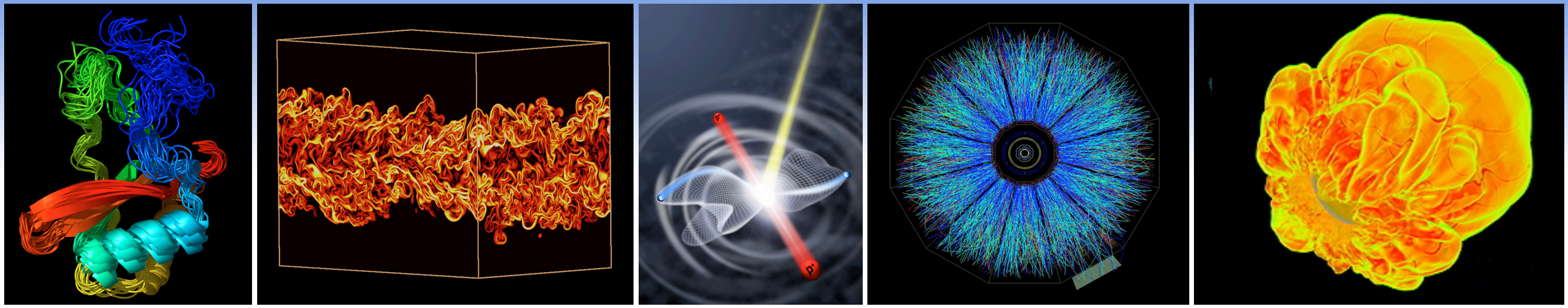


SU2009 Procurement

Brent Draney
NERSC Systems Department
October 7, 2009

SU2009 Procurement

- Scaleable Unit (SU) cluster procurement
  - Reasonable sized building blocks to assemble clusters
  - Best Value Source Selection (BVSS) procurement
  - Expected to scale up to 100TF
- NERSC program replacement for Bassi and Jacquard
  - Budget comparable to extending maintenance on Bassi and Jacquard for 3 years
  - Targeted 15 to 20 TF
- ARRA capable contract vehicle

Procurement Team

• Shane Canon,	Data Systems GL
• Tina Declerck,	Computational Systems
• Brent Draney,	Networking, Security, Servers GL
• David Paul,	Computational Systems
• Lynn Rippe,	Procurement
• David Turner,	User Services

Carver Cluster

- NERSC NCS-C production cluster
  - Bassi/Jacquard replacement
- Named after George Washington Carver
  - Botanist and inventor
- Vital Statistics
  - 34.2 Teraflop, peak
  - 5 scalable units
  - 400 compute nodes (3200 cores) with Nehalem quad-core
  - 9.6 TB DDR3 memory (3 gigabytes/core)
  - QDR InfiniBand fabric
  - Center-wide NGF (GPFS) file system for all storage needs

Dalton Cluster

- Magellan Cloud Test Bed
- Named after John Dalton
  - Father of atomic theory
  - Meteorologist who investigated the physics of clouds
- Vital Statistics
  - 61.5 Teraflop, peak
  - 9 scalable units
  - 720 compute nodes (5,760 cores) with Nehalem quad-core
    - Including 160 extended capability nodes
      - 6GB/core memory & 1 TB local disk
  - 21.1 TB DDR3 memory
  - QDR InfiniBand fabric
  - Center-wide NGF (GPFS) file system for most storage needs
  - Flash storage for data-intensive applications

Dalton Availability

- Dalton is a research vehicle
  - For experiments contributing to cloud research
  - For special projects initiated by joint agreement between ASCR and NERSC

- Dalton cycles will be available to NERSC users through a special queue on Carver
  - Performance data will be collected during runs to help characterize mid-range workloads

