NERSC Platforms

**Edison: Cray XC-30**
- 1,376 nodes, 133K, 2.4GHz Intel “IvyBridge” Cores, 357TB RAM
- 7.6 PB Local Scratch 163 GB/s

**Cori: Cray XC-40**
- Ph1: 1630 nodes, 2.36GHz Intel “Haswell” Cores, 203TB RAM
- Ph2: >9300 nodes, >60cores, 16GB HBM, 96GB DDR per node
- 1.5 PB “DataWarp” >1.5 TB/s

**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

**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
New System in 2020: Perlmutter

• **Target:** 3x Cori processing capacity
  – Assuming GPU utilization!
1 of 5 Theory Contributions

Computing

Data

Experiment Contribution

Computing

Data
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 ✓ scaled to 15k jobs

• **Typically restrictive access policies**
  – ALICE Grid services run by local user ✓
  – ALICE has a history at NERSC

• **Our tests of normal (Grid) simulation payloads & STAR reconstruction**
  – CPU Performance competitive with ALICE batch farms
  – Excellent I/O Performance to Lustre
Where we may be going

• 2019 Allocation requested in September, began Jan 2019
  – 500 slots ~ 24x365

• Storage capacity when needed – e.g. few 100TBs
  – 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
    • Network between NERSC & HPCS?
  – Large scheduled production with CPU reservations + data pre-staging?
    • Reservations can manage large processing (20k jobs) over short periods (days)
  – Dynamic Nano-AOD analysis facility?
    • data is ephemeral by definition and leverage I/O capabilities
Where we are now

• **ALICE VOBox**
  – On a Cori Workflow node ==> reserved Cori login nodes
    • Do not allow incoming network connections
    • CVMFS is not available in a submit environment
  – On a HPCS VM
    • LQ/SSH_style.pm \(\rightarrow\) qqueued.sh, qrunning.sh, qsubmit.sh
    • Can be moved into NERSC VM environment when they are in production

• **Processing Environment:**
  – Shifter image from PDSF where ALICE currently runs
  – /cvmfs mount of NFS export, tested as scale by STAR, CMS & ATLAS
  – Test queue options:
    • Serial queue
    • Single node queue, but 1 job/node, move to whole-node

• **Biggest Issue**
  – No dedicated time to work on this task!
Summer students tasks?

• Harden Vobox method for job-submission
  – Independent jobs:
    • Serial jobs
    • Multi-core
    • Multi-node
  – Interdependent processes:
    • DDS on HPC?

• Port to jAlien instead of PERL-based systems
• XRootD on data transfer nodes?
• Scheduled production?
• Use of accelerator architectures?
• CHEP2019 talk/poster?