STAR Data Reconstruction at NERSC/Cori

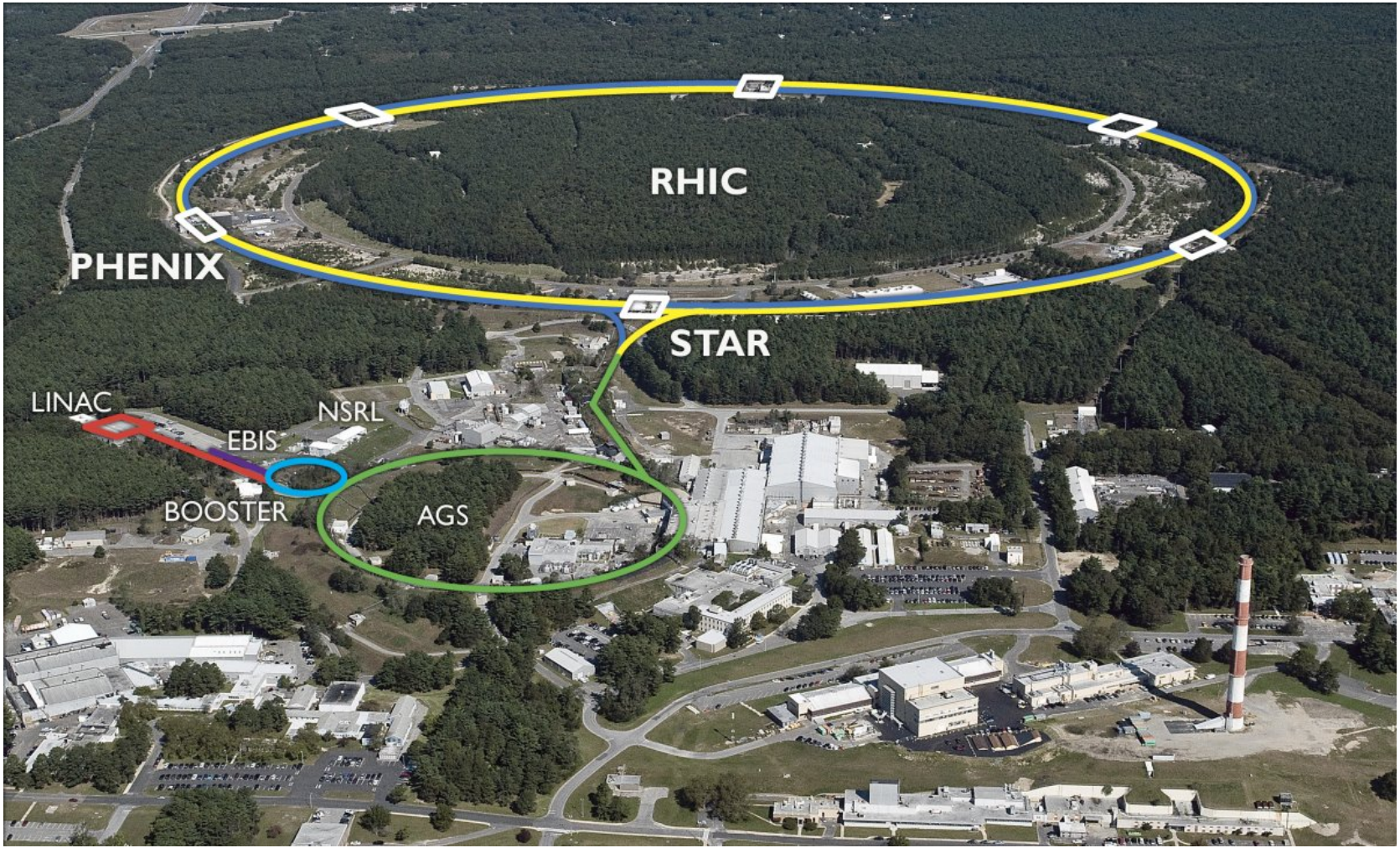
an adaptable Docker container approach for HPC

Mustafa Mustafa, Jeff Porter, Jan Balewski
Lawrence Berkeley National Laboratory

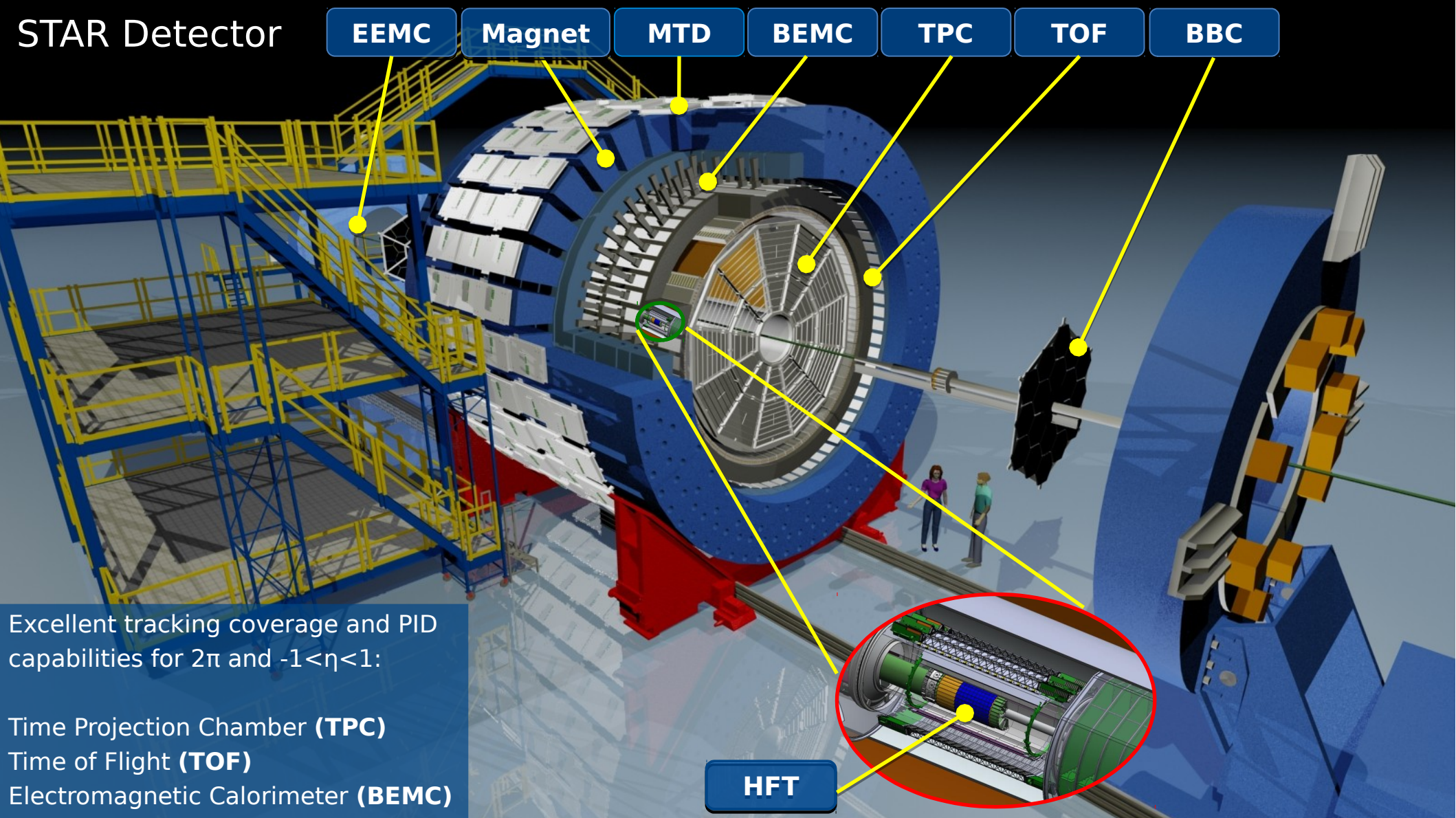
Jérôme Lauret
Brookhaven National Laboratory

2017 March 7th
Open Science Grid All Hands Meeting – La Jolla, CA

Relativistic Heavy Ion Collider – Mass Production of Quark Gluon Plasma



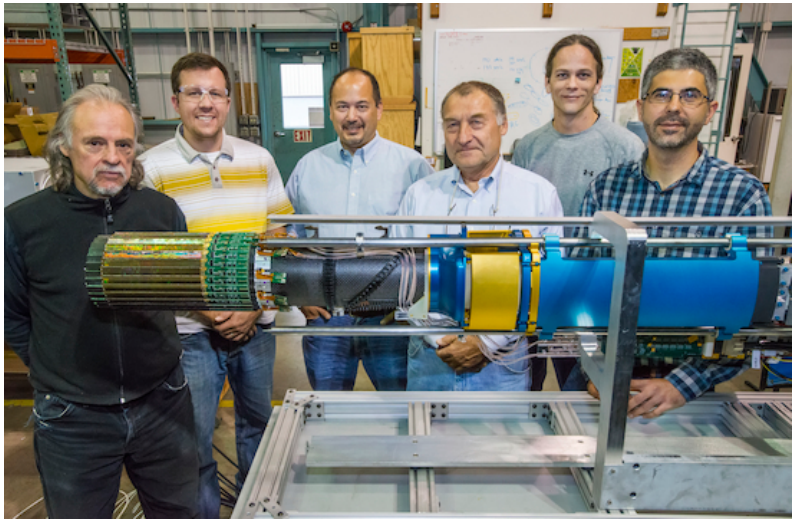
STAR Detector



Excellent tracking coverage and PID capabilities for 2π and -1<η<1:

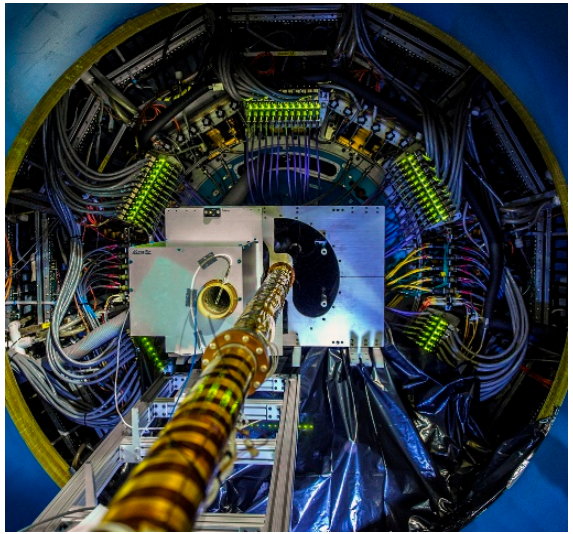
Time Projection Chamber **(TPC)**
Time of Flight **(TOF)**
Electromagnetic Calorimeter **(BEMC)**

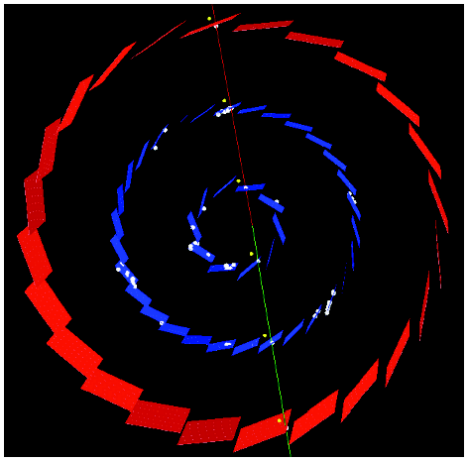
Precision Measurement of Charm Quarks Production – Heavy Flavor Tracker



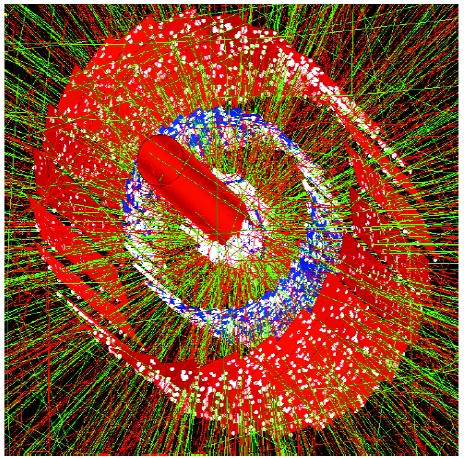
Heavy Flavor Tracker

- State of the art Pixel sensors technology
- First time in heavy-ion collisions
- Fully installed at STAR in year 2014



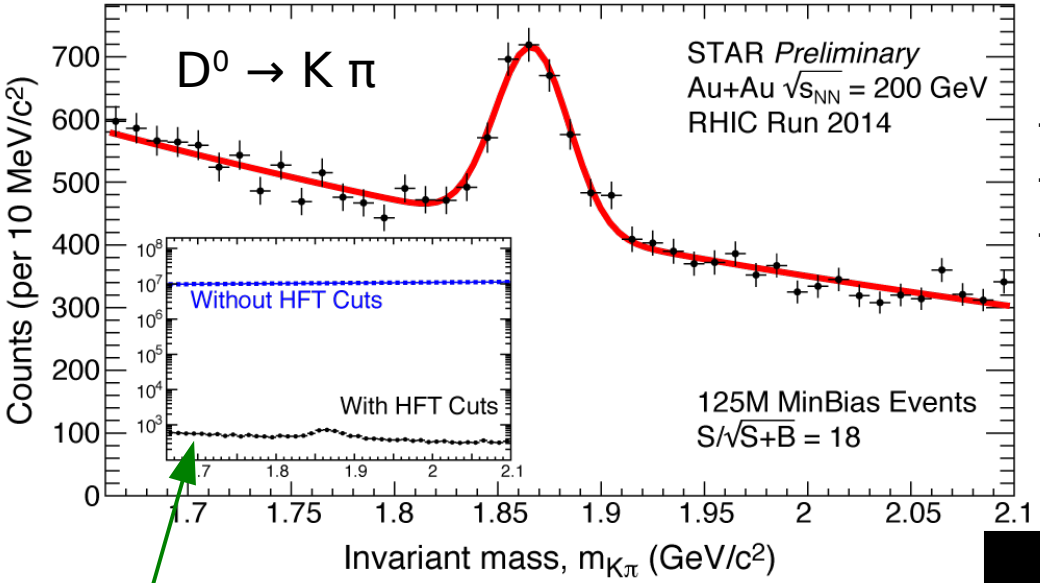


Cosmic ray event



Au+Au event

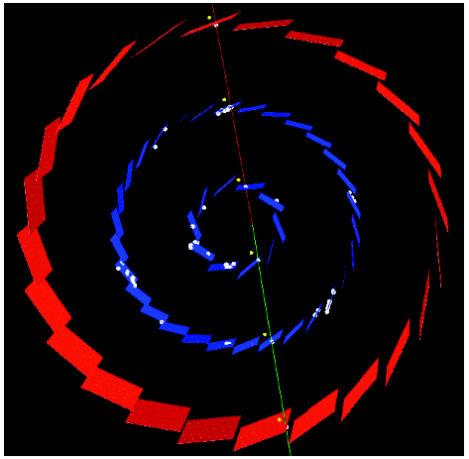
Precision Measurement of Charm Quarks Production – Heavy Flavor Tracker



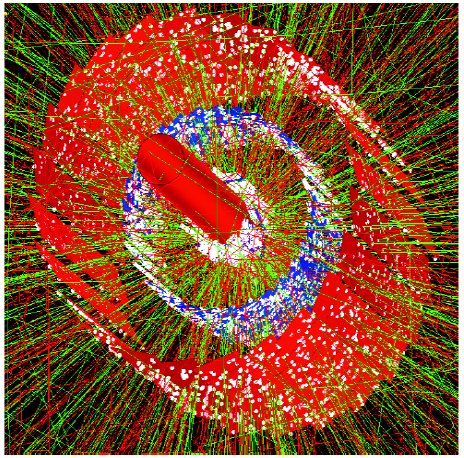
Heavy Flavor Tracker

- State of the art Pixel sensors technology
- First time in heavy-ion collisions
- Fully installed at STAR in year 2014

+4 orders of magnitude reduction in combinatorial background



Cosmic ray event



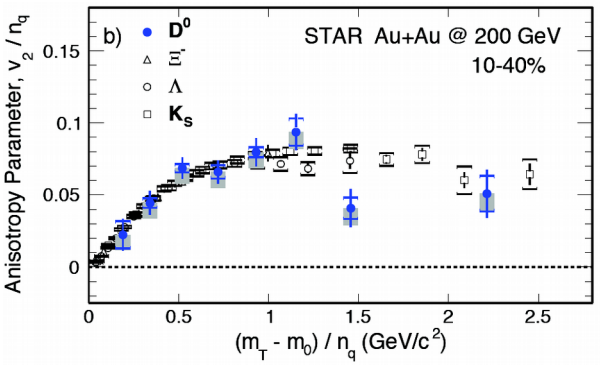
Au+Au event

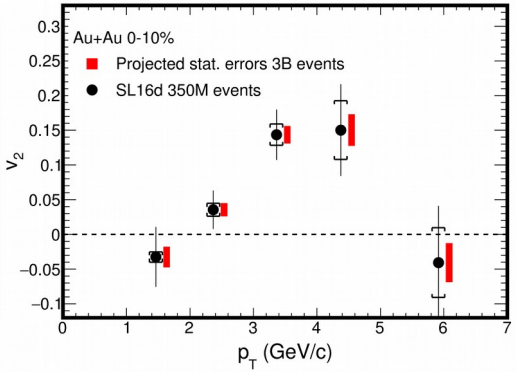
HFT Datasets and Timeline

HFT Au+Au data production and resource needed:

- Au+Au 200 GeV Run 2014: 1B minimum-bias events
  → ~1.2 Pb & 30M CPU hours → ~3 RCF months

- Au+Au 200 GeV Run 2016: 2B minimum-bias events
  → ~3 Pb & 50M CPU hours → ~6 RCF months





Measurements publications time-line:

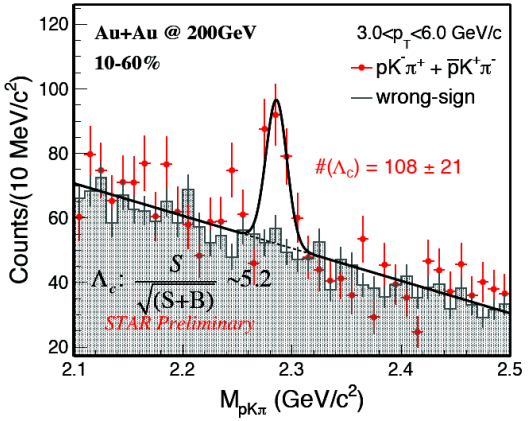
→ D^0 v_2 – Submitted to PRL - **Run14 data**

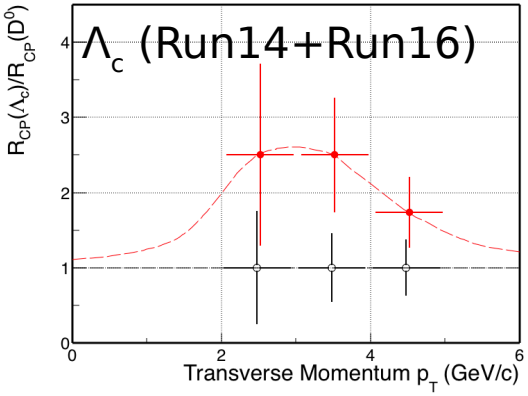
→ D^0 production and energy loss – **Run14 data**

→ D^0 centrality dependent v_2 / v_3 - **Run14+Run16**

→ Λ_c cross-section, **Run14+Run16**

→ Many more measurements …

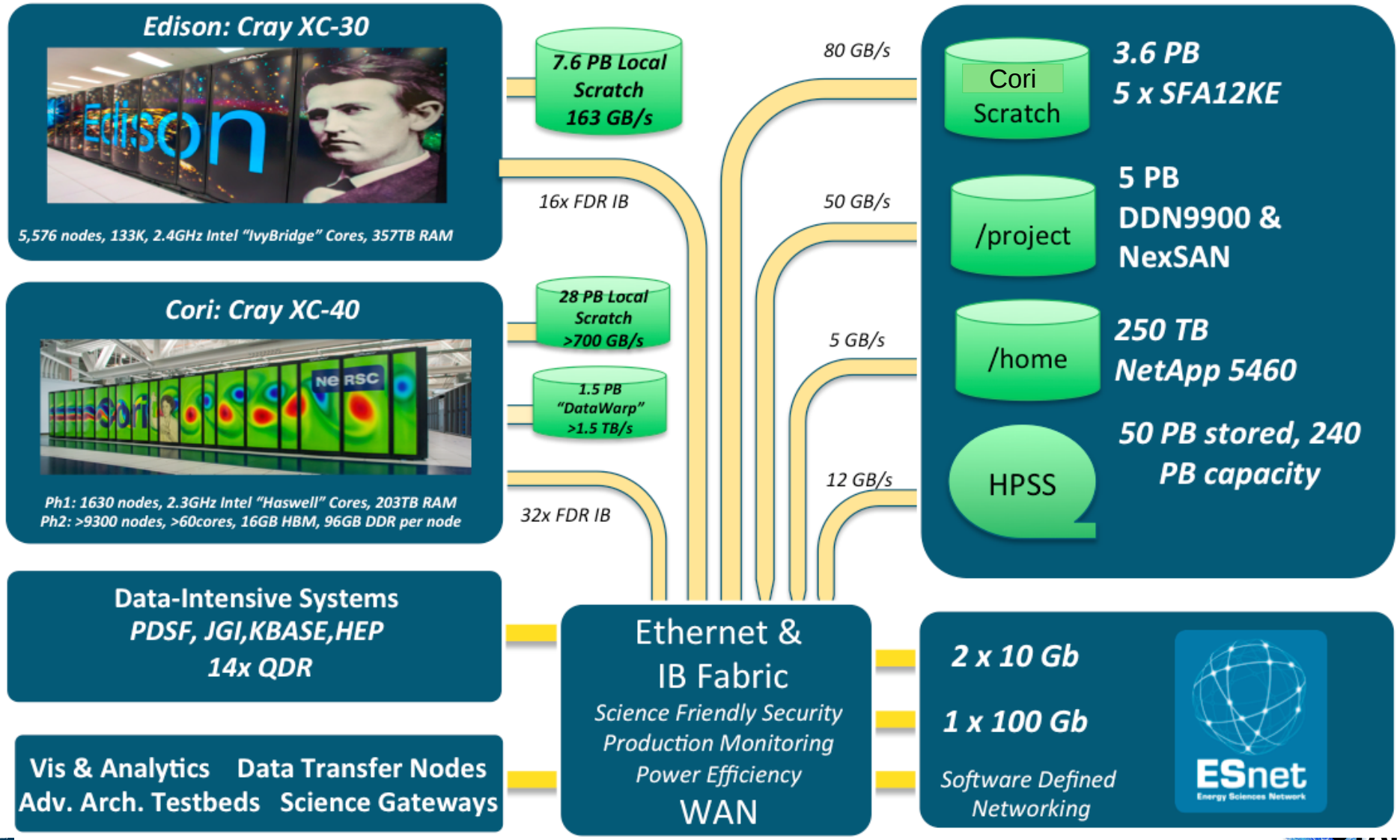




STAR proposal to NERSC:

To leverage Cori & ESnet resources to process ~50% of year 2016 dataset.

Cori is NERSC's newest supercomputer system (NERSC-8). It is named after American biochemist Gerty Cori.


Edison: Cray XC-30
Edison
5,576 nodes, 133K, 2.4GHz Intel "IvyBridge" Cores, 357TB RAM
7.6 PB Local Scratch 163 GB/s
16x FDR IB
Cori: Cray XC-40
NERSC
Cori
Ph1: 1630 nodes, 2.3GHz Intel "Haswell" Cores, 203TB RAM
Ph2: >9300 nodes, >60cores, 16GB HBM, 96GB DDR per node
28 PB Local Scratch >700 GB/s
1.5 PB "DataWarp" >1.5 TB/s
32x FDR IB
80 GB/s
50 GB/s
5 GB/s
12 GB/s
Cori Scratch
3.6 PB
5 x SFA12KE
/project
5 PB
DDN9900 & NexSAN
/home
250 TB
NetApp 5460
HPSS
50 PB stored, 240 PB capacity
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
ESnet
Energy Sciences Network
BERKELEY

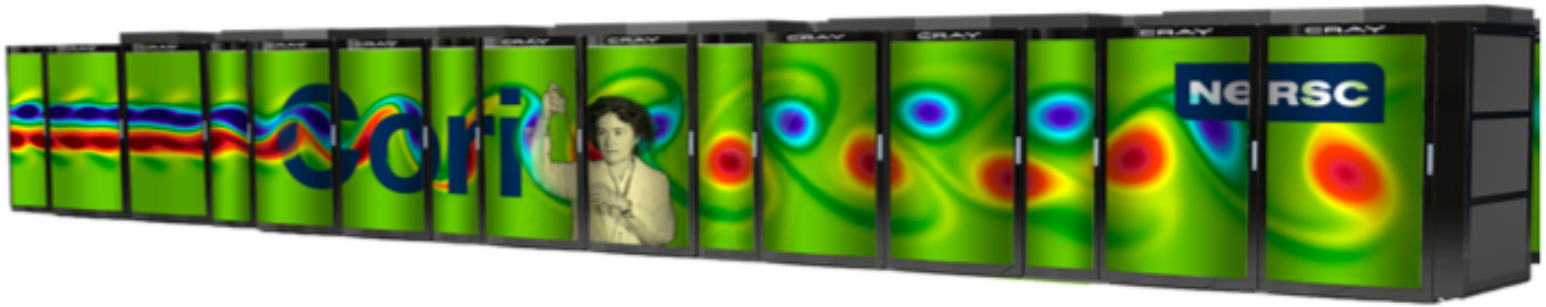
Cori @ NERSC

- **Phase I system architecture:**
  - 2,004 compute nodes
  - Two 2.3 GHz 16-core Haswell processors per node
  - 128 GB of memory per node → 4GB / core
  - Lustre file system with 30 PB of disk and > 700 GB/second I/O bandwidth

- **Phase II system architecture:**
  - Over 9,300 single-socket nodes in the system (Next-generation Intel® Xeon Phi™ Knights Landing (KNL))
  - 68 cores per node with support for four hardware threads
  - 96 GB DDR4 memory per node using 6 channels → 1.4 GB / core