- Details will be announced at a later time

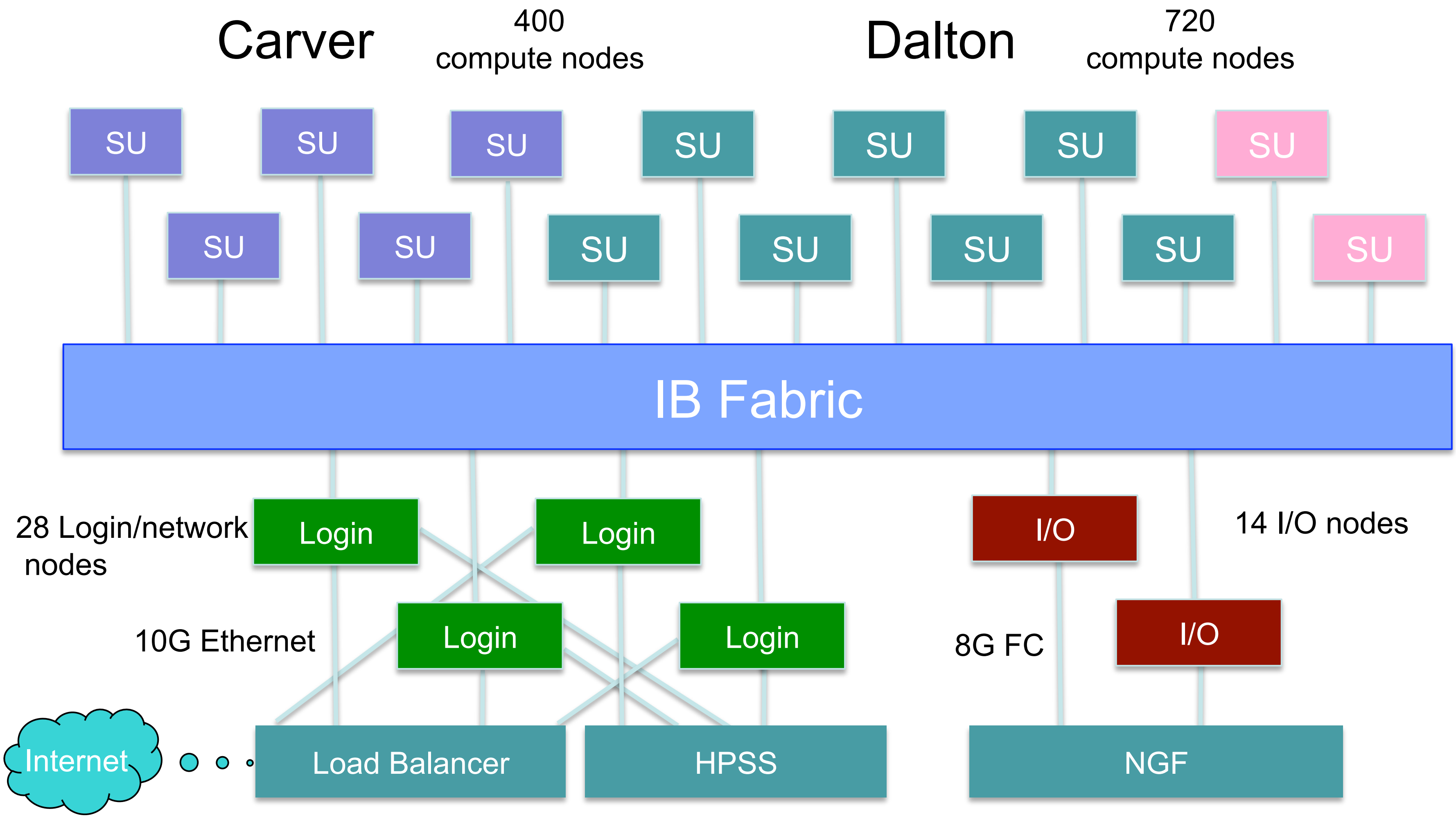
Key Benefits

- Carver cost is comparable to extending maintenance for Bassi and Jacquard for 3 years
- Carver is 3.5X more powerful than Bassi and Jacquard
  - Bassi + Jacquard = 9.9 TF
  - Carver = 34.2 TF
- Carver + Dalton power usage is less than Bassi and Jacquard
  - Bassi + Jacquard = 700KW
  - Carver + Dalton = 500KW
- Carver+Dalton are 1/3rd smaller than Bassi and Jacquard
  - Liquid cooled
    - Reduces cooling costs by as much as 1/2
    - Reduces floor space requirements by 30%

Availablility

- Installation: Nov. 2009
- Early users: Dec. 2009
- Production use: Jan 12, 2009
- Bassi / Jacquard retirement: Jan. 2009

