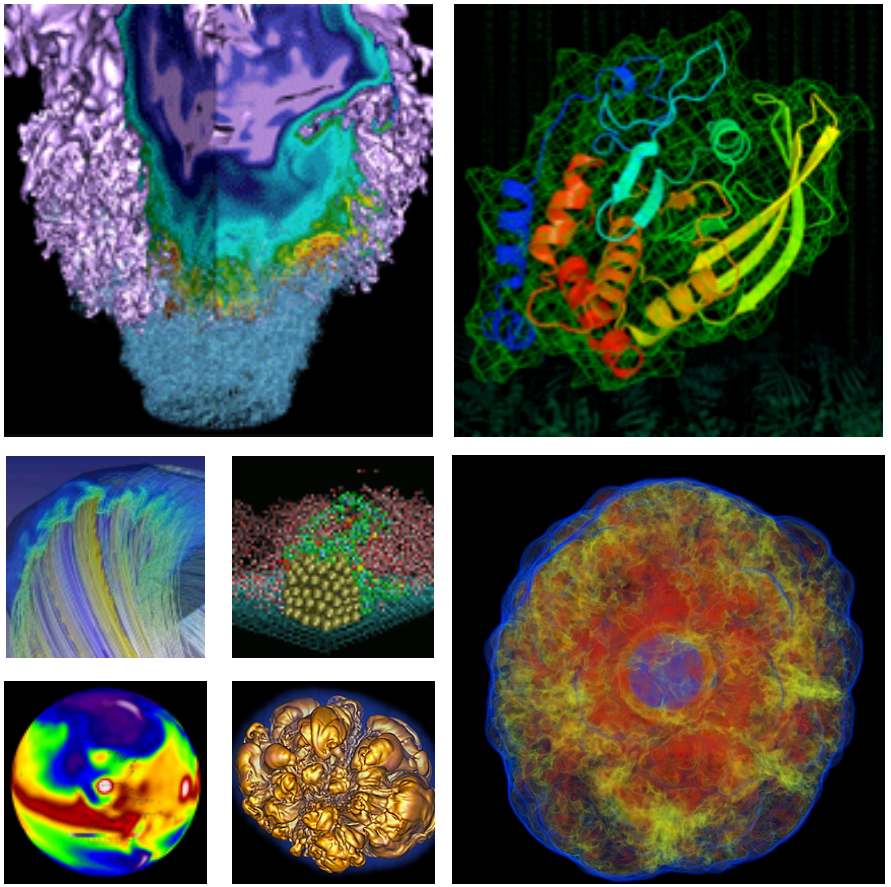
Cori: A Cray XC Pre-Exascale System for NERSC



Cray Signs $70 Million Supercomputer Contract with the National Energy Research Scientific Computing Center (NERSC)

NERSC 40 YEARS at the FOREFRONT 1974-2014

Katie Antypas, Nicholas Wright, Nicholas Cardo, Allison Andrews, Matthew Cordery

U.S. DEPARTMENT OF ENERGY | Office of Science

Cray User Group 2014
May 8, 2014

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC-8 Project Goals

- **NERSC directly supports DOE's science mission; we focus on the scientific impact of our users**
- **Need to provide a significant increase in computational capabilities for DOE SC computational research; at least 10x increase in sustained performance over NERSC-6 (Hopper)**
- **Begin transitioning user code base to energy- efficient manycore architectures and programming environments**
  - Only way to continue to provide compute speed improvements that meet user need; attempt to do this only once
- **Integrate the system into the NERSC environment, enabling user productivity**

NERSC-8 system named after Gerty Cori (1896 – 1957): Biochemist

- First American woman to win a Nobel Prize in science (1947)
- Born in Prague; US naturalized 1928
- Shared the Nobel Prize in Medicine or Physiology with her husband + 1 other
- Recognized for work involving enzyme chemistry in carbohydrates: how cells produce and store energy.
- Breakdown of carbohydrates and mechanism of enzyme action are of fundamental importance in renewable bioenergy
  (cf. DOE Complex Carbohydrate Research Center)

Cori Configuration

- **50 Cabinets of Cray XC System**
  - Approximately 9300 'Knights Landing' compute nodes
    - Self-hosted, (not an accelerator) MPI + OpenMP programming model
    - Greater than 60 cores per node with multiple HW threads each
    - 64-128 GB memory per node
    - High bandwidth on-package memory
  - 14 external login nodes
  - Aries Interconnect
  - 10x Hopper SSP
- **Lustre File system**
  - 28 PB Disk
  - 432 GB/sec
- **Option for a Burst Buffer**
- **5 FTE years of Cray Center Of Excellence staff**
- **Intel training and support**
- **Delivery in mid-2016**

Edison, a Cray XC-30 plays a key role in NERSC's strategy

- NERSC assessed that our broad workload was not ready for GPUs and procured Edison, with Ivy Bridge Intel CPUs
- Workloads that have difficulty moving to NERSC-8 can still work productively on Edison while the code is adapted
- In 2016 Edison will likely provide ~20% of NERSC's cycles

