full featured viz apps
  - TeraScale Browser – Specialized viz apps
  - MIDAS – Merlot based remote rendering
  - Blockbuster – Tiled animation display

MCR simulations used to design NIF early light diagnostics

Simulation and experiment working together to determine what physics processes are dominate.



Details of laser beam is of great interest to NIF program as it determines efficiency and backscatter

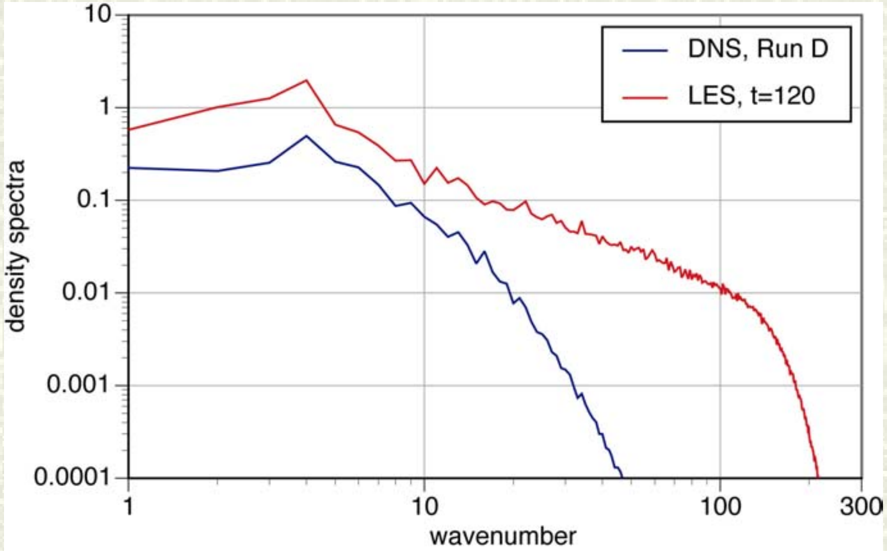


- PF3D performance
  - An MCR processor is twice as fast as an ASCI Q processor for simulations that only include filamentation.
  - A processor on Q is about the same speed as an MCR processor for simulations that include SRS and SBS.
  - A processor on Q is 1.5-2 times faster than an ASCI white processor for most pf3d simulations.
- Simulation
  - The grid had 1536x1536x2880 = 6.8B zones
  - 1,920 Pentium 4 processors of the MCR cluster
  - Ran for 10 days to simulate 35 ps (two light crossing times).

What was learned from this simulation?

