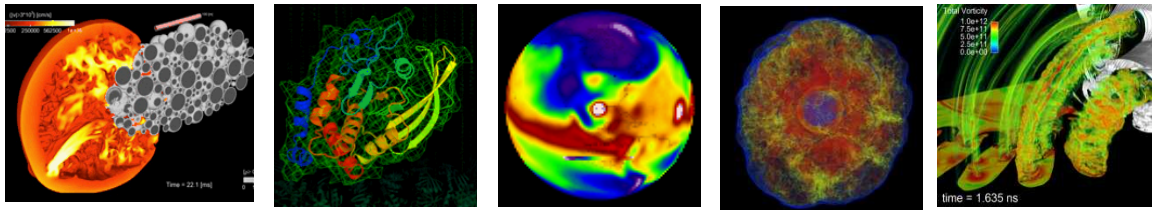
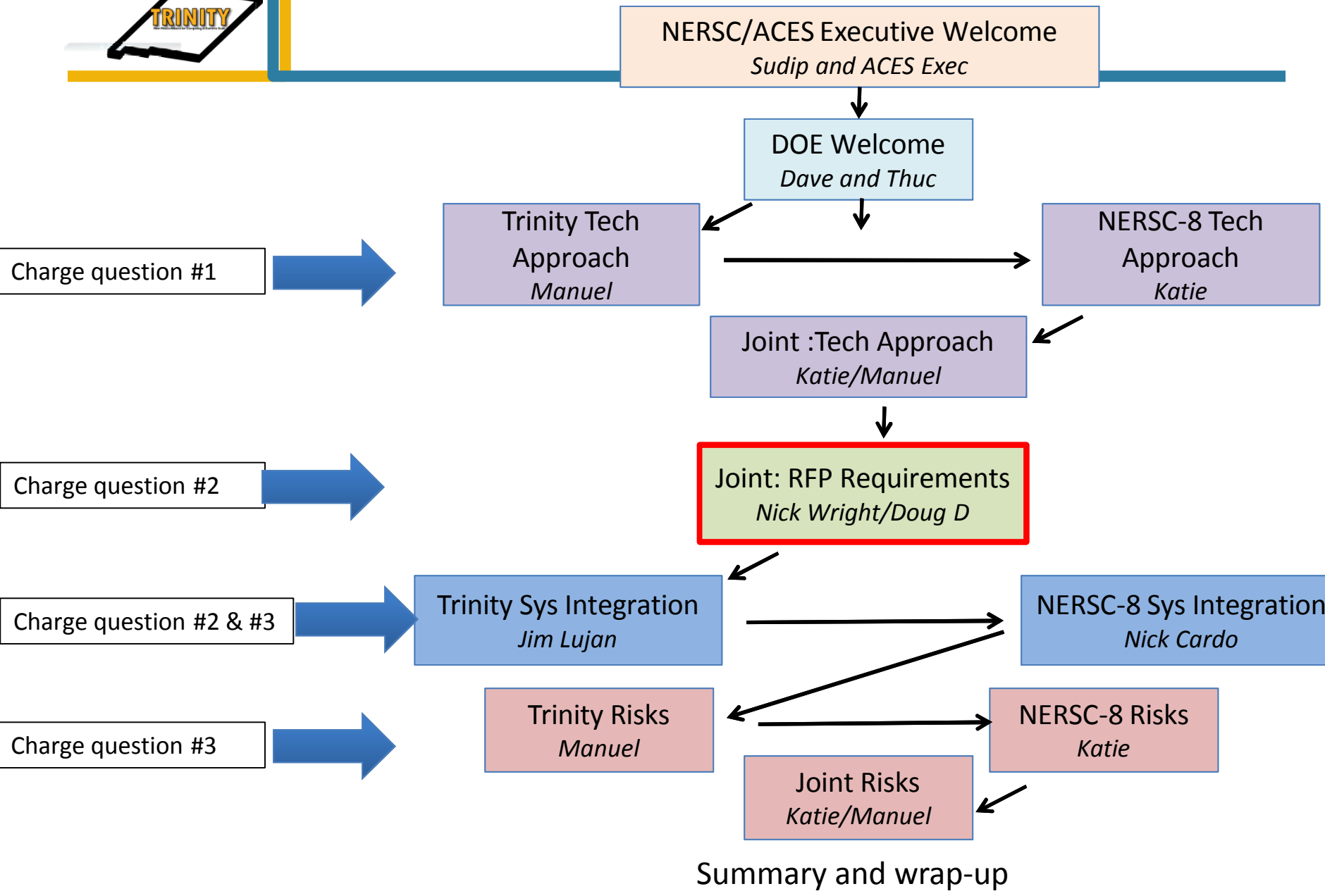
Are the RFP technical requirements reasonable, clear, and consistent with the goals and objectives for NERSC-8 and Trinity projects?

SAND2013-2434P



Nicholas J. Wright, NERSC
Douglas Doerfler, Sandia

Plan for the Day



Questions for the Panel

1. Is the technical approach appropriate to support the NERSC-8 and Trinity Mission Need Requirements?
2. Are the RFP technical requirements reasonable, clear, and consistent with the goals and objectives for NERSC-8 and Trinity projects?
3. Have the major technical risks and appropriate mitigation strategies been correctly identified at this stage of the project?

