Update on Cori Integration into the ALICE grid

Markus Fasel

Lawrence Berkeley National Laboratory

ALICE Offline Week,
March 31, 2016

Introduction

Goals

- Utilize resources available on Cori for ALICE
- Integrate Cori into the ALICE Computing infrastructure
- Initial payload: Simulation jobs

Requirements

- Access to payload / executable, output location
- ALICE software stack
- Condition Database

Limitations

- Optimized for parallel jobs
  → Whole-node scheduling
- Limitations in network access
- Job execution time needs to be provided during job submission
- No swap

Tasks

- Translator MPI - serial
- Grid payload assignment to different cores
- Software handling

Reminder: ANALISA

Tool which runs multiple serial jobs as a MPI job

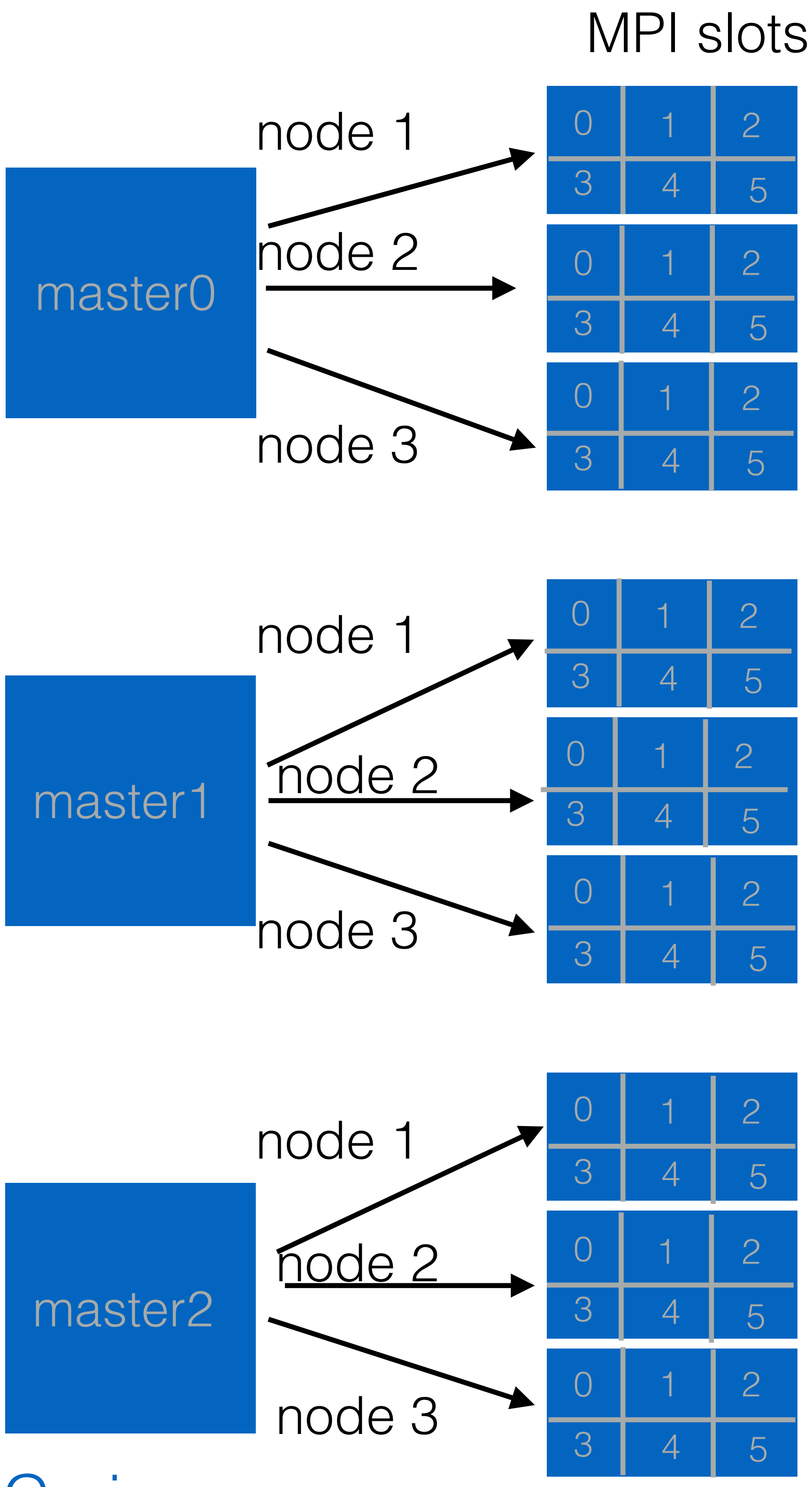
- Submitter:
  - Splits a master into n sub jobs
- Worker (MPI):
  - Runs the subjobs (payload)
- Job description: config, json, xml