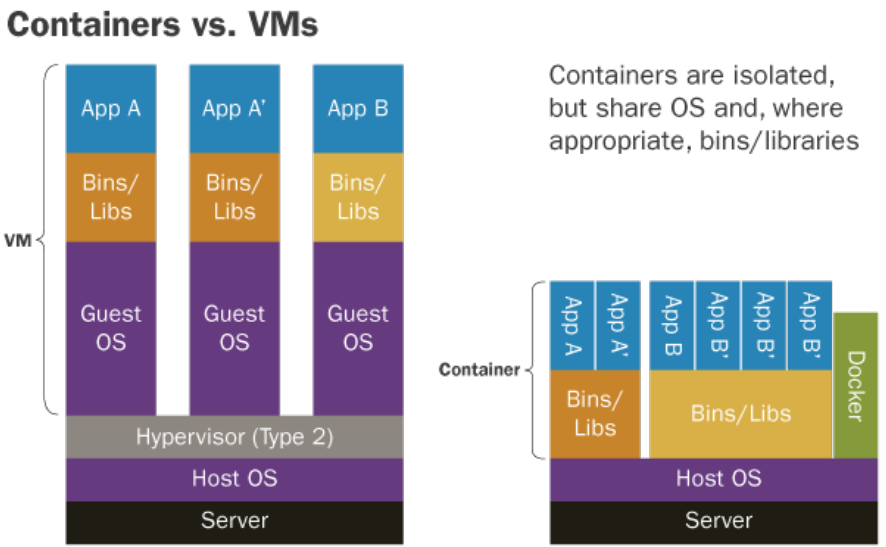
Cori Phase I – Data Features

- NVRAM Flash Burst Buffer as I/O accelerator
- High bandwidth external connectivity to experimental facilities from compute nodes
- More login nodes for managing advanced workflows
- Support for real time and high-throughput queues with SLURM
- Virtualization capabilities with Shifter



Linux Containers - Docker

HPC systems provide vast resources for computation and data intensive applications. However, for technical and logistic reasons, their software is not easily customizable for specific project needs.



With a Docker enabled system we can push STAR software/environment with the job to the computing node.

Shifter: containers for HPC

Challenge and Opportunity

- Data Intensive computing often require large, complex software stacks that are difficult to support in HPC.
- Docker is rapidly becoming a new way to package and run applications.

Innovation

- Shifter is a NERSC R&D effort, in collaboration with Cray, to support User-created Application images.
- Shifter provides “Docker-like” functionality for HPC

Impact and Early Successes

- Shifter has already enabled multiple projects to quickly make use of NERSC (e.g. STAR, LCLS, LHC)
- Shifter can improve job-startup times and application performance
- Shifter will be supported by Cray and is already being evaluated by other HPC centers
- *Submitted application for R&D 100 Award*

The Register®
Biting the hand that feeds IT

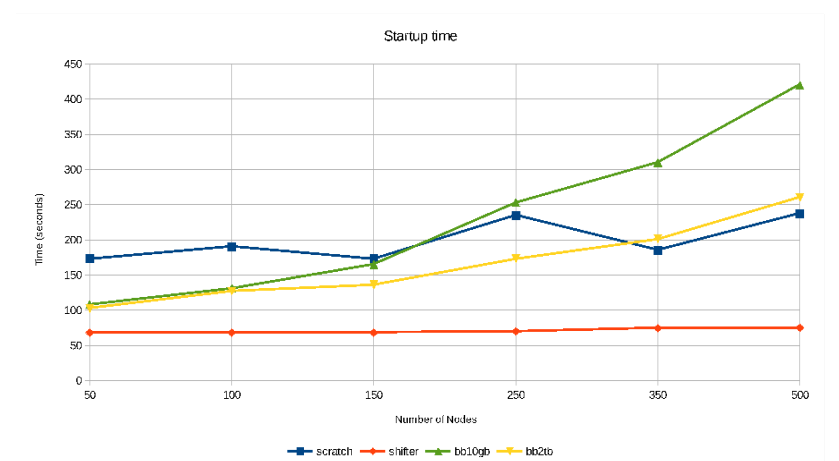
DATA CENTER SOFTWARE NETWORKS SECURITY INFRASTRUCTURE DEVOPS BUSINESS HARDWAR

Data Center ▸ HPC

Cray hoists Docker containers into supercomputers

Productivity gains without performance hits

18 Nov 2015 at 00:01, Drew Cullen 25 35



STAR Docker Image

To allow for layers reuse we break the build of the environment into 5 layers:

- STAR-Base-Dockerfile
  Login scripts + RPMs + CERN library

- STAR-OPT-Dockerfile → 3rd party software and libraries, rarely modified (mostly when OS changes)
- STAR-ROOT-Dockerfile → CERN/ROOT version used by STAR (changes are seldom)
- STAR-SL-Dockerfile → STAR Software Library (the only layer that changes a few times a year)
- STAR-PDSF-Dockerfile → mpi + mysql-server + pipeline relevant macros

```
> sudo docker images
REPOSITORY                    TAG                 IMAGE ID            CREATED             SIZE
mmustafa/sl64_sl16c           v1_pdsf             fda69d6f00a3        41 hours ago        13.34 GB
sl64_sl16c                    v1                  241fdbde9d25        3 days ago          13.34 GB
sl64_root                     v1                  c290d8178cca        3 days ago          7.19 GB
sl64_opt                      v1                  36d2d0db307c        9 days ago          4.553 GB
sl64_base                     v1                  f5239e6a0e70        9 days ago          1.243 GB
ringo/scientific              6.4                 b0d9e5931f6d        22 months ago       322.1 MB
```

Planning to optimize docker image size in the upcoming weeks

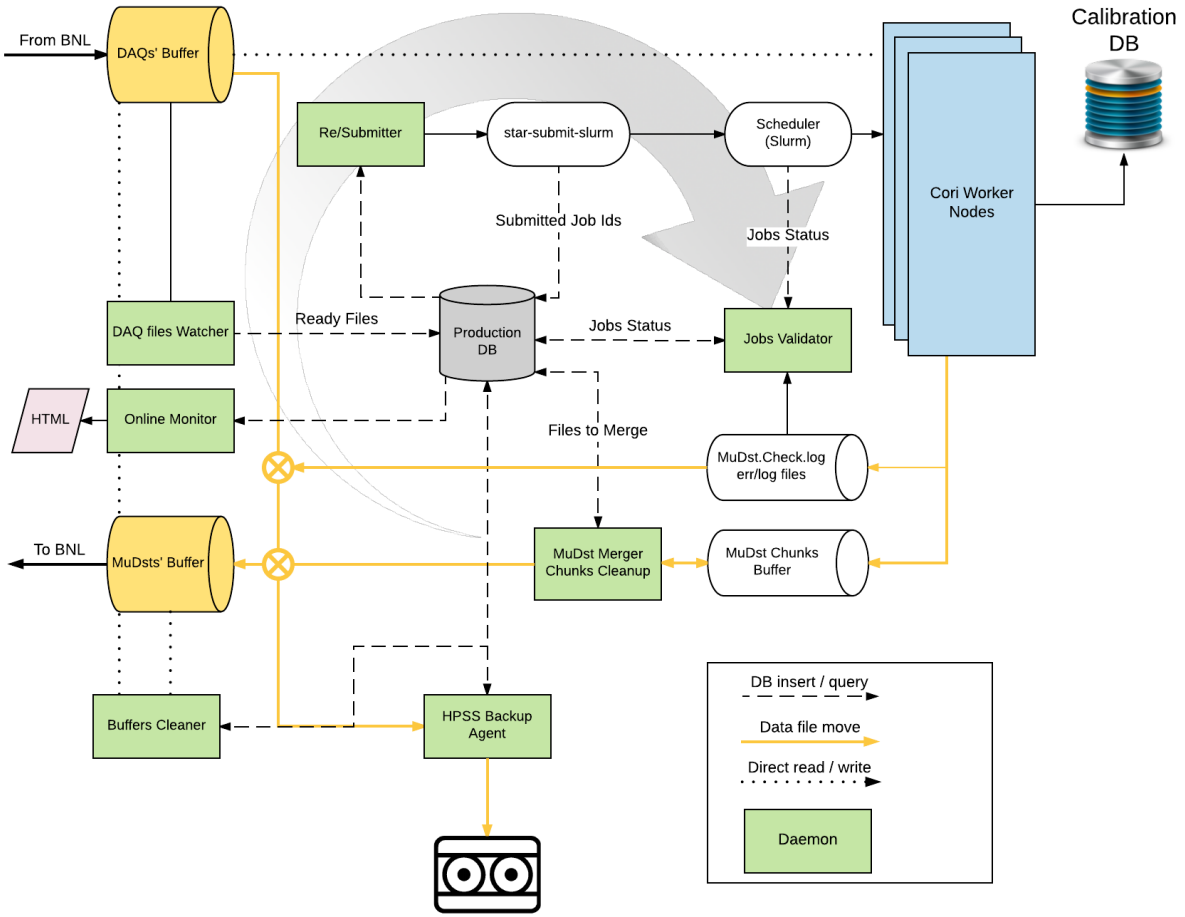
For example, current configuration has:

- OPT and ROOT: follow symbolic links

Data Production Pipeline

- Input/output data transfer on ESnet

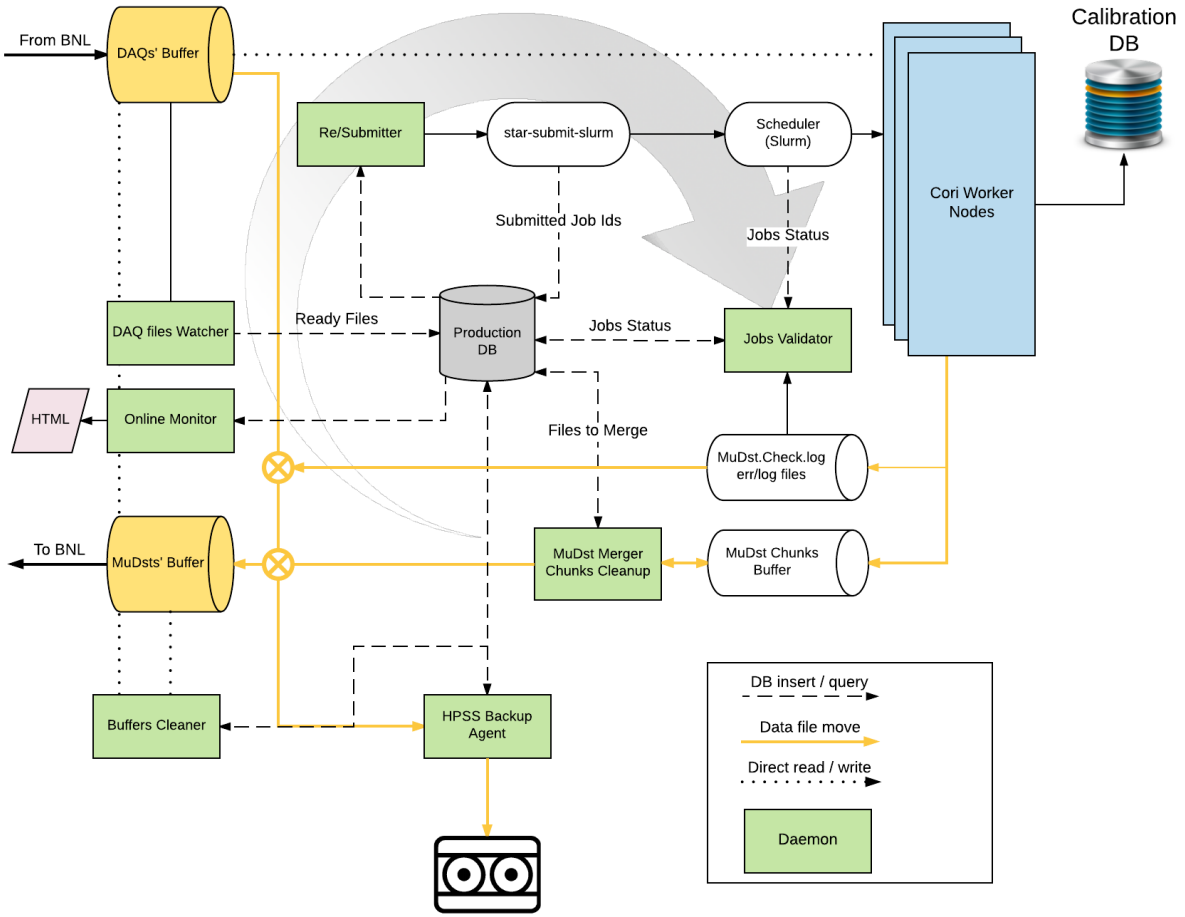
  ~100Tb / 10k cores / week
- Automated finite state workflow (a simple map-reduce pattern)
- MongoDb
  - ➔ communicate/control system states
  - Provides persistent storage of pipeline states (all daemons are stateless)
- Fault-tolerance is a must in HPC environment. Need to be ready for all Cori downtime scenarios
  - ➔ Jobs cold start capability
  - ➔ daemons die/resume capability
- Continuous online pipeline monitoring
- Archive input and output files to HPSS



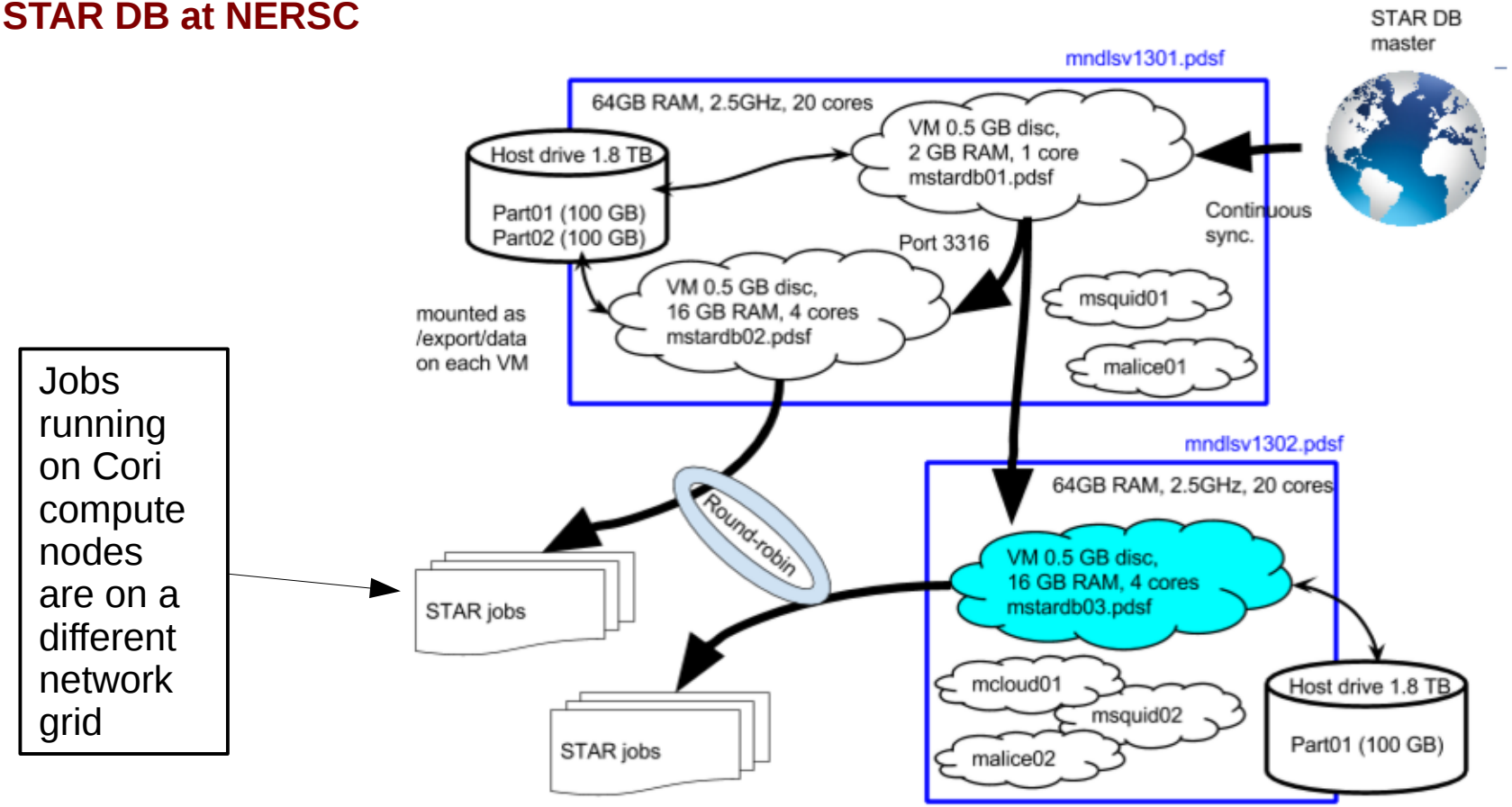
Data Production Pipeline

- Workflow pipline
  - 12 daemon threads

    (6 heartbeat daemons)
  - Jobs resources utilization gathered from slurm and log files to allow monitoring and further tunning
  - Slurm for job submission and states (no heartbeat from jobs)
  - All daemons are configurable using json files
- Monitoring:
  - Python Flask app running on portal-auth.ners.gov responds to web-base users queries to MongoDB
  - JavaScript runs in browser end renders numbers into either tables or graphs



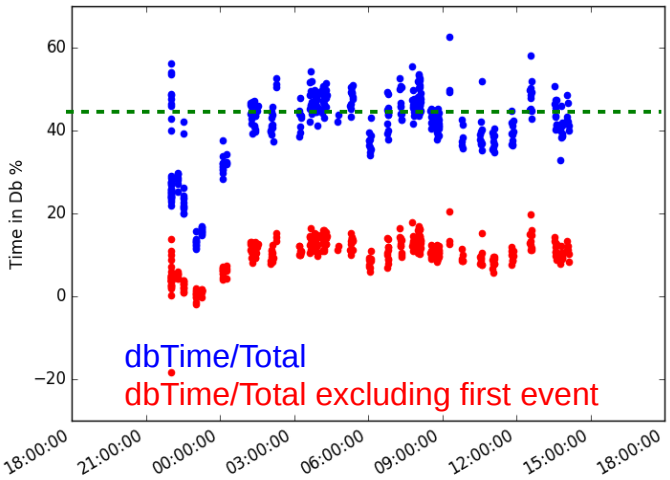
STAR DB at NERSC



Configuration of hardware and software for STAR DB mirror at PDSF:

- Master continuous syncing with BNL.
- Two slave VMs serving jobs. 4 cores, 16GB RAM per slave.

