Sequoia Sets New FLOP/s:Watt Standard

Salishan 2009

Mark Seager
Lawrence Livermore National Laboratory

Version 1
28 April 2009

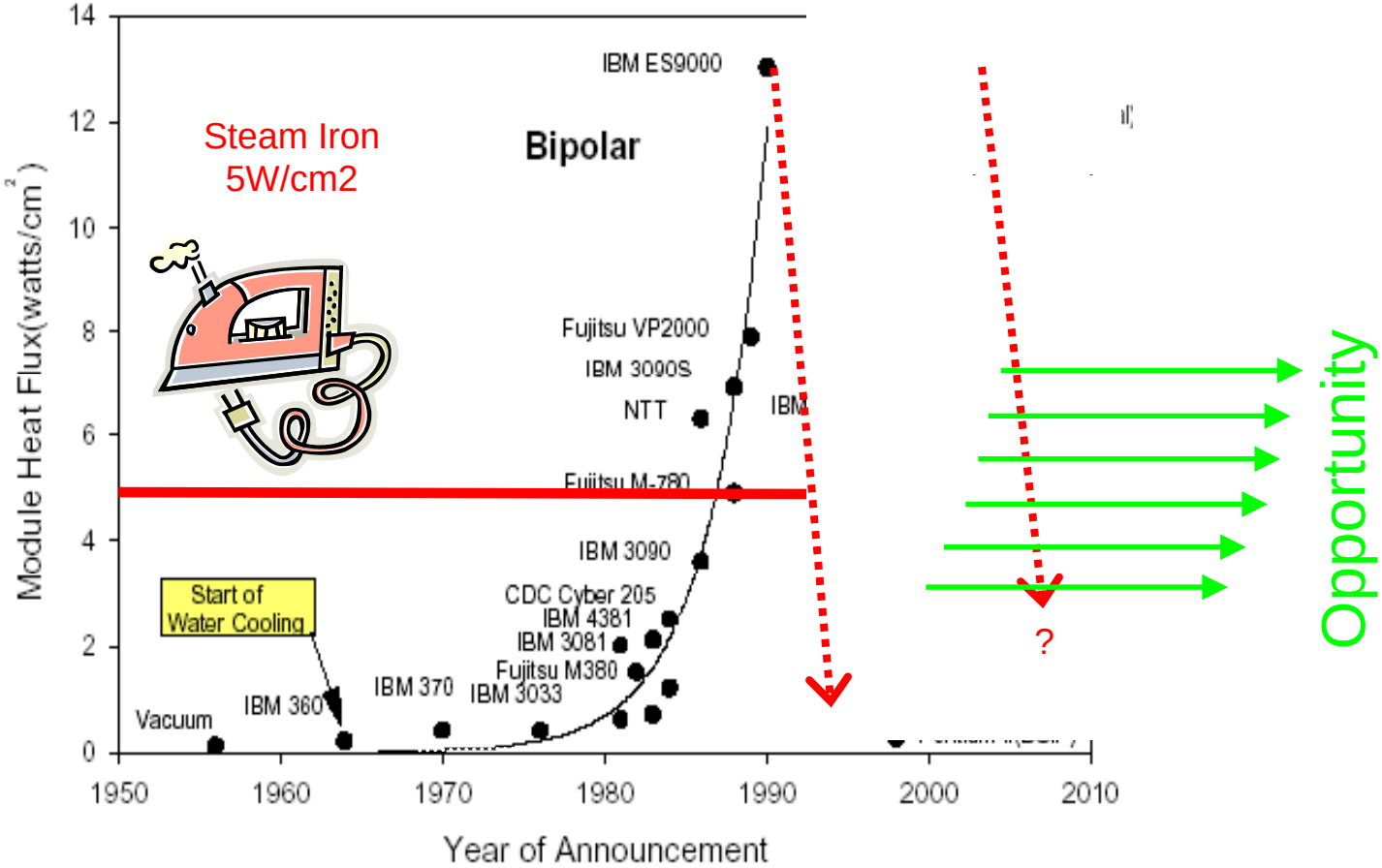
LLNL-PRES-TBD
This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Overview

- The case for a low power approach
- High level Sequoia programmatic drivers
- A high level overview of the Sequoia target architecture and multi-petascale applications strategy
- Applications work has already catalyzed new ways of thinking about parallelism and applications development model changes

