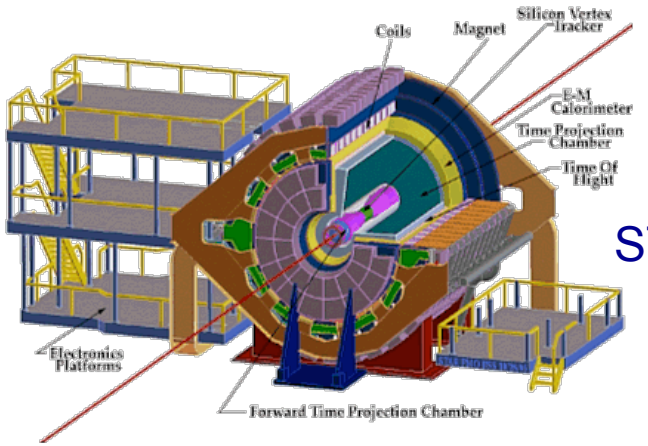
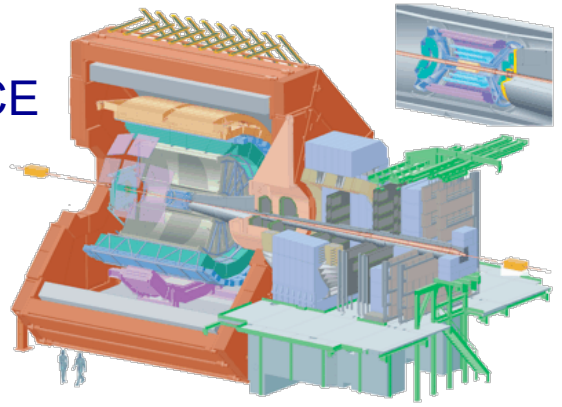
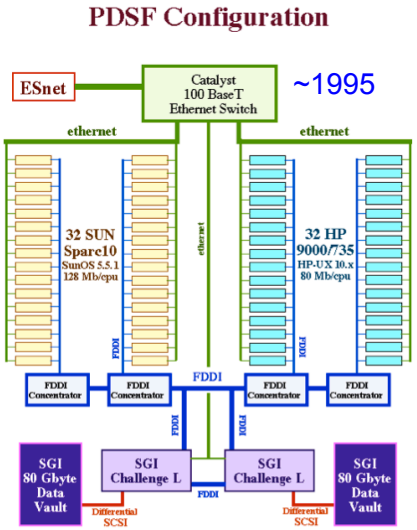


STAR

ALICE



Present and Future Computing Requirements for Heavy-ion collider experiments: Focus on ALICE experiment at the LHC



R. Jeff Porter

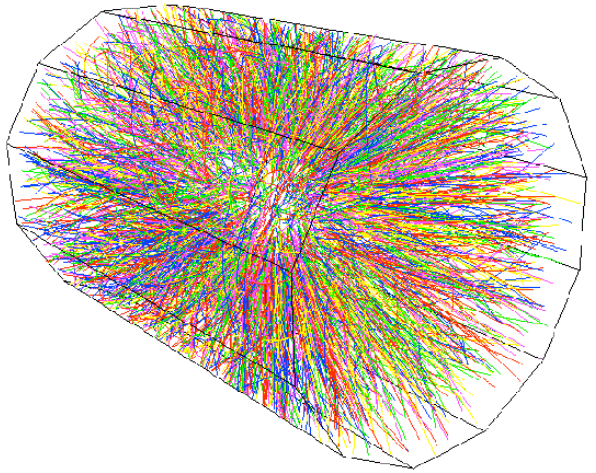
(LBNL)

