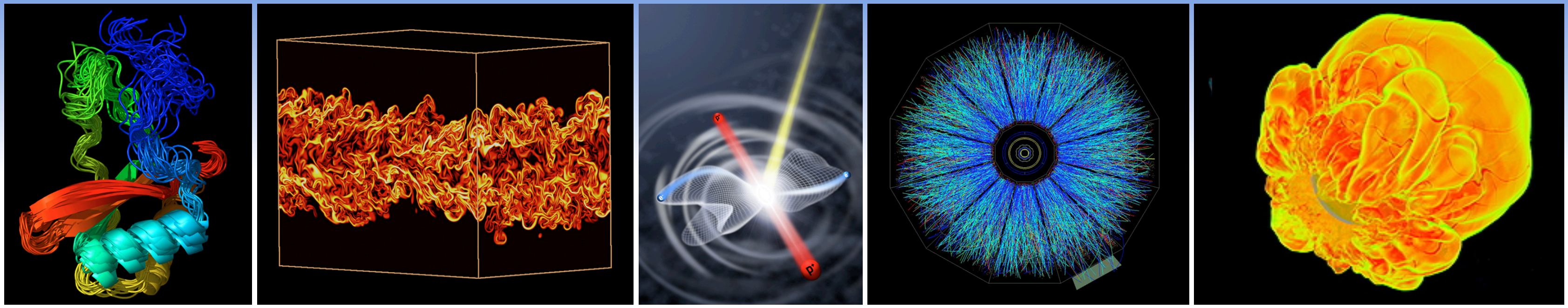


NERSC-6 Plans

Bill Kramer

NERSC-6 Project Manager

Lawrence Berkeley National Laboratory

August 8, 2008

BERKELEY LAB

NERSC

Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC is the Production Facility for DOE SC

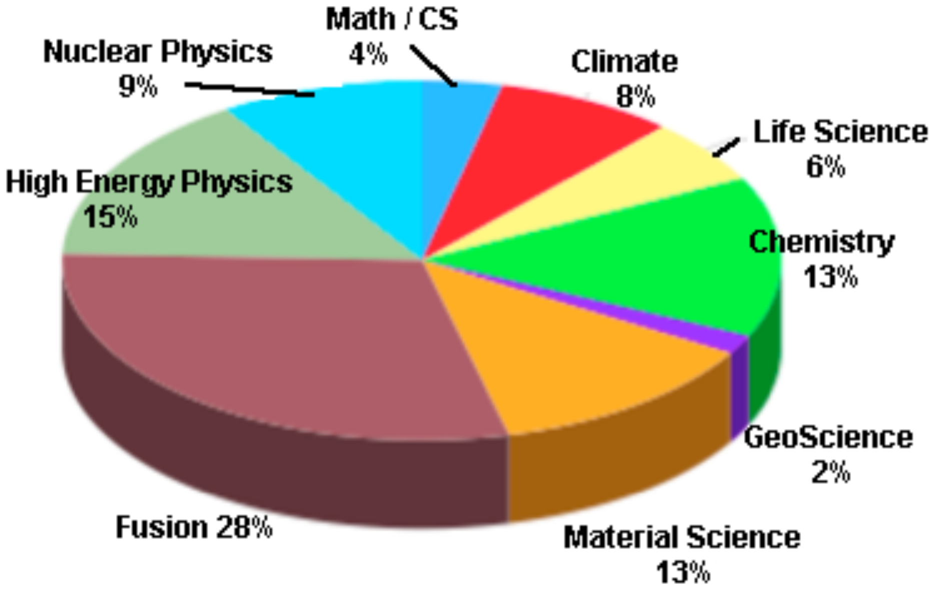
- **NERSC serves all areas**

  ~3000 users, ~400 projects

- **Allocations managed by DOE**
  - 10% INCITE awards:
    - Large allocations, extra service
    - Used throughout SC; not just DOE mission
  - 70% Production (ERCAP) awards:
    - From 10K hour (startup) to 5M hour
    - Only available at NERSC
  - 10% each NERSC and DOE/SC reserve

- **Award mixture offers**
  - High impact through large awards
  - Broad impact across science domains

Awards by Science Areas

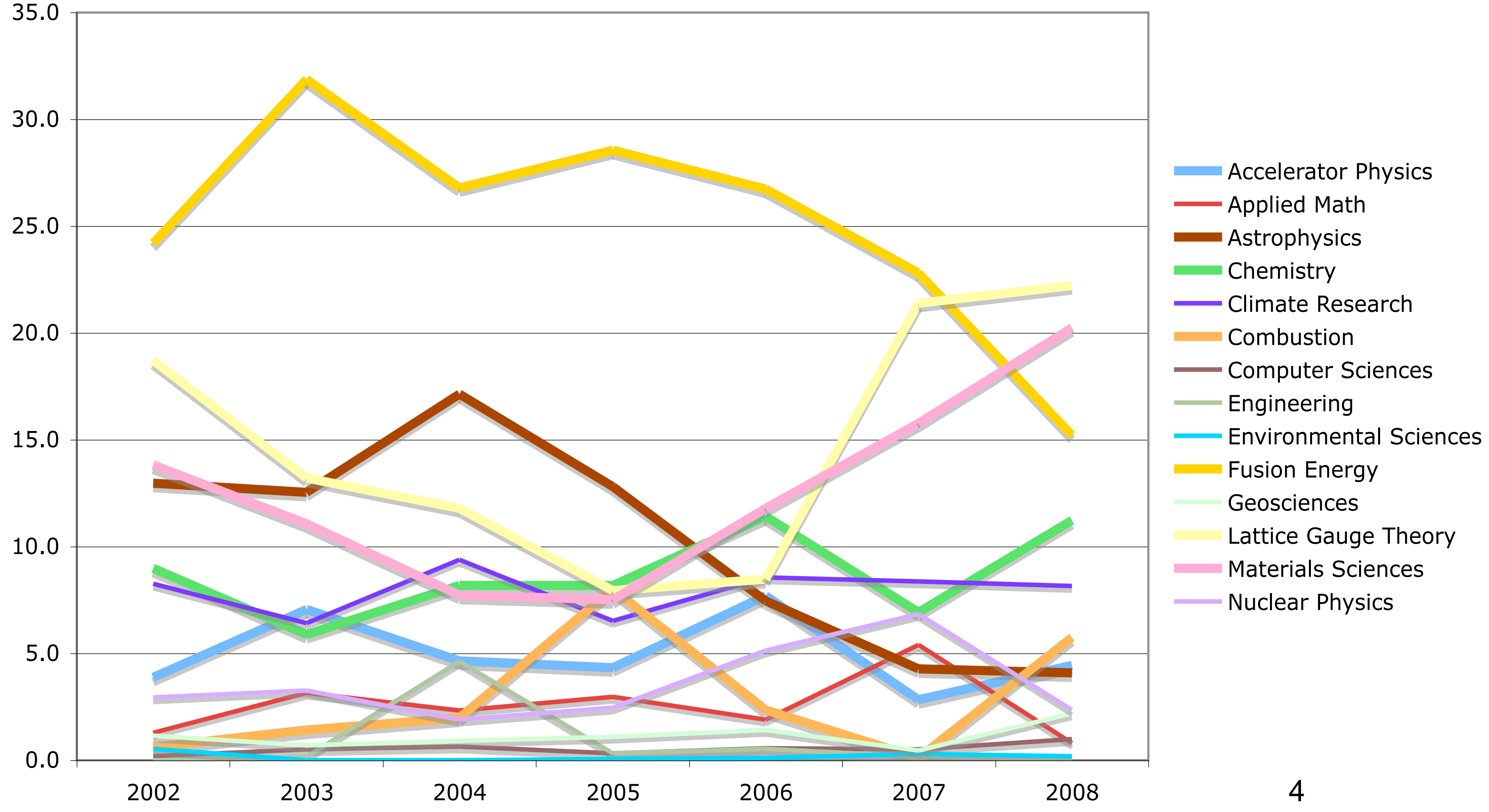


Outline

- Brief NERSC Workload and Usage Overview
- NERSC-6 Goals
- RFP Details
- Highlights between NERSC-5 and NERSC-6
- Benchmarks
- IO and NGF
- System Integration

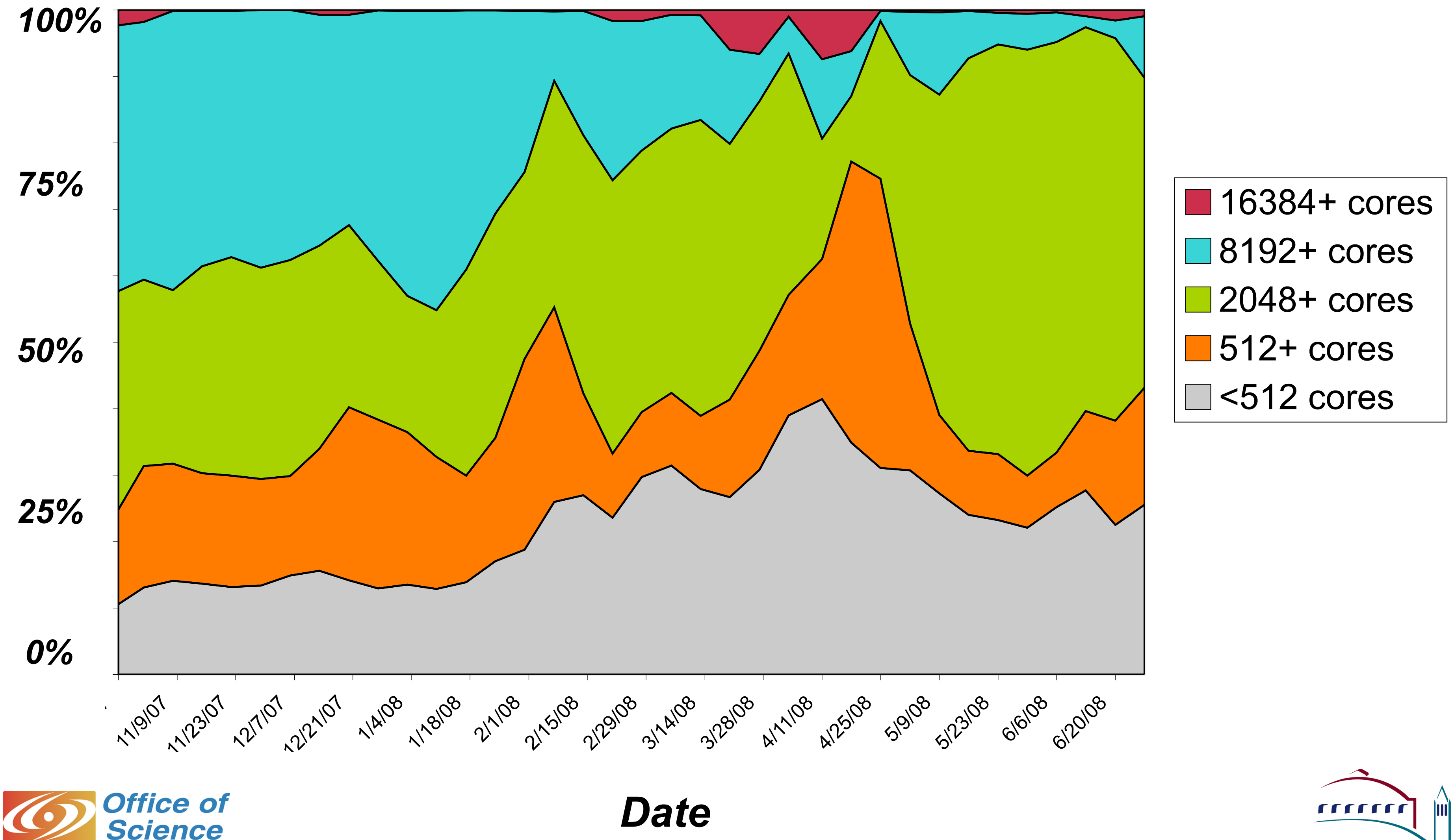
Science Priorities are Variable

Usage by Science Type as a Percent of Total Usage



Computing at Scale

Concurrency on Franklin by % of Raw Hours



NERSC-6 System Goals

- A complete, integrated computing system for a multi-user, multi-application parallel scientific workload.
  - Reliable compute resources
  - Meet the needs of the entire computational workload
  - NERSC applications often need, in different ratios,
    - Memory bandwidth
    - Low latency and high bandwidth interconnect
    - High performance networks
    - High speed CPUs
  - A standards-compliant application program development environment
  - Robust and scalable system administration environment
  - Capabilities integrated into complete and supported product by vendor

