



# Software and Computing at BNL: 2022 Highlights and 2023 Outlook

Torre Wenaus (BNL NPPS), Ofer Rind (BNL SDCC), and contributions from many

Software and Computing Round Table  
January 17 2023

# This talk

Scope is experimental nuclear and particle physics in the BNL Physics Department

- Selective and incomplete!
- A (necessarily) harder cut this year on material in the talk
  - More information in supplementary slides
- Emphasis on common software aspects
- Many fewer slides than last year, such discipline ;-)

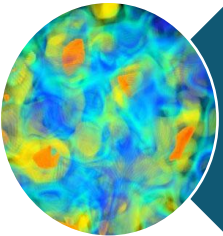
## Outline

- Computing facilities: SDCC
- Software: NPPS
- HEP experiments (going first this year)
- NP experiments
- Data and analysis preservation
- AI/ML and complex workflows
- Diversity, equity, inclusion, community
- 2023 outlook
- Supplementary slides

# Scientific Data and Computing Center (SDCC)



**RHIC**  
Relativistic Heavy Ion  
Collider



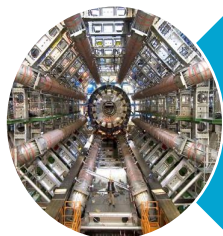
**LQCD**  
Lattice Quantum  
Chromo-Dynamics



**NSLS II**  
National Synchrotron  
Light Source II



**CFN**  
Center for Functional  
Nanomaterials



**ATLAS**  
Large Hadron  
Collider



**Belle II**  
Experiment in Japan



**ARM:**  
Atmospheric Radiation  
Measurement



**EIC**  
Electron-Ion  
Collider



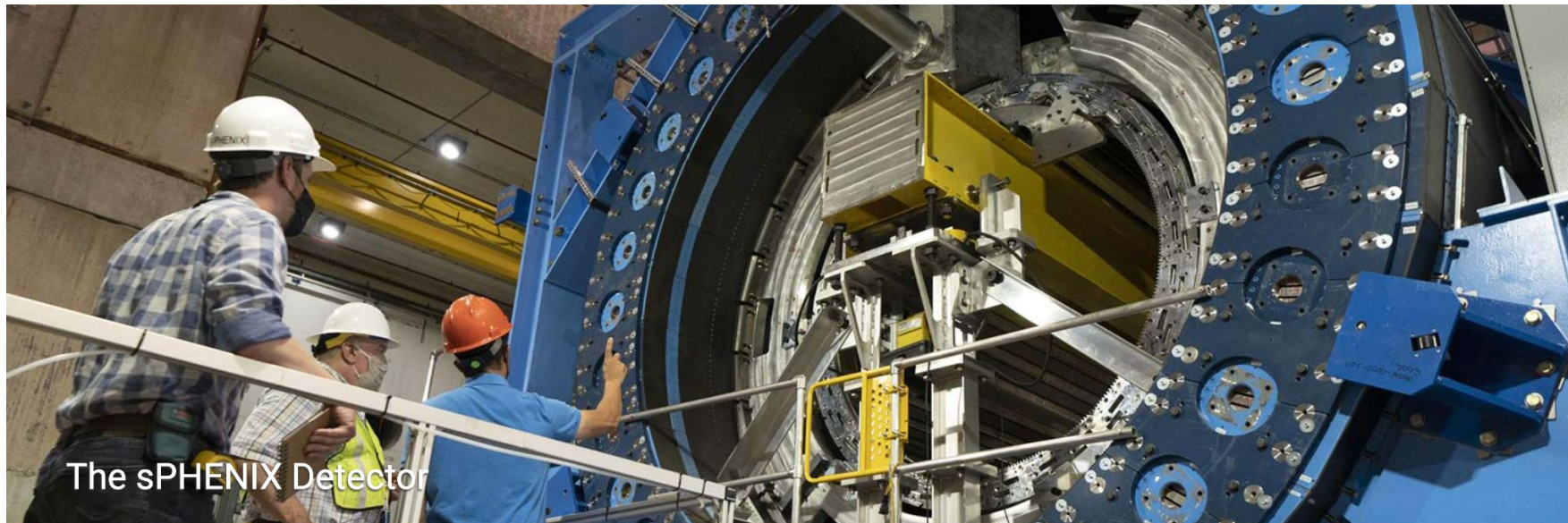
**NVBBCC**  
National Virtual Biosecurity  
for Bioenergy Crops Center



**Brookhaven**  
National Laboratory

# The sPHENIX Challenge

- Top priority: on-time sPHENIX readiness and datataking 2023-2025
  - The last RHIC running! STAR also running concurrently 2023-2025
- sPHENIX is one of the main drivers for several SDCC developments
  - High rate (20-35 GB/s) **data streaming** from DAQ to SDCC
  - Built a **new data center** with ~3x more space and ~5x more electrical power capacity
  - Increasing **tape storage** usage by a factor of ~3 (500+ PB)
  - Provisioning **~200k computing cores** by 2025
- Valuable experience for future support of EIC and HL-LHC programs



# sPHENIX Developments

## Online

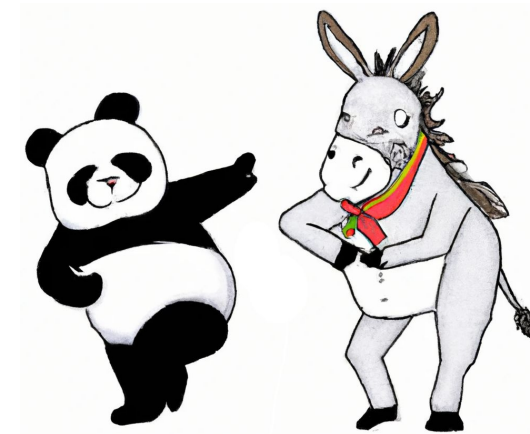
- Collaborative effort to **measure and optimize I/O** at counting house
- Support for **Lustre filesystem**

## Offline

- Dedicated **CVMFS repository**
- **Conditions DB** provisioning and support
- **PanDA** (DB, submit servers, authentication, etc)
- **Rucio** (including support for integration with PanDA)
- **Hardware** (cpu, disk, tape, network) planning and acquisition

## Collaborative tool support

- Federated **Indico** (in coordination with BNL ITD)
- **Shift sign-up** (adapting software originally used in STAR)



Courtesy of DALL·E

# Distributed and Central Disk Storage

A number of storage services for a variety of NP and HEP experiments

**dCache** for LHC-ATLAS, BELLE2 and DUNE

dCache.org 

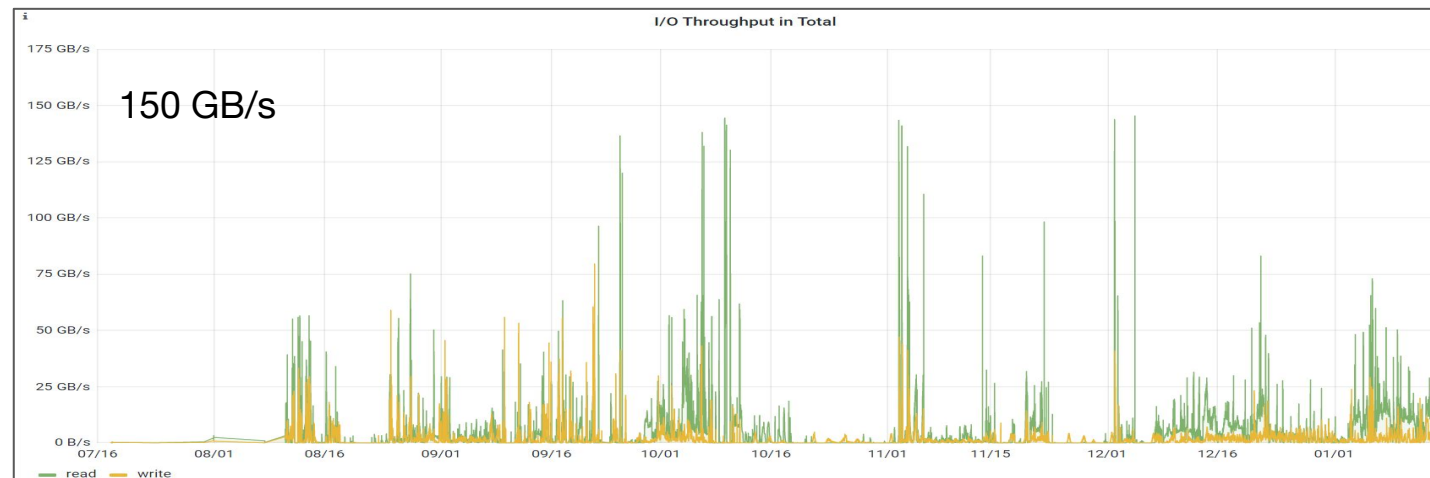
- 76PB (30% DISK) in around 122M files
- **Lustre** support for NSLS-II, sPHENIX, ATLAS, LQCD, STAR/EIC and CFN
  - Stable running with total capacity 60.8 PB and 572 million files
  - Upgrade coming this year
    - Deploy a new sPHENIX Lustre for workgroup users
    - Add 28PB to existing sPHENIX Lustre
      - Excellent streaming performance, aggregate throughput of 210 GB/s