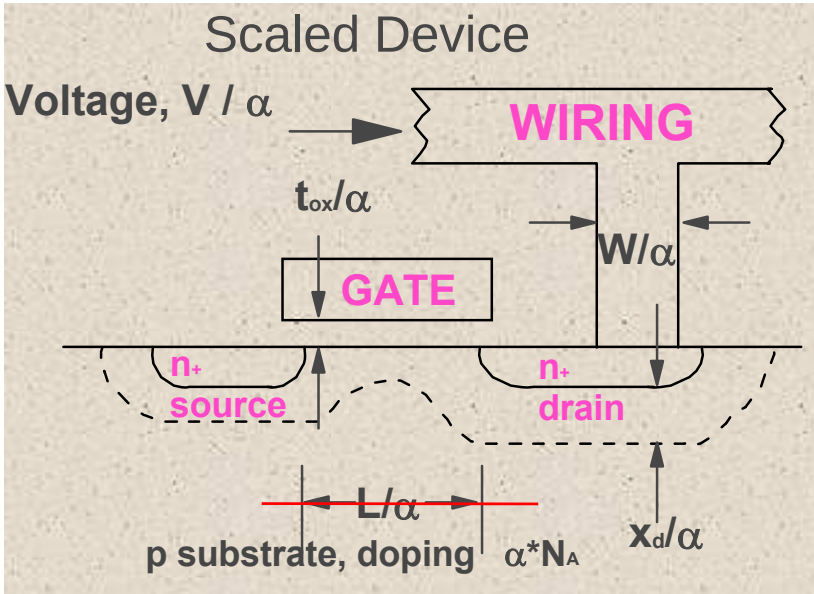
System power is again THE problem

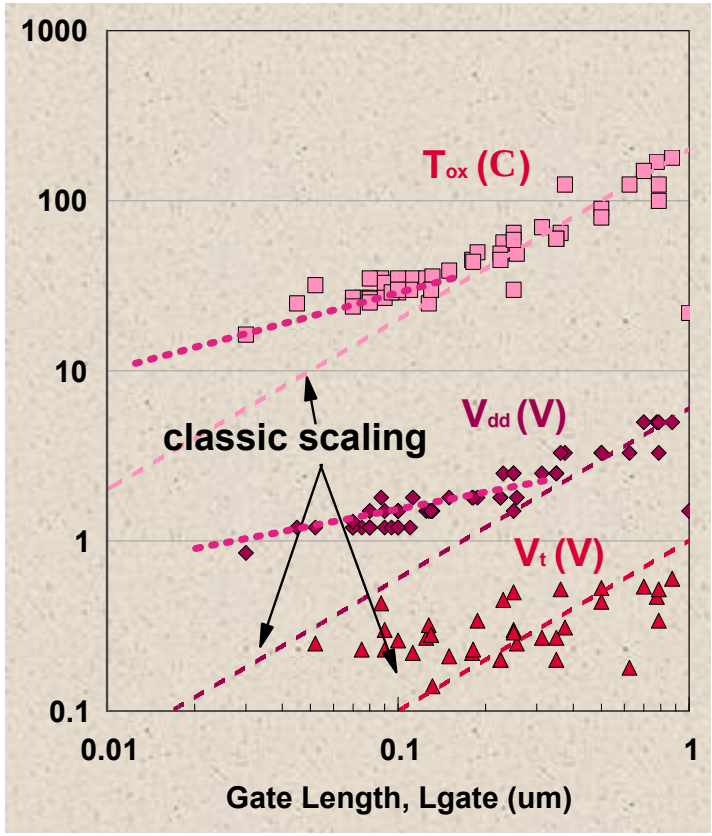
Single thread focus has resulted in power inefficient design



Source: IBM Research

How does one know that classical CMOS scaling is really dead?





SCALING:

Voltage:	**V/α**
Oxide:	t_{ox}/α
Wire width:	W/α
Gate width:	L/α
Diffusion:	x_d/α
Substrate:	$\alpha * N_A$

RESULTS:

Higher Density:	$\sim\alpha^2$
Higher Speed:	$\sim\alpha$
Power/ckt:	$\sim 1/\alpha^2$
Power Density:	**~Constant**

- Why deviate from "ideal" scaling
  - unacceptable gate leakage/reliability
  - additional performance at higher voltages
- What is the consequence of this deviation?
  - **a dramatic rise in power density**

Source: IBM Research

In the face of adverse changes in the base silicon technology, BlueGene dramatically improves power efficiency

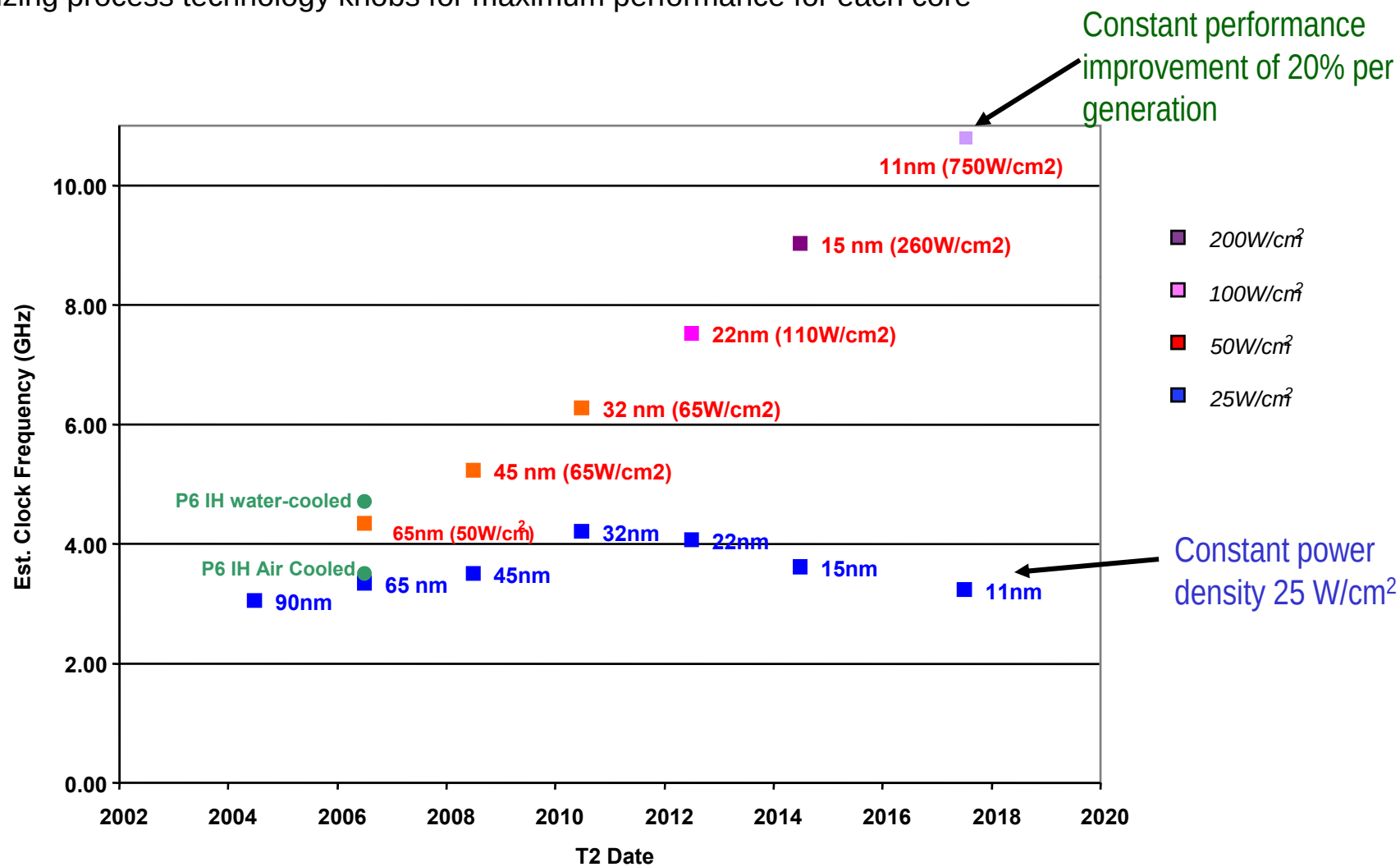
	Rack TFLOP /s	Width (ft)	Depth (ft)	Height (ft)	Density GFLOP/s ft^3	Linpack Perf / Watt (MFlops/Watt)
Earth Simulator	0.12	3.2	4.5	6.5	1.33	2.7
Purple	0.73	4	6	7	4.34	15.7
Dell Xanadu Nehalem	5.5	2	4	6	115	181.
Blue Gene/L	5.7	3	3	7	91	208.
Blue Gene/P	13.9	4	3	7	166	371.
Sequoia	209.7	4	3	7	2,497	2000.§

General Approach to Low Power

- High level of integration
  - fewer parts each with significant power overhead, also impacts RAS
- Focus on simplicity and address power-performance trade-offs at all levels
  - Simple power efficient processing core.
  - SIMD floating point to exploit data reuse optimally from a power perspective
  - Utilize concurrency to avoid climbing non-linear power-performance curve.
  - Large low power on chip cache based on eDRAM
  - Glueless network design to reduce chip count and increase scalability (no layer-like limits , predictable, repeatable performance
  - On-board memory controller, direct attached commodity lowest power memory
  - Power distribution focus on reducing system droop allowing for lower supply voltage.
  - Water cooled design (Sequoia) holds temperatures lower (reduces CMOS leakage currents), increases compute density (allows for shorter signaling distances requiring lower power) and improves supply efficiency.