NERSC-6 System Goals

- Significantly increase computational resources available to users using measured performance criteria
  - Arrives FY 2009
  - Significant impact for NERSC users in 2010 allocation year
    - January 2010 to January 2011
  - What is significant?
    - Webster - “a noticeably or measurably large amount”
  - When is does it start?
    - That depends on how much of an impact the new system has.
    - The requirement does not mean it has be all there January, 2010.
    - It does mean the system cannot just squeak in on December 31, 2010.

NERSC-6 System Goals

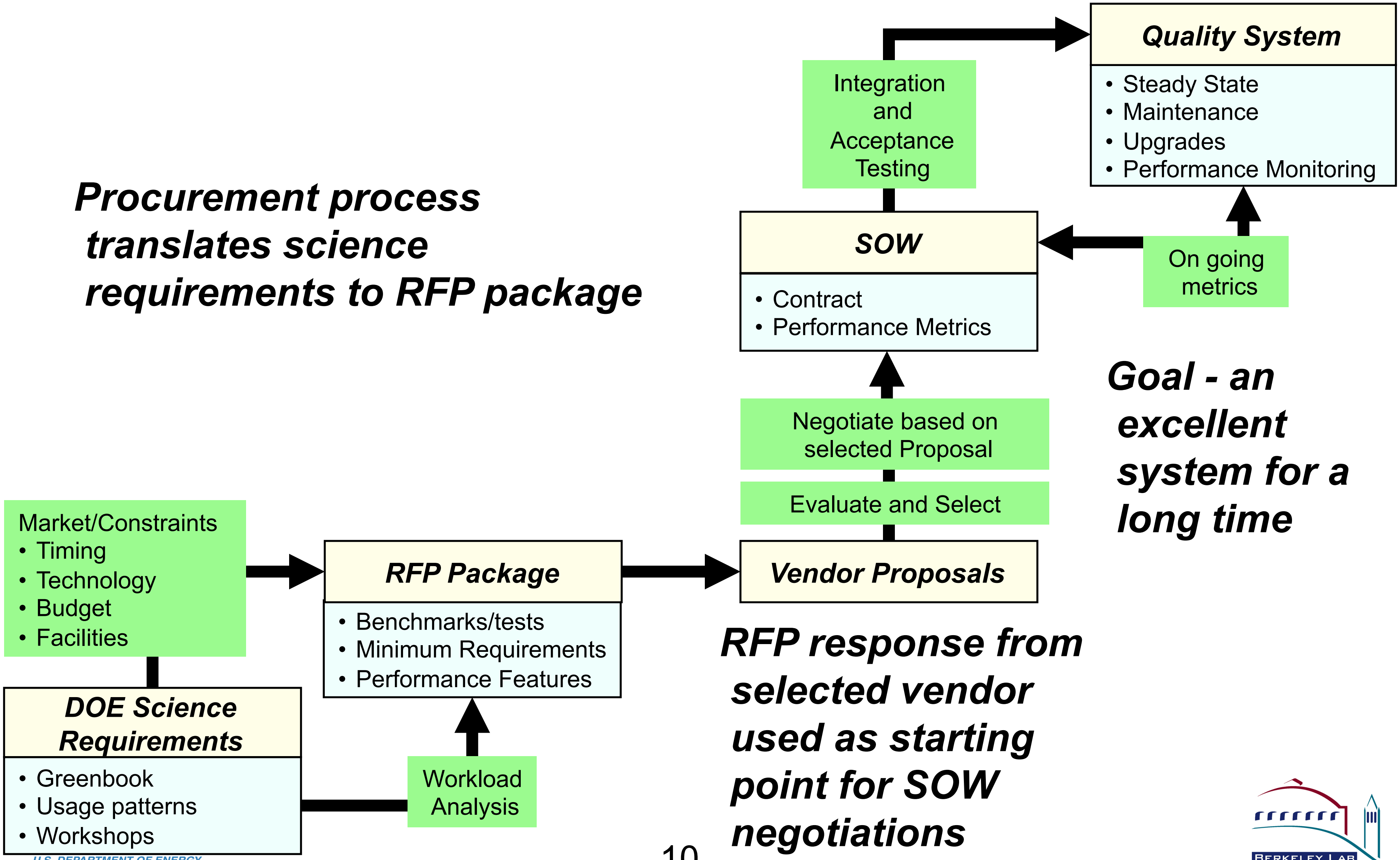
- Sustained System Performance (SSP) over 3 years
  - Goal is 70-100 Tflops/s average over the first three years based on the NERSC-6 SSP Suite
- System balance is emphasized as part of BVSS
  - Aggregate memory
  - Global usable disk storage
  - Interconnect bandwidth and latency
  - Storage capacity and bandwidth
  - Network bandwidth
- Integrate with the NERSC environment
- Installed at current OSF
- Add best value to existing NERSC resources

NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

Key Aspects of the NERSC-6 Approach

- Algorithm targets for today and in the future
  - New thrust areas for workload targets
    - e.g., One sided communication, AMR, sparse problems, implicit solves
- Exclusive emphasis on sustained performance for the breadth of the science workload
- Emphasis on user viewpoints
- PERCU for holistic assessment of system
- Acknowledges and addresses system complexity
- Designed to deliver a system ready to run the entire production workload by the time acceptance is complete – "The End Game"
- Provides flexibility for NERSC and for vendors to be innovative in technology, in time, in risk, and in cost based on their time line

The Process: Translate DOE Science Requirements to a Great System



DOE Science Requirements
- Greenbook
- Usage patterns
- Workshops

Market/Constraints
- Timing
- Technology
- Budget
- Facilities

Workload Analysis

RFP Package
- Benchmarks/tests
- Minimum Requirements
- Performance Features

Vendor Proposals

Evaluate and Select

Negotiate based on selected Proposal

SOW
- Contract
- Performance Metrics

Integration and Acceptance Testing

Quality System
- Steady State
- Maintenance
- Upgrades
- Performance Monitoring

On going metrics

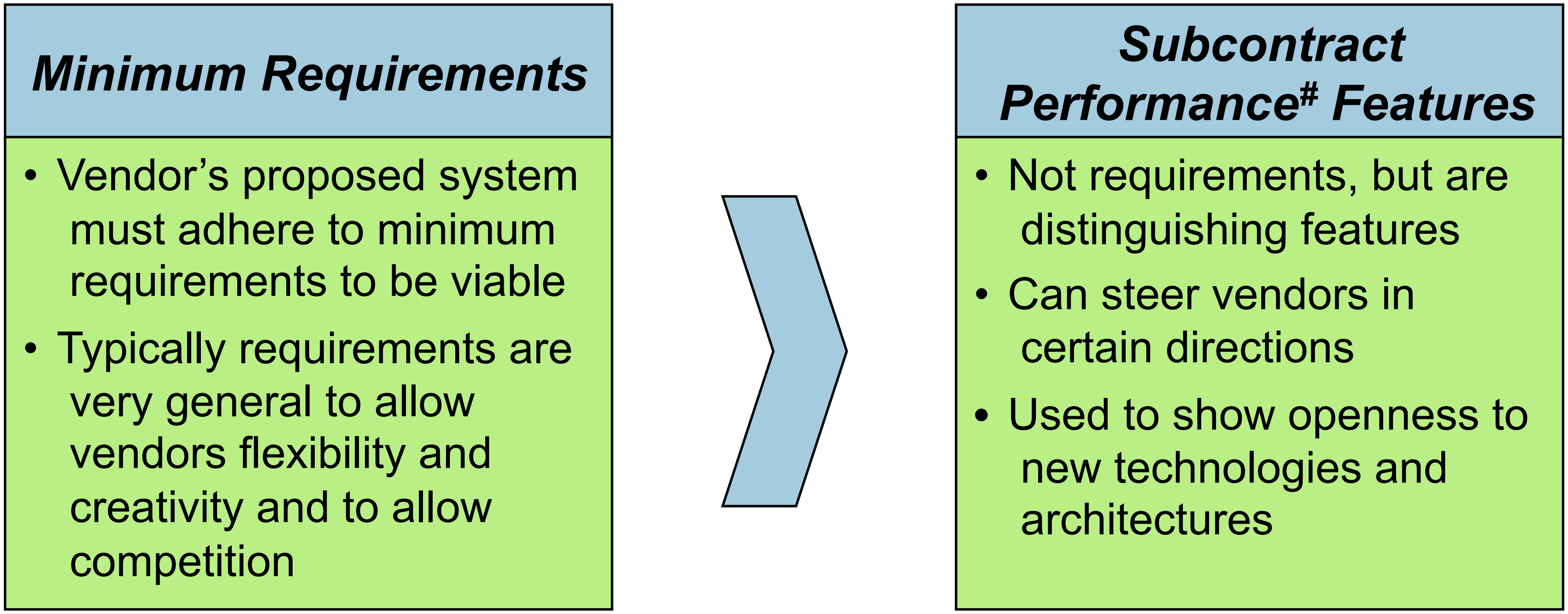
Procurement process translates science requirements to RFP package

Goal - an excellent system for a long time

RFP response from selected vendor used as starting point for SOW negotiations

BVSS RFP Process

NERSC's procurement process allows vendors to propose their best solution given a set of requirements and performance features.