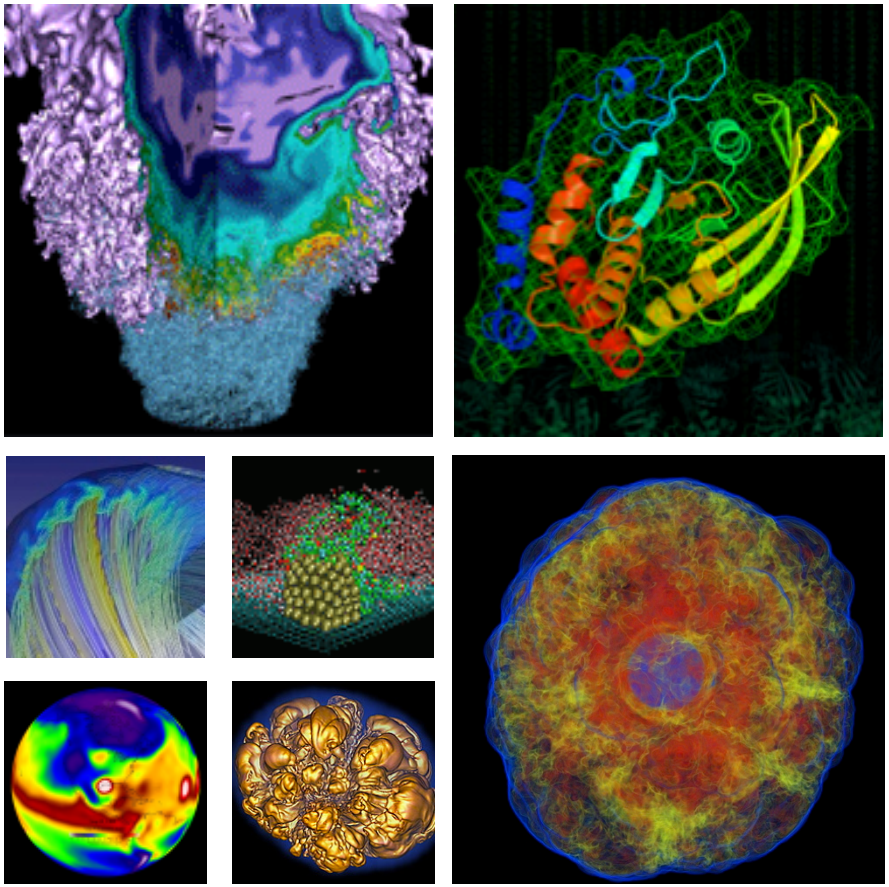
NERSC's Key Challenges

- **Application Readiness**
  - We must prepare the broad user community for manycore architectures, not just a few codes
  - Will require deep collaboration with select code teams
  - Finding the additional parallelism in some applications may be difficult.
  - Unclear how to use on-package memory, as explicit memory or cache
- **Burst Buffer**
  - How to integrate and monitor in a production environment?
  - Which applications are best suited to use the Burst Buffer?
  - How to make the Burst Buffer user friendly
- **Integration into NERSC environment in CRT**
  - Mounting NERSC-8 file system across other systems, (Edison)
  - Integration into a new facility

The Computational Research and Theory (CRT) building will be the home for Edison and Cori

- **Four story, 140,000 GSF**
  - 300 offices on two floors
  - 20K -> 29Ksf HPC floor
  - 12.5MW -> 42 MW to building
- **Located for collaboration**
  - CRD and ESnet
  - UC Berkeley
- **Exceptional energy efficiency**
  - Natural air and water cooling
  - Heat recovery
  - PUE < 1.1
  - LEED gold design

