Addressing the Challenges of Petascale Systems Deployment

Presented to Salishan 2010 Conference

Mark Seager

Lawrence Livermore National Laboratory

29 April 2010

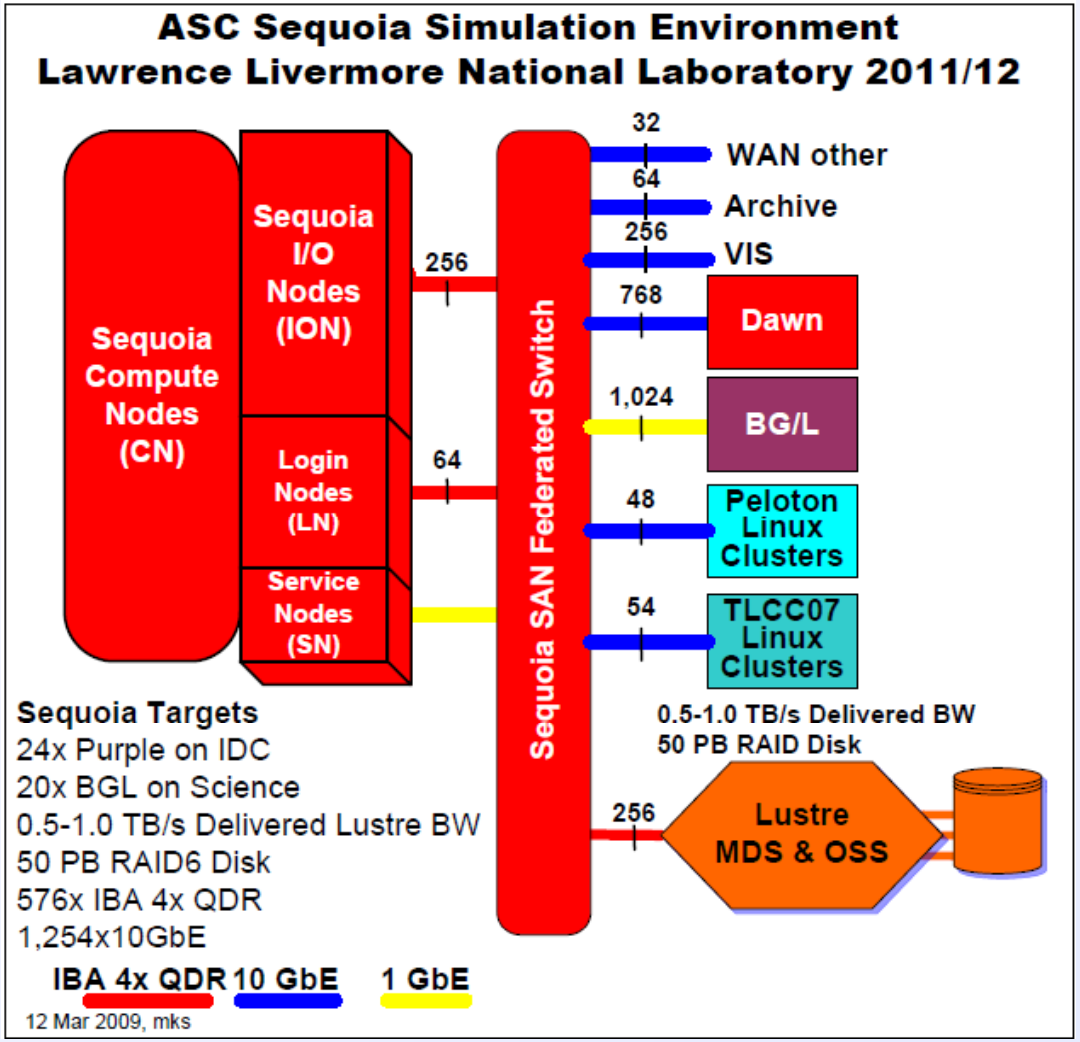
This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 LLNL-PRES-426142

Talk Overview

- Progress on Sequoia
- LLNL Facilities upgrade for Sequoia and Exascale
- The Hyperion partnership expanding to include a Data Intensive Testbed

Artist's rendition of Sequoia

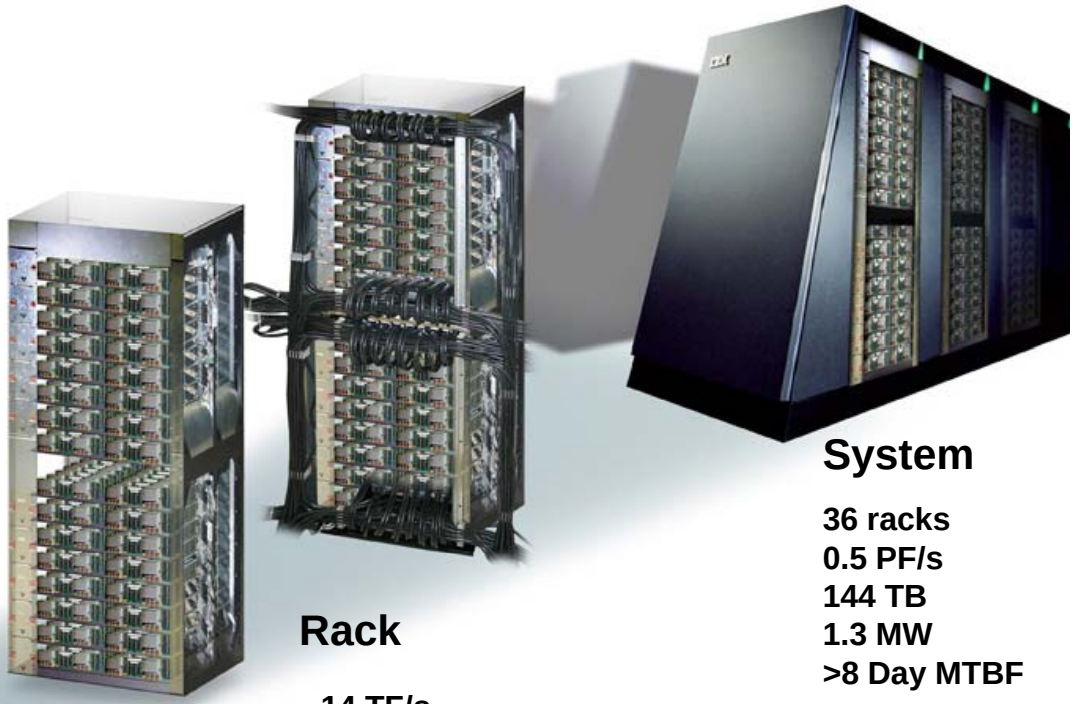
Sequoia Hierarchal Hardware Architecture in Integrated Simulation Environment



- Sequoia Statistics
  - 20 PF/s target
  - Memory 1.6 PB, 4 PB/s BW
  - 1.5M Cores
  - 3 PB/s Link BW
  - 60 TB/s bi-section BW
  - 0.5-1.0 TB/s Lustre BW
  - 50 PB Disk
- 8.0MW Power, 3,500 ft^2
- Third generation IBM BlueGene
- Challenges
  - Hardware Scalability
  - Software Scalability
  - Applications Scalability

DAWN

Sequoia Initial Delivery
Second Generation BlueGene

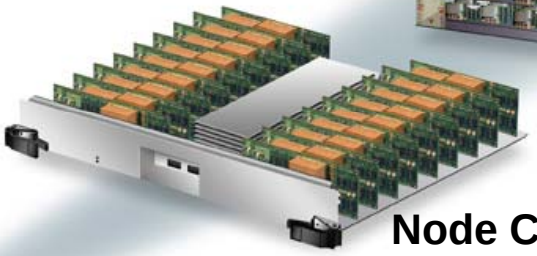


System

36 racks
0.5 PF/s
144 TB
1.3 MW
>8 Day MTBF

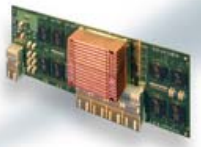
Rack

14 TF/s
4 TB
36 KW



Node Card

435 GF/s
128 GB



Compute Card

13.6 GF/s
4.0 GB DDR2
13.6 GB/s Memory BW
0.75 GB/s 3D Torus BW

Chip

850 MHz PPC 450
4 cores/4 threads
13.6 GF/s Peak
8 MB EDRAM

National Nuclear Security Administration

Dawn now in Classified Service and delivering to the program

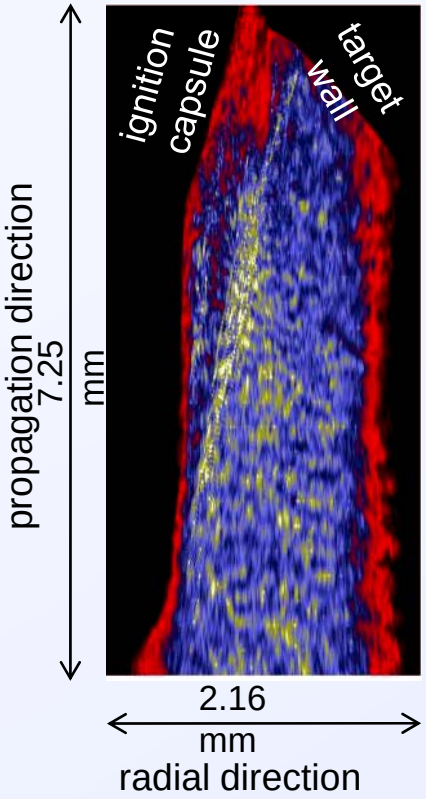
- Dawn hardware delivery started 19 Jan 2009. Rapid deployment of 36 racks completed ahead of an aggressive schedule
- Full Synthetic Workload acceptance test successfully completed 26 March 2009
- Twelve codes from Tri-Lab community ran on system during science runs
- Dawn Dedication 27 May 2009
- Now in classified service

The first half of DAWN (initial delivery of Sequoia) was received at the TerascaleSimulation Facility in late January, 2009