l.u.s.t.r.e.®

Studies on several storage technologies inform production services

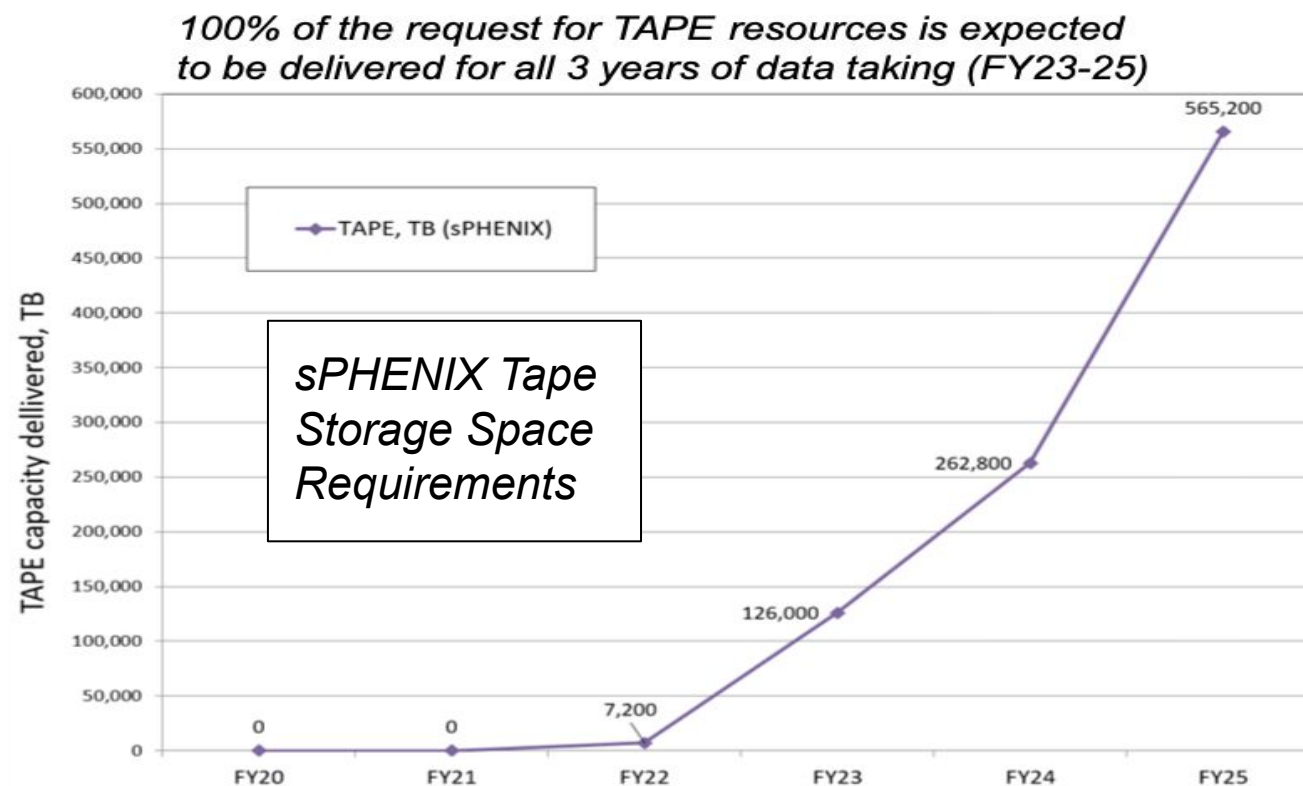
- Lustre, Xrootd server, dCache, ZFS studies and performance comparisons

*sPHENIX Lustre throughput  
last six months  
Up to 150 GB/s*



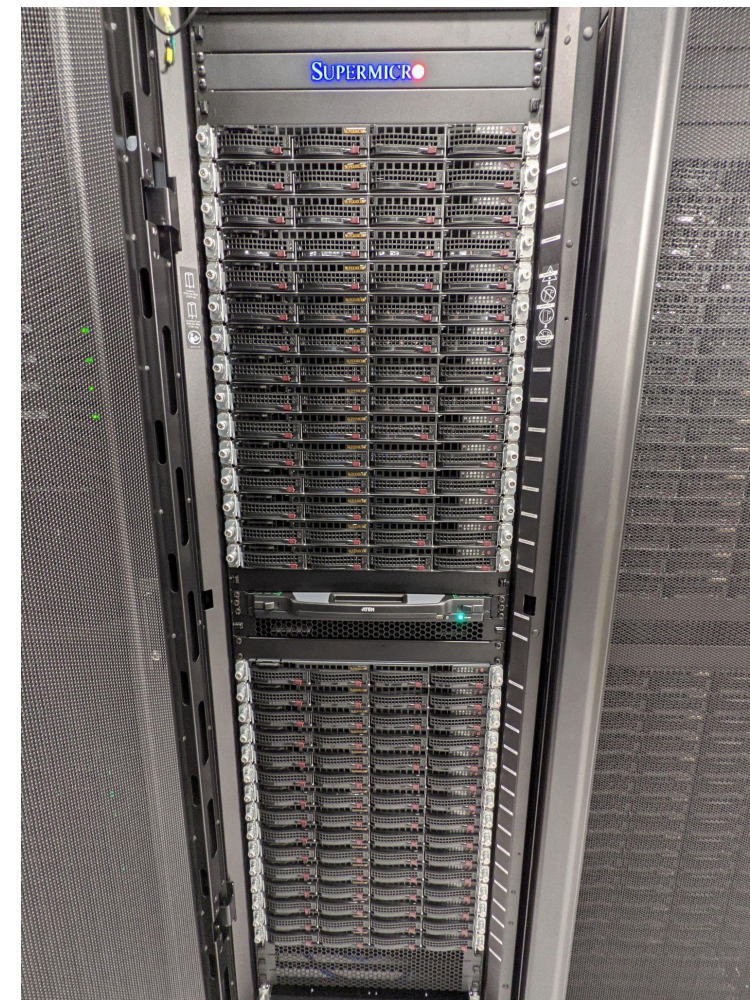
# HPSS Tape Storage

- Currently ~200 PB of data in HPSS with ~70k tapes
- New HPSS hardware in the newly commissioned data center for sPHENIX
  - 10GB/sec data injection requirement
  - High performance/capacity disk cache (2.1PB, 330 HDDs)
  - Two new IBM TS4500 tape libraries
  - Total of 64 new LTO-9 tape drives in the TS4500 libraries