Cori architecture

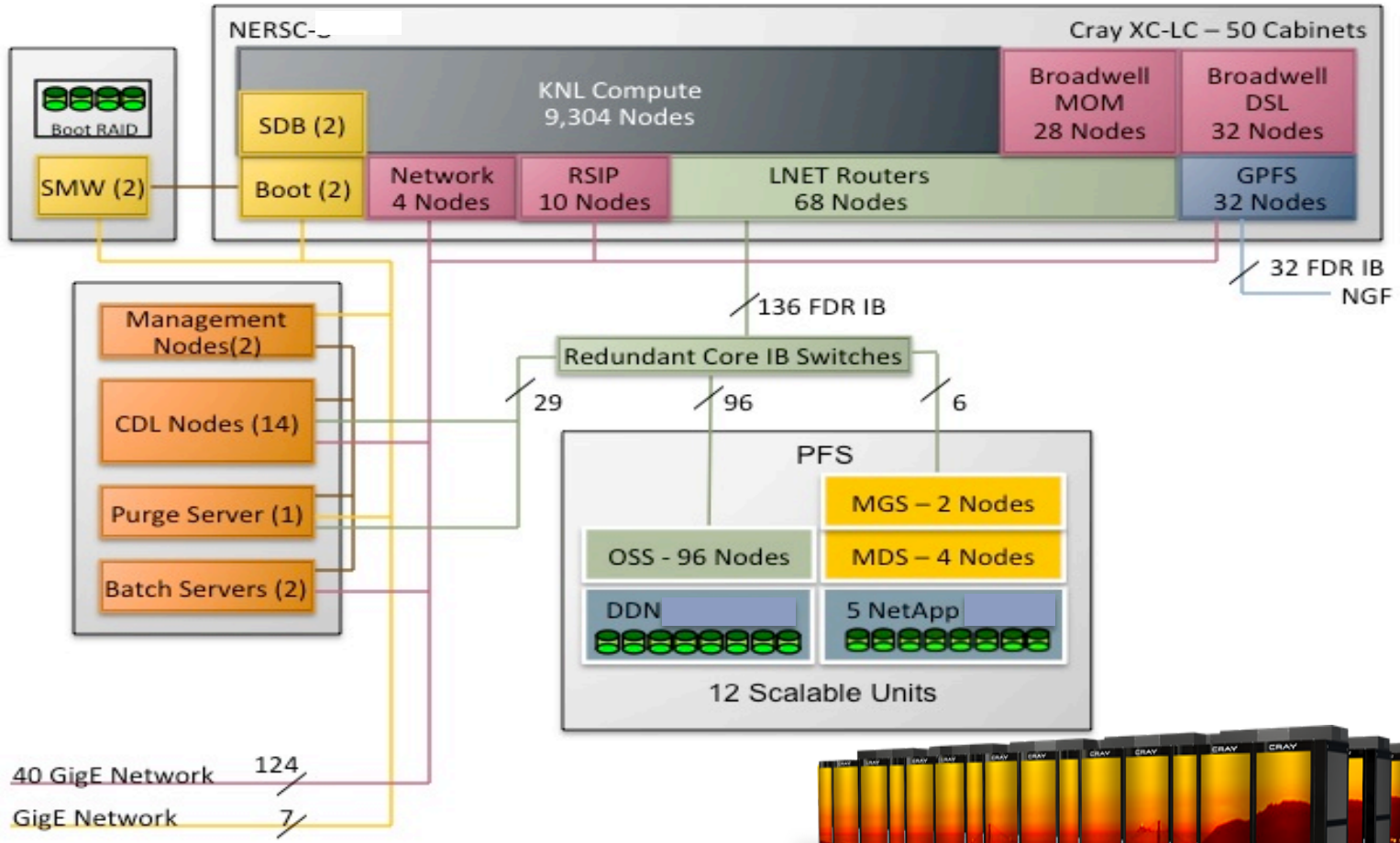


NERSC 40 YEARS at the FOREFRONT 1974-2014

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

The Cori System



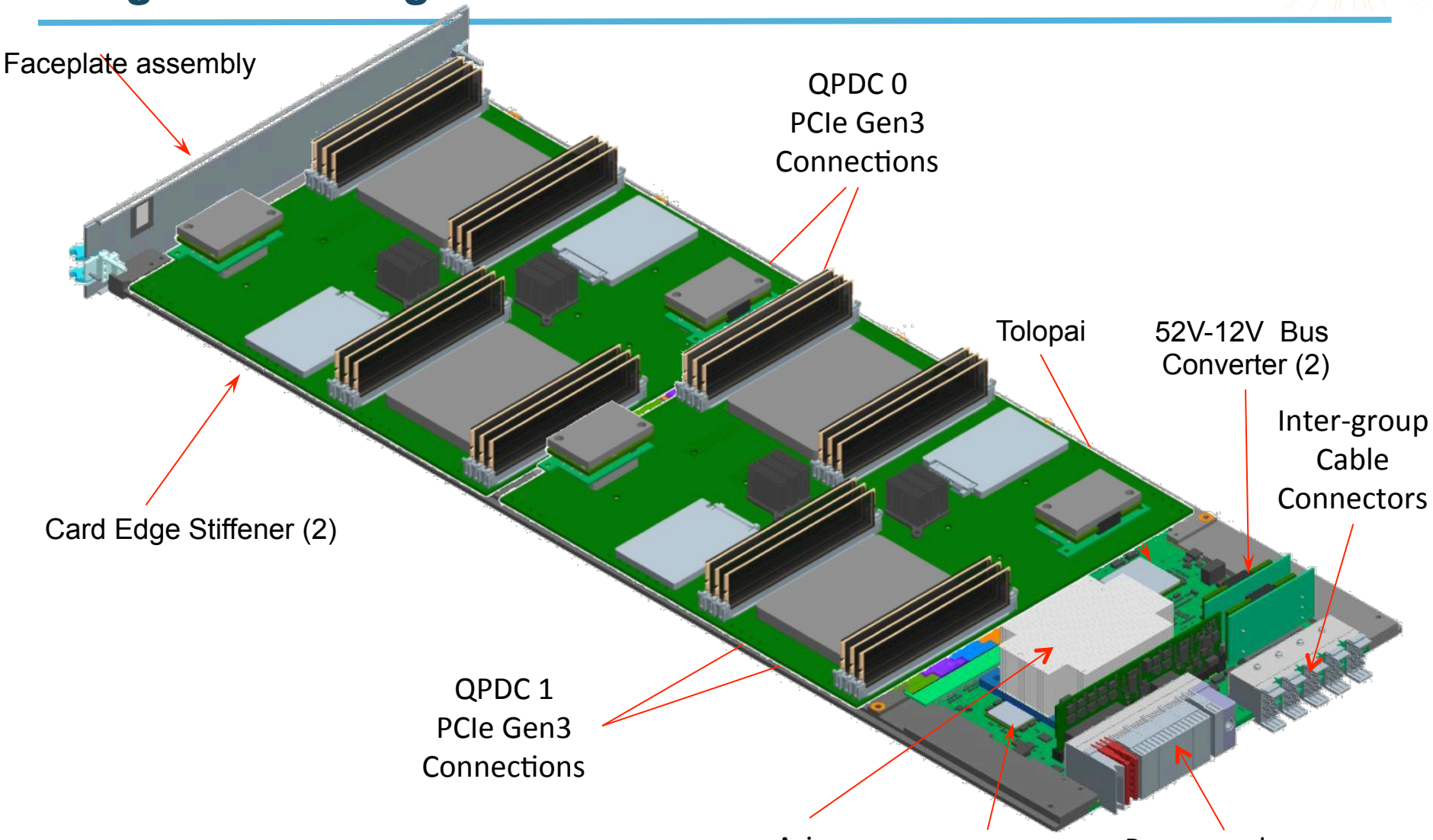
Intel "Knights Landing" Processor

- Next generation Xeon-Phi
- Single socket processor >3TF peak
- Self-hosted, not a co-processor
- Greater than 60 cores per processor
- Multiple threads/core
- 512b vector units (32 flops/clock – AVX 512)
- Improved single thread performance improvement over current generation Xeon-Phi co-processor
- Higher performance per watt
- High bandwidth on-package memory of similar capacity or greater than in current Xeon-Phi co-processor local memory (> 8GB)