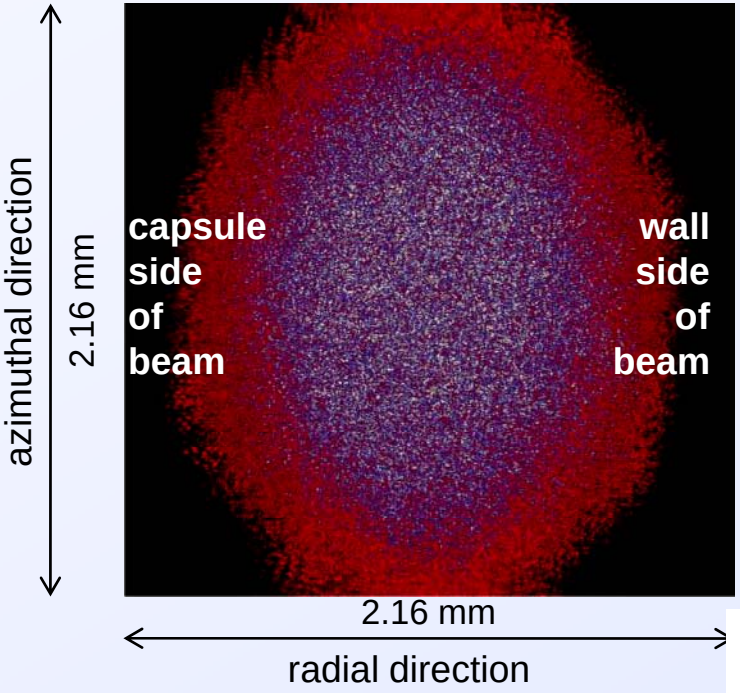
As an example of the interdependence of theory and experiment, NIF recently simulated an entire 30^0 beam quad with improved physics in preparation for Ignition

Ignition Design
30^0 Quad Intensity (W/cm^2)

Quad Propagation



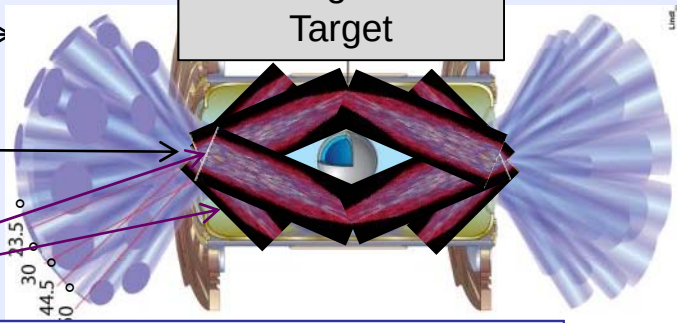
Quad Cross Section



- our simulations:
  -- resolve laser speckles
  -- include improved physics
  -- "more of the problem"
  -- show 9.5% reflectivity (within spec)

- The quad of beams we simulated is in this cone
- We plan to simulate two crossing quads later this year

NIF Ignition Target



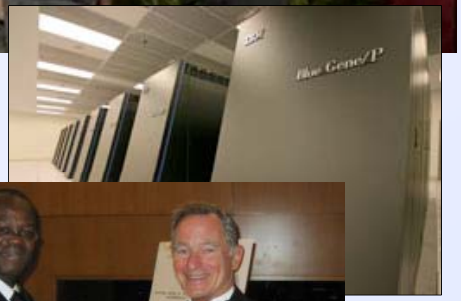
A mammoth four-week calculation completed June 10 using all of the 500 TF Dawn to support first ignition experiments…

2009 National Medals of Science and of Technology recognize LLNL accomplishments and collaborations

- **Berni Alder, computational pioneer**
  - Founder of molecular dynamics
  - Recognized for large-scale simulations to solve quantum mechanics problems

White House

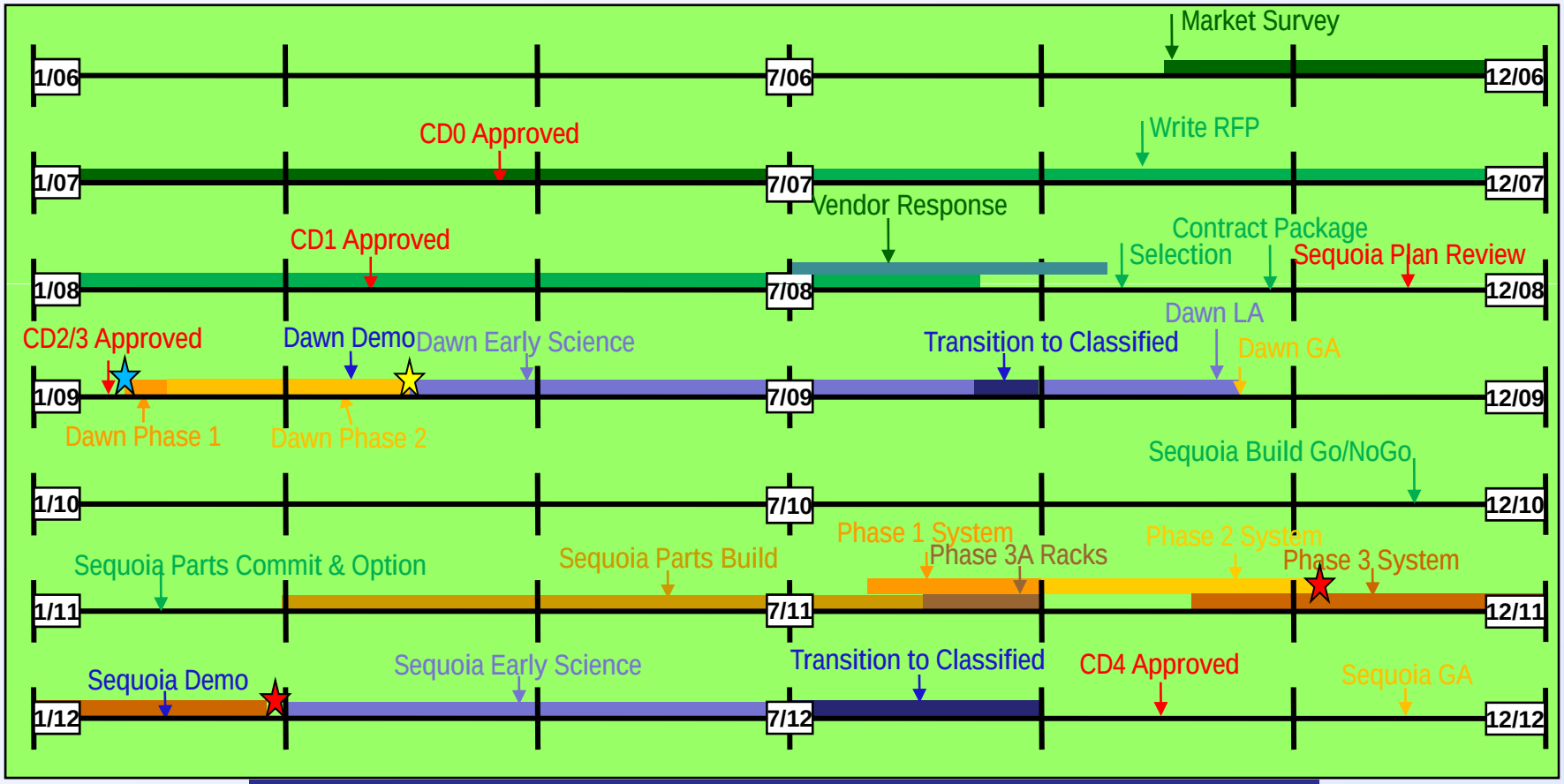
- **IBM - Blue Gene**
  - Series of energy-efficient supercomputers
  - LLNL and ANL partnership strongly impacted extreme-scale design and DOE supported IBM R&D



Awards Dinner

President Obama presented the Medals to Berni Alder and Sam Palmisano (IBM CEO) at a White House Ceremony on October 7, 2009