Goals and Objectives for the Trinity Project

1. Acquire right size platforms to meet mission needs for ASC codes in support of Stockpile Stewardship
2. Invest in prioritized R&D technologies to explore and enable new and incoming technologies
3. To help prepare and begin the transition the ASC program to future advanced technologies & programming environments
4. The full integration of Trinity into the LANL classified environment and enable a productive user environment

Goals and Objectives for the NERSC-8 Project

1. Provide a significant increase in computational capabilities for DOE SC computational research, with at least a 10 times increase in sustained performance over the NERSC-6 Hopper system in the 2015/2016 timeframe
2. Begin transitioning the broad SC user code base to advanced manycore architectures and programming environments
3. Provide an environment that enables user productivity

RFP Technical Requirements : Guiding Principles

- Do not prescribe a solution – let the vendor propose one that meets our requirements
- The solution needs to provide increased capability to the NERSC and ACES mission
  - Focus is upon science – not peak flops
- Not a one off solution – continuity in programming model
- Advanced technology architecture
- In places where Trinity and NERSC8 may differ ensure that vendor can configure solution to meet differing needs

Basic Structure of the RFP technical requirements

- Mandatory Requirements
- Target Requirements
- Technical options
- Additional system options
- Delivery and Acceptance Requirements*
- Technical Services, Documentation and Training
- Vendor Capabilities and Risk Management*

* Covered later

RFP: Mandatory Requirements

- A single proposal that where needed describes how the Trinity and NERSC8 systems differ
- Detailed architectural description
- Description of how proposed system fits into long term roadmap
- Address all the technical options

System Configurations

	Trinity	NERSC-8
Memory Capacity	2 PB to 4 PB	1 PB to 2 PB
Sustained System Performance (SSP)	20 to 60x over Hopper	10x to 30x over Hopper
Capability Improvement	8 to 10x over Cielo	N/A
JMTTI	> 24 hours	> 35 hours
JMTTI/delta	> 30	> 30
File System BW metric – time to dump 80% RAM	20 mins	30 mins
File System disk capacity	> 30x main memory	> 20x main memory
Power	< 12+3 MW	< 6 MW
Off-platform I/O	> 140 GB/s	> 180 GB/s

System Configurations

	Trinity	NERSC-8
Memory Capacity	2 PB to 4 PB	1 PB to 2 PB
Sustained System Performance (SSP)	20 to 60x over Hopper	10x to 30x over Hopper
Capability Improvement	8 to 10x over Cielo	N/A
JMTTI	> 24 hours	> 35 hours
JMTTI/delta	> 30	> 30
File System BW metric – time to dump 80% RAM	20 mins	30 mins
File System disk capacity	> 30x main memory	> 20x main memory
Power	< 12+3 MW	< 6 MW
Off-platform I/O	> 140 GB/s	> 180 GB/s

Expected Sizing Depends on Budget

- Sizing is primarily driven by aggregate memory capacity and application performance requirements
  - Memory capacity requirement only includes that memory which is used as the primary store for the problem.
  - Nominally, this means the amount of DDR SDRAM used as main processor memory
  - It does NOT include memory associated for scratch pad use and/or accelerator memory
- We expect a node for Trinity/NERSC-8 to be priced consistently in $/node with historical acquisitions

The Sizing Requirements do Align with the Budget: Memory

- We nominally expect a node to contain 128 to 256 GB of main memory capacity
  - 4,096 to 8,192 nodes per aggregate PB
- Trinity: (2 PB) 8,192 to (4 PB) 32,768 total nodes
  - Trinity will be ~2X the number of nodes of Cielo
  - Trinity Budget is ~2X Cielo's budget for the baseline system
- NERSC8: (1 PB) 4,096 to (2 PB) 16,384 total nodes
  - NERSC8 will be ~1X the node count of Hopper
  - NERSC8 budget is ~1X Hopper's budget for the baseline system

Capability Improvement Definition

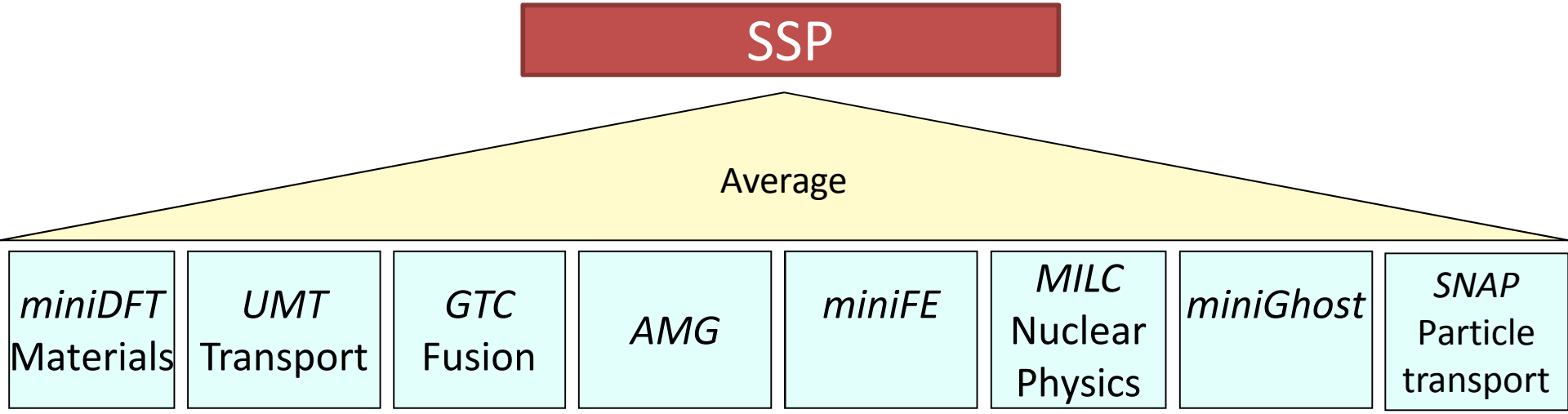
- Requirement: Capability Improvement (CI) factor 8x to 10x of Cielo
- Capability Improvement (CI) is defined as the product of an increase in problem size/complexity and an application specific runtime speedup factor. E.g.:
  - Size or complexity (app dependent) increases by a factor of 8
  - Runtime figure of merit (app dependent, e.g. time/iteration) improves by a factor of 1.2
  - Capability Improvement = 8x * 1.2 = 9.6x
- ASC Application Suite
  - PartiSN (LANL)
  - Sierra/Aero (SNL)
  - ???/(LLNL)
- **Application performance, in conjunction with high memory capacity, allows Trinity to meet the mission need.**