Minimum Requirements

- Vendor's proposed system must adhere to minimum requirements to be viable
- Typically requirements are very general to allow vendors flexibility and creativity and to allow competition

Subcontract Performance[#] Features

- Not requirements, but are distinguishing features
- Can steer vendors in certain directions
- Used to show openness to new technologies and architectures

[#] "Performance features" is a BVSS procurement term that relates to criteria in a Performance Based Contract which are beyond minimum requirements (sometimes also known as "desired" or "value-added" features)

The PERCU Method
What Users Want

- **Performance**
  - **How fast will a system process work if everything is working really well**
- **Effectiveness**
  - **The likelihood users can get the system to do their work when they need it**
- **Reliability**
  - **The likelihood the system is available to do the user’s work**
- **Consistency**
  - **How often the system processes the same or similar work correctly and in the same length of time**
- **Usability**
  - **How easy is it for users to get the system to process their work as fast as possible**

PERCU

RFP Details

Minimum Requirements

- General
  - The system shall consume no more than 3.3 MW of electrical power including cooling.
  - The system shall be capable of using 480v, 3 phase power.
- Performance
  - The proposal must state a minimum Sustained System Performance (SSP) for the proposed system, as measured by the SSP metric.
  - A high-performance, high-bandwidth, low-latency, fault -tolerant interconnect with scalable performance characteristics over the entire system.
  - 10 Gigabit Ethernet connectivity to NERSC infrastucture.
- Effectiveness
  - An application development environment consisting of at least: standards compliant Fortran, C, and C++ compilers, and MPI and MPI-IO libraries.
  - Ability to run a single application instance over all the compute nodes in the system.

Minimum Requirements

- Reliability
  - Comprehensive maintenance and 24x7 support for all hardware and software components including providing all replacement / spare parts.
- Consistency
  - Consistent and reproducible execution times in dedicated and production mode.
  - Correct, consistent and reproducible computation results.
- Usability
  - Compliance with 64-bit IEEE 754 floating point arithmetic.
  - An external parallel filesystem, administered independently from the computational nodes, with a single, unified namespace and high parallel I/O performance for all user data. All system shared storage and storage fabric shall be standards-based and packaged independently. Acceptable standards are Fibre Channel, Ethernet, and/or InfiniBand.
  - All components comprising the external filesystem supplied by the system Offeror shall be compatible with NERSC's NGF system.

Performance Features

- General
  - An integrated system providing for a concurrent multi-user, multi-application parallel scientific workload.
  - Energy efficient computing, power distribution and cooling, including the ability of the system to operate permanently and successfully at the higher levels of the American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE) Recommended Thermal Range and within the ASHRAE allowable rate of change.
  - Ease and minimal cost of integration into the existing NERSC facility, including the use of ISO-Base technology for seismic protection.
  - Credible roadmap for future hardware and software products and support thereof.
  - Ease of expandability and configurability in terms of CPUs, memory, storage and interconnect.

Performance Features

- Performance
  - Documented performance characteristics and RFP benchmark results.
  - Large amount of aggregate user addressable memory with automated error detection and correction.
  - Support for advanced programming languages such as UPC, CAF, the emerging HPCS languages, shared memory abstractions such as Global Arrays through one-sided messaging (e.g., put/get remote memory access semantics), efficient RDMA support, and/or global addressing.
  - High sustained parallel and single stream filesystem I/O bandwidth to and from global shared storage system.
  - A global shared storage system of at least 1 PB user usable disk space, providing at least 70 GB/s of measurable, sustained aggregate filesystem I/O bandwidth between the external parallel filesystem and the computational nodes.
  - High efficiency for both large and small block I/O for shared file access with high concurrency. This includes support for new parallel I/O interfaces such as pNFS.
  - Tight integration with 3rd party hardware and software, in particular the filesystem software, to improve MPI-IO and POSIX performance. This could include auto tuning parameters.
  - High sustained aggregate external network bandwidth including upgrade path to 100 Gigabit Ethernet.

Performance Features

- Effectiveness
  - A scalable, robust, effective and comprehensive system administration and resource management environment.
  - Ability to effectively manage system resources with high utilization and throughput under a workload with a wide range of concurrencies.
  - Minimal intrusion upon memory available to application data structures by system libraries, daemons, operating system and/or kernel. Document the amount of memory used by these various system components on the compute nodes of the system including jobs running at full system concurrency, and indicate if the space is permanently resident in physical memory.
  - High performance, well-integrated MPI implementation that delivers a high percentage of hardware performance for the NERSC workload.
  - Support within MPI (and underlying hardware) for accelerated collective operations such as dedicated collective networks and other acceleration technology.
  - Advanced resource management functionality; e.g., checkpoint-restart, job migration, backfill, advanced and persistent reservations, job preemption and architecture aware job placement.

Performance Features

- Reliability
  - Demonstrated ability to produce and maintain the proposed system.
  - Commitment to achieving specific quality assurance, reliability and availability goals.
  - A clear plan documenting how the Offeror will effectively respond to software defects and system outages at each severity level, and how a problem or defect will be escalated if not fixed in a timely manner.
  - An effective methodology for system upgrades, repairs and testing. Provide a description of how issues of system availability and user productivity are addressed by the methodology.
  - Fault resilience and detailed fault monitoring, reporting, and prediction. Capability of the system to fail gracefully with failures in one part of the system not impacting jobs running on other portions of the system.
  - Demonstrated ability to track errors and analyze failures for all software and hardware components.
  - In-house testing and problem diagnosis capability, including hardware resources at appropriate scale.

Performance Features

- Consistency
  - Minimal intrusion on CPU resources available to application processes by system libraries, daemons, operating system and/or kernel. Document the major sources of CPU use by these various system components on the compute nodes of the system, and provide an estimate on the percentage of CPU time used by such functions when applications are running on the system.
    - For the purposes of this procurement, CPU = core = processor.
  - Architectural features to improve application scaling and decrease system jitter or synchronization mismatches across socket and node boundaries.

Performance Features

- Usability
  - User access to performance counters on the CPU, node, storage subsystems and interconnect via a documented API including a PAPI interface to the performance counters.
  - Support for centralized configuration management and change management.
  - Capability for remote administration including hardware reset, power management, booting, and remote console.
  - Fully featured application development environment, including: vendor optimized serial and parallel scientific libraries (e.g., LAPACK, BLAS); MPMD MPI; Python; GNU tools and utilities; a parallel debugger such as Totalview and Allinea’s DDT; Standards compliant MPI-2 and OpenMP (if appropriate); performance profiling and tuning tools.
  - Accounting and activity tracking functionality, e.g., job containers, which assist in job, session and Unix process tracking for security and resource management purposes.

Performance Features

- Usability (continued)
  - Online documentation of all system software and hardware available to NERSC staff and online documentation of all user-visible system features available to all NERSC users.
  - On-site training for NERSC system management and user support staff.
  - Ability to integrate with grid environments running current software implementations, for example, the Virtual Data Toolkit (http://vdt.cs.wisc.edu).
  - A plan for integrating, supporting, achieving and maintaining high performance parallel access to the NGF system.
  - Engineering assistance with the re-allocation of storage hardware from the NERSC-6 system to the NGF system. Maintenance and required licenses will continue on the storage hardware.

Highlights: New or Changed from NERSC-5

- **Minimum Requirements**
  - *No more than 3.3 MW of electrical power including cooling.*
  - *Use 480v, 3 phase power.*
- **Performance Features**
  - General
    - *Ability of the system to operate permanently and successfully at the higher levels of the ASHRAE* Recommended Thermal Ranges and rate of change.*
  - Performance
    - Advanced programming *languages such as UPC,* CAF, the emerging HPCS languages….
    - *High efficiency for both large and small block I/O for shared file access at high concurrency.*
    - *New parallel I/O interfaces such as pNFS.*

** American Society of Heating, Refrigerating and Air Conditioning Engineers*

Highlights: New or Changed from NERSC-5

- **Performance Features (Cont.)**
  - **Reliability**
    - ***An effective methodology for system upgrades, repairs and testing.***
    - ***Ability to track errors and analyze failures for all software and hardware***
    - ***In-house testing and problem diagnosis capability, including resources at appropriate scale.***
  - **Consistency**
    - **Minimal intrusion on CPU resources from system related SW.** ***Document the major sources … and estimate on the percentage of CPU time used by such functions***
  - **Usability**
    - **Fully featured application development environment, including: …** ***Python***

- **Supplier Attributes**
  - **Management and corporate capability,** ***including identifying and managing risk throughout the NERSC-6 project.***

Benchmarks

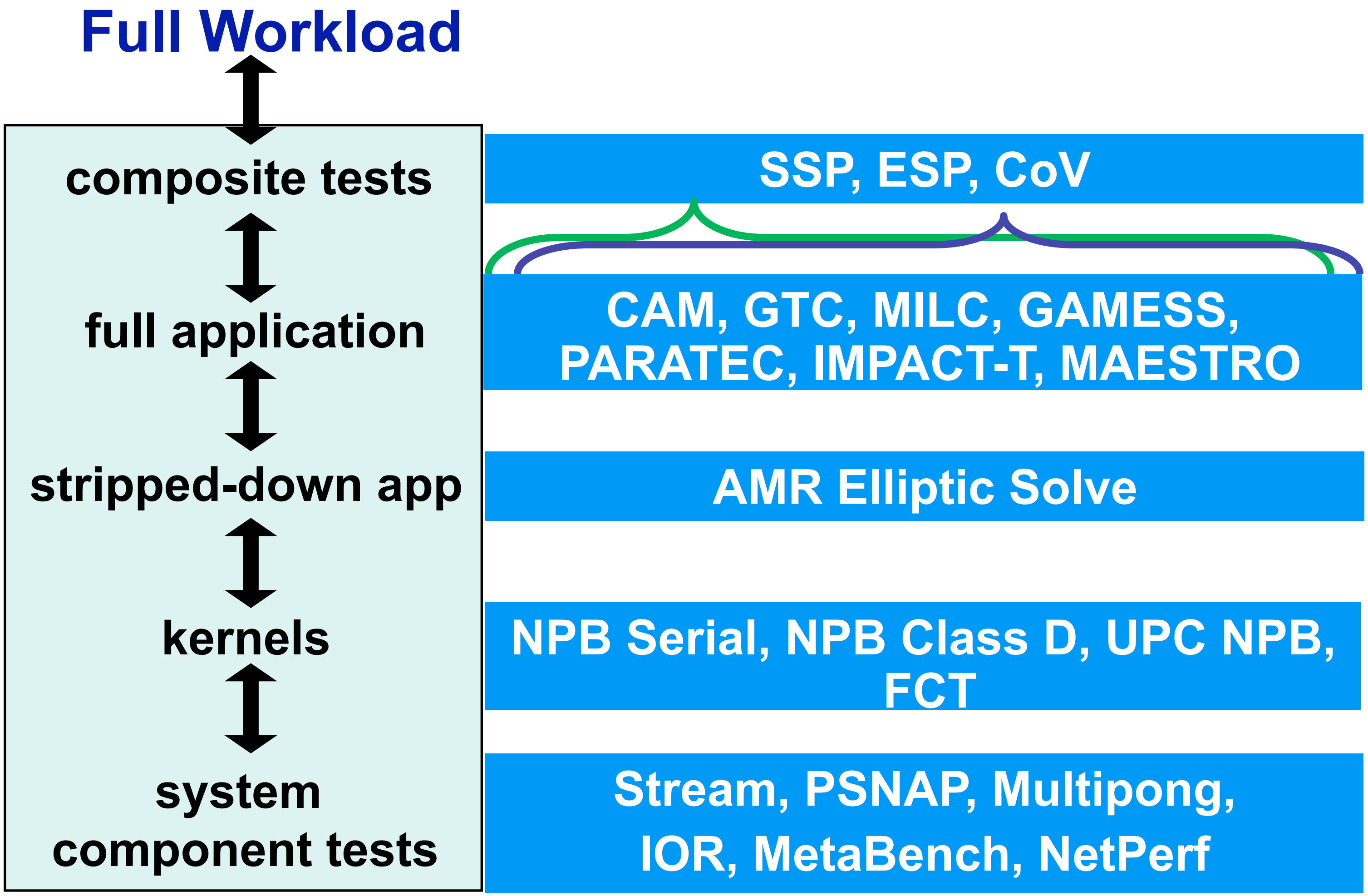
Benchmarks Serve 3 Critical Roles

- Carefully chosen to represent characteristics of the expected NERSC-6 workload

- Give vendors the opportunity to provide NERSC with concrete performance and scalability data.
  - Measured or projected.