Cluster architecture

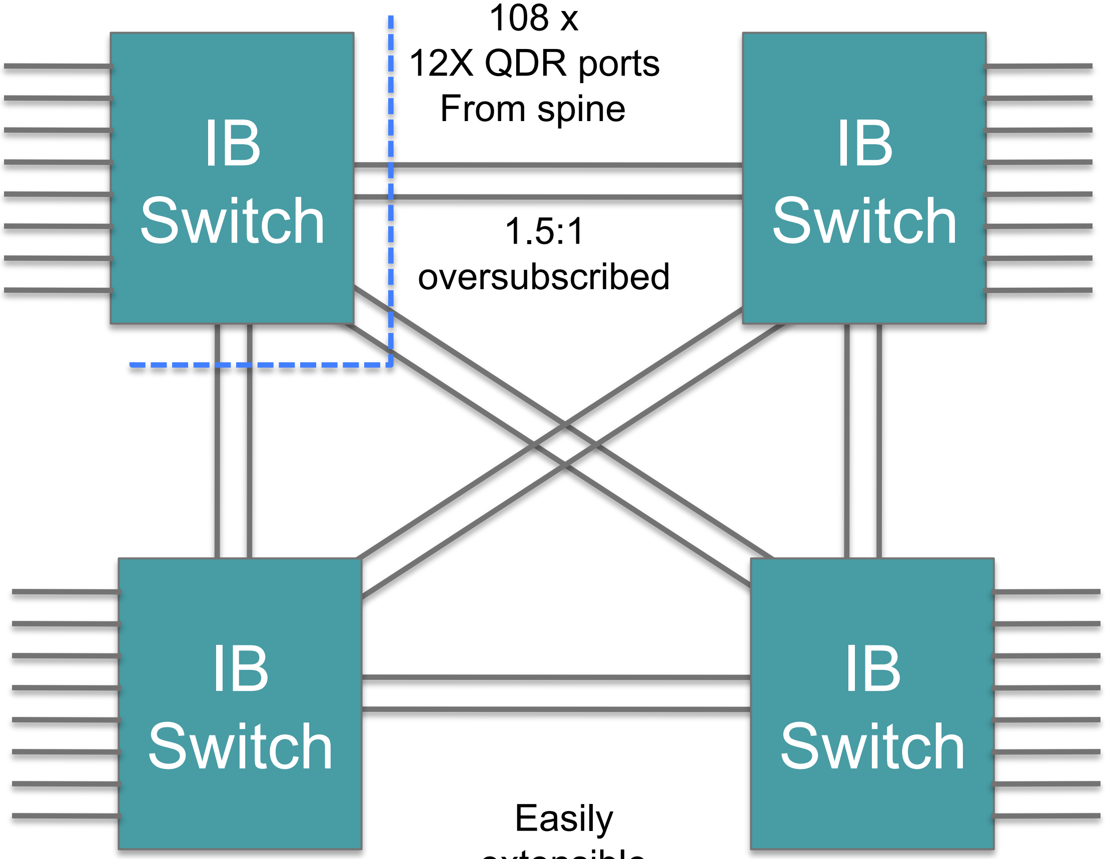


Scalable units

- Scalable Units
  - 80 compute nodes per rack (6.835 TFlop)
  - 2 login nodes
  - 1 I/O node

- IBM iDataplex chassis
  - High density
  - Front-access cabling
  - Liquid-cooled using rear-door heat exchangers

InfiniBand Fabric



324 x
4X QDR ports

3-hop FBB
within switch

4 switches
1296 ports

4-hop mesh

Easily
extensible

Node Types

	Standard Compute Node	Extended Compute Node	Login/Network Node	I/O node
Processor	Dual Intel Xeon 5550 (Nehalem) 2.67GHz Quad-Core Processors	Same	Same	Same
Memory	24GB Memory (DDR3 Chipkill ECC RDIMMs @ 1333MHz)	48GB @ 1066Mhz	48GB	24GB
PCI	1 x PCIe Gen2 x16 slot	Same	4 x PCIe Gen2 x8 slots	Same
Cluster interconnect	1 x Mellanox 4x QDR InfiniBand PCIe 2.0 x8	Same	Same	Same
Management network	1Gb Ethernet	Same	Same	Same
Internal disk	None	1 x 1TB SAS Disk	2 x 146GB SAS Disk	2 x 300GB SAS Disk
External interfaces	None	None	2 x Chelsio 10GigE NIC with SFP+	2 x Single-Port 8Gb Qlogic Fibre Channel HBAs with SR optics