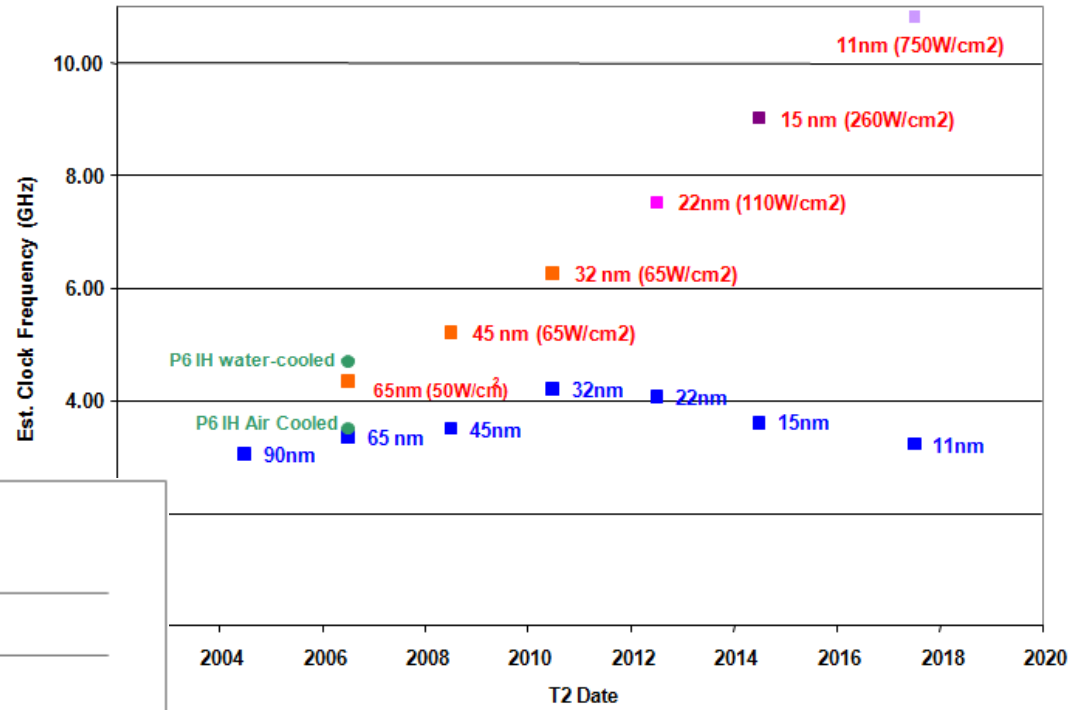
IBM Research PDSOI optimization results indicate lower power approaches provide better power efficiency

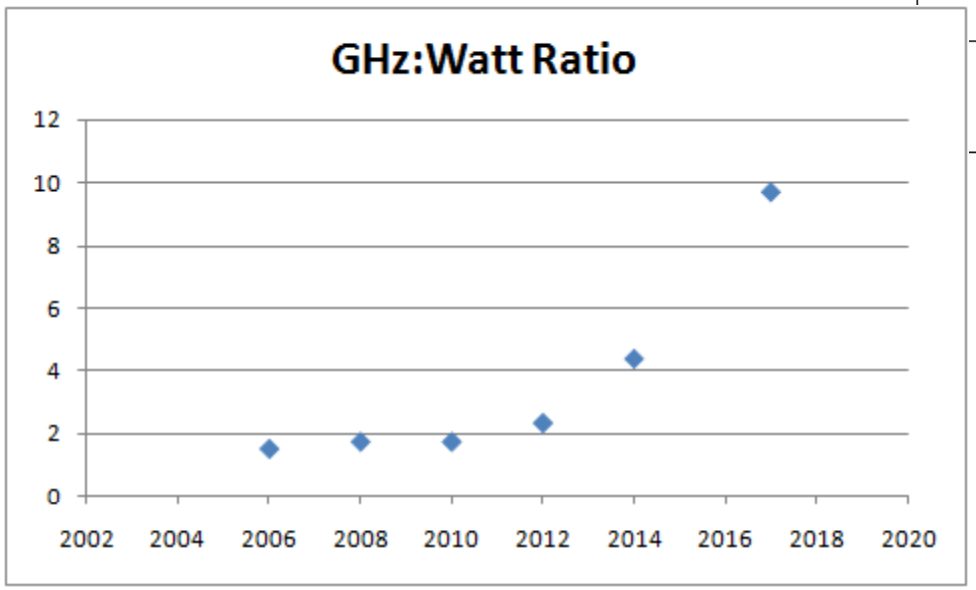
Optimizing process technology knobs for maximum performance for each core



D. Frank, C. Tyberg
Source: IBM Research

High power processors are even less attractive for HPC systems in the future…





For a given silicon process technology and target performance level, the maximum frequency designs will yield dramatically higher power require-ments than low power designs.

Source: IBM Research

Sequoia will be a key simulation tool for keystones and uncertainty quantification for stockpile stewardship

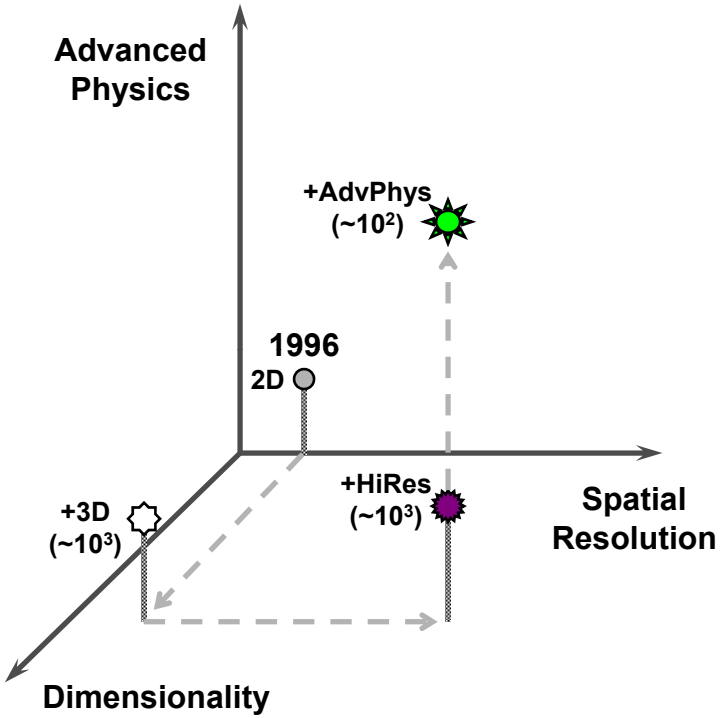
- ***ASC Strategy* and *ASC Roadmap* provide a vision for and keystones leading to "predictive simulation" or prediction with quantified uncertainties**