- Part of the acceptance test and a measure of performance throughout the operational lifetime of NERSC-6.

NERSC-6 Benchmark Implementation



Full Workload

composite tests

full application

stripped-down app

kernels

system component tests

SSP, ESP, CoV

CAM, GTC, MILC, GAMESS, PARATEC, IMPACT-T, MAESTRO

AMR Elliptic Solve

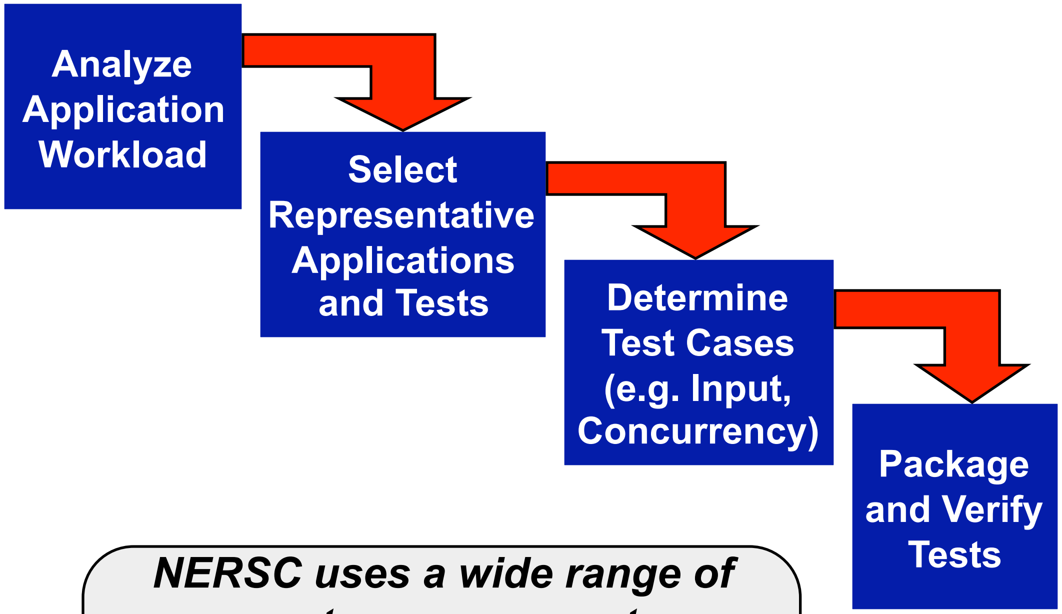
NPB Serial, NPB Class D, UPC NPB, FCT

Stream, PSNAP, Multipong, IOR, MetaBench, NetPerf

Performance Obligations

- Selected Offeror is required to meet benchmark performance levels reported in the RFP response as a condition of acceptance
  - and throughout the life of the subcontract.
- Includes all SSP apps (with all inputs), all lower-level tests, SSP, ESP, FCT, and dedicated & production CoV.
- Offeror may be required to demonstrate other performance metrics as part of a negotiated SOW.
- Benchmark run rules are incorporated into SOW.

Benchmark and Test Hierarchy



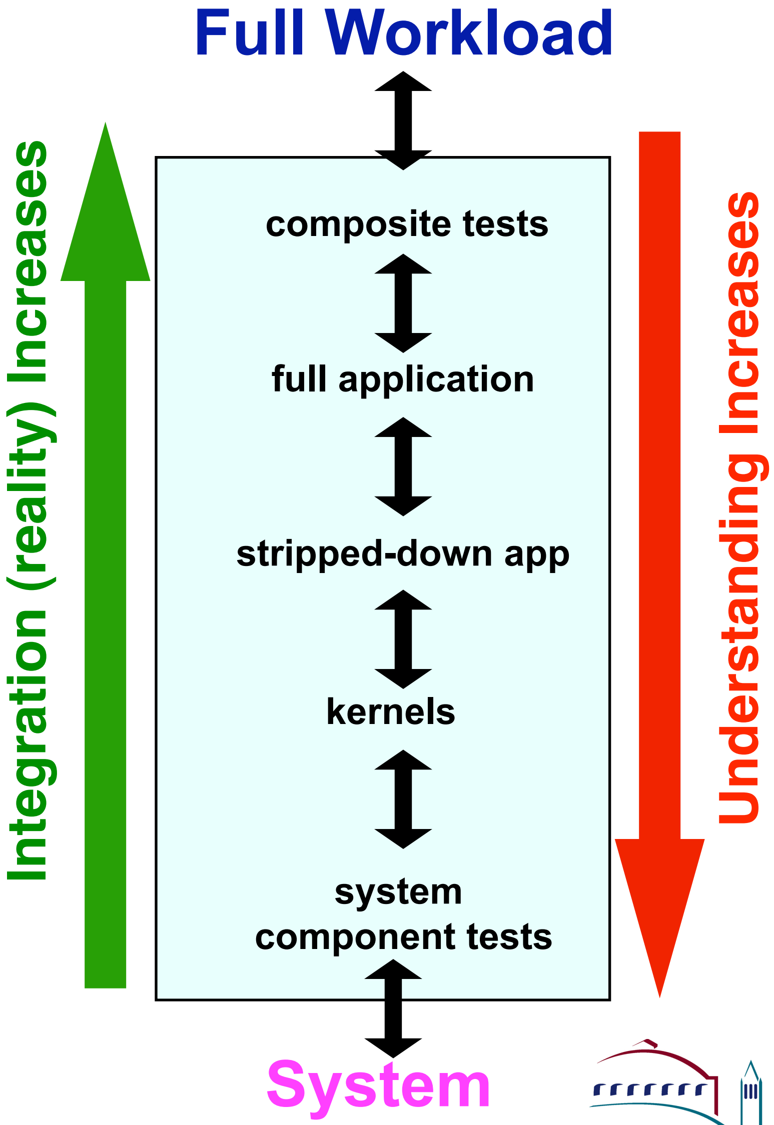
Analyze Application Workload

Select Representative Applications and Tests

Determine Test Cases (e.g. Input, Concurrency)

Package and Verify Tests

NERSC uses a wide range of system component, application, and composite tests to characterize the performance and efficiency of a system



Full Workload

composite tests

full application

stripped-down app

kernels

system component tests

System

Integration (reality) Increases

Understanding Increases

Lower-Level Benchmarks

CODE	PURPOSE / DESCRIPTION
STREAM	Single- and multi-core memory bandwidth.
FCT	Full-Configuration Test, run a single app over all cores; FFT mimics planewave DFT codes.
PSNAP	FWQ operating system noise test.
NAS PB serial & 256-way MPI	Serial application performance on a single **packed** node; measures memory BW/ computation rate balance and compiler capabilities. **Packed** means all cores run.
NAS PB UPC	Measure performance characteristics not visible from MPI for FT benchmark.
Multipong	NERSC MPI PingPong for "latency" and BW, nearest- and furthest nodes in topology; also intra-node.
AMR Elliptic	C++/F90 LBNL Chombo code; proxy for AMR Multigrid elliptic solvers; 2 refinement levels; weak scaling with geometry replication; very sensitive to OS noise;

Full Applications

Algorithm Diversity

Algorithm *Science areas*	*Dense linear algebra*	*Sparse linear algebra*	*Spectral Methods (FFT)s*	*Particle Methods*	*Structured Grids*	*Unstructured or AMR Grids*	*Data Intensive*
Accelerator Science		X	X	X	X	X	
Astrophysics	X	X	X	X	X	X	X
Chemistry	X	X	X	X			X
Climate			X		X	X	X
Combustion					X	X	X
Fusion	X	X		X	X	X	X
Lattice Gauge		X	X	X	X		
Material Science	X		X	X	X		

NERSC users require a system which performs adequately in all areas

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Algorithm Diversity

Algorithm Science areas	Dense linear algebra	Sparse linear algebra	Spectral Methods (FFT)s	Particle Methods	Structured Grids	Unstructured or AMR Grids	Data Intensive
Accelerator Science		X	X	X	X	X	
Astrophysics	X	X	X	X	X	X	X
Chemistry	X	X	X	X			X
Climate			X		X	X	X
Combustion					X	X	X
Fusion	X	X		X	X	X	X
Lattice Gauge		X	X	X	X		
Material Science	X		X	X	X		

High Flop/s rate

High performance memory system

High bisection bandwidth

High performance memory system

High flop/s rate

Low latency, efficient gather /scatter

Storage, Network Infrastructure

NERSC users require a system which performs adequately in all areas

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Full Application Characteristics

Benchmark	*Science Area*	*Algorithm Space*	*Base Case Concurrency*	*Problem Description*	*Memory*	*Lang*	*Libraries*
CAM	Climate (BER)	Navier Stokes CFD	56, 240 Strong scaling	D Grid, (~.5 deg resolution); 240 timesteps	0.5 GB *per MPI task*	F90	netCDF
GAMESS	Quantum Chem (BES)	Dense linear algebra	256, 1024 (Same as TI-09)	DFT gradient, MP2 gradient	~2GB *per MPI task*	F77	DDI, BLAS
GTC	Fusion (FES)	PIC, finite difference	512, 2048 Weak scaling	100 particles per cell	.5 GB*per MPI task*	F90	
IMPACT-T	Accelerator Physics (HEP)	PIC, FFT component	256,1024 Strong scaling	50 particles per cell	1 GB *per MPI task*	F90	
MAESTRO	Astrophysics (HEP)	Low Mach Hydro; block structured-grid multiphysics	512, 2048 Weak scaling	16 32^3 boxes per proc; 10 timesteps	800-1GB *per MPI task*	F90	Boxlib
MILC	Lattice Gauge Physics (NP)	Conjugate gradient, sparse matrix; FFT	256, 1024, 8192 Weak scaling	8x8x8x9 Local Grid, ~70,000 iterations	210 MB *per MPI task*	C, assem.	
PARATEC	Material Science (BES)	DFT; FFT, BLAS3	256, 1024 Strong scaling	686 Atoms, 1372 bands, 20 iterations	.5 -1GB *per MPI task*	F90	Scalapack, FFTW

For strong scaling cases memory per MPI task is for the small problem. Each problem represents one of a spectrum of inputs.

