



ALICE

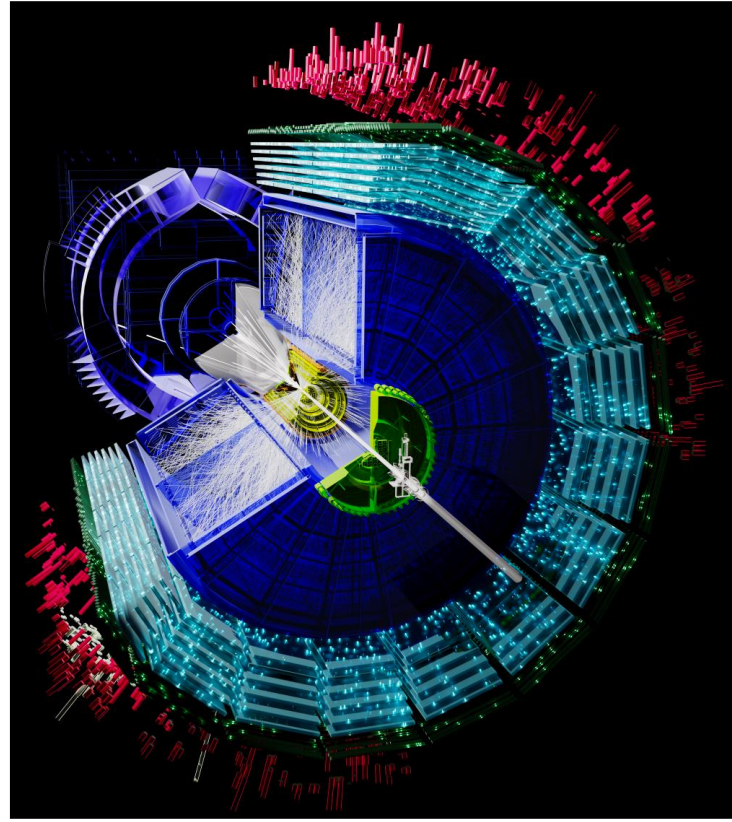
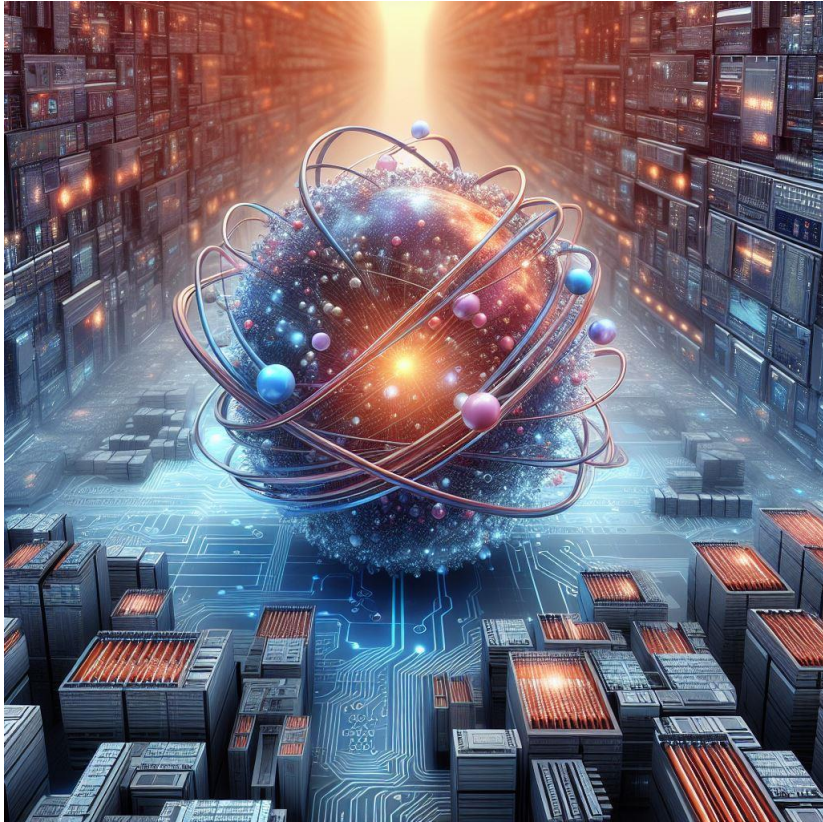
ALICE Computing

The 39th Winter Workshop on Nuclear Dynamics

16 February 2024

Irakli Chakaberia

Lawrence Berkeley National Laboratory



Premise

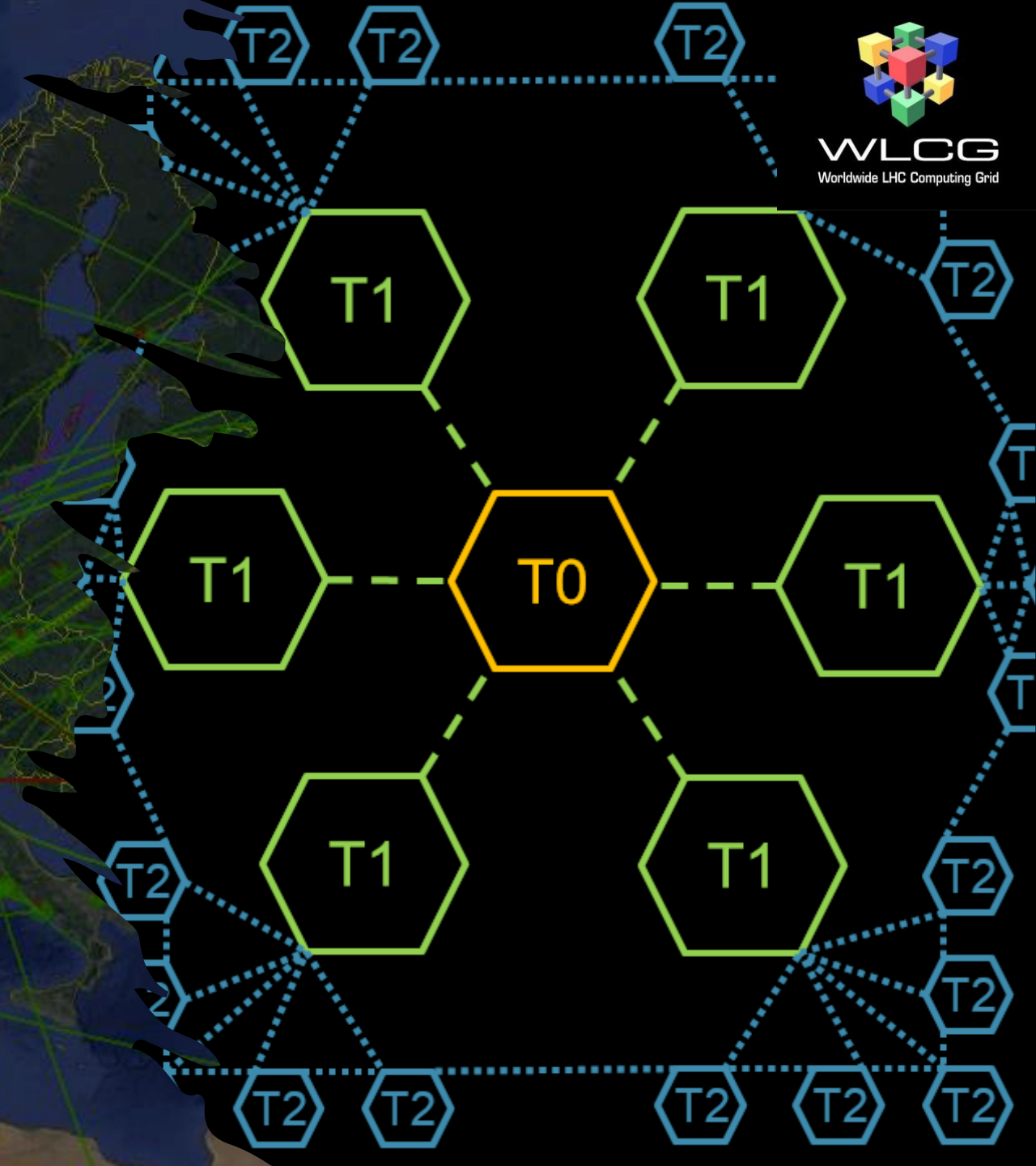
To learn about exciting processes of a subatomic world we need to build *sophisticated detectors* and *highly advanced computing infrastructure*

Excerpt from the Long Range Plan

“As we enter the era of exascale computing, with increasing numbers of communities within nuclear physics poised to take advantage of HPC, enhanced support will maximize scientific progress.”