Capability Improvement Requirements are Reasonable

- 2011-2015: 8x Application performance per node
- Trinity 2-4x Cielo nodes
- Each Node Memory Capacity: Trinity 4-8x Cielo
- To solve a 8x larger problem than today in the same time need 8x memory & x performance = 8x capability improvement
- 2x #nodes of Cielo = 16x capability improvement
- 4x #nodes of Cielo = 32x capability improvement

Sustained System Performance

Because hardware peak performance does not necessarily reflect real application performance



- Trinity and NERSC8 have defined a common application performance metric based on a diverse set of applications
  - By composing the suite with applications from both missions, performance will be applicable to both program's needs
- This is consistent with all other key sizing requirements, in particular memory capacity
  - A common application metric also provides a comparison point for the two platforms at acceptance. i.e. results are comparable and consistent

SSP Improvement Estimates are Reasonable

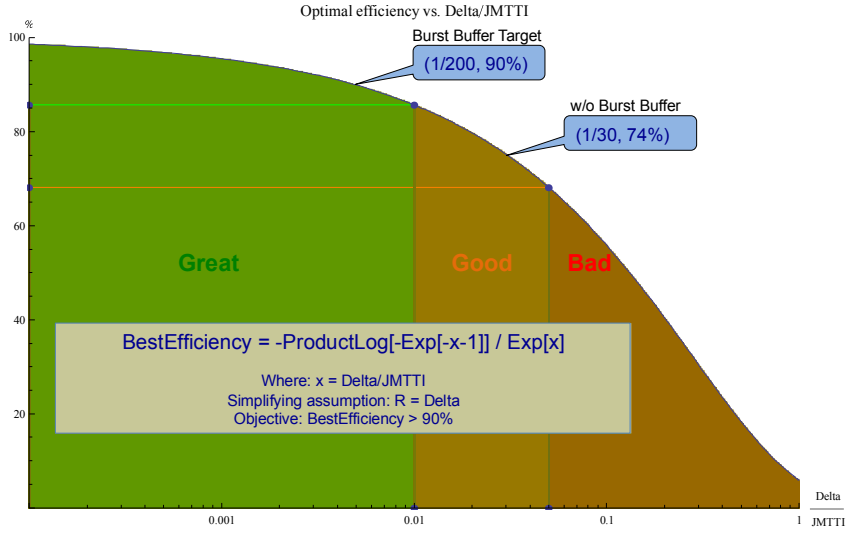
- 2011-2015: 8x Application performance per node
- Trinity 2.8-5.6x number of Hopper nodes
- NERSC-8 1.4-2.8x number of Hopper nodes
- The same problem should be solved in the same time and use 1/8 less nodes
- 8*(2.8-5.6)x = 22.4-44.8x Trinity (20-60 required)
- 8*(1.4-2.8)x = 11.2-22.4x NERSC-8 (10-30 required)

System Configurations

	Trinity	NERSC-8
Memory Capacity	2 PB to 4 PB	1 PB to 2 PB
Sustained System Performance (SSP)	20 to 60x over Hopper	10x to 30x over Hopper
Capability Improvement	8 to 10x over Cielo	N/A
JMTTI	> 24 hours	> 35 hours
JMTTI/delta	> 30	> 30
File System BW metric – time to dump 80% RAM	20 mins	30 mins
File System disk capacity	> 30x main memory	> 20x main memory
Power	< 12+3 MW	< 6 MW
Off-platform I/O	> 140 GB/s	> 180 GB/s

JMTTI & Productivity

- Requirements
  - JMTTI > 24 hours (Trinity) >35 NERSC-8
  - JMTTI/delta > 30 using PFS
  - JMTTI/delta > 200 using Burst Buffer
- 24 hours is a practical metric to provide a productive, efficient environment for the user, but still relying on checkpoint/restart and an adequate PFS.
- JMTTI/delta is a new metric that lets us quantify efficiency
  - 30 is an efficiency of ~75%
  - 200 is an efficiency of ~90%
- **Reliability of the platform is key to meeting productivity requirements and hence mission need.**