Database access from Cori – The challenge of a network isolated DB server "A bad event instance"



← ~45% of production time in DB

- First event takes more than 30-60 minutes while DB Loads/Procs ~ 160 per VM
- Some events take ~3 hrs
- A few events never managed to connect to DB within 3 hours

Conclusions:

1) Network conditions are unpredictable
2) Very challenging to scale to our target throughput

Local MySQL Server at Compute Nodes

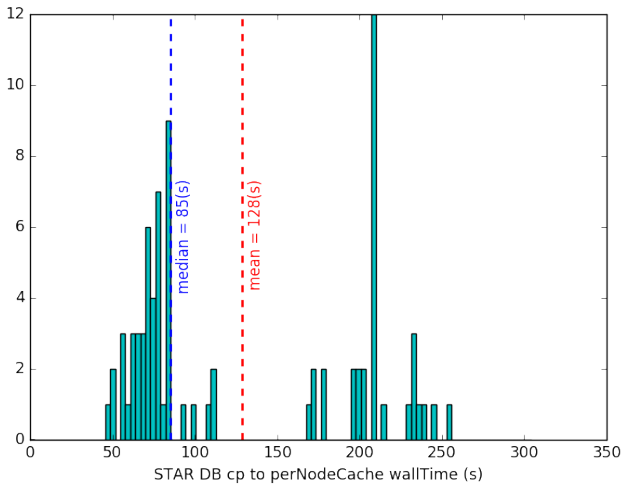
- Pros:
  - No network routing
  - Straightforward scaling with number of jobs
- Cons:
  - Have to start the server from scratch for every job
    - Lustre filesystem is not ideal for transactional load → some tables take a long time to cache (a particular table makes ~30k queries)

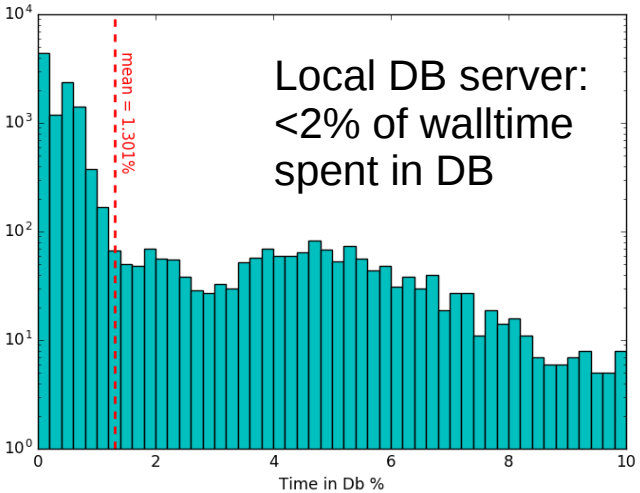
Shifter perNodeCache proves to be an ideal solution to this problem. It creates a loopback mounted XFS file on Lustre with metadata residing on the local node memory. DB caching dramatically drops from +30 minutes to a few seconds. The overhead is the copy time of the DB load to the XFS file.

Local MySQL Server at Compute Nodes

- Pros:
  - No network routing
  - Straightforward scaling with number of jobs
- Cons:
  - Have to start the server from scratch for every job
    - Lustre filesystem is not ideal for transactional load → some tables take a long time to cache (a particular table makes ~30k queries)

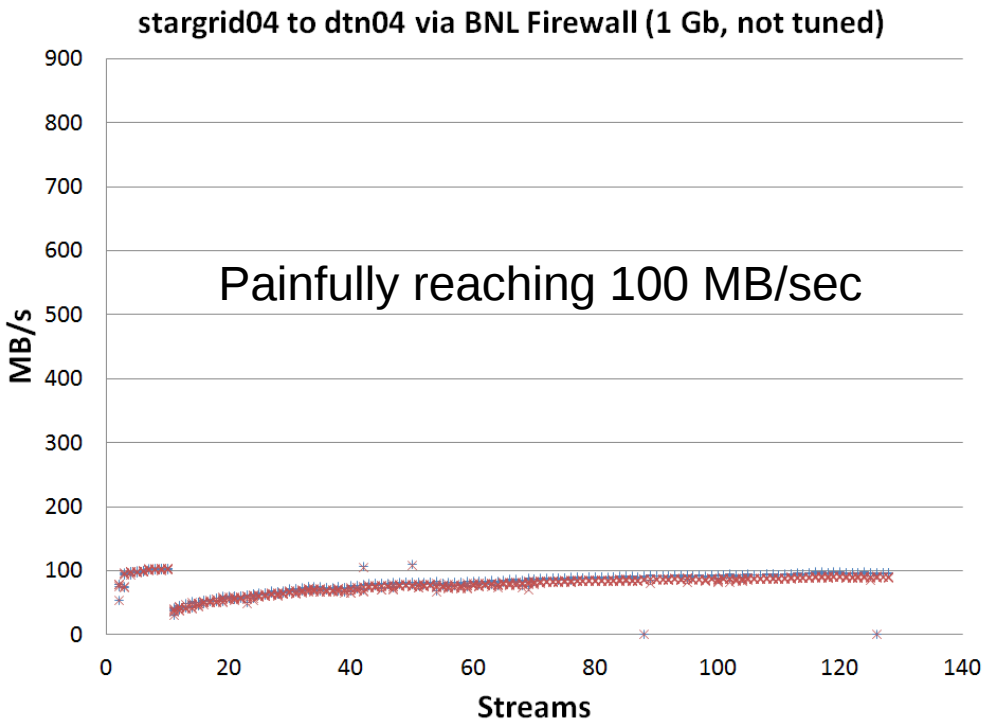
Shifter perNodeCache proves to be an ideal solution to this problem. It creates a loopback mounted XFS file on Lustre with metadata residing on the local node memory. DB caching dramatically drops from +30 minutes to a few seconds. The overhead is the copy time of the DB load to the XFS file.



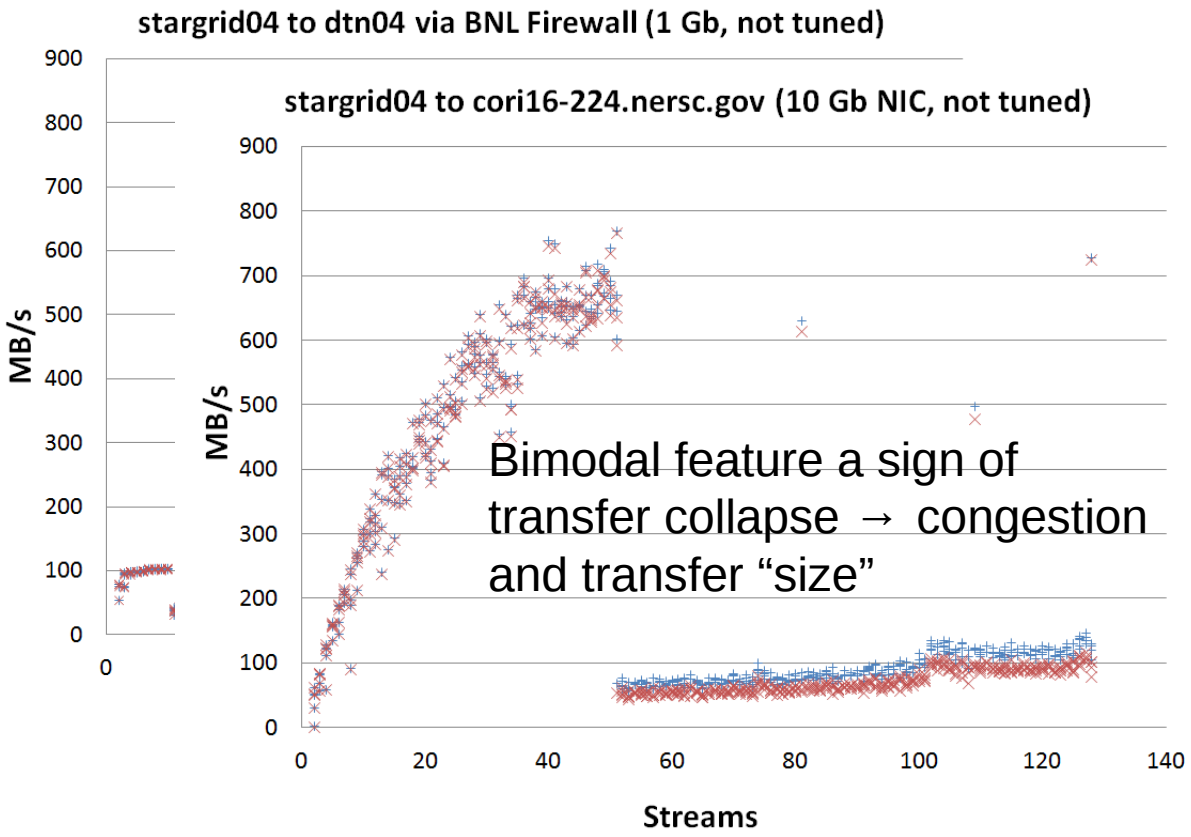


Problem solved! Virtually scalable solution to any number of jobs.