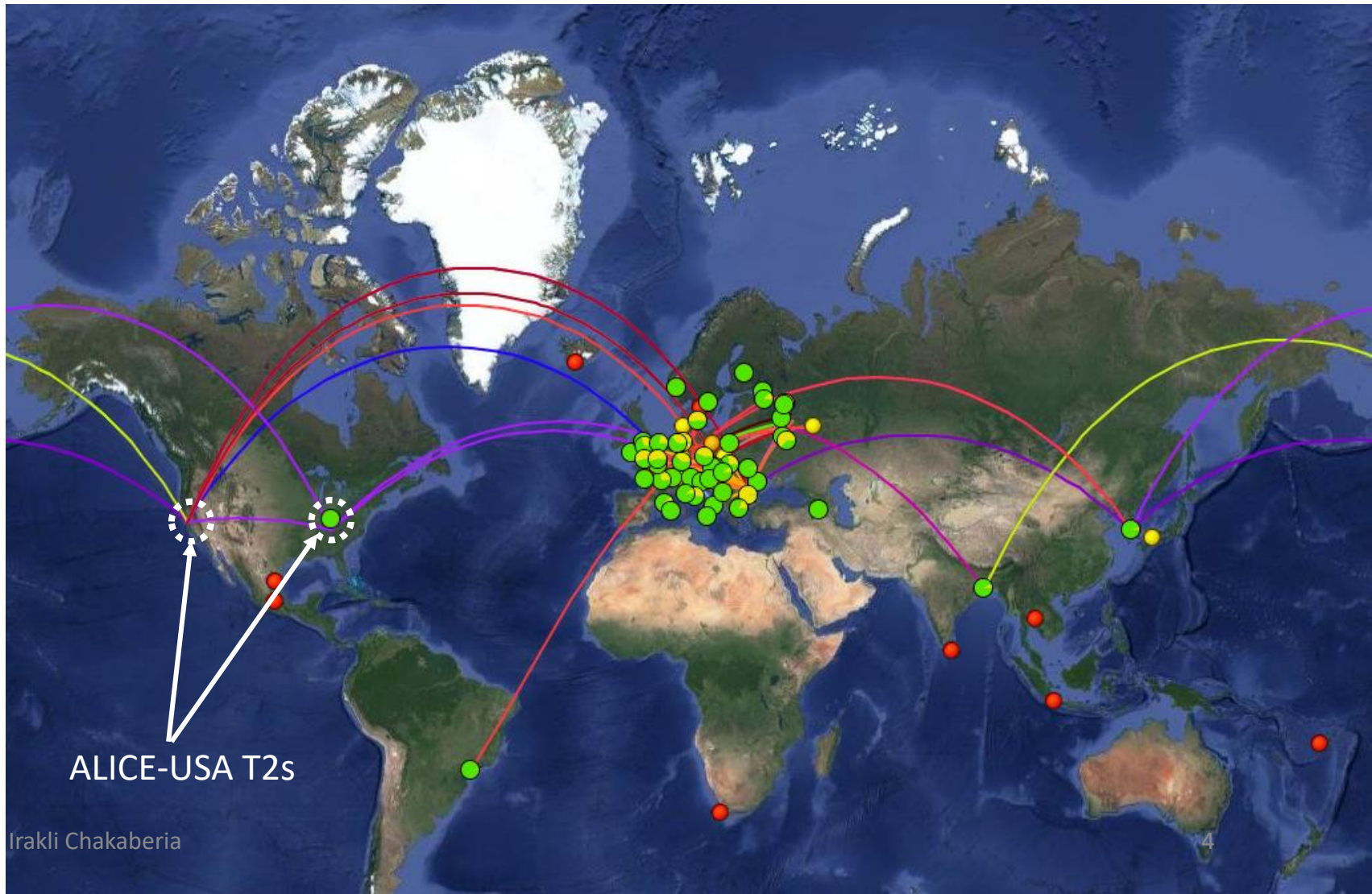
WLCG Computing infrastructure

- The Worldwide LHC Computing Grid (WLCG) project is a global collaboration
- The mission of the WLCG project is to provide global computing resources to store, distribute and analyse the ~200 Petabytes of data expected every year of operations from the Large Hadron Collider (LHC) at CERN
- It operates around 170 computing centers in more than 40 countries
- Globally distributed system of computing centers:
 - Configured in a tiered architecture that functions as a single coherent system
 - Tier 0 – Tier 1 – Tier 2 – Tier 3
 - Each center provides Grid-enabled gateways to CPU and storage
 - some of the centers also provide Analysis Facilities
 - Extensive high-quality network allows for communication among all computing centers



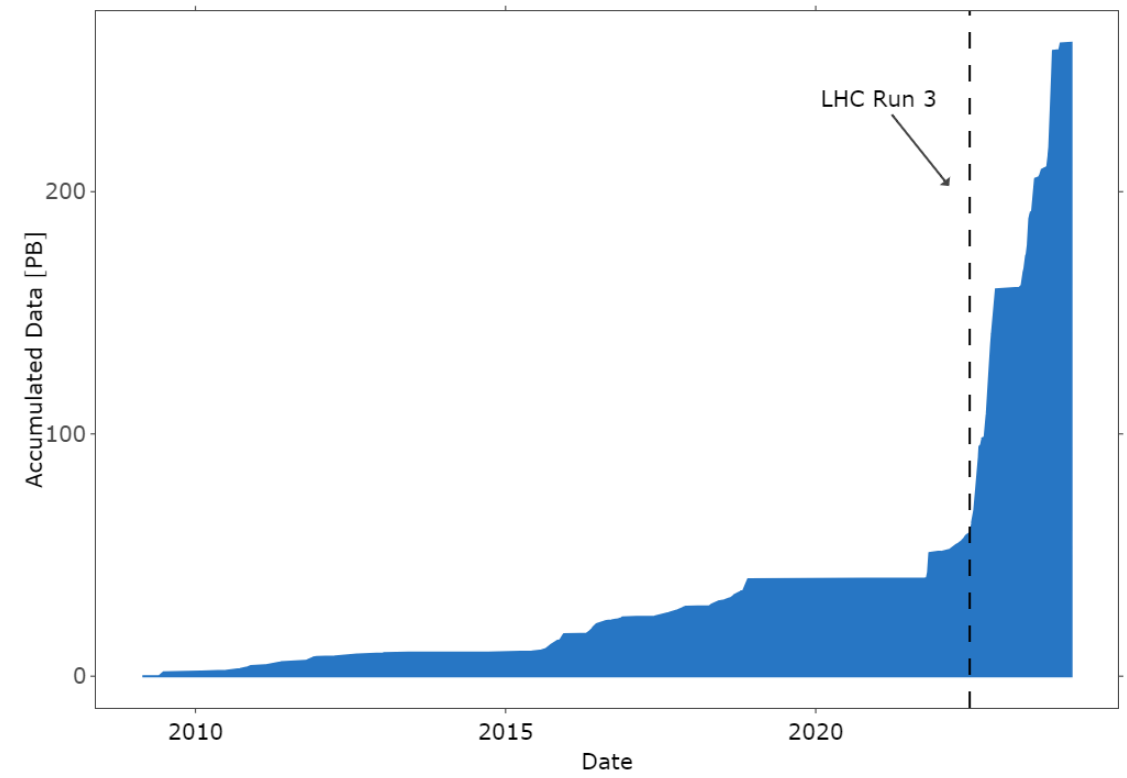
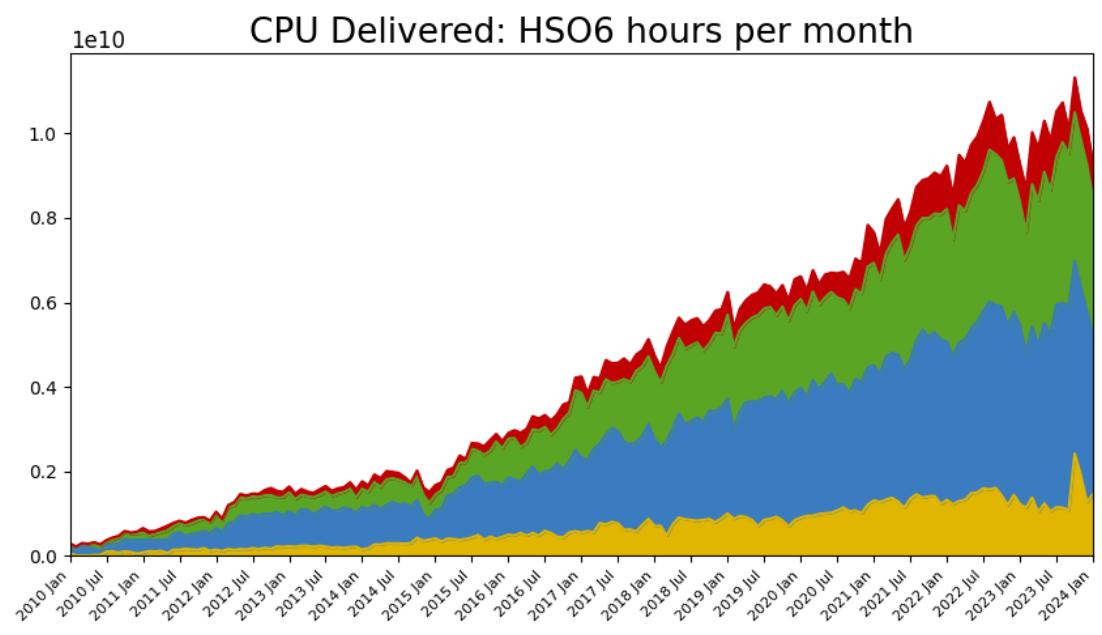
ALICE Grid

- ALICE computing philosophy is grid distributed computing between collaboration institutions around the world
- Grid consists of:
 - ALICE Tier-0 site at CERNs
 - 7 Tier-1s and about 100 Tier-2 sites around the world
 - US operates two Tier-2 sites, currently at LBL and ORNL

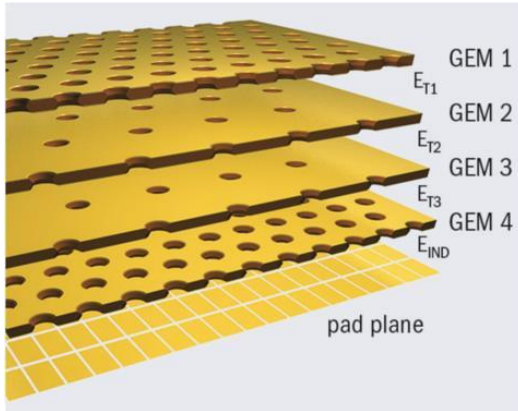


History in one slide

- In order maintain sustainable computing model with ever increasing data rates constant improvement in software and computing infrastructure is paramount