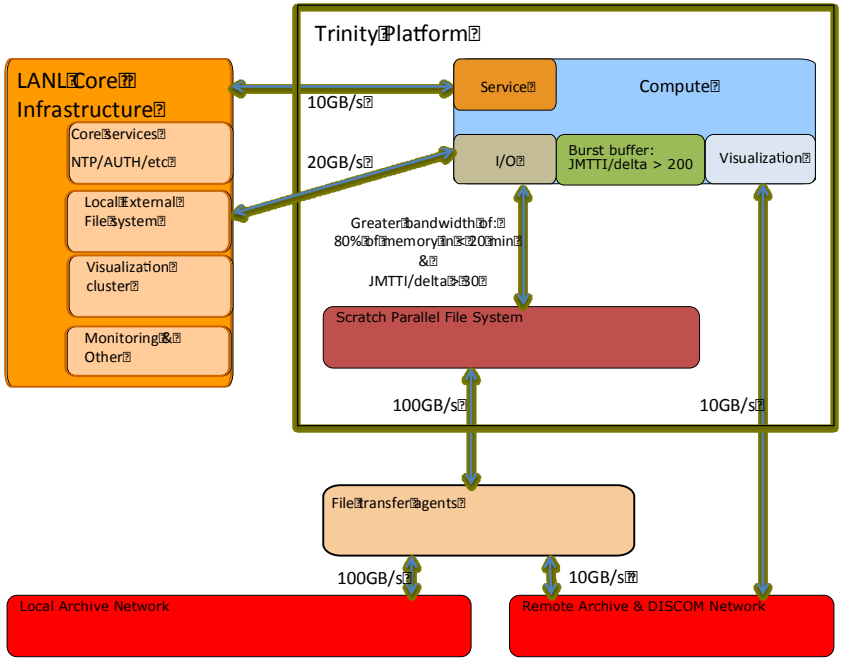


System Configurations

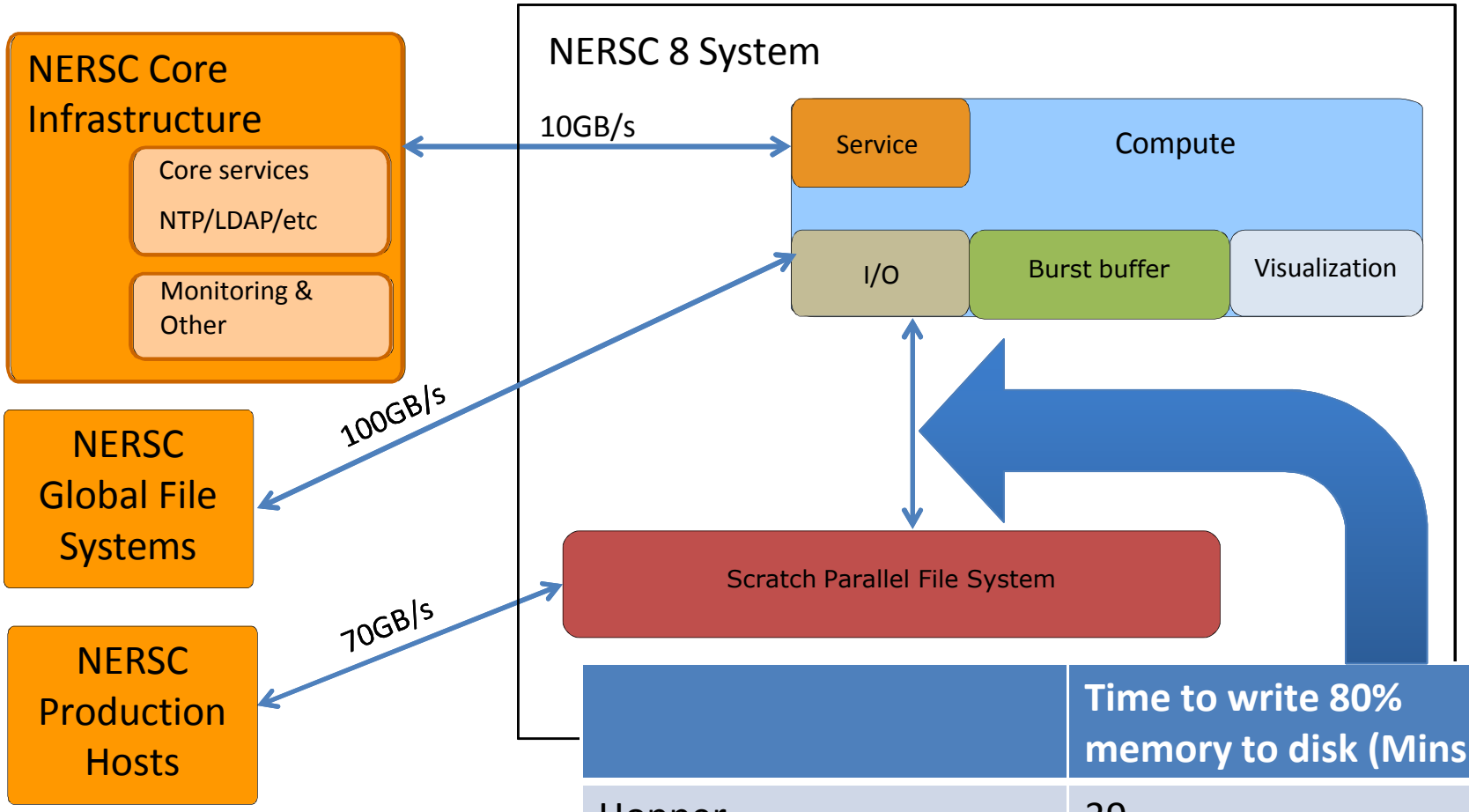
	Trinity	NERSC-8
Memory Capacity	2 PB to 4 PB	1 PB to 2 PB
Sustained System Performance (SSP)	20 to 60x over Hopper	10x to 30x over Hopper
Capability Improvement	8 to 10x over Cielo	N/A
JMTTI	> 24 hours	> 35 hours
JMTTI/delta	> 30	> 30
File System BW metric – time to dump 80% RAM	20 mins	30 mins
File System disk capacity	> 30x main memory	> 20x main memory
Power	< 12+3 MW	< 6 MW
Off-platform I/O	> 140 GB/s	> 180 GB/s

Trinity: Platform I/O Balance

- Parallel File System BW: 16x Cielo
  - /scratch BW of 2.7 TB/s
  - (80% of 4 PB in 20 min)
- Parallel File System Capacity: 16x Cielo
- General I/O: 1x Cielo
  - External login
  - Local WAN
- Local File Systems: 2.5x Cielo
  - NFS, /home, /projects, etc.
- File Transfer Agents: 2.5x Cielo
  - DISCOM Network
  - Remote Archive
  - Local Archive
  - Remote archive & file transfer
- Visualization: 10x Cielo
  - To the desktop