NERSC-6 Benchmarks Coverage

Science areas	*Dense linear algebra*	*Sparse linear algebra*	*Spectral Methods (FFT)s*	*Particle Methods*	*Structured Grids*	*Unstructured or AMR Grids*
Accelerator Science			IMPACT-T	IMPACT-T	IMPACT-T	
Astrophysics		MAESTRO			MAESTRO	MAESTRO
Chemistry	GAMESS					
Climate			CAM		CAM	
Combustion					MAESTRO	AMR Elliptic
Fusion				GTC	GTC	
Lattice Gauge		MILC	MILC	MILC	MILC	
Material Science	PARATEC		PARATEC		PARATEC	

Base Case for Application Runns

- A basis for comparison among proposed systems.
- Limits the scope of optimization.
  - Modifications only to enable porting and correct execution.
- Limits allowable concurrency to prescribed values.
- MPI only for all codes (even if OpenMP directives present).
- Fully packed nodes.
- Libraries okay (if generally supported).
- Hardware multithreading okay, too.
  - Expand MPI concurrency to occupy hardware threads.

Optimized Case for Application Runs

- Allows the Offeror to highlight features of the proposed system.
- Applies to seven SSP apps only, all test problems.
- Examples:
  - Unpack the nodes;
  - Higher or lower concurrency than corresponding base case;
  - Hybrid OpenMP / MPI;
  - Source code changes for data alignment / layout;
  - Any / all of above.
- Caveat: number of tasks used to calculate SSP must use the total number of processors blocked from other use.

Highlight: New or Changed Benchmarks from NERSC-5

- Two new application benchmarks address the evolution of the workload and algorithms
  - MAESTRO and IMPACT-T
- High concurrency
  - Largest increases from 2,048 to 8,196
  - New problem sets
  - Increased focus on strong scaling
- Benchmarks for emerging programming models and algorithms
  - UPC, AMR, implicit and sparse methods
- Two ways for vendors to run benchmarks, as mentioned in previous slides
  - Optional Optimized Case (“Full Fury”) allows vendors to achieve their best case performance
    - Allows dramatic changes (e.g any amount of code change, concurrency, libraries…)
    - As long as users can achieve similar performance with the same methods in a production mode
    - Requires same inputs, same answers

PARATEC: Parallel Total Energy Code

- **Authors**: LBNL + UC Berkeley.
- **Relation to NERSC Workload**
  - Represents / captures the performance of a wide range of codes (VASP, CPMD, PETOT, QBox).
  - 70% of NERSC MatSci computation done via Planewave DFT codes.
- **Description**: Planewave DFT; calculation in both Fourier and real space; has custom 3-D FFT to transform between.
- **Coding**: 50,000 lines of Fortran90; uses SCALAPACK / FFTW / BLAS3; vectorizable version.
- **Parallelism**: fine-grain parallelism over DF grid points via MPI.
- **NERSC-6 tests**: strong scaling on 256 and 1024 cores.
- **Profile**: all-to-all data transpositions dominate communications time; good differentiation between systems.
- **Special**: Also used for NSF Trac-I/II benchmarking.

CAM3: Community Climate Model

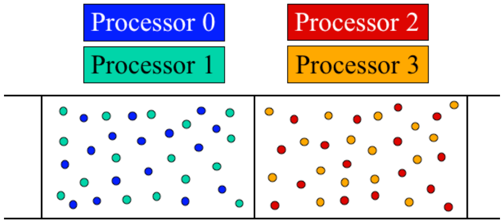
- **Authors**: NCAR + substantial DOE scientific and software input.
- **Relation to NERSC Workload**
  - Atmospheric part of CCSM; most timing consuming part.
  - Wide American and foreign scientist usage for climate research.
    - e.g., Carbon, bio-geochemistry models are built upon (integrated with) CAM3.
    - IPCC predictions will use CAM3 (in part).
- **Description**: explicit time integration dynamics + subgrid-scale physics + data movement between.
- **Coding**: ~100K lines Fortran90 + 7K lines of C.
- **Parallelism**: Two, 2-D processor decompositions; 0.5 deg (60km) resolution grid; hybrid parallelism via MPI; OpenMP possible for optimized runs.
- **NERSC-6 tests**: strong scaling, 56 and 240 cores; 6 GB of input data files; finite-volume dycore.

GAMESS: Computational Chemistry

- **Authors**: DOE Ames Lab, Iowa St + many others.
- **Relation to NERSC Workload**
  - Quantum chemical computations for Chemistry / Mat Sci / Life Sci.
  - Representative of codes exposing communication performance characteristics not visible from MPI.
- **Description**: General purpose electronic structure code.
- **Coding**: ~ 500,000 lines Fortran77; Can use highly optimized vendor libraries for communications and for BLAS; NERSC-6 updated to R6 (March 2007) source.
- **Parallelism**: SPMD + its own comm library for GA abstraction.
- **NERSC-6 tests**: Chose two electronic structure methods.
  - DFT energy and gradient calculation on 256 processors;
  - MP2 energy and gradient calculation on 1024 processors;
  - Intended to be same as DOD HPCMP TI-09 GAMESS benchmark;
  - Extremely useful during the implementation phase of the acquisition due to its complexity.

GTC: 3-D Gyrokinetic Toroidal Code

- **Authors**: S. Ethier, et al. (PPPL); Z. Lin (UC Irvine)
- **Relation to NERSC Workload**
  - Physics of burning plasmas in magnetically confined fusion experiments such as TFTR and NSTX, and ITER.
  - Especially for modeling turbulent transport, one two-orders-of-magnitude regime of the 10^14 fusion timescale range.
- **Description**: 3D particle-in-cell, toroidal geometry, iterative Poisson solver.
- **Coding**: Using version obtained from S. Ethier in Feb 07.
  - Portable RNG for initial particle distribution.
- **Parallelism**: 1-D domain decomposition plus 1-D particle decomposition within each domain.
- **NERSC-6 tests**: weak scaling on 512 & 2048 cores; 100 particles per cell; inputs for other sizes included; scales to 16k cores.



IMPACT-T: Accelerator Science

- Author: J. Qiang, et al., LBNL Accelerator & Fusion Research Div.
- Relation to NERSC Workload
  - DOE High Energy Physics (HEP) and Nuclear Physics (NP) programs, plus SciDAC COMmunity Petascale Project for Accelerator Science and Simulation.
  - Part of a suite of codes, IMPACT-Z, Theta, Fix2d/3d, others.
  - Wide variety of science drivers/approaches/codes: Accelerator design, electromagnetics, electron cooling, advanced acceleration.
- Description: 3-D PIC, quasi-static, integrated Green Function, moving beam frame; FFT Poisson solver.
- Coding: 33,000 lines of object-oriented Fortran90.
- Parallelism: 2-D decomposition, MPI; frequent load-rebalance based on domain.
- NERSC-6 tests: photoelectron beam transported through a photoinjector similar to one at SLAC; strong scaling on 256 and 1024 cores; 50 particles per cell.

MAESTRO: Low Mach Number Flow

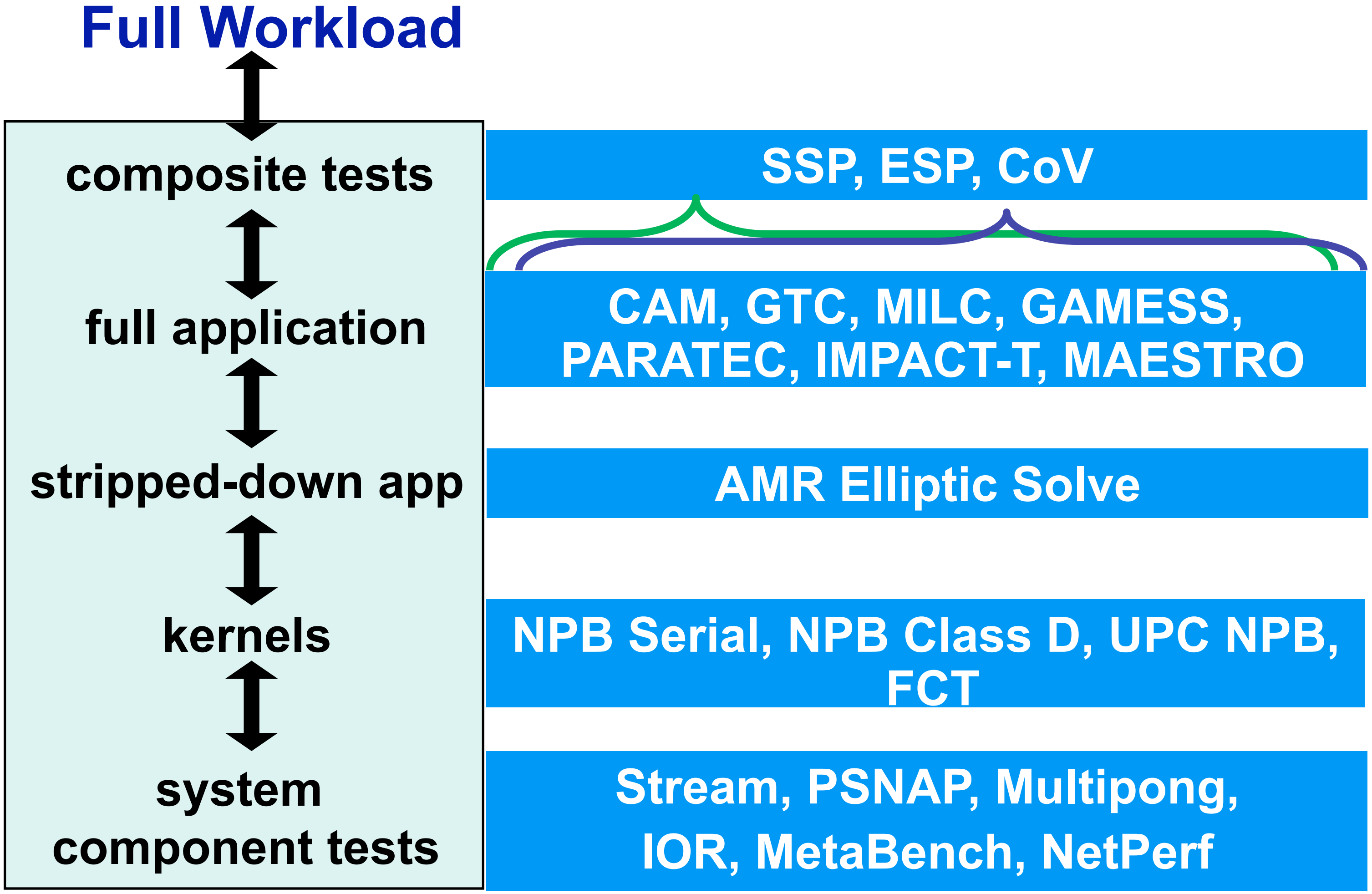
- **Authors: LBNL Computing Research Division; SciDAC07**
- **Relation to NERSC Workload**
  - **Model convection leading up to Type 1a supernova explosion;**
  - **Method also applicable to 3-D turbulent combustion studies.**
- **Description: Structured rectangular grid plus patch-based AMR (although NERSC-6 code does not adapt);**
  - **hydro model has implicit & explicit components;**
- **Coding: ~ 100,000 lines Fortran 90/77.**
- **Parallelism: 3-D processor non-overlapping decomposition, MPI.**
  - **Knapsack algorithm for load distribution; move boxes close in physical space to same/close processor.**
  - **More communication than necessary but has AMR communication characteristics.**
- **NERSC-6 tests: weak scaling on 512 and 2048 cores; 16 boxes (32^3 cells each) per processor.**

MILC: MIMD Lattice Gauge QCD

- **Authors**: MILC collaboration, especially S. Gottlieb
- **Relation to NERSC Workload**
  - Funded through High Energy Physics Theory
  - Understand results of particle and nuclear physics experiments in terms of Quantum Chromodynamics
- **Description**: Physics on a 4D lattice, CG algorithm, sparse 3x3 complex matrix multiplies - highly memory bandwidth intensive.
- **Coding:**
  - V7; ~ 60,000 lines of C; POWER and x86 assembler (Cray redid for Opteron DC & QC); wants gcc.
  - Extensive hard-coded prefetch;
  - CG algorithm with MPI_Allreduce
- **Parallelism**: 4-D domain decomposition, MPI.
- **NERSC-6 tests**: weak scaling, 8x8x8x9 local lattice, emphasize CG iterations.