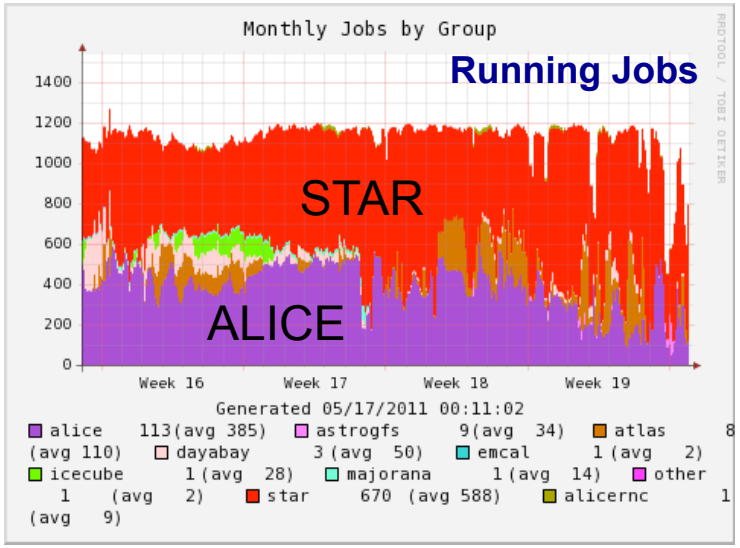
PDSF at a Glance

Interactive Nodes	
4	pdsf.nersc.gov pdsf[1-4].nersc.gov
Compute Nodes	
205	1100 Cores
GPFS Filesystems	
641TB	Eliza[1-18]
Local Disk	
450TB	
Batch System	
SGE	Sun

Processing model for ALICE & STAR on NERSC/PDSF

- Event-based processing → "pleasantly parallel"
  - Event = independent "collision" registered in detector
  - Task = process an event collection, set of independent jobs
  - Naturally distributed: on cluster, grid of clusters …
- Data Intensive
  - Annual multi-PB sized datasets – HPSS is a critical NERSC facility
  - ALICE minimum site recommendation ~ 0.5TB disk/core
    - PDSF: 1000 cores & 2PB disk, ~2TB/core
    - NERSC: 200k cores & 5PB disk, ~0.025TB/core
- Large Software Infrastructures
  - 100s k lines of C++ code
  - Built on C++ ROOT Class Library
    - http://root.cern.ch/drupal/
  - With GEANT for building detector simulations
    - http://geant4.web.cern.ch/geant4/
  - No HPC parallel programming methods
  - Code portability can be a challenge



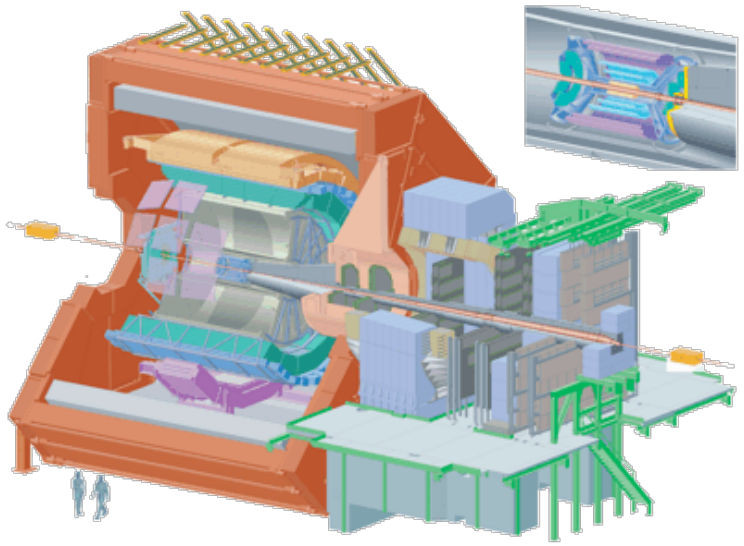


ALICE & ALICE-USA

- ALICE Experiment
  - Dedicated heavy ion experiment at the LHC

    study physics of strongly interacting matter
    at extremely high energy densities