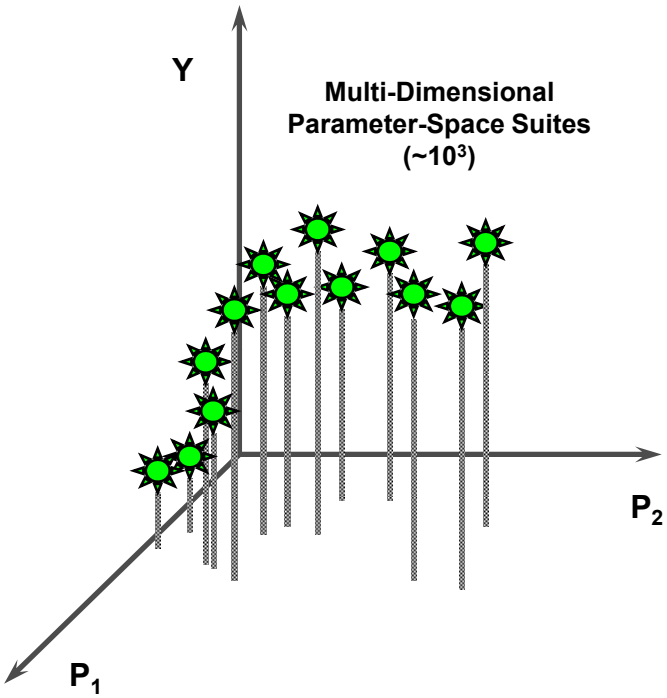
- ***Thermonuclear Burn Initiative*, *National Boost Initiative* and *Predictive Capability Framework* represent Stockpile Stewardship Program (SSP) planning to coordinate on the key issues impeding predictive simulation**

- **Sequoia is intended to address requirements coming from this planning in the period between 2012 - 2017, focusing on UQ and materials science, related to boost and certification**

- **To demonstrate it can meet these objectives, Sequoia will:**
  1. **Achieve 12X-24X Purple throughput for integrated weapons calculations related to Uncertainty Quantification (stretch goal >> 24X)**
  2. **Achieve 20X BG/L (stretch goal 50X) on a science materials effort**
  3. **Single RFP mandatory was Peak + Sustained ≥ 40**

Predictive simulation requires hero runs, but also large ensemble of calculations



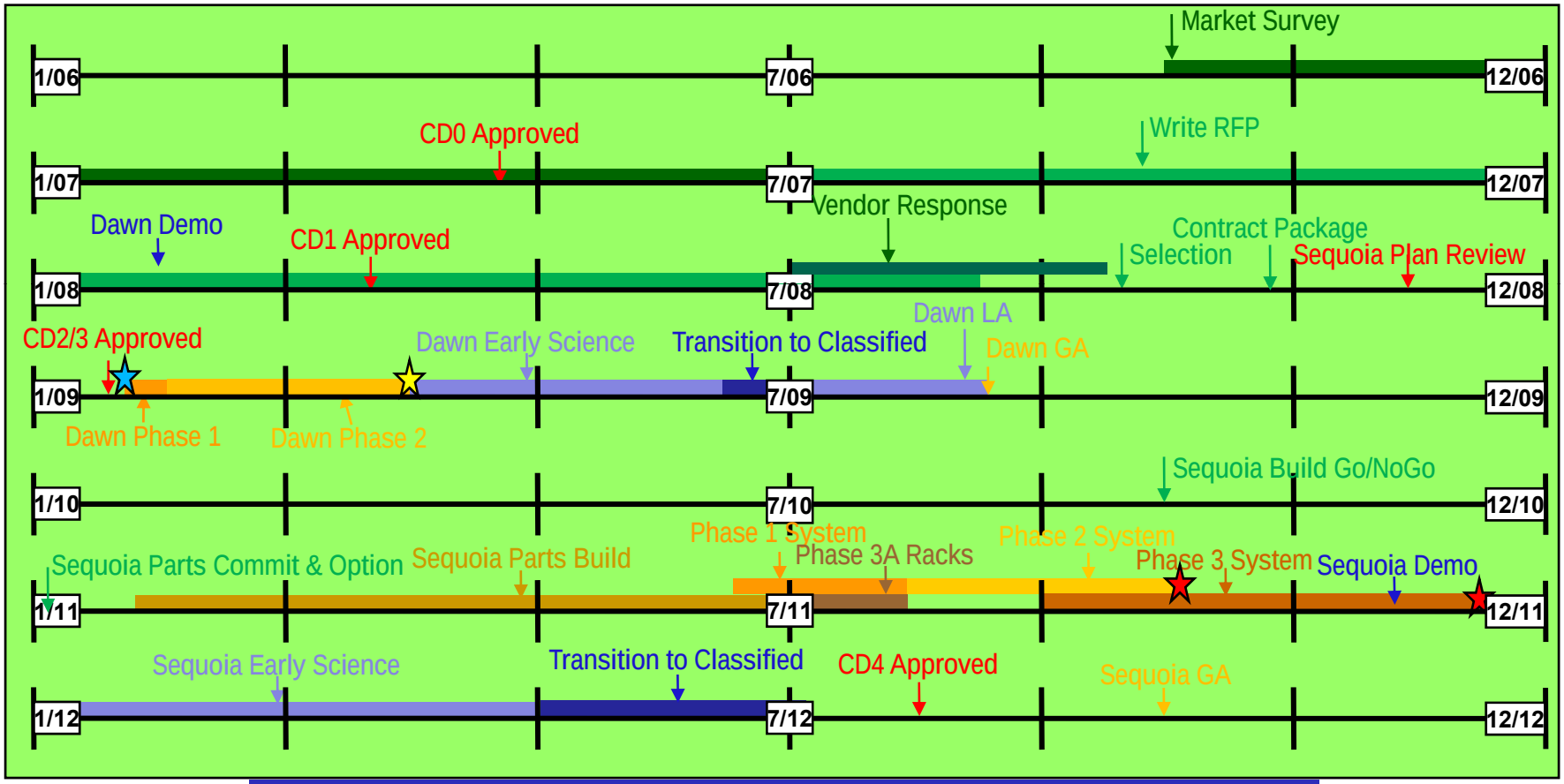


A path to a single 3D high-resolution high-fidelity calculation

Assessing the uncertainty in our predictions requires a suite of 3D high-resolution high-fidelity calculations

Sequoia will initiate the transition to 3D UQ suites with advanced models for energy balance and boost