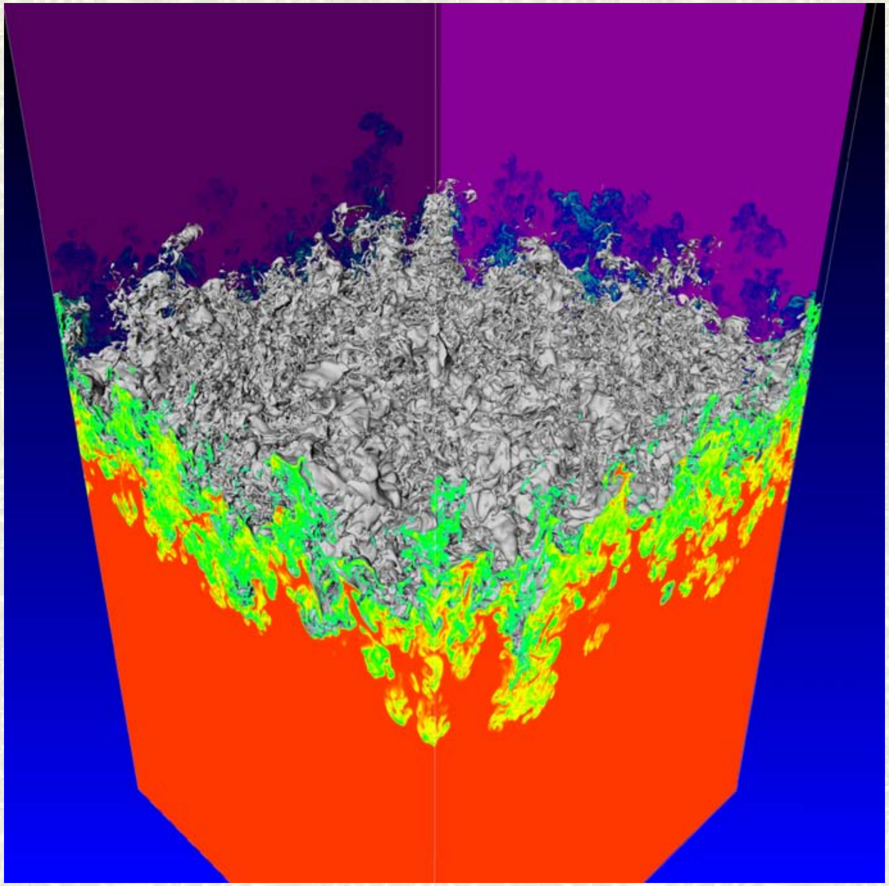
- Interactions between the laser beam and the plasma control the flow of laser energy
- The nominal laser intensity is 10^{15} W/cm^2.
- The beam is passed through a continuous phase plate to produce "speckles" with intensities up to 10^{16} W/cm^2.
- The most intense parts of the beam self-focus to even higher intensity and then break up and "spray" into several directions.
- Roughly 60% of the laser energy is absorbed by the plasma.
- Laser light scatters off fluctuations in the electron density in a process called Stimulated Raman Scattering (SRS). Roughly 25% of the light is back-scattered by SRS in this simulation.
- Laser light scatters off ion density fluctuations in a process called Stimulated Brillouin Scattering (SBS). SBS is weak in this simulation.
- Roughly 17% of the laser energy exits at the back of the simulation volume.

What is Rayleigh-Taylor Instability?
(and why do large numerical simulations of it?)