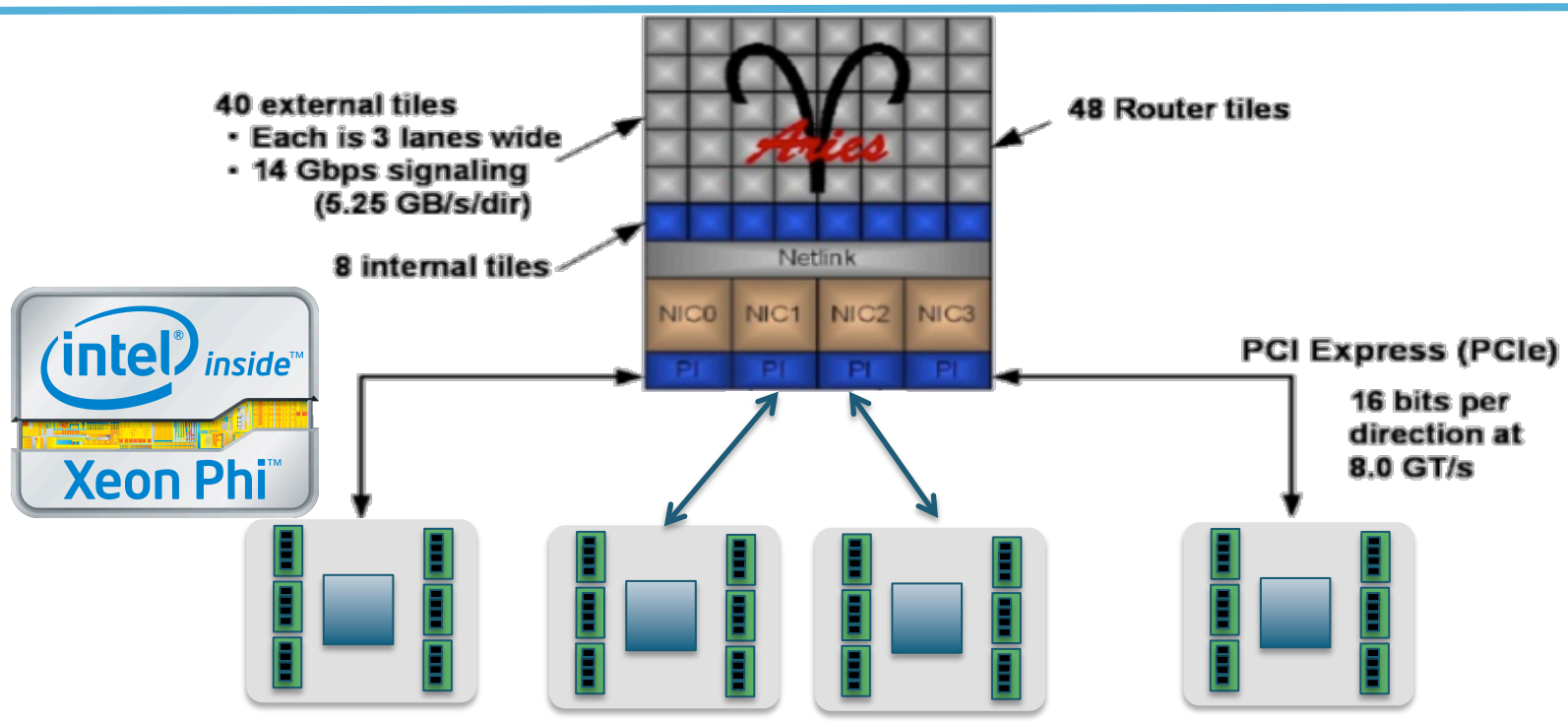
Programming Model Considerations

- **Knight's Landing is a self-hosted part**
  - Users can focus on adding parallelism to their applications without concerning themselves with PCI-bus transfers
- **MPI + OpenMP preferred programming model**
  - Should enable NERSC users to make robust code changes
- **MPI-only will work – performance may not be optimal**
- **On package MCDRAM**
  - How to optimally use ?
    - Explicitly or implicitly ??

