Designing, Building and Using a 1.5 Million Core System

2012 Smoky Mountains Conference

September 5, 2012

Bronis R. de Supinski, CTO, Livermore Computing

Lawrence Livermore National Laboratory

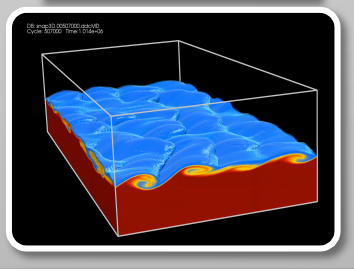
LLNL-PRES-579072

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Sequoia provides unprecedented levels of capability and concurrency

- Sequoia statistics
  - 20 petaFLOP/s peak
  - 16 petaFLOP/s LINPACK
  - Memory 1.5 PB, 4 PB/s bandwidth
  - 98,304 nodes
  - 1,572,864 cores
  - 3 PB/s link bandwidth
  - 60 TB/s bi-section bandwidth
  - 0.5−1.0 TB/s Lustre bandwidth
  - 50 PB disk
- 9.6MW power, 4,000 ft^2
- Third generation IBM BlueGene
- Challenges
  - Hardware scalability
  - Software scalability
  - Applications scalability

- Schedule
  - Deliveries completed on April 16 (very tight schedule)
  - Unclassified science as possible throughout 2012
  - Acceptance: September 2012
  - Classified: January 2013



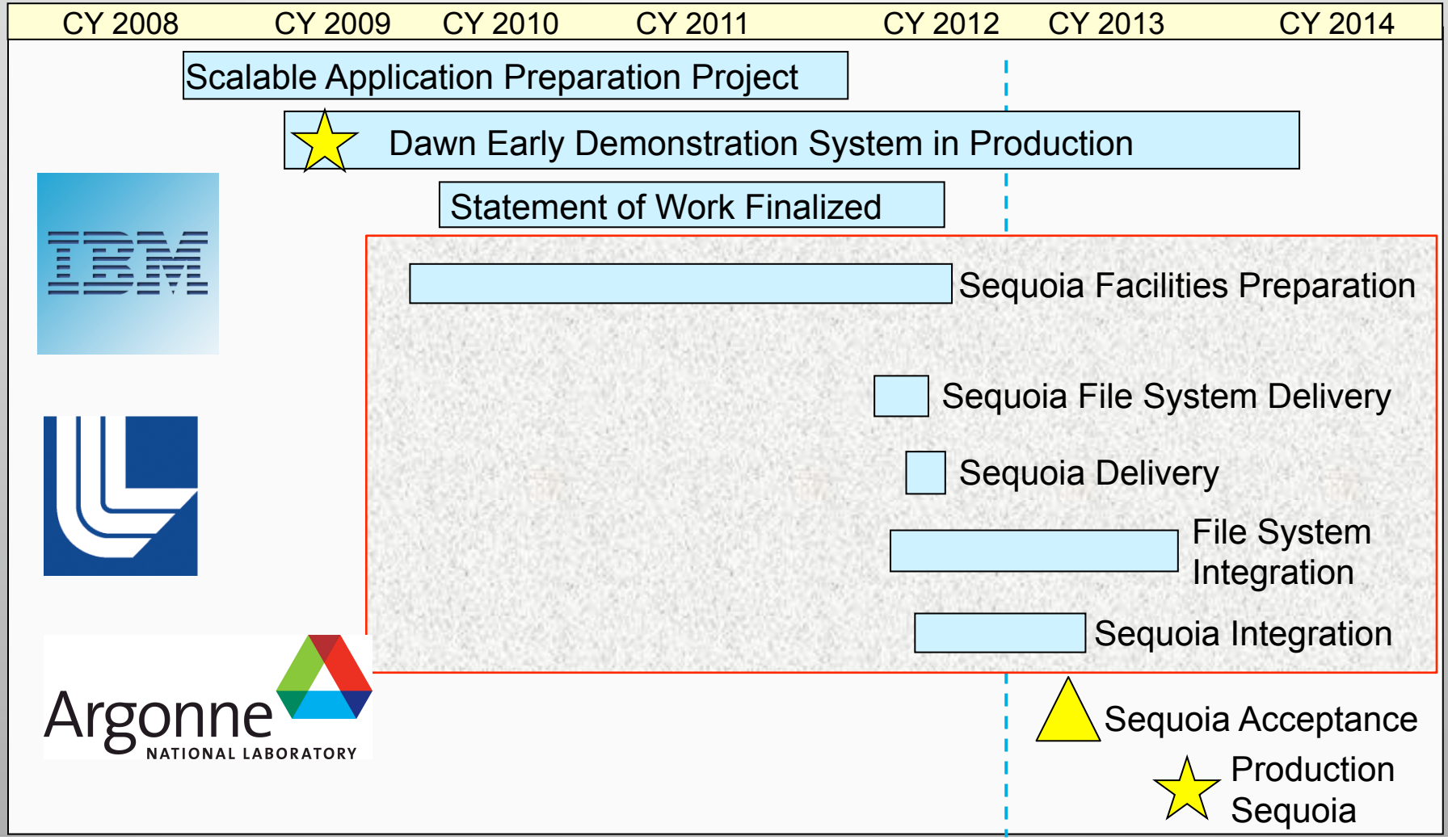
Acceptance Criteria:

UQ: Run 24 simultaneous Purple-class Integrated Design Code calculations

while also running…

Weapons science: at 4 PF sustained

It's a long road to a production 20PF machine



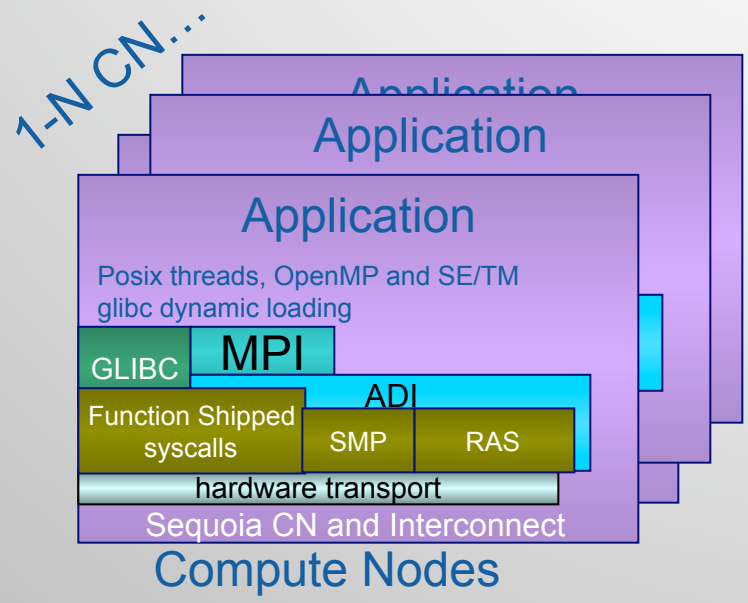
Sequoia installation is complete; acceptance testing is in progress

- 96 racks are running at Livermore
- 96-rack Linpack run with 81% efficiency 16.3PF/s
- Sequoia acceptance planned for September 2012
- Significant system software testing remains

Sequoia statistics:

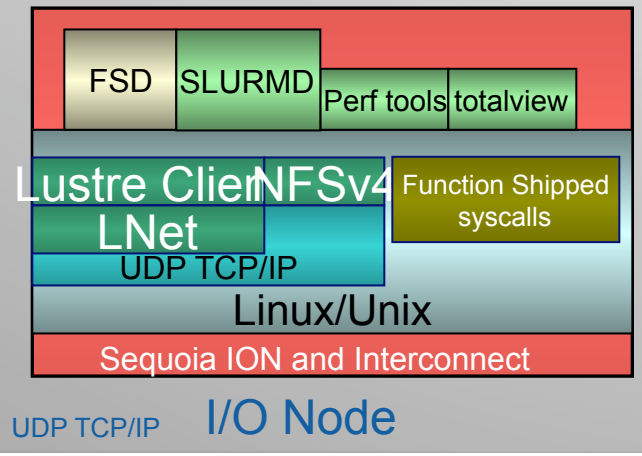
20 PF/s target
Memory 1.6 PB,
1.5M cores, 6M threads
50 PB disk, 512GB/s

Sequoia continues the Blue Gene system software strategy



Light weight kernel on compute node

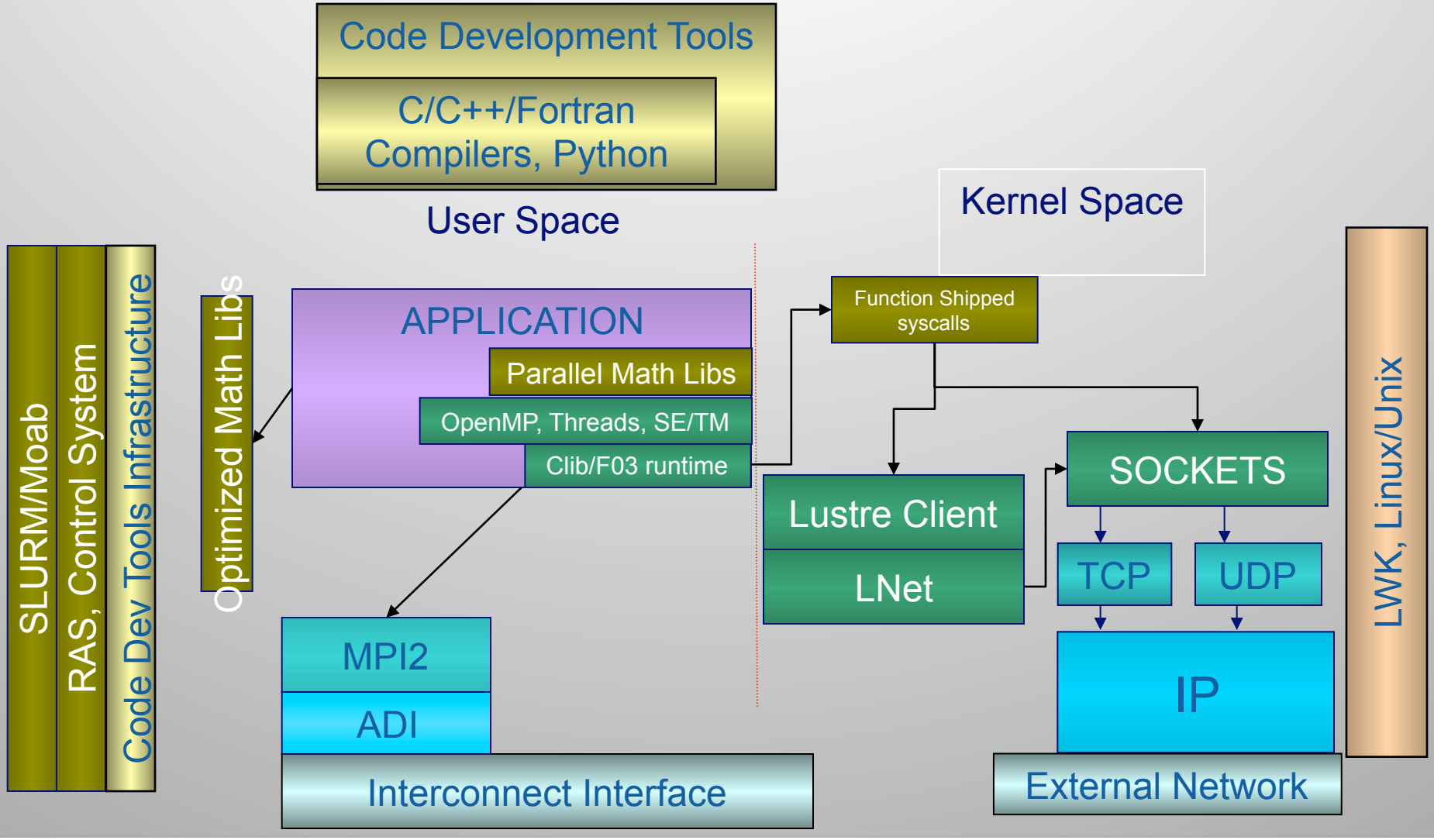
- Optimized for scalability and reliability
  - As simple as possible. Full control
  - Extremely low OS noise
  - Direct access to interconnect hardware
- OS features
  - Linux/Unix syscall compatible with IO syscalls forwarded to I/O nodes
  - Support for dynamic libs runtime loading
  - Shared memory regions
- Open source



Linux on I/O Node

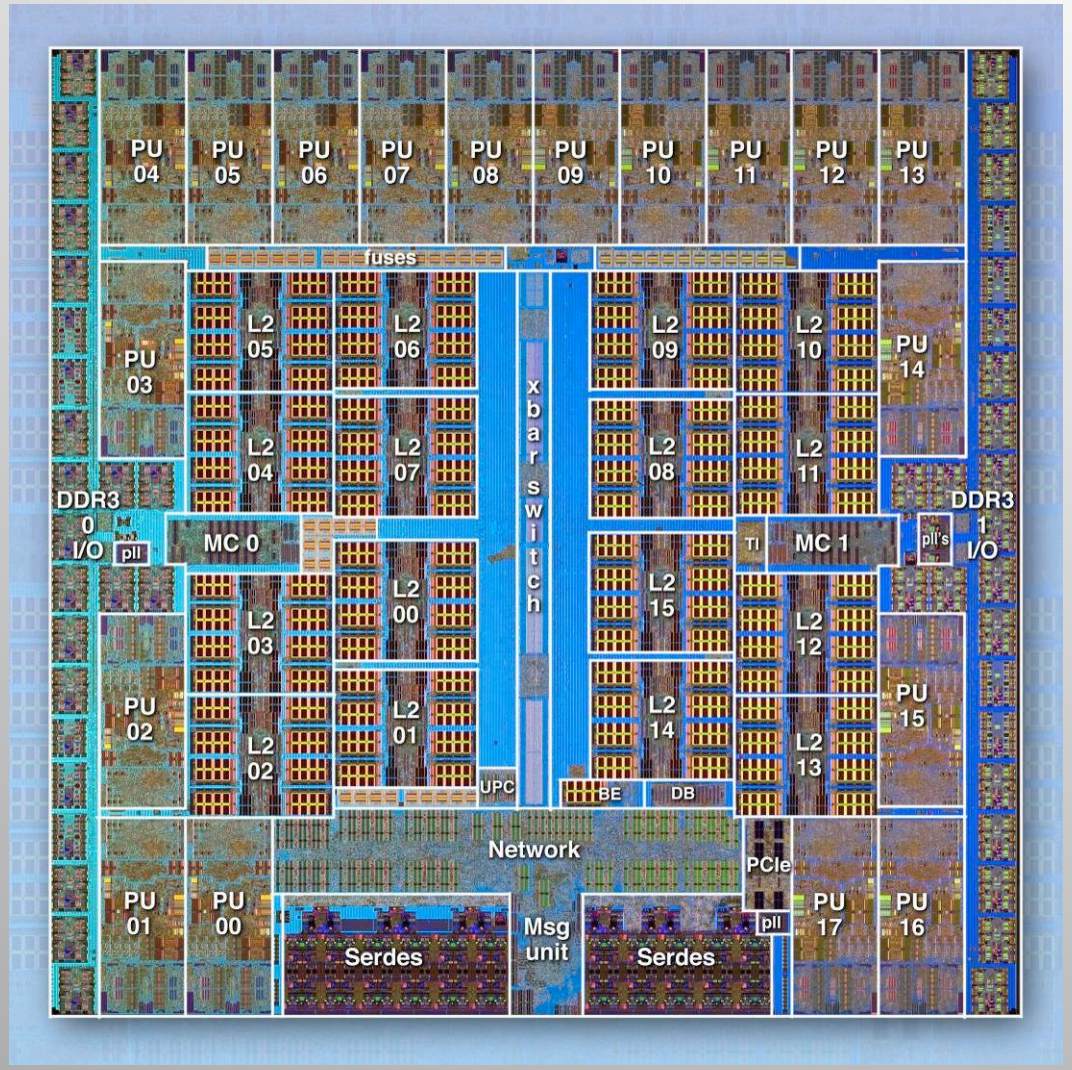
- Leverage huge Linux base and community
- Standard file systems (Lustre, NFSv4, etc.)
- Factor to simplify: Aggregates N CN for I/O & admin
- Open source

