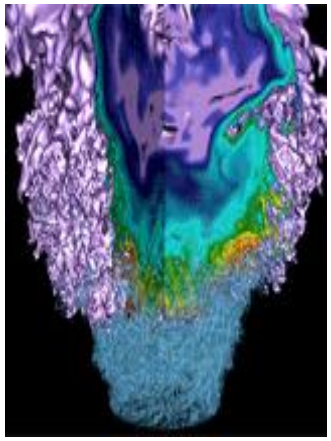
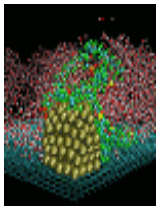
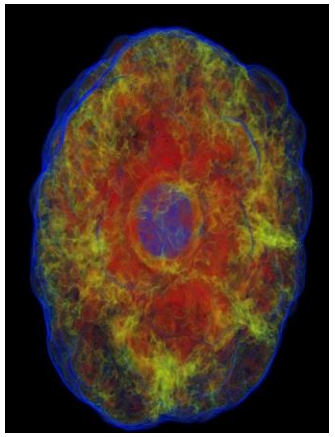
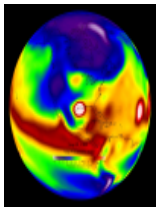
Cori – and data




NERSC 40 YEARS at the FOREFRONT 1974-2014

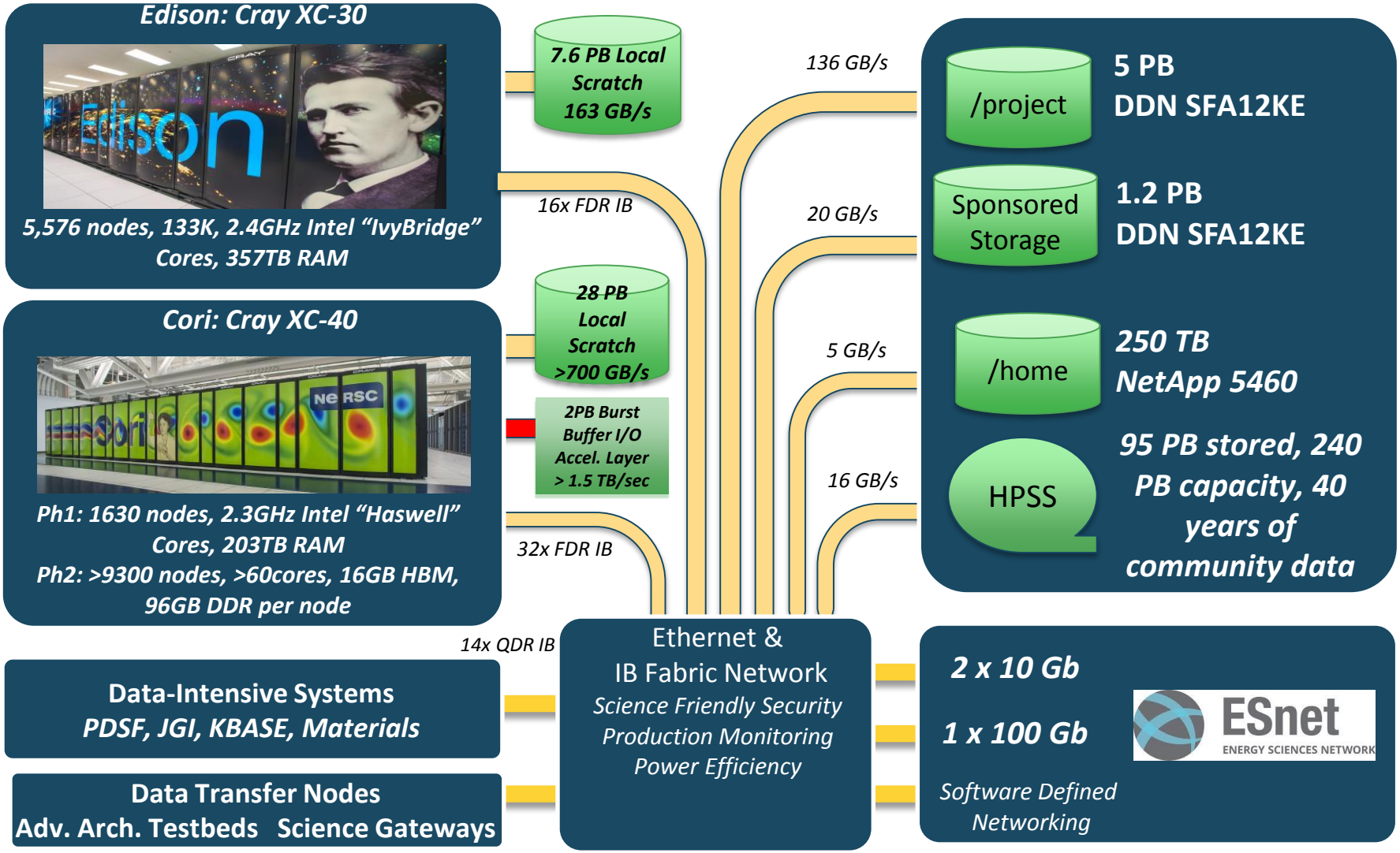
Wahid Bhimji

Alice Visit
Mar 15th, 2016

U.S. DEPARTMENT OF ENERGY | Office of Science

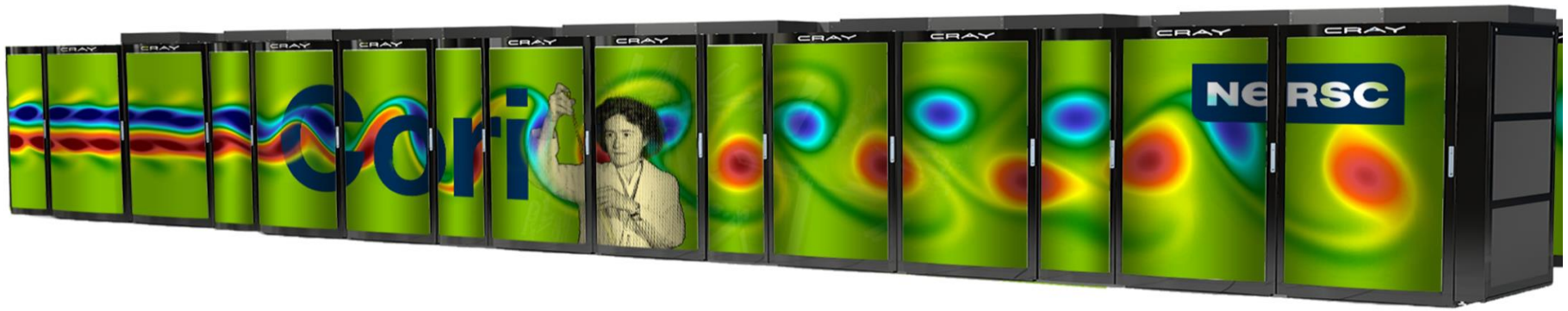
BERKELEY LAB

NERSC - 2016


Edison: Cray XC-30
5,576 nodes, 133K, 2.4GHz Intel "IvyBridge" Cores, 357TB RAM
Cori: Cray XC-40
Ph1: 1630 nodes, 2.3GHz Intel "Haswell" Cores, 203TB RAM
Ph2: >9300 nodes, >60cores, 16GB HBM, 96GB DDR per node
7.6 PB Local Scratch 163 GB/s
28 PB Local Scratch >700 GB/s
2PB Burst Buffer I/O Accel. Layer > 1.5 TB/sec
16x FDR IB
32x FDR IB
14x QDR IB
136 GB/s
20 GB/s
5 GB/s
16 GB/s
/project
5 PB DDN SFA12KE
Sponsored Storage
1.2 PB DDN SFA12KE
/home
250 TB NetApp 5460
HPSS
95 PB stored, 240 PB capacity, 40 years of community data
Ethernet & IB Fabric Network
Science Friendly Security
Production Monitoring
Power Efficiency
Data-Intensive Systems
PDSF, JGI, KBASE, Materials
Data Transfer Nodes
Adv. Arch. Testbeds Science Gateways
2 x 10 Gb
1 x 100 Gb
Software Defined Networking
ESnet
ENERGY SCIENCES NETWORK

