RHIC/LHC heavy ion program requirements

NERSC NP Requirements Workshop
Apr 29-30, 2014
High Temperature Phases of QCD
Rich program over ~14 years

Top Energy: $^{\text{AuAu}} \sqrt{s} = 200 \text{ GeV}/c^2$

Polarized: $^{\text{pp}} \sqrt{s} = 200 \& 500 \text{ GeV}/c^2$

Reference data: $^{\text{pp}}, ^{\text{dAu}}, ^{\text{CuCu}}, ^{\text{CuAu}}$

Beam Energy Scan: $^{\text{AuAu}} \sqrt{s} = 7, 11, 15, 19, 27, 39 \text{ GeV}/c^2$

Largest data set to date
$^{\text{AuAu}} \sqrt{s} = 200 \text{ GeV}/c^2$

700 million minimum bias events
500TB analysis ready data files
ALICE LHC Run 1: 2010-2013

pp @ 0.9 – 8. TeV/c²
PbPb @ 2.6 TeV/c²
pPb @ 5.02 TeV/c²

ALICE Run 1
7PB of Raw Data
16PB of derived data
Characteristics of computing in collider-based experiments

- **Event-based processing ➔ “pleasantly parallel”**
  - Each collision = independent event
    - Dataset = event collection
    - Distributed in independent files
  - Computing task
    - process an event collection
    - set of independent jobs, 1 job ➔ N files
  - Natural fit with distributed processing
    - On nodes, cluster of nodes, grids of clusters

- **Large, complicated detectors ➔ software infrastructure**
  - Requires algorithm expertise per subsystem
  - Common framework with reliance on common toolsets: ROOT, GEANT, ...
    - 10s of millions of lines of code

➢ **HPC methods are **not** typically used**
Characteristics of computing in collider-based experiments

- **HEP/NP data intensive science**
  - High precision measurements require statistically large samples
  - Experiments continuously operate over long running periods (6-10 months)
  - ALICE + STAR ~ 75% of PDSF share
    - PDSF: 2500 cores & ~4PB disk storage

- **Processing task: data reduction with pattern recognition**
  - Raw signals processed into detector ‘hits’
  - Detector hits into physics entities: particle tracks, energy deposition
  - User analysis: entities into spectra, correlations, ...

- **International collaborations: large scientific user base**
  - Operate vast, loosely-coupled resources over a sustained periods of time: ➔ High Throughput Computing (HTC)
STAR Physics Plans → 2017

• Precision measurement of heavy flavor (charm) production
• Two new major detector systems
  – Heavy Flavor Tracker (HFT)
  – Muon Telescope Detector (MTD)
• ~5x data increase
  – 2014 : 2+ billion AuAu @ 200 GeV/c²
  – 2015 : 200 million pp, 500 million pAu @ 200 GeV/c²
  – 2016 : 2+ billion AuAu @ 200 GeV/c²
  ➢ Note: data is analyzed for years
• Future program goal → high statistics Beam Energy Scan
ALICE LHC Run 2: 2015-2017

- Exploring LHC Energy Regime
  - accumulate statistics
  - pp, PbPb & pPb

- Detector upgrades
  - DCal → back-to-back calorimeter
  - Precision measurements at very large $p_T$
    - Jet-hadron correlations

- 2-3x data increase
  - 7 billion p+p events @ 8-14 TeV/c$^2$
  - 0.7 billion PbPb events @ 5.5 TeV/c$^2$
  - 0.4 billion pPb events @ ? TeV/c$^2$
STAR Computing Model

• Non-distributed model
  – 85% of work done at RACF at BNL
  – 15% at NERSC/PDSF + KISTI

• Data management in single instance STAR File Catalog
  – two-way mirror at NERSC

• Rely on site-specific data storage, GPFS, XRootD, ...

• STAR @ NERSC
  – Software built and maintained locally
  – Users log into PDSF and submit jobs on local batch system
  – heavy use of STAR purchased PDSF file systems
  – NGF is critical for migrating to other NERSC system, primarily Carver
  – Large HPSS allocation for archival of derived data
• Distributed model organized within WLCG Collaboration

• ALICE Grid Facility
  – Tier 0 at CERN
  – Several Tier 1 sites
    • Includes archival storage
      – None in US – by choice
  – ~80 Tier 2 sites
    • CPU & stable disk storage
    • 3 active in US: NERSC, LLNL/LC, & OSC
      – ORNL to replace LLNL

• Software distributed via CVMFS

• High Throughput Computing (HTC)
  – 500 million cpu-hrs/yr
  – >10 million jobs

Average 36K jobs/day
ALICE Data model

• Data distributed at generation & registered in FileCatalog
  – 1<sup>st</sup> copy at site of processing
  – 2<sup>nd</sup> copy at nearby site
  – 3<sup>rd</sup> copy - hot data only

• Jobs go to site with data
  – Can pull from WAN on error

• Data Access patterns
  – 10/1 read/write
  – Hot data is much higher.

• Analysis “Lego” trains reduce read access
  – Many analyses connected to same input
  – More than doubled <cpu/wall>
    • 10:1 is an improvement!