Sequoia presents application teams with a familiar development environment

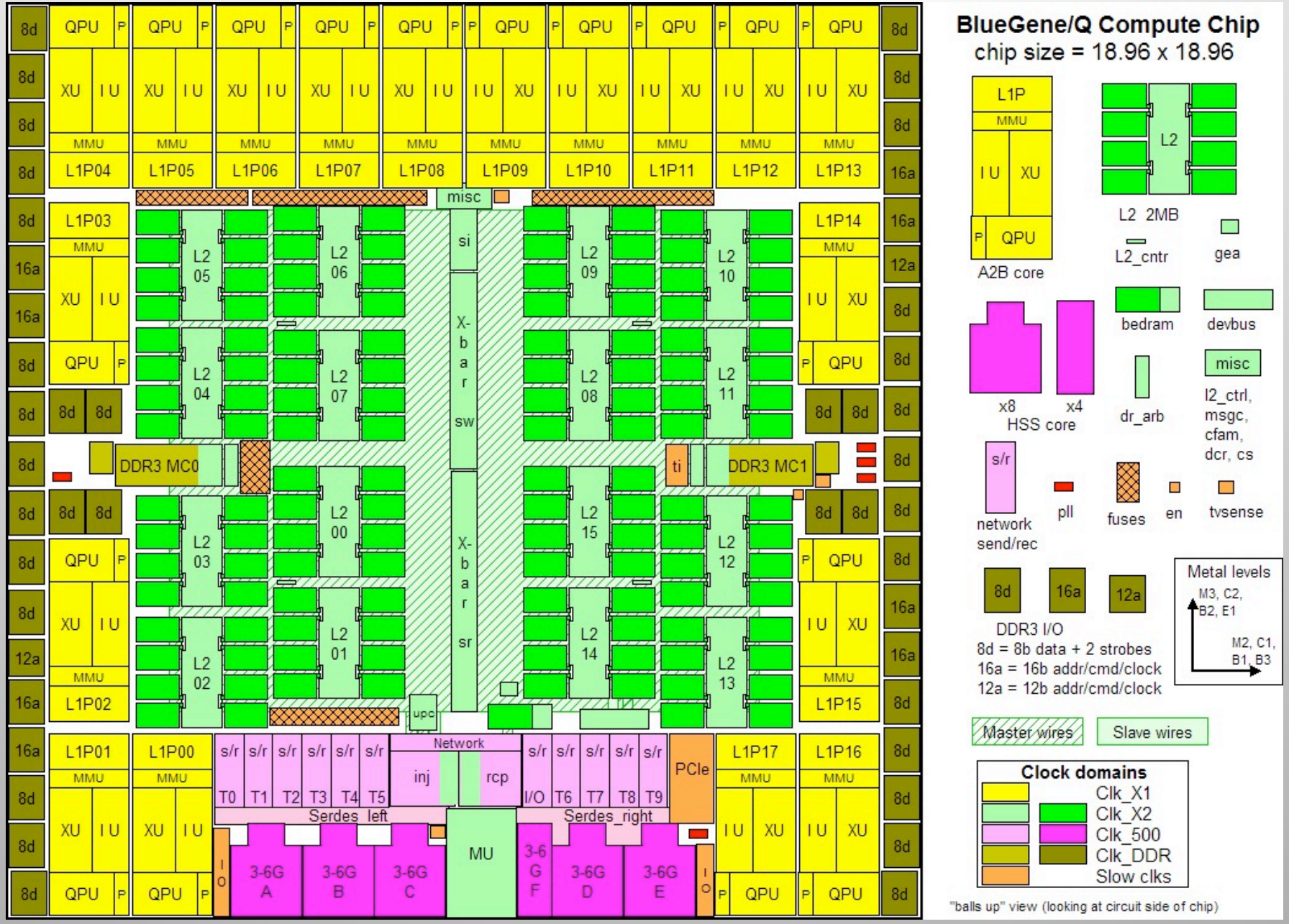


BlueGene/Q Compute chip

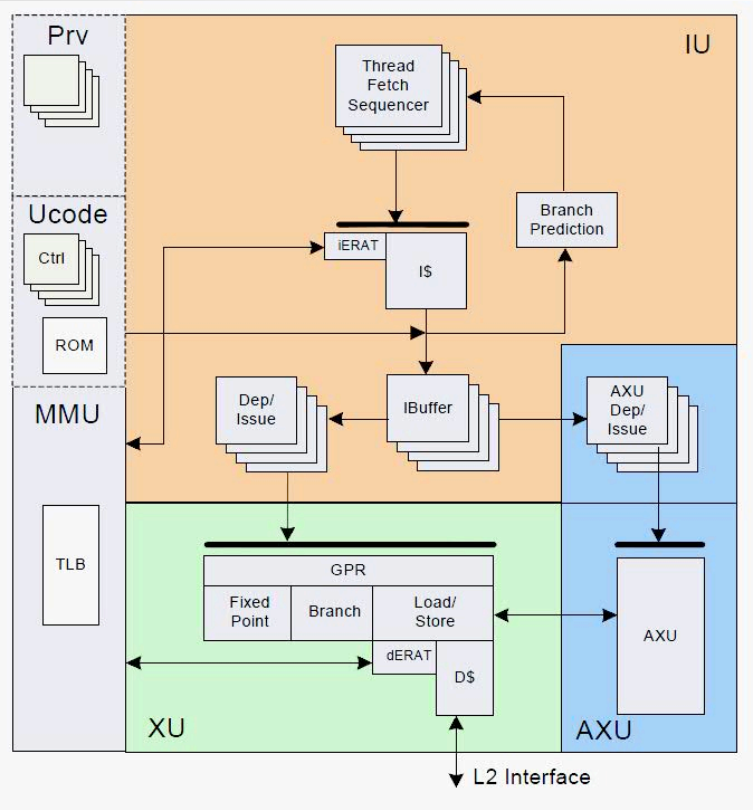
System-on-a-Chip design: integrates processors, memory and networking logic into a single chip