# High Throughput Computing

- Providing users with ~1,900 HTC nodes
  - ~90,000 logical cores, ~1050 kHS06
  - Managed by HTCondor 9.0
- Purchased ~650 Supermicro SYS-610C-TR nodes for ATLAS and RHIC (~62k logical cores total)
  - Expected delivery of 20 racks in Feb/Mar 2023
  - System specs:
    - Dual Intel Ice Lake Xeon Gold 6336Y 24-core processors
    - 12x32 GB 3200 MHz ECC DDR4 RAM (384 GB total)
    - 4x2 TB SSD drives
    - 1U form factor
    - 10 Gbps NIC
- All nodes running Scientific Linux 7
- **Testing/preparations for an OS upgrade to Alma Linux 9 in progress**
- Beginning preparations for Condor 10.0 update



*2021 Supermicro SYS-6019U-TR4 Servers*

# High Performance Computing

## Currently supporting 5 HPC clusters

- **Institutional Cluster gen1 (IC)**
  - 216 HP XL190r Gen9 nodes with EDR IB
  - 108 nodes with 2x Nvidia K80
  - 108 nodes with 2x Nvidia P100
- **Skylake Cluster**
  - 64 Dell PowerEdge R640 nodes with EDR IB
- **KNL Cluster**
  - 142 KOI S7200AP nodes with dual rail Omnipath Gen.1 interconnect
- **ML Cluster**
  - 5 HP XL270d Gen10 nodes with EDR IB
  - Each node has 8x Nvidia V100
- **NSLS2 Cluster**
  - 32 Supermicro nodes with EDR IB
  - 13 nodes with 2x Nvidia V100

## New cluster coming: Institutional Cluster gen2 (IC)

Order placed with the following specs:

- 2x Intel Xeon (Ice Lake)
- 512GB DDR4-3200 CPU nodes
- 1TB DDR4-3200 GPU nodes
- NDR200 InfiniBand interconnect (200Gbps per uplink)
- 4x Nvidia A100 80GB

We are looking at a **performance boost of 3x from current IC node**: ~8 TF to ~25TF IC gen2 node.



*NSLS2 HPC Cluster*

# OKD Clusters

Two production OKD clusters brought online in 2022

- **ATLAS cluster**

- Primarily for **Analysis Facility** services that require kubernetes
  - [REANA](#) reusable and reproducible data analysis platform
  - [ServiceX](#) analysis data transformation system
  - N.B. JupyterHub in the analysis facility does not require k8s, uses batch (as can REANA also)

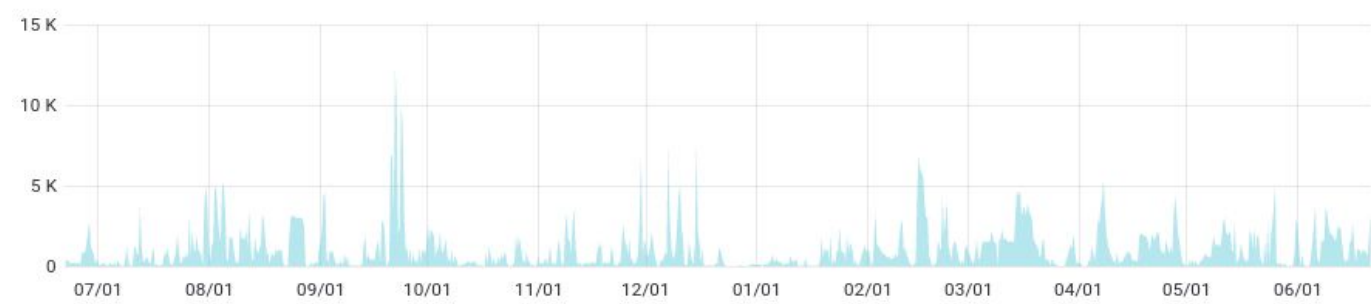
- **sPHENIX cluster**

- Primarily for **PanDA service** and **conditions database**
  - PanDA developers have created numerous helm charts
- Example of a developer/user-deployed service in OKD
  - Collaboration between SDCC and NPPS (Nuclear Particle Physics Software) groups
  - SDCC managing the OKD software/hardware
  - Conditions DB and PanDA deployments maintained/managed by the NPPS group, with SDCC support

reana



# ATLAS Analysis Facility



SDCC Shared HTCondor Pool Usage by the ATLAS AF

- SDCC shared pool used by ATLAS users for batch analysis
  - Users occasionally making opportunistic use of more than **10k cores**
  - Kerberos-authenticated **CERN EOS** mounts deployed on interactive hosts in 2022
- **JupyterHub** service with eight interactive VMs provided for ATLAS user analysis
  - Added **federated login** capability in 2022, accepting user MFA credentials from CERN, FNAL and SLAC, and implementing “lightweight” SDCC accounts
  - New **Lustre** deployment to address user storage need for analysis files
- Outlook for 2023
  - Expanded performance and capability testing using **Dask/HTCondor** and **IRIS-HEP** tools, such as ServiceX (upcoming [Analysis Grand Challenge](#))
  - Development of an **automated build service** for custom user containers

SDCC custom Jupyterhub interface

# Nuclear and Particle Physics Software (NPPS) Group

The Nuclear and Particle Physics Software (NPPS) Group consolidates much (not all) of the NPP software development in the Physics Department

23 NP and HEP members working on

- NP: EIC, PHENIX, sPHENIX, STAR
- HEP: ATLAS, Belle II, DESC, DUNE, Rubin Observatory, LuSEE@Night

physics → detector → **software** → performance → physics

**Emphasis: cross-experiment common efforts across HEP and NP**

Shared personnel, expertise, software

19 members are working on >1 experiment

Strong CERN based team

- 8 ATLAS, 3 Intensity Frontier (mainly DUNE)

Many collaborations

- SDCC/CSI at BNL
- WLCG (LHC)
- HSF (everyone)
- IRIS-HEP (NSF)
- HEP-CCE (DOE)

See <https://npps.bnl.gov>



# ATLAS Distributed Computing: Towards HL-LHC

**Wen Guan, Eddie Karavakis, Alexei Klimentov, Tadashi Maeno, Paul Nilsson, Torre Wenaus, Zhaoyu Yang, Xin Zhao**

## The flagship: PanDA workload manager

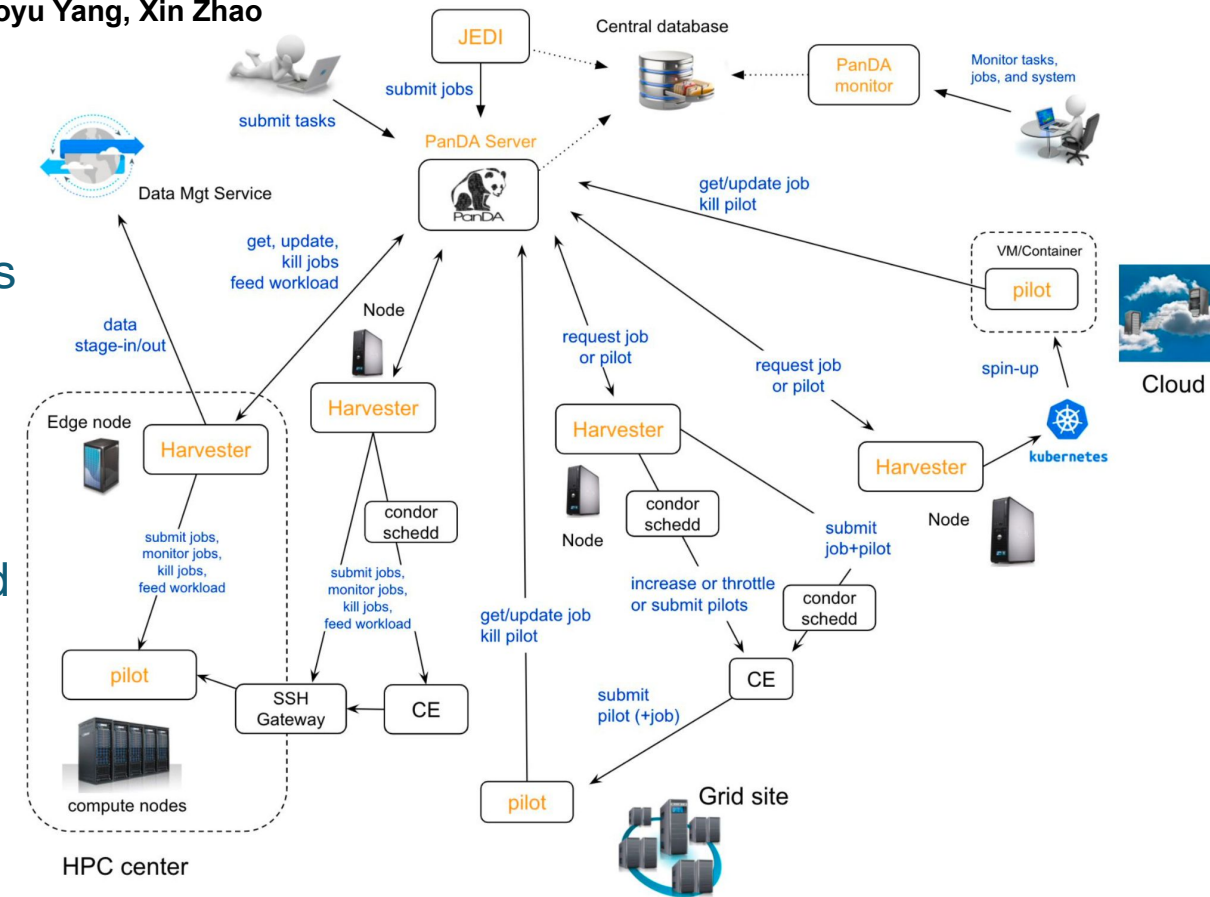
- Developed by BNL and UT Arlington 2005-present
- 24x365 processing on ~800k cores globally for ATLAS
- All workloads, production and analysis, at ~150 facilities
- ~1500 users, ~1M jobs/day, Exabyte throughput/year
- All resource types: clusters, grids, clouds, HPCs
- Collaborations with ANL, LBNL, SLAC, SBU, Madison