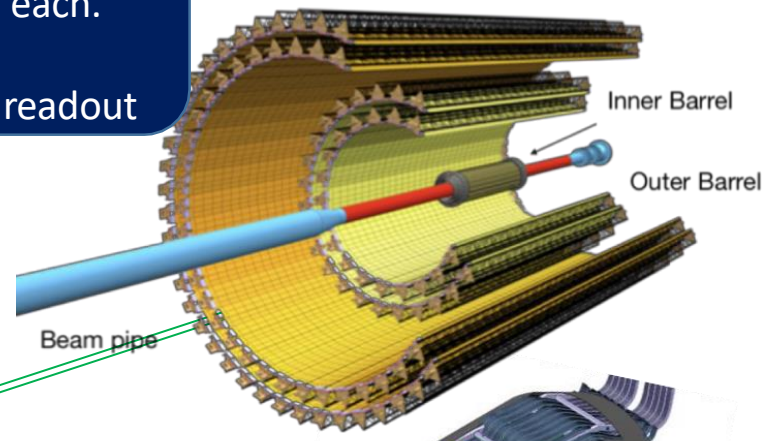


ALICE in Run 3

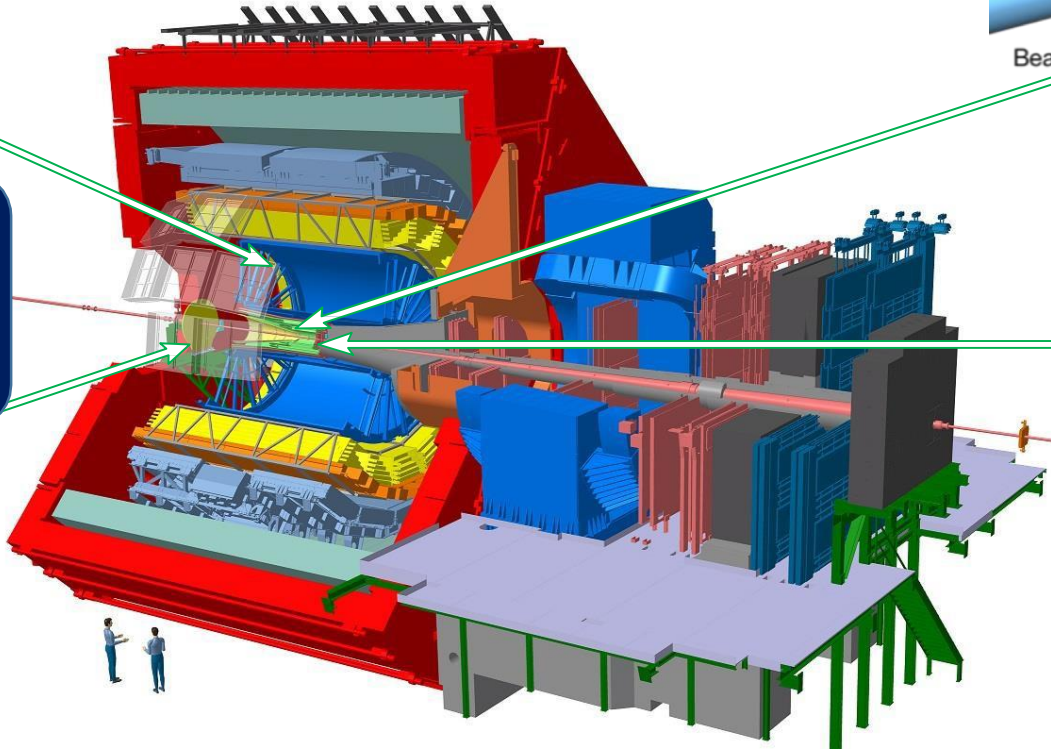
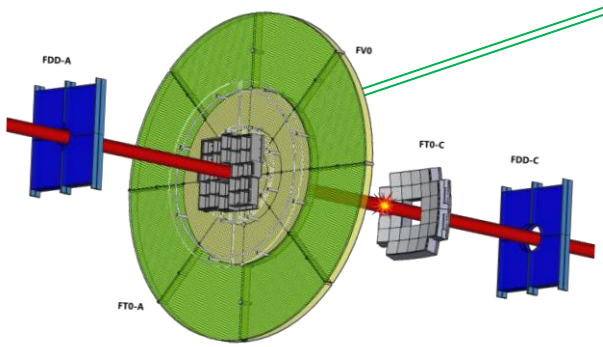


TPC MWPC readout → 4 layer GEM
(Intrinsic ion backflow ~99% blocking)
5MHz continuous sampling

New Inner Si Tracker: 10 m² of
MAPS with 29x27μm² pixel size
3 inner layers ~0.3% X0 each.
Closer to the beam
50-500 kHz continuous readout

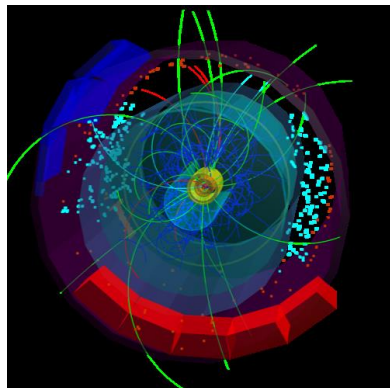


Fast Interaction Trigger (FIT) detector
Scintillator (FV0, FDD) + Cerenkov (FT0)
detectors to provide Min.Bias trigger
for detectors with triggered R/O

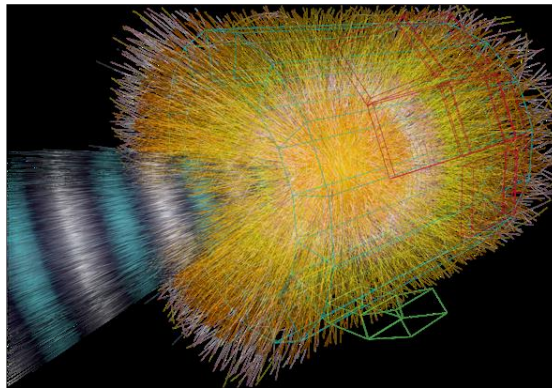


Muon Forward Tracker
to match muons before
and after the absorber.
Same Si chips as new ITS

