Running ALICE application on NERSC supercomputers

Markus Fasel
Jeff Porter
(Lawrence Berkeley National Laboratory)

Outline

- NERSC supercomputer systems
- Running ALICE jobs on NERSC HPC System
- First experience with ALICE jobs on NERSC supercomputers

NERSC Systems Today

NERSC 40 YEARS at the FOREFRONT

Edison: 2.58PF, 357 TB RAM

Cray XC30 5,576 nodes, 134K Cores

Hopper: 1.3PF, 217 TB RAM

Cray XE6 6,384 nodes 150K Cores

7.6 PB Local Scratch 168 GB/s

16 x FDR IB

2.2 PB Local Scratch 70 GB/s

16 x QDR IB

80 GB/s

50 GB/s

5 GB/s

12 GB/s

Global Scratch — **4.0 PB 5 x SFA12KE**

/project — **5 PB DDN9900 & NexSAN**

/home — **260 TB NetApp 5460**

HPSS — **70 PB stored, 240 PB capacity, 20 years of community data**

Data Intensive Clusters
Carver, PDSF, JGI
56x FDR, 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 G*b*

Software Defined Networking

ESnet

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Cori, the next NERSC Workhorse

- **Named for Biochemist Gerty Cori**
  - 1st American Woman to be awarded Nobel Prize
- **Cori Design includes**
  - Next-generation Intel® Xeon Phi™ Knights Landing (KNL) product with improved single thread performance targeted for highly parallel computing
- **Full Deployment in late 2016**
  - http://www.nersc.gov/users/computational-systems/cori/
- **But first … Cori Data**

Cori : the Deployment plan

Late 2016

- 20+ PF
- ~9K Knights Landing nodes
- 60+ Cores / node
- Massively Parallel Code

Late 2015

- 1+ PF
- ~2K Haswell nodes
- 32 Cores 128GB RAM/node

Scratch File System
28 PB, 400+ GB/Sec

0.5 TB/core
6MB/s/core

Burst Buffer
~1.5 PB, ~1.5TB/Sec

slide by Yushu Yao
Cori Data services lead

Limitations

- Outgoing network access
- MPI jobs
- Allocation per host
- accounting in Walltime, needed to be defined before job starts
- Jobs become killed after allocation is exceeded
- No swap, very little local tmp

