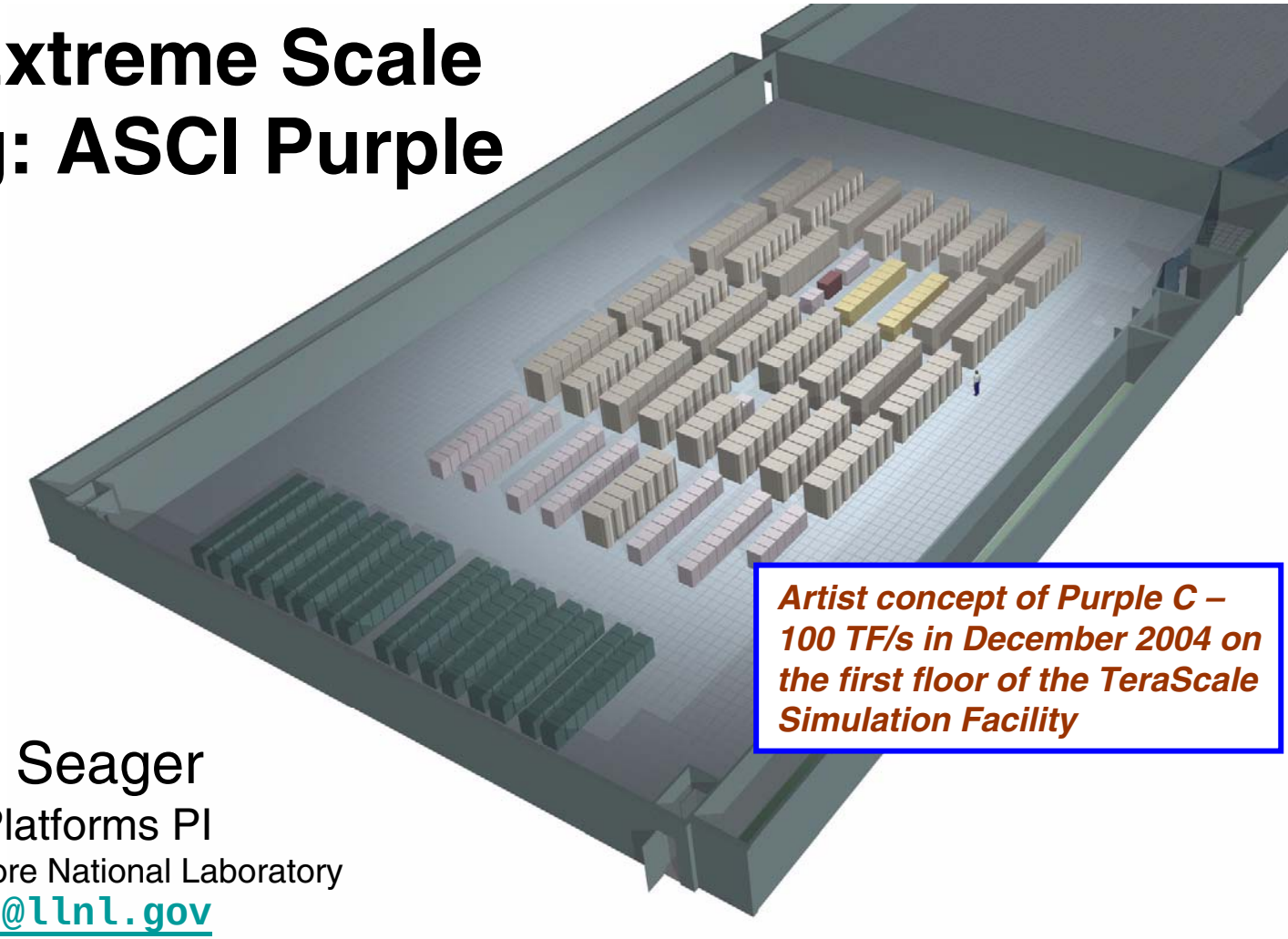
Defining Extreme Scale Computing: ASCI Purple



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

Mark Seager

ASCI Platforms PI

Lawrence Livermore National Laboratory

seager@llnl.gov

925-423-3141

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

Overview

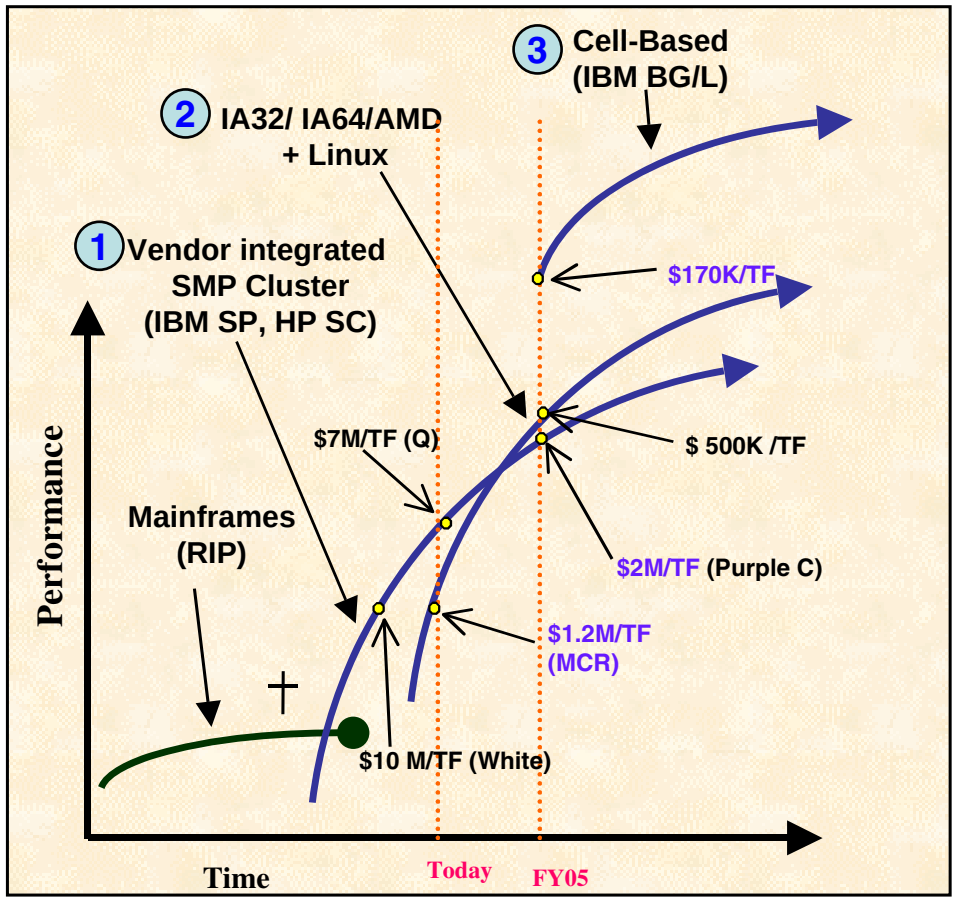
- Platform Strategy
  - Balance risk and benefit to deliver best cost performance to meet programmatic objectives
- Advanced architectures
  - BlueGene/L and friends
- Commodity clusters and OSS
  - Lustre
  - R&D collaboration
- ASCI Purple partnership status

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
   - Capability system for risk tolerant programs
   - Capacity systems for risk averse programs
3. Delivers affordable path to petaFLOP/s
   - Research environment, leading transition to petaflop systems?
   - Are there other paths to a breakthrough regime by 2006-8?