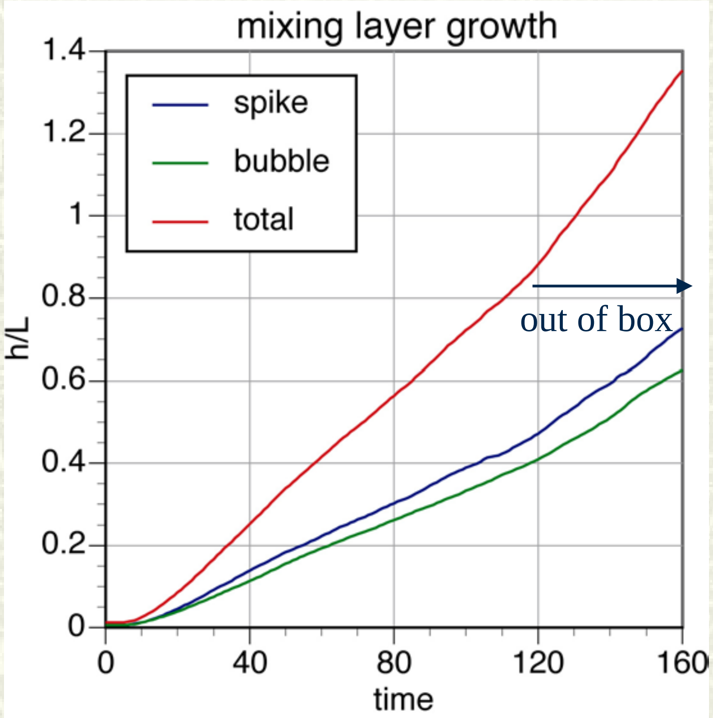
Simulation stats

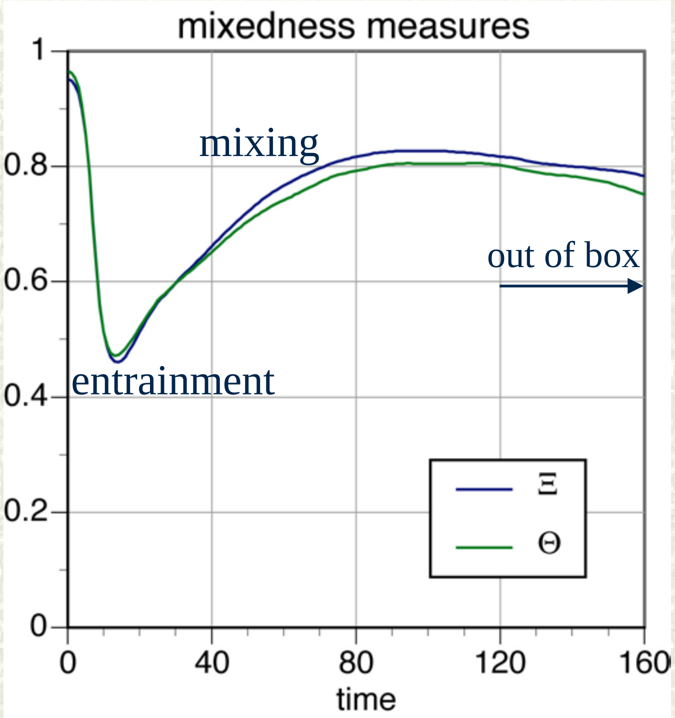
- The *Miranda* code uses 10th-order compact (spectral-like) scheme in all directions to compute derivatives and FFTs for the Poisson solve, which requires lots of global MPI communication for transposes (MPI_ALLTOALL)
- The very large number of CPUs were used to optimize *Miranda*'s I/O efficiency
- I/O model: each CPU wrote/read its own restart and graphics data (lots of files!)
- We ran a case with 720 x 720 x 1620 grid points using 810 nodes (1620 CPUs)
- The simulation performed 44,500 cycles in 21 CPU days of "quality time" (27 days real time, 24/7)
- A total of 37 TB of graphics and restart data were written
- 811 graphics dumps > 1.3 million files (21 TB) were generated and archived
- Dealing with the large amount of data storage needed to make movies presented a challenge



Simulation results: Growth rate of mixing layer

- Growth rate of mixing layer: more linear than quadratic, consistent with some recent “α-group” simulations
- Loss of horizontal resolution coincides with more rapid growth, less mixedness





The Aftermath ...

- Unanswered science questions:
  - What mechanisms control linear v. quadratic growth?
  - What's the sensitivity to initial conditions?
  - What's the high Re limit? Is there a mixing transition?
- A new, even larger run is under way with 1152^3 grid points and different (small-scale, narrow-band) initial conditions

… and a Wish List

- More, more, more … The bigger the better for turbulence simulations (and there's a long way to go)
- Improved data storage infrastructure
- There needs to be large, state-of-the-art ASCI machines dedicated to doing *BIG* science runs, instead of being allocated in small pieces to the maximum number of users

MDCASK Code and Parallelization

- The MDCASK code solves Newton's equations of motion (a set of coupled ordinary differential equations) for a large number of atoms interacting trough a given Embedded Atom Method (EAM) interatomic potential. It uses a 4th order predictor corrector solver, with linked cells and a neighbor list within each cell.
- MDCASK has about 16,000 lines of Fortran code (94%) and 1000 lines of C code (6%, used for efficient IO). This is an streamlined version of MDCASK optimized for this MCR run, which can only deal with metals and single component systems. A much longer versions exist to tackle semiconductors and organics, together with metals.
- The parallelization is simple domain decomposition with MPI utilized between domains. MPI is used to communicate information along the boundary or "skin region" of each cell to its direct neighbors at each time-step.
- MDCASK achieves Y% of single CPU peak (a-b GF/s) on MCR with Z% parallel speed up for N CPUs (N/2 nodes).
- Expect single CPU performance to improve to 25-35% peak