## Built to scale smoothly to HL-LHC

- Efficient optimized workflows to economize storage and processing at the HL-LHC

## Emphasis now on analysis functionality

- **Ease of use:** Improved interactivity through Jupyter and interactive PanDA monitor extensions
- **Speed:** Latency reduction throughout the system, more direct message-driven internal communication
- **Effective resources:** PanDA queues and sites ranking, leveraging non-grid platforms like k8s, clouds
- **Complex workflows: PanDA as an engine for large scale AI/ML** and other complex workflows



## Production and Distributed Analysis (PanDA) workload management system

PanDA's companion iDDS essential to complex workflows described on supplementary slide

# ATLAS Distributed Computing Highlights and Plans

- **Containerization** of all components and helm charts for k8s service deployment
  - Automatic image building and publication through github actions
  - Serving two k8s-based PanDA infrastructures: Rubin/SLAC and sPHENIX/BNL
- New **PostgreSQL** database back end implemented and moving towards production
  - New DB schema versioning to keep consistency among multiple instances and back end technologies
  - Code cleaning to get rid of legacy database tables, improve maintainability
- Development team organization and **inclusivity**
  - JIRA-centric project activity tracking, Mattermost channels for dev team, core meeting at US compatible time
- **Data carousel**: in production and entering a **new R&D** phase
  - **Smart writing** to tape by **grouping associated data**
  - Deleting less popular data and **recreating on demand**
- Complex workflows
  - Extending **AI/ML services**, developing advanced **active learning** algorithms
  - Planning **funcX/parsl** integration particularly to leverage LFCs for large scale workflows
  - Extending **workflow definition** support: integration of Snakemake's Python-based workflow description
- Sustainable computing
  - Plan to show the **CO2e emissions** per user/job/task
  - Exploring **elastic resource usage** based on electricity price, integrating green resource providers like Lancium
- Publication
  - The long-awaited (requested) **PanDA Paper** is in advanced preparation! **iDDS paper** will follow

# ATLAS Google R&D Project

Johannes Elmsheuser, Alexei Klimentov, Tadashi Maeno, Paul Nilsson, Xin Zhao

New BNL-led R&D project with Google is starting  
Building on a successful prior round

- PanDA grid site equivalent with processing and Rucio storage
- All production workflows work fine with very high efficiency
- See supplementary slide for Phase 1 projects

Particular focus on using and evaluating the Google cloud for analysis

- Leverage tools and capabilities of Google Cloud Platform (GCP)
  - ARM, GPUs, FPGA, large memory/CPU, large databases like BigQuery
  - Kubernetes engine with on-demand scaling
  - Jupyter notebooks with DASK backend and Cloud storage
  - Cloud storage with S3
- Ideal for bursty work and requests, leverage the elasticity

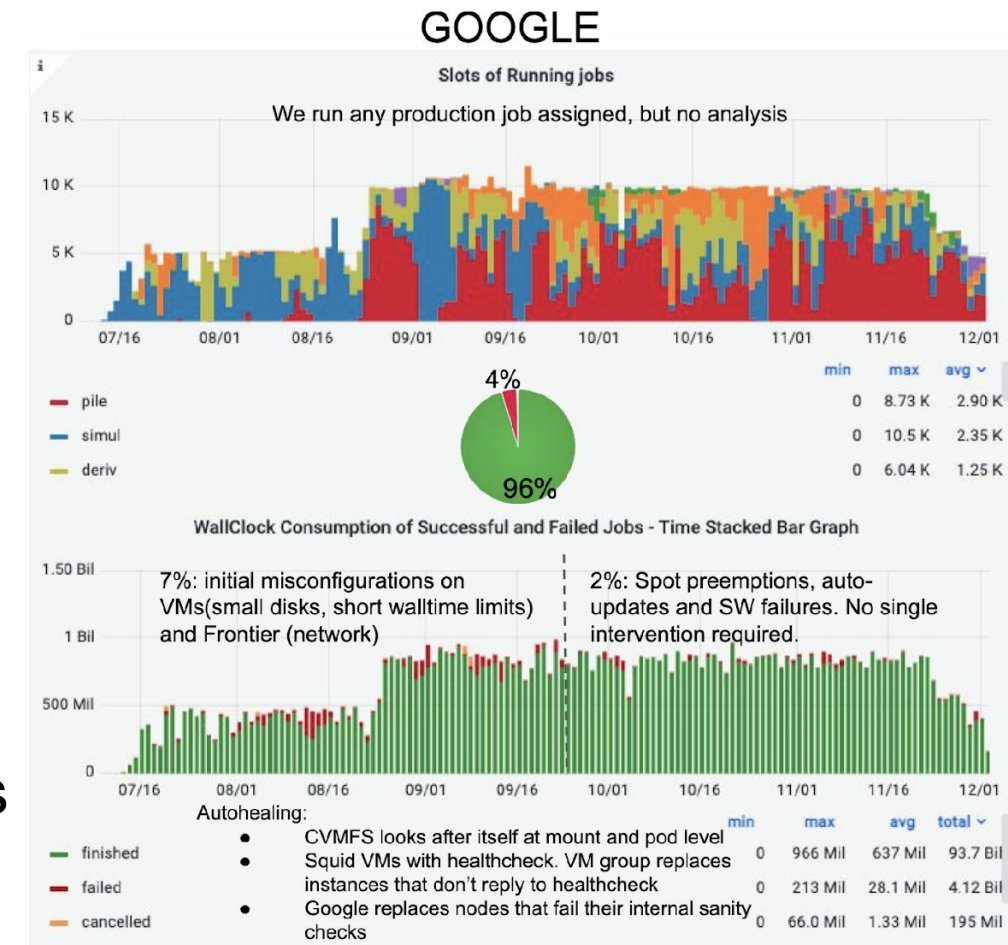
Also make a fair comparison to traditional resources

- Evaluate Google as an ATLAS site able to run all workflows
- Provide a measure of total cost of ownership

Resource scale 10k virtual CPUs, duration 18 mo



## Stability



# ATLAS Offline Software

Johannes Elmsheuser, Marcin Nowak, Scott Snyder (Omega group)

Our small team has made important impactful contributions to a large activity area