Key facts:

- PYTHON, mpi4py
- BSD-type license
- https://bitbucket.org/berkeleylab/analisa

Hiding complexity of resource management for the user

Started on Hopper, running in production on Edison and Cori



cvmfs

cvmfs not directly available on Cori

- Shifter:
  - Docker container with full copy of cvmfs content running on compute node
- Parrot:
  - Tool mounting a copy of the cvmfs file catalogue located on persistent file system under original path

a) Shifter:

- Minimal SLC6 docker container
- 2 Images:
  - Only Software
  - Software + condition database

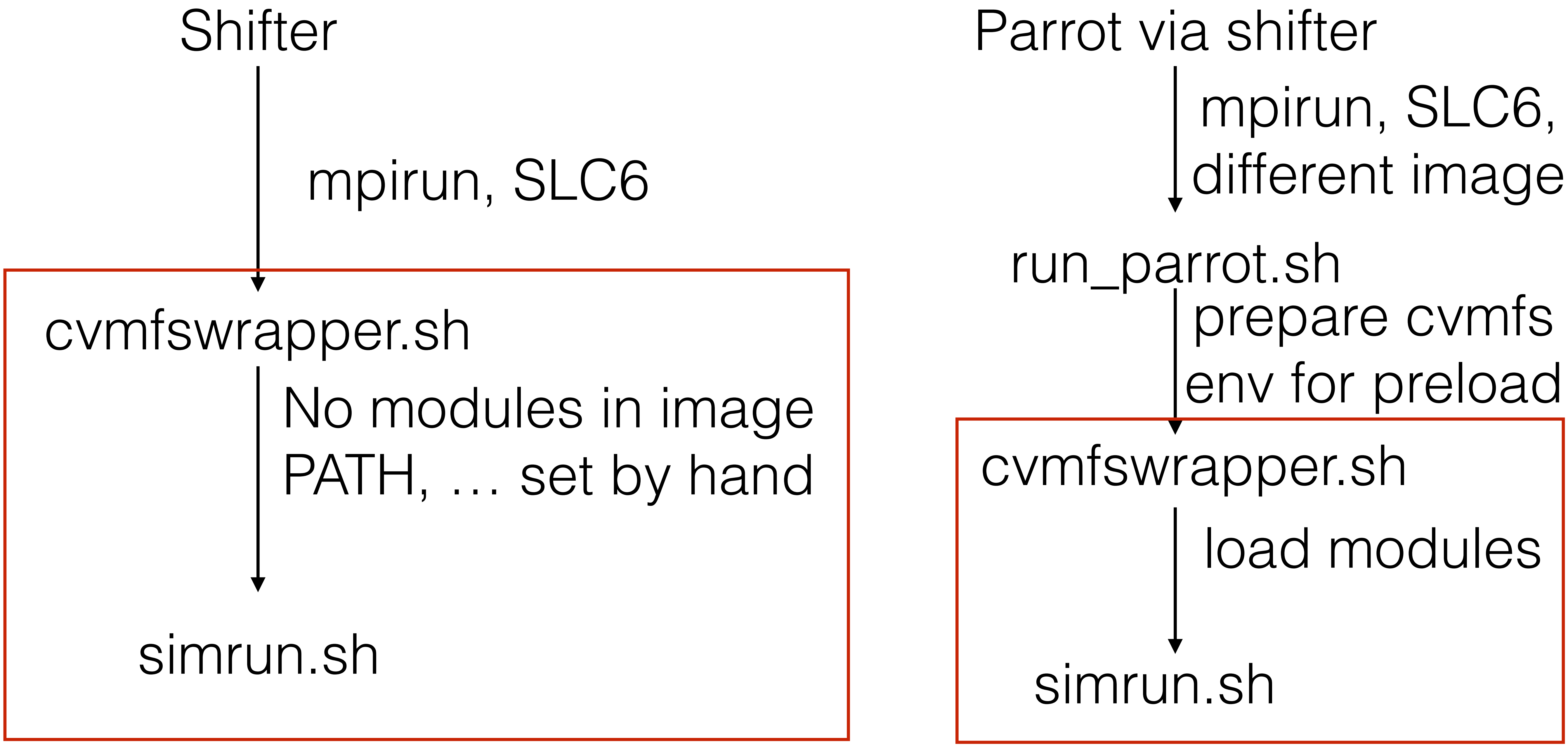
Data (software, condition database) part of the image!