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



Scientific simulation at scale is qualitatively different. We do not yet understand the full implications of this technological development

By being on the leading edge at the 10 teraFLOP/s level with a toehold into the 100 teraFLOP/s → petaFLOP/s computing era, we are fundamentally changing the nature and scope of the scientific method.

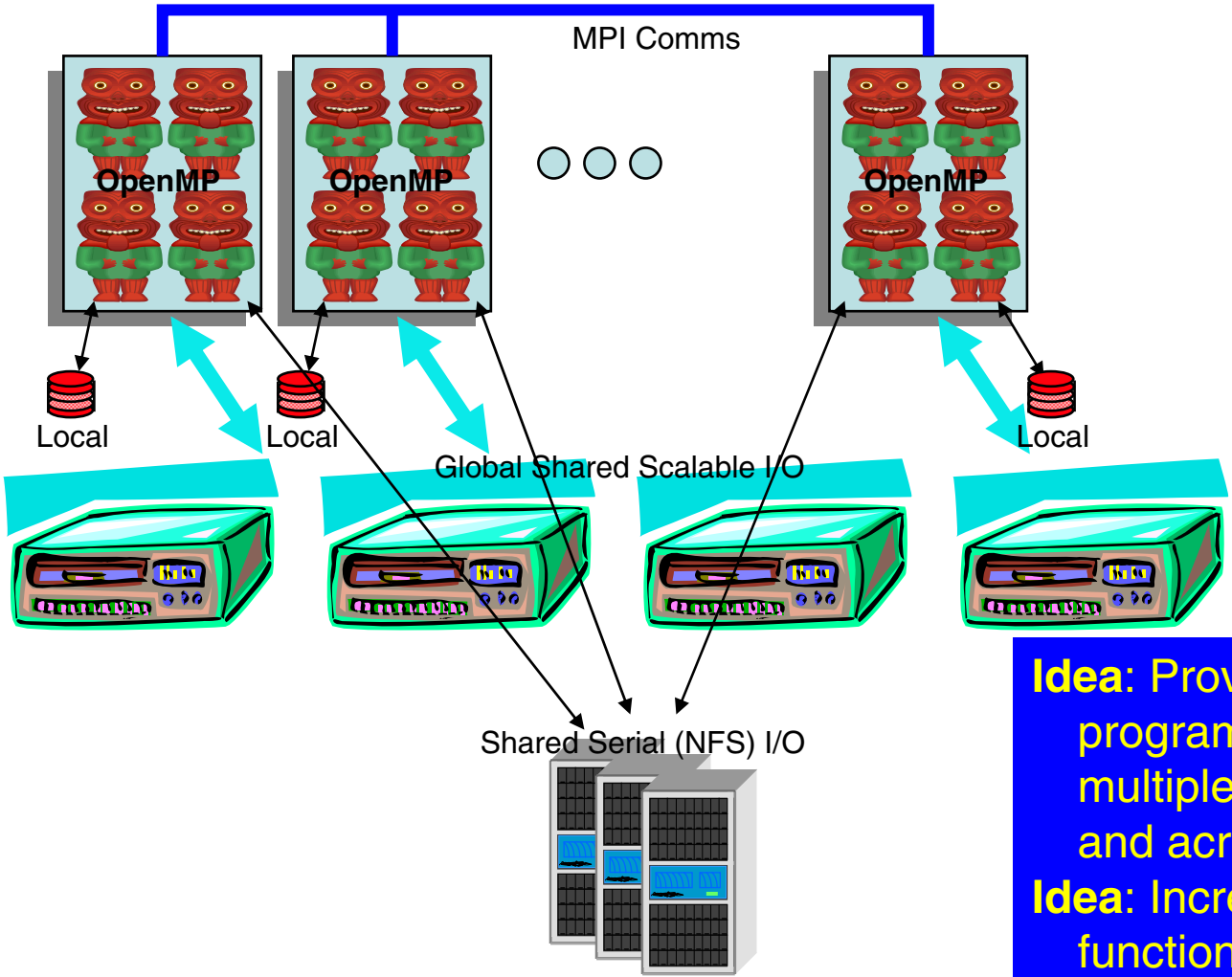
- Edsger Dijkstra: "A quantitative difference is also a qualitative difference, if the quantitative difference is greater than an order of magnitude.

- A quantitative example in transportation
  - 1 Mi/Hr is the speed of a baby crawling
  - 10 Mi/Hr is the speed of a top marathon runner
  - 100 Mi/Hr is the speed of a fast automobile
  - 1,000 Mi/Hr is the speed of a fast jet
- Qualitative ramifications of this transportation example
  - Driving allows people to go to places they could not reach on foot.
  - Flying allows people to go to places they could not reach in time.

Ramifications of this strategy

- Benefits
  - Able to maximize cost performance & adapt quickly to changes
  - Offer options to programmatic customers that match their requirements, not a computing dogma
- Costs
  - Requires expertise in multiple technologies
    - Organization must be capable of simultaneously fielding systems on multiple technology curves
  - Requires constant attention to new technology
    - Must correctly asses
      - Longevity of technology
      - Maturity (risk) of technology
      - Usability of technology
- Issues
  - Programming model and environment must be made as consistent as possible