Cori, a Cray XC40 system

- **Cori Phase 1 – partition to support data intensive applications**
  - 1630 Intel Haswell nodes
  - Two Haswell processors/node,
    - 16 cores/processor at 2.3 GHz
    - 128 GB DDR4 2133 Mhz memory/ node
  - NVRAM Burst Buffer to accelerate data intensive applications
  - Lustre Filesystem 28 PB of disk
  - Cray Aries high-speed “dragonfly” topology interconnect
- **Cori Phase 2 delivery with over 9,300 Intel Knights Landing compute nodes – mid 2016**
  - Self-hosted, (not an accelerator) manycore processor with up to 72 cores per node
  - 16GB On-package high-bandwidth memory at ~400 GB/sec
  - 96 GB DRAM memory per node



CORI Phase I 'data features': Now, Soon, Soonish

Filesystems

- Burst Buffer for high bandwidth, low latency I/O
- High-performance Lustre Filesystem: 28 PB of disk, >700 GB/sec I/O bandwidth
  - Multiple lustre metadata servers (DNE)
- Cross mounting of filesystems (Cori scratch on Edison and DTNs)
- Large amount of memory per node (128 GB/node) as well as high-mem nodes (775GB/node)

Queue Configuration (SLURM batch system)

- Large number of login/interactive nodes to support applications with advanced workflows
  - Used for Spark, JupyterHub (ipython), and long running workflow software
- Immediate access (realitme) queues for jobs requiring real-time data ingestion or analysis
- High throughput ,serial (shared) queues for analysis, screening, UQ, etc.
- Killable (preemptable) queue
- Internal sshd (CCM mode) in any queue