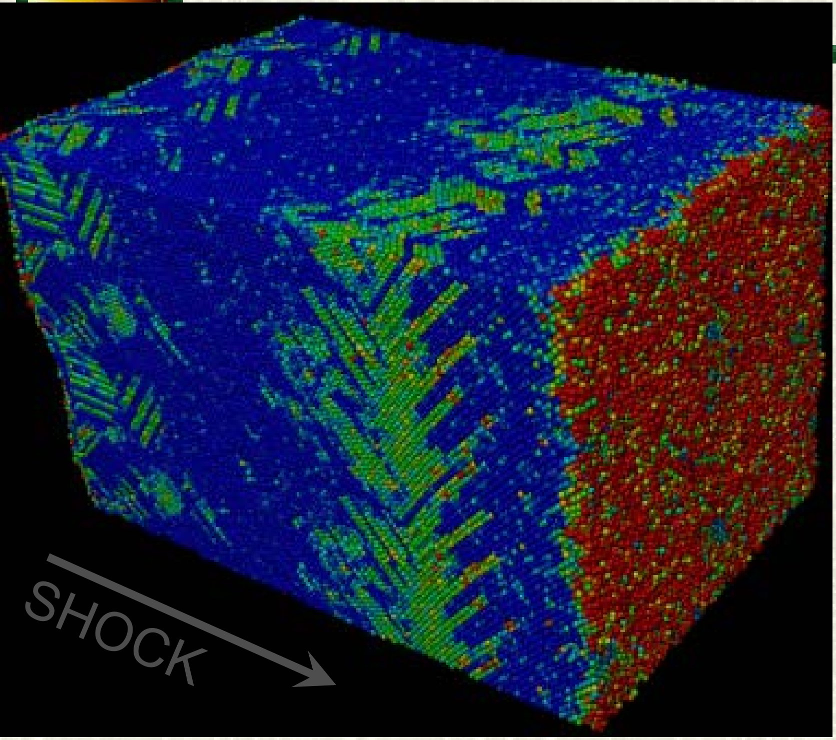
MDCASK Run Details

- Uniform grid and processor layout for the calculations
  - Utilized 400 nodes (800 CPU's) and 2GB of memory per node for 0.8 TB
  - The decomposition was 32 (CPU's) in the x direction (fastest or vector direction), 5 in the y direction, and 5 in the z direction.
  - This was selected through, minimal inter-process communication, and benchmark tests with different numbers of CPU's.
- 4 independent simulations completed so far. Quality of runs = excellent.
- The wall clock time was 4 hours (per run).
  - The compute time was basically the same since I/O was minimal (110 GB per run).
- The verdict on how MCR worked excellent. There were virtually no problems with MCR.