Provide consistent programming model across scalable platforms

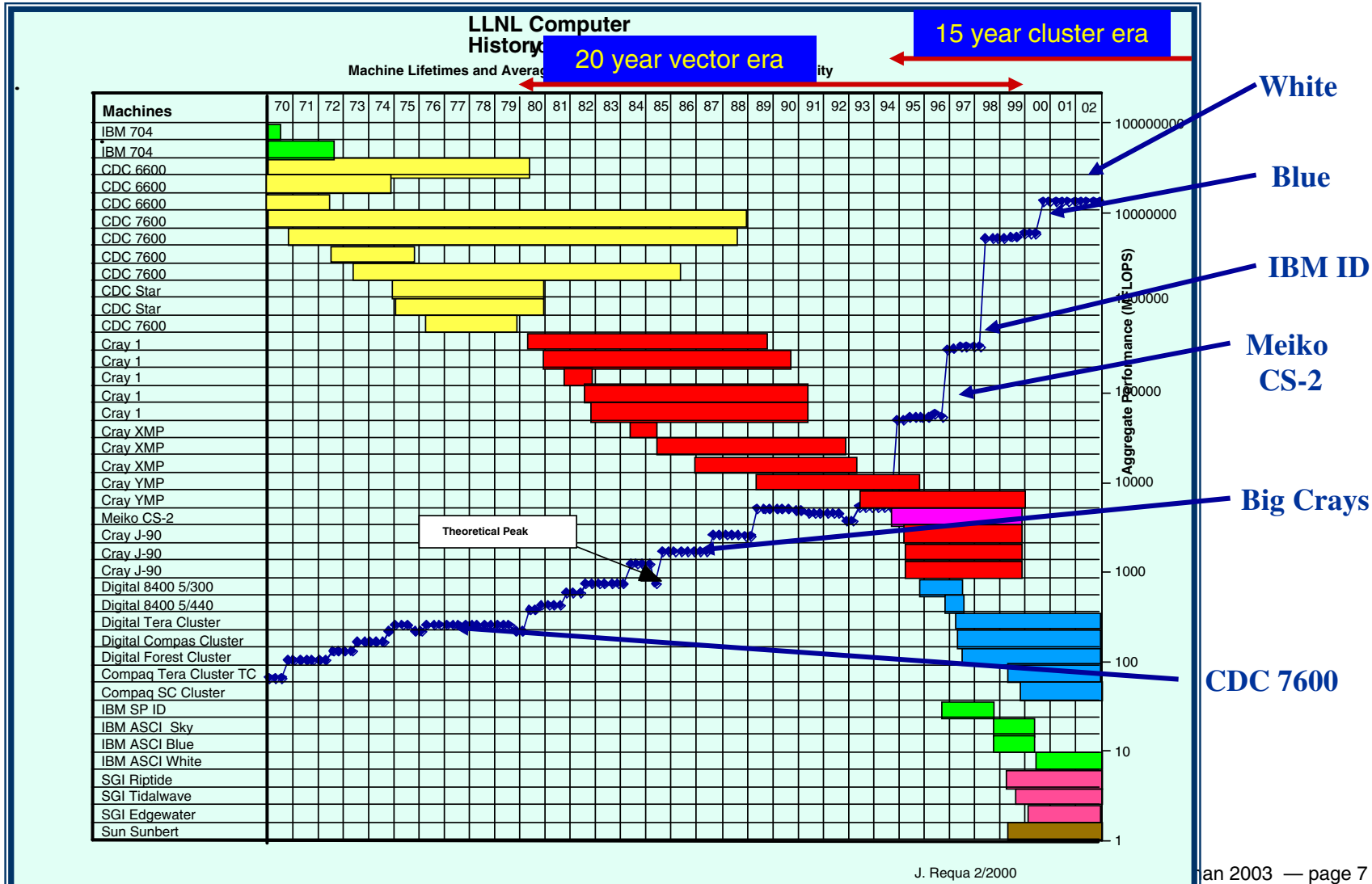


Idea: Provide a consistent programming model for multiple platform generations and across multiple vendors!

Idea: Incrementally increase functionality over time!

LLNL Computer History

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.

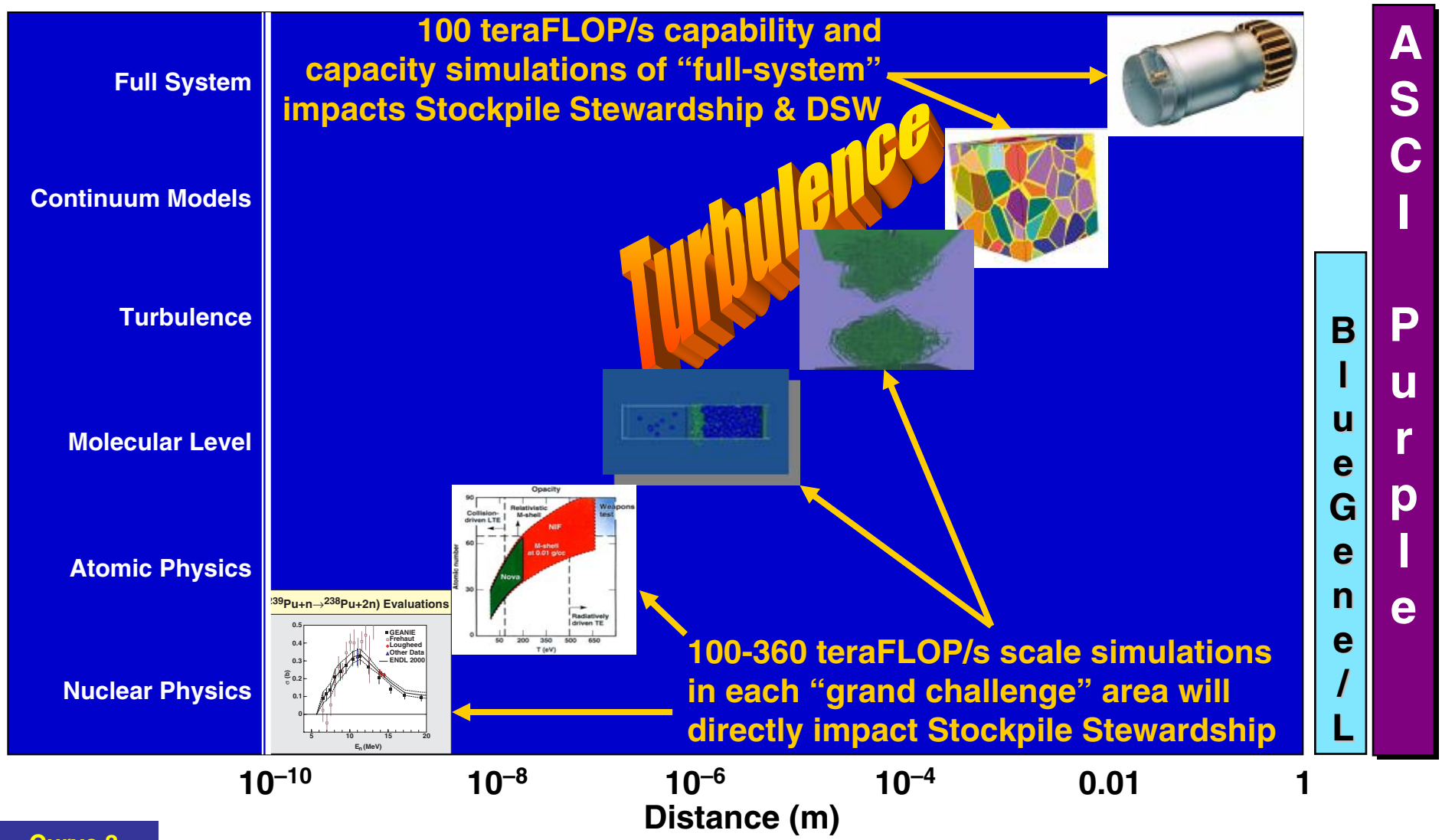


BlueGene/L is an important part of our strategy because the project is aggressively addressing the five critical issues blocking access to petaFLOP/s scale computing

- The five issues
  - Power
  - Floor space
  - Cost
  - Single processor performance
  - Network scalability

- Max Baron, Micro Processor Report, V7A2
  - "During the next few years, the search for energy-efficient computing will become more important than the drive for [single processor] performance; its results will enrich the portable computing experience and keep desktop processors from emitting energy in the visible spectrum."
- And we would add …
  - "and provide an affordable path to petaFLOP/s scale computing…"