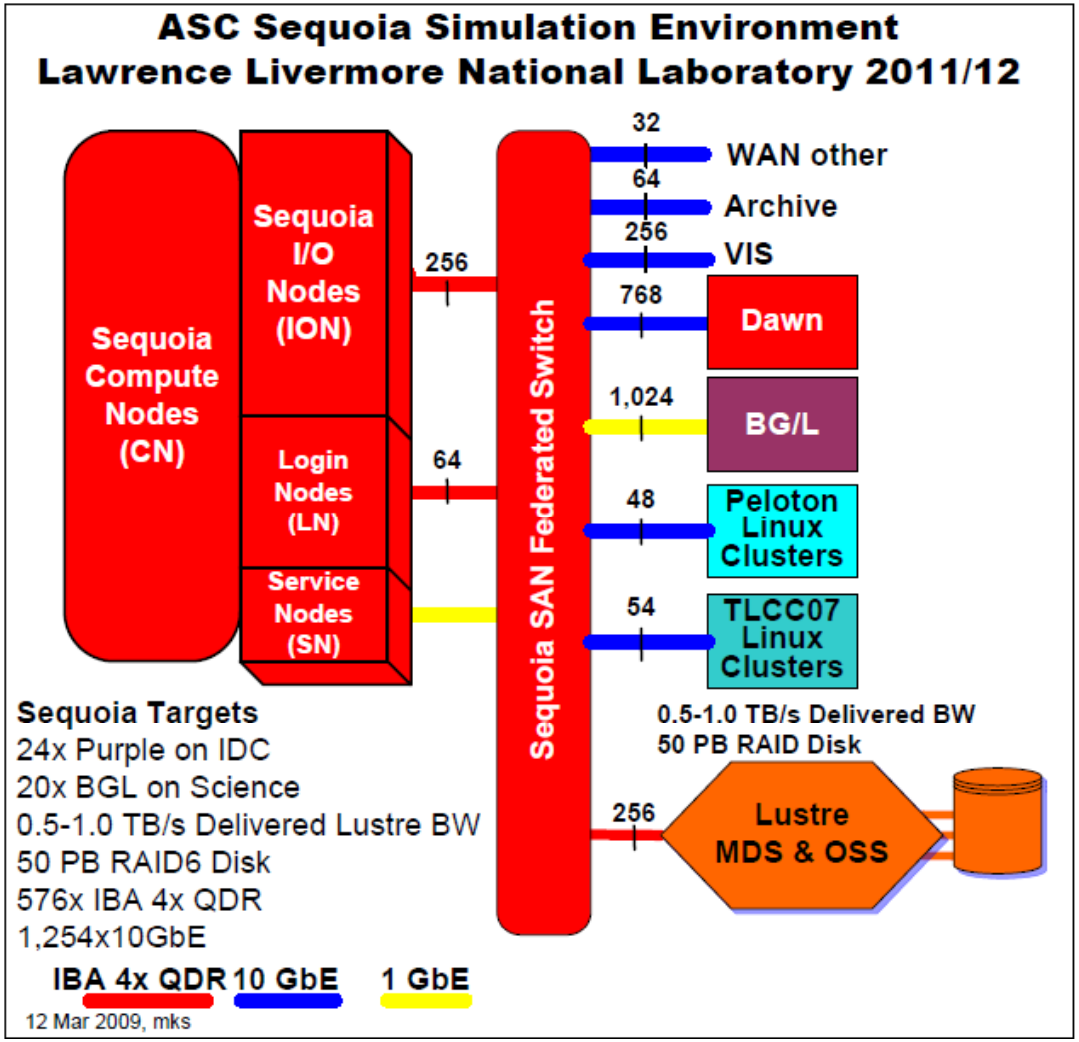
Sequoia Timeline Delivers Petascale Resources to the Program



Sequoia Five Years Planned Lifetime Through CY17

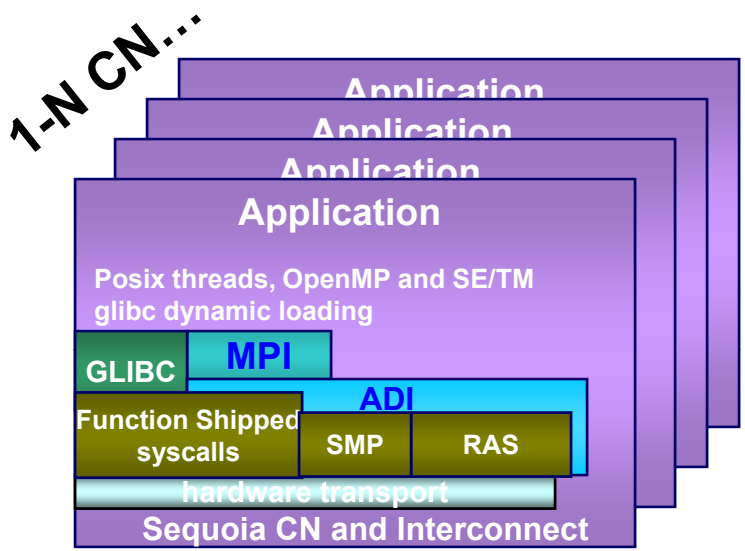
Sequoia contract award | Dawn system acceptance | Sequoia phase 2 & final system acceptance

Sequoia Hierarchal Hardware Architecture in Integrated Simulation Environment



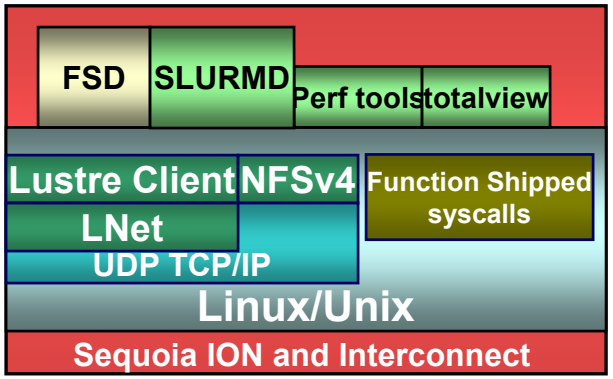
- Sequoia Statistics
  - 20 PF/s target
  - Memory 1.6 PB, 4 PB/s BW
  - 1.5M Cores
  - 3 PB/s Link BW
  - 60 TB/s bi-section BW
  - 0.5-1.0 TB/s Lustre BW
  - 50 PB Disk
- 6.0MW Power, 3,500 ft^2
- Third generation IBM BlueGene
- Challenges
  - Hardware Scalability
  - Software Scalability
  - Applications Scalability