Knight's Landing Blade

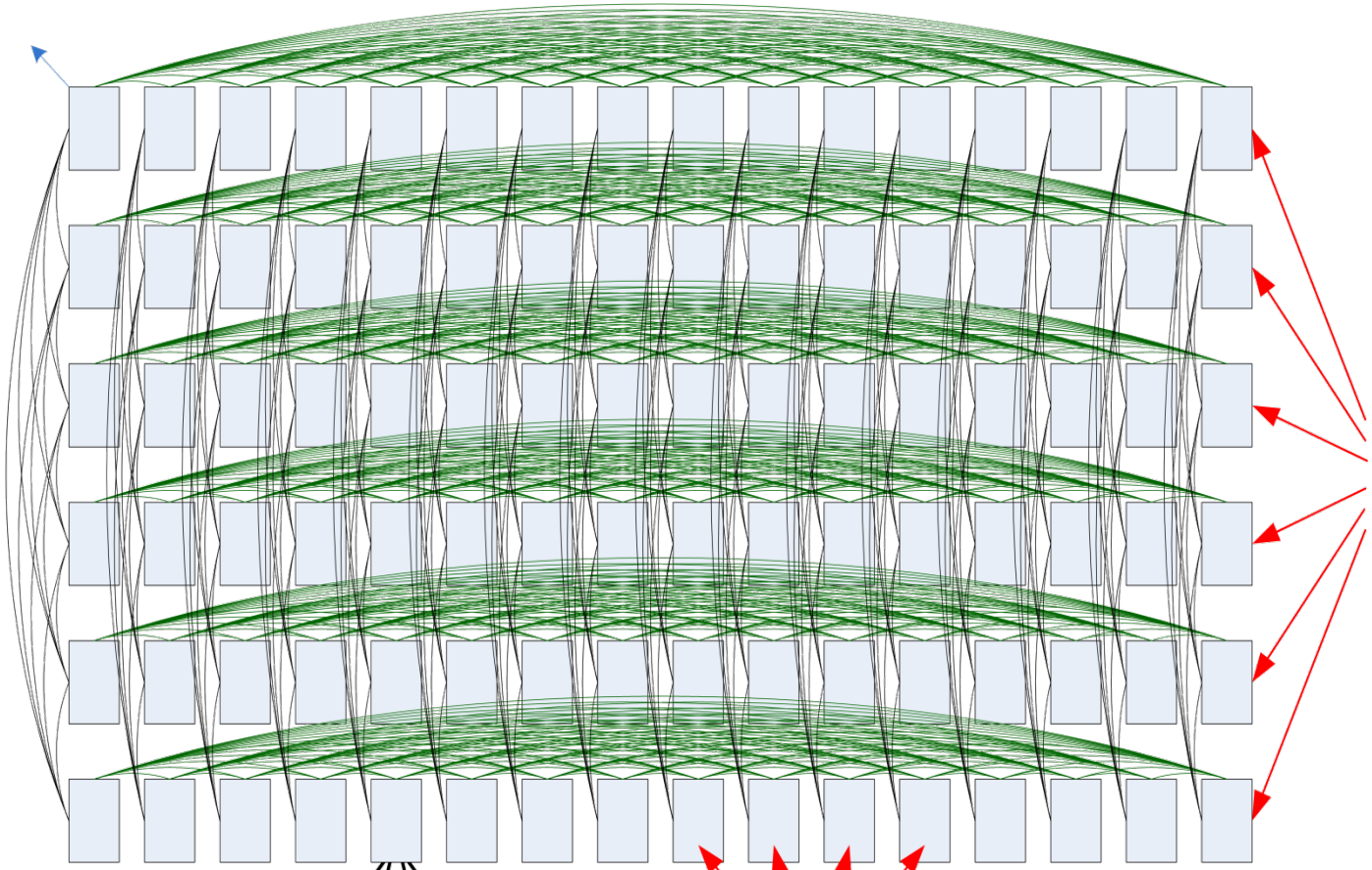


Overview of Aries Node Architecture



- **Aries**
  - PCIe Gen3 x16, 4 NICs per Aries
  - 48 Tile Router – 8 Internal links and 40 External links
  - 14 Gbps electrical and 12.5 Gbps optical links

Dragonfly "Group"



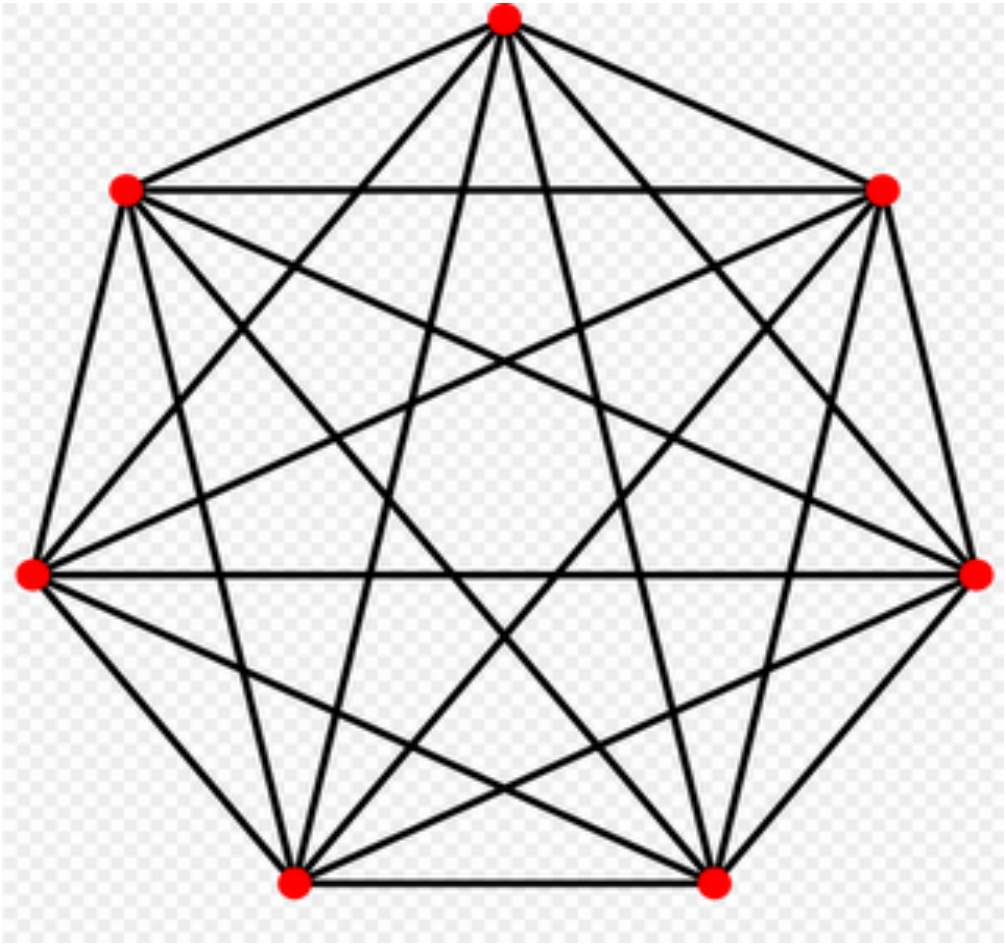
Local network can handle twice the aggregate injection bandwidth