Many important physics issues can be addressed by BlueGene/L

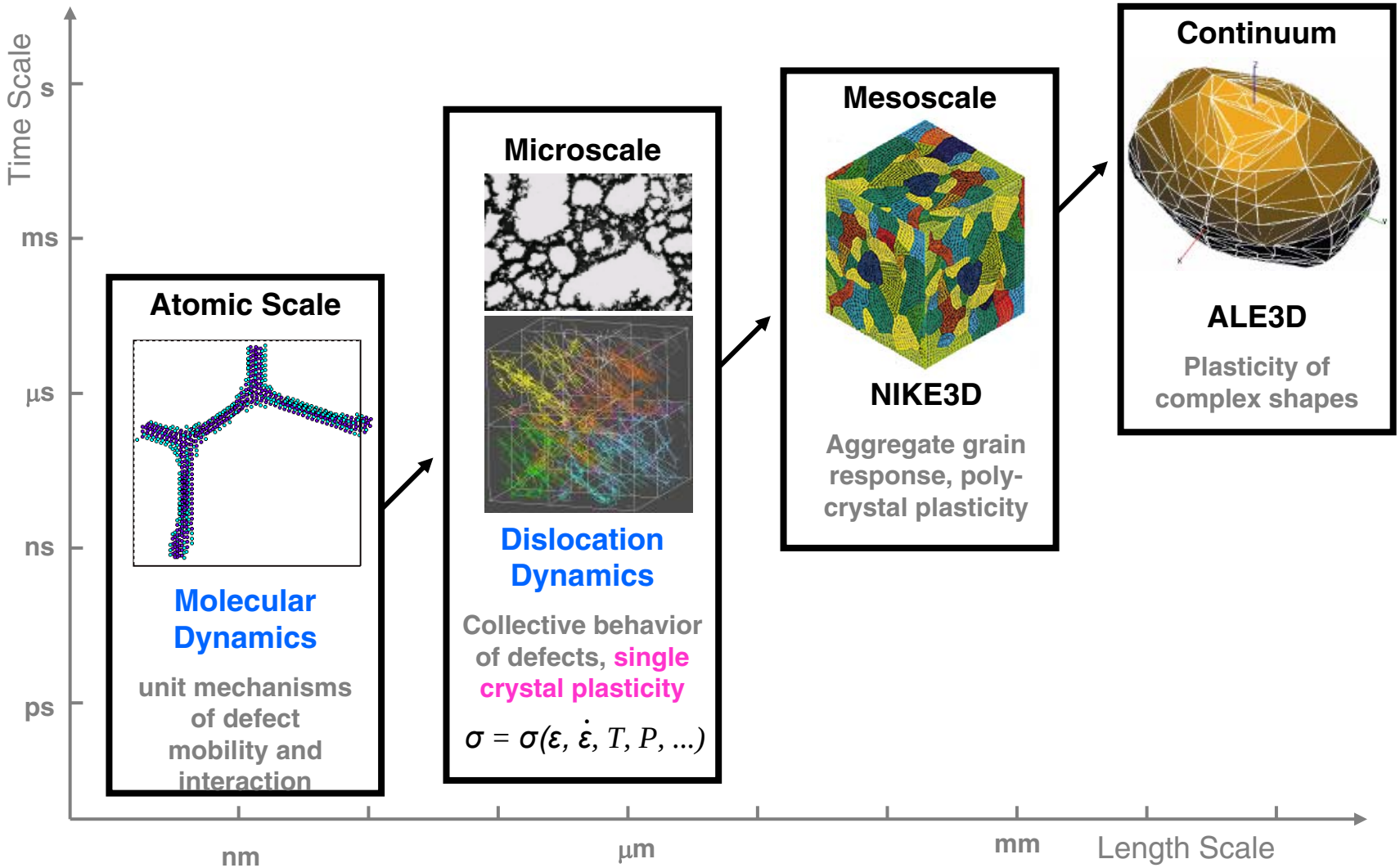


Curve 3

By bridging time and length scales, material sciences is the "killer applications" for BlueGene/L

- Compiling a list of applications of programmatic interest expected to effectively utilize BlueGene/L
  - Currently at 40 applications and growing – applications from LLNL, LANL, SNL and ASCI Academic Alliances
- Earliest adopter applications include
  - first-principles molecular dynamics
  - dislocation dynamics
  - atomistic materials models
- BlueGene/L will allow materials simulations at time and length scales that allow
  - overlap calculations of models from different scales
  - materials properties at a scale allowing direct comparison with experiment – 1 μm needed for NIF experiments, obtainable with BlueGene/L

By contributing at every length and time scale, BlueGene/L will first time allow for overlapping evaluation of models



Direct impact of a few targeted BlueGene/L applications on Stockpile Stewardship and DSW

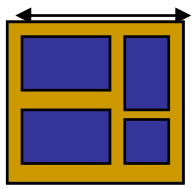
- GP – first-principles molecular dynamics
  - develops and verifies models at lowest end of multiscale modeling
  - extensive work for scaling to large processor count
- DD3d – dislocation dynamics
  - predictive simulation of crystal plasticity – currently the largest uncertainty in our multi-scale modeling of materials with strength – need 100x in space and 100x in time scales, possible with BlueGene/L
- ALE3D – large-scale continuum model
  - grain-scale modeling of detonation meterial
  - simulations with 2-3 materials possible on BlueGene/L
- Other exciting areas for early adoption
  - interaction of dislocations with grain boundaries
  - rapid resolidification of molten metal
    - forefront in grain formation, need large enough system to see formation of isolated grains – needed for next-level formalism
  - nanomechanics, a new frontier
    - current multiscale modeling based on bulk mechanics

Important scientific questions that will be addressed with BlueGene/L

- Why do atomistic calculations always get a mobility that is too low (by a factor of about 5)? What is the actual mobility?
  - BlueGene/L will allow direct overlap calculations to understand discrepancy
- How can we attain high density of dislocation lines--*ie* stop dislocation line loss across computation cell boundary?
  - BlueGene/L will allow simulation of large systems, allowing dislocation dynamics calculations at for volume sizes up to about 10 μm on a side
- How do we model the interaction between grains--*ie* when do dislocation lines cross grain boundaries?
  - BlueGene/L will provide simulations large enough to allow formation of multiple isolated grains within the simulation volume
- How do we closely couple multiscale model development with experiment (such as those on the National Ignition Facility)
  - BlueGene/L is a cost effective resource for creating simulations with the length and time scales needed for direct comparison with experiments

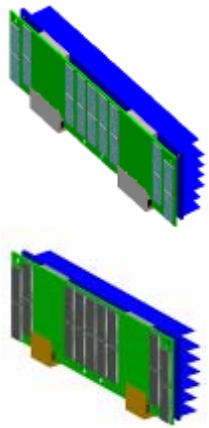
These are some of the qualitative differences in Stockpile Stewardship we expect from the two orders of magnitude differences in capability

Building BlueGene/L



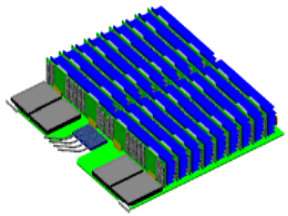
Compute Chip

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



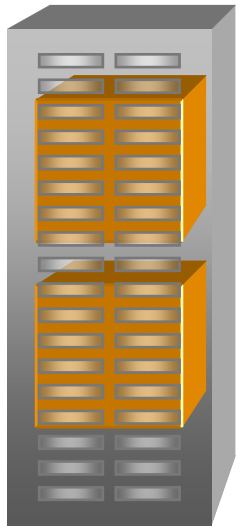
Compute Card
I/O Card

FRU (field replacable 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

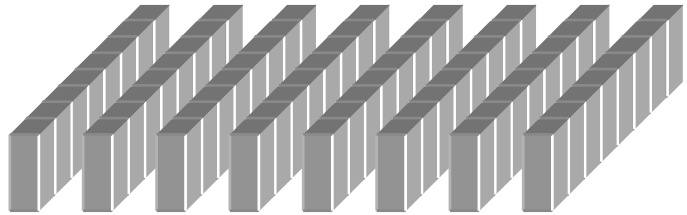


Midplane

SU (scalable unit)
16 node boards
512 nodes
(1,024 CPUs)
(8x8x8)
1.4/2.9 TF/s
128 GiB* DDR
7-10 kW

