Integration of NERSC HPC with the ALICE Grid

R. Jeff Porter

ALICE Tier-1/Tier-2 Workshop

April 16, 2018

U.S. DEPARTMENT OF ENERGY | Office of Science

Outline

- **A bit about NERSC**
- **Why not HPC?**
- **Using NERSC HPC**
- **Where we are & where we may go**

A bit about NERSC

- **NERSC: DOE Office of Science Flagship High Performance Scientific Computing**
  - Available to all DOE Office of Science sponsored research
  - Allocations reflect Program Offices Science priorities

- **Computing for Scientific Research**
  - Large HPC systems
  - Multi-PB global & scratch file systems
  - Large archival storage (HPSS)
  - Extensive user support services
  - External Interfaces:
    - Data Transfer, Science Gateway & OSG/Grid Services

- **Review feedback → support data analytics**
  - Creation of Data Department in NERSC
  - Cori & future machines will include data intensive requirements

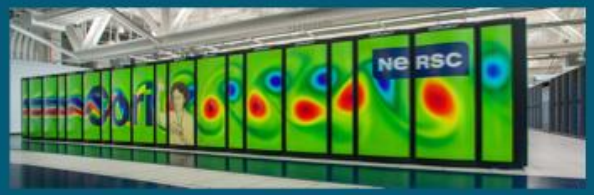
NERSC Platforms

ALICE

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

16x FDR IB

28 PB Local Scratch >700 GB/s

1.5 PB "DataWarp" >1.5 TB/s

32x FDR IB

80 GB/s

50 GB/s

5 GB/s

12 GB/s

Global Scratch — 3.6 PB 5 x SFA12KE

/project — 5 PB DDN9900 & NexSAN

/home — 250 TB NetApp 5460

HPSS — 50 PB stored, 240 PB capacity

Data-Intensive Systems
PDSF, JGI,KBASE,HEP
14x QDR

Vis & Analytics Data Transfer Nodes
Adv. Arch. Testbeds Science Gateways

Ethernet & IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
WAN

2 x 10 Gb

1 x 100 Gb

Software Defined Networking

ESnet
Energy Sciences Network

Section II

- **Why not HPC?**

Why not HPC?:
Well … some challenges for using HPC

- **Special Proprietary Internal network**
  - External network connection is discouraged (blocked) as it might interfere with tightly coupled processes.
- **Special OS and limited use of non-HPC software tools**
  - no local disk space for CVMFS
  - no FUSE for CVMFS
- **Restrictive access policies**
  - Multi-factor Authentication
  - Export control restrictions – limits who can access system

So Why HPC?

- **HPC is growing dramatically**
  - US DOE Exascale Initiative

U.S. DEPARTMENT OF ENERGY

Secretary of Energy Rick Perry Announces $1.8 Billion Initiative for New Supercomputers