Composite Tests

NERSC-6 Benchmark Implementation



Full Workload

composite tests	SSP, ESP, CoV
full application	CAM, GTC, MILC, GAMESS, PARATEC, IMPACT-T, MAESTRO
stripped-down app	AMR Elliptic Solve
kernels	NPB Serial, NPB Class D, UPC NPB, FCT
system component tests	Stream, PSNAP, Multipong, IOR, MetaBench, NetPerf

Sustained System Performance (SSP)

- Geometric Mean of the processing rates of seven applications multiplied by *N*, # of cores in the system.
  - Largest base case concurrencies used.
- Aggregate, un-weighted measure of computational capability relevant to achievable scientific work.
- Uses floating-point operation count predetermined on a reference system by NERSC.

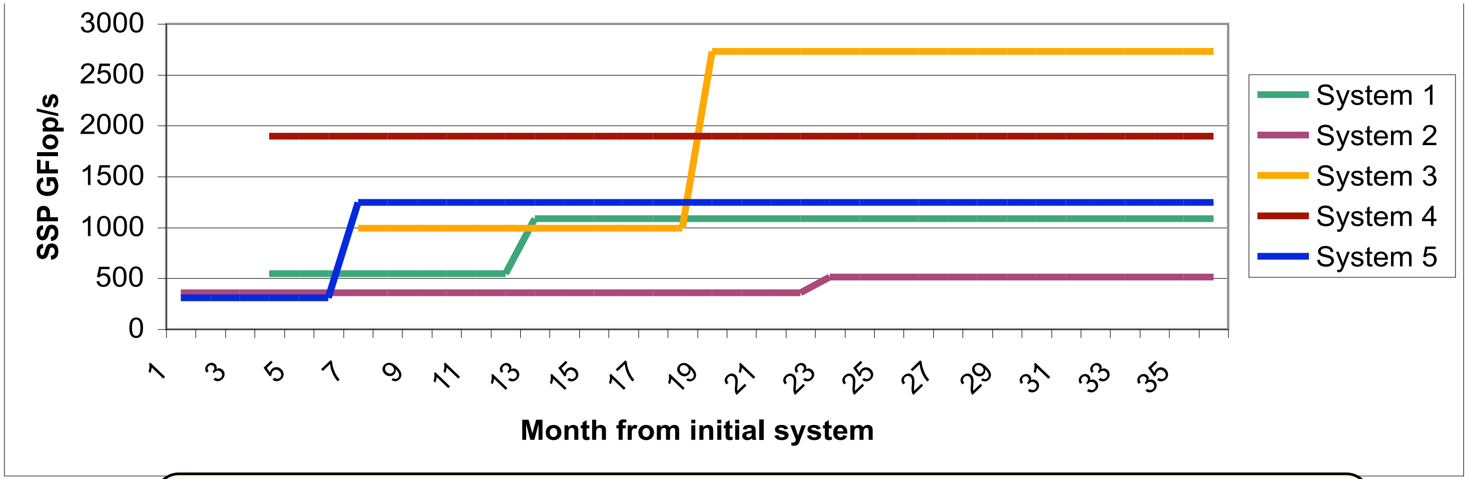
$$\text{SSP in TFLOPS} = \frac{N * \sqrt[7]{\prod_i P_i}}{1000}$$

Key Point - Sustained System Performance (SSP) Over Time

- Measures mean flop rate of applications integrated over time period
- SSP can change due to
  - System upgrades, Increasing # of cores, Software Improvements
- Allows evaluation of systems delivered in phases
- Takes into account delivery date
- Produces metrics such as SSP/Watt and SSP/$

$$Value_s = \frac{Potency_s}{Cost_s}$$

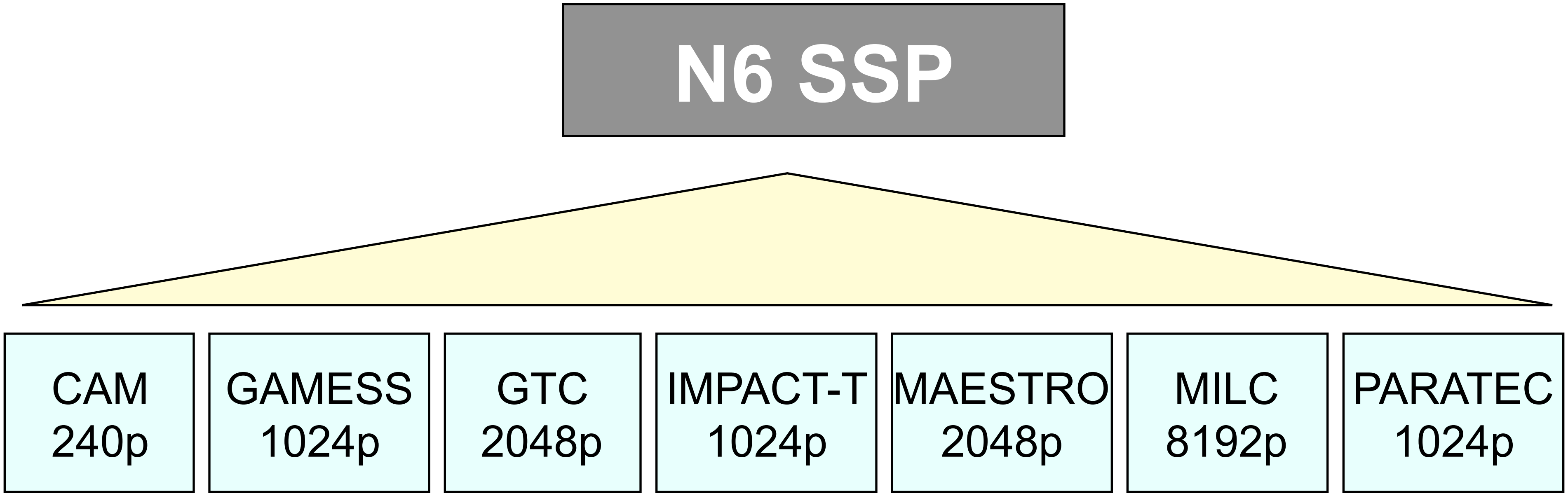
SSP Over 3 Year Period for 5 Hypothetical Systems



Area under curve, when combined with cost, indicates system 'value'

N6 Composite SSP Metric

The largest concurrency run of each full application benchmark is used to calculate the composite SSP metric



N6 SSP

CAM 240p

GAMESS 1024p

GTC 2048p

IMPACT-T 1024p

MAESTRO 2048p

MILC 8192p

PARATEC 1024p

For each benchmark measure

- *FLOP counts on a reference system*
- *Wall clock run time on various systems*

Example of N6 SSP on Hypothetical System

Hypothetical N6 System			Results	
	Tasks	System Gflopcnt	Time	Rate per Core
CAM	**240**	**57,669**	**408**	0.589
GAMESS	**1024**	**1,655,871**	**2811**	0.575
GTC	**2048**	**3,639,479**	**1493**	1.190
IMPACT-T	**1024**	**416,200**	**652**	0.623
MAESTRO	**2048**	**1,122,394**	**2570**	0.213
MILC	**8192**	**7,337,756**	**1269**	0.706
PARATEC	**1024**	**1,206,376**	**540**	2.182
GEOMETRIC MEAN				0.7

Rate Per Core = Ref. Gflop count / (Tasks*Time)

Flop count measured on reference system

Measured wall clock time on hypothetical system

Geometric mean of ‘Rates per Core’

SSP (TF) = Geo mean of rates per core * # cores in system/1000

N6 SSP of 100,000 core system = 0.7 * 100,000 /1000 = 70

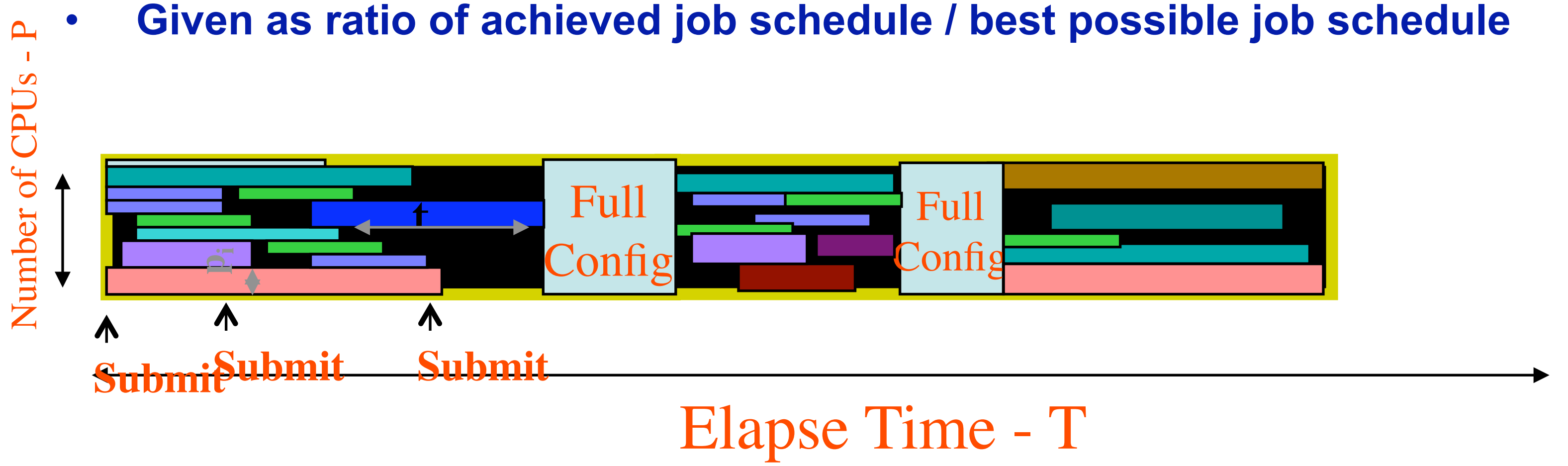
N6 SSP of 200,000 core system = 0.7 * 200,000 /1000 = 140