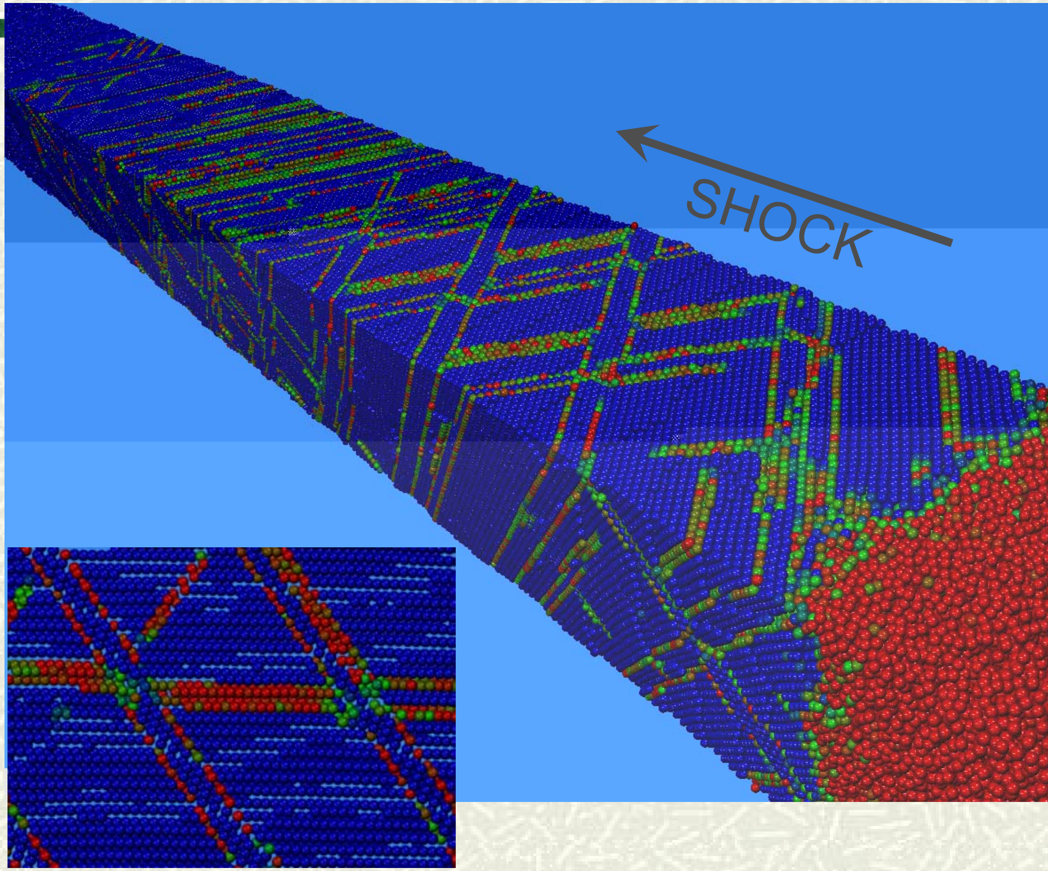
MDCASK Science Run Details

- Simulation cell and processor layout for the planned calculations
  - The sample had 200x200x2200 fcc cells (~350 million atoms). This represents a volume of 723x723x7953 Å^3. Before the shock is applied each atom interacts simultaneously with 54 neighbors, but this number can increase up to 50% depending on the shock strength. Simulated time: 100 ps (~5 10^5 time steps)
  - Plans to utilize 9720 nodes (1944 CPU's) and 1 GB of memory per node for 5 TB of final (compressed) output, during 7 days.
  - The decomposition will be 6x6x54 CPU's in the x,x,z directions. Shock propagates along the z direction.
  - This was selected to achieve minimal inter-process communication and load balance between the nodes.
- Only 2 relatively small test simulations completed so far with 1000 and 128 nodes
- The wall clock time was 10 hours (per run). I/O took ~10% of the run, producing 150 GB/run.
- MCR performed flawlessly. Two unexpected problems were found with version of MDCASK, since simulated scenario was beyond anything done beforehand. Problems are now solved by incorporating new physics and software solutions into MDCASK.

Shocks in single crystal Cu <100>



A 50 Gpa shock has bounced back at the free surface and meets the shock traveling forward. Large dislocation activity can be seen.
Some dislocation activity can also be seen near the piston, at the back of this picture.