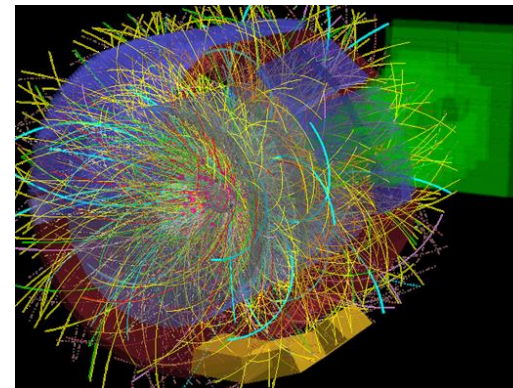
Run 3



p-p



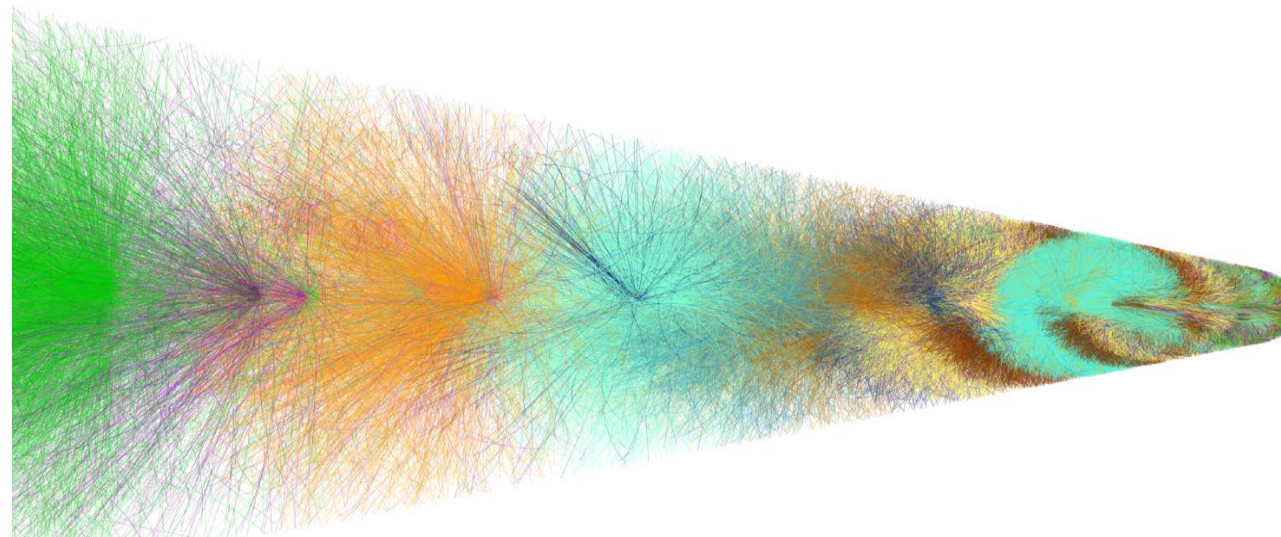
p-Pb



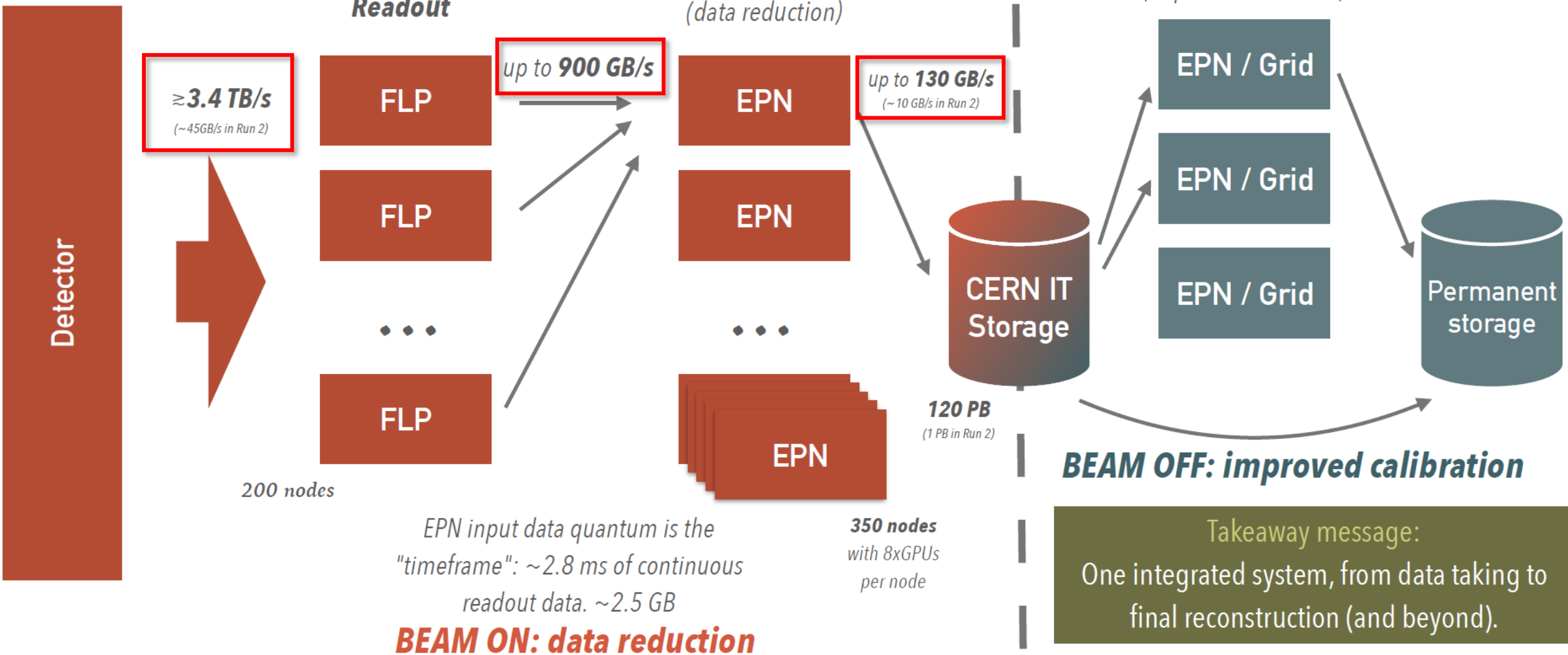
Pb-Pb

Run 2 – less than 1 kHz event rate

- Completely new detector readout
- Substantial detector upgrades
- Reconstruct TPC data in continuous readout in combination with triggered detectors
- Reconstruct 100 times more events online
- Store 100 times more events (needs factor 36x compression for TPC)
- Cannot store all raw data, use GPUs to do compression online
- Experiment has to cope with 4 times increase in resources over the next 10 years while handling 100 times more events. Use GPU farm to speedup processing.



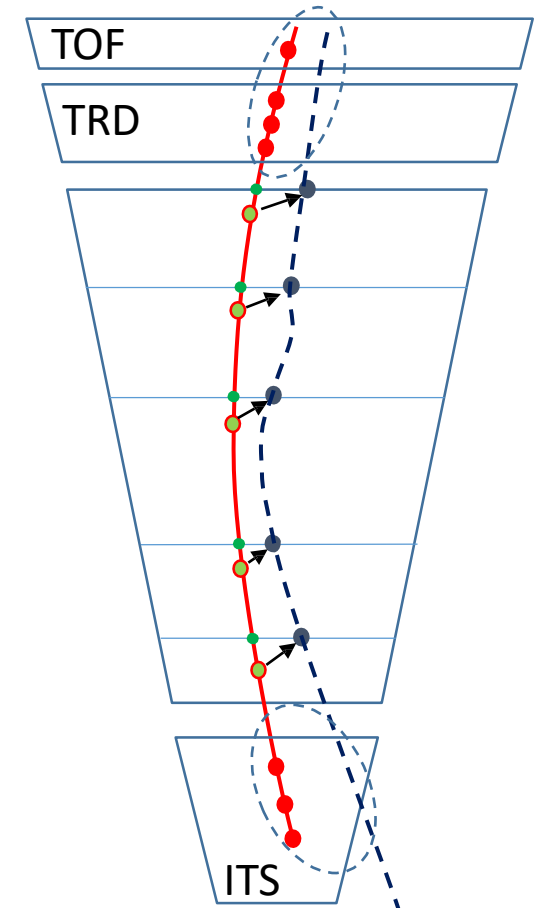
Overlapping events in TPC with realistic bunch structure @ 50 kHz Pb-Pb
 Timeframe of 2 ms shown (will be 2.8 ms in production)
 Tracks of different collisions shown in different colour





TimeFrame synchronous processing on EPNs

- Full TPC clustering and tracking (GPU)
- Full ITS+MFT clustering (CPU)
- Full FIT & ZDC reconstruction (may be considered also on FLP)
- Partial ITS tracking + ITS/TPC/TRD/TOF matching (CPU, GPU possible) as much as needed for **QC** and calibration:
 - ~2 kHz of 50-100% centrality events ($\langle \text{mult} \rangle = \sim 14\%$ of MB) provide enough statistics for per 1 minute Run2-like TPC distortion calibrations (most statistics hungry)
 - ⇒ 4% of collisions ($\sim 0.6\%$ of all tracks) @ IR=50 kHz





TPC Data Reduction and Compression

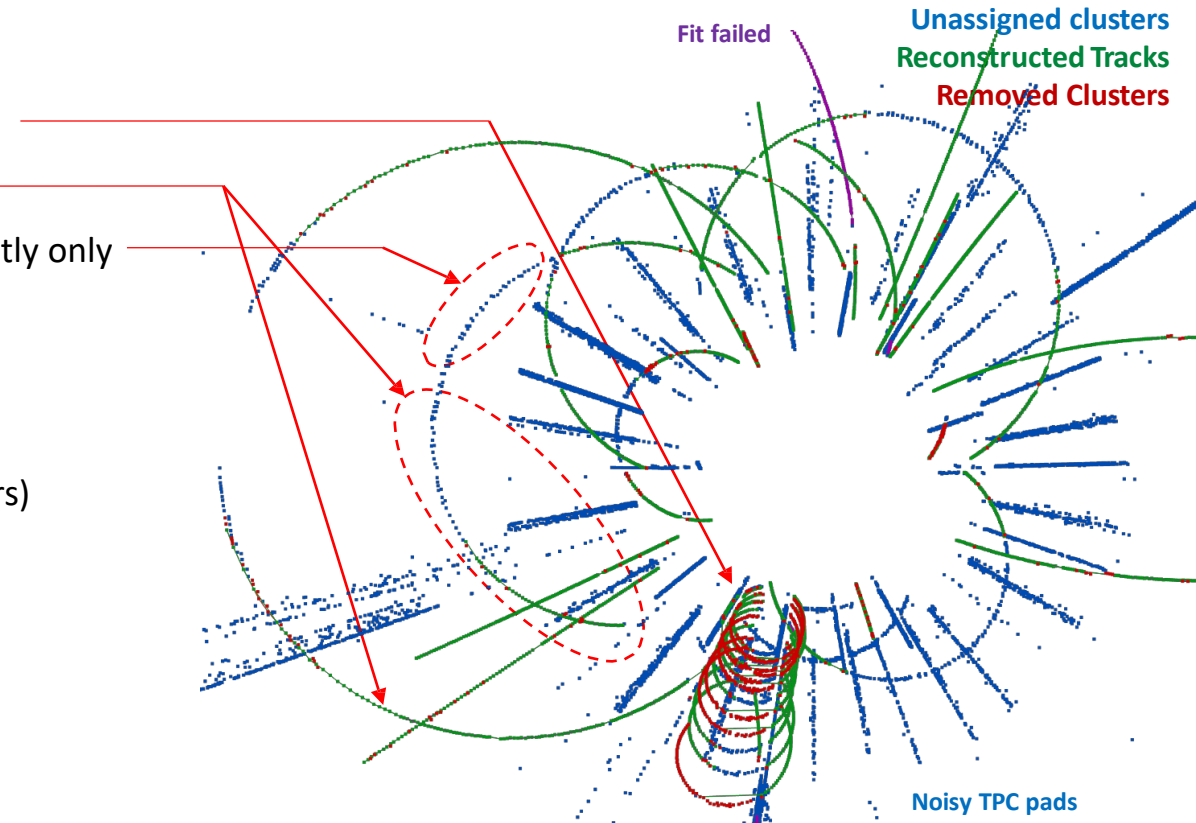
- Target in ideal case: reject ~50% of clusters
- Two alternative scenarios:

A: Keep all clusters except those from identified

1. (looping) tracks $p_T < 50$ MeV/c (not needed for physics)
 2. extra legs of loopers $50 < p_T < 200$ MeV/c
 3. segments of tracks with high inclination to pad-rows ($\varphi > 70^\circ$) currently only
~13% rejection rate achieved
- About 39% rejection achievable if
 - merging of looper legs improved
 - looper tagging can be extended to $p_T < 10$ MeV (~15% of clusters)
(track radii < 6 cm, Hough transform is tested)

B: Keep only clusters attached or in the vicinity of tracks

- interesting for physics ($p_T > 50$ MeV/c, principal leg for loopers)
- Currently ~37% rejection achieved
- About 52% rejection achievable in case of ideal loopers' legs merging



