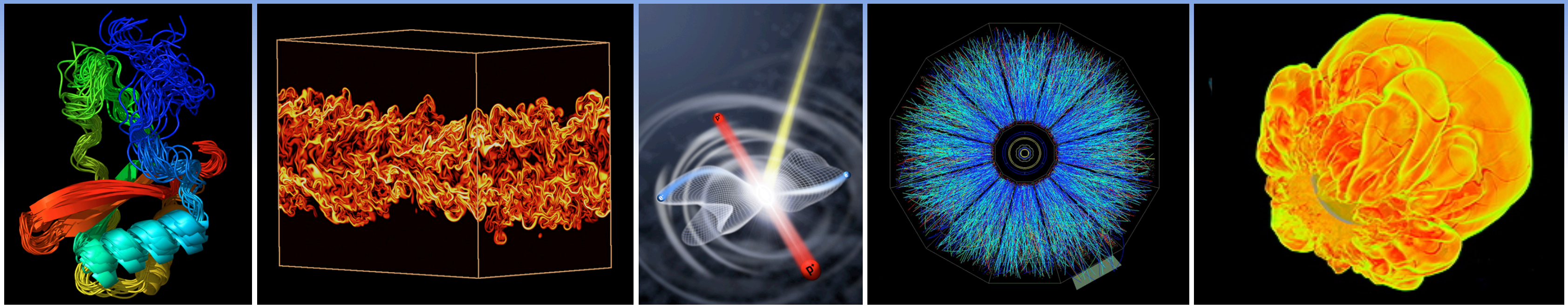


Hopper, the New NERSC-6 System

Jonathan Carter

October 7th, 2009

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC

BERKELEY LAB

Timeline

- **Initial Project approval Mar 08**
- **Lehman Review Jul 08**
- **RFP released Sep 08**
- **Responses received Oct 08**
- **Evaluation conducted Nov 08**
  - Jonathan Carter replaced Bill Kramer as Project Lead
- **Negotiations conducted Dec 08 - Mar 09**
- **Final Project approval Apr 09**
- **Contract was signed Jul 09**
- **Factory Test of Phase 1 System Sep 09**

Cray Proposal is the Best Value

- Best application performance per dollar
- Highest sustained application performance commitment
- Best sustained application performance per MW
- Excellent in-house testing facility and benchmarking/performance/support expertise at Cray
- Easy to integrate into our facility
- Acceptable risk

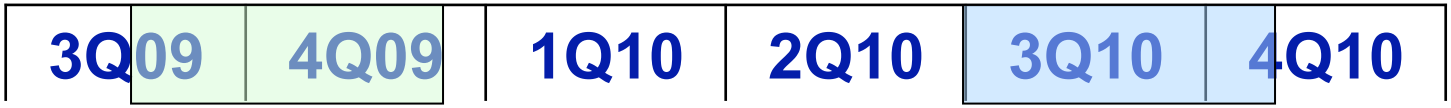
Hopper System

Phase 1 - XT5

- 668 nodes, 5,344 cores
- 2.4 GHz AMD Opteron (Shanghai, 4-core)
- 50 Tflop/s peak
- 5 Tflop/s SSP
- 11 TB DDR2 memory total
- Seastar2+ Interconnect
- 2 PB disk, 25 GB/s
- Air cooled

Phase 2

- >6000 nodes, >150,000 cores
- AMD Opteron (Magny-Cours, 12-core)
- >1.0 Pflop/s peak
- >100 Tflop/s SSP
- >200 TB DDR3 memory total
- Gemini Interconnect
- 2 PB disk, 80 GB/s
- Liquid cooled