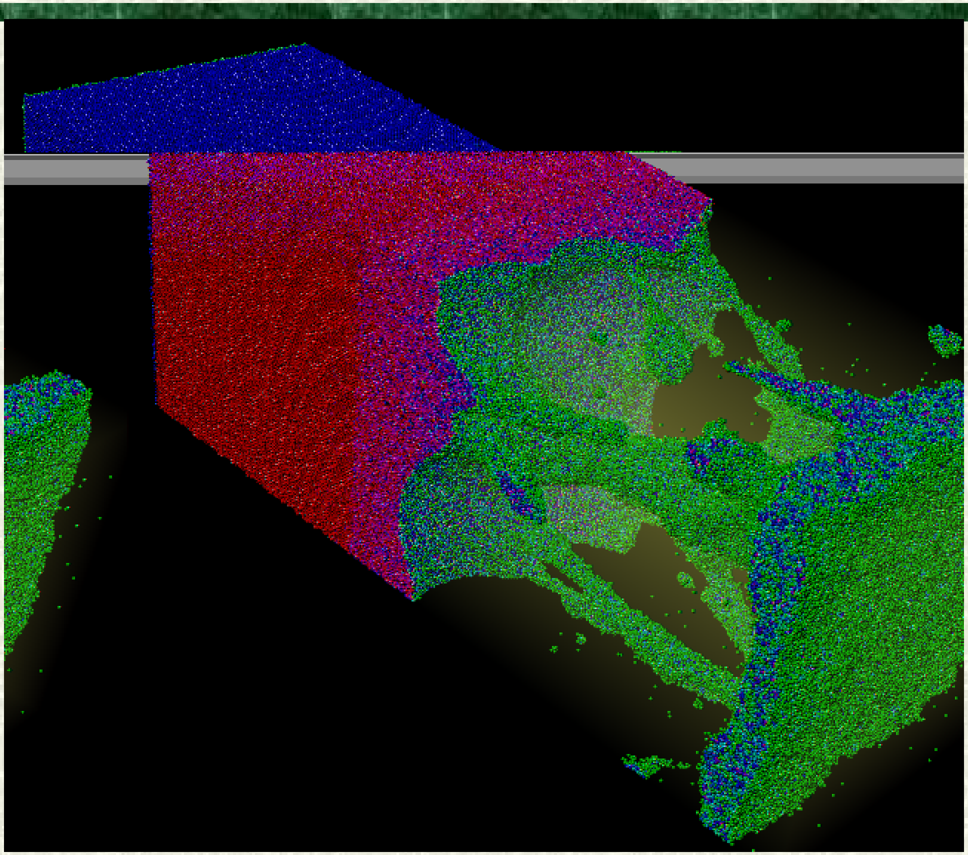
A 40 Gpa shock traversing 0.15 microns can create a pattern of criss-crossing stacking faults.
Micro-twins (a particular type of lattice deformation) are seen in experiments for 40 GPa shocks. In our simulations nano-twins, as the ones in the inset, form at the same pressure.

Colors indicate the “order” in the lattice, with red being very disordered atoms, and blue normal atoms

These shock simulations are directly related to large international experimental and theoretical efforts on NIF (LLNL) and Omega (U. Rochester) lasers.

MOVIES AVAILABLE!

Laser Ablation of Copper



This simulation was done in support of the “Fabrication of Mesoscale Objects” project, The purpose is to develop laser ablation methods for machining NIF targets. The colors indicate the instantaneous potential energies of the Cu atoms; from blue for low energies to red for the highest. In the figure we show ablation near the threshold laser pulse energy for removing significant material The issue is the roughness of the ablated surface. The size of the craters left after ablation is significantly larger and more complex than previous simulations and hence the MCR system is required.

The previous results conclusively demonstrate that 10 teraFLOP/s scale computing is qualitatively different from workstation science

- Dijkstra – A quantitative change is also a qualitative difference, if the quantity has changed by an order of magnitude.
- A Quantitative example
  - A baby crawls at 1 mph
  - A marathoner runs at 10 mph
  - A fast car can drive at 100 mph
  - A fast jet can fly at 1,00 mph
- Qualitative change
  - Driving allows one to go place you cant get to on foot
  - Flying allows one to go to places in time that driving would not allow

Conclusions

- Standing today with multiple 10 teraFLOP/s computational science platform and anticipating about 0.5 petaFLOP/s in 2005
  - We see that simulations are qualitatively different from those done on previous supercomputers
  - We see that this is changing the way science is done
  - Computing is now on-par with theory and experiment in terms of impact on new scientific discoveries!
- It is amazing to think about the impact of the new affordable 10 TF/s scale Beowulf computing
  - Scientists everywhere have already figured this out!
  - This will drive scientific computing and HPTC market
- The key to success is
  - Scalable cluster design, with key component replication
  - Scalable highly integrated multi-cluster simulation environment