ALICE-USA Computing Project

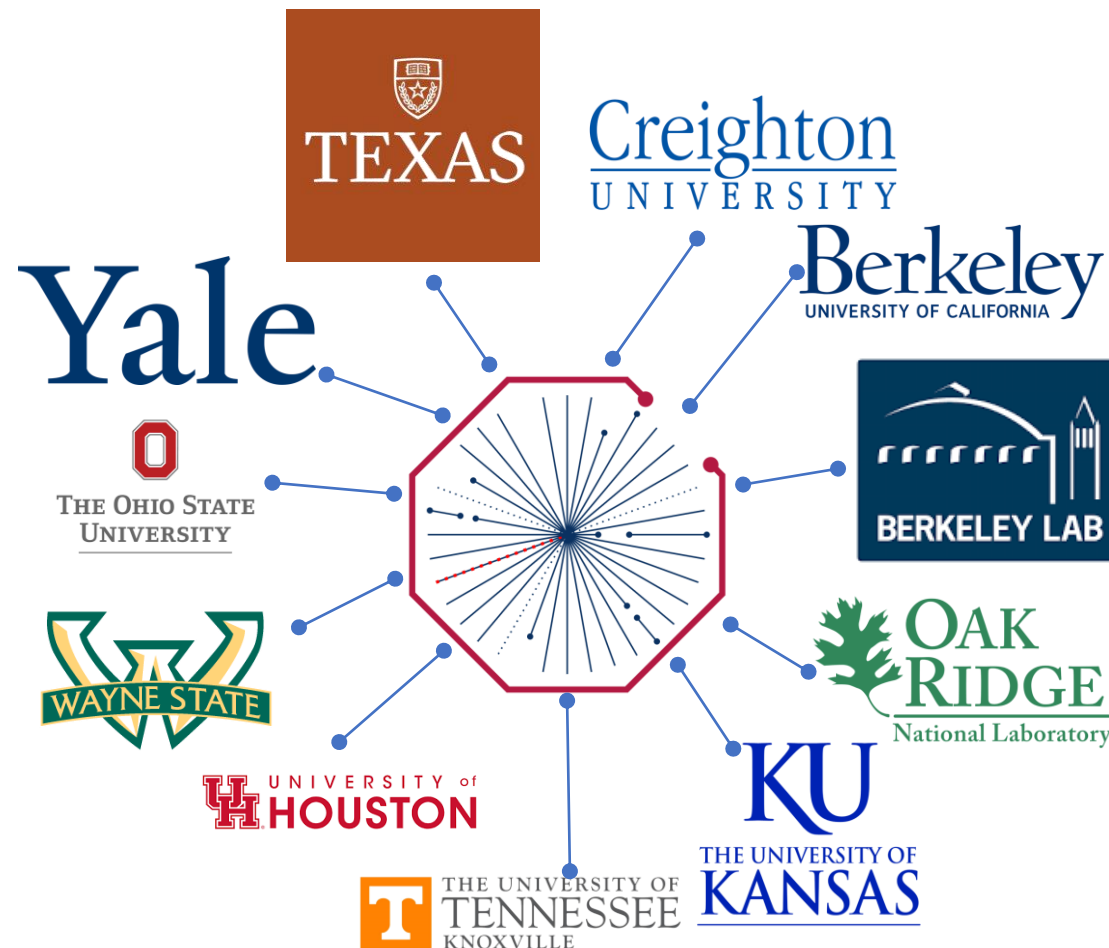
- ALICE-USA Computing project provides and maintains compute and storage resources for ALICE
- Fulfills DOE funded MoU-based ALICE USA obligations for compute and storage resources to ALICE
- Operate ALICE grid facilities at 2 DOE labs



Irakli Chakaberia
Mateusz Ploskon
Jeff Porter
John White
Karen Fernsler

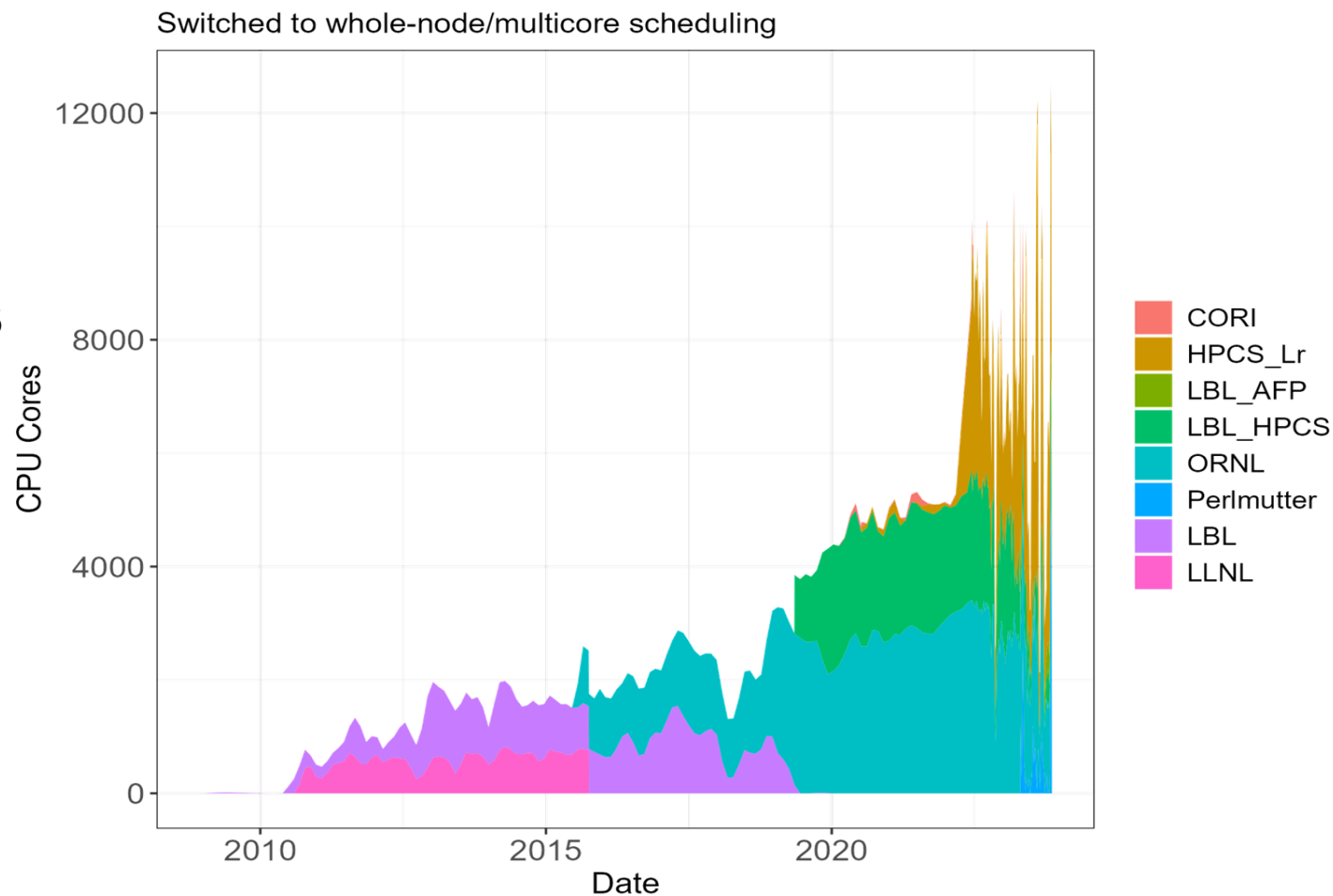


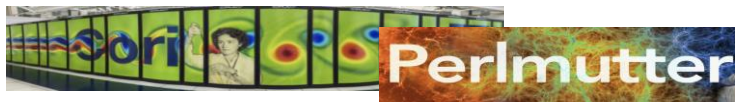
Ken Reed
Pete Eby
Steve Moulton



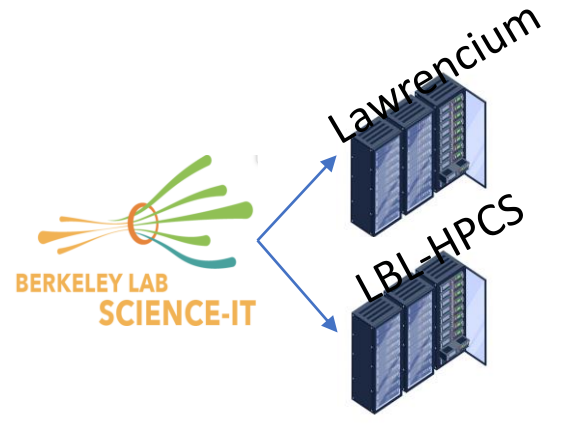
Project History

- Project was proposed in 2009
- Continuous operations since 2010
- Replace LLNL/LC with a new facility at ORNL
- ORNL T2 operational & LLNL shutdown by FY2016
- Operational changes 2017-2021
 - ORNL T2 moved within ORNL, personnel retained
 - New LBNL/ITD cluster - HPCS, PDSF retired in 2019
 - Report to the WLCG under the US_LBNL_ALICE federation
 - US R&D use of HPC resources, extends the ALICE grid model
 - Opportunistic use of Lawrence Livermore supercomputer at LBNL