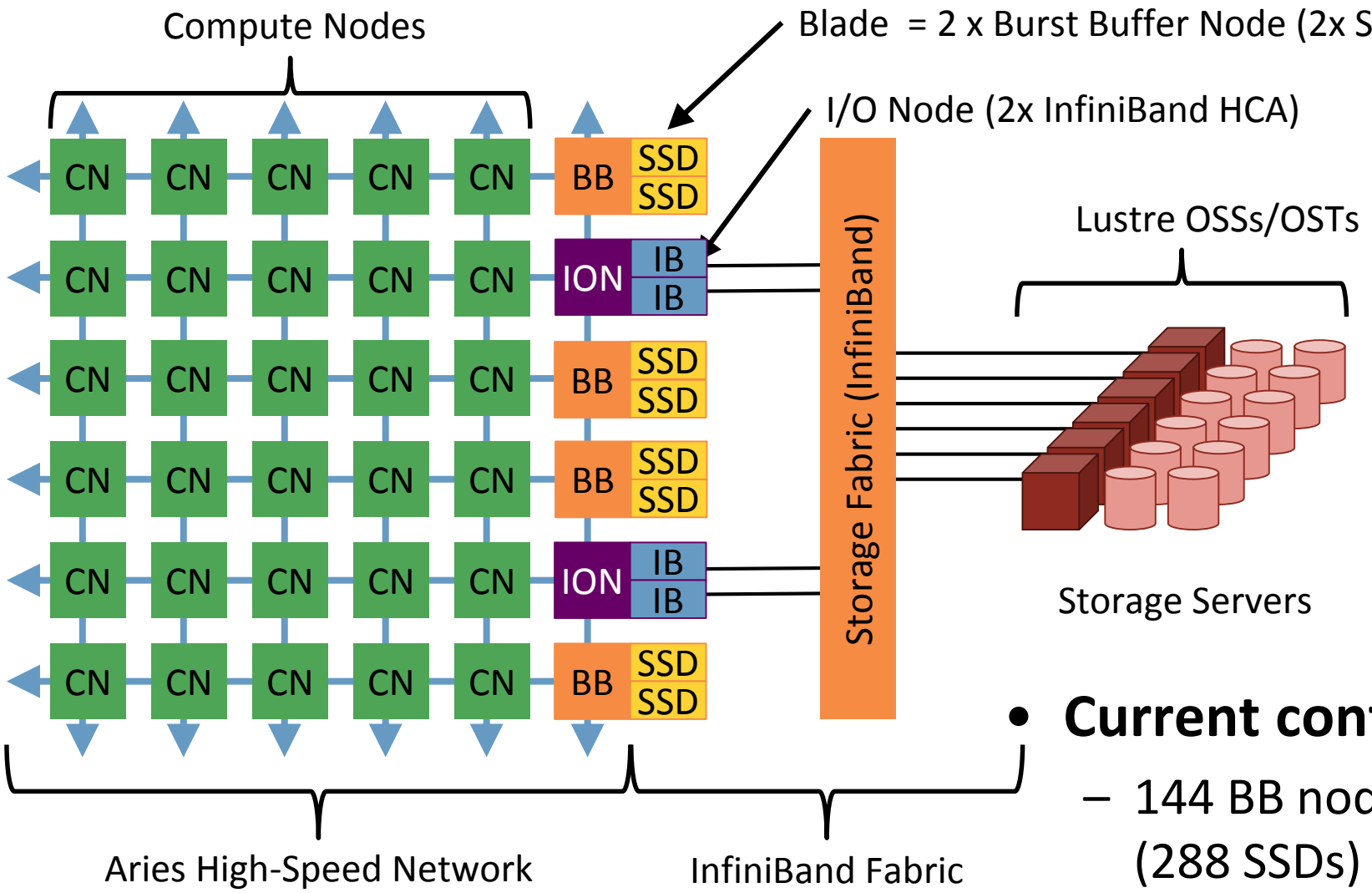
Networking

- Improved outbound Internet connections (eg. to access a database in another center.)
- Software Defined Networking R&D for high bandwidth transfers in and out of the compute node

On Node software:

- Improved shared library performance
- User-defined images/Shifter

Burst Buffer Architecture



- **Current config**
  - 144 BB nodes (288 SSDs)

NVRAM based – 'Burst Buffer'

https://www.nersc.gov/users/computational-systems/cori/burst-buffer/

Configuration:

- ~1.5PB capacity, ~1.5TB/s for full Cori System
- Half with Phase 1 (144 nodes with 2x3.2TB SSD modules)
- Available via SLURM batch system integration with Cray 'Data Warp' Software

```
#!/bin/bash
#SBATCH -N 1
#SBATCH -p regular
#SBATCH -t 1:00:00
#DW jobdw type=scratch access_mode=striped capacity=50GiB
#DW stage_in source=/global/cscratch1/sd/glock/bigdata.txt
             destination=$DW_JOB_STRIPED/bigdata.txt
             type=file

srun -N 1 ./myjob.x -input=$DW_JOB_STRIPED/bigdata.txt
```

Applications:

- **IO improvements: high bandwidth** reads and writes, e.g. checkpoint/restart (> 900 GB/s measured); **high IOP/s**, e.g. non-sequential table lookup; (> 12.5 m measured)
- **Workflow performance improvements: coupling applications,** using the BB as interim storage; Optimizing node usage by changing node concurrency part way through a workflow (using a persistent BB reservation)
- **Analysis and Visualization: In-situ / in-transit; Interactive** (using a persistent BB reservation)