Sequoia Timeline Delivers Petascale Resources to the Program

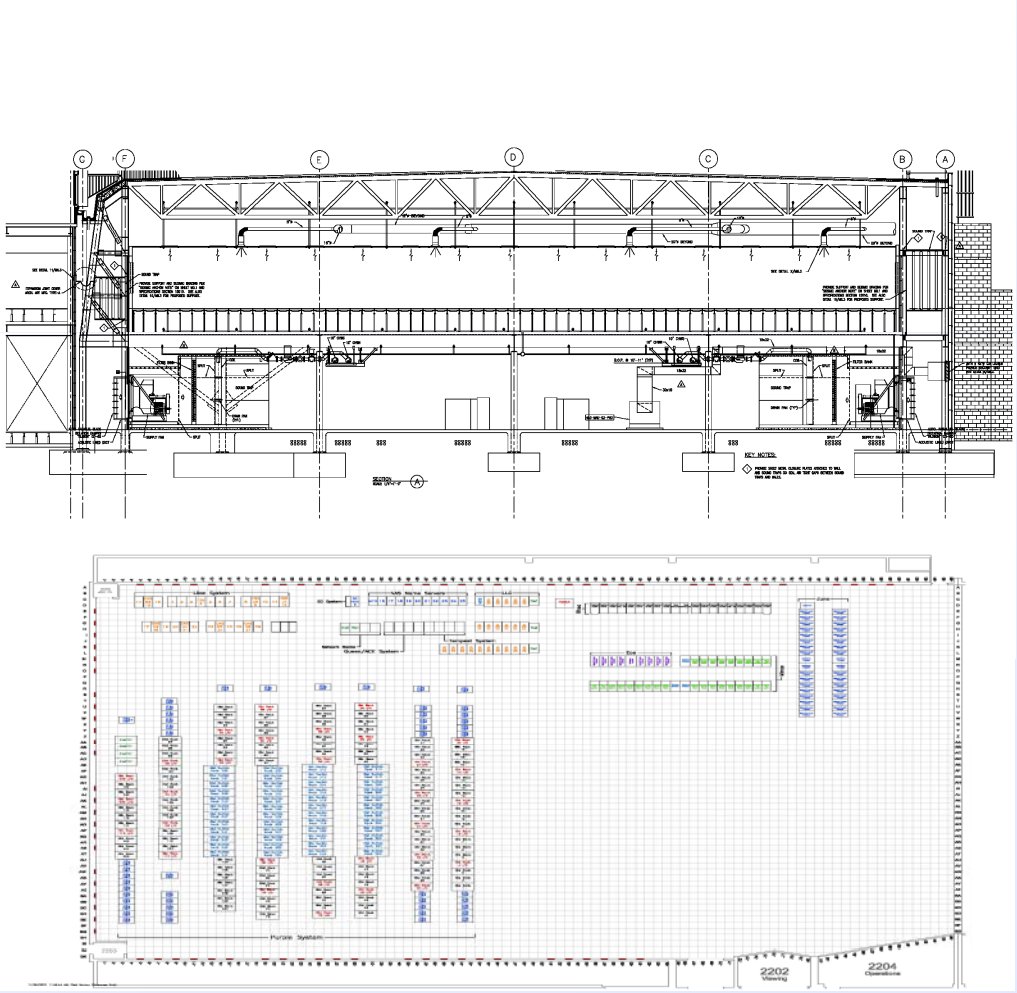


Sequoia Five Years Planned Lifetime Through CY17

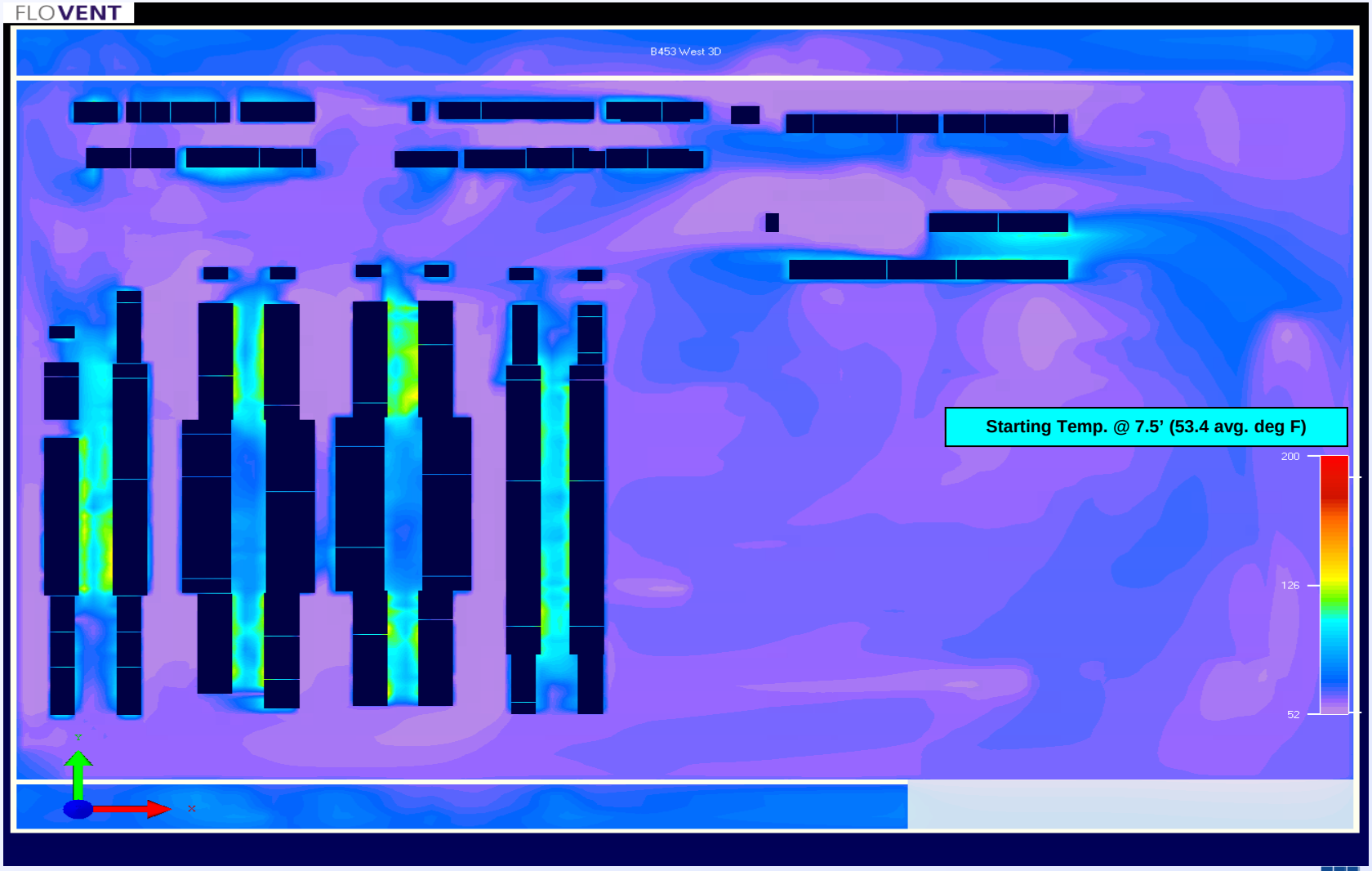
Sequoia contract award | Dawn system acceptance | Sequoia phase 2 & final system acceptance

A comprehensive computational fluid dynamic (CFD) model was performed to analyze airflow patterns in the TSF

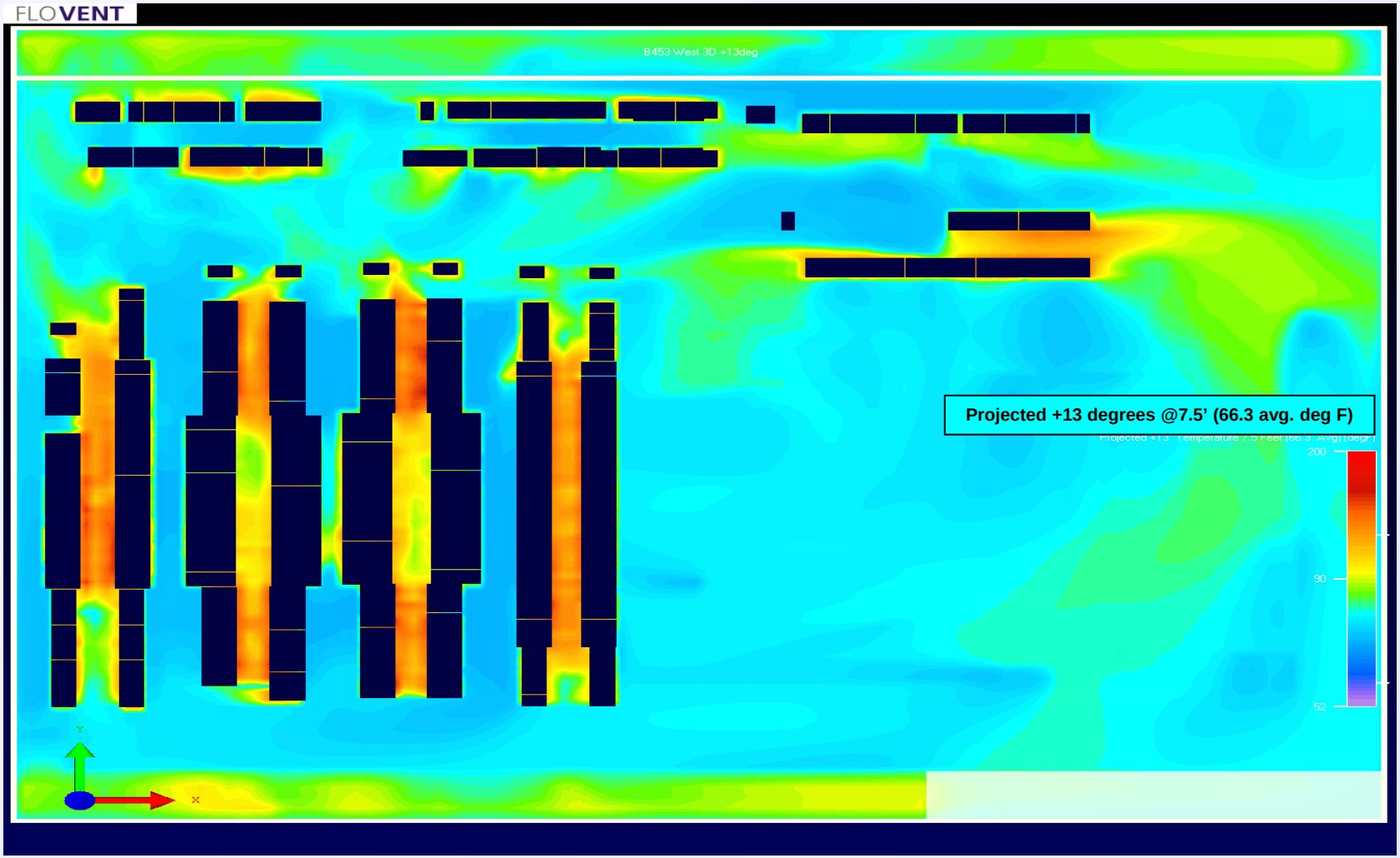
- Physical layouts imported
- Baseline CFD
  - Starting temperature 53.4°F
- Modeled airflow
  - 2” above finished floor (AFF) – inlet of racks
  - 7.5’ AFF – above the racks
  - 10.5’ AFF – ceiling



The TSF west room was baselined with a starting temp of 53.4°F average and modeled at 2", 7.5', and 10.5' AFF



The TSF west room was modeled with a projected temperature of 66.3°F average at 7.5’ AFF



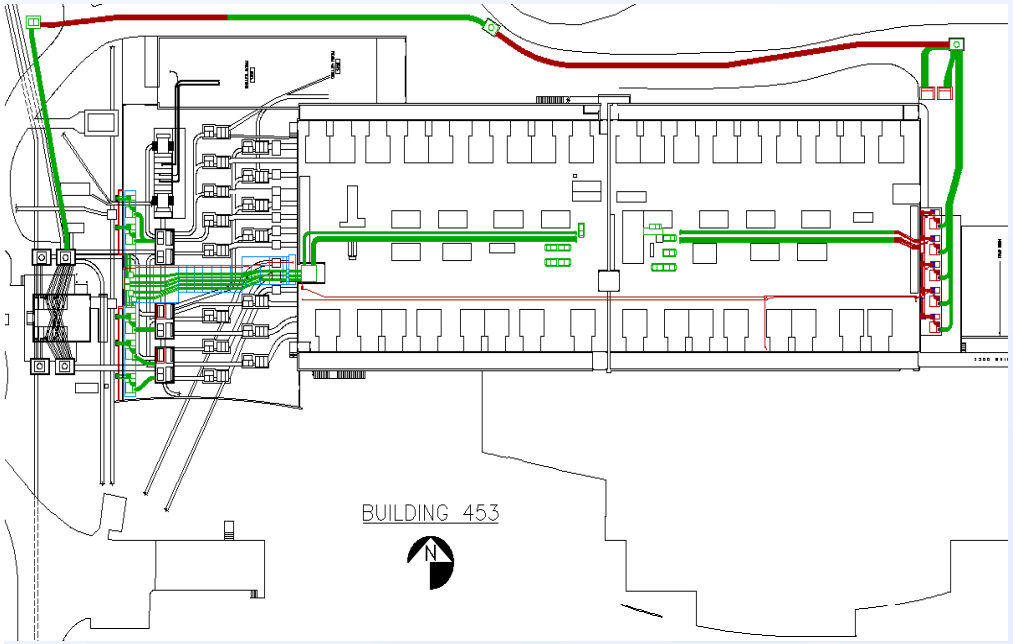
After benchmarking and CFD modeling, changes were made in the B453 without negative operational impact