69 SEs, 29PB in, 240PB out, ~10/1 read/write
Storage Technology: XRootD

- Wide use by LHC exp.
  - 3 different models
  - 4\textsuperscript{th} including STAR
  - Distributed structure
    * Internal data discovery
    * Network access protocol

- Plug-in architecture
  - protocols, authN/authZ, ...
  - e.g. LSST parallel query

- Requires data management layer
  - Experiment-specific tools

Andy Hanushevsky’s NERSC talk*

Current Large Deployments

- LHC ALICE
  - Data catalog driven federation
- LHC ATLAS
  - Regional topology
- LHC CMS
  - Uniform topology with some regionalization
- LSST (Large Synoptic Sky Telescope)
  - Clusters MySQL servers for parallel queries

Each with 10s PBs distributed WW

- RHIC STAR
  - >5 PB w/ local disk on >500 of WNs
  - Local access only

\[\text{http://www.slac.stanford.edu/~abh/nersc/NERSC1311.pptx}\]
XRootD at NERSC

**ALICE Grid Enabled Storage Element @ NERSC**
- Part of global data storage system
  - Both WAN and local access
- 10 data servers → 0.72 PB
- ALICE supplied data management layer
  - SE Discovery @ ALICE Global FileCatalog
  - Data discovery locally with XRootD
  - Monitoring by ALICE MonaLisa module

**STAR XRootD@NERSC**
- Independent of system @ BNL
  - LAN access only → servers on compute nodes
  - 200+ servers → 1.0 PB
- Local data management layer
  - Scripts walk data & load local MongoDB
    - Includes XRootD, GPFS, NGF & HPSS
  - Users query for file lists, access w/ ROOT
Next Generation for ALICE: EOS

- CERN IT project built on top of XRootD
  - Dynamic life-cycle management with simple operation model
  - Significant system administration features on top of XRootD
  - Works well with cheap hardware
STAR & ALICE Baseline: → 2017

• ALICE-USA Computing Project
  – Manages disk and cpu procurements and deployment
    • Review of new 3-year proposal (FY15-FY17) this June
  – Expect modest growth at 2017:
    • 2x cpu capacity: ~10 kHEPSPEC06 → 20 kHEPSPEC06
    • 3x disk space: 0.7 PB → 2.0 PB
  – No large change in workflow
  – Little HPSS usage

• STAR Resources needs
  – Expect modest growth: cpu & disk → ~2x-3x
  – HPSS backlog due to manpower shortage ~ 1-2 PB
  – Computing Lead asked to investigate data preservation scheme with NERSC
    • 100% of STAR data moved to NERSC HPSS… several 10s PB

• Both ALICE & STAR leverage NERSC Grid enabled resources
  – Open Science Grid
Hitting HTC Limits

- HTC works best for near homogenous systems

- We do have specific tasks that fit accelerator architectures
  - But that breaks homogenous workflow structure

- Work is hitting on-board bottlenecks
  - 10GigE per worker node
  - 5GB/core memory

- Community needs to better use of whole node processing
  - HEP Colleagues @ ATLAS & CMS are further along
  - Multi-threaded GEANT and ROOT is critical
LHC High Luminosity Era: Run 4 3

Big Data Outlook

100 PB

2015-2017  2019-2021
ALICE High Luminosity Era: Run 3

- Run 3 (2019): ALICE to operate in continuous readout mode
  - Data rate off the detectors: ~TB/s → 1PB/day
    - Overwhelms predicted bandwidth and permanent storage capacities
    - Real-time online data reduction methods – Not triggered data, minimum bias!
  - Large & complex online compute facility
    - Must leverage trends in many-core
    - Offline quality event reconstruction

- Online/Offline ($O^2$) Project
  - Full/fast Offline processing in Online
    - Fast detector calibration, reco & QA
    - Final store ONLY reduced data
    - 100x data reduction with 100x event rate
O² Era Challenges

- **Online/Offline Reconstruction @ 50kHz Event Rate**
  - continuous data stream
  - Real-time event reconstruction
  - Data buffering
    - real-time 2\textsuperscript{nd}/final pass calib
    - event reconstruction

- **Large new Monte Carlo needs**
  - Currently 60% of grid resources
  - Scales approx \sim \#-real-events
    - New strategies underway
  - event sample increase \sim 100x !!

- Move some MC production onto HPC facilities?
  - ALICE hopes to leverage US opportunity ... of proximity?
Both Titan and Stampede use accelerator technologies.
ALICE Use of HPC Resources

- **ALICE**: base code already ported to Cray
  - Geant4 version 10
    - includes support for multi-threading
    - ALICE port to Geant4 exists
  - New rewrite project: Geant V
  - Root 6.x under development

- Workflow is a challenge
  - Working with ASCR PanDA project
  - Becomes trivial with one requirement
    - Outgoing connection from compute node → NAT
Summary

• STAR & ALICE have ‘pleasantly parallel’ event-based processing
  – HTC not HPC modes
• ALICE relies heavily on distributed processing & grid-enabled resources, STAR less so
• Requirements from both groups show modest (~3x) growth
  – STAR HPSS has a backlog for HPSS
• High luminosity running for ALICE in 4 years → HPC solutions
  – ALICE O2 project is underway to prepare
  – Code base evolution, GEANT & ROOT, are also underway
  – Could leverage NERSC resources for workflow & processing model
    • scale of need is modest