6 chassis connected by cables to form a single group

4 nodes connected to each Aries router

16 Aries routers connected by chassis backplane

Top-down view of Inter-group connections



- **Each group is connected to every other**
- **Adaptive routing**
- **Note NERSC-8 will have 25 groups**
  - Each link will be a bundle of 4 cables ~1.7 TB/s
  - 40% of the optical ports will be populated

I/O Features

- **The NERSC-8 system will include several features which provide for the Storage, Access, and Transfer of user data sets at high performance**
  - 28PB Lustre "scratch" file system capable of 432GB/s
  - I/O gateway nodes to existing NGF GPFS file systems
  - Remote access to Edison's Parallel File System
  - 40Gb/s External Network connectivity for accessing archival storage(HPSS) and WAN resources

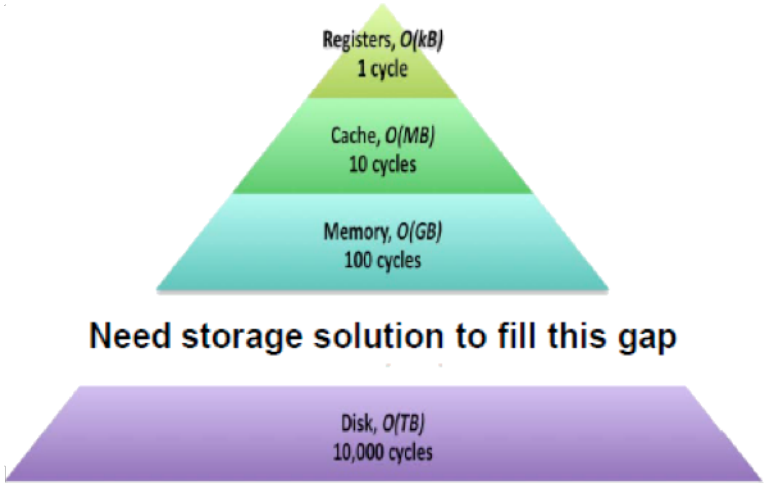
Parallel file system comparison

	Cori	Hopper
Bandwidth	432GB/s	35/35GB/s (70)
Metadata ops (creates/s)	77K/s	17/17 K/s (34)
Capacity	28.5PB	1.1/1.1PB
Delta-PFS*	29min	44min

Delta-PFS: Time to write 80% of memory to the Parallel File System

Burst Buffer

- **Flash storage which would act as a cache to improve peak performance of the PFS.**



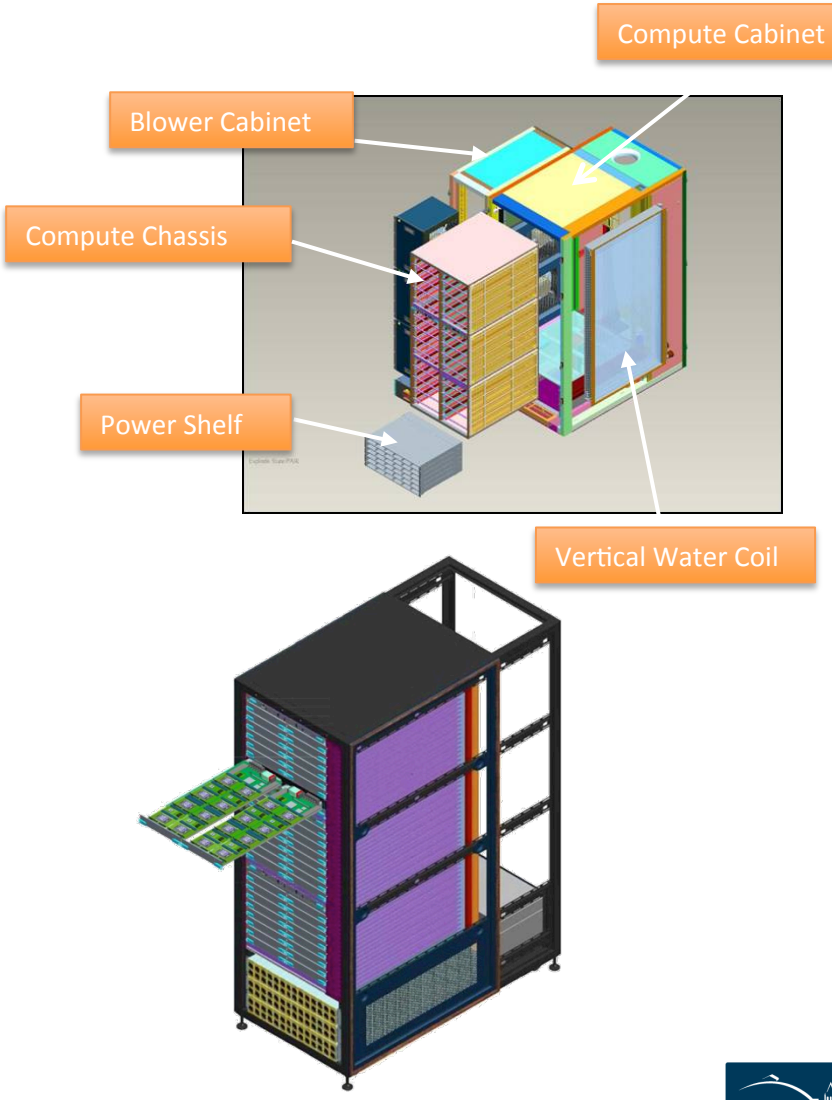
- **Flash is currently as little as 1/6 the cost of disk per GB/s bandwidth and has better random access characteristics(no seek penalty).**