3Q09 4Q09 1Q10 2Q10 3Q10 4Q10

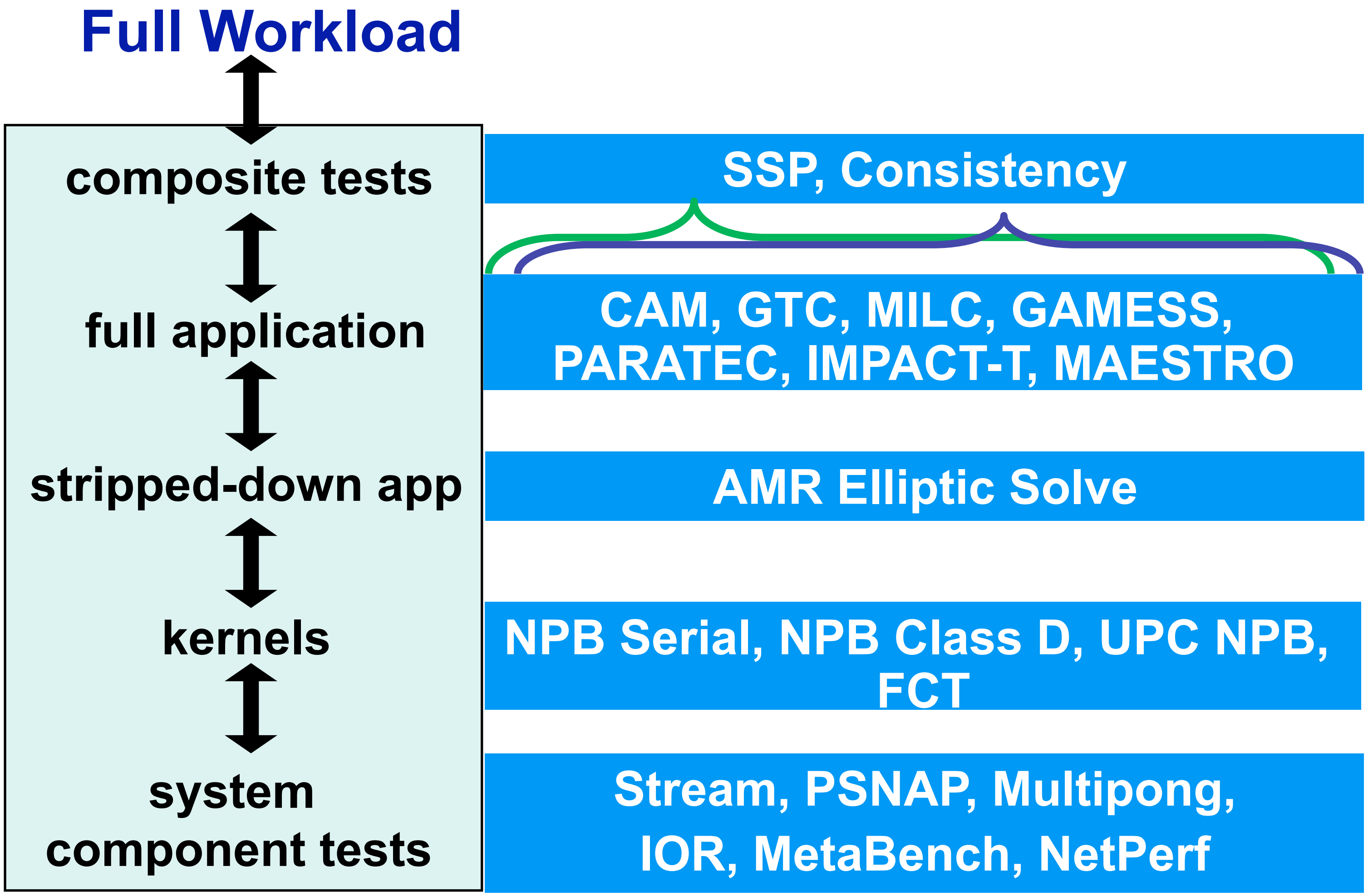
Project Goals

- Deploy a complete, integrated computing environment for a multi-user, multi-application, parallel scientific workload
- Support entire DOE Office of Science Workload
- Greatly increase computational resources available to users using measured performance criteria
- Integrate into the NERSC environment