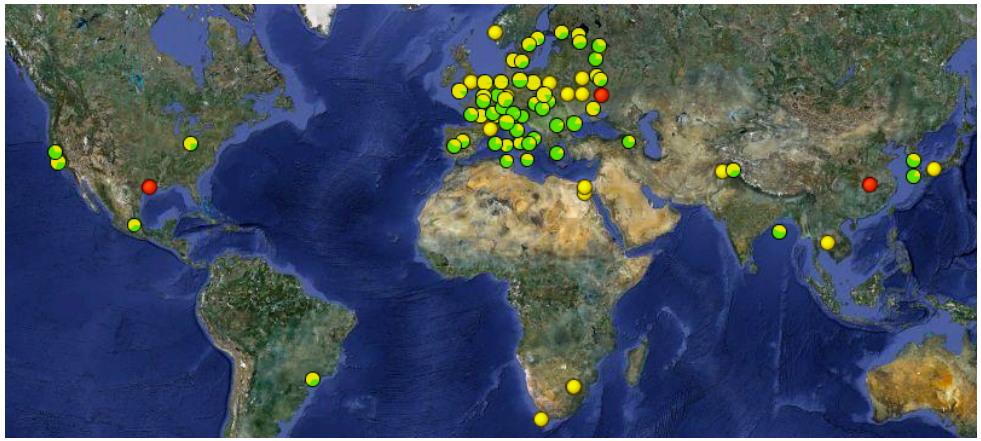
  - 1000 physicists, 33 countries
  - ALICE-USA: ~50 physicists, 11 institutions



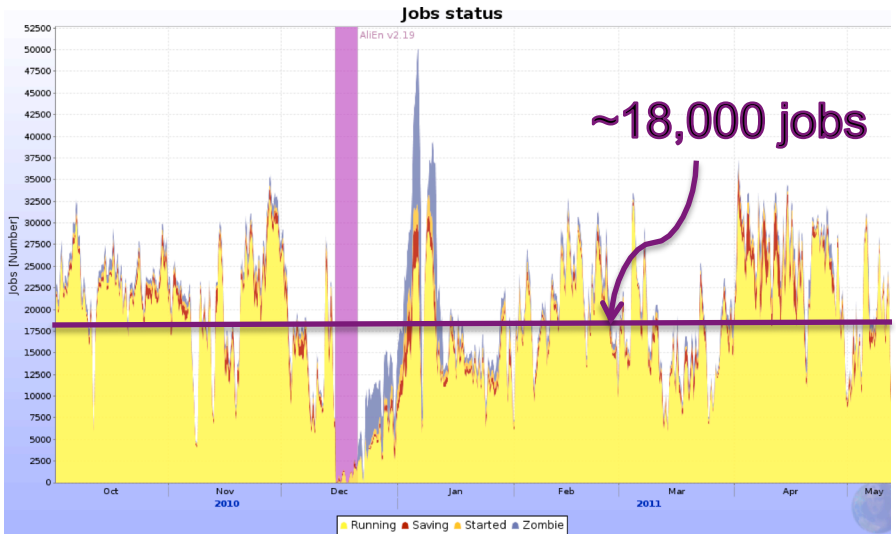
- ALICE-USA Computing Project
  - Build a facility to meet ALICE-USA computing obligations to ALICE
  - Proposal developed in 2008/2009 & project approved in 2010:
    - Locate facility at two sites: LLNL/LC and LBNL NERSC/PDSF
    - LBNL as project host lab:
      - James Symons NSD Division Director & Peter Jacobs NSD PI
      - Ron Soltz (LLNL) Computing Coordinator
      - Jeff Porter (LBNL) Project Manager

ALICE Grid Facility

- ALICE activities at NERSC are primarily via Grid
- AliEn "Alice Environment" Grid :
  - VOBox → site POC to manage
    - Job submission onto site
    - software deployment
    - site monitoring
  - Central Task Queue (TQ) at CERN
  - Job agents pull jobs from TQ
    - Matches jobs with local data
    - Verifies local environment
    - Request client software versions
  - Jobs run where data exists