b) Parrot:

Shifter used to provide a native SLC6 from which parrot is run

Data (software, condition database) external!

Shifter workflow



Shifter

mpirun, SLC6

cvmfswrapper.sh

No modules in image
PATH, … set by hand

simrun.sh

Parrot via shifter

mpirun, SLC6,
different image

run_parrot.sh

prepare cvmfs
env for preload

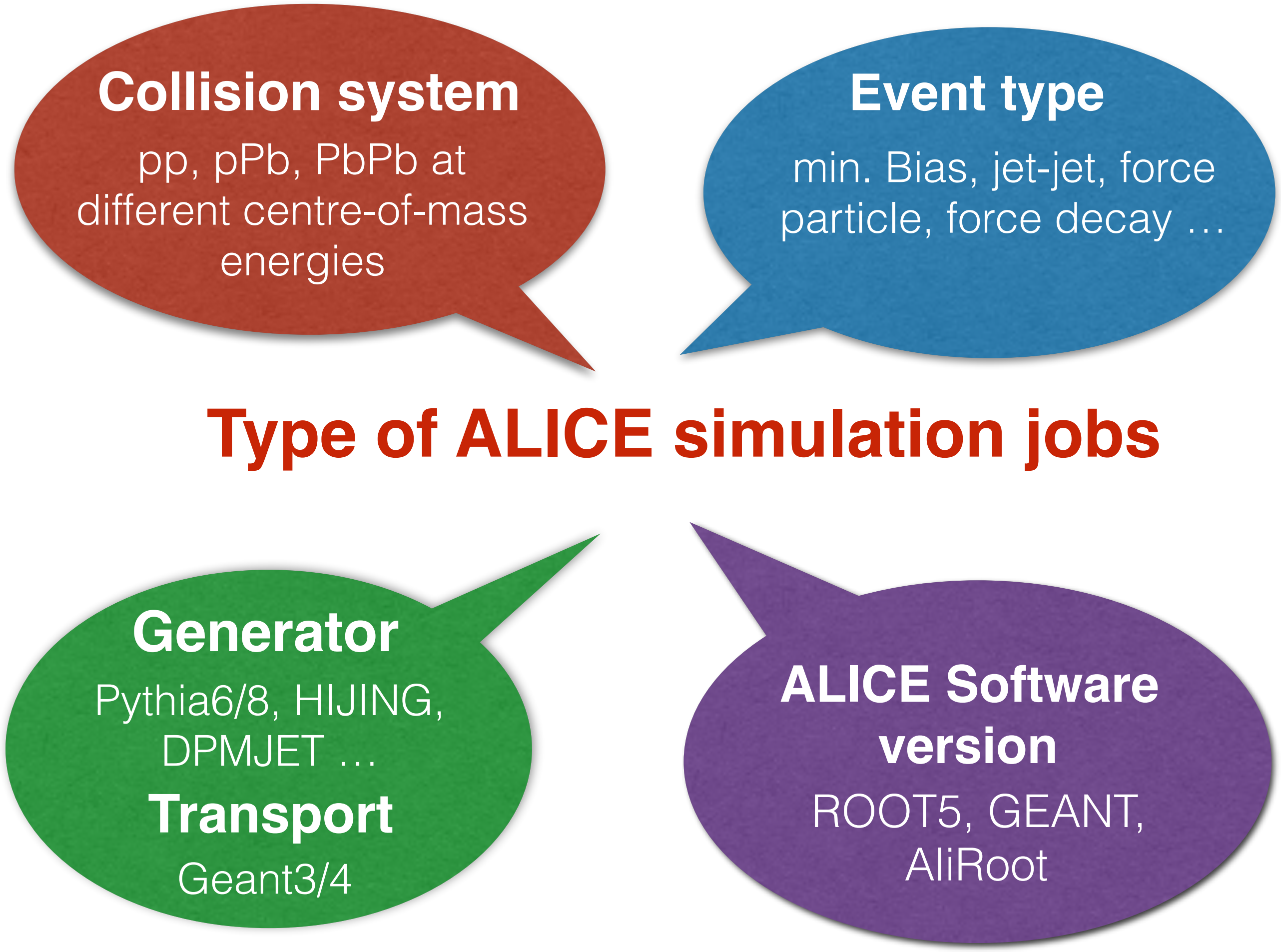
cvmfswrapper.sh

load modules

simrun.sh

Red box: subshell with cvmfs mount

Test of ALICE simulation jobs on NERSC HPC platforms



Collision system

pp, pPb, PbPb at different centre-of-mass energies

Event type

min. Bias, jet-jet, force particle, force decay …

Type of ALICE simulation jobs

Generator

Pythia6/8, HIJING, DPMJET …

Transport

Geant3/4

ALICE Software version

ROOT5, GEANT, AliRoot

Job Parameters:

- Cori:
  - 20 Nodes, 32 jobs / Node
- Edison:
  - 26 Nodes, 24 jobs / Node
- PDSF:
  - 400 jobs / use case

Payload exactly as it runs on the grid!