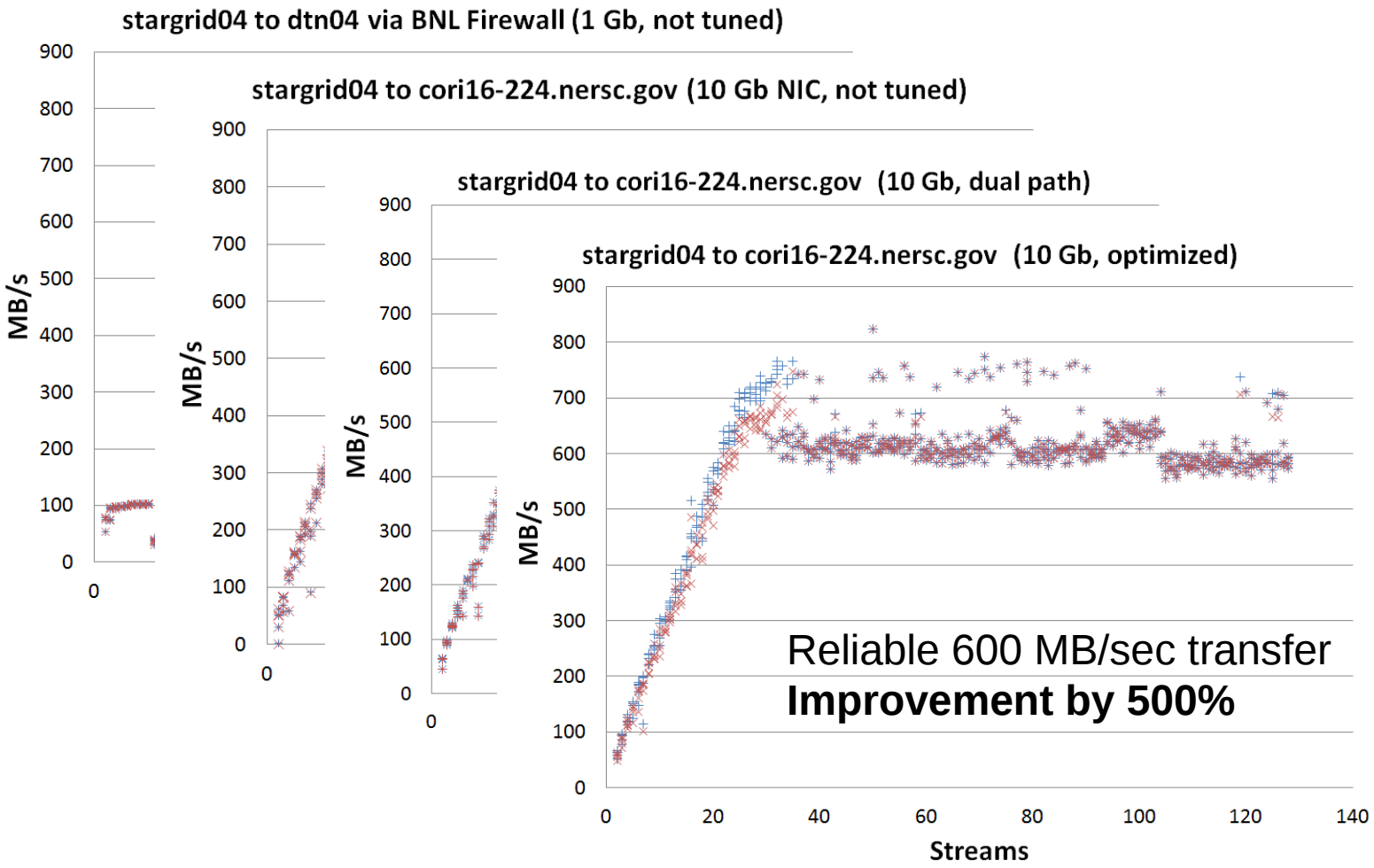
BNL – NERSC Data Transfer Optimization



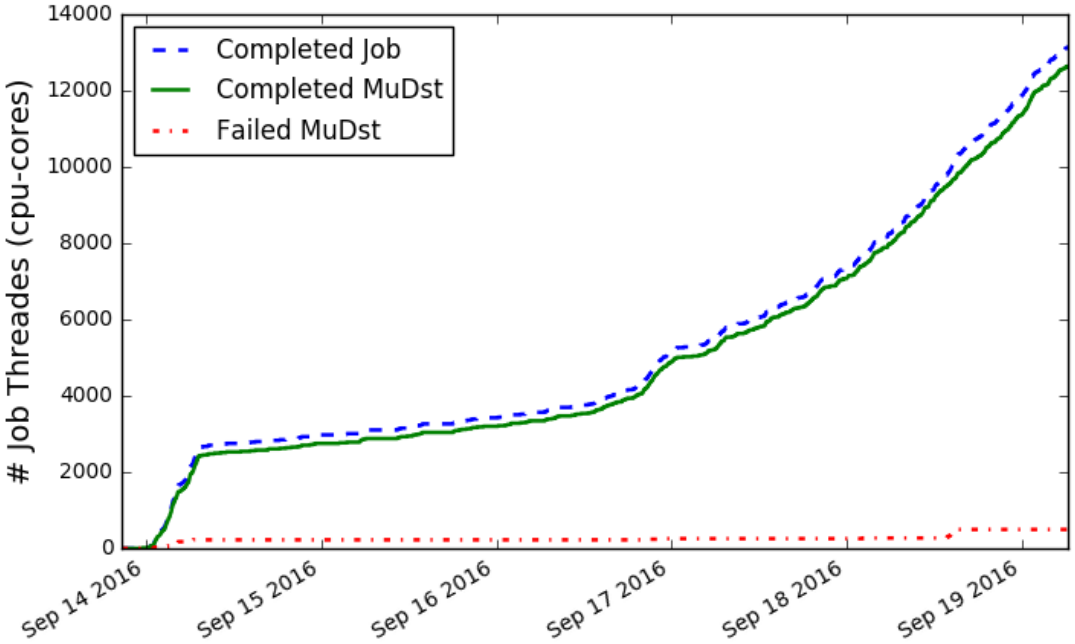
BNL – NERSC Data Transfer Optimization



BNL – NERSC Data Transfer Optimization



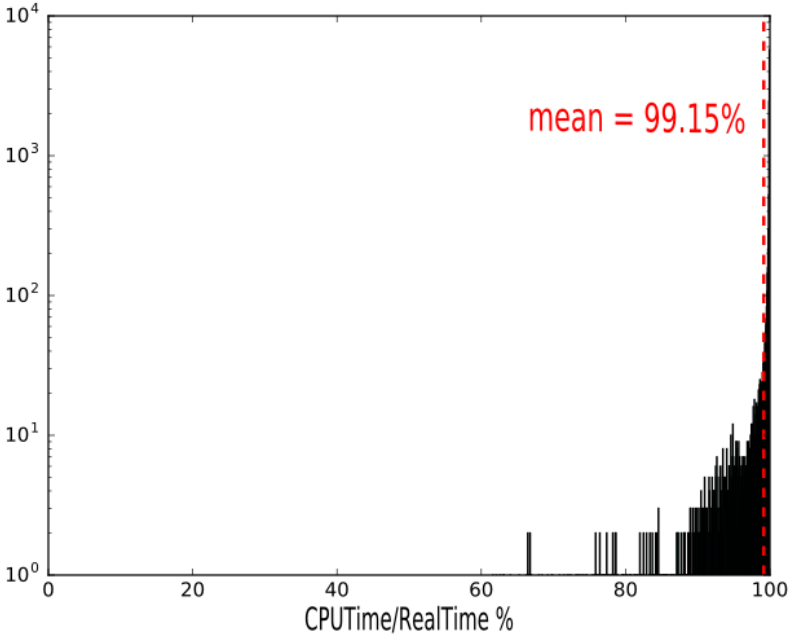
Performance Test

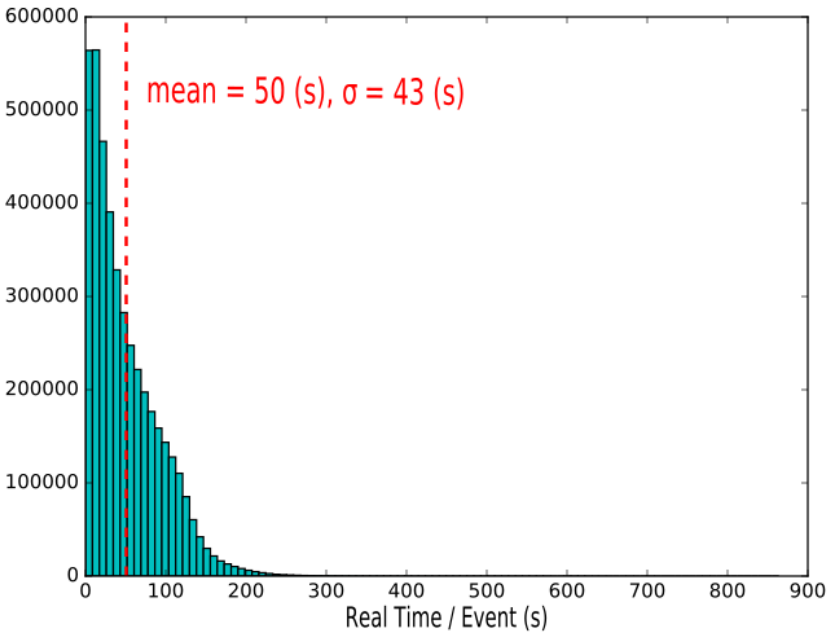


+120k core-hours real data reconstruction test shadowing a production at BNL

- < 4% failure → Success Rate > 96% → Qualifies this tier-2 facility to production ready
- A solution has been identified to increase the Success Rate to the +99% range
- This test demonstrates the trivial DB scalability with local MySQL servers

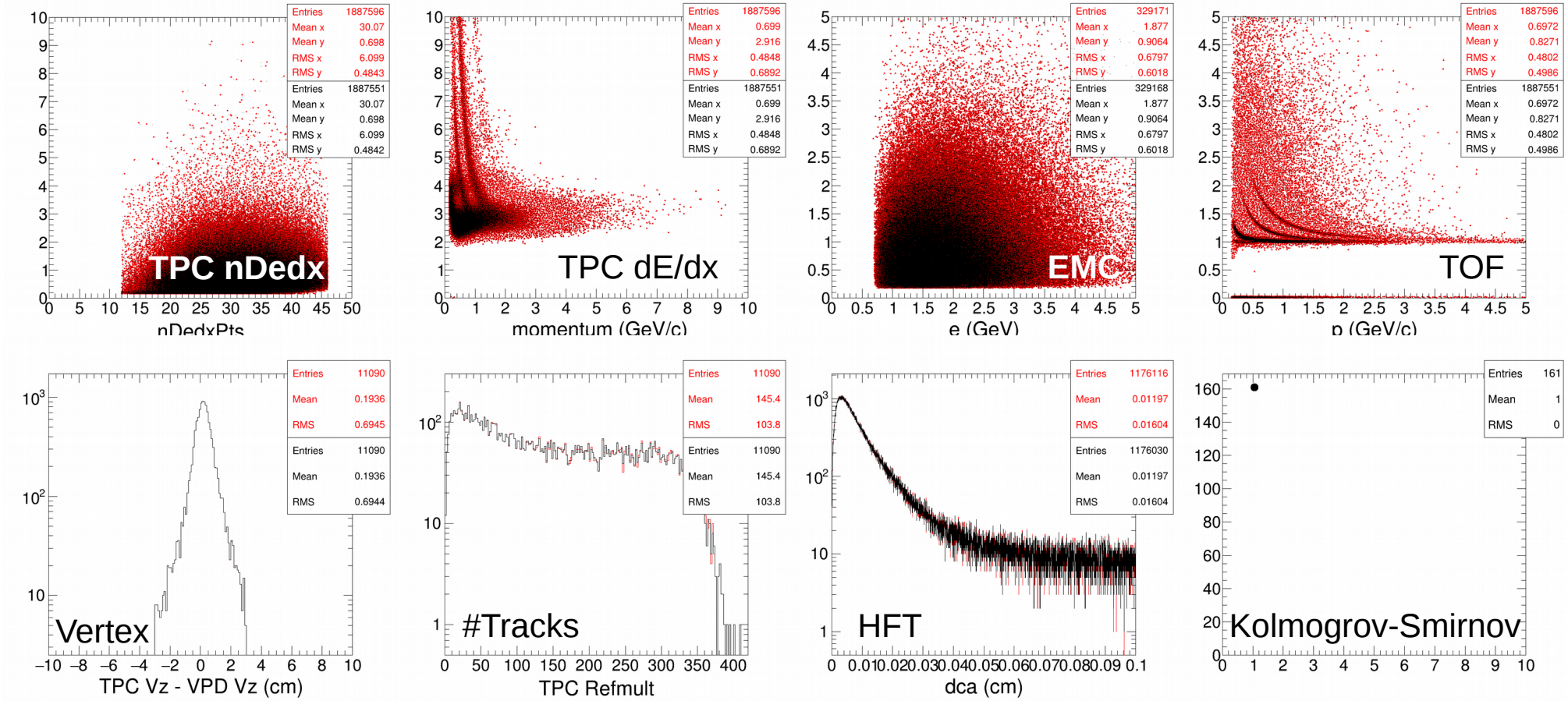
Performance Test





- 16 production processes + 1 MySQL server per 16 CPU cores → +99% walltime efficiency
- 50s / Au+Au event is comparable to 48s without a local MySQL CPU overhead

Reconstructed Data Integrity



Examples of histograms comparison of production at RCF and at Edison.

Kolmogorov-Smirnov test of 161 (1D and 2D) histograms show that distributions are identical.

Corollary: files output merging works!

Summary & Outlook

- Fault tolerant (finite state) pipeline built and tested on +120k core-hours
  - ~99% walltime efficiency demonstrated
  - +96% overall pipeline success rate achievee
- End-to-end optimization:
  - Optimized NY-CA data transfer over Esnet
  - Scalable access to read-only conditions database in an HPC environment demonstrated

    Shifter perNodeCache feature is critically important for DB workloads
- We will start a 25M MPU hours (a few months long) full scale production at Cori

  Stay tuned for lessons learned from full production ...

More details of our setup can be found at arXiv:1702.06593

STAR Data Reconstruction at NERSC/Cori, an adaptable Docker container approach for HPC

Mustafa Mustafa, Jan Balewski, Jérôme Lauret, Jefferson Porter, Shane Canon, Lisa Gerhardt, Levente Hajdu, Mark Lukascsyk

(Submitted on 21 Feb 2017)

As HPC facilities grow their resources, adaptation of classic HEP/NP workflows becomes a need. Linux containers may very well offer a way to lower the bar to exploiting such resources and at the time, help collaboration to reach vast elastic resources on such facilities and address their massive current and future data processing challenges. In this proceeding, we showcase STAR data reconstruction workflow at Cori HPC system at NERSC. STAR software is packaged in a Docker image and runs at Cori in Shifter containers. We highlight two of the typical end-to-end optimization challenges for such pipelines: 1) data transfer rate which was carried over ESnet after optimizing end points and 2) scalable deployment of conditions database in an HPC environment. Our tests demonstrate equally efficient data processing workflows on Cori/HPC, comparable to standard Linux clusters.