- **16 user + 1 service + 1 redundant cores**
  - Each 4-way multi-threaded, 1.6 GHz 64-bit PPC
  - L1 I/D cache = 16kB/16kB
  - L1 prefetch engines
  - Each has Quad FPU (4-wide double precision, SIMD)
  - Peak performance 204.8 GFLOPS @ 55 W
- **Central shared L2 cache: 32 MB**
  - eDRAM
  - Multiversioned cache, supports atomic ops
- **Dual memory controller**
  - 8–16 GB external DDR3 memory
  - 1.33 Gb/s
  - 2 * 16 byte-wide interface (+ECC)
- **Chip-to-chip networking**
  - 5D Torus topology + external link
    → 5 x 2 + 1 high speed serial links
  - Each 2 GB/s send + 2 GB/s receive
  - DMA, remote put/get, collective operations
- **External (file) IO -- when used as IO chip**
  - PCIe Gen2 x8 interface (4 GB/s Tx + 4 GB/s Rx)
  - Re-uses 2 serial links
  - Interface to Ethernet or Infiniband cards

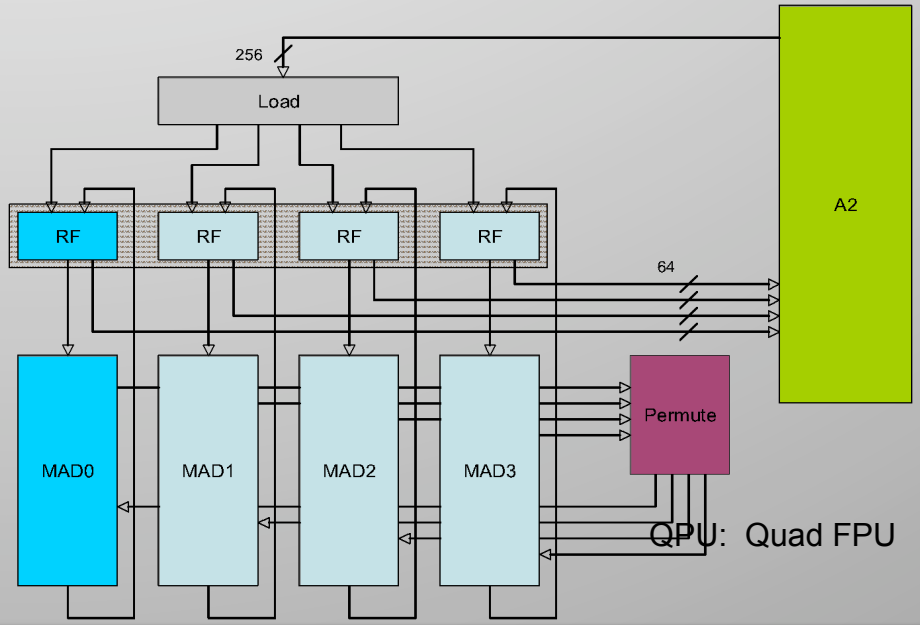

BlueGene/Q Compute Chip
chip size = 18.96 x 18.96
L1P
MMU
I U
XU
P
QPU
A2B core
L2
L2 2MB
L2_cntr
gea
bedram
devbus
misc
x8
x4
HSS core
dr_arb
l2_ctrl, msgc, cfam, dcr, cs
s/r
network send/rec
pll
fuses
en
tvsense
8d
16a
12a
DDR3 I/O
8d = 8b data + 2 strobes
16a = 16b addr/cmd/clock
12a = 12b addr/cmd/clock
Metal levels
M3, C2, B2, E1
M2, C1, B1, B3
Master wires
Slave wires
Clock domains
Clk_X1
Clk_X2
Clk_500
Clk_DDR
Slow clks
"balls up" view (looking at circuit side of chip)
L1P04
L1P05
L1P06
L1P07
L1P08
L1P09
L1P10
L1P11
L1P12
L1P13
L1P03
L1P14
L1P02
L1P15
L1P01
L1P00
L1P17
L1P16
L2 05
L2 06
L2 09
L2 10
L2 04
L2 07
L2 08
L2 11
L2 03
L2 00
L2 15
L2 12
L2 02
L2 01
L2 14
L2 13
misc
si
X-bar sw
X-bar sr
ti
upc
DDR3 MC0
DDR3 MC1
Network
inj
rcp
T0
T1
T2
T3
T4
T5
I/O
T6
T7
T8
T9
Serdes_left
Serdes_right
PCIe
MU
3-6G A
3-6G B
3-6G C
3-6G F
3-6G D
3-6G E
IO

BlueGene/Q Core



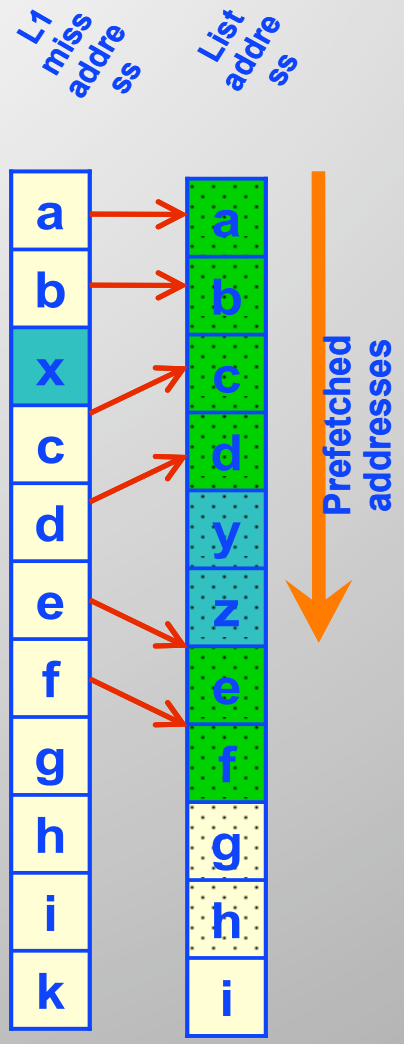
- **A2 processor core**
  - Implements 64-bit PowerISA™
  - Optimized for aggregate throughput:
    - 4-way simultaneously multi-threaded (SMT)
    - 2-way concurrent issue 1 XU (br/int/l/s) + 1 FPU
    - in-order dispatch, execution, completion
  - L1 I/D cache = 16kB/16kB
  - 32x4x64-bit GPR
  - Dynamic branch prediction
  - 1.6 GHz @ 0.8V

- **Quad FPU**
  - 4 double precision pipelines, usable as scalar FPU, 4-wide SIMD or 2-wide complex SIMD
  - 6 stage pipeline
  - 2W4R register file (2 * 2W2R) per pipe
  - 8 concurrent FLOPs (FMA) + load + store