- JE as Software Co-Coordinator, MN as event I/O expert, SS as framework/C++ expert

Key objectives achieved for Run-3 which began in 2022

- **Multithreaded framework** in production with 3x reduction in memory use
  - Essential prerequisite for leveraging accelerators
- New **python based configuration** system for the C++ framework
  - Draws on 20 years of experience with python based C++ config

Full ATLAS software stack ported to **ARM**

- Successful physics validation of full simulation workflow
- Ready for large **HPCs based on ARM** (they are coming)
- Works on your Apple M1, M2 laptop/desktop too :-)

New **compact formats** to reduce disk storage footprint

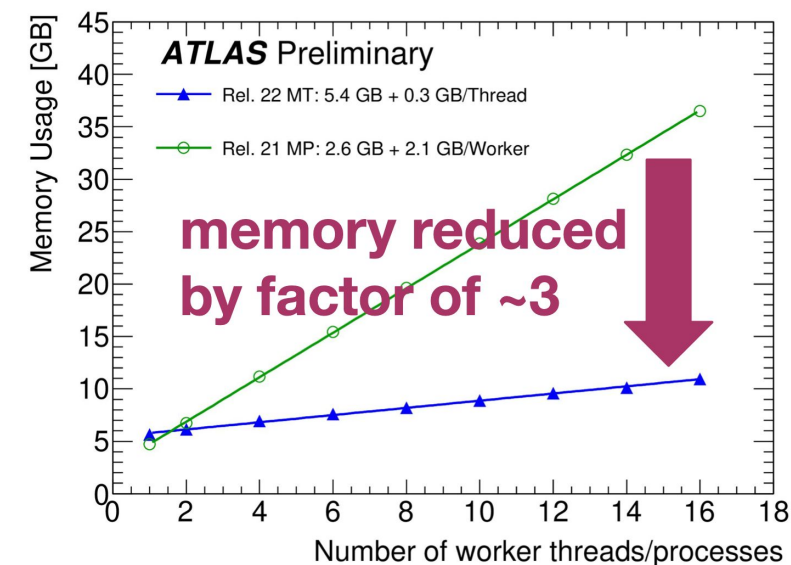
- Important to addressing HL-LHC storage challenge
- Almost all ATLAS disk storage is analysis formats

**RNTuple** prototype for analysis formats

- Expanded to full scale persistent storage technology backend
- Handles all Athena data types
- Collaborating with RNTuple developers on improvements/extensions

Integrating **AI/ML** capability in framework

- Framework support for ML enables ML based algorithms

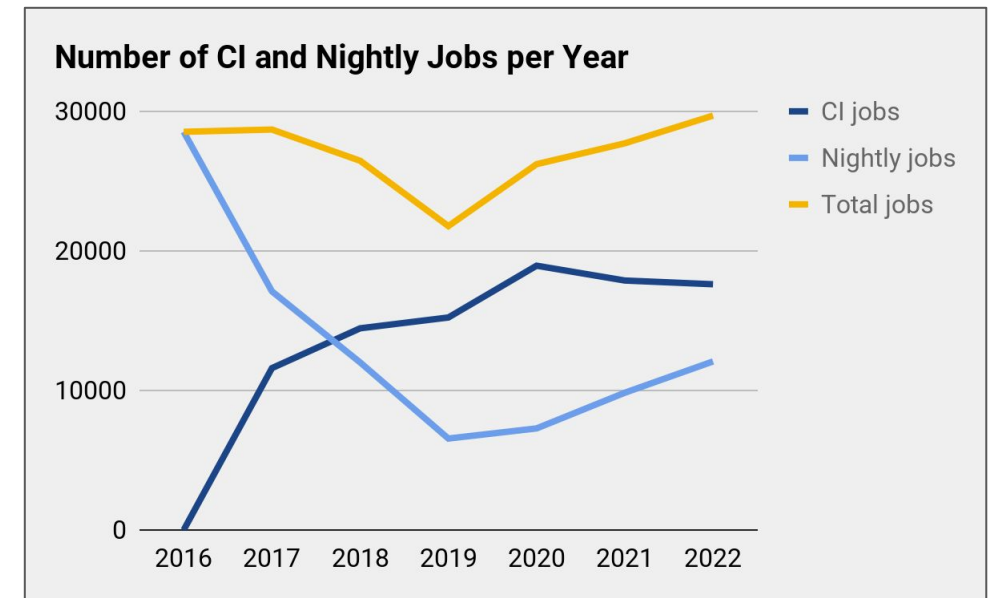
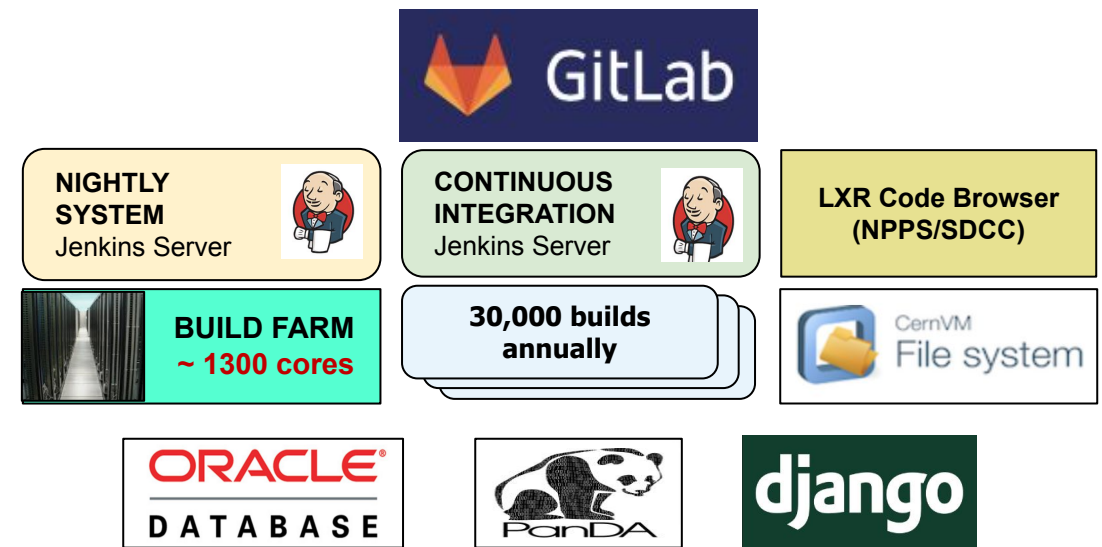