Comments: 7 pages, 4 figures, The 22nd International Conference on Computing in High Energy and Nuclear Physics, CHEP 2016
Subjects: **Data Analysis, Statistics and Probability (physics.data-an)**; High Energy Physics - Experiment (hep-ex)
Cite as: **arXiv:1702.06593 [physics.data-an]**
(or **arXiv:1702.06593v1 [physics.data-an]** for this version)

Summary & Outlook

Linux containers enable use of cloud-type and HPC systems to provide vast elastic resources suitable for HEP/NP real data and simulation productions. This has the potential to address current and future challenging data processing needs of experiments.

STAR experiment demonstrated equally efficient data production workflows on Cori/HPC, comparable to standard Linux clusters.

We hope to bring this experience to the wider HEP/NP community through the HEP Center for Computation Excellence initiative (HEP-CCE @ hepcce.org)

ABOUT THE CCE

Find out about our organization, projects, opportunities, and how you can become an HEP-CCE partner. Learn more...

CURRENT PROJECTS

Data Transfer Challenge
Edge Services for HPCs
Generator Scaling
Containers and Virtual Machines
Mini-Apps for testing
I/O and Storage
Scalable Deep Learning

HEP-CCE: Promoting Computational Excellence

HEP-CCE Coordinators: Salman Habib (Argonne), Rob Roser (Fermilab), and Peter Nugent (LBNL)

The HEP-CCE is a cross-cutting initiative to promote excellence in high performance computing (HPC) including data-intensive applications, scientific simulations, and data movement and storage. Enhancing connections with DOE's Advanced Scientific Computing Research (ASCR) program is an important part of the Center's activities. This includes promoting future-looking R&D initiatives in exascale architectures and systems, intelligent networking, and new data management and data analysis tools. Although the HEP-CCE is not a service-oriented entity, limited resources are available to support collaborative computing efforts for the HEP community, including a common GitHub repository for open source codes, a website for aggregating useful information, and expertise within and without HEP for solving computational problems via the Expert Forum. The HEP-CCE also sponsors topical workshops and student training programs.