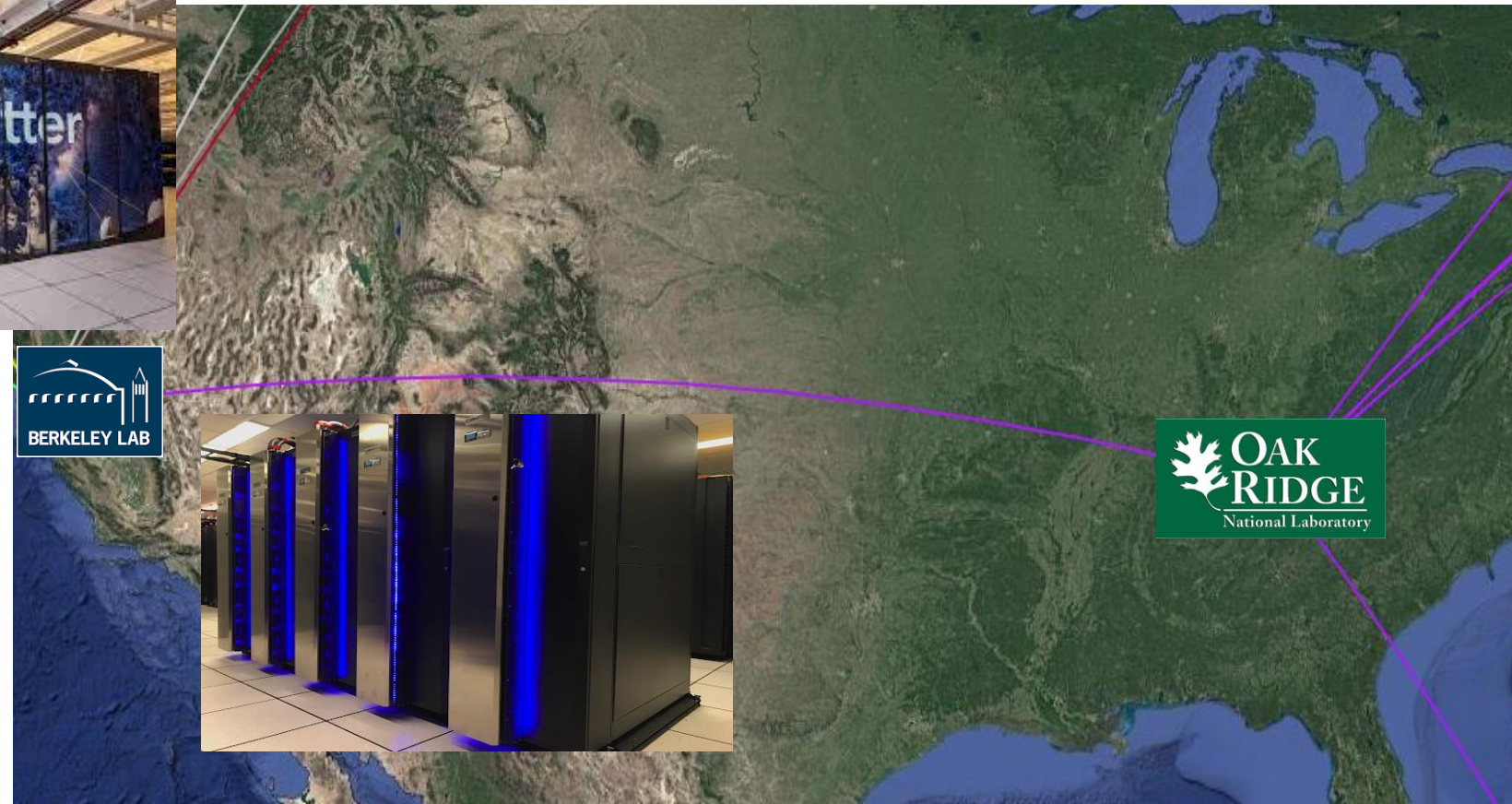




ALICE-USA T2 Sites and more



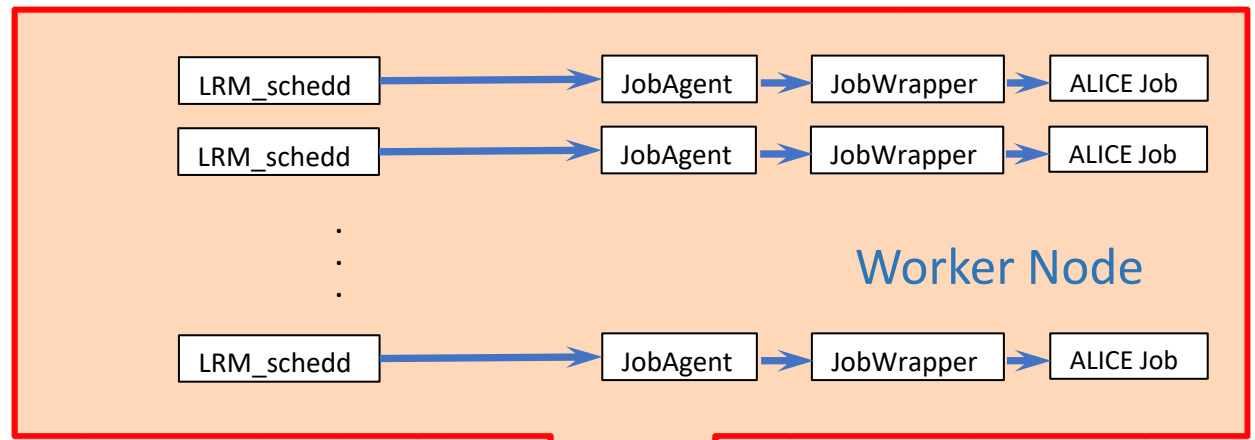
- Project currently operates two sites at ORNL and LBNL
- In addition, we provide resources on Lawrencium (opportunistic) and Perlmutter HPCs



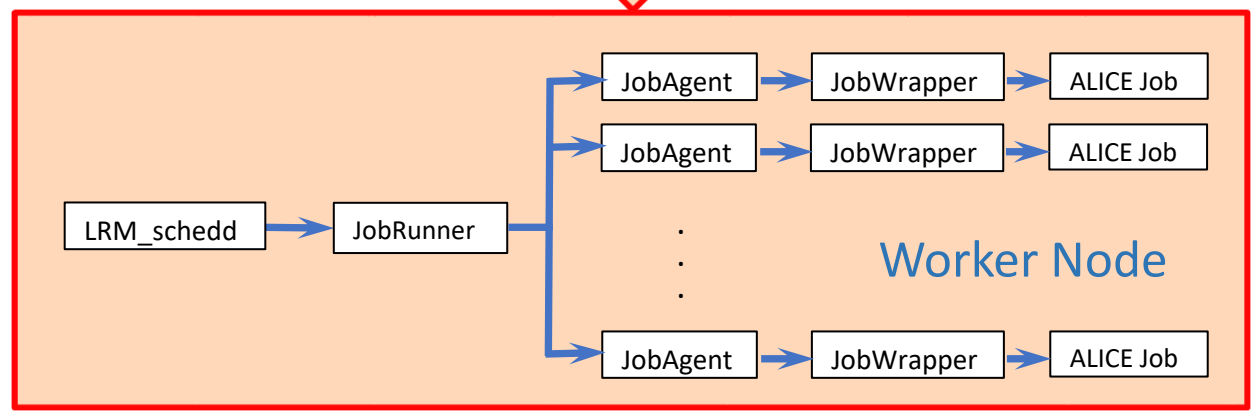
JAlIEn on HPC

- JAlIEn is ALICE grid computing environment that is responsible for job management on the grid (identifying resources, scheduling jobs, managing running jobs, reporting status, etc.)
- HPC provides access to a very high-performance multi-core CPUs
- To utilize these capabilities, we added multi-core job submission option into JAlIEn JobRunner
- JobRunner optimizes jobs for a particular node thus making use of a particular architecture matching it to the most appropriate job
- JobRunner model is being adopted for all JAlIEn deployments
- This work was performed as part of the ALICE-USA R&D project by the Student funded by the project – Sergiu Weisz