- Supply air temperature increased from 53.4º F to 60.4º F (West) and 64º F (East)
- Supply chilled water temperature increased: 43º F to 50º F
- Air leakage addressed
  - Building penetrations sealed
  - Pillows and panels installed at rack level
  - Louvered perforated tiles replaced with solid tiles
    - 450,000 cubic feet per minute (CFM) leakage

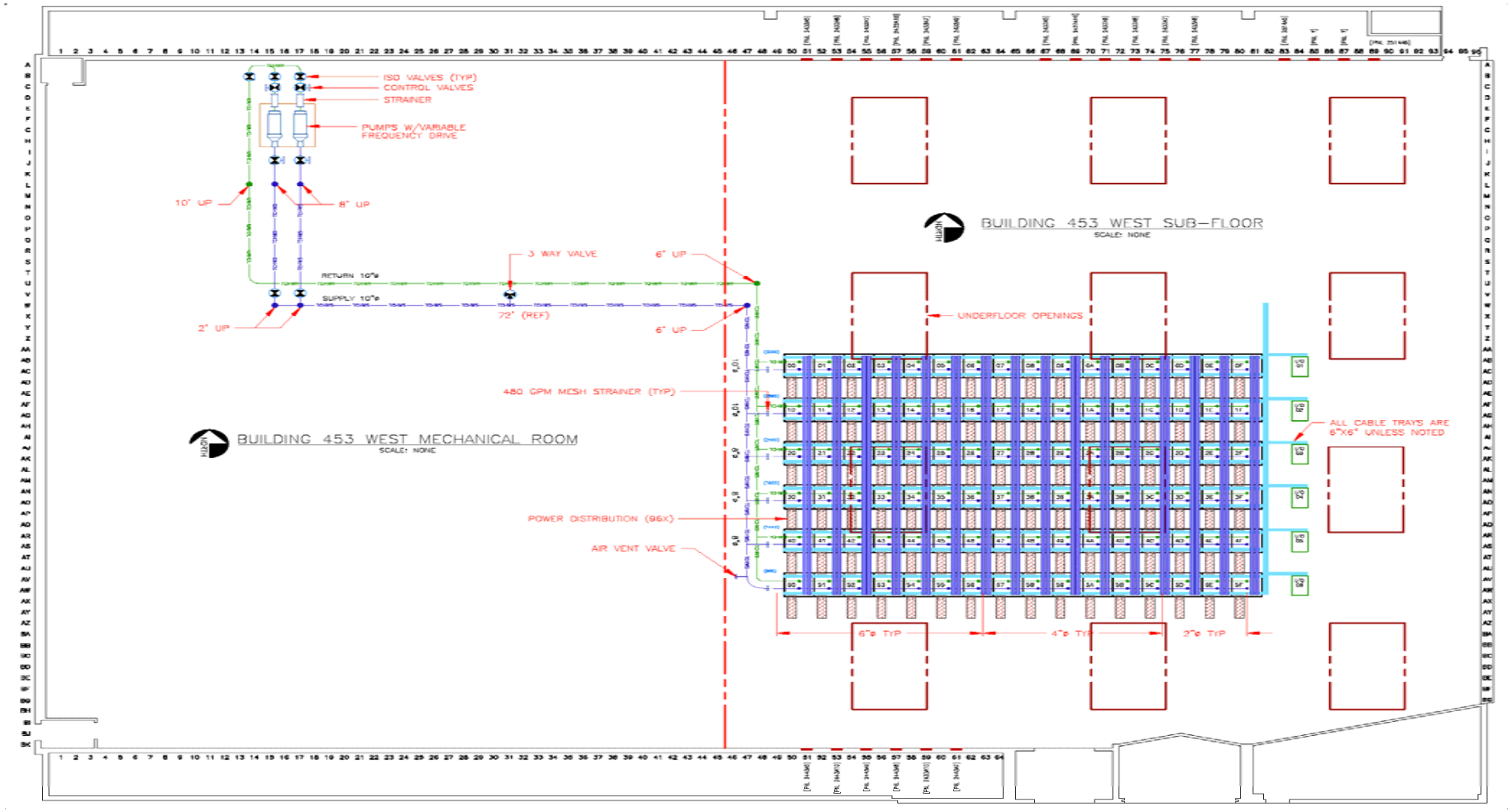
Initiative	Savings (kwh/yr.)	$/kwh	Total Savings ($)/yr.
Raise Air Supply Temperature	32,307,692	0.065	$ 2.100,000
Raise Chilled Water Supply Temperature	15,677,704	0.065	$ 1,019,051
Air Leakage (2) Air Handlers Off Line	542,211	0.065	$ 35,244
Total			$ 3,154,295

B453 computer room power is being scaled from 15MW to 30MW

- Capitalize on the computational efficiencies (TF/MW and SF/TF)
- Capitalize on the electrical/mechanical system efficiencies
- Adding an additional 15MW into the TSF