Chip includes novel list-based prefetcher

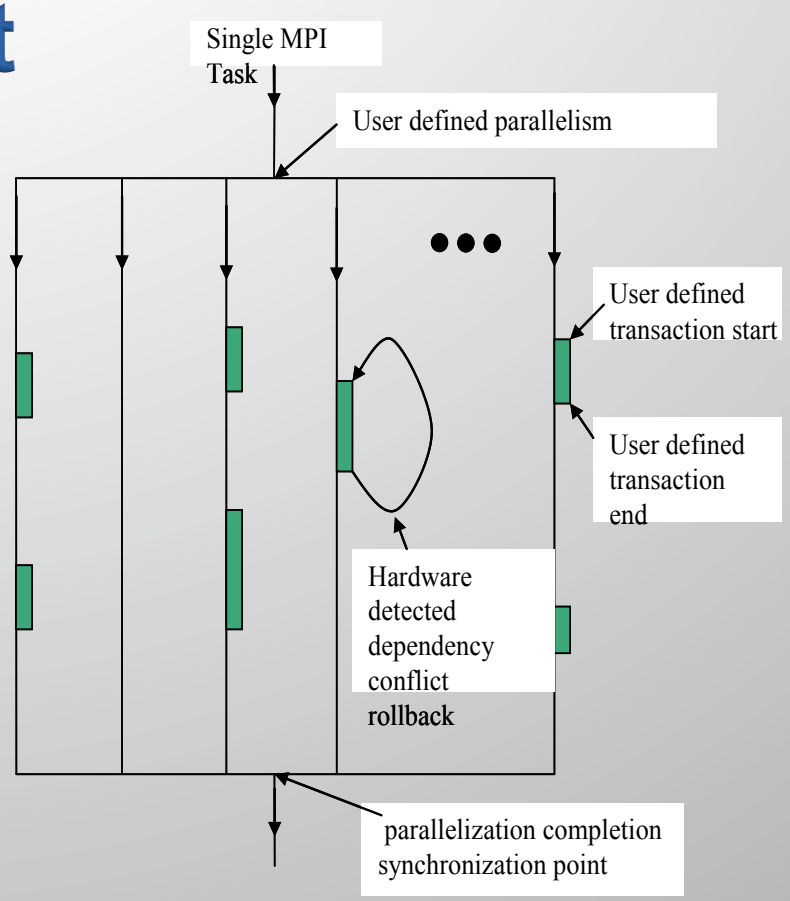
- Prefetches to depth of 8 L2 cache lines (128B)
- Normal mode: traditional stream prefetching
- Additional: 4 list-based prefetching engines
  - One per thread
  - Activated by program directives, e.g., bracketing complex set of loops
  - Used for repeated memory reference patterns in arbitrarily long code segments
  - Record pattern on first iteration of loop; playback for subsequent iterations
  - On subsequent passes, list is adaptively refined for missing or extra cache misses (async events)



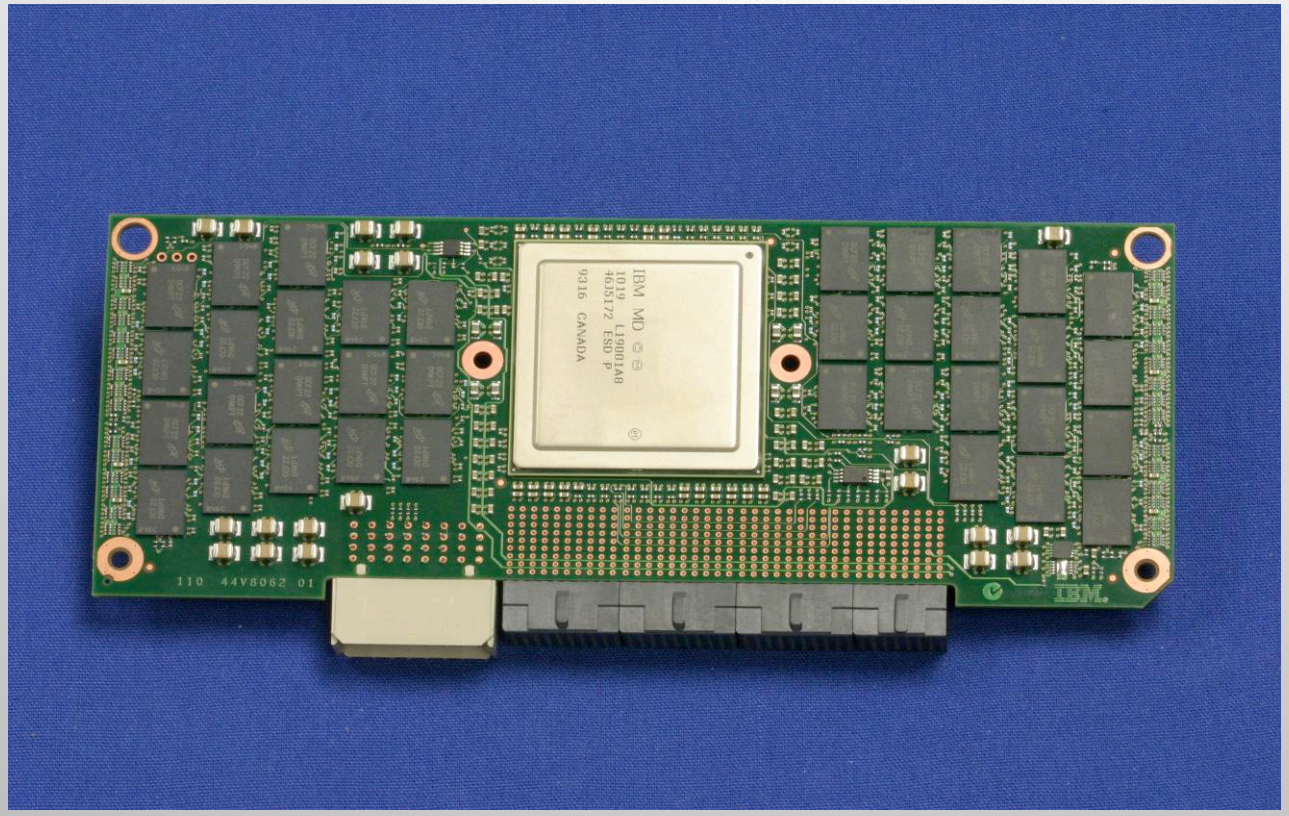
tolerance for missing or extra cache misses

L2 cache provides unique parallelization support

- Basics: point of coherency
  - 32 MB
  - 16 way set-associative
  - 128B line size
  - 16 slices @ 2MB each
    - 8 * 2.25 Mb eDRAMs (data+ECC)
    - Directory SRAMs
    - Buffers
    - Control logic.
- Multi-versioned
  - Data tagged, tracked by score board
  - Used for transactional memory
    - Guarantees “atomicity” of user-defined transactions
    - Eliminates need for locks
  - Also used for speculative execution (ordered transactions)