Reduction in memory utilization by switching to multi-threaded software

# ATLAS Software Infrastructure

Alex Undrus, Shuwei Ye

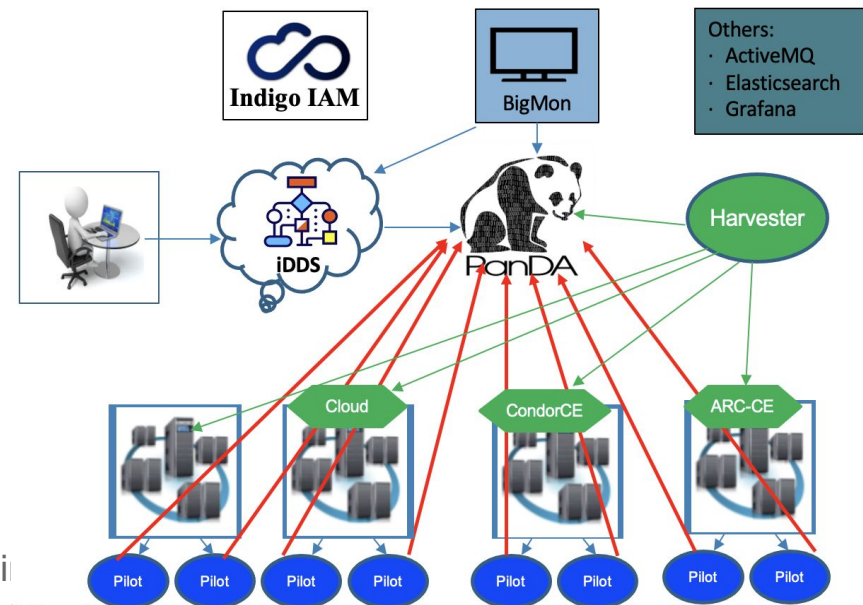
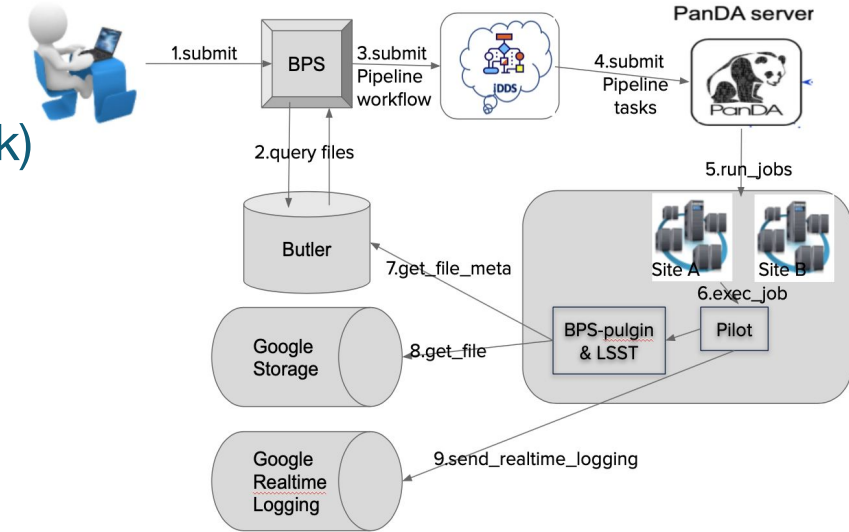
- Supporting 3000 participants worldwide
- CI builds for each GitLab MR
- 30 nightly branches
  - Testing code and tools updates
  - Migration to new platforms and compilers
- Local and larger-scale grid testing
- Runtime environments for all use cases
- Resource-efficient, intelligent operations
  - Hardware improvements, parallelization techniques resulted in 3x job acceleration
  - Operating costs lowered by node sharing between the CI and Nightly systems





# Rubin Development

- PanDA/iDDS improvements for Rubin's large scale DAG workflows
- Improved Rubin + PanDA integration
  - Rubin production software stack integration with PanDA
  - Scaling to multiple sites (most of the work required is on Rubin stack)
    - UK, France in addition to SLAC, Google
- Kubernetes based deployment at SLAC now in place
  - Based on PostgreSQL database
  - Performing scaling tests and preparing for pre-production
  - Will soon take over from CERN PanDA instance
- Improved user experience
  - Server performance, failure processing, latency reduction
  - Improved UI (command line + monitor)
  - Further tailoring the PanDA monitor for Rubin
  - Grafana based performance/health monitoring (CPU, walltime etc.)
- Campaign Management
  - Working on higher level interface to manage production campaigns



# Astrophysics experiments added this year

David Adams, Maxim Potekhin

## Dark Energy Science Collaboration (DESC) at Rubin Observatory

Computing operations: Study, optimize, improve production performance at NERSC

- Parsl-based scheduling on Perlmutter 256-CPU nodes
- Performed study of scaling/performance, with detailed monitoring
- Dramatic improvement in launch rates working with parsl and DESC developers
- But scheduling is a bottleneck for multi-node jobs; need multiple schedulers and something managing them
- Hence the **DESCprod** service in development
  - Intuitive user interface to configure, submit and monitor jobs
  - Schedulers run inside the service, parcel out tasks to workers

## LuSEE at Night

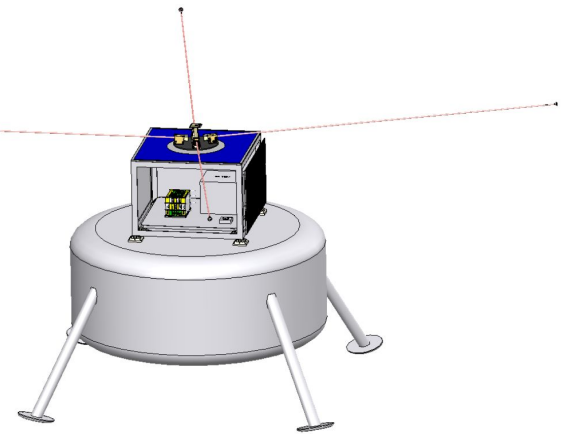
A NASA radio telescope to be landed in the quiet environment of the moon's dark side in 2025

- Galactic and extragalactic radio astronomy at very low frequencies (0.1-50 MHz)
- ESA's Lunar Pathfinder Communication Satellite to communicate with Earth

Contributing to several aspects of software

Experience flowing both ways - deriving benefit for other activities

- Implemented Python bindings for LuSEE C++ software
- Now informing investigating Python bindings for ePIC and eAST EIC C++ software



# DUNE

David Adams, Doug Benjamin (SDCC, DUNE data management co-lead), Lino Gerlach, Paul Laycock

## **dataprep** plug-in based mini-framework for initial processing of data

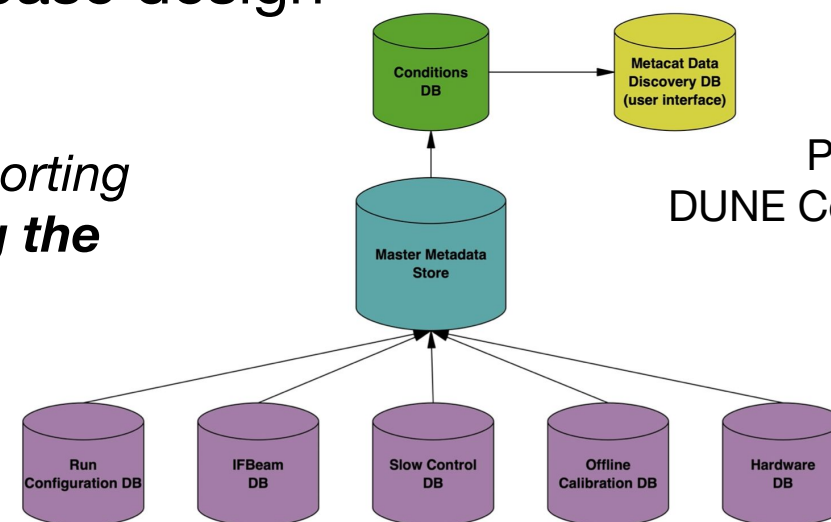
- Handle processing for the DUNE prototypes and full-scale detectors
- New context-switching tools, context is run conditions, e.g. specified by run number
  - Job configuration expressed with such tools can be used for different run conditions, e.g. prototype designs
- Design underway to support multi-threaded environment
  - One context manager instance per thread reporting back to parent in closeout
  - Includes support for stateful tools (e.g. filling counters or histograms)

## Conditions DB for Proto-DUNE: **NoPayloadDB**

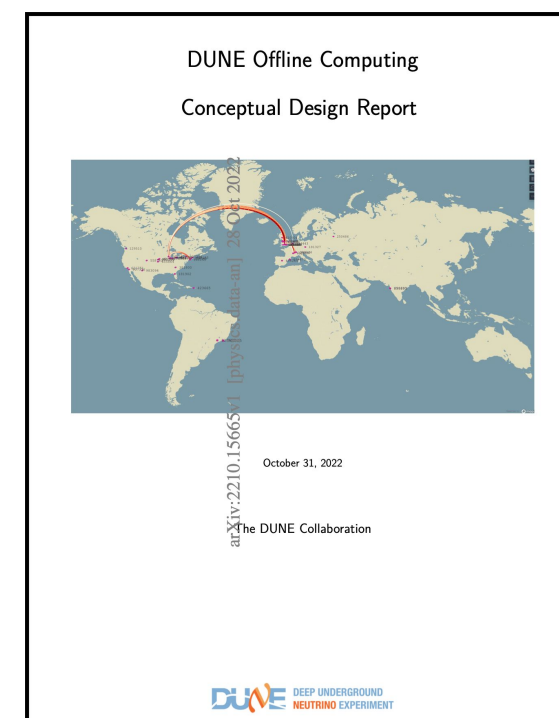
- Developing experiment agnostic client-side tool to communicate with NoPayloadDB
- Write DUNE-specific code in LArSoft service that uses this tool
- Deploy as solution, gather feedback & learn lessons for DUNE