RFP

- 13 ‘Minimum Requirements’ (e.g., 24x7 support) that absolutely must be met
  - Proposals that don’t meet are not responsive and are not evaluated further
- 38 ‘Performance Features’ (e.g., fully featured development environment) wish list of features
  - Evaluated qualitatively via in-depth study of Offeror narrative.
- Benchmarks
  - Kernel tests and full applications
  - Sustained application performance (measured by SSP benchmarks)
- Supplier attributes (ability to produce/test, corporate risk, commitment to HPC, etc.)
- Cost of ownership (incl. life-cycle, facilities, base, and ongoing costs) and affordability

NERSC-6 Benchmarks



Full Workload

composite tests — SSP, Consistency

full application — CAM, GTC, MILC, GAMESS, PARATEC, IMPACT-T, MAESTRO

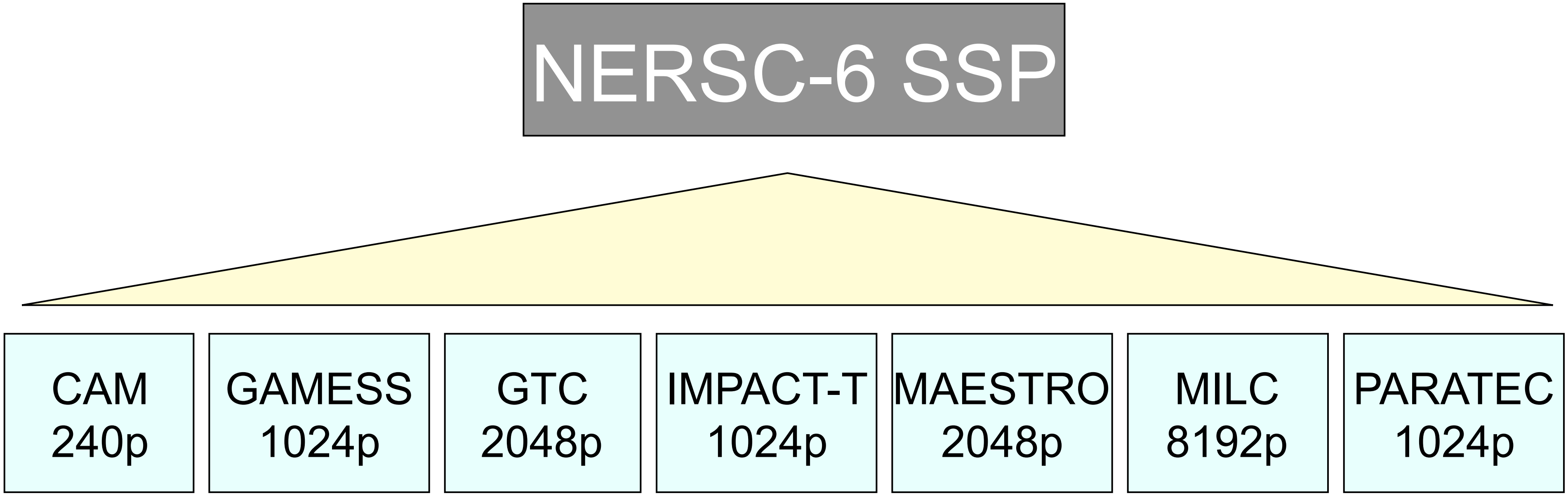
stripped-down app — AMR Elliptic Solve

kernels — NPB Serial, NPB Class D, UPC NPB, FCT

system component tests — Stream, PSNAP, Multipong, IOR, MetaBench, NetPerf

NERSC-6 SSP Metric

The largest concurrency time of each full application benchmark is used to calculate the SSP