Burst Buffer Option

- **Solid state storage(flash) presents an opportunity to significantly improve the I/O experience of NERSC users for several reasons.**
  - Lower cost for bandwidth(cost/capacity becomes limiting factor)
  - Better random I/O performance(no seek penalty)
  - Better small block(4K) performance.
- **Wide variety of use cases at NERSC**
  - Accelerating I/O
  - Checkpoint/restart
  - Reading large datasets
  - Launching area for shared libraries

Cabinet Design

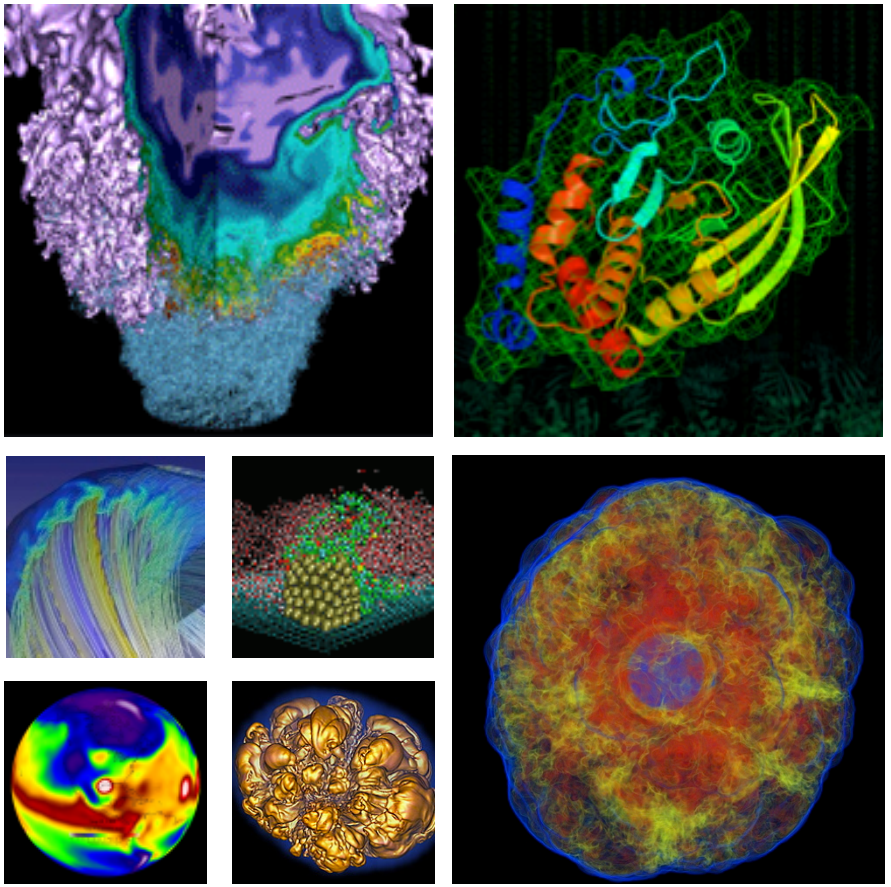
- **Rack**
  - 3 chassis / cabinet
  - Up to 16 blades/chassis
    - Up to 8 I/O blades
  - 4 Nodes/compute blade
    - 1 sockets/node
  - 2 single socket nodes/ service blade
    - 2 PCIe gen3 x8 slots/node
  - Water coil on right side of each cabinet



Advance Cooling Technology

- Primarily water cooled
- One blower assembly for each cabinet pair (group) + one at the end of row
- Compute rack cooling coil valve adjusts flow rate to maintain outlet air temp.
- Exhaust air can be room neutral or require residual cooling
- N+1 blower configuration
- Hot swap blower assembly
- Low noise

NERSC-8 programming environment and application readiness



NERSC 40 YEARS at the FOREFRONT 1974-2014

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Programming Models Strategy

- **The necessary characteristics for broad adoption of a new programming model is**
  - Performance: At least 10x-50x performance improvement
  - Portability: Code performs well on multiple platforms
  - Durability: Solution must be good for a decade or more
  - Availability/Ubiquity: Cannot be a proprietary solution
- **Our near-term strategy is**
  - Smooth progression to exascale from a user's point of view
  - Support for legacy code, albeit at less than optimal performance
  - Support for a variety of programming models
  - Support optimized libraries

Running on Cori

- Codes will probably run on NERSC-8 without any changes.
- To take advantage of the Knights Landing architecture, applications must
  - Exploit more parallelism
  - Express thread-level parallelism
  - Exploit data level parallelism
  - Manage data placement and movement
  - Accommodate less memory per process space

Application Readiness Approach

- **Start early; profile workload; set user expectations**
- **Enable a significant fraction of NERSC workload to run on NERSC-8.**
- **Be technology agnostic in optimizations**
  - (*i.e*. target parallelism, memory use, algorithms) to maintain investment going forward.
- **Performance portability to greatest extent possible**
- **Educate NERSC users**
- **Early test beds to learn about the issues as soon as possible**
- **Transfer lessons learned to/from broader DOE community**
  - Collaborate with ACES/ALCF/OLCF Application Readiness Teams