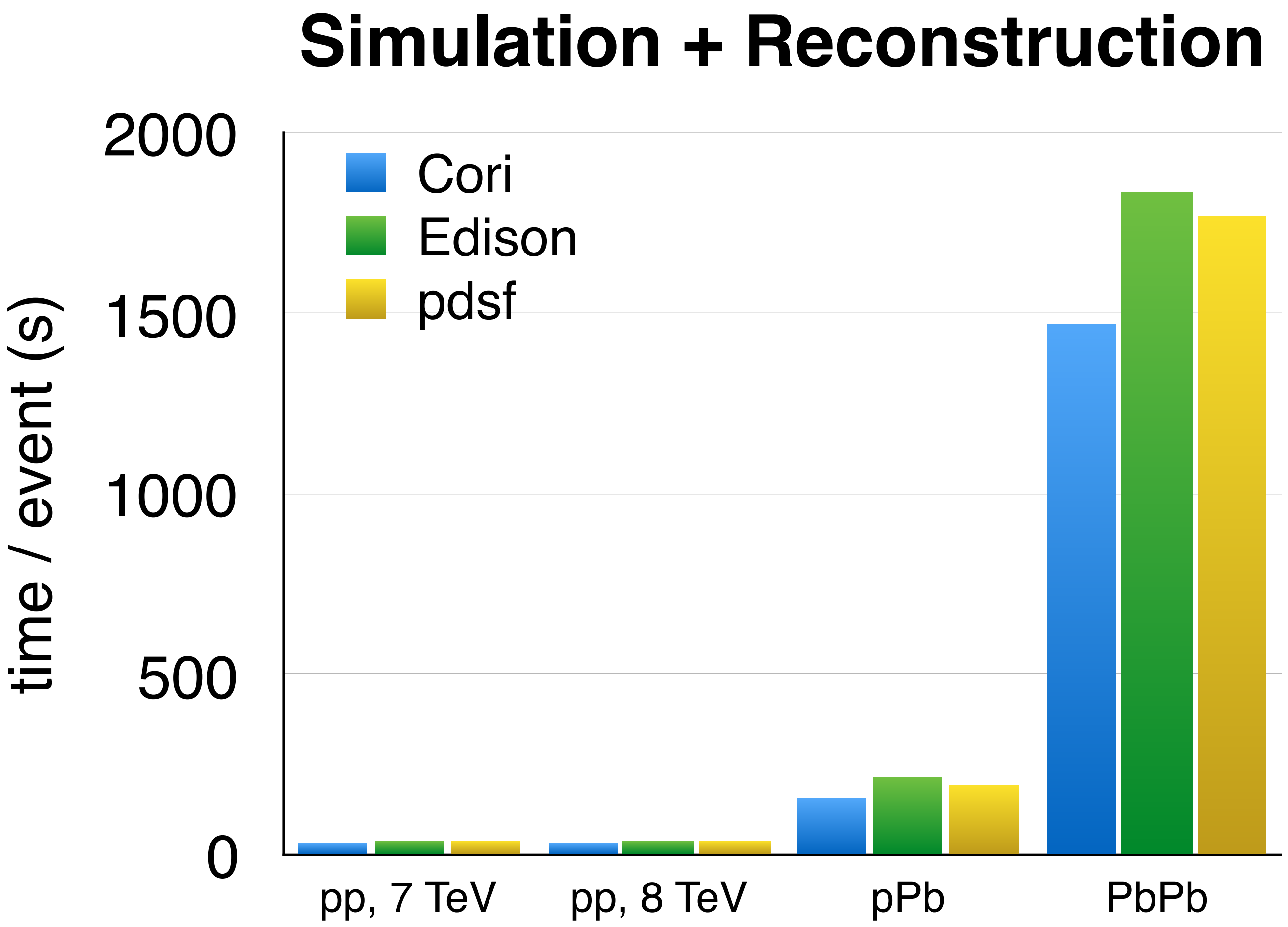
4 Scenarios

- pp, $\sqrt{s}$ = 7 TeV:
  - PYTHIA6
  - Min. Bias
  - Tune Perugia 2011
- pp, $\sqrt{s}$ = 8 TeV:
  - PYTHIA8
  - Min. Bias
  - Tune Monash2013
- p-Pb, $\sqrt{s_{NN}}$ = 5.02 TeV:
  - DPMJET
  - Min. Bias
- Pb-Pb, $\sqrt{s_{NN}}$ = 5.02 TeV:
  - HIJING
  - Min. Bias