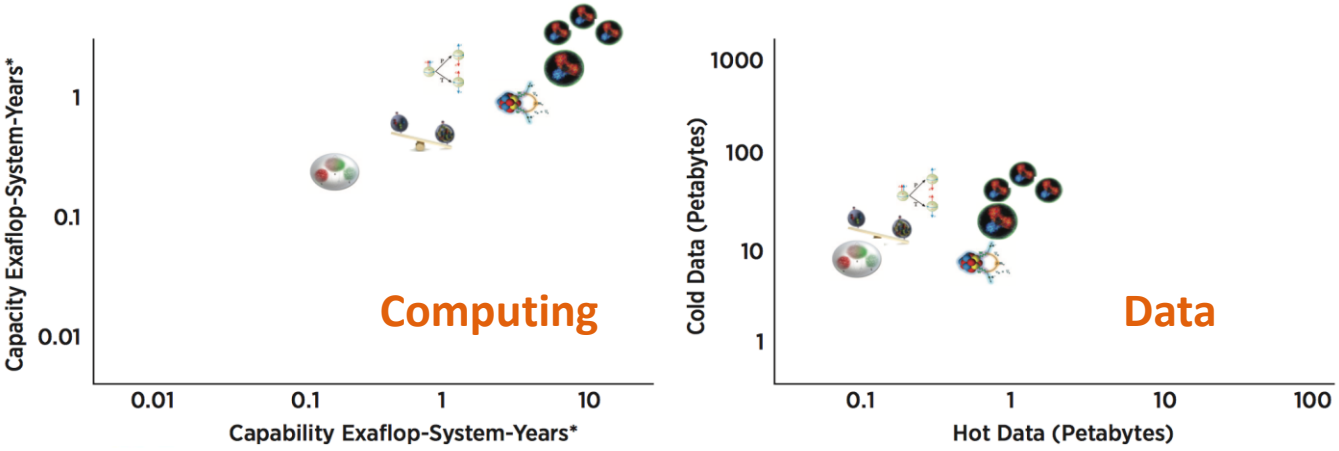
April 10, 2018

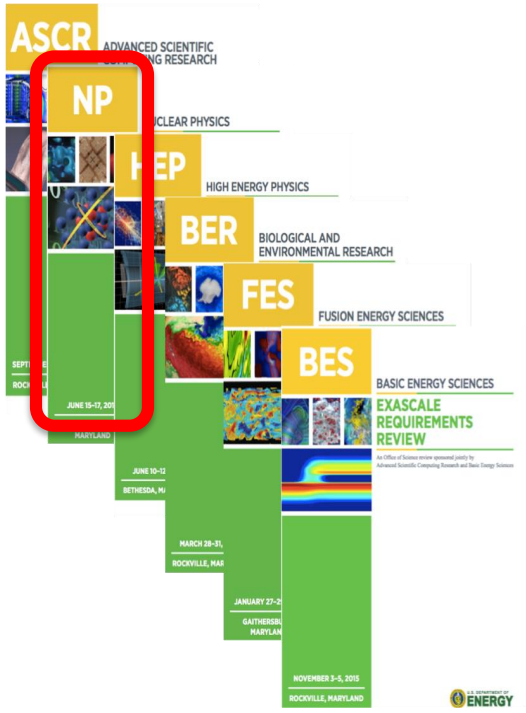
- **NERSC HPC Resources are allocated / given to researches**
  - CPU needs of experiment are small relative to theorists' simulation plans

➢ **Next NERSC machine (N9) arrives in 2020, in time for ALICE O2 era**

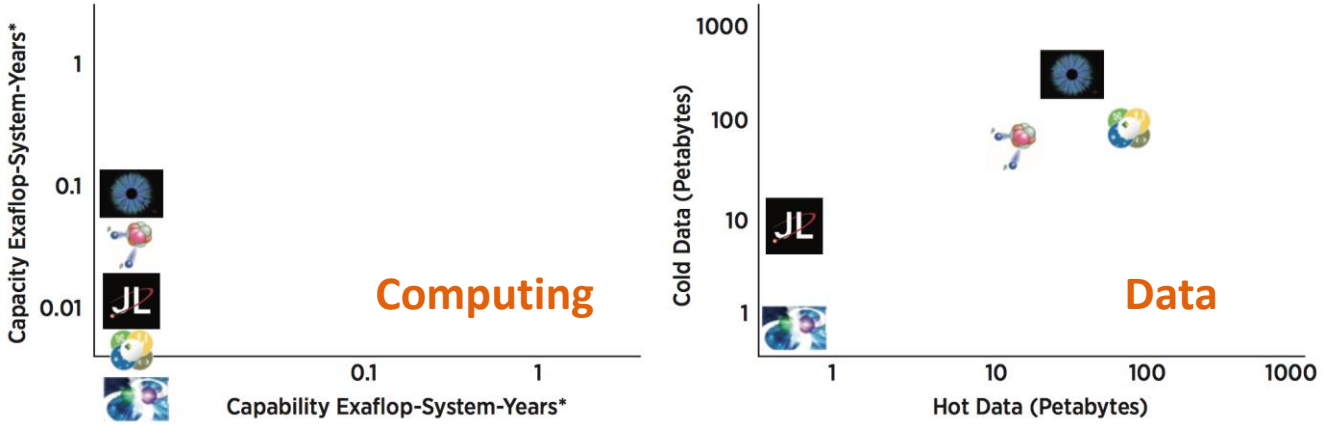
DOE NP/ASCR Exascale Requirement Revew Report