NERSC App Readiness Team formed in 2011 to look at manycore programming challenges


Katerina Antypas
NERSC-8 Project Lead

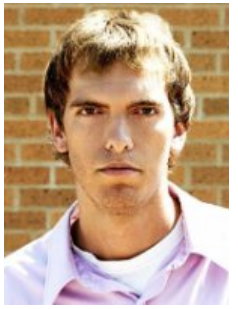

Nick Wright
ATG Group Lead

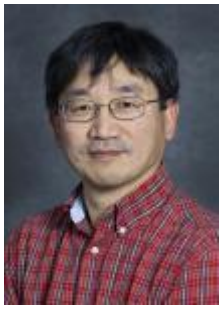

Richard Gerber
USG Group Lead


Harvey Wasserman
Chemistry

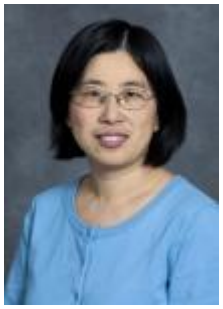

Brian Austin
Quantum Chemistry


Zhengji Zhao
Materials Science


Jon Rood
Applied Math/Bio-informatics


Woo-Sun Yang
Climate


Jack Deslippe
Materials Science


Helen He
Climate


Matt Cordery
Climate


Kirsten Fagnan
Bio-Informatics


Christopher Daley
Astrophysics/ Adaptive Mesh

- ***Team studied thread parallelism and vectorization in key codes on local test beds.***
- ***Result: compendium of case studies reflecting porting effort, performance results, best practices, common issues on real codes***

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

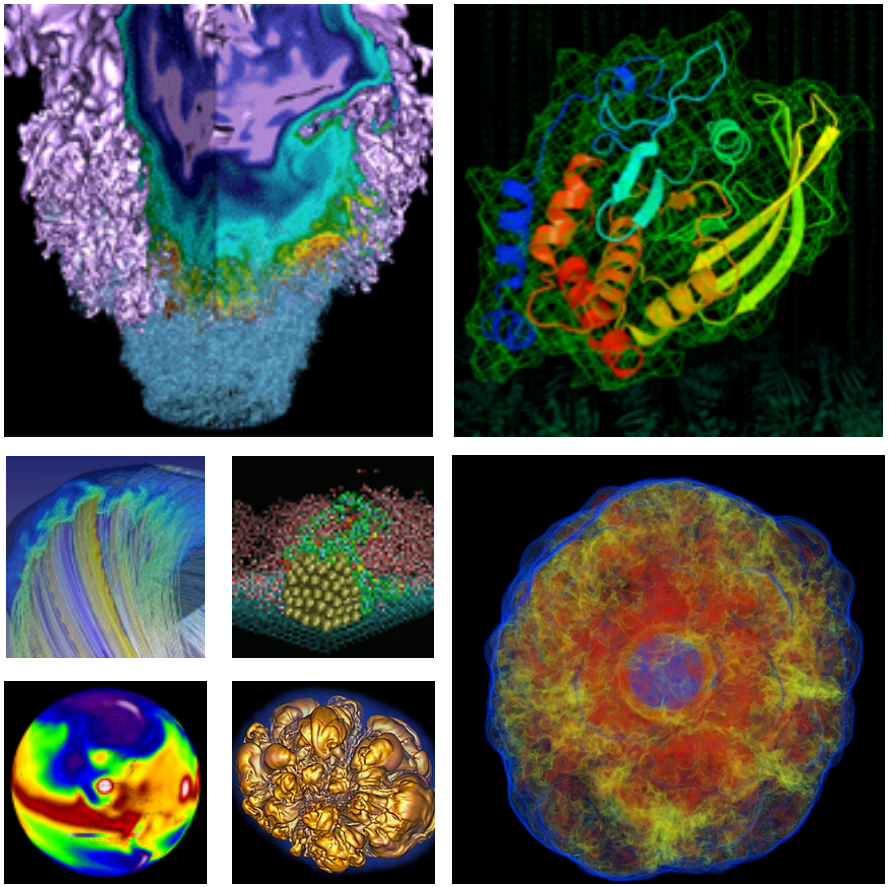
Some initial lessons learned

- Improving a code for advanced architectures can improve performance on traditional architectures.
- Inclusion of OpenMP may be needed just to get the code to run (or to fit within memory)
- Some codes may need significant rewrite or refactoring; others gain significantly just adding OpenMP and vectorization
- Profiling/debugging tools and optimized libraries will be essential
- Vectorization important for performance

What all this means for users

- **We will urge users to *begin preparing for N8 now*, using Edison and the "Babbage" Intel MIC test bed system**
- **Users can profile codes; examine vectorization levels and loop lengths; begin to transform loops**
- **Can also examine OpenMP parallelism**
  - Difficult to estimate performance effects, though
  - If code performs well on Babbage, it will probably perform well on N8
- **More Application Readiness training, announcements and plans will be coming soon**

Conclusions



Babbage Knights Corner Testbed

- **Babbage is a 45 node cluster with Knights Corner co-processor**
- **Can use co-processors in 'native' mode to be more representative Knights Landing architecture**

- **https://www.nersc.gov/users/computational-systems/testbeds/babbage/**

Conclusions

- **Programming model changes are coming and will affect computing at all levels; not just about preparing for Cori; is really about preparing for Exascale computing**
- **NERSC's goal is to provide *usable Exascale computing***
- **NERSC is dedicated to helping our users make this transition**