All except Pb-Pb: 100 events / job

Pb-Pb: 5 events / Job

Test results

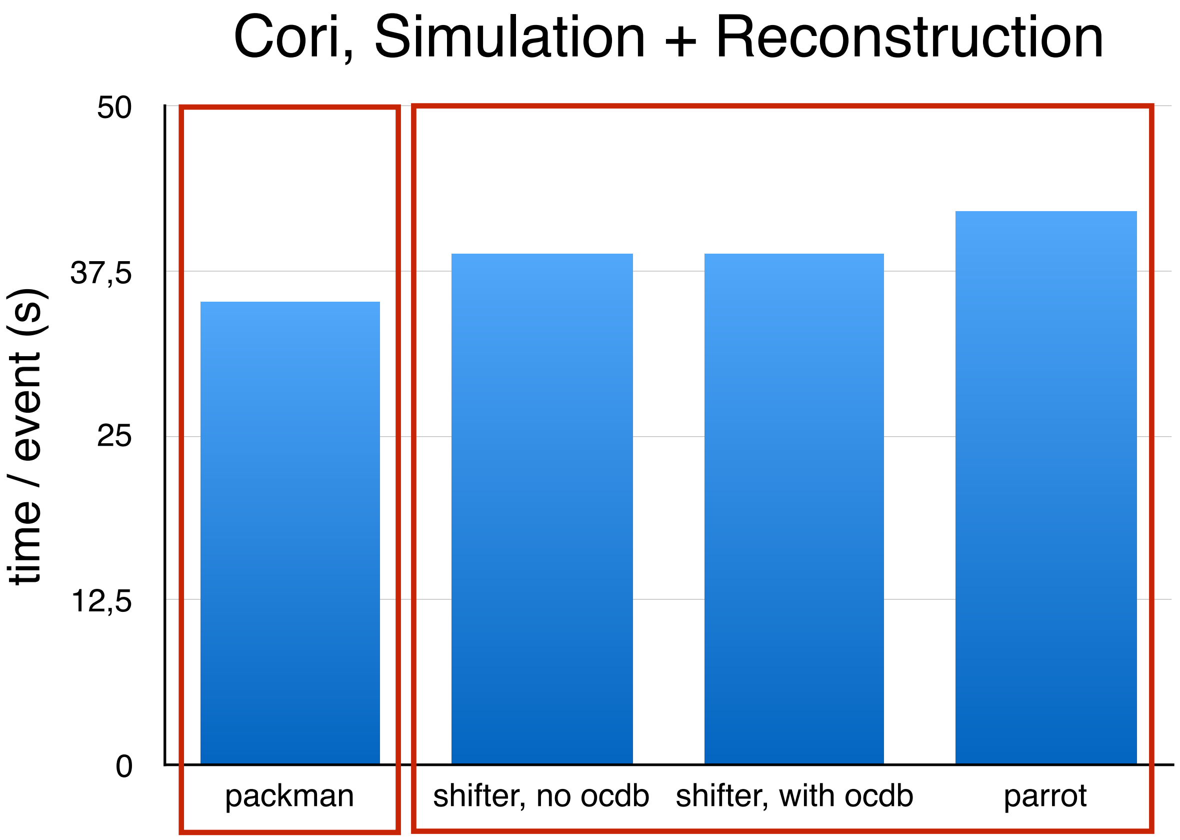


High performance cluster are competitive compared to standard batch farms

PDSF has a mixture of different CPU types

- Same performance to Cori for jobs on same CPU type

cvmfs test



local build system

cvmfs mimicing

First tests show that cvmfs be provided on Cori - optimizations ongoing

pp, $\sqrt{s}$ = 7 TeV Perugia2011 in all cases

Burst buffer

File system for I/O intensive jobs

- Cray Data Warp technology
- SSD based
- 800 GB/s peak I/0
- Size
  - At Phase 1: 750 TB
  - At Phase 2: ~1.5 PB

Ideas / Tests

- Condition Database
- Software stack via preload
- Job sandbox (ongoing)

Planned tests

cvmfs via parrot

- Preload on burst buffer
  - Needs persistent allocation on the burst buffer
- Local squid instead of preload
- Stratum-1 at Fermilab instead of preload

Cori has network access, but limited

Summary

- Tool ANALISA submitting multiple serial jobs as MPI job
  - Demonstrating capabilities to run ALICE simulation jobs on Cori
- Several methods for cvmfs on Cori available
  - More (natural) ways to be tested
- Further integration ongoing
  - Usage of the Burst Buffer
  - Running of the grid pilot
  - …