1 of 5 Theory Contributions



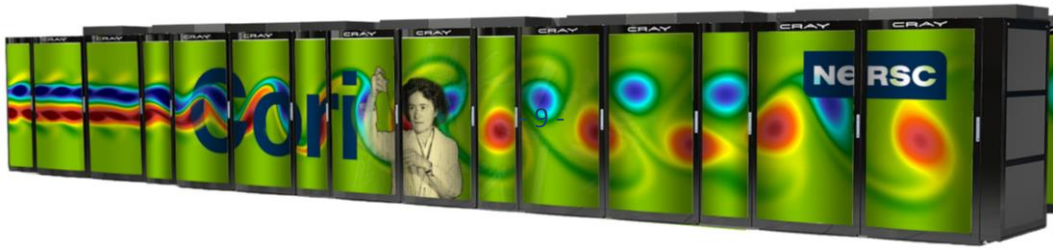


Experiment Contribution



NERSC and Cori

- **NERSC at LBL, production HPC center for US Dept. of Energy**
  - 1000 projects and 7000 diverse users across science domains
- **Cori – NERSCs Newest Supercomputer – Cray XC40**
  - Phase 1 (since Nov 2015): >2300 Intel Haswell dual 16-core nodes 128 GB
    - ~75000 cores or 1400 kHS06
  - Phase 2 (since late 2016): >9,600 Intel Knights Landing nodes
    - ~600,000 cores or ~ 4,000 kHS06
  - Cray Aries high-speed “dragonfly” topology interconnect
- **Lustre Filesystem: 27 PB ; 248 OSTs; 700 GB/s peak performance.**



U.S. DEPARTMENT OF ENERGY | Office of Science

Cori faces the HPC Data Challenge

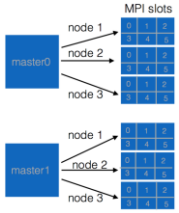
- **Special Proprietary Internal network**
  - External network connection is allowed, resilient at ~MB/s per job slot. ✓
- **Special OS and limited use of non-HPC software tools**
  - Container support via NERSC Shifter (similar to Singularity) ✓
  - CVMFS currently supported by NFS export ✓ **we'll see if it scales!**
- **Typically restrictive access policies**
  - Grid services run by local user ✓
  - ALICE has a history at NERSC

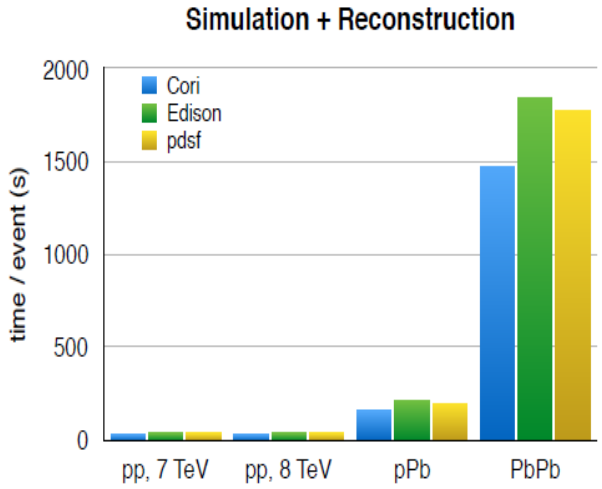
Section III

- **Using NERSC HPC**

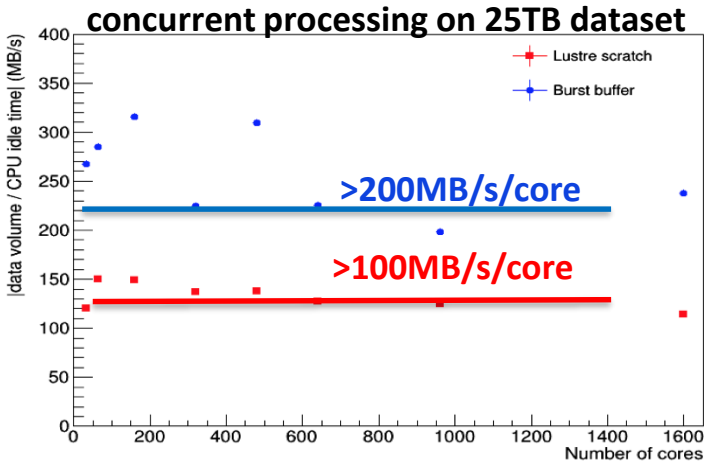
ALICE processing on NERSC HPC

- **Analisa Tool – Markus Fasel as primary developer**
  - Python-based tool developed here in **2015** to run multiple serial jobs as MPI jobs
    - Submitter splits master job into N subjobs
    - Workers (MPI) process subjobs (payload)
  - Hides resource complexity from users
    - Sub-modules for Edison & Cori
  - Software build tool
    - Optimized with local builds
    - Only used because of lack of CVMFS