Cabinet

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



System

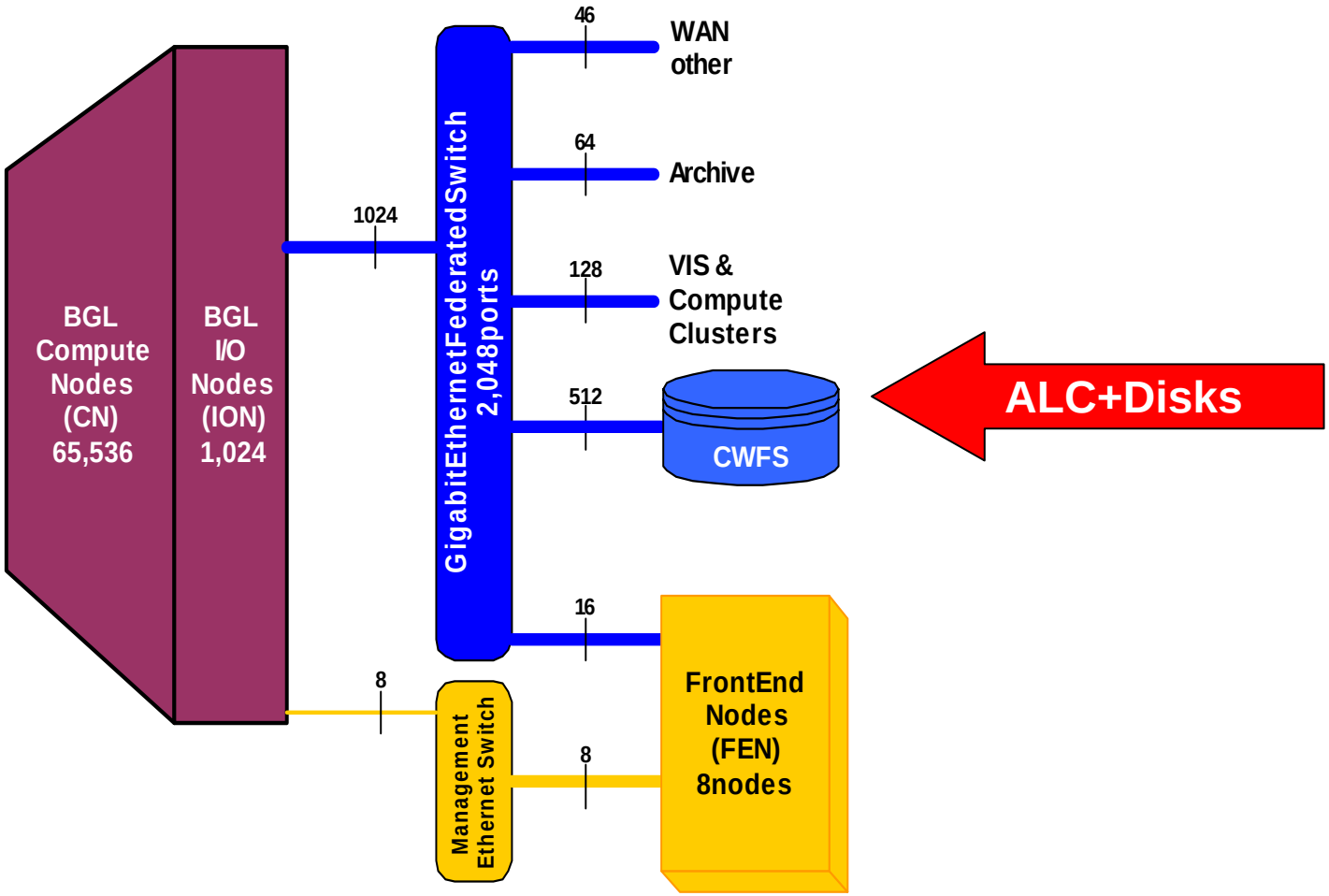
64 cabinets
65,536 nodes
(131,072 CPUs)
(32x32x64)
180/360 TF/s
16 TiB*
1.2 MW
2500 sq.ft.

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

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

Curve 3

BGL OCF Simulation Environment



Curve 3

BlueGene/L collaboration involves the Tri-Lab with a growing list of industry and academia

Lawrence Livermore National Laboratory 1952-2002

Los Alamos NATIONAL LABORATORY

Sandia National Laboratories

Applications
File Systems
Batch system
Kernel Evaluation
Programming Models
Debugger & Vis

IBM®

Hardware design and build
Network design and build
OS and system software

CACR

California Institute of Technology 1891

Network simulator
MPI tracing
Application scaling

ICL UT

PAPI - performance monitoring

Argonne National Laboratory · University of Chicago

MPI – message passing interface

TU WIEN Technische Universität Wien Vienna University of Technology

Optimized FFT

ATM

STAPL – standard adaptive template library

ETNUS

Debugger

pallas

Performance analysis
Vampir/GuideView

Beckman Institute for Advanced Science and Technology

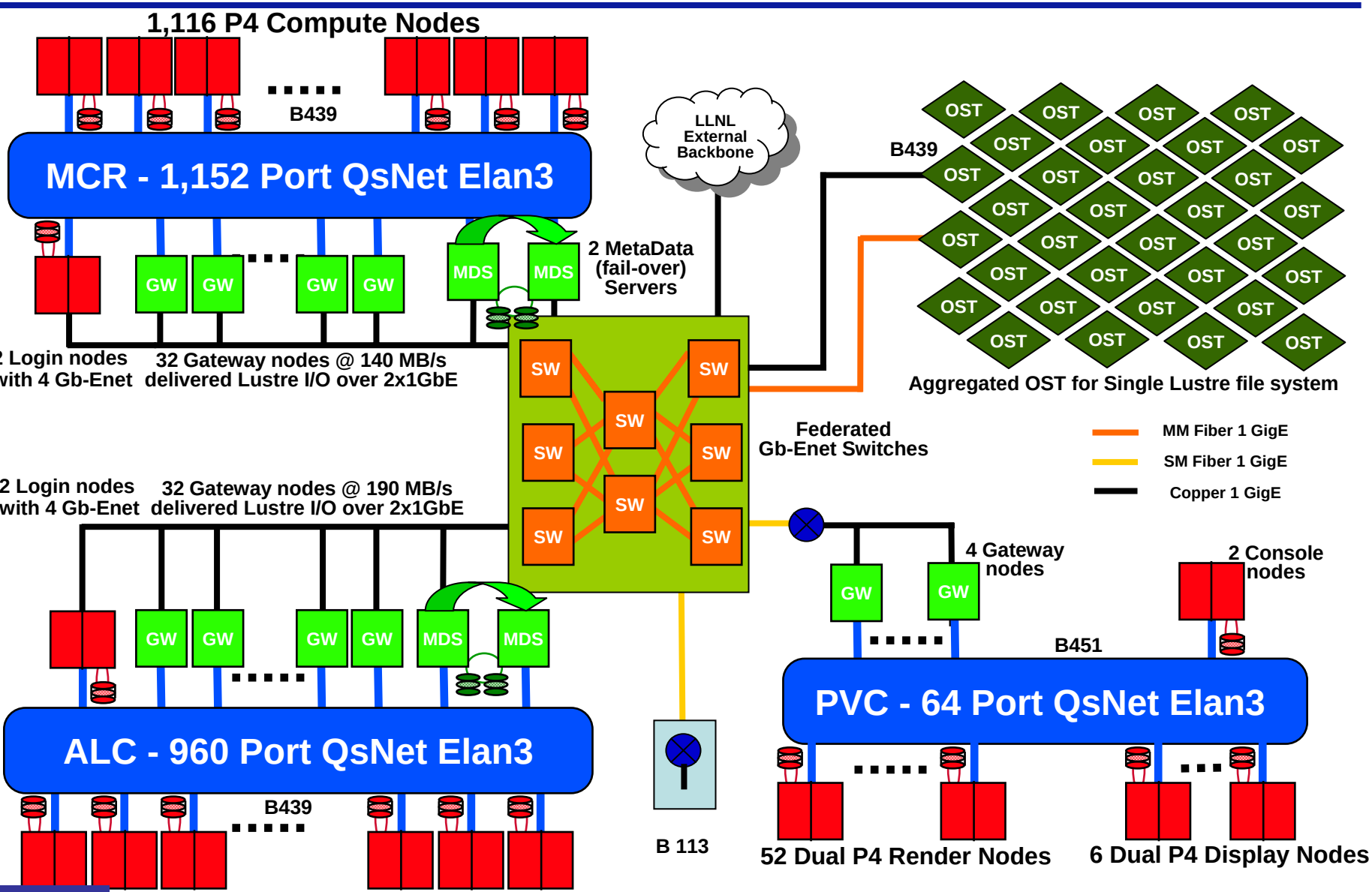
Parallel Objects
CHARM++

PMaC Performance Modeling and Characterization

SDSC

Applications
Application Tracing & Performance

Multi-cluster Simulation Environment based on a single Lustre File System is already impacting every program at LLNL