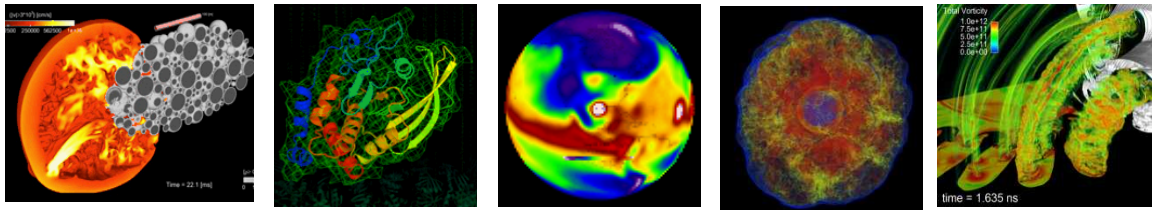


NERSC: Platform I/O



	Time to write 80% memory to disk (Mins)
Hopper	39
Edison	32
NERSC-8	30

Application performance - Benchmarking Strategy



Vendor Surveys have formed the basis of our requirements development

- The Trinity/NERSC-8 teams had several formal and informal (e.g. telecons) interactions with vendors over the last 15 months
  - Will continue these interactions leading up to the RFP release
- Surveys have focused on major prime and technology provider candidates:

Processor Technology Choice is Key

- Processor becoming more of a defining architectural feature
  - Dictates programming environment and models
  - Prime vendors becoming more dependent on processor vendor for the entire tool chain
  - Driving development of features necessary to meet reliability and power demands of future systems
  - First round of Fast Forward is focused on processor vendors
- High level landscape
  - x86 multi-core: Intel, AMD
    - greater vector level parallelism, increasing core counts
  - Homogenous manycore: Intel Phi
    - greater process, thread and vector level parallelism with a homogenous core design
  - Heterogeneous manycore: Nvidia Tesla and AMD APU
    - Greater process and thread level parallelism using a heterogeneous architecture of throughput and latency optimized core designs
  - Power: IBM Power architecture
    - High throughput cores with high bandwidth and the potential for acceleration
  - Other
    - Technologies being developed for HPC but not ready for our time frame: E.g. ARM processors; on-socket or on-die high speed interconnects; etc.

Processor technology choice is key cont.

- MPI+X: X programming models and environments are key to choice of processor
  - At the highest level, abstractions such as OpenMP and OpenACC will provide high performance, portable code
  - However, to extract maximum performance it may be necessary to use less portable languages or directives.

All the potential technology options will require significant changes to traditional notions of benchmarking strategy

High Speed Interconnect Defines Scalability

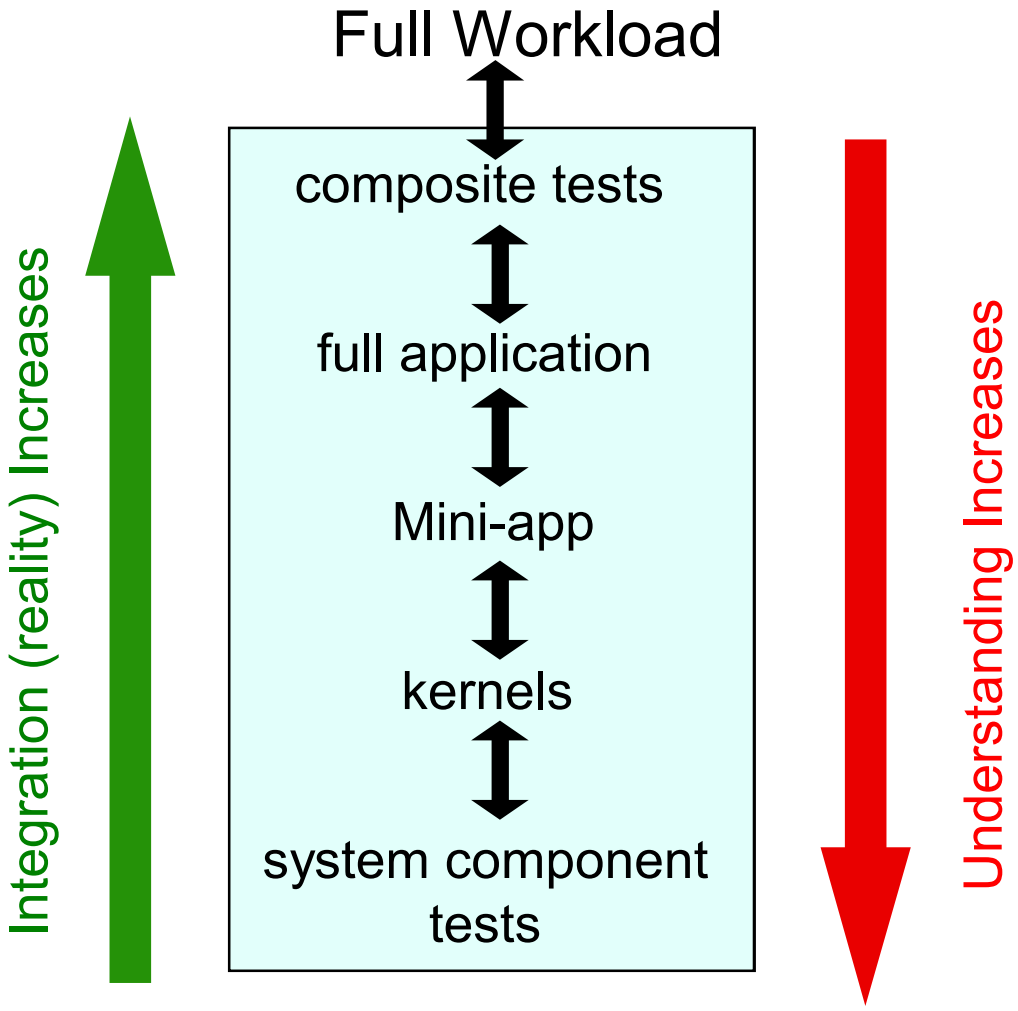
- High-speed interconnect solutions are lagging processor technology development
- 2013 Technologies
  - Mellanox InfiniBand FDR
  - Cray Aries
  - IBM Blue Gene Q
- 2015 Technologies
  - Mellanox InfiniBand EDR
  - Cray Aries
  - And others…..
- B/F ratios are decreasing by a factor of 5 to 10 between now and 2015
  - This gap is being actively worked by the ACES and NERSC teams, but alternative options are scarce to non-existent
- In addition to high bandwidth, we require high message rates (throughput) and low latencies
- Topology choices
  - Topology is less of a concern, our codes have shown to perform well across a variety of topologies
  - Fat Tree topology using InfiniBand is the most mature, 3D torus and hypercubes have also been deployed
  - Cray's Aries interconnect supports the Dragon Fly topology
  - IB supports various topologies