Sequoia will scale utilizing hierarchal software model



Light weight kernel on compute node

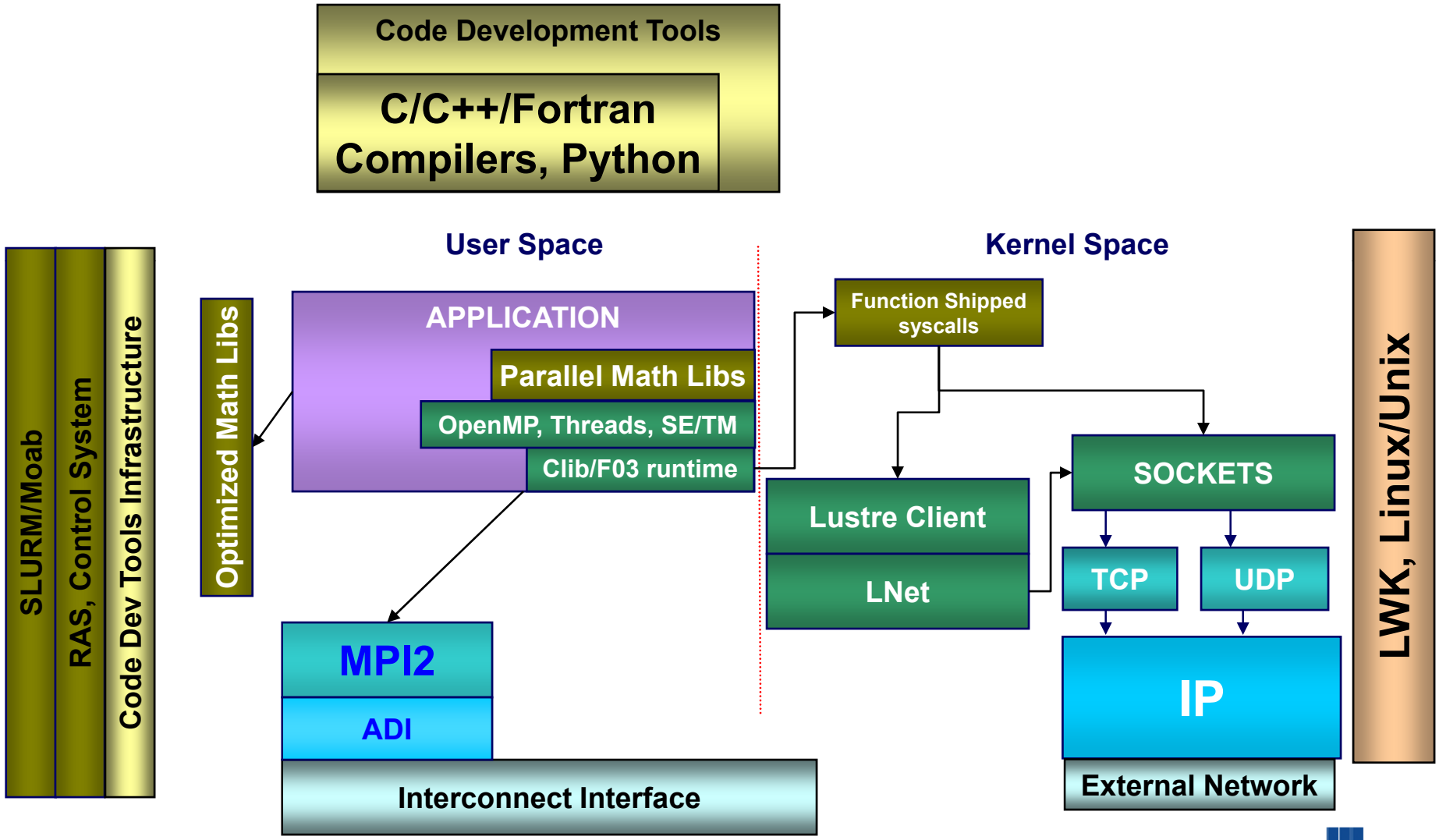
- Optimized for scalability and reliability
  - As simple as possible. Full control
  - Extremely low OS noise
  - Direct access to interconnect hardware
- OS features
  - Linux syscall compatible with IO syscalls forwarded to I/O nodes
  - Support for dynamic libs runtime loading
  - Shared memory regions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



Linux on I/O Node

- Leverage huge Linux base & community
  - Enhance TCP offload, PCIe, I/O
- Standard File Systems Lustre, NFSv4, etc
- Factor to Simplify:
  - Aggregates N CN for I/O & admin
- Open source

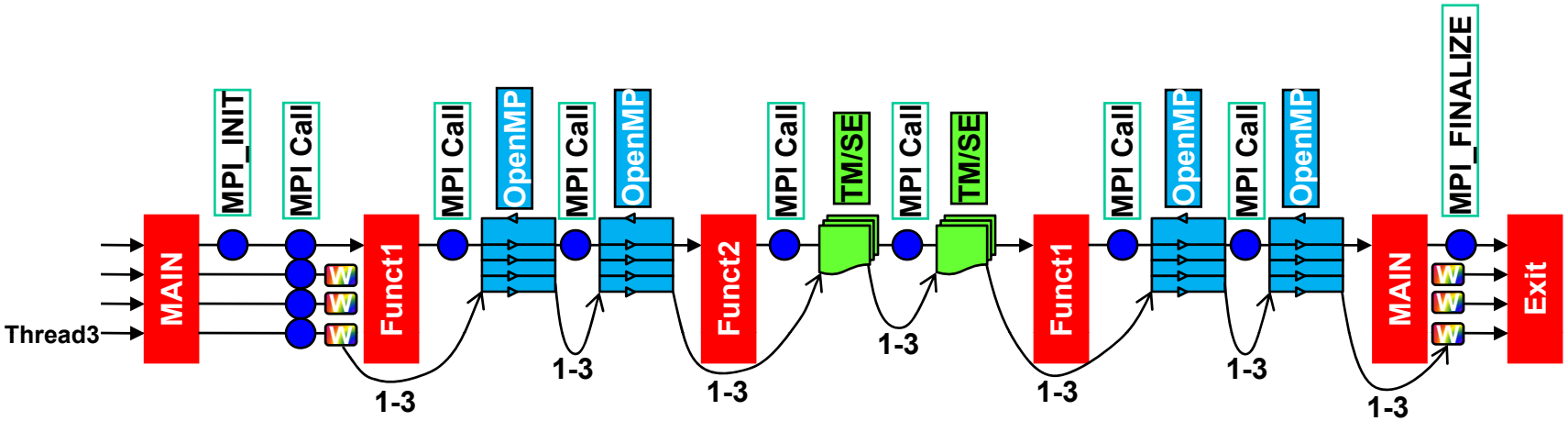
Sequoia Software Stack – Applications Perspective



Application Programming Model Requirements

- MPI Parallelism at top level
  - Static allocation of MPI tasks to nodes and sets of cores+threads
- Effectively absorb multiple cores+threads in MPI task
- Support multiple languages: C/C++/Fortran03
- Allow different physics packages to express node concurrency in different ways

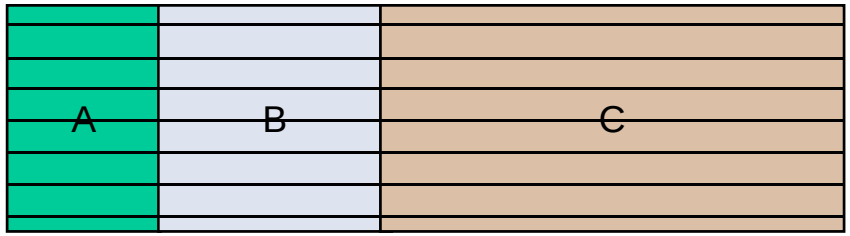
Sequoia's programming model is a simple extension beyond MPI with flexible mechanisms to absorb cores and threads