Our tool: ANALISA

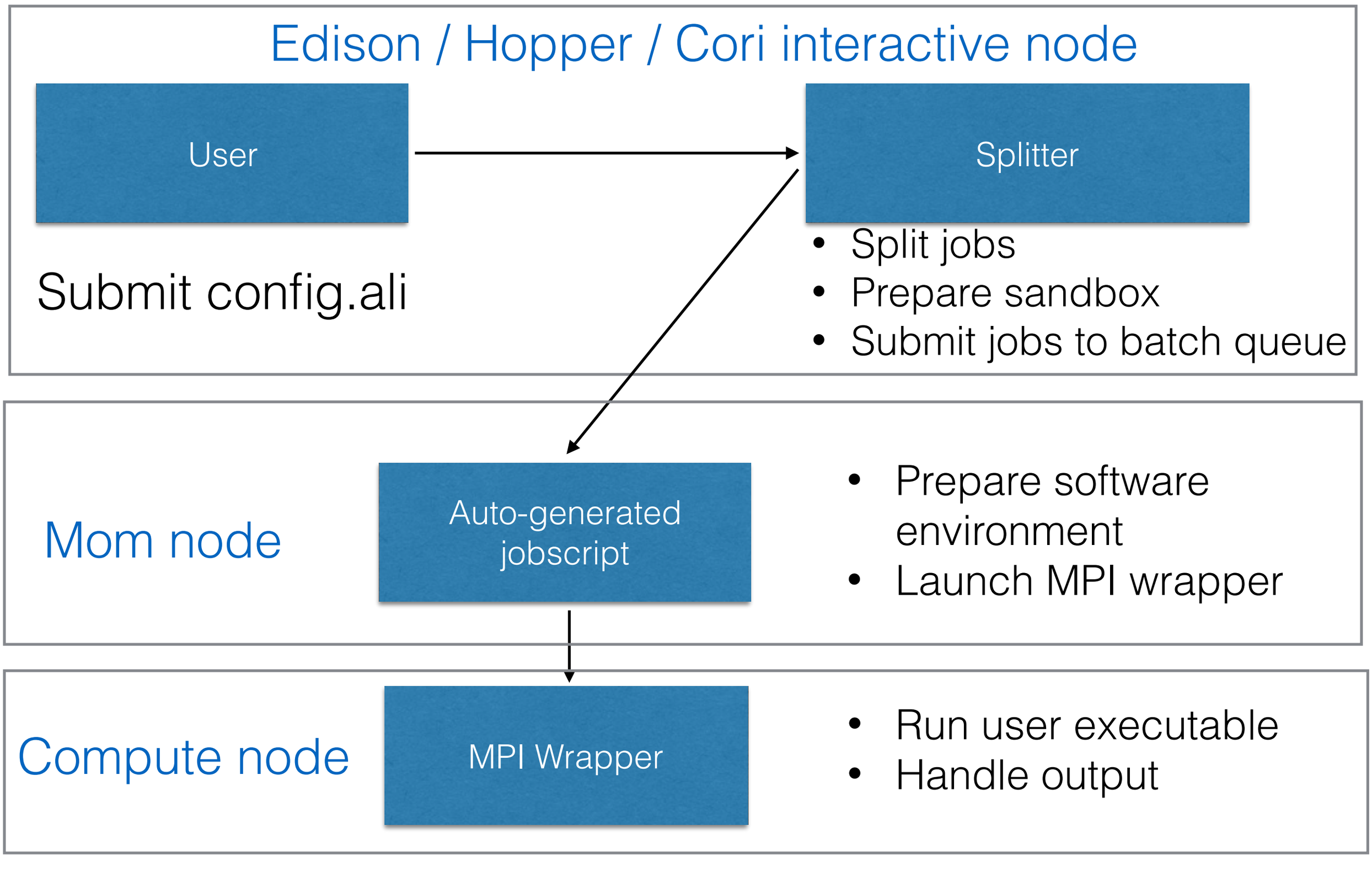
- Acroym: A NERSC ALICE Submitter Agent
- Takes care about:
  - Allocating slots on the HPC system
  - Distribution of payload to the jobs
- Adaptable to other supercomputers

Details:

Language:	Python
License:	BSD
Dependencies:	python >= 2.7 mpi4py

https://bitbucket.org/berkeleylab/analisa

Flow chart


Edison / Hopper / Cori interactive node
User
Splitter
Submit config.ali
• Split jobs
• Prepare sandbox
• Submit jobs to batch queue
Mom node
Auto-generated jobscript
• Prepare software environment
• Launch MPI wrapper
Compute node
MPI Wrapper
• Run user executable
• Handle output

Job submission

- User submits a config file in which payload, queue, I/O is specified
- Syntax: Strongly inspired by the alien JDL syntax
    - Large overlaps

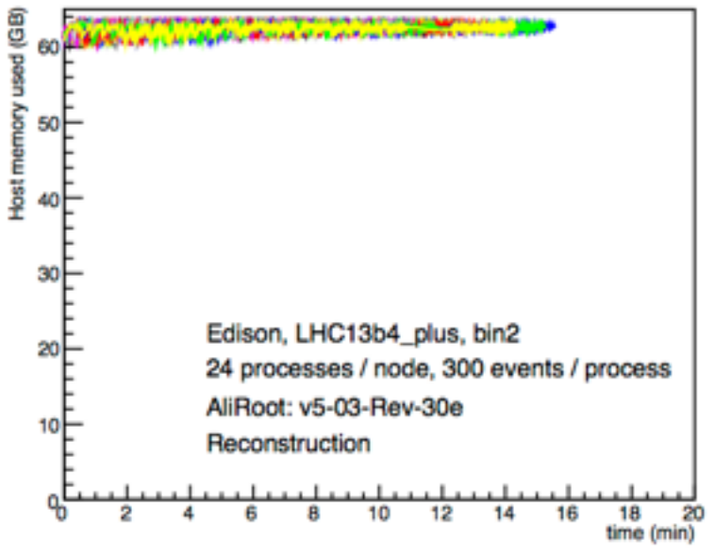
Submission command:

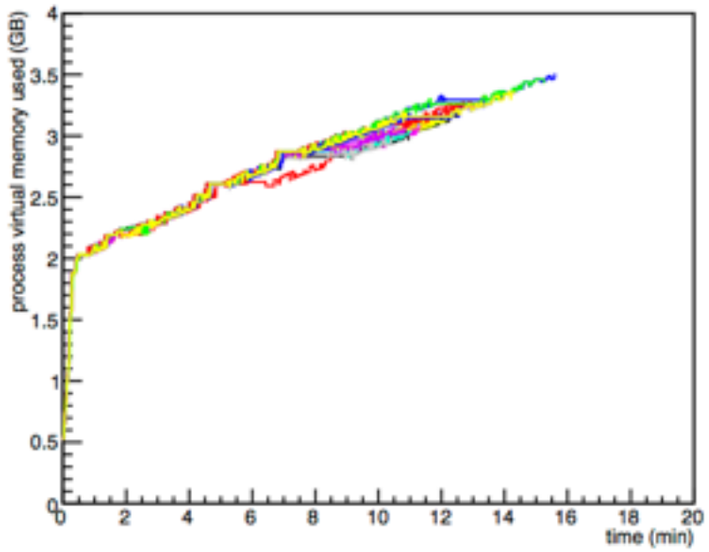
analisa submit -c config.ali

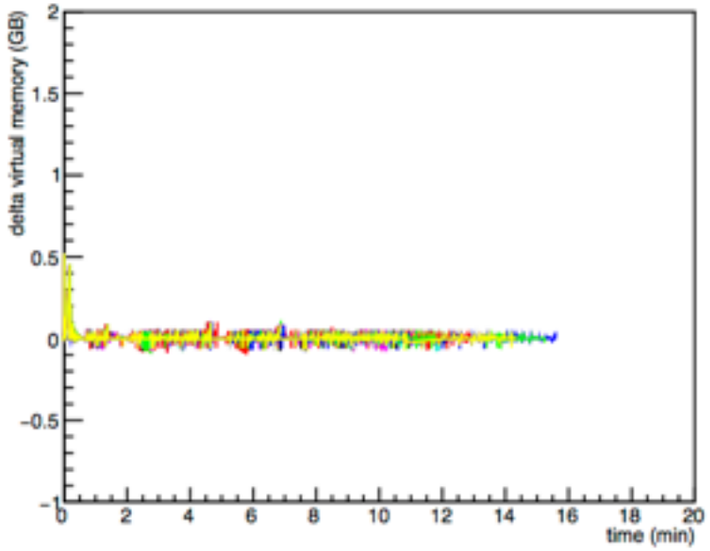
Test (production) done with ALIROOT payload

- Setup:
  - LHC13b4_plus (p-Pb, jet-jet)
  - AliRoot: v5-03-Rev-30 (compiled locally, extracted to scratch fs before job execution)
  - Number of events per job: 300
  - System: Edison

Memory issue in reconstruction







Memory Limits:

- Hopper: 1.5 GB/Core
- Edison: 2.5 GB/Core
- Cori: 4 GB/Core

Attention: All cores share the full memory

Summary

- Simulation jobs within the AliRoot framework have been running on NERSC HPC systems
- Possibly we can also find solutions for other kinds of jobs
- Software deployment possible using cvmfs together with parrot
- Possibly integration as grid side in order to use resources opportunistically.