## Part of the DUNE Database design

*SDCC (DB) collaborating in supporting  
**Data Management and running the  
Data Challenge***



Published in the  
DUNE Computing CDR



# Belle II

Paul Laycock, Ruslan Mashinistov, Cedric Serfon, Zhaoyu Yang

## Distributed data management: Rucio

Belle II is a major stakeholder in Rucio with the second biggest instance after ATLAS

- Bigger than CMS !
- Running on PostgreSQL, **O**(1 TB) database

Strong code contributions to the Rucio community

- Integration with (Belle)DIRAC as a FileCatalog plugin
- Simplification of monitoring ecosystem
- Chained subscriptions (automatically migrate new data to tape after copying to disk for processing)
- *Tape staging à la ATLAS Data Carousel is in progress*

Investigating Rucio as a primary metadata store

- Adds ~1kB per file, ~30% increase in DB size
- Read/write tests scale well beyond Belle II case
- End users poll a secondary copy in ElasticSearch

## Software infrastructure: From ATLASSIAN (JIRA, git stash, etc) to GitLab

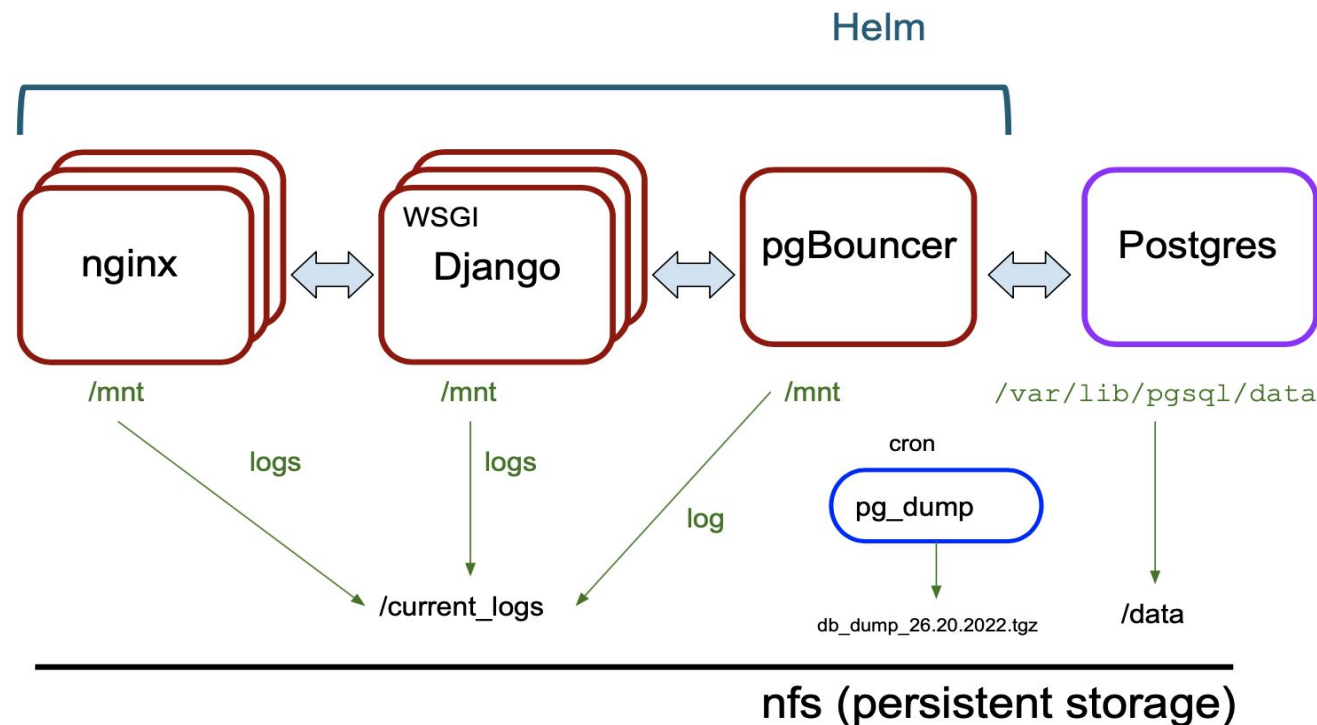
A new project Belle II gave us this year, to save licensing \$

- Non-trivial particularly for ticket migration, main repo successfully migrated before the holidays
- Migration enables CI/CD in particular automatic build and unit-tests that were not available in stash
- Big positive impact expected on development, code quality and speed up releases

# Conditions Database: From Belle II to sPHENIX ... and HSF Reference Implementation

Lino Gerlach, Paul Laycock, Ruslan Mashinistov, Dmitri Smirnov

- This year we generalized the CDB we did for Belle II to **experiment-agnostic NoPayloadDB**
- Leveraged useful discussions in HSF CDB activity, opportunity for HSF reference implementation
- In production for **sPHENIX**, use case is 200k cores running reco at BNL
  - **ProtoDUNE** will try this too
  - Expect to migrate **Belle II** to the same software
  - **Key4HEP** has also expressed interest
- SDCC hosts Belle II and sPHENIX CDB instances and is the calibration center of Belle II



- Postgres DB: persistent storage on nfs
- [pgBouncer](#): DB connection pooler
- NoPayloadDB: Django application running under [gunicorn](#) wsgi server
- nginx: web server
- Helm-defined universal deployment procedure, fully automated
- Horizontal build-out in back end provides scalability, under test
- Deployed on OKD at SDCC

# sPHENIX Tracking based on Acts

Joe Osborn

sPHENIX datataking begins this Spring!

Complete reco workflows worked out

- Alignment based on [Millepede2](#) + Acts
- Full integration test circa March (yes it's tight :-)

Memory and timing constraints under control

- Bringing these within budget one of the major challenges

Ongoing efforts on improving performance

- e.g. moving TPC seeding to operate in phi-z space instead of phi-eta for significantly improved off vertex track reconstruction in streaming readout mode
  - Triggered readout in 2023, streaming in 2024