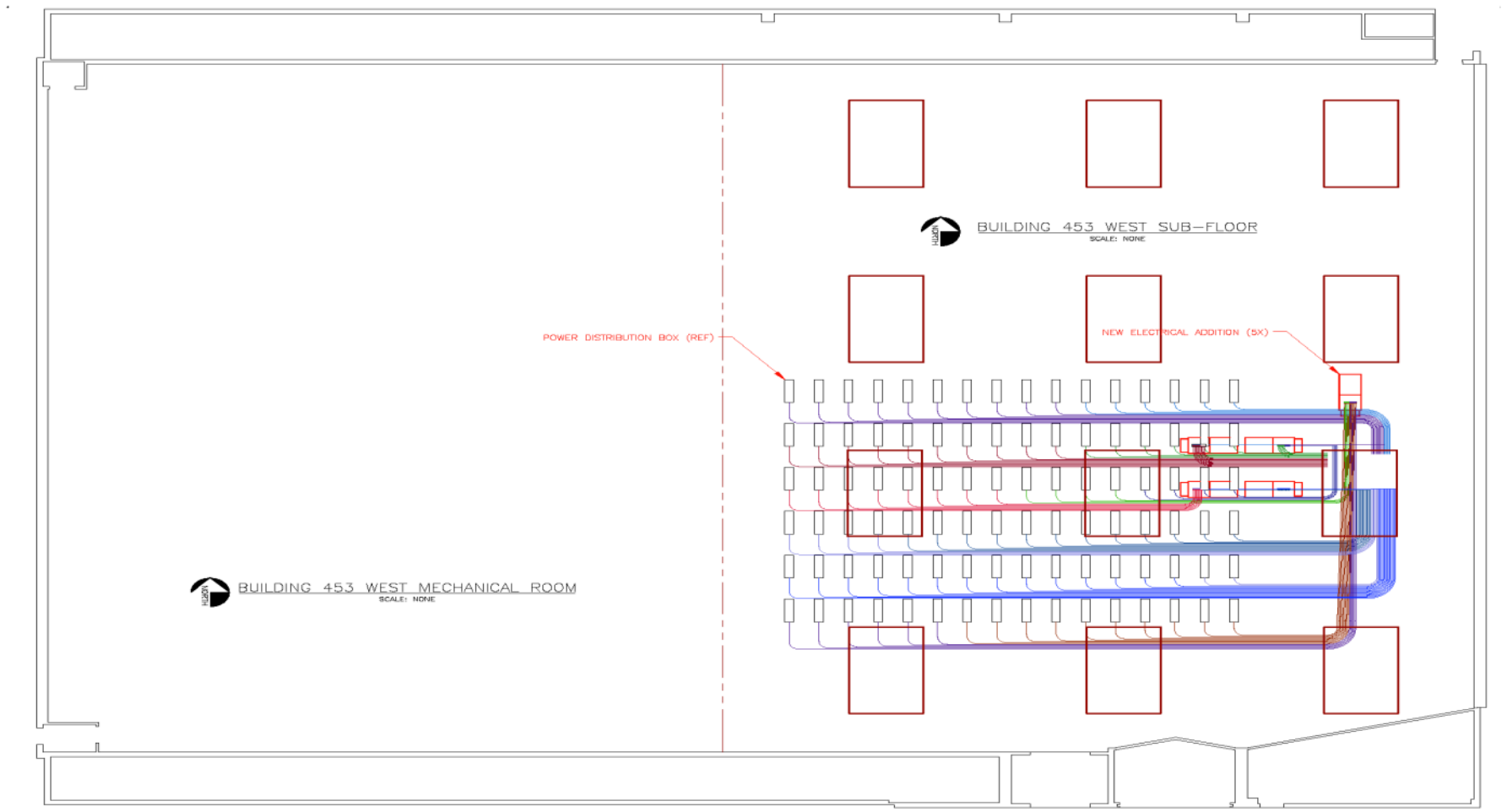


Sequoia Overall Facilities System Layout in B453 West Room

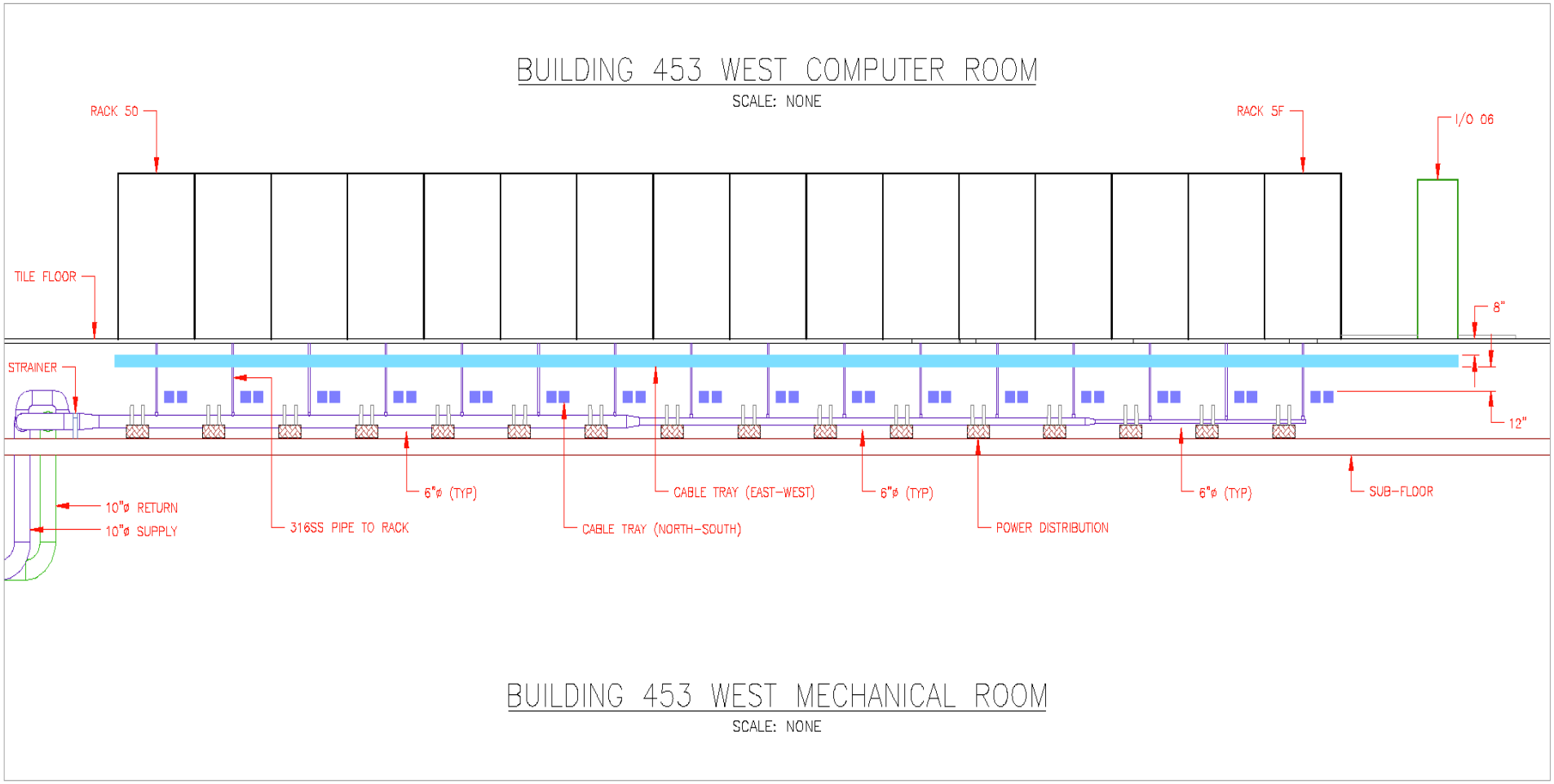


Locally designed under rack power cord consolidation saves significant installation costs and increases facilities efficiency

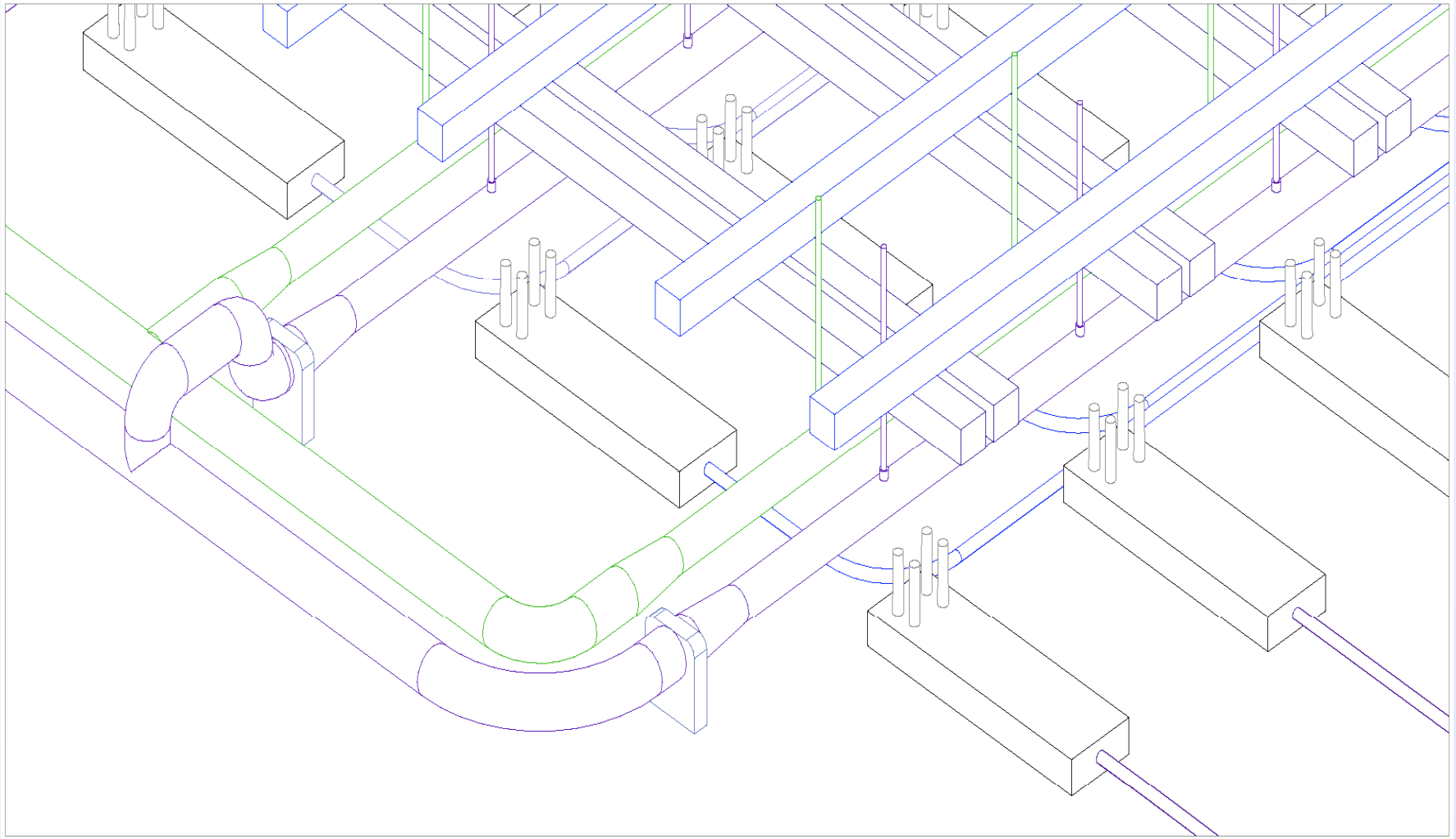
Sequoia Electrical Distribution in B453 West Room



Sequoia Facilities System Layout – Profile View

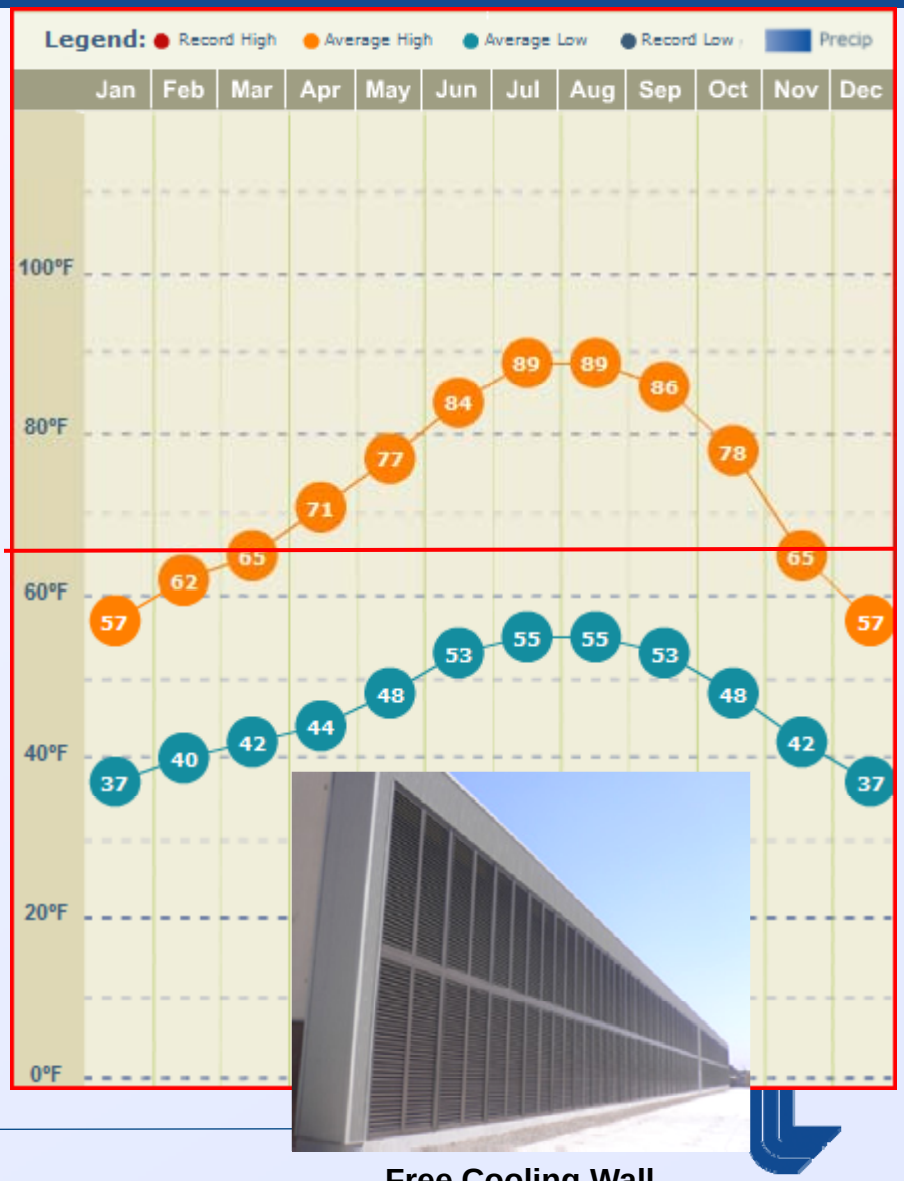


Sequoia Facilities System Layout – Under Floor Isometric



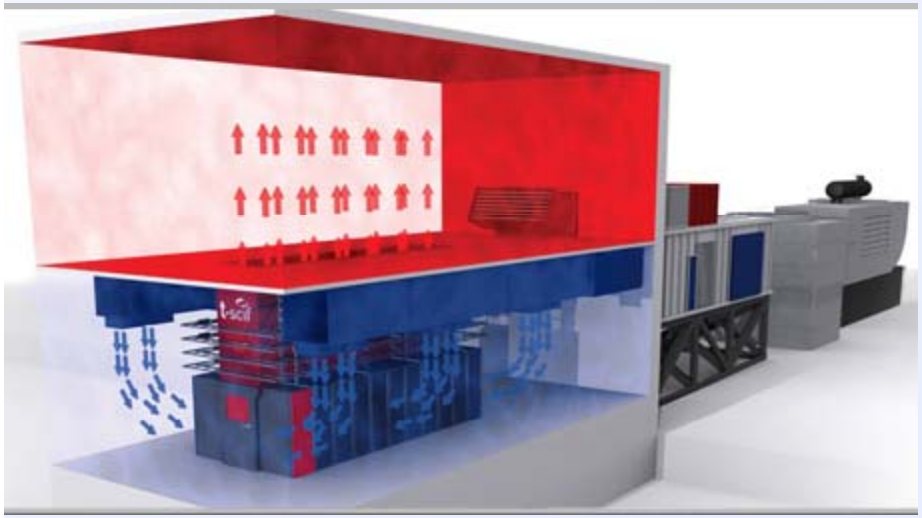
Facilities for future systems will continue to be a challenge, still need to further improve PUE