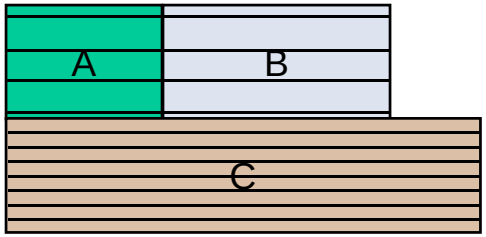
Weapon physics codes can use most efficient style of multi-core programming for each package and nest them

New approach to parallelization: apply multiple approaches to parallelism at the code and package levels

- Utilize the optimal parallelism methodology for each package
  - Nested Node Concurrency programming model allows different packages to exploit SMP parallelism differently
  - OpenMP, Pthreads and SE/TM available
- Run packages within a code in parallel
  - Can absorb appropriate number of nodes to load level the application



Inefficient: packages A,B,C are run sequentially, with sub-optimal level of parallelism

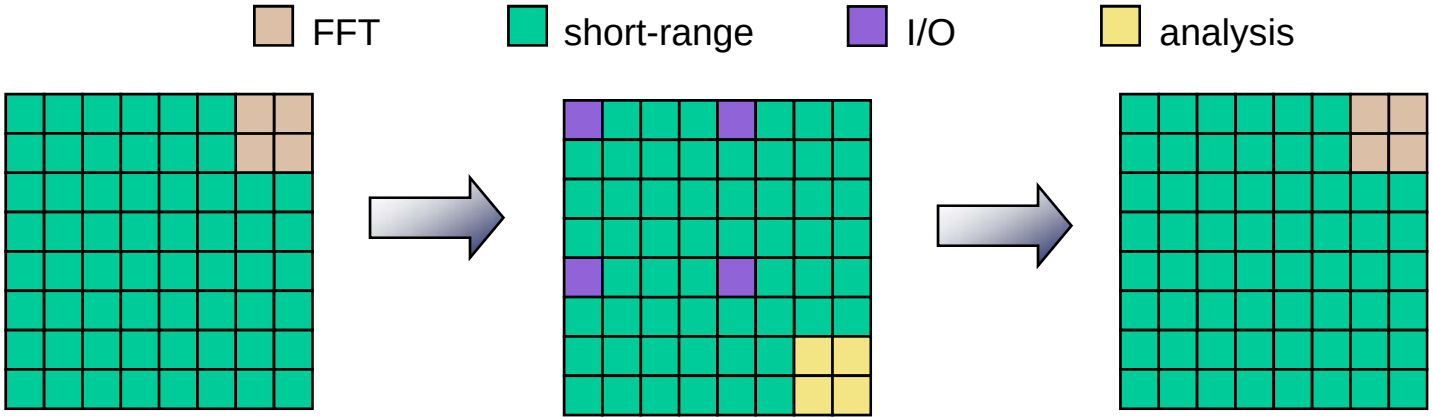


High-performance: compute-heavy package C is run concurrently with packages A and B

Availability of Dawn as a robust testbed and support of IBM collaboration allows development of novel parallelization model and implementations

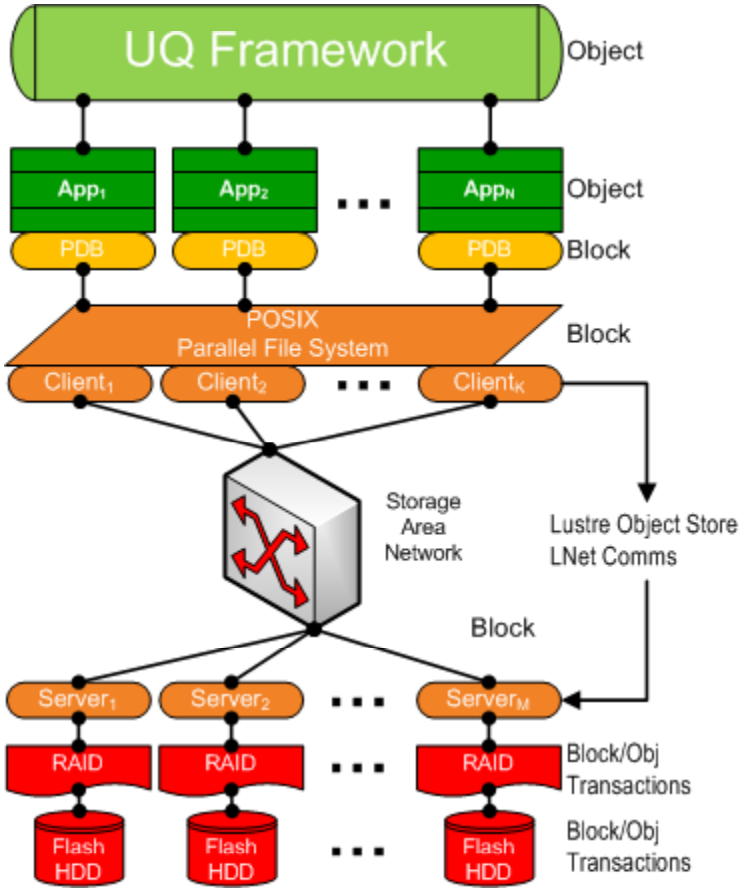
One example: novel parallelization in ddcMD/plasma

- Plasma modeling requires efficient handling of both short-range and long-range interactions
- Long range interactions are typically calculated using reciprocal space (FFT) methods, which do not scale well



Solution:
- majority of nodes are used for short-range forces that scale extremely well
- use just enough MPI tasks to minimize the communication bottleneck for FFT
- use local threads to efficiently calculate FFT on small number of nodes

Thinking about applications in the UQ Framework allows new approaches to solving old problems



- Contextual information within the UQ framework and applications could be transmitted through IO path to provide a systematic approach to scientific data management
- Current parallel file systems with block oriented interfaces currently don't encourage accumulating contextual information

Sequoia Platform Target Performance is a Combination of Peak and Application Sustained Performance

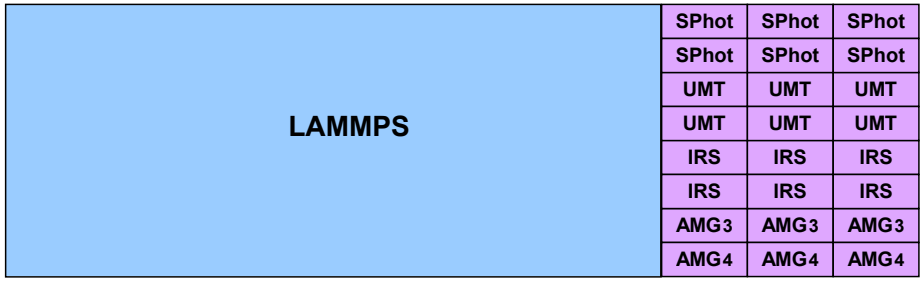
- "Peak" of the machine is absolute maximum performance
  - FLOP/s = FLoating point OPeration per second
- Sustained is weighted average of five "marquee" benchmark code "Figure of Merit"
  - Four IDC package benchmarks & one "science" benchmark from SNL
  - FOM chosen to mimic "grind times" factors out scaling issues
- Three purposed for Sequoia Benchmarks
  - RFP selection
  - Bone fides for Marquee & other requirements in the contract
  - Synthetic Workload for machine pre-ship and acceptance testing