'Early User Program' Including ALICE underway now

Extras

Intel “Knights Landing” Processor

- **Next generation Xeon-Phi, >3TF peak**
- **Single socket processor - Self-hosted, not a co-processor, not an accelerator**
- **>60 cores per processor with support for four hardware threads each; more cores than current generation Intel Xeon Phi™**
- **512b vector units (32 flops/clock – AVX 512)**
- **3X single-thread performance over current generation Xeon Phi co-processor (KNC)**
- **High bandwidth on-package memory, up to 16GB capacity with bandwidth projected to be 5X that of DDR4 DRAM memory**
- **Higher performance per watt**
- **Presents an application porting challenge to efficiently exploit KNL performance features**

Upgrading Cori's External Connectivity

Enable 100Gb+ Instrument to Cori

- Streaming data to the supercomputer allows for analytics on data in motion
- Cori network upgrade provides SDN (software defined networking) interface to ESnet. 8 x 40Gb/s bandwidth.
- Integration of data transfer and compute enables workflow automation

Cori Network Upgrade Use Case:

- X-ray data sets stream from detector directly to Cori compute nodes, removing need to stage data for analysis.
- Software Defined Networking allows planning bandwidth around experiment run-time schedules
- 150TB bursts now, LCLS-II has 100x data rates

We currently deploy separate Compute Intensive and Data Intensive Systems

Compute Intensive

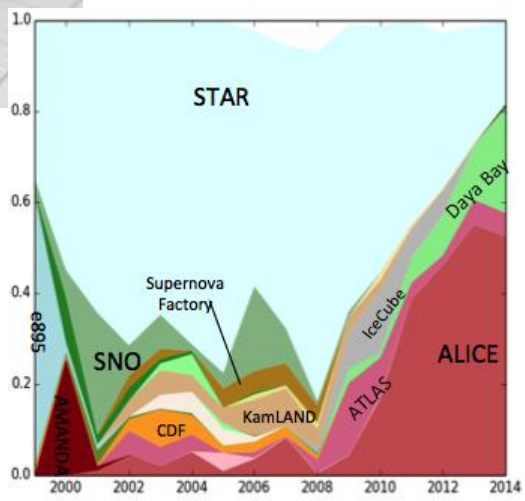
Data Intensive

PostgreSQL

SciDB

mongoDB

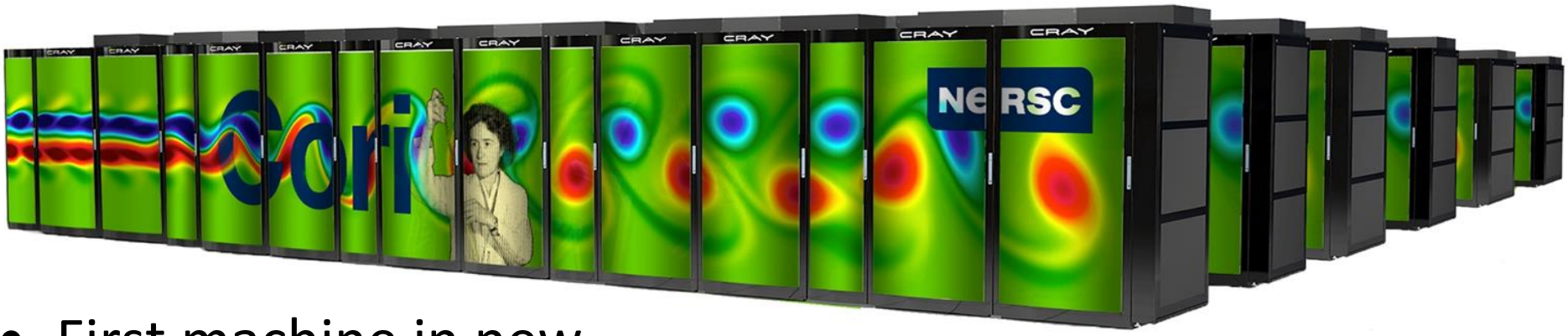
Genepool



PDSF

The Cori System

- **Cori will cater for HPC and HTC, support existing users of HPC (e.g. Edison) and HTC (e.g. Carver) but also enable new data workflows**



- First machine in new Computational Research and Theory (CRT) building

U.S. DEPARTMENT OF ENERGY | Office of Science