NERSC-6 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*

Technology Observations

- Multi-core continued its progression:
  - Most proposals had more than 2X number of cores as current largest NERSC system
  - All Offers had two sockets per node – interconnect becoming more sparse and NUMA becoming more important
  - Clock speeds remained the same or showed modest increase
- Several commodity-based systems (Nehalem / IB + Linux) packaged for HPC
- Systems with open-source software stacks were offered
- No accelerator- or GPU-based systems proposed
- Several different Infiniband topologies were offered
- Vendors responded to request to comply with stricter thermal (ASHRAE recommended) standards with innovative solutions

Feedback from NERSC Users was crucial to NERSC6 negotiations

User Feedback from Franklin	*NERSC6 Enhancement*
Login nodes need more memory	8 external login nodes with 128 GB of memory (with swap space)
Shared libraries are not supported	Shared libraries are supported.
Need more disk space	Includes a 7x increase in disk space over Franklin (2PB)
Increase I/O bandwidth	Includes a 3x increase in I/O bandwidth over Franklin (70 GB/sec)
Connect NERSC Global FileSystem to compute nodes	/project file system will be available to compute nodes
Workflow models are limited by memory on MOM (host) nodes	•Increased # and amount of memory on MOM nodes •Phase II compute nodes can be repartitioned as MOM nodes