Software Tools & Programming Environment

- MPI and MPI+X environments
  - MPI-only may not be high performance
  - MPI+X will most likely be necessary for most codes
- OpenMP and/or OpenACC are anticipated to be the “entry” point for X
- Other, perhaps less portable methods may be necessary to achieve high performance
  - E.g. CUDA, Thread Building Blocks, compiler directives and extensions, intrinsics, etc.
- High level languages include C, C++, Fortran77 and Fortran 2003
  - Desirable to have C++11 features, such as lambda functions
  - Fortran 2008, including CoArray features, is also desirable

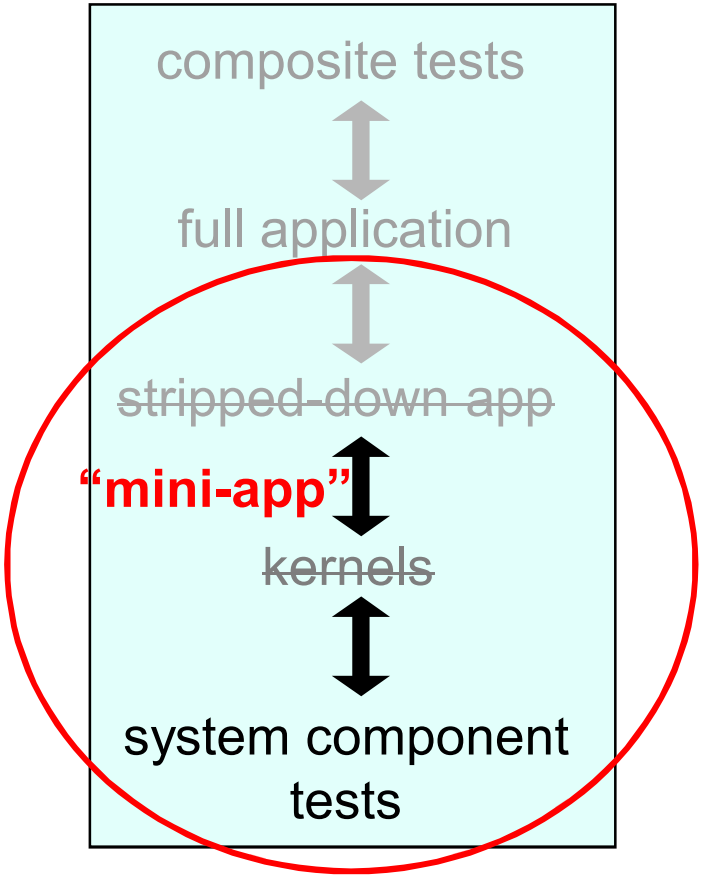
Past procurements have released benchmarks of various levels of complexity

- **Distinguish performance of systems**
- **Represent scientific workload on system**
- **Give confidence that chosen system will perform well for a given workload**



"mini-apps", full apps and micro-benchmarks will be used for system evaluation

- Technology trends are driving a move to MPI+X
  - Unreasonable to expect vendors to port large application to X
- Mini-app
  - Stripped down version of a scientific application that contains key kernels
  - Significantly less lines of code than full application – allows experimentation with new programming models
  - Performance obtained representative of performance of full application



NERSC-8/Trinity plan to use "mini-apps", some full apps for system evaluation

MiniApp	Description
miniDFT (Quantum Espresso)	Density Functional Theory (DFT)
MILC	Lattice Quantum Chromodynamics (QCD). Sparse matrix inversion, CG
GTC	Particle-in-cell magnetic fusion
AMG	Algebraic Multi-Grid linear system solver for unstructured mesh physics packages
UMT	Unstructured-Mesh deterministic radiation Transport
miniFE	Unstructured implicit finite element
miniGhost	Finite difference stencil
SNAP	Neutral particle transport application

Benchmark Run Rules

- Three problem sizes defined: small, large and extra large
  - Small is sized for at a single Trinity/NERSC-8 node
  - Large is sized for 1,000 to 2,000 nodes (also used for SSP)
  - Extra large is sized for nominally 10,000 nodes (NNSA mini-applications only)
- Desired that full memory hierarchy is utilized, e.g. 50% of main memory
- Base and Optimized
  - Base case is MPI-only, this is to gauge performance of legacy codes w/o major modifications
  - Optimized case is MPI+X, modifications are allowed so long as all techniques are fully documented and code changes are made available. Algorithms fundamental to the application are not to be modified.
- SSP
  - NERSC's Hopper platform is the baseline reference point
  - Large problem size, Optimized results will be used