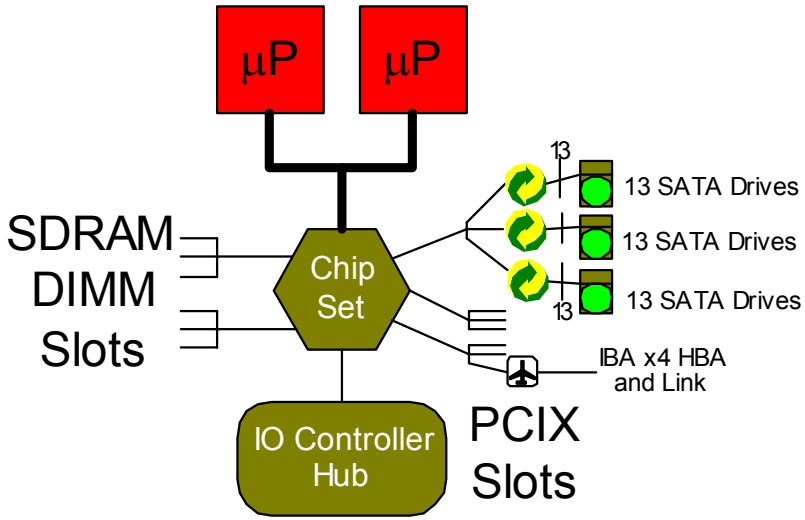


Curve 2

Commodity cluster wide file system based on Lustre Open Source file system

4U OST Based on Serial ATA RAID

280 GB drives, 33+3P drives = 9.24 TB per OST and 150 MB/s
10 OST/rack = 92.4 TB and 1.5 GB/s
25 racks is 2.31 PB and 37.5 GB/s
250 OST @ $15-$20K (depending on config) = $3.75-5.0M
This is $1.6-2.16M/PB



92.4 TB/rack

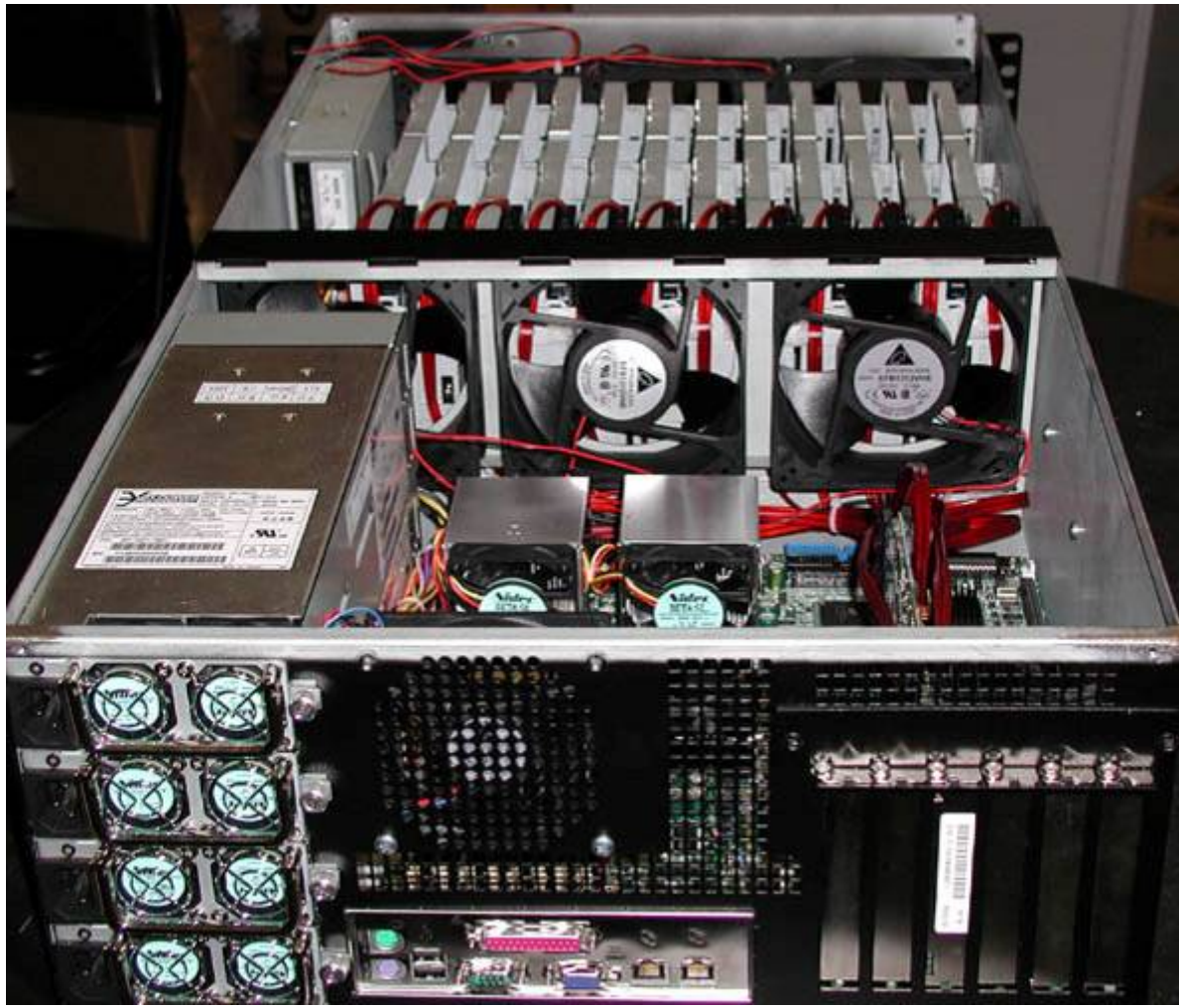
OST Serial IDE RAID
OST Serial IDE RAID
OST Serial IDE RAID
OST Serial IDE RAID
OST Serial IDE RAID
OST Serial IDE RAID
OST Serial IDE RAID
OST Serial IDE RAID
OST Serial IDE RAID