Feedback from NERSC users was crucial to NERSC6 negotiations

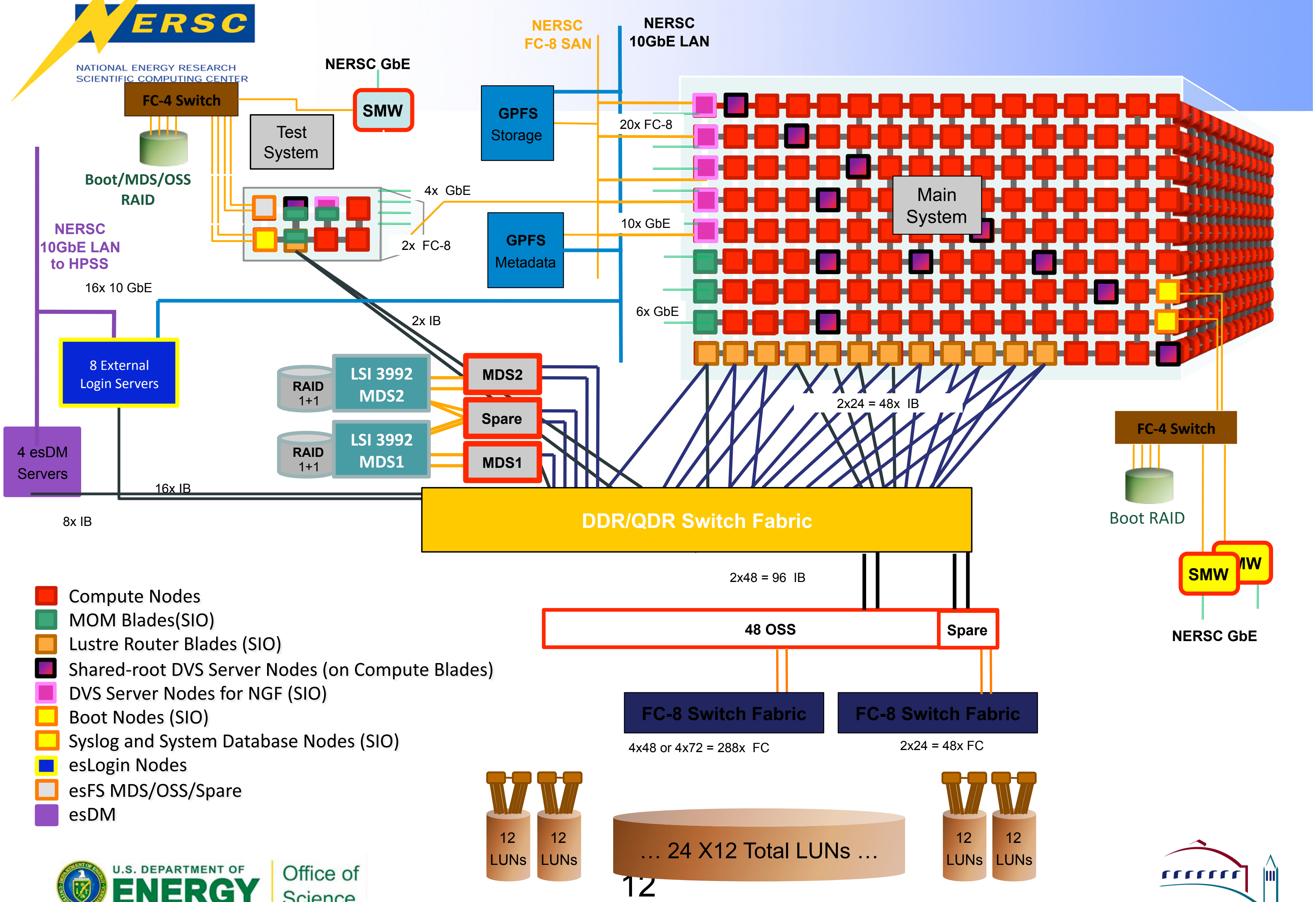
User Feedback from Franklin

Improve Stability and Reliability

NERSC6 Enhancement

- External login nodes will allow users to login, compile and submit jobs even when computational portion of the machine is down
- External file system will allow users to access files if the compute system is unavailable and will also give administrators more flexibility during system maintenances
- Gemini interconnect has redundancy and adaptive routing. (System will survive a down link.)

(All will still require some shakeout!)



NERSC GbE

NERSC FC-8 SAN

NERSC 10GbE LAN

FC-4 Switch

SMW

Test System

GPFS Storage

20x FC-8

Boot/MDS/OSS RAID

4x GbE

2x FC-8

GPFS Metadata

10x GbE

Main System

NERSC 10GbE LAN to HPSS

16x 10 GbE

6x GbE

2x IB

8 External Login Servers

RAID 1+1

LSI 3992 MDS2

MDS2

Spare

RAID 1+1

LSI 3992 MDS1

MDS1

2x24 = 48x IB

FC-4 Switch

Boot RAID

4 esDM Servers

16x IB

8x IB

DDR/QDR Switch Fabric

SMW

SMW

NERSC GbE

2x48 = 96 IB

48 OSS

Spare

FC-8 Switch Fabric

FC-8 Switch Fabric

4x48 or 4x72 = 288x FC

2x24 = 48x FC

12 LUNs

12 LUNs

… 24 X12 Total LUNs …

12 LUNs

12 LUNs

- 🟥 Compute Nodes
- 🟩 MOM Blades(SIO)
- 🟧 Lustre Router Blades (SIO)
- ■ Shared-root DVS Server Nodes (on Compute Blades)
- 🟪 DVS Server Nodes for NGF (SIO)
- 🟨 Boot Nodes (SIO)
- 🟨 Syslog and System Database Nodes (SIO)
- 🟦 esLogin Nodes
- ⬜ esFS MDS/OSS/Spare
- 🟪 esDM

Software and Compilers

- Software will be very similar to Franklin but with shared library support
- Four different compilers
  - Portland Group
  - PathScale
  - Cray Compilers
  - GNU
- Some codes see significant performance improvements with a specific compiler
- NERSC will provide guidance and support to help users choose