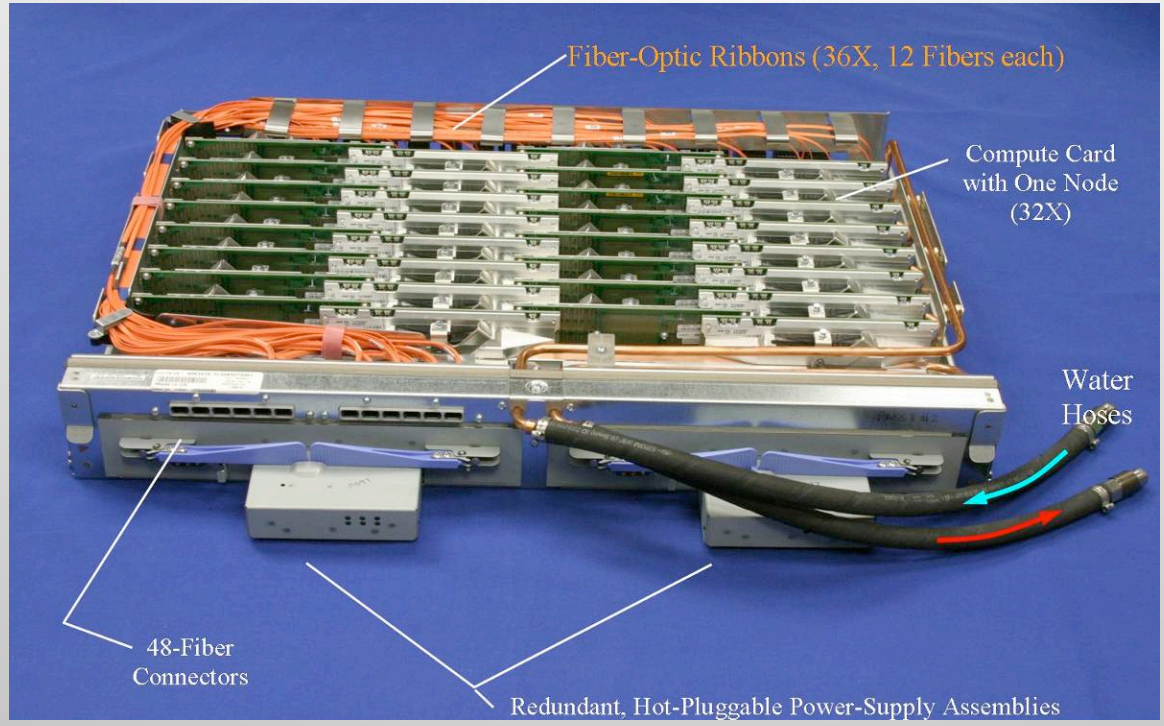
Blue Gene/Q compute card is the basic field replaceable unit



- Compute Card has 1 BQC chip + 72 SDRAMs (16GB DDR3)
- Two heat sink options: water-cooled → "Compute Node" / air-cooled → "IO Node"
- Connectors carry power supplies, JTAG etc, and 176 HSS signals (4 and 5 Gbps)

Courtesy of IBM

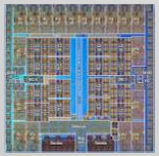
Blue Gene/Q node card assembly provides dense packaging



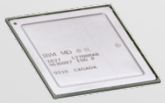
- High bandwidth, low latency electrical interconnect on-board
- 18+18 (Tx+Rx) 12-channel optical fibers @10Gb/s
  - 8*48-channel fibers for rack-to-rack (Torus) and 4*12 for Compute-to-IO interconnect
- Compute Node Card assembly is water-cooled (18-25ºC – above dew point)
- Redundant power supplies with distributed back-end ~ 2.5 kW

Tradtitional BlueGene overall system integration results in small footprint

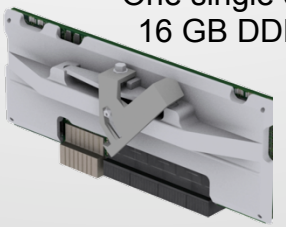
1. Chip
16 cores



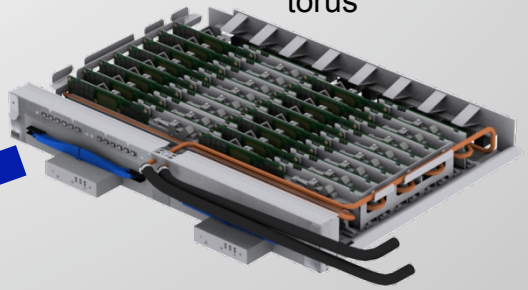
2. Module
Single chip



3. Compute card
One single chip module,
16 GB DDR3 memory



4. Node card
32 compute cards,
Optical modules, link chips,
torus

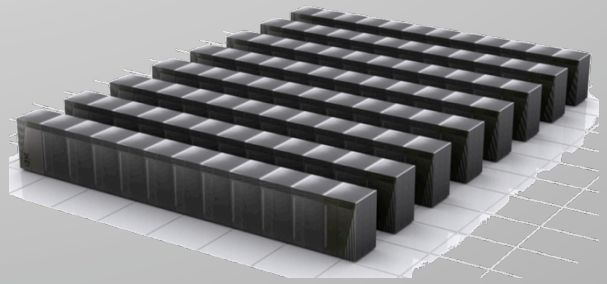


5b. I/O drawer
8 I/O cards
8 PCIe Gen2 slots

5a. Midplane
16 node cards

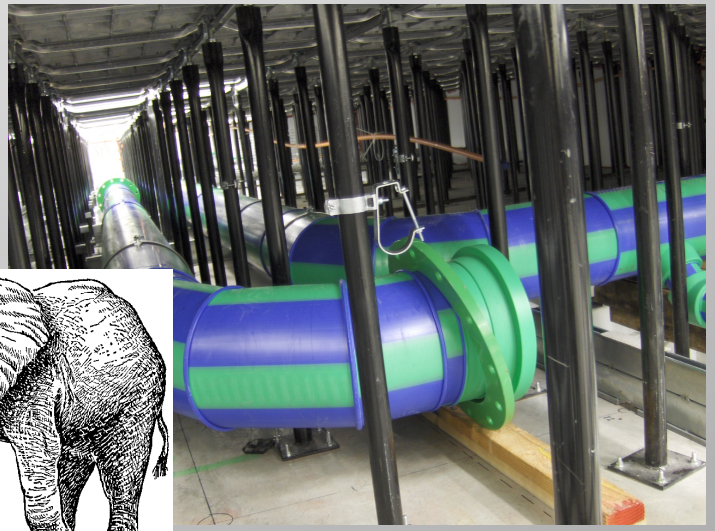
6. Rack
2 midplanes
1, 2, or 4 I/O drawers

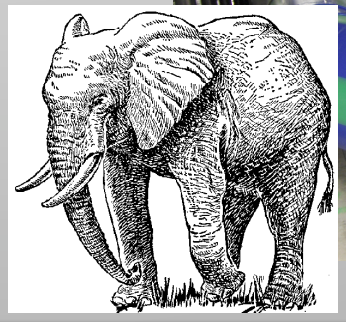
7. System
20 PF/s



Facilities work was a grand challenge

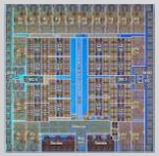
- Custom designed receptacles
  - Reduced under-floor congestion
  - All 4 feet of subfloor are packed
- Mechanical room requirements:
  91% liquid cooled and 9% air cooled
  - New tertiary CHW loop needed to meet requirements
  - Air-cooling requirements:
    - 1700 CFM/rack= 163,000 CFM total
- Electrical requirements:
  100 kW/rack = 9.6MW
- Physical requirements
  - 96 racks in 4,000 ft^2
  - 4,500 lbs/racks = 210 tons total
    - Approximately equivalent to 30 adults elephants