- Large fraction of time outside air near or below inlet air requirement
- Free air cooling gets you to a PUE of 1.04 or better, if one includes reductions in fans in systems

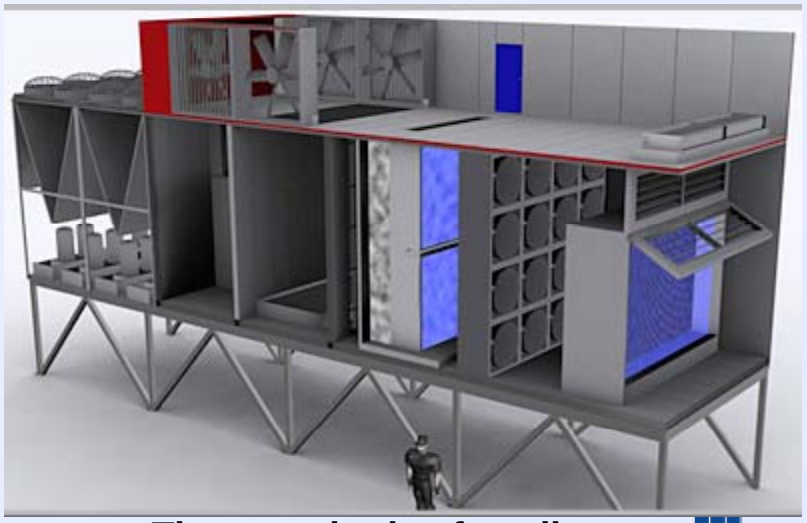


Free Cooling Wall

To minimize capital costs and maximize PUE, large scale data centers moving away from traditional raised floor configurations



Hot & Cold isle containment, slab floor



Three methods of cooling

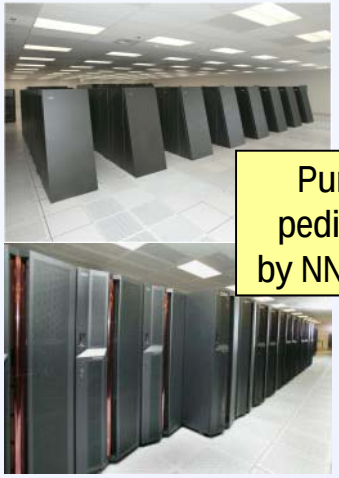
Free Cooling Evaluations for B453 indicate improvement of PUE to 1.17 or better

- Currently working with Johnson Controls to develop the mechanical and controls modifications required
  - Existing supply/return system is built into the north and south walls of B453
  - Modify to install louvers, intake dampers, humidifiers, filters and an array of multiple fans
  - Evaluating the use of airside economizers to take advantage of seasonal and nighttime outside air variations to provide cooling

B-453 Exterior Elevations

Future Free Cooling Wall

Sequoia reduction strategy is multifaceted and intended to provide a production multi-petaflop simulation environment



Purple and BG/L pedigree leveraged by NNSA User Facility

Dawn

IBM BG/P

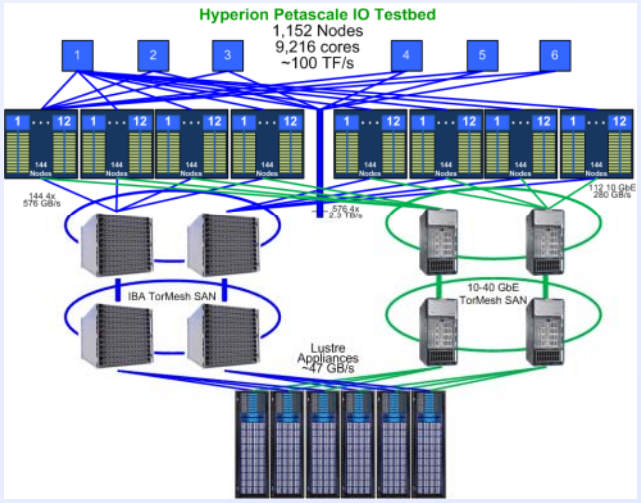
Weapons codes transition to multicore architecture 2009-2012

SEQUOIA

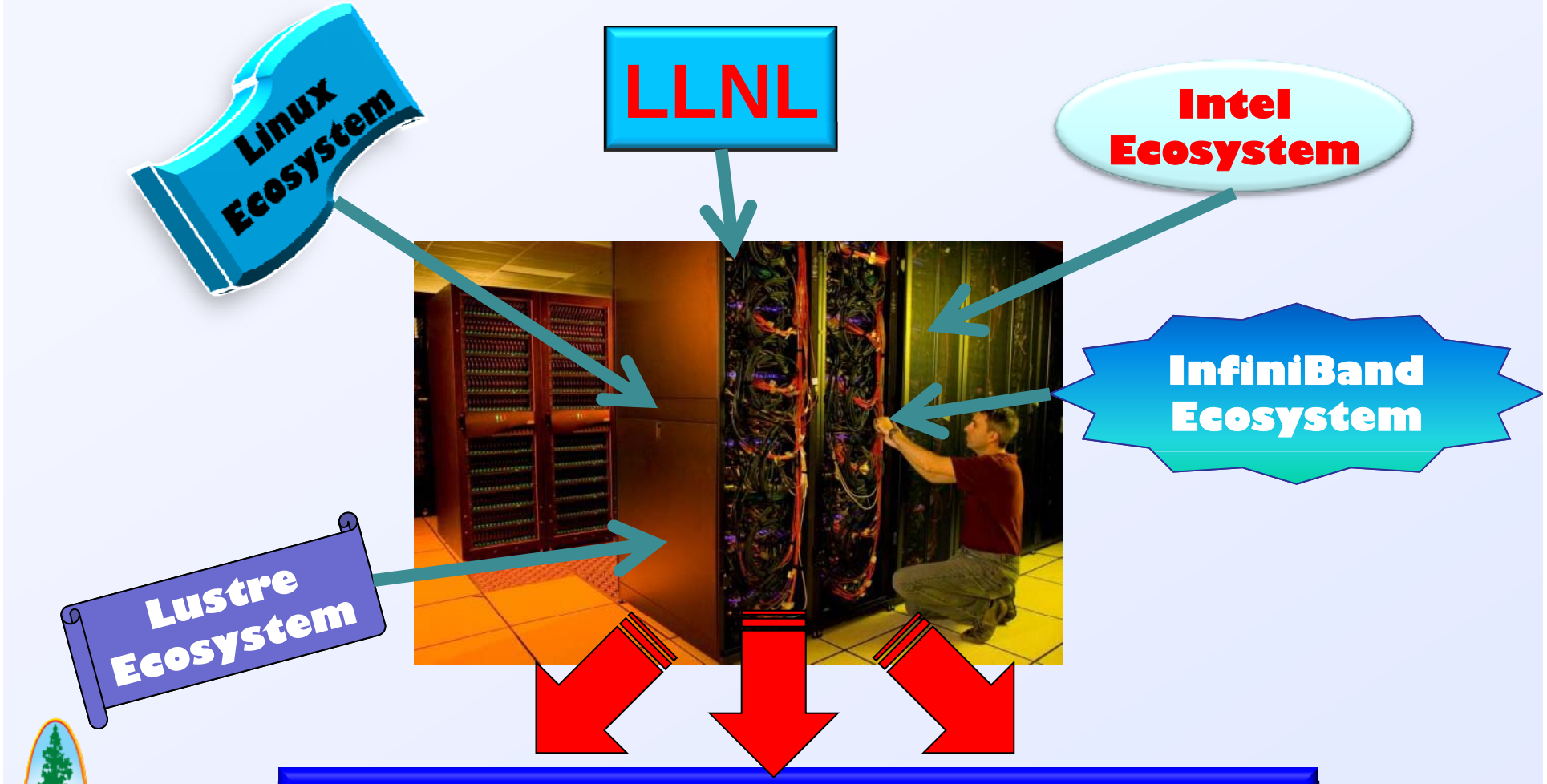
IBM BG/Q

Innovative cost-shared test bed, "Hyperion" to prepare for the massive I/O demand (1+TB/s)

Cost-sharing model supported by LLNL Industrial Partnerships Organization (IPO)