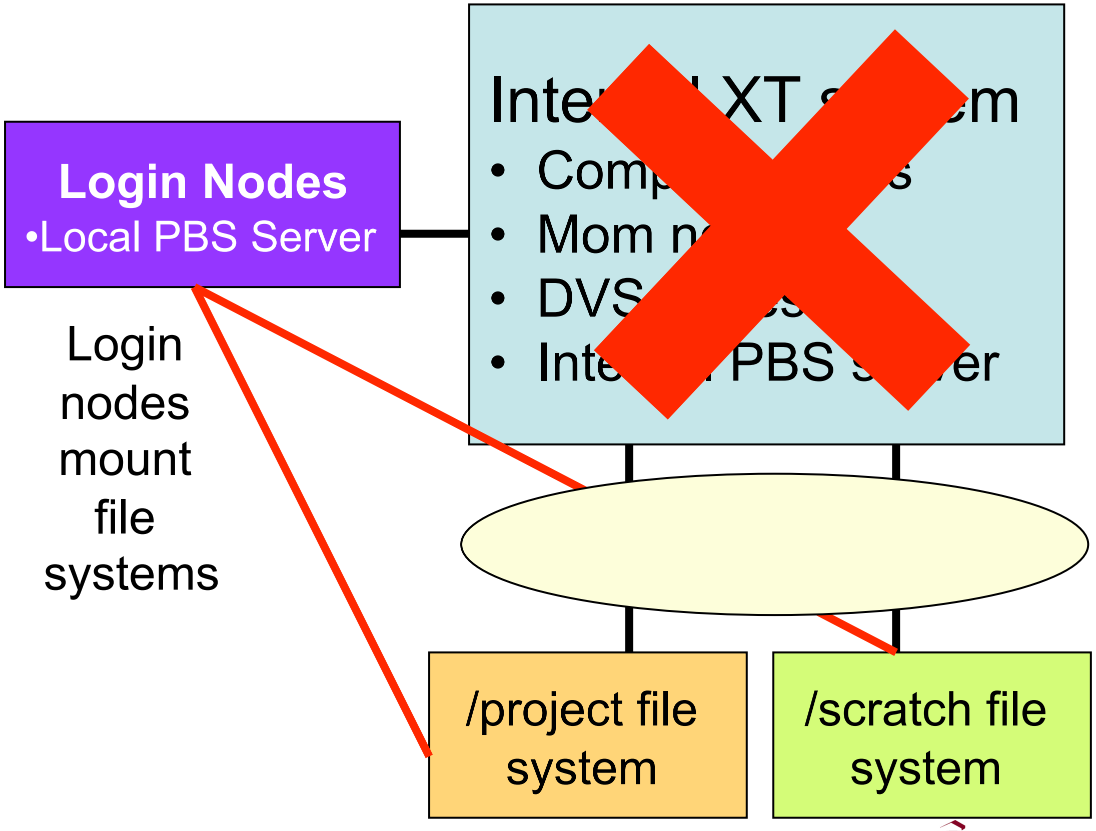
Hopper Login Nodes

- 8 login nodes external to main XT system
- Quad socket, quad-core AMD Opteron 2.4GHz
- 128 GB of memory with swap space
- Load balanced for more optimal usage
- Ability to run more intensive tools on login nodes, IDL, debuggers, etc.
- Available when XT is down

Access to data and login nodes even when XT is unavailable

- **Submit jobs when XT down**
- **Local PBS server on login nodes**
- **Holds jobs while XT is down**
- **Jobs forwarded to internal XT PBS server when XT available again**