Serial Architecture



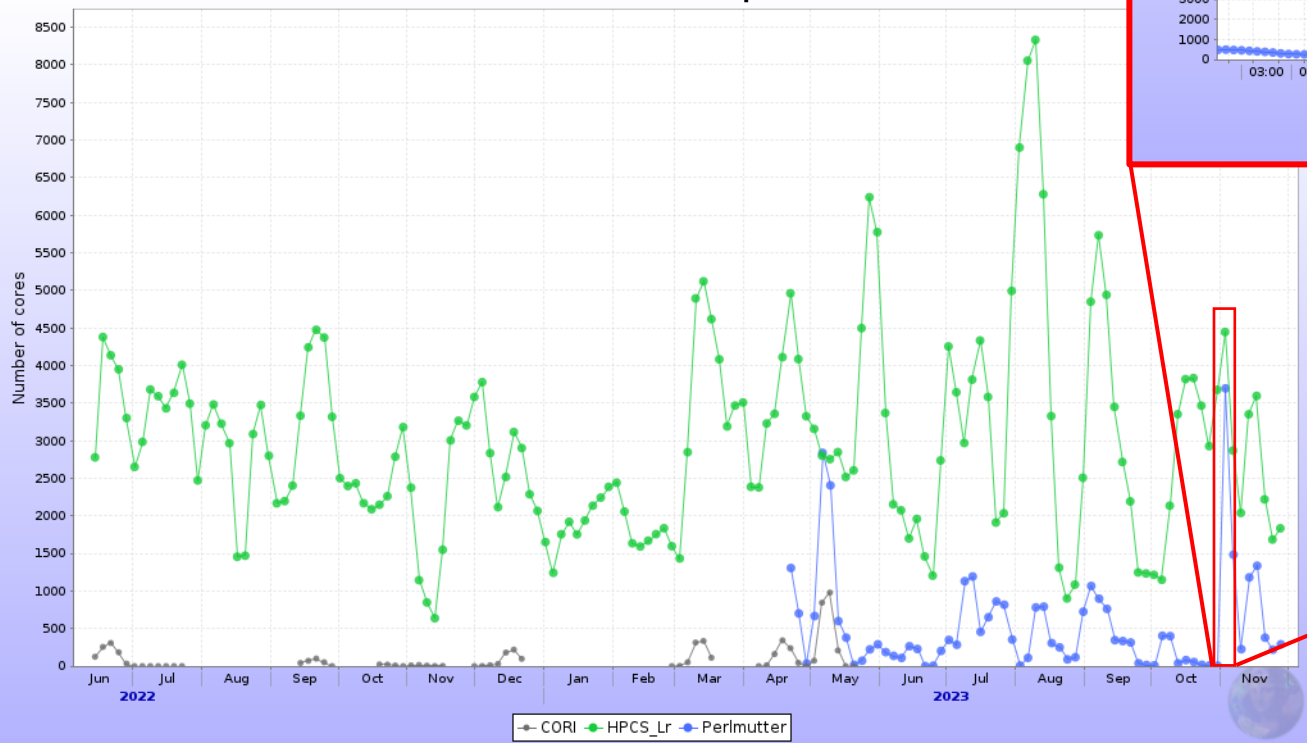
Node-level Architecture



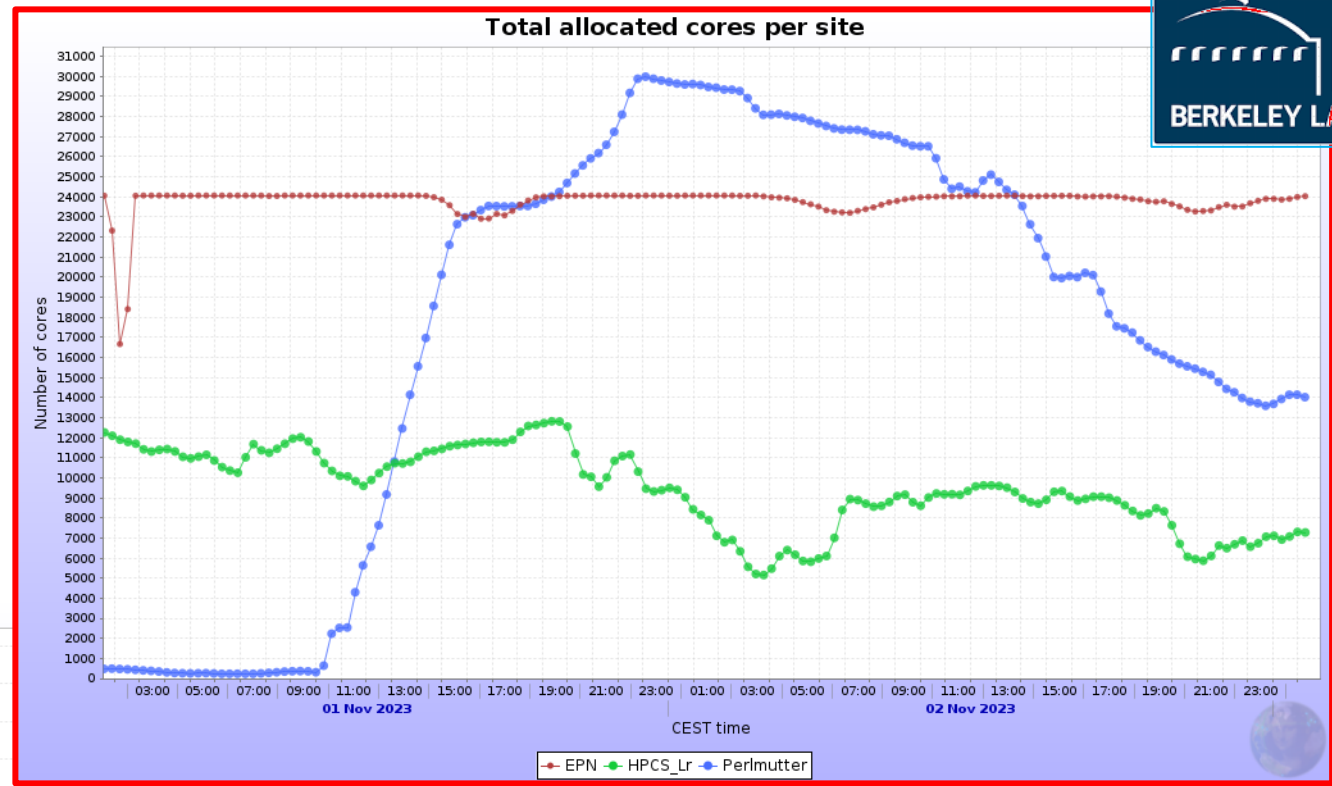
HPCs @ LBL

- As a result of the R&D work (and consequent grid-wide deployment of JAliEn) we can successfully use the HPC resources available at LBL

Total allocated cores per site



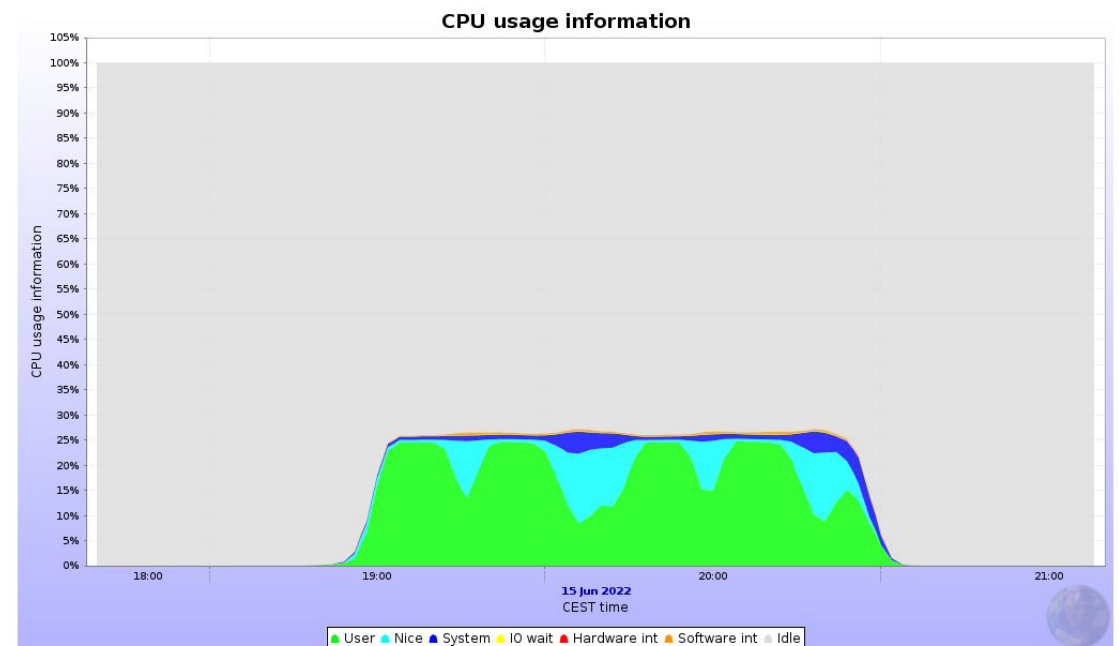
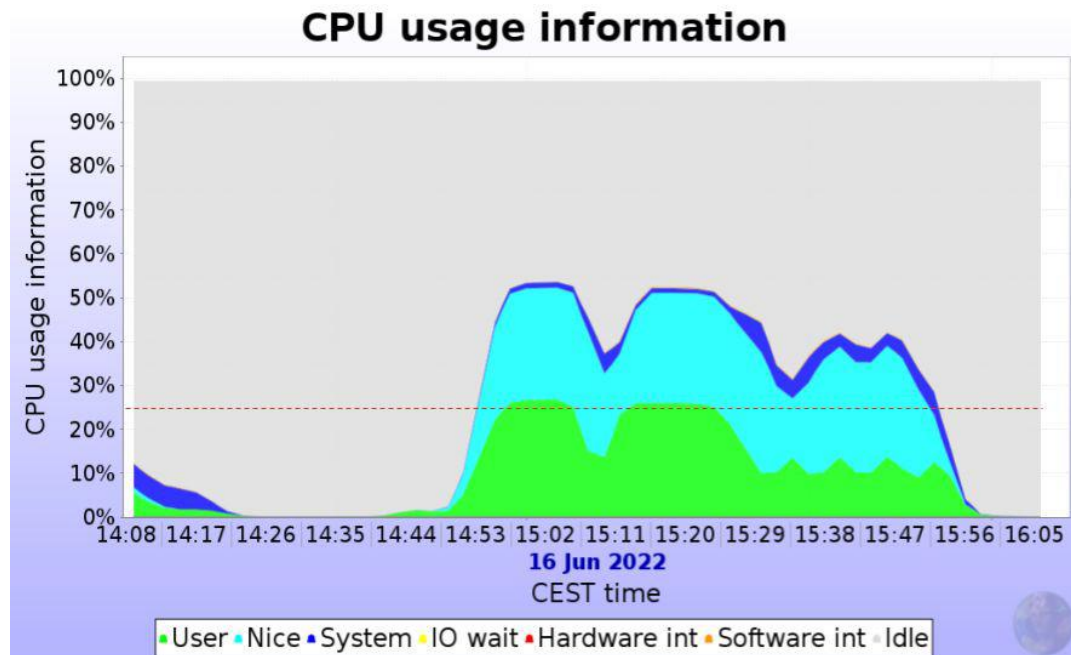
Total allocated cores per site



- Perlmutter is able to handle tremendous amount of computing power

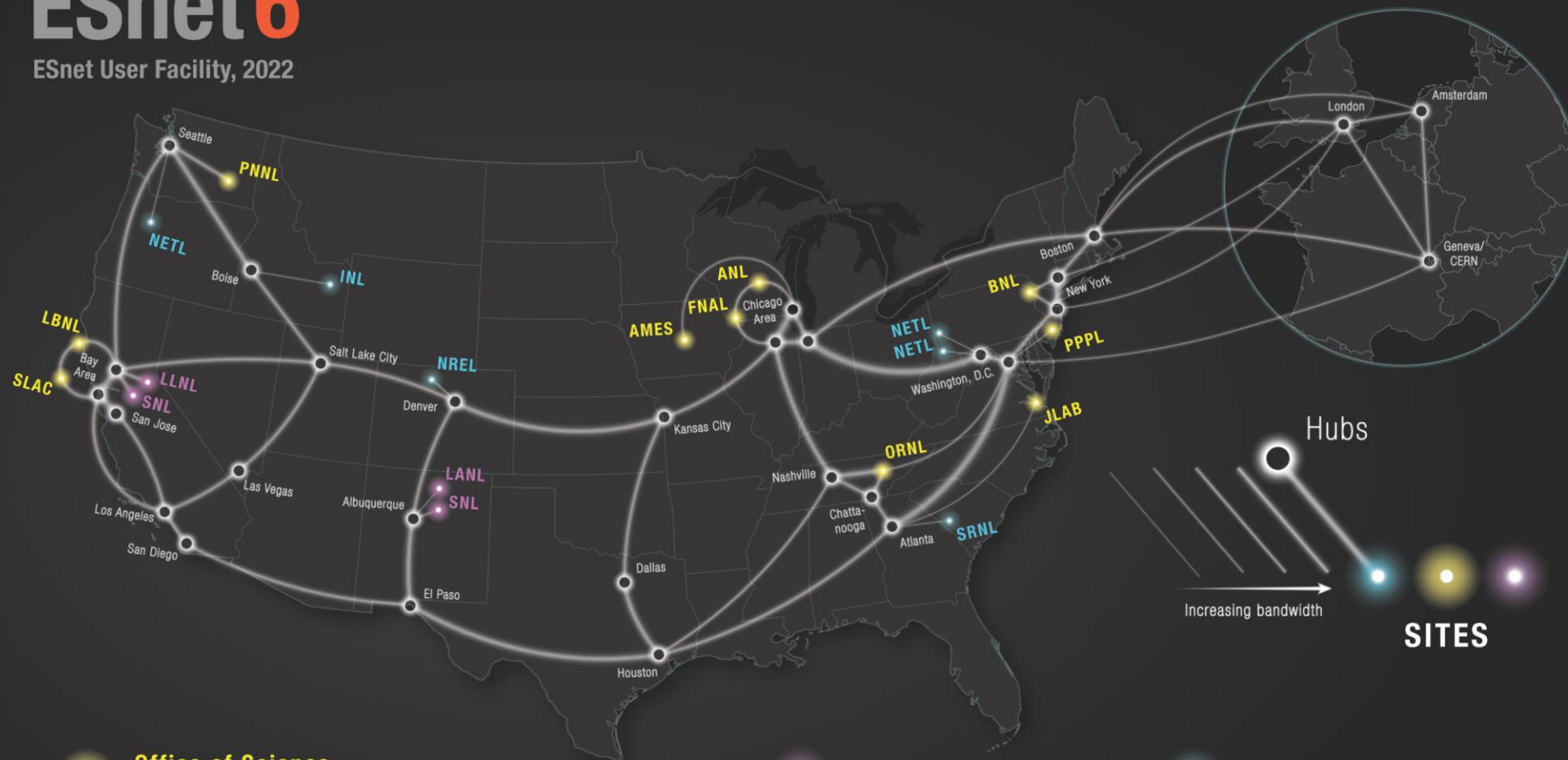
Improvements in whole-node submission scheme