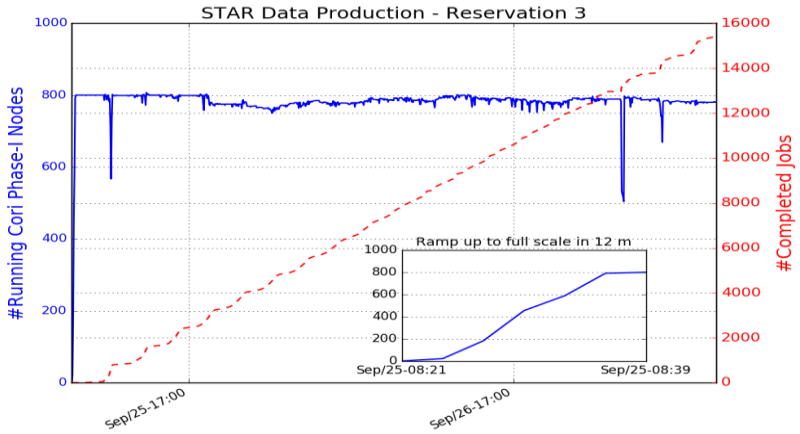
- **Tested normal (Grid) simulation payloads**
  - CPU Performance competitive with ALICE batch farms
  - Excellent I/O Performance



Markus Fasel, ACAT 2016

2017 STAR Raw Data Reconstruction on NERSC Cori

- **Project Targeted Serious Backlog of Raw data for reconstruction**

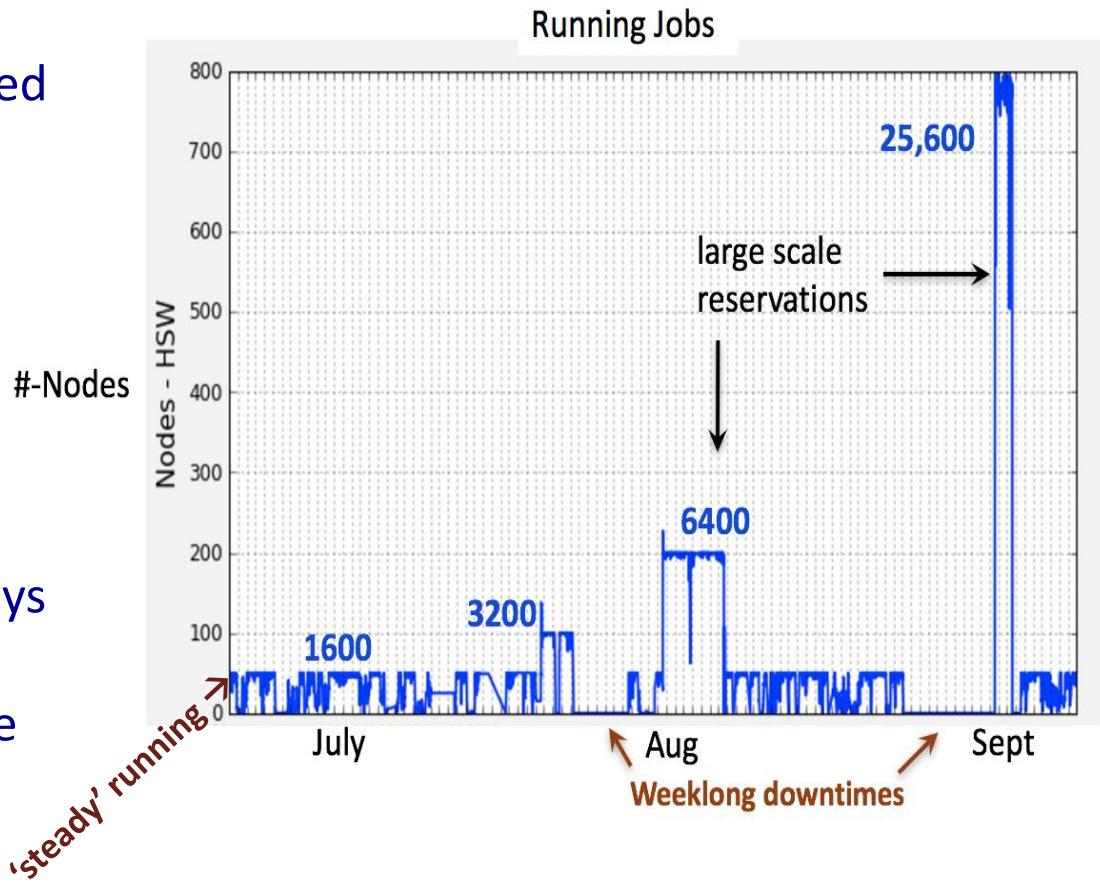


- **Demonstration project**:
  - Stable & efficient processing
  - Scaled to >25000 cores
    - 0-to-25k in 12 minutes
  - 98% efficiency