Methods covered by NERSC-8/Trinity Benchmarks

Codes	Dense Linear Algebra	Sparse Linear Algebra	FFTs	Particle Methods	Structured Grids	Unstructured Grids/AMR
miniDFT	X		X		X	
MILC		X		X	X	
GTC				X	X	
UMT		X				X
AMG		X				X
miniFE		X				X
miniGhost		X			X	
SNAP					X	X

Science Area coverage by NERSC-8/Trinity Benchmarks

Codes	Accel Sci	Astro physics	Chem	Climate	Combustion	Fusion	Lattice Gauge	Material Science
miniDFT			X					X
MILC							X	
GTC						X		
UMT		X				X		
AMG		X			X	X		
miniFE	X	X		X	X	X		
miniGhost	X	X	X	X	X	X	X	X
SNAP		X						X

Machine Stresses - NERSC-8/Trinity Benchmarks

Codes	Flops	Memory Bandwidth	Memory Latency	Network Bandwidth pt-to-pt	Global Network Bandwidth	Network Latency
miniDFT	X				X	X
MILC		X		X		X
GTC			X		X	
UMT		X				X
AMG		X			X	X
miniFE		X	X	X		X
miniGhost				X	X	
SNAP		X			X	

Communication Benchmarks

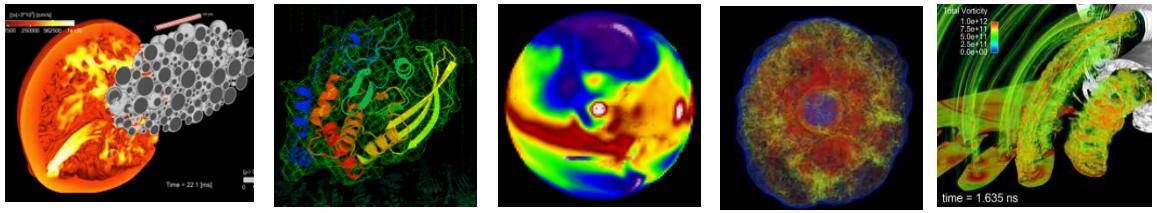
- OSU MPI benchmarks
  - Interconnect performance
- SMB
  - Message passing host processor overhead
- MPIMEMU
  - MPI node memory usage
- ZIATEST
  - MPI startup time
- UPC-FT
  - PGAS functionality and performance

I/O benchmarking

- Bandwidth & IOPs - IOR tests - MPI-I/O & POSIX I/O
  - File per process (N-to-N)
  - Shared file (N-to-1)
  - Max read/write for transfer sizes 10 KB, 100 KB, 1MB
  - Where N is determined by
    - Peak result on single node
    - Using all cores on node, number of nodes that yield peak results on test system
    - Using all cores on node, all nodes in test system
- Metadata – Mdtest
  - Create/remove 2^{20} files
    - By 1 process in 1 directory
    - By N processes in 1 directory
    - By N processes in N directories
  - Create/remove 1 file by N processes
  - Where N is determined by
    - Best result on a single node
    - Best result on multiple nodes
    - Using all nodes on the test system
    - Using all nodes on delivered system

Bandwidth and Metadata performance and scalability

Technical Options and Non Reoccurring Engineering

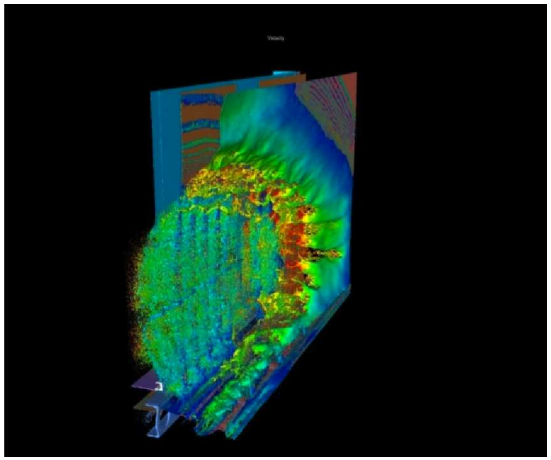


Options in the final RFP will allow each site to further customize a system

- Visualization partition
- Burst Buffer
- Advanced power management
- Application transition support
- Early access development systems and testbeds

Visualization Partition

- Requirement: On-platform visualization is necessary to meet user productivity requirements
  - Need high speed access to data generated by simulations, and hence need to be a peer on all high speed data networks including access to the parallel file system
  - Post-processing: analyzing data stored on the PFS
  - In-situ: analyzing data in memory as it is generated
  - In-transit: analyzing data off-node as it is generated
  - Ensemble: large ensembles of data, in-transit or post-processing



Reference: Steve Attaway, Shivonne Haniff, Joel Stevenson and Jason Wilke, "Cielo CCC-1 Summary: Lightweight, Blast Resistant Structure Development", SAND2011-6477P, Unclassified, Unlimited Release.