Allows vendors to size systems based on benchmark performance

Key Point - Effective System Performance (ESP) and Consistency

- Resource managers as important to efficient system utilization as sustained computational performance; risks are
  - Ability to respond to operational priority changes, Scheduler ability to make decisions based on limited data, Job Start Overhead, Interconnect performance with job fragmentation, etc.
- ESP measures resource manager performance in terms of
  - Efficiency in scheduling available resources
  - Job priority management
- Given as ratio of achieved job schedule / best possible job schedule

$$E_{s,k} = \frac{\sum_{i=1}^{I} (\chi_i * T_i)}{\left[N_{s,k} * \left(T - BEST_s \right) \right]}$$



- Consistency measured by Coefficient of Variation (CoV)
  - Standard deviation/mean
  - Measured in dedicated and general use time
  - Measured on applications, IO and other aspects

$$CoV = \frac{\sqrt{\frac{1}{O} \sum_{o=1}^{O} \left(t-obs_o - \overline{t-obs} \right)^2}}{\overline{t-obs}}$$

NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

I/O and NGF

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

NERSC-6 Filesystem

- Vendors shall provide a hardware and software solution for an external parallel filesystem as part of NERSC-6 system.
  - All storage hardware and fabric shall be standards-based (FC, IB, Ethernet) and be compatible with NERSC's existing NGF system.
- Vendor shall provide on-going support for the hardware and software in their proposed solution.
- The Vendor-provided filesystem will be benchmarked as part of the NERSC-6 system.
- After NERSC-6 is accepted, at NERSC's discretion, Vendor will provide engineering assistance to relocate the storage hardware from the NERSC-6 system to the NGF system.

RFP I/O Highlights

- NERSC-6 RFP has I/O bandwidth and capacity targets but allows Vendors to determine specific configuration:
  - 70 GB/s sustained aggregate bandwidth
  - 1 PB usable filesystem capacity
  - Software layer as important as hardware layer
- Performance features
  - Tight MPI-IO integration with file system
  - High sustained parallel and single stream filesystem I/O bandwidth to and from global shared storage system
  - High efficiency for both large and small block I/O for shared file access with high concurrency
- Consistency
  - Document performance variation in dedicated and production modes

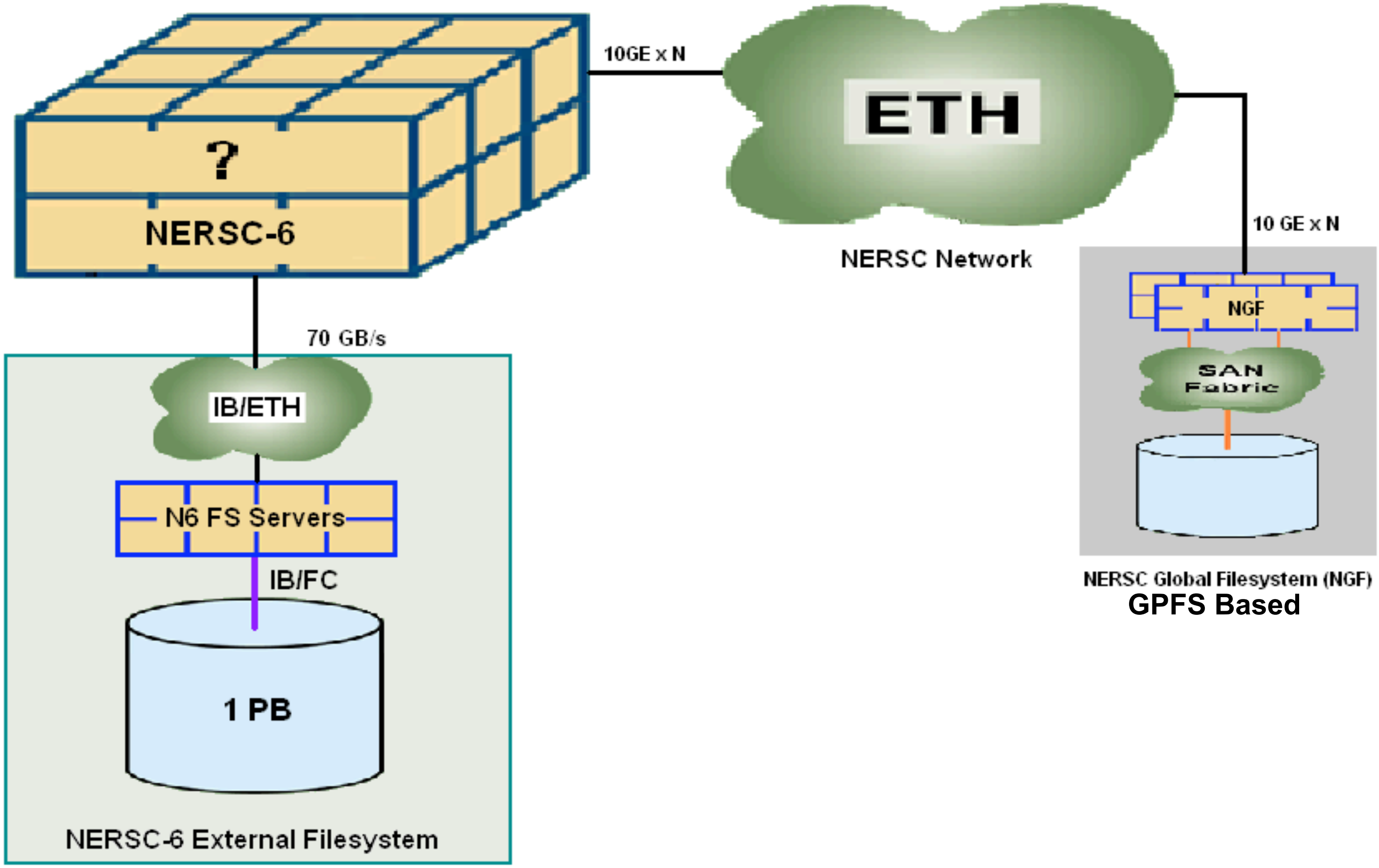
NERSC-6 I/O Benchmarks

- Both synthetic benchmarks, (IOR and Metabench) and a full application benchmark with I/O are used
- Recognize Vendors do not always have full I/O capabilities in test systems so committed projections are acceptable
- IOR measures a variety of access patterns, interfaces and concurrencies
- Metabench measures metadata performance
- MAESTRO application benchmark runs without I/O and subsequently with I/O turned on
  - Intention is to measure % I/O time in applications
  - SSP is calculated with I/O turned off

NERSC-6 I/O Tests

- IOR (developed by LLNL for I/O stress test)
  - Access Pattern - (writes and reads)
  - File Type - (binary)
  - Programming Interfaces (POSIX, MPI-IO)
  - Block size (1 MB)
  - Transfer Size (10KB, 100KB, 1MB)
  - Concurrency (1, node size, 64, 512)
- Metabench (developed by LBNL for metadata test)
  - Create, delete files
  - On 8 and 512 processors
  - Single and multiple directories

Storage Configuration for NERSC-6 Acceptance



GPFS Based

NGF Current Production Configuration

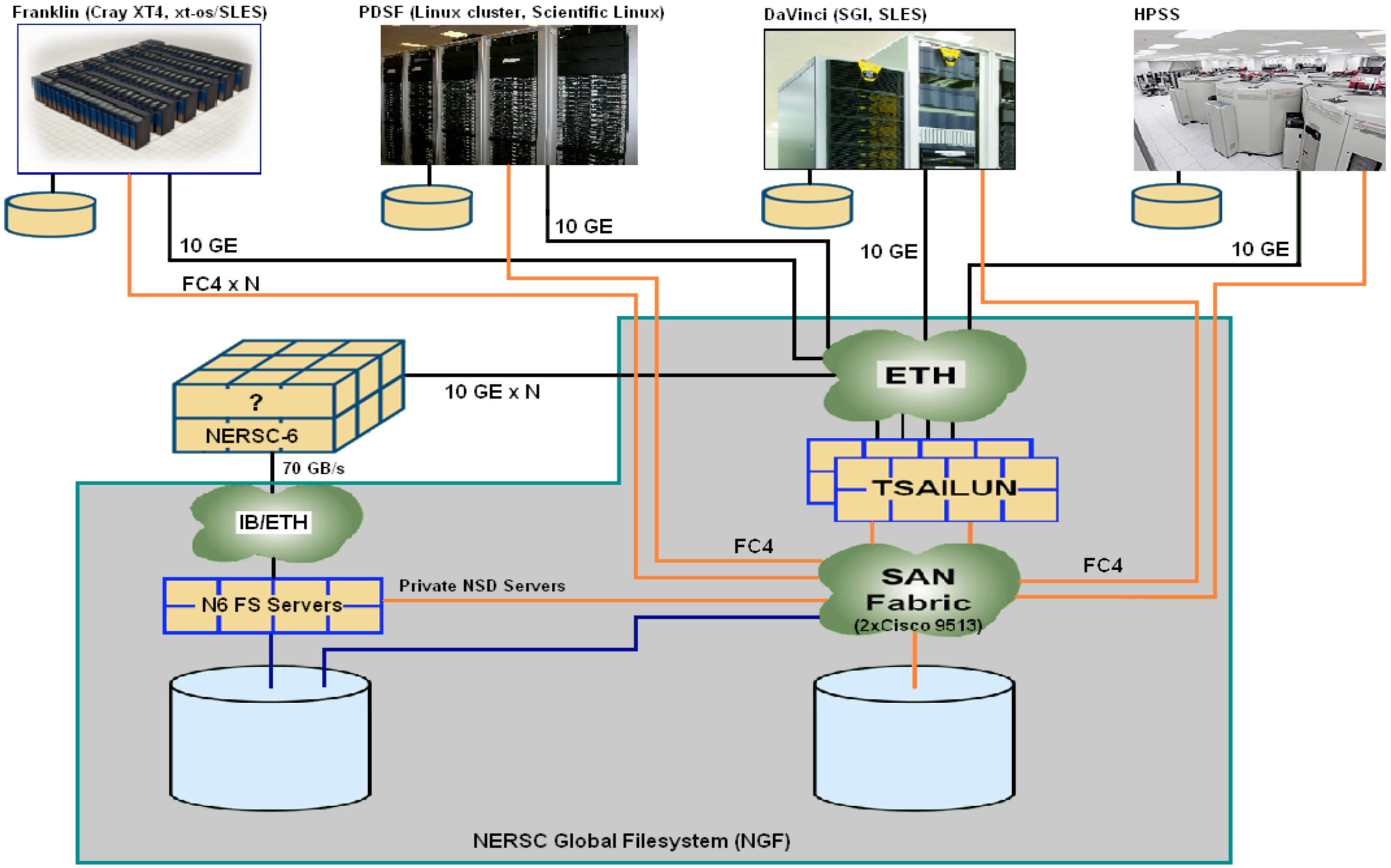
- In October 2005, NERSC deployed a production NERSC Global filesystem (NGF), /project, across multiple NERSC systems.
- The current NGF instance includes:
  - 26 I/O Server Nodes, Linux SLES9 SP3, GPFS 3.1 PTF20
  - 230 TB usable end user storage
    - DDN 9500 with SATA drives and FC drives
    - IBM DS4500 SATA drives and FC drives
    - Sun 6140 Storage Array with SATA drives
  - 50 million inodes
  - 5.5+ GB/s bandwidth for streaming I/O
  - Storage and servers external to all NERSC systems
  - Distributed over 10 Gigabit Ethernet and FC infrastructure
  - Single filesystem instance providing file and data sharing among multiple NERSC systems
    - Both large and small files
    - Persistent data, not scratch
    - Backed up to HPSS