Software Environment

- Full Linux OS on every node
  - Scientific Linux, a RedHat repackage
    - Full support for scripting languages (python, perl, R, etc.)
    - Full support for shared objects
  - Torque/Moab
  - OpenFabrics
  - OpenMPI
- Programming Environment
  - PGI compiler suite, gcc/gfortran (will evaluate PathScale later)
  - Libraries: OpenMPI, NAG, PETSc, FFTW, NCAR, NetCDF, HDF, (Parallel) NetCDF, HDF5, LAPACK, ScaLAPACK, SuperLU, etc.
  - TotalView debugger
- Applications - full suite will be available
  - Not on Franklin: Gaussian, full version of NAMD, Matlab, mathematica, q-chem, wien2k
  - Popular apps: Amber, Gamess, IDL, Molpro, MySQL, VASP, etc.
  - Utilities: bbcp, hsi, htar, pftp, getnim, OSG grid stack, nx, etc.

Flash Storage

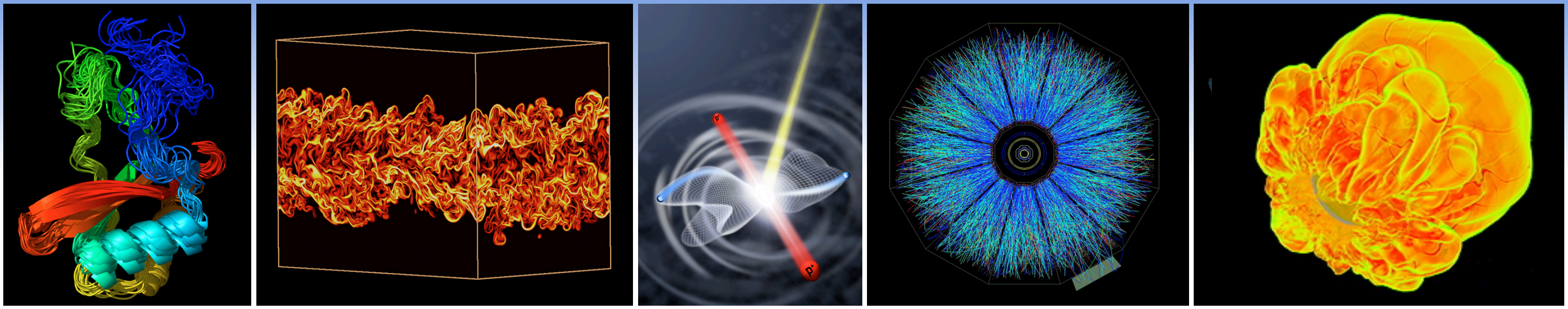
- ~10TB will be deployed in NGF
  - High bandwidth, low-latency storage class
  - Metadata acceleration
- ~16TB will be deployed as local SSD in one SU
  - Data analytics
  - Local read-only data
  - Local temp storage
- ~2TB will be deployed in HPSS
  - Metadata acceleration

Proposed Carver Queues

Submit Queue	Execution Queue	Nodes	Cores	Time Limit	Relative Priority	Charge Factor	User Run Limit
interactive	interactive	1-8	1-64	30 mins	1	1	1
debug	debug	1-32	1-256	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	17-32	129-256	36 hrs	3	1	3
	reg_big	33-64	257-512	24 hrs	3	1	3
	reg_long	1-4	1-32	168 hrs	3	1	1
low	low	1-32	1-256	12 hrs	4	0.5	5
dalton	variable	1-128	1-1024	24 hrs	TBD	TBD	TBD

Notes

- 5 running jobs/user (system-wide limit)
- 4 “queued” (eligible for scheduling) jobs/user (unlimited submits)
- reg_long: 1 running job/user, 1 queued job/user, 4 running jobs max
- Dalton queue will run on the cloud, as available. Performance data collection may be turned on.



Thank you!

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC

BERKELEY LAB

NUG Town Hall Questions

- How do you see using Carver/Dalton?