Curve 2

See http://www.lustre.org

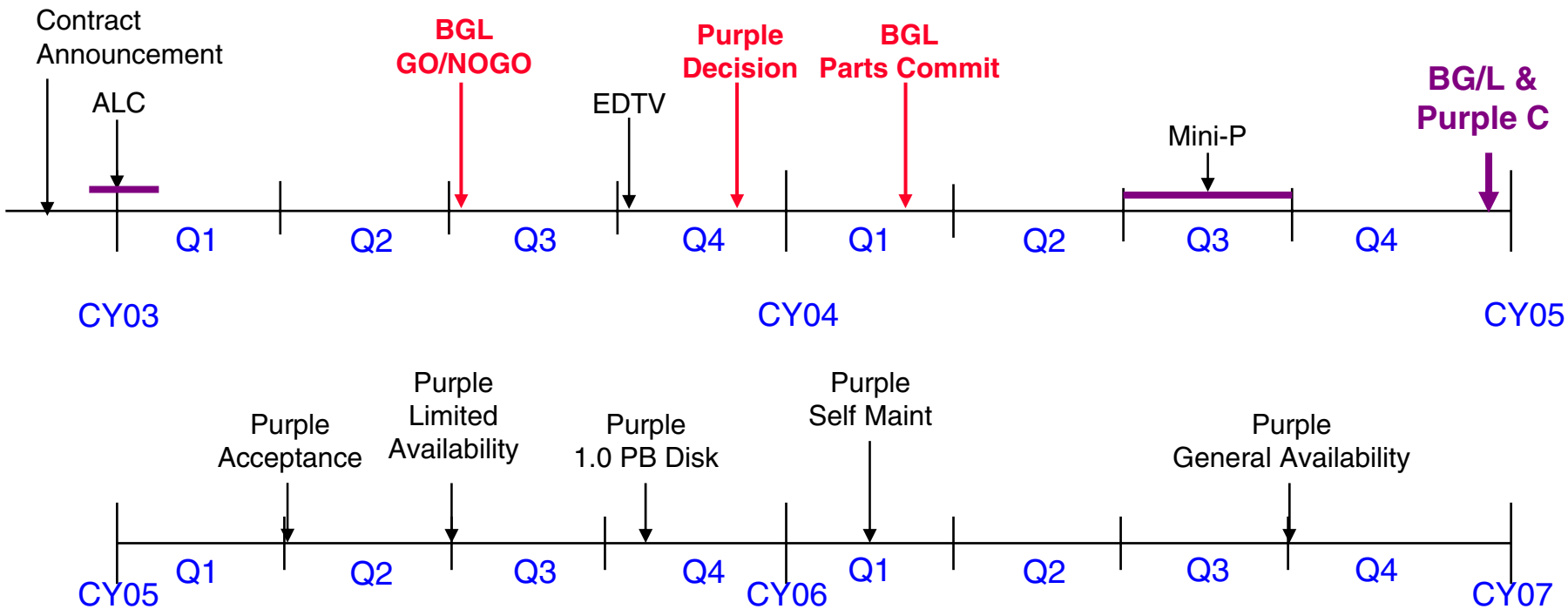
SATA OST is currently undergoing alpha testing

- Working with multiple partners
  - Some based on commodity nodes
  - Some based on proprietary RAID front ends
  - All offer 2x-4x $/B improvement
- Pictured unit includes
  - 36 SATA 180 GB drives
  - Choice of Dual Xeon Motherboard
  - Choice of interfaces
    - Up to three GbE
    - Up to six GbE TOE
    - Up to six FC2
    - QsNet Elan3
    - IBA x4



Curve 2

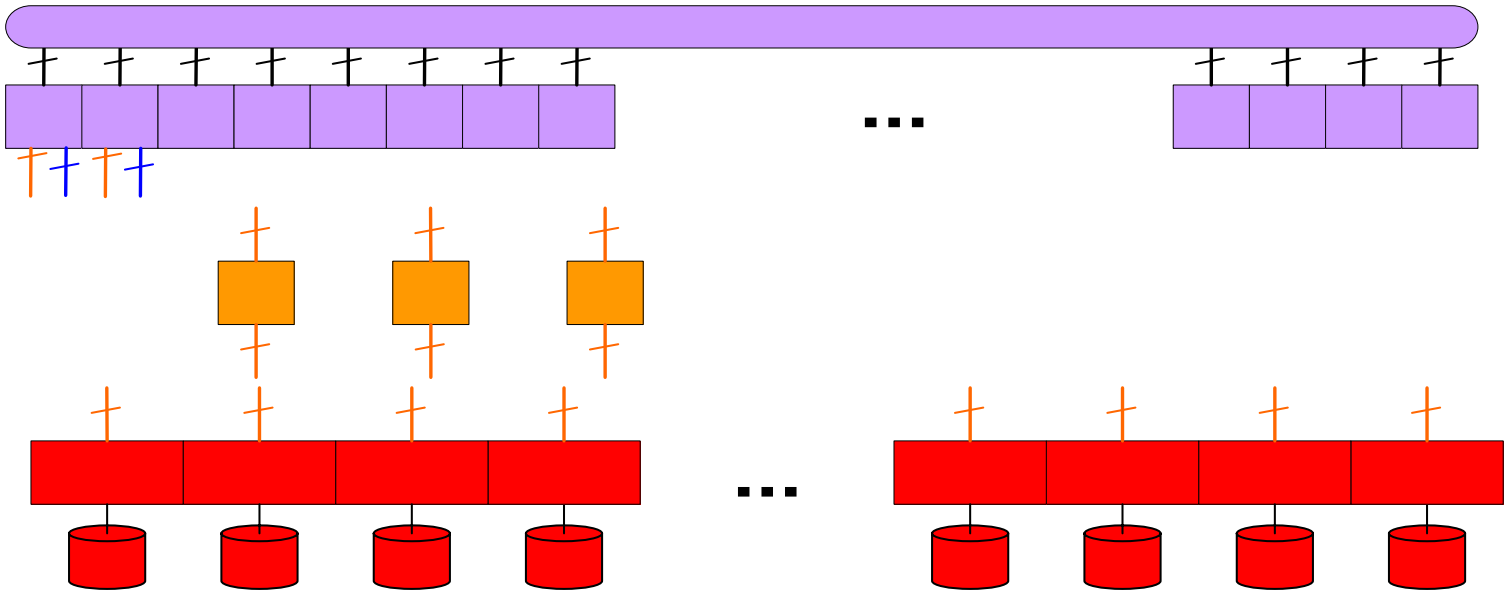
Purple Timeline



- Planned retirement
  - BG/L is December 2007 (three years after acceptance)
  - Purple C is December 2009 (five years after acceptance)

Curve 1, 2, 3

Purple Early Demonstration of Technology Vehicle†



- 32 Regatta H+ at 7 TF/s peak
  - 32-way SMP with Power4+ at 1.7GHz
- 4 TB of memory
- 147 TB global disk @ 7 GB/s
- Delivery October 2003
- First exposure to FEDERATION

- Federation
  - 8x4.0GB/s links (2.0GB/s per direction)
  - Peak B:F = 32/218 = 0.147
  - Delivered bandwidth will improve
    - 11.2 (19.2) GB/s non-striped
    - 7.52 (15.04) GB/s striped
  - Latency 10-15 µs

32 Regatta H+

2 GPFS IO @ 3.5 GB/s

3/s

8 8 8

2 3

Evaluating other options in conjunction with NERSC for BluePlanet type improvements