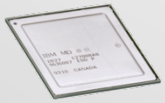


We have encountered integration issues but still moving forward

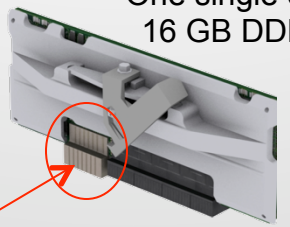
1. Chip
16 cores



2. Module
Single chip

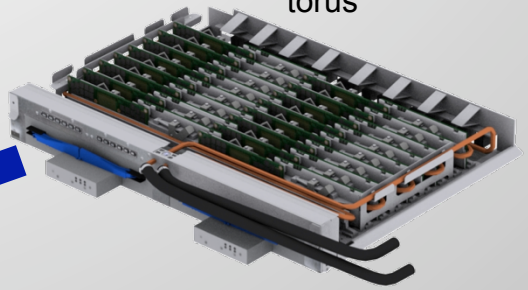


3. Compute card
One single chip module,
16 GB DDR3 memory



Bent pins on node cards

4. Node card
32 compute cards,
Optical modules, link chips, torus

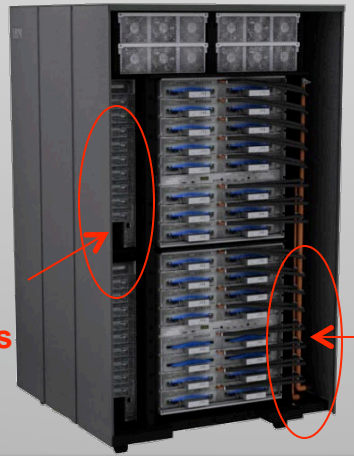


5b. I/O drawer
8 I/O cards
8 PCIe Gen2 slots

5a. Midplane
16 node cards

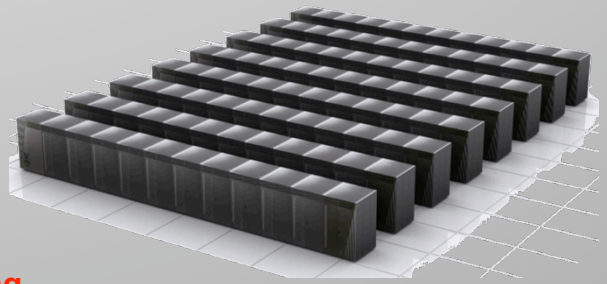
Bulk power module gaskets replaced

6. Rack
2 midplanes
1, 2, or 4 I/O drawers



Cooling monitor issues

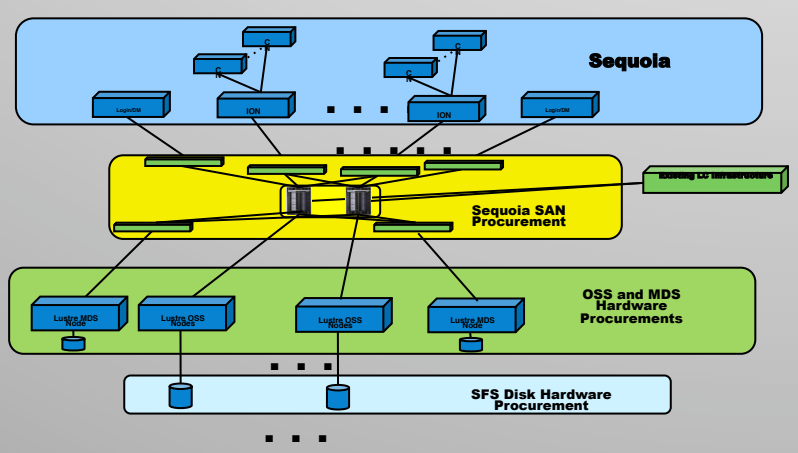
7. System
20 PF/s



Sequoia file system is massive; risk mitigation is that hardware was proven when we signed the contract

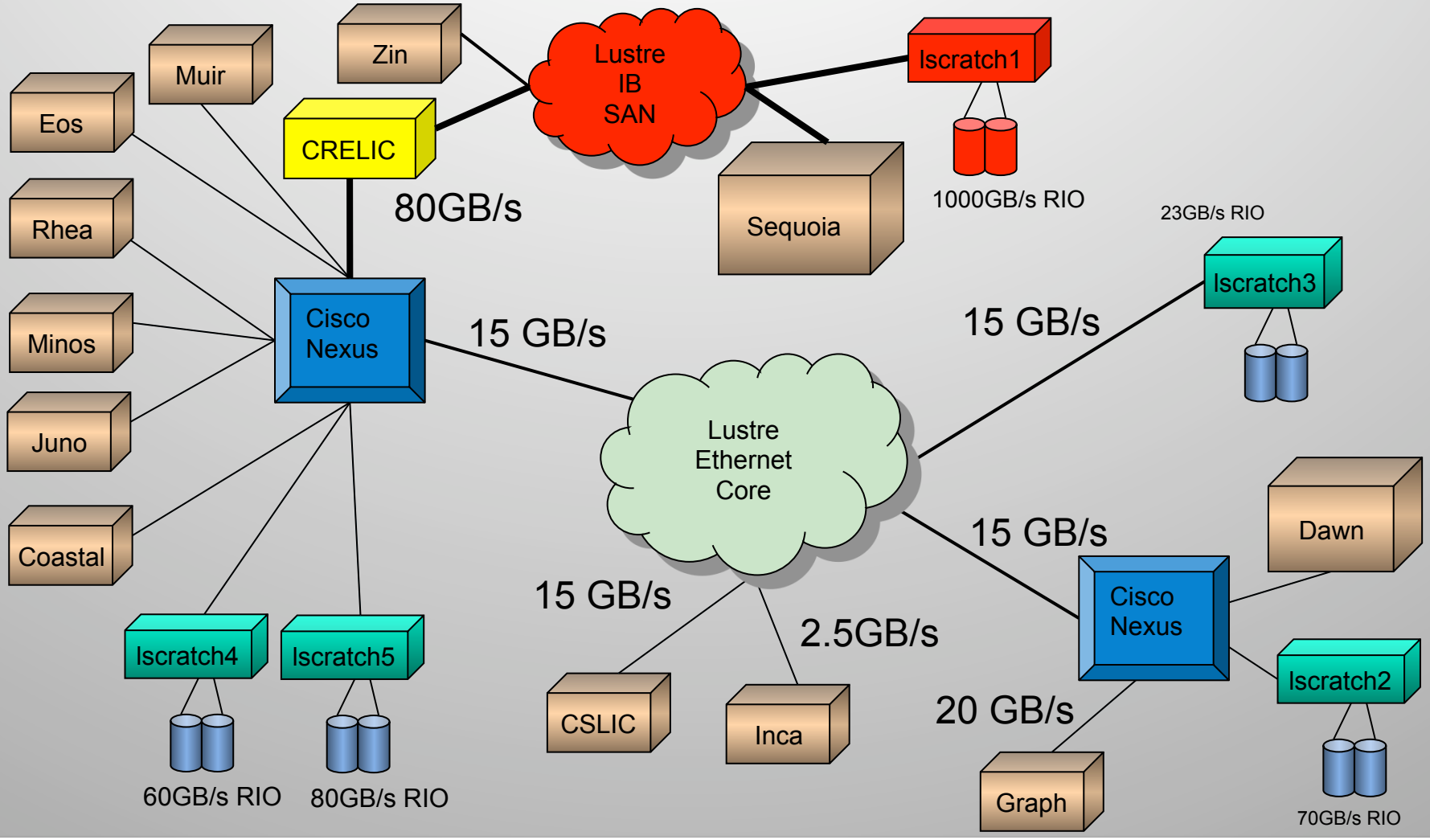
- Requirements
  - 50PB file system
  - 500GB/s minimum, 1TB/s stretch goal
  - QDR InfiniBand SAN connection to Sequoia—
    768 IB links into core
  - Must integrate with existing Ethernet infrastructure

- 384 Netapp E5400s
  - 60 3TB SAS drives per bay, 4U
  - Dual RAID controllers



- 768 Appro GreenBlade OSS Nodes
  - Intel SandyBridge, TLCC2
  - Dual socket, 8 core @2.6GHz

Sequoia file system is bridged to existing Ethernet infrastructure



Facility and file system work was simultaneous; we worked smartly to coexist peacefully