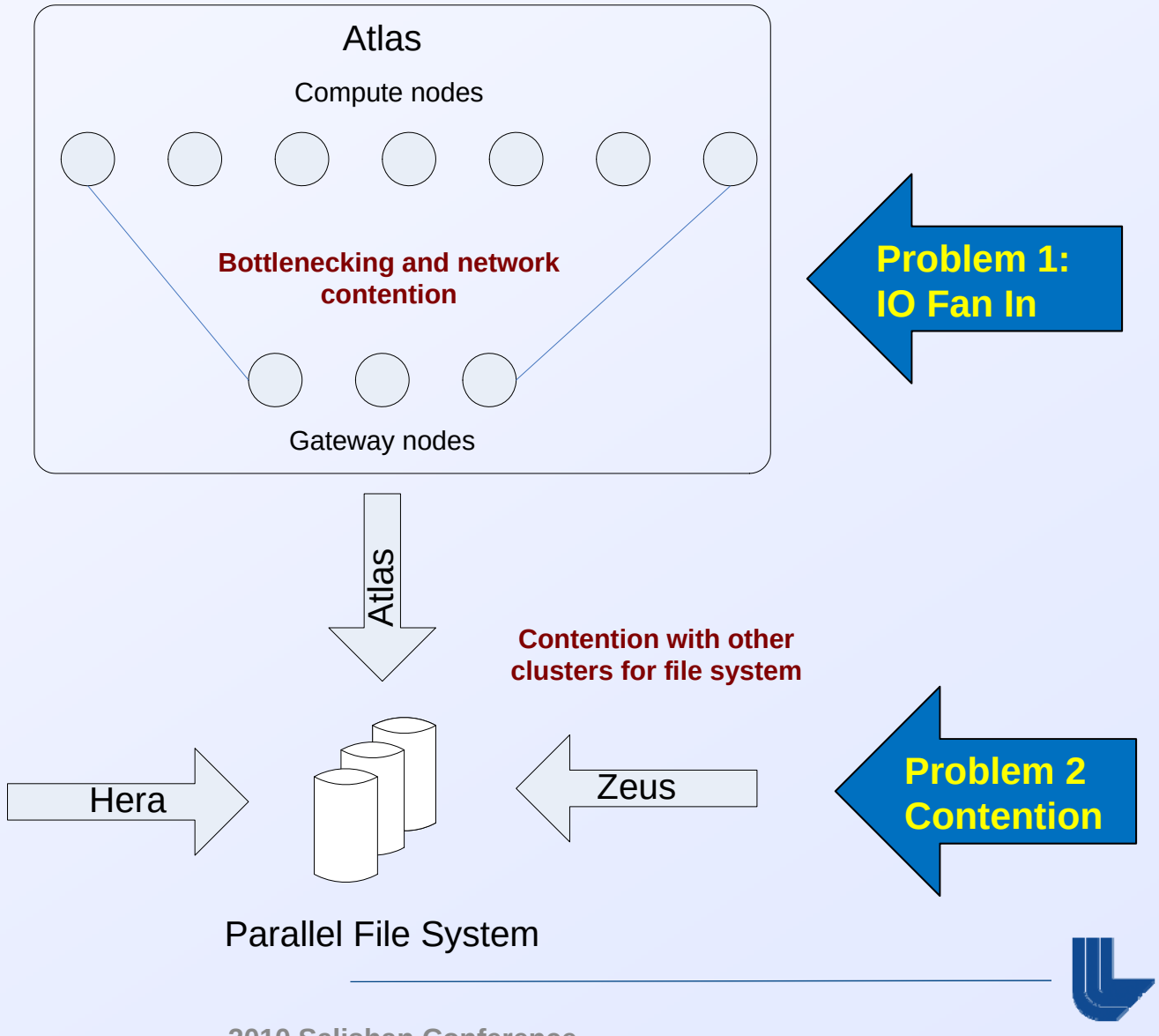
As part of the Sequoia risk reduction strategy, we have catalyzed the Hyperion collaboration of 11 partners to build something unique and beyond what any one partner can achieve



Hyperion Partnership Update

- 2009 HPCWire Award for best “Government & Industry Partnership”
- Moved system to green network and have foreign national collaborators on the machine
- IBM/Houston considering joining partnership
  - Test next release of HPSS at scale
  - Collaborators (Sun) to test Lustre HSM back-end
- Collaboration wants to develop outreach activity to ISV community
- Major IO expansion planned for FY10 for scale testing in preparation for Sequoia

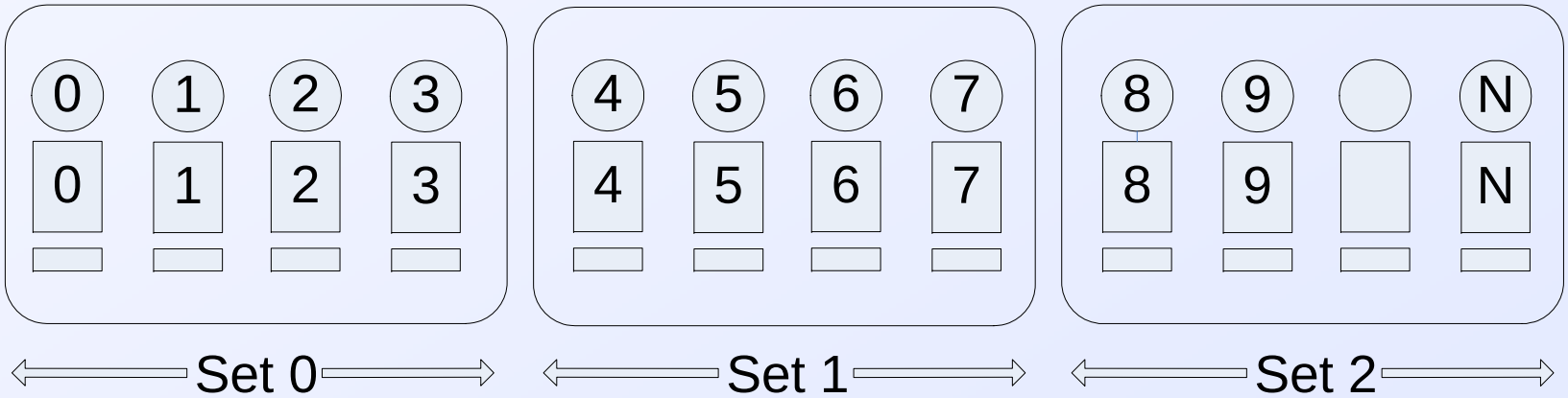
By writing to local file systems, Scalable Ceckpoint/Restart (SCR) Library Avoids Two Problems



From Adam Moody

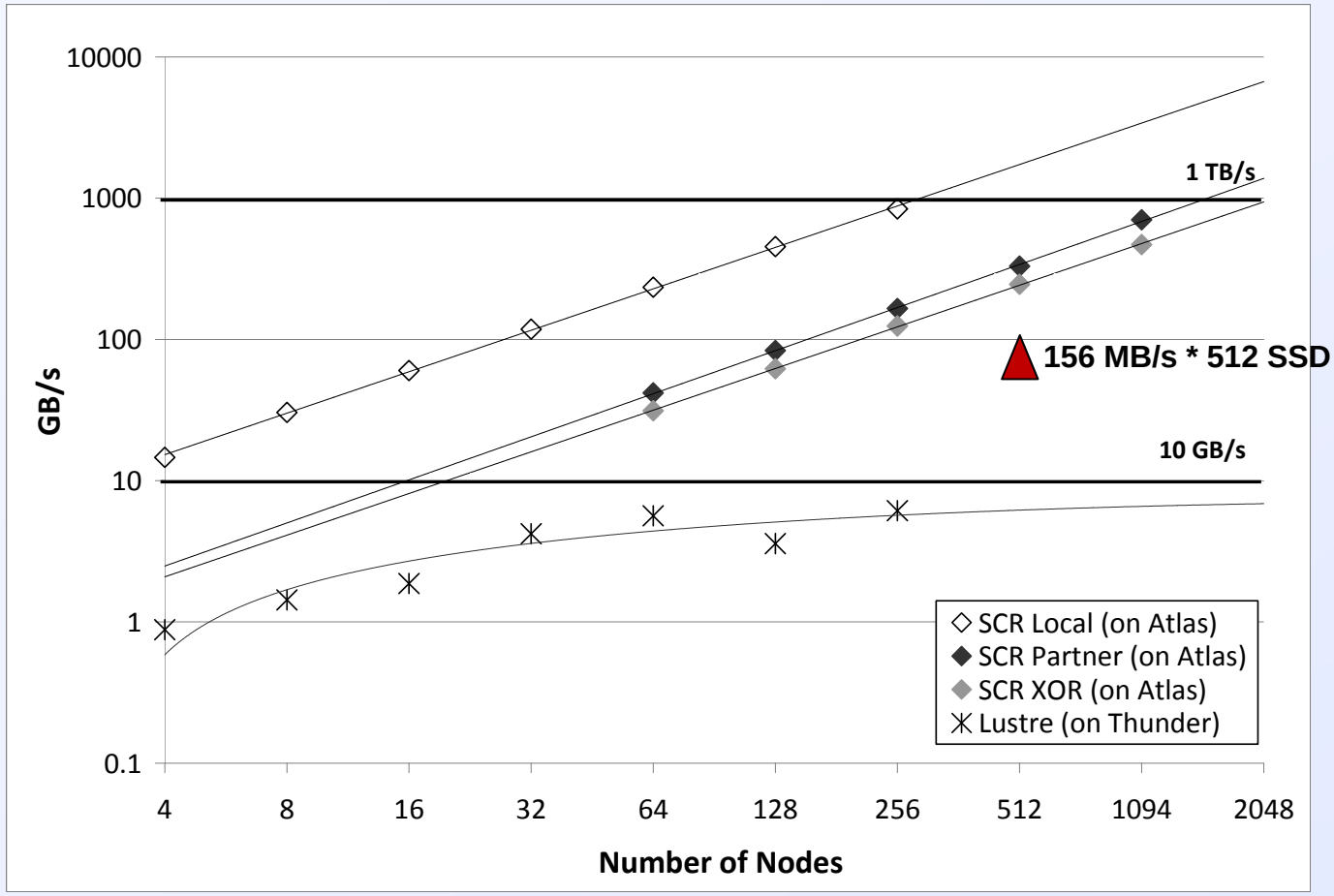
SCR utilizes a sophisticated XOR redundancy to distribute data and reduce overheads

- Break nodes for job into smaller sets, and execute XOR reduce scatter within each set.
- Can withstand multiple failures so long as two nodes in the same set do not fail simultaneously.



From Adam Moody

Benchmark checkpoint times to RAM disk and local SSD provide scalable bandwidth to applications



From Adam Moody

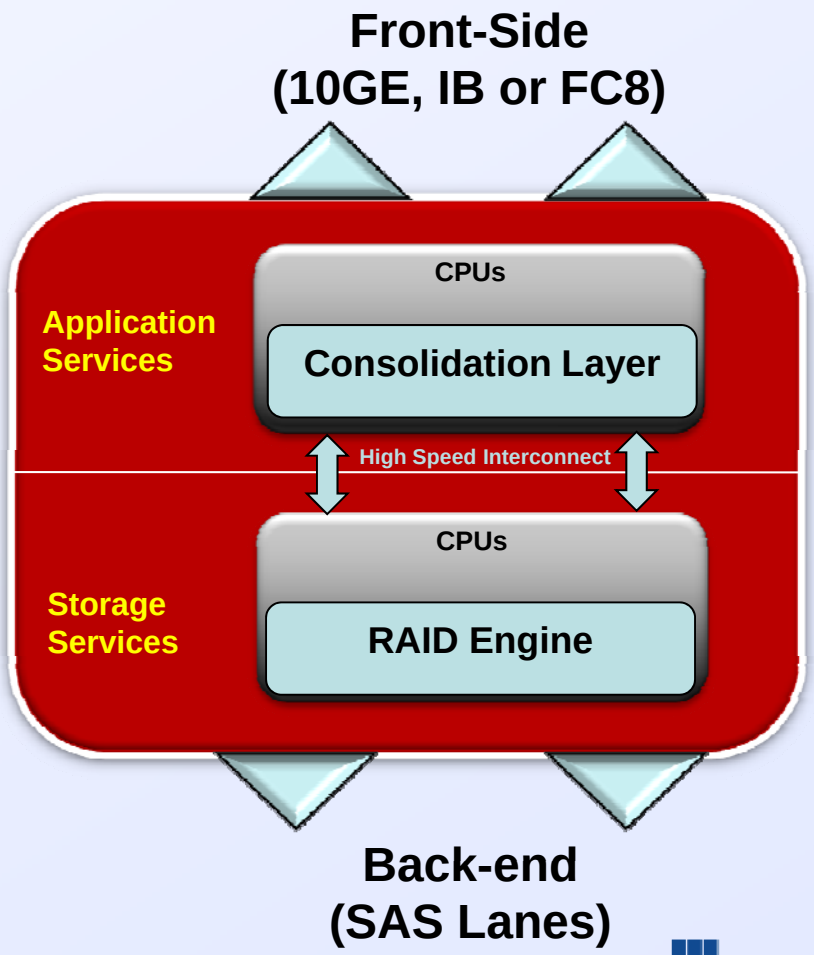
Storage Fusion Architecture is a Disruptive Technology for Storage