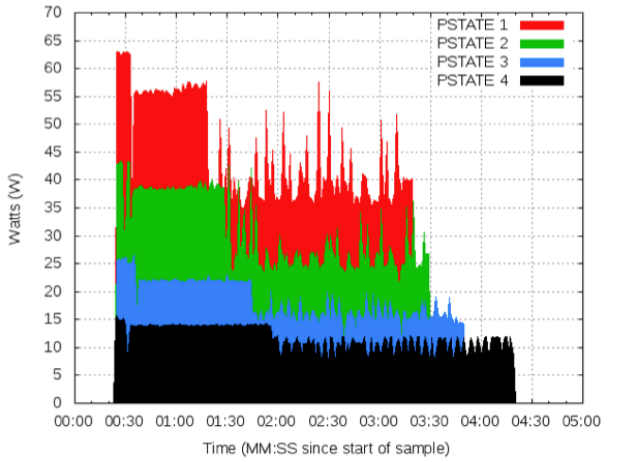
- Viz partition may or may not use common hardware as the main compute partition
- Either way, Viz partition is managed independent of the compute partition

Burst Buffer

- The Burst Buffer (BB) technology development will improve the efficiency of I/O operations on the platform
  - Checkpoint/restart is the primary use case
    - With C/R, application efficiency increases to > 90%
  - Data analytics is also a key use case: post processing and in-transit visualization, bioinformatics, etc.
- The BB will be designed in conjunction with the primary parallel file system, but failure of the BB shall not impact the parallel file system and storage
  - Independent data paths and failure domains
- The BB reliability will be assessed as part of the overall system reliability

Active Power Management

- Power is a constraining factor in the operation of Trinity
- Need to understand & control power
  - Cabinet & component level I & V measurements
  - Scalable collection infrastructure
  - Tunable collection fidelity: cabinets to components
  - Administrative & user accessible interface for feedback and tuning
- Need to manage power at the platform, runtime & application level
  - Policy driven
  - Weighted combination of performance & energy
  - Energy caps based on time of day, physical capacity, etc.



Single node capture of watts over time for each run of AMG2006, varying P-states

P-states (Frequency/Voltage States)

- P1: 2.1 GHz , 1.25V
- P2: 1.7 GHz, 1.1625V
- P3: 1.4 GHz, 1.125V
- P4: 1.1 GHz, 1.1V

AMG demonstration on 6,144 nodes of ORNL's Jaguar shows that managing P-States allows for a 32% decrease in energy used while only increasing time to solution by 7.5%

Application transition support

- The establishment of a collaboration between the Labs, the chosen OEM, and key technology providers, e.g. processor, is essential to meet the goals of the making efficient use of the platform in a timely manner
  - SSP metric applications
  - Capability Improvement metric applications
  - Selected applications expected to use the machine shortly after operational readiness

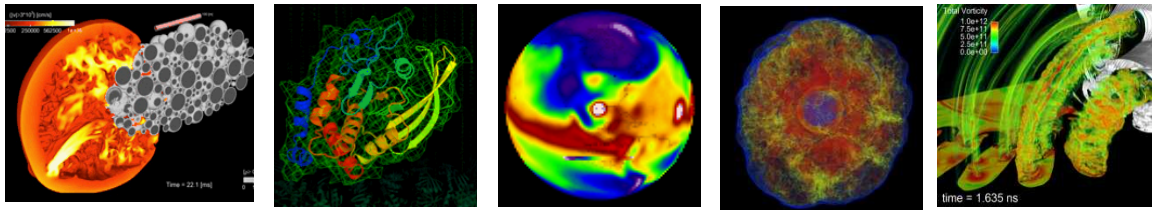
Early Access Development System

- Early access to key technologies and programming environments is essential for application transition
  - Programming environment is crucial
- Early Access Development System is tightly coupled to the success of the Application Transition Support initiative
- Desirable to have deployed at least 1 year before final system
- Sized at 2% to 10% of final system
- Also asking for proposals of smaller development test beds for advanced technology areas, e.g. power management and burst buffer

Draft Technical Requirements– Summary of Vendor Comments

- Provided to vendors Dec 13 2012
- Ten Responses received by Jan 17 2013 deadline
- Requests to clarify some definitions
  - Glossary moved to front of document and augmented
- Clarify Burst Buffer use cases
  - Document produced and will be released with the RFP
- Sample Acceptance test plan produced
- More benchmark info – Run Rules published
- Vendors requested to be allowed to submit supplemental technical information to augment the prime vendors response
  - Mechanism created

Summary



The RFP technical requirements are reasonable, clear and consistent with the goals and objectives of the Trinity project

1. Acquire right size platforms to meet mission needs for ASC codes in support of Stockpile Stewardship
   - Problem size/complexity capabilities 8x to16x that of Cielo
   - And performance sized to achieve a capability improvement 8x to 32x Cielo
2. Invest in prioritized R&D technologies to explore and enable new and incoming technologies
   - Raise efficiency to 90% using Burst Buffer strategy and technology
   - Allow accelerate of power management and control
3. To help prepare and begin the transition the ASC program to future advanced technologies & programming environments
   - Application transition support will make selected applications productive and efficient early in the platforms life
   - Programming environment will allow the larger application base to transition to Trinity and future advanced technology platforms
4. The full integration of Trinity into the LANL classified environment and enable a productive user environment
   - Facility upgrade on schedule
   - Trinity provides balanced I/O to ensure a productive user environment

The RFP technical requirements are reasonable, clear and consistent with the goals and objectives for the NERSC-8 Project

1. Provide a significant increase in computational capabilities for DOE SC computational research, with at least a 10 times increase in sustained performance over the NERSC-6 Hopper system in the 2015/2016 timeframe
   - 10-30x requirement is reasonable and consistent
   - The benchmarks represent the workload
2. Begin transitioning the broad SC user code base to advanced manycore architectures and programming environments
   - Application readiness effort will make DOE SC applications productive and efficient early in the platforms life
   - Programming environment will allow the larger application base to transition to NERSC-8 and future advanced technology platforms
3. Provide an environment that enables user productivity
   - The system will be fully integrated in the NERSC infrastructure
   - Many other RFP technical requirements are focused upon productivity