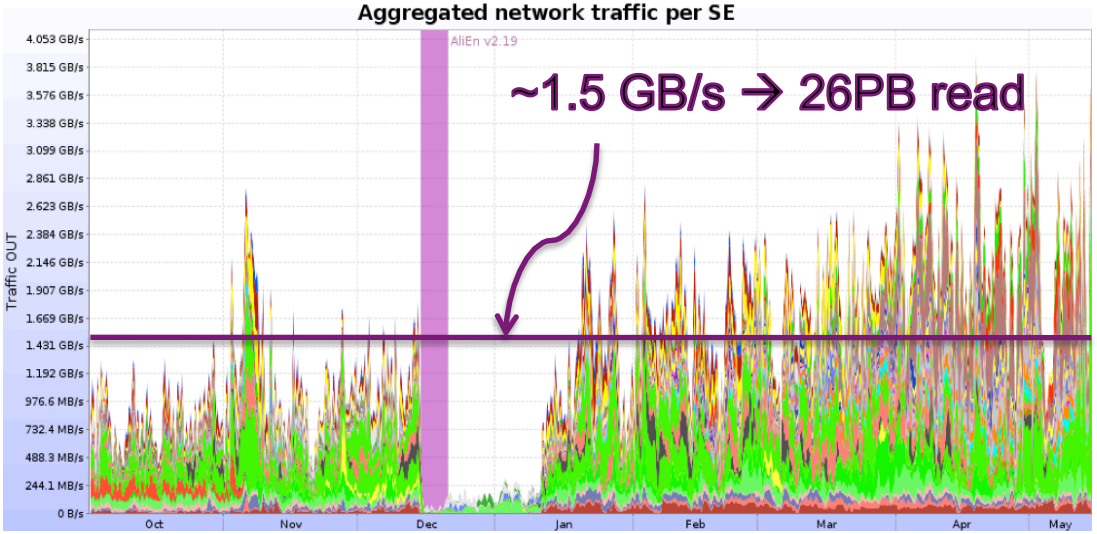


- Monitoring with MonALISA
  - Jobs Statistics:
    - Total numbers & by site
    - CPU, memory usage, … extendable
  - Storage capacities & availability
  - Network topology



Data Management on ALICE Grid Facility

- Global AliEn File Catalog – a tolerable single point failure
  - Logical File Name, multiple physical instances
  - Currently 150 million file instances, representing ~15PB total volume
- Grid-Enabled XROOTD based Storage Elements
  - XROOTD provides a distributed file system
  - Storage Elements at each site with local manager/redirector
  - Grid-Enabled + global redirector provides world-wide distributed file system



ALICE on NERSC/PDSF – Next Three Years

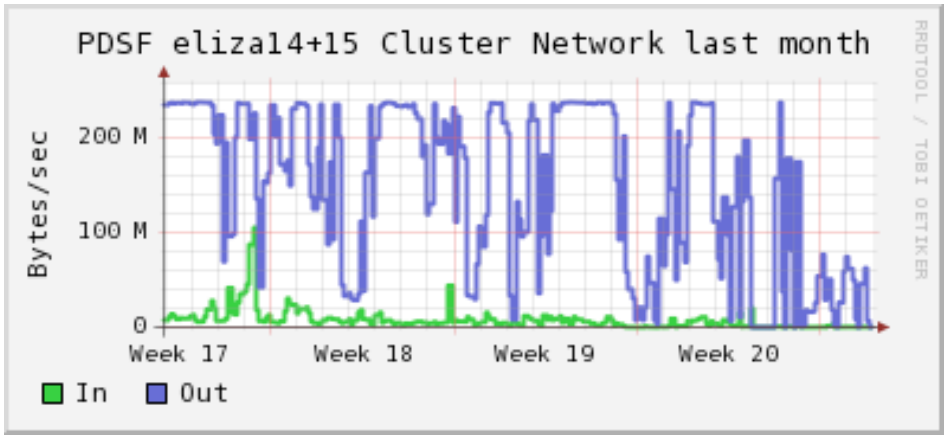
- Procurements based on Computing Project Plan w/ annual targets
  - 500 new compute cores
  - 500TB new XROOTD disk storage
  - 500TB new HPSS allocations with XROOTD enabled front end
- CPU & disk stable after 3 years (2013): 1500 cores & 1.5 PB disk
- HPSS new allocation requests will continue

Fiscal Year		FY10	FY11	FY12	FY13
NERSC Procurements					
CPU (nodes)		40	40	40	40 30
New CPU (kHEP06)		3.8	5.7	7.7	7.7 5.8
New Disk (PB)		0.65 0.35	0.65 0.4	0.8 0.5	0.6 0.4

ALICE job-level usage requirements

- Parallel approach is done at job submission not programming level
  - Experiment infrastructure splits jobs up by input data
  - Sends job to data on the grid when possible