NGF-compatible Components

- NGF is based on IBM's General Parallel File System (GPFS) and will work with any hardware or software that is listed in the GPFS FAQs (www.ibm.com/clusters/software/gpfs.html).
  - Platform: AIX, Linux (SLES, RHEL)
  - Interconnects:
    - Linux: Ethernet, 10-GE, Myrinet, InfiniBand
    - AIX: Ethernet, 10-GE, Myrinet, InfiniBand, HPS
  - Storage arrays and disks: most FC-based storage; new IB-based storage is expected to be compatible also. NGF can work with all FC-based storage and NERSC will work with IBM for support for disks that are not on the IBM supported list.
- Configurations that may not be able to migrate to NGF:
  - Filesystems that use disks that are locally attached to compute nodes.
  - Filers or appliances with proprietary hardware (e.g., NetApp, Panasas)

NERSC Global File System Vision

NGF (after NERSC-6 is accepted)



GPFS Based

Filesystem Options

- Vendors are free to choose the filesystem hardware and software solution that they believe provides the best overall system solution for NERSC.
  - There is the initial software cost and the support costs
  - There are multiple filesystem software packages where the software is free
    - For example, Open Source Filesystem Software such as Lustre, PVFS
    - NERSC's GPFS license covers the NERSC-6 system for any number of clients
      - Includes before and/or after Acceptance and factory testing
  - The Vendors are responsible
    - for assuring the filesystem proposed works well with the rest of their system
    - for providing on-going support for their proposed filesystem

Filesystem Requirements (Backup)

- **Minimal Requirements:**
  - An external parallel filesystem, administered independently from the computational nodes, with a single, unified namespace and high parallel I/O performance for all user data. All system shared storage and storage fabric shall be standards based and packaged independently. Acceptable standards are Fibre Channel, Ethernet, and InfiniBand.
  - All components comprising the external filesystem supplied by the system vendor shall be compatible with NERSC's NGF system.
- **Performance Features:**
  - High sustained parallel and single stream filesystem I/O bandwidth to and from global shared storage system.
  - A global shared storage system of at least 1 PB user usable disk space, providing at least 70 GB/s of measurable, sustained aggregate filesystem I/O bandwidth between the external parallel filesystem and the computational nodes.
  - High efficiency for both large and small block I/O for shared file access with high concurrency.
  - A plan for integrating, supporting, achieving and maintaining high performance parallel access to the NGF system.
  - Engineering assistance with the re-allocation of storage hardware from the NERSC-6 system to the NGF system. Maintenance and required licenses will continue on the storage hardware.

NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

Statement of Work Negotiations

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Statement of Work (SOW)

- Offeror's proposal is the foundation of the SOW included in the Subcontract
- Negotiations involve:
  - clarifying performance elements
  - possible deletion of features which do not add value as determined by the evaluation
  - and potential improvement of features considered to be weaknesses by the evaluation
- A detailed description of the hardware and the software features of the selected system are included in the SOW
  - Fleshing out the details of the proposed system

Statement of Work (SOW)

- Upon selection for negotiation:
  - Draft SOW will be sent to Vendor
  - Face to face meetings
    - First kick-off meeting scheduled (usually couple weeks notice after selection)
      - Begin review process of draft SOW
    - Schedule next meeting in a week or two to allow each party to include respective experts in outside review
    - Iterative process and meetings at each other's sites (or mutually agreed alternate site) until complete

Statement of Work (SOW)

Statement of Work (SOW)

Statement of Work (SOW)

Statement of Work (SOW)

Statement of Work (SOW)

Statement of Work (SOW)

Statement of Work (SOW)

Statement of Work (SOW)

Statement of Work (SOW)

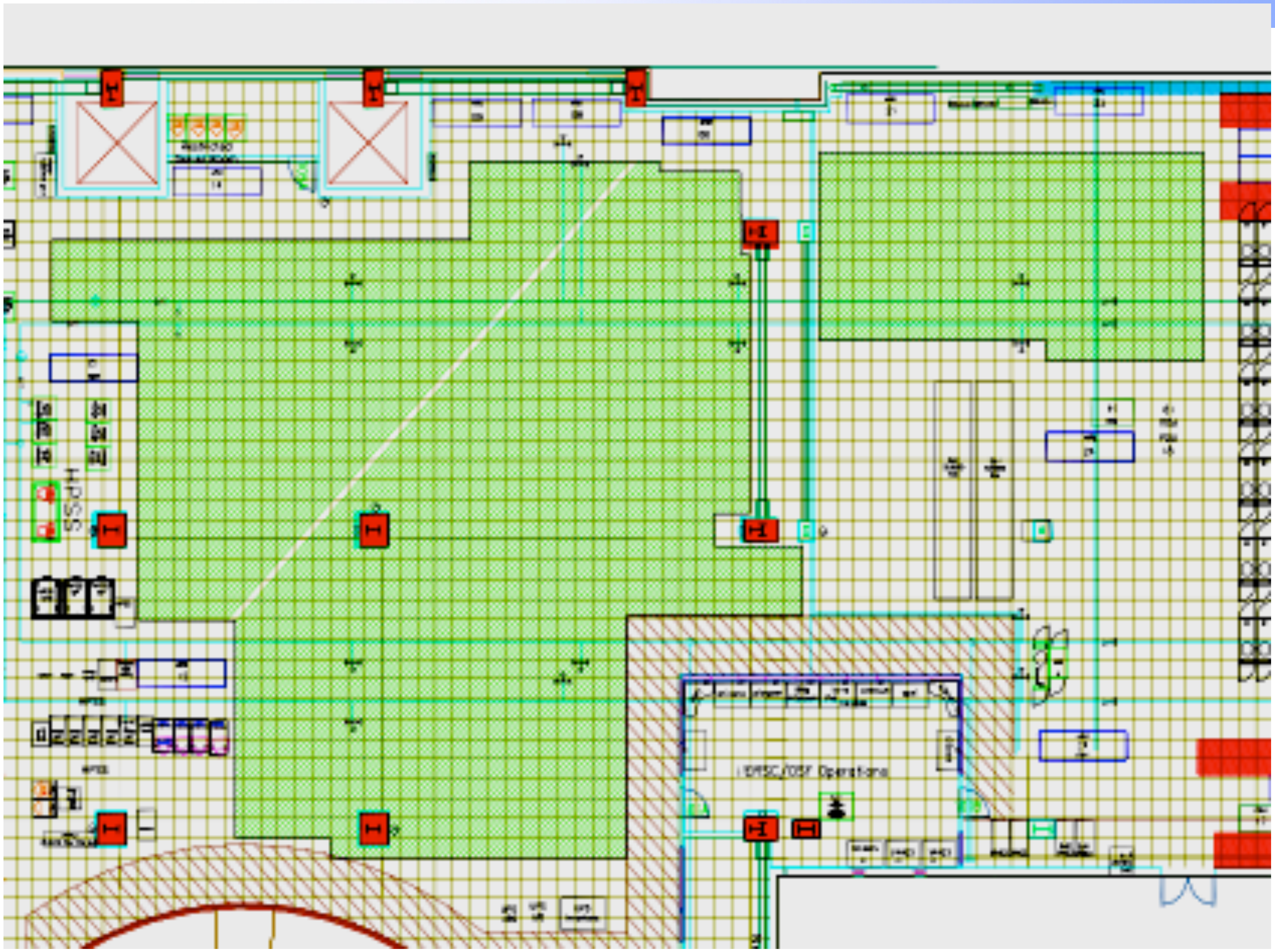
Statement of Work (SOW)

Approval and Award

- Upon completion of an agreeable SOW
  - Contract Award Package completed and ready for signatures
  - LBNL must conduct a Contract Review Board (CRB) award package (Procurement) before Lab can sign contract
  - DOE Review of Award Package
  - Contract sent to Vendor for their signature

Facility Integration

NERSC-6 Area (5000 sq. ft.)



Power and Cooling

	Power (including cooling)	Cooling
OSF	6MW	1450 tons
N6	3.3MW	700 tons

- Did major upgrade to facility power and cooling in 2005
- OSF can support
  - Air cooling using chilled water air handling units
  - Liquid cooling using direct chilled water connections

Power Required for Cooling

Cooling (Tons)	Cooling (KW)	Remaining Power (MW)
700	219	3.08
600	188	3.11
500	156	3.14
400	125	3.18
300	94	3.21
200	63	3.24
100	31	3.27

The OSF chillers consume 0.3125 KW/Ton of cooling. If a liquid cooled system requires 700 tons of cooling then 0.3125 * 700 = 219KW of power which leaves about 3.1MW for computer power. Air cooled systems will require additional power for the computer room CRAH units.

Seismic Protection

- Work Safe Technologies: ISO-Base

System Delivery, Installation, Integration and Testing

- Test system delivered and installed
- Formal factory test to assess readiness of system for shipment
  - Includes both functional and performance testing
- System is installed, seismically protected, interconnect cabled
- Vendor stabilizes system and initial performance testing takes place
- System is configured for production and integrated with NERSC software infrastructure
  - System Security - some special points
    - System is isolated from the outside and other systems
    - System is examined and hardened
    - System is scanned for vulnerabilities before general access
- Performance tuning to demonstrate committed performance levels
- Three phase Acceptance Test
  - Functionality and System Tests
  - Performance Tests
  - Reliability and Availability Test

Software Integration

- Center Infrastructure
  - Grid
    - OSG software stack
  - NIM
    - Centralized account and allocation management
  - HPSS Archive
    - Hsi, htar, pftp
  - LDAP
    - OpenLDAP infrastructure for authentication
  - Nagios
    - Centralized system, fabric and storage monitoring

- User Software environment
  - Batch scheduler configuration
  - User development environment
    - Compilers
    - Debuggers
    - Profiling and performance analysis
  - Libraries
  - Third-party applications

NERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

Questions?

Tours?

Office of Science
U.S. DEPARTMENT OF ENERGY

BERKELEY LAB