Sketch of Hopper



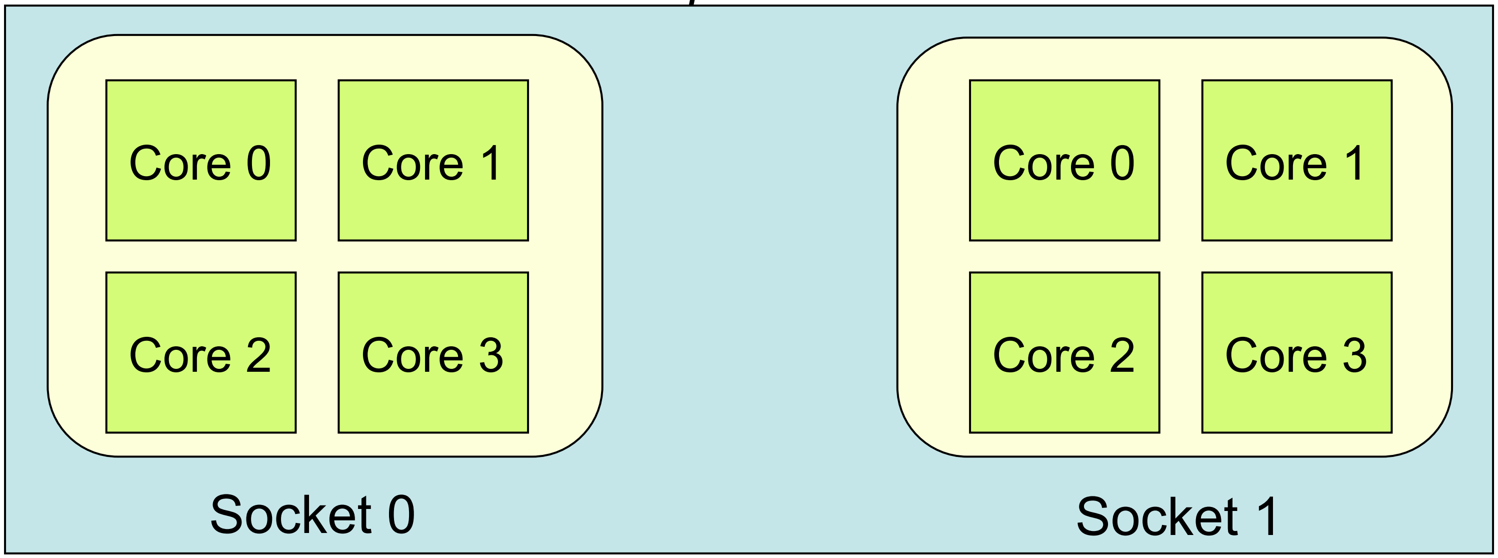
Dynamic and Shared Libraries

- All user software has a shared library version (mpich, acml, libsci, etc.)
- Static binaries is default environment
- Use the -dynamic compiler and linker flag
- In batch script set environment variable CRAY_ROOTFS=DSL which enables shared root file system

aprun Options

- Hopper has 2 sockets per core, increasing the aprun options, particularly for openMP codes

Compute Node



Core 0 Core 1

Core 2 Core 3

Socket 0

Core 0 Core 1

Core 2 Core 3

Socket 1

- New options to specify, how many sockets, which socket, cores per socket, strict memory containment between sockets
- NERSC will provide guidance on the options

Hopper Phase I Target Users

- Application performance will be similar to Franklin
- All users welcome to run on Hopper, but target users who need additional functionality
  - I/O intensive applications
  - Shared and dynamic libraries support
  - Heavy use of login nodes
  - Heavy use of MOM (host) nodes

Proposed Hopper Queues

Submit Queue	Execution Queue	Nodes	Cores	Time Limit	Relative Priority	Charge Factor	User Run Limit
interactive	interactive	1-16	1-128	30 mins	1	1	1
debug	debug	1-64	1-512	30 mins	2	1	1
regular	reg_short	1-16	1-128	4 hrs	3	1	5
	reg_small	1-16	1-128	48 hrs	3	1	3
	reg_med	17-64	129-512	36 hrs	3	1	3
	reg_big	65-256	513-2,048	24 hrs	3	1	3
	reg_long	1-4	1-32	72 hrs	3	1	1
low	low	1-64	1-512	12 hrs	4	0.5	5

Limits

- 5 running jobs/user (system-wide limit)
- 4 queued (eligible for scheduling) jobs/user
- reg_long: 1 running job/user, 1 queued job/user, 4 running jobs max