Sequoia target is P + S ≥ 40

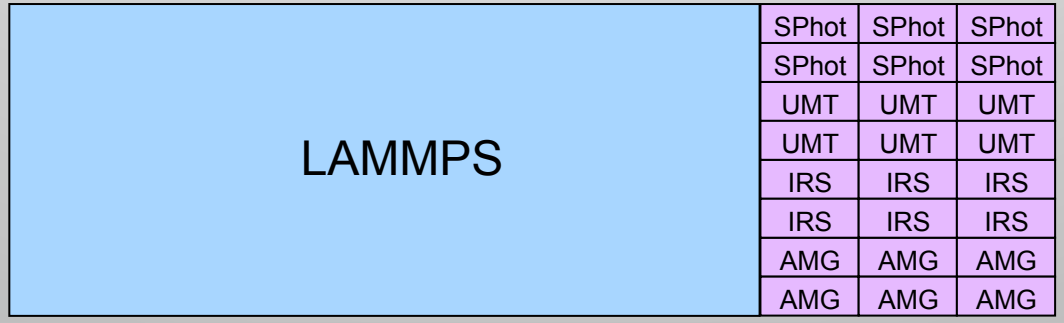
- P is "peak" of the machine measured in petaFLIN/s
  - FLIN/s = FLoating point INstructions per second
  - More accurate performance predictive metric than FLOP/s for ASC "Integrated Design Codes" (IDC)
- S is weighted average of five "marquee" benchmarks
  - Four IDC package benchmarks
    - UMT, IRS, AMG, and Sphot
    - Program goal is 24x Purple capability throughput
  - One "science workload" benchmark from Sandia: LAMMPS
    - Program goal is 20x-50x BGL for science capability workload

Ninety -six Sequoia racks powered up as of April 26, 2012.

Acceptance testing ensures machine meets functionality, performance and stability requirements

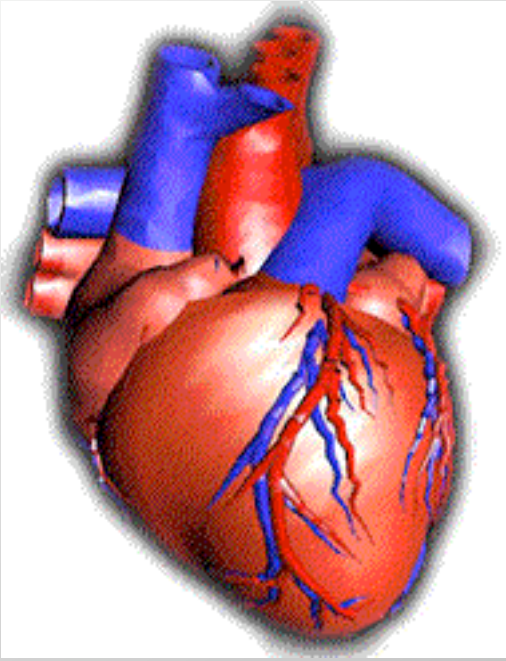
- 7 full days of Synthetic Workload testing that includes performance, functionality, and stability tests
  - MTBF is measured during the test
- Sustained performance will be measured on Sequoia CNs with simultaneous problem execution
  - FOM calculations only time the "compute" portion of the run
  - The sustained, aggregate measurement runs for four hours

$$FOM_{LAMMPS} + 6*(FOM_{UMT} + FOM_{IRS} + FOM_{AMG} + FOM_{SPhot}) \geq 2.0e+16*$$



This benchmarking strategy ensures that Sequoia will deliver *both* UQ and science to the Stockpile Stewardship Program

Modeling cardiac electrophysiology on the Sequoia



- Aim is to understand mechanisms that lead to arrhythmia so that mitigation strategies can be applied
  - Contraction of heart is controlled by electrical behavior within and across heart cells
- LLNL, with IBM, has developed cardiac electrophysiology code (*Cardioid*) designed to take advantage of Sequoia
- Cardioid, when run on full Sequoia, is several hundred times faster than other known 2011-published HPC efforts
  - Higher resolution enables more detailed representation of infarcts
  - Ability to run thousands instead of tens of heart beats enables detailed study of drug effects

Cardioid, a 2012 Gordon Bell finalist, is achieving outstanding performance

- Replace time-consuming computations (involving exponentials, logs) with validated rational function approximates
- Make effective use of registers and cache
- Apply different cores to different phases of computation
- Apply hardware threading (64 threads per node) and SIMDization
- Use low-level communications library (SPI) instead of MPI
- Use specially-designed fast barrier in place of OpenMP synchronization
- Resulting code is 50 times faster per CPU
- Code weak-scales perfectly through 48k nodes

Planning, innovation and hard work have been key elements of the Sequoia journey

- It has been a long road, and we aren't "there" yet
- Sequoia's success will be measured by the simulation and science achievements it enables
  - Cardoid already achieving scientific advances
  - Marquee benchmarks are performing well
    - LAMMPS job of over 1 million MPI processes ran successfully
      - 16-rack BG/Q: 16*1024*16*4=1,048,576 MPI processes (64 per node)
      - Scaling was nearly perfect: time spent in communication increased 0.5 seconds out of 65.2 seconds as code scaled from 1 to 16 racks!
  - Initial ASC code performance also scales excellently

We are on track to deliver two major BG/Q systems in 2012

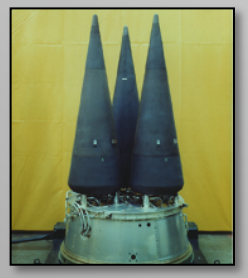
Classified: Sequoia 20PF

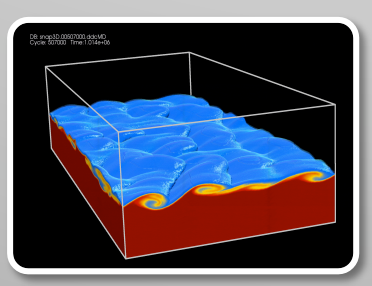
Acceptance criteria relate to the machine's mission:

UQ ➔ Run 24 simultaneous Purple-class Integrated Design Code (IDC) calculations

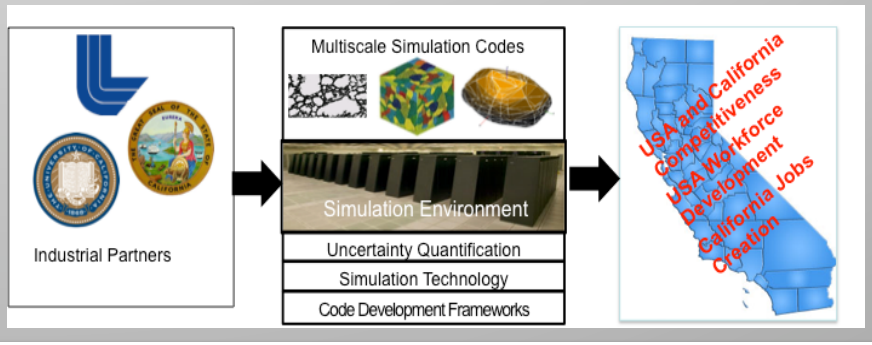
while also running….

Weapons science ➔ at 4 PF sustained





Unclassified Collaboration Zone: 5-10PF

Mission ➔ institutional computing, ASC Alliances, WCI codesign & industrial partners

- Baseline change ➔ Don Cook & Bob Meisner
- Encouragement for our industrial partnership strategy within HQ and LSO