Features

- Dedicated resources provided to Storage Services and Application Services
- High-speed internal connections and shared memory architecture
- Protocol conversions eliminated
- Massive and balanced front-side and back-end bandwidth

Benefits

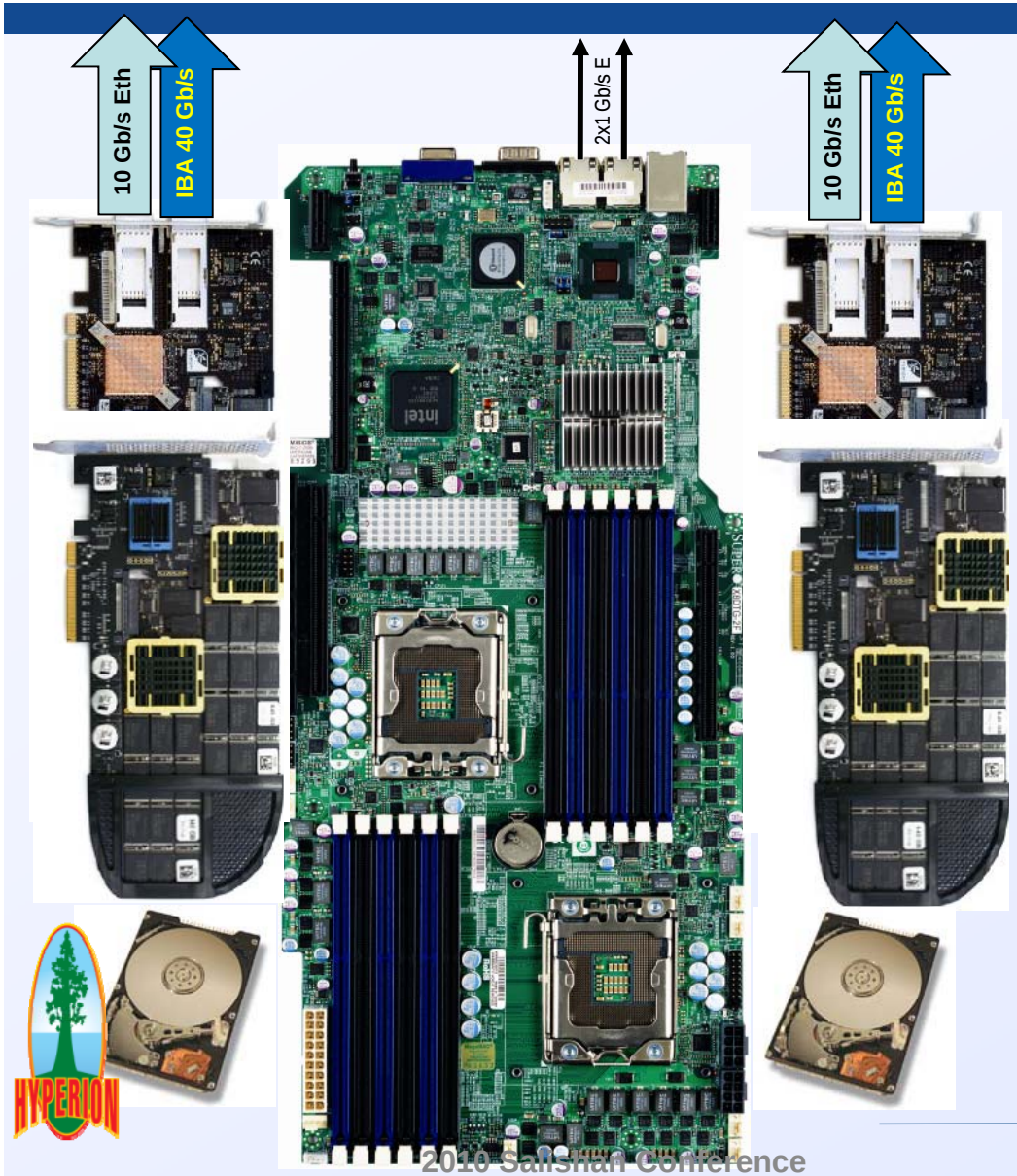
- High performance bandwidth ***and*** IOPS
- Stable performance for both Applications and Storage Services
- Reduced latency between application servers and storage
- Reduction in infrastructure and complexity
- Reduced number of individual storage systems required to scale capacity



FusionIO FLASH Memory & Networking

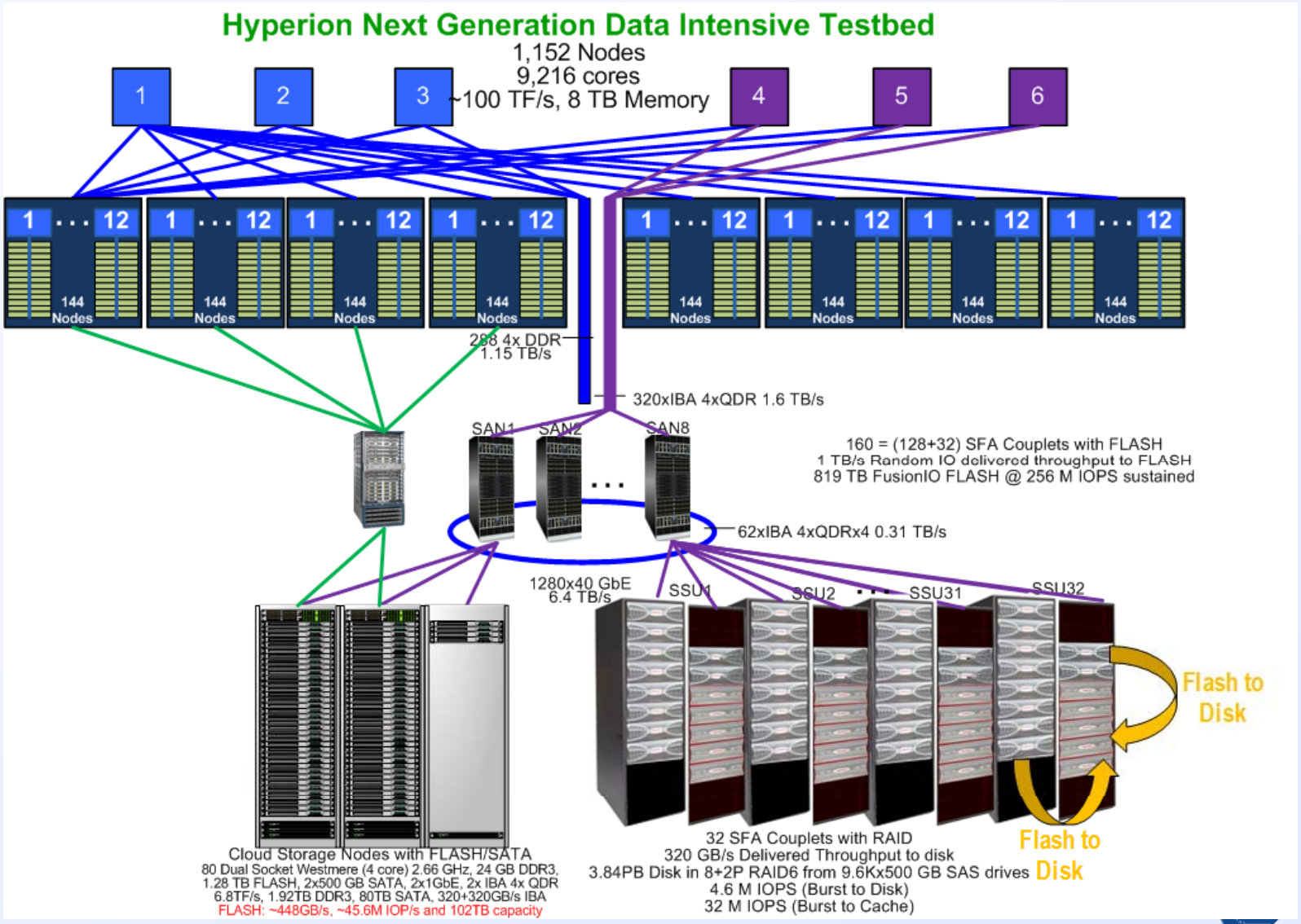
- ioSAN Card Physical
  - Full height/length PCIe x8 Gen2
  - 35 Watts nominal
- (2) ioMemory modules
  - 640 GB NAND FLASH Capacity
  - 1.6 GB/s Bandwidth
  - 200,000 IOPS
  - 30 us latency
- (2) 10 GigE / 20 GBps DDR Infiniband
  - 3.2 GB/s Bandwidth, 1 µs latency
  - CX-4 connectors
- (4) SAS / SATA Ports (Up to 200 HDD's)
  - Option in place of second ioMemory module
  - 200 TB HDD capacity
  - 800 MB/s Bandwidth 4,000 IOPS

Cloud Computing Testbed 1U node



- Dual socket Westmere (12 cores)
- 24 GB DDR3
- 4x PCIe2 x8 sots
- 2x ioSAN, 2x ioDUO
- 2x IBA 4x QDR
- 2x 10 Gb/s Eth
- 2x 1 Gb/s Eth
- 2x 500 SATA

Hyperion Next Generation DIT With Virtualized RAID/Application and FLASH is a Disruptive Technology for Multiple Applications Areas



Summary

- Sequoia project has made significant progress in the last year
  - Dawn delivering to the program
  - Sequoia development progressing toward prototype this summer and GO/NOGO in October 2010
  - TSF facilities 15MW ➔ 30 MW (to the computer) upgrade nearing completion
  - Site planning for Sequoia ready for construction
- Award winning Hyperion project is delivering results and will expand to include a data intensive testbed