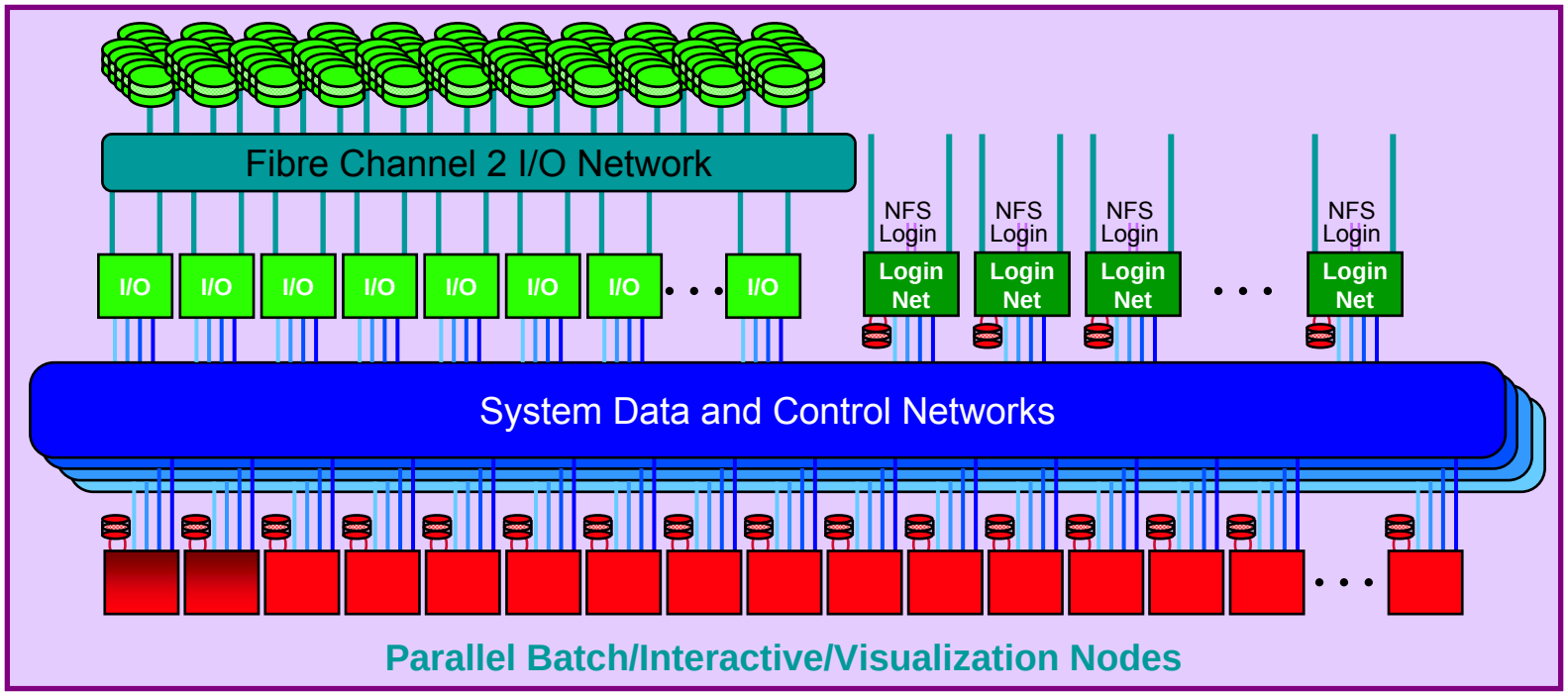
36 36

12 12

24 1Gb Links

Curve 1

Purple hardware architecture includes I/O, login and compute resources



Purple System

- Parallel batch/interactive nodes
- 4 Login/network nodes
  - Login/network nodes for login/NFS
  - 8x10 Gb/s for parallel FTP on each Login
  - All external networking is 1-10 Gb/s Ethernet
- Clustered I/O services for cluster wide file system
  - Fibre Channel2 I/O attach does not extend

Programming/Usage Model

- Application launch over all compute nodes up to 8,192 tasks
- 1 MPI task/CPU and Shared Memory, full 64b support
- Scalable MPI (MPI_allreduce, buffer space)
- Likely usage
  - multiple MPI tasks/node with 4-16 OpenMP/MPI task
- Single STDIO interface
- Parallel I/O to single file, multiple serial I/O (1 file/MPI task)

Squadron Improvements

- 64-way SMP
- Peak is 0.512 TF/s
- 256 GB of memory (B:F = 0.5)
  - Four outstanding loads per bank (rather than one)
- 682 GB/s of memory bandwidth (B:F = 1.33)
  - Large cache line size
- 16 Federation links with peak 64 GB/s (B:F = 0.125)
- Faster GX bus to RIO and Federation

- Purple Peak of 100TF/s
  - 30 TF/s sustained on sPPM+UMT2000
  - 2FTE effort to reach 45 TF/s on sPPM+UMT2000
  - Anticipate that 50-70% of time in UMT2000 will be divides…

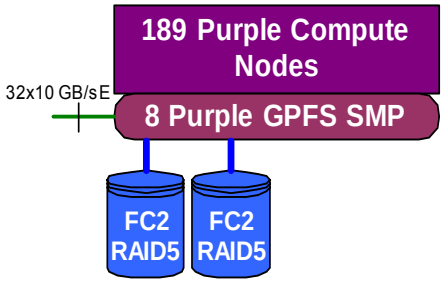
IBM Proprietary Information

IBM SP software stack for Purple will be very familiar to White users

- Code Development Environment
  - IBM Fortran90, C and C++, GNU g77, gcc and PGI?
  - OpenMP and MPI to 8,192 tasks
  - TotalView debugger
  - Scalable HPM and MPI tracing
- Resource Management
  - DPCS/LoadLeveler
  - ESP rating of >85%
- Cluster Software Management (CSM)
  - Dual boot with improvements in boot, install times
- GPFS parallel file system

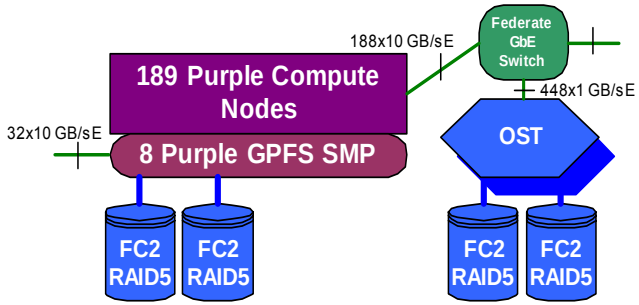
Curve 1

Migrating Purple C to Lustre

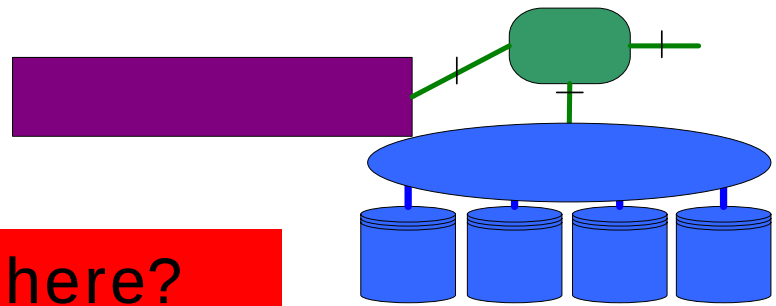


Stage 1: GPFS only @ 50 GB/s, pftp external access ~300 MB/s

Stage 2: Mixed GPFS and Lustre Each @ 50 GB/s



Stage 3: Lustre only @ 100GB/s



How do we get there from here?
Leverage work on BlueGene/L!

Curve 1

Summary

- We described a extreme scale platform strategy that balances risks and benefits to provide cost effective platforms for a range of uses.
- The straddle curve strategy requires extensive collaborations and the execution of extremely complex procurement and integration processes.
- This platform strategy has the potential to carry us from the 10 teraFLOP/s of today's platforms to 100 teraFLOP/s in 2004 and to 1 petaFLOP/s in 2006-2007 timeframe.
- This quantitative change will usher in another qualitative change in the science of stockpile stewardship and DSW