- **Pipeline & Framework makes use of:**
  - Globus toolkit for data transfers
  - Shifter images with STAR software
  - Snapshot of MySQL Calibration DB on each processing node
  - Central MongoDB for process orchestration

STAR Processing on Cori

- **Observations:**
  - Insufficient resources devoted to serial queue:
    - Need to use single or multi-node queues
  - Stable processing but single node limited to 1600 cores:
    - Average ~800 jobs
  - Occasional scheduled maintenances for several days
  - NERSC willing to schedule large scale reservations once shown to be efficiently used



LBNL Resources: LDRD with Physics At LBNL

- **Develop use of HPC for HEP/NP data analysis**
  - Goal is to prepare for running on NERSC N9 Machine
  - Focus to support ATLAS, ALICE, DayaBay, Lz, STAR, ….
  - In practice – we have some effort to help ALICE work at NERSC
- **Year 1 identified several items to be addressed:**
  - Scalable use of CVMFS for distributed software deployment
    - Currently available in a NFS mount
  - Automated data-delivery in HPC that optimizes network & storage
  - Multi-node/multi-core job manager
- **Year 2 provide a testbed for analysis models**

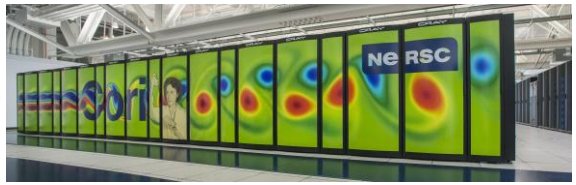
Automated Data Access Model

- **I/O done directly via XRootD**
  - Used by ALICE, ATLAS and STAR
  - Extend use XRootD caching to dynamically manage local disk cache?

~10s MB/s

NERSC Data Transfer Nodes
XRootD
Proxy Cache

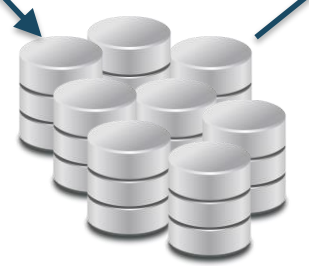
data request



Developed for use by OSG/ATLAS→

100s GB/s

- How best to use?
  - ALICE prefers a static SE
  - Resource likely limited to ~250TB & Tied to annual allocation



← Work done at GSI for ALICE

Cori Scratch

Section IV

- **Where we are & where we may go**

Where we are now

- **ALICE VOBox on a Cori Workflow node:**
  - Workflow nodes are equivalent to Cori login nodes but reserved for special use cases
- **Processing Environment:**
  - Use Shifter image from PDSF where ALICE currently runs
  - /cvmfs bind mount of NFS export, in use by CMS & ATLAS
  - Test queue options:
    - Serial queue
    - Single node queue, but 1 job/node, move to whole-node
- **2018 Goal is to demonstrate routine use of Cori**
  - NERSC is willing to augment our allocation during this transition year with PDSF to test use cases.

Where we may be going

- **2019 Allocation request in September, begins Jan 2019**
  - 500 slots are likely, >1000 possible – calculated as 24x7x365
- **Will request storage capacity – e.g. few 100TBs – but it will not be permanent**
  - 1 year, subject to renewal → non-standard ALICE SE
  - Larger storage allocation may be allowed if included in purge policy,
    - delete data not touched for 12 weeks.
- **How best to use in 2019 given non-standard SE allocation?**
  - Normal T2 site?
    - Ignore non-standard storage, use nearby SE maintained at LBNL/HPCS cluster
  - Raw data reconstruction with reservations and data pre-staging?
    - Reservations can manage large processing over short periods
  - Nano-AOD analysis facility?
    - data is ephemeral by definition and leverage I/O capabilities

Summary

- **HPC is a huge focus for US DOE Scientific program**
  - NERSC will be a part of that largess!
- **NERSC is very interested in supporting data-intensive science on their production HPC systems**
  - ATLAS & CMS are already using NERSC for simulations
- **Our task is to figure out how best to use the facility for ALICE, taking into account that the next big machine will come online just before the O2 ERA**