The marquee benchmark strategy for aggregating performance incentivizes IBM in two ways: scalability and throughput

AMG	wFOM = A x “solution vector size” * iter / sec
IRS	wFOM = B x “temperature variables” * iter / sec
SPhot	wFOM = C x “tracks” / sec
UMT	wFOM = D x corners*angles*groups*zones * iter / sec
LAMMPS	wFOM = E x atom updates / sec

Aggregate wFOM = $wFOM_{AMG} + wFOM_{IRS} + wFOM_{SPhot} + wFOM_{UMT} + wFOM_{LAMMPS}$

- **Applications weights**
  - **Normalize the benchmarks to each other on reference systems**
  - **All benchmarks are of equal importance**
  - **Based on the targets of 24X Purple IDC & 20X BG/L Science**



This benchmarking strategy assures Sequoia will deliver *both* UQ and Science to the Stockpile Stewardship Program

Performance Projections for Marquee Benchmarks Delivers on Program Goals for Sequoia

M = P + S = 20.0 + 28.3 = 48.3

Benchmark	Raw FOM ratio of BGQ relative to Purple/BGL node	Est. Speed Factor From 6 copies
AMG	1.07	6.4
IRS	1.95	11.7
SPhot	1.82	10.9
UMT	1.23	7.4
LAMMPS	18.0	20.3

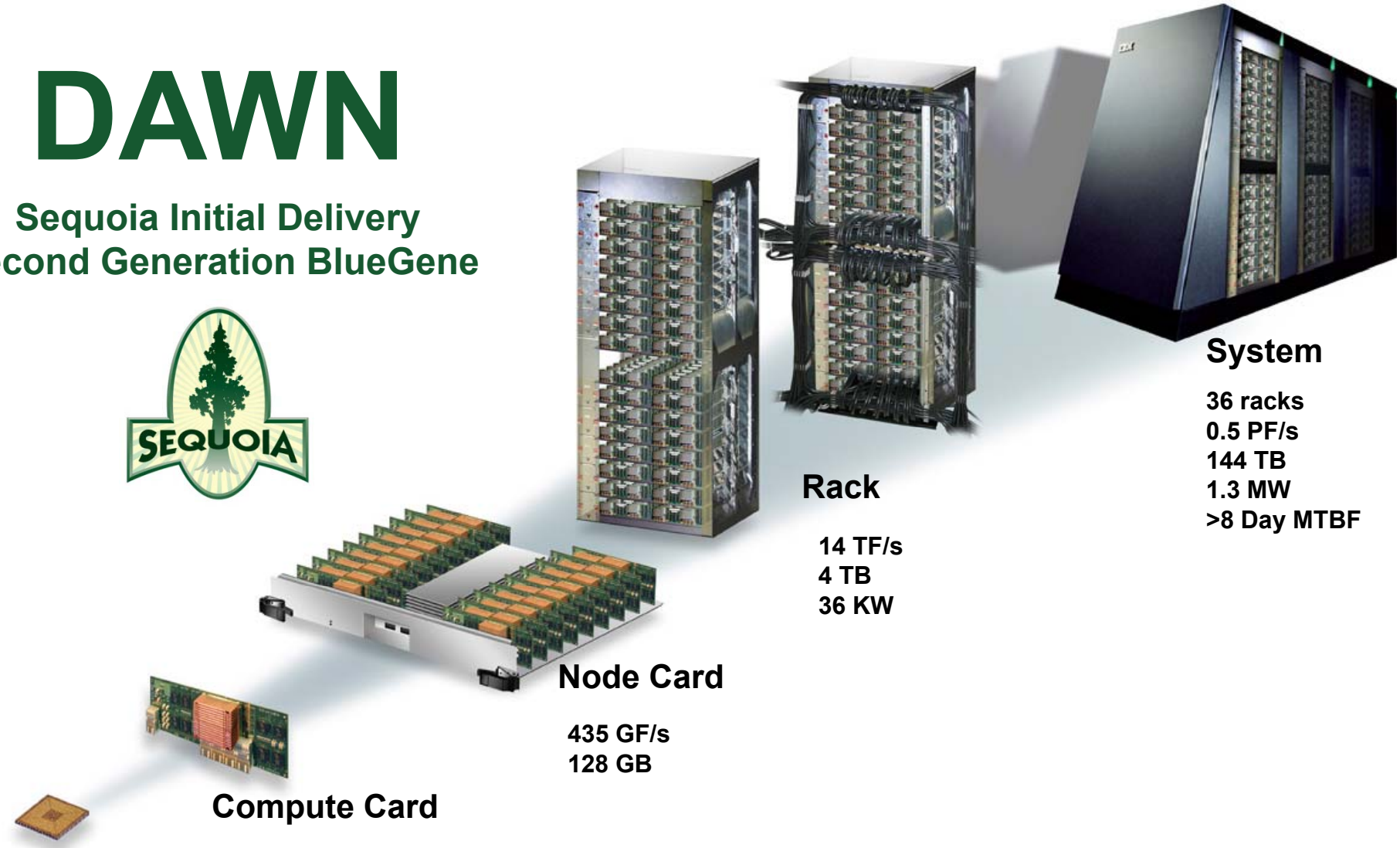
Aggregate Target 24X Est. 36.4X

Target 20X

Multi-physics runs on 1,024 BGQ nodes with 8MPI/8OMP and 1,024 Purple nodes w/8MPI
LAMMPS run on 72K BGQ nodes with 16MPI/4OMP and 64K BGL nodes w/2MPI

DAWN

Sequoia Initial Delivery
Second Generation BlueGene



Dawn acceptance is complete and early science runs have commenced

- 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
- Initial full system science runs currently underway
- Transition to classified service mid July, depending on science run demands

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

Scalable Applications Preparation (SAP) Project assists code teams in fully exploiting Sequoia capability

- Training and seminars on key technologies for multi-core programming

- Leverage: PCET LDRD, LLNL/ANL R&D partnership to accelerate Sequoia First Wave Applications

- User Guide and Performance Tuning Documentation developed by LC User Training and Hotline staff

- Engagement of Tri-Lab code teams with site visits for training, workshops, and regular video conferences

- Use Dawn, hardware and software simulators to provide early access to new technologies for Sequoia