- Memory needs
  - Productions jobs ~2GB/core
  - Some analysis jobs ~4-5GB/core



- I/O requirements per core
  - Production jobs ≤ ~0.1 MB/s
    - Simulation jobs ~10s kB/s
    - Event reconstruction jobs ~0.1 MB/s
  - Analysis jobs can use several ~MB/s
    - Efficiencies as cpu-time/walltime drop significantly

Usage requirements for distributed computing

- Grid Infrastructure
  - Support Job submission & data transfer protocols
  - Support Large Virtual Organizations: user authorization & authentication
- WAN data processing, dynamically move data to compute resource
  - Data access decisions based on real time network monitoring
  - Grid-enabled peer-to-peer data access technologies?
- Cloud/VM Technology?
  - STAR: Interested
    - Successful running CPU-bound production on Cloud: simulations & reconstruction
    - VMs provide a solution to portability challenges
    - Add new data management issues
  - ALICE: Not very interested
    - I/O performance cost appears too high for general use
    - Code portability is a higher priority (e.g. code runs on Carver)
  - User analysis is generally I/O bound & not found to be a good match

ALICE & Future HPC needs

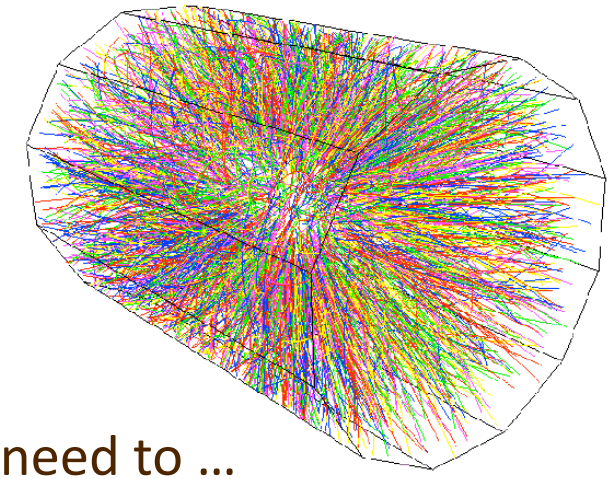
- Grid approach works well for production tasks
  - Highly managed input data, job properties are well known
  - Processing times of ~months are expected

- User analysis more challenging
  - requires quick turn around, ~hours or ~day to process
  - Re-access the same input dataset many times

- ALICE promotes separate analysis facilities based on PROOF
  - http://root.cern.ch/drupal/content/proof
  - Normal (non-grid) user login access
  - Highly optimized I/O on pre-staged target datasets of ~10s TB scale
  - Infrastructure splits job into N parallel jobs & merges output
  - Interest in porting to NERSC – perhaps PDSF

Many-Core Future

- The field recognizes the need to adapt
  - Talks & posters at Computing in High Energy & Nuclear Physics (CHEP) 2010
    - http://event.twgrid.org/chep2010/
  - Examples mostly target multi-core issues & scaling up



- Reducing multi-core resource contention
  - Jobs share memory for common data
    - Geometry, Calibrations, Conditions
  - File reader / Job launcher on node to fill the core space
- Lot of concurrency in the data and tasks, but still need to …
  - Prototype common processing tasks & structures for using parallel techniques
  - Develop techniques to target GPU and other specialized parallel architectures
  - Learn how to expose parallelism in the frameworks

Summary

- Heavy-ion collider experiment computing model
  - Pleasantly parallel → easily distributed
  - Data intensive → large storage (Disk & Tape) & I/O requirements
  - Very large code base → can limit portability
- For our use of NERSC
  - Large data storage capacities, multi-PB on HPSS
  - Maintain high bandwidth capacity between NERSC & ESNet
  - Maintain support for grid services
  - Different support model for large Virtual Organizations (VOs) of ~1000 users
- Field recognizes value in learning to use new HPC architectures
  - Can't be blind to cpu performance increases, even on commodity HW
  - May allow expansion onto other NERSC resources
    - Elastic expansion for periods of high demand
    - New analysis "facilities" on standard NERSC Hardware
- Field is rich with data, scientific advancement from analysis techniques not feasible today
  - Many-particle correlation techniques