- To optimize usage of heterogeneous ALICE grid infrastructure ALICE is going towards the full control over the compute node
- CPU pinning is one of the approaches in order to increase efficiency on an individual compute node
- To further optimize CPU efficiency JAliEn is looking to implement usage of “*cgroups v2*” to gain full control over workflow resource allowance



ESnet6

ESnet User Facility, 2022



Office of Science National Laboratories

- AMES** Ames Laboratory (Ames, IA)
- ANL** Argonne National Laboratory (Argonne, IL)
- BNL** Brookhaven National Laboratory (Upton, NY)
- FNAL** Fermi National Accelerator Laboratory (Batavia, IL)
- JLAB** Thomas Jefferson National Accelerator Facility (Newport News, VA)
- LBL** Lawrence Berkeley National Laboratory (Berkeley, CA)
- ORNL** Oak Ridge National Laboratory (Oak Ridge, TN)
- PNNL** Pacific Northwest National Laboratory (Richland, WA)
- PPPL** Princeton Plasma Physics Laboratory (Princeton, NJ)
- SLAC** SLAC National Accelerator Laboratory (Menlo Park, CA)

NNSA Laboratories

- LANL** Los Alamos National Laboratory (Los Alamos, NM)
- LLNL** Lawrence Livermore National Laboratory (Livermore, CA)
- SNL** Sandia National Laboratory (Albuquerque, NM; Livermore, CA)

Other DOE Laboratories

- INL** Idaho National Laboratory (Idaho Falls, ID)
- NETL** National Energy Technology Laboratory (Morgantown, WV; Pittsburgh, PA; Albany, OR)
- NREL** National Renewable Energy Laboratory (Golden, CO)
- SRNL** Savannah River National Laboratory (Aiken, SC)

Accelerating Scientific Discovery




Office of Science

ALICE Data Reconstruction @ LBL

- Thanks to our use of HPCs it is possible to even run data reconstruction at LBL
- Improvements in network infrastructure allows us to stream data directly from CERN
- We have recently successfully tested data reconstruction at our Lawrencium cluster while accessing data from CERN

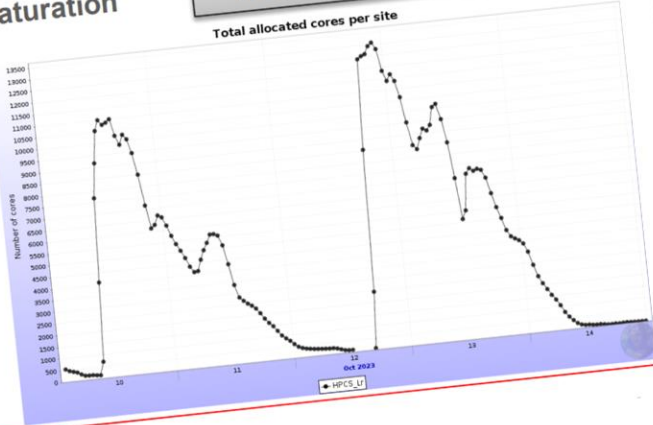
CTF Remote Reconstruction at US HPC Sites



- Tested the use of the US HPCs, Lawrencium (Lr) and Perlmutter, for remote Pb-Pb reconstruction
- Fully opportunistic resources
- **Jobs were run with remote access to the data**
- **Adequate network - data pulled from CERN, no saturation**
- Possible to expand temporary T0 with additional 6-10k cores
- Not entirely new - ALICE did remote Pb-Pb reco in Run 2, with limited radius due to network limitation
- Since the progress in the network over the past 5 years has been remarkable, eligibility increased
- Evaluating other potential candidates

Tests at Lr were successful
We are evaluating Perlmutter now

Total allocated cores per site



Number of cores

Oct 2023

HPC@L

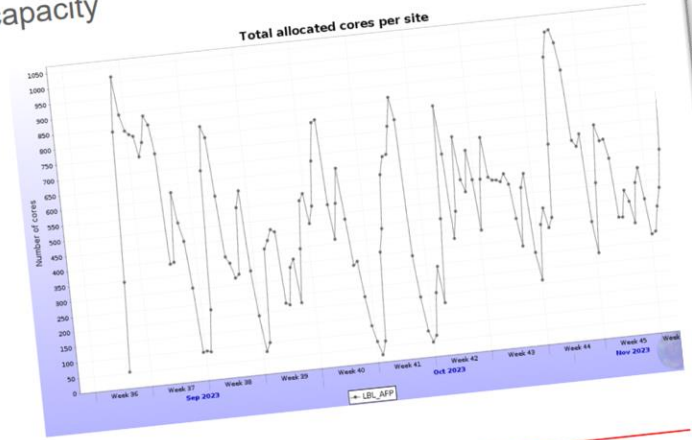
ALICE Analysis Facility

- ALICE Collaboration has introduced the concept of Analysis Facility (AF) in its Run 3 computing model
 - ALICE AF is meant to be a Grid Facility like T2s
 - AFs will provide resources dedicated for the AOD level physics analysis
 - Faster turnover for analysis tuning on the data subset

Large Ion Collider Experiment

ALICE Analysis Facilities

- In addition to the existing AFs (GSI and Wigner), added a new US-based AF
- US AF resources in addition to pledged capacity
- Started operations on September 7th
- The current setup provides
 - 1.1 PB of storage
 - 640 CPU cores
- Total computing capacity of the 3 AFs:
 - 10.5k cores
 - 8.4 PB storage

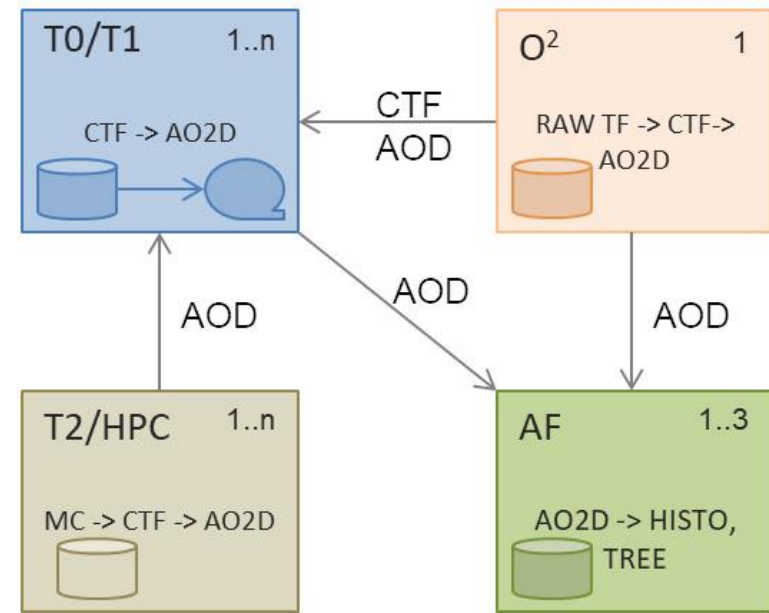


Reconstruction
Calibration
Archiving
Analysis

Simulation

Calibration
Reconstruction
Compression

Analysis



Hyperloop



- Organized job submission and monitoring system
 - Train system for O2 , replacing the LEGO train system
 - Allows organized analysis on the GRID and Analysis Facilities
 - Fully integrated with O2, allowing task configuration
 - Individual workflows - known as wagons - are combined into trains
 - Skimmed / Derived data stored for further processing in subsequent trains
 - Data available: converted Run 2 data, Run 3 data and MC, Derived data
 - Dedicated views for regular users and operators
- 24/5 Operation (3 different timezones)
 - Institutes: multiple ALICE-USA institutions, 2 in Europe, 1 in Asia
 - Shift-type support during working hours
 - Organized feedback sessions

USER



MyAnalyses

AllAnalyses

Dashboard

OPERATOR



Train Submission

Train Runs

Datasets

Derived Data

DPG runlists

Trains with issues



Summary

- LHC Run 3 and ALICE upgrades were challenging from the computing point of view
- Grid infrastructure and ALICE O2 software improvements allowed to handle 100-fold increase in the readout
- ALICE computing is growing and implementing novel approaches to:
 - Reconstruction
 - Reduction
 - Compression
 - Job optimization
 - Identifying and adding additional resources
 - Improving user experience for data analysis
 - etc.
- We are already looking forward to the HL-LHC related computing challenges to make sure the infrastructure and software will be ready

A background image showing a particle detector visualization with a dark blue and purple color scheme, featuring a central bright spot and radiating patterns of light points and lines, resembling a starburst or a complex network of data points.

Can't Computing
Watking