Integrating with online data collection

Developing commissioning tools

- e.g. K0s for testing calibration/alignment

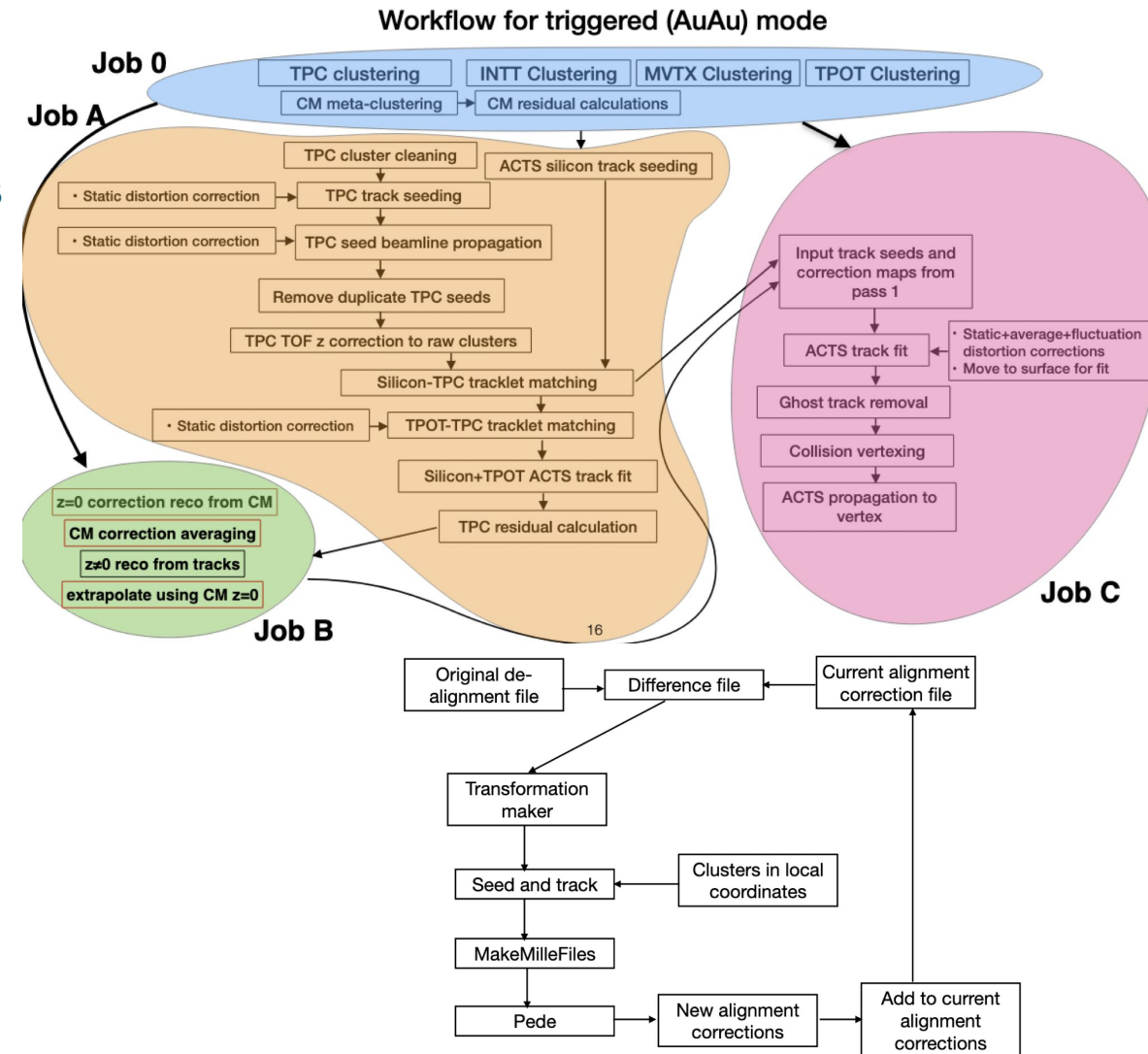
Work with Acts team on new tools/features

- Gaussian Sum Filter, KD tree based seeder, ...
- Deployment strategies for an operational experiment!

Most recent work focused on commissioning!

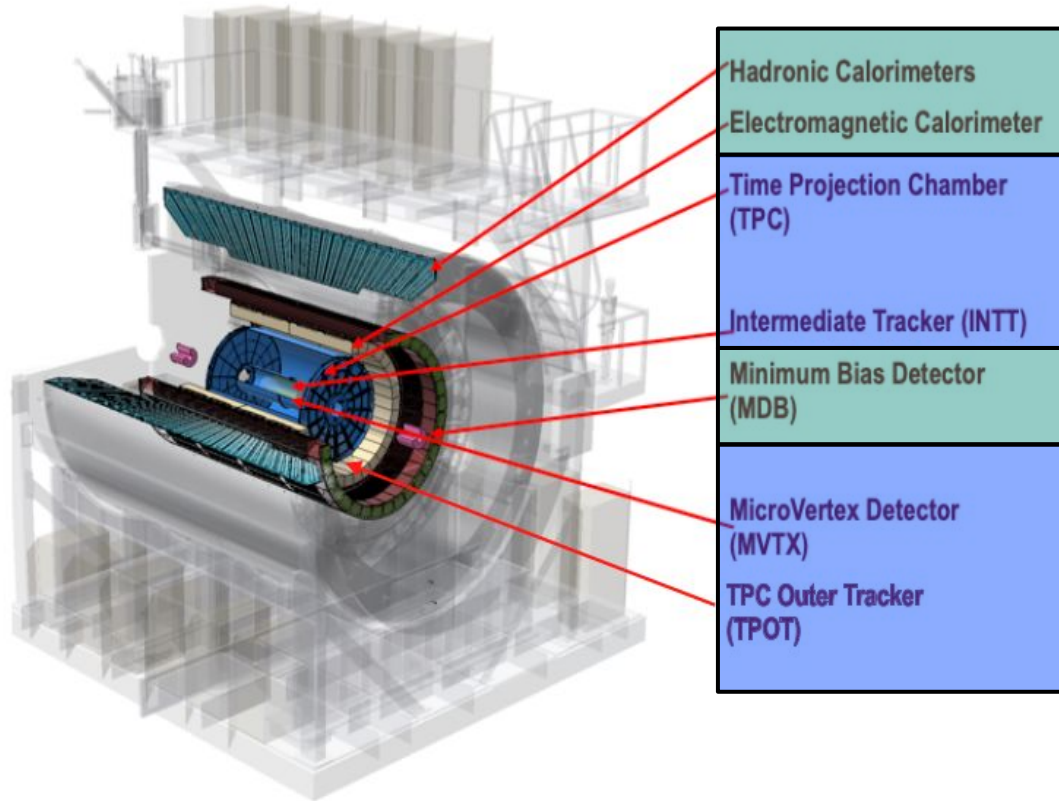


TPC installation this week!



# sPHENIX Processing

Jason Webb, Wen Guan, Tadashi Maeno, Xin Zhao



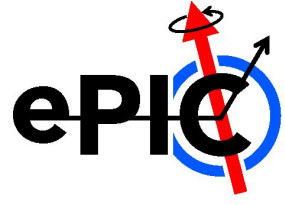
## sPHENIX 3 year run plan

- 2023: Commissioning and AuAu physics
- 2024: pp and pAu reference data
- 2025: High statistics AuAu run

- Mix of **triggered** (calorimeters and minbias detector) and **streaming** (tracking) readout.
- Online: Events saved across ~60 files per run
- Offline: Read event data sequentially from files to build the events for reconstruction
- **Reconstruction to be performed in near real time with data taking, from buffered data files prior to tape archival**
- NPPS has mapped sPHENIX reconstruction chain onto PanDA's flexible DAG-described workflows using SHREK
  - Discovers new data in Rucio and dispatches reconstruction
  - Full simu (evgen through reco) workflows demonstrated
- PanDA instance installed in SDCC and under test
  - Based on k8s (OKD) and PostgreSQL
  - Awaits production-level PostgreSQL
- Rucio instance in place, integration in progress

# EIC Software

Kolja Kauder, Joe Osborn, Maxim Potekhin



## EIC now has a detector collaboration, ePIC!

## 2022 was a busy year for EIC S&C

- Proto-collaboration phase ended with reference design in March
- Aggressive schedule: a full simu/reco stack in October
  - Impressively rapid, yet thorough and harmonious, software decision process during Spring/Summer (cf. the last round table)
  - Guided by software principles developed and endorsed by ~all
  - Followed by a software stack development blitz
  - First simulation campaign with ePIC stack launched in November
- Will feature strongly in the JLab talk I'm sure, won't preempt it
- BNL contributed to all of the phases with more to come
- Less urgent aspects postponed to discussion this year
  - Conditions database
  - Distributed computing tools (workflow and data management)
- Ongoing collaborative tools work supporting sw and EIC generally
  - Websites, mailing lists, calendars, collaborative storage, wikis, youtube

		Discussion topic(s)	Decision topic(s)
May	4	A/WG	
	11	Transition Period	Present procedure. Decide on list and order of decision topics
	18	No meeting (Streaming Readout X Workshop)	
	25	Code Repository	Repository: - Location (GitHub, GitLab+Host) - Admins - Access
Jun	1	Discussion Schedule	Schedule: - Decide most critical decisions to make before July 27th EICUG meeting - Schedule of topic discussions
	8	Geometry	Geometry: - Package (e.g. DD4HEP)
	15	Data Model	Data format - Generated events - Simulated data - Processed data (e.g. ROOT w/ specific tree format)
	22	Data Model	
	29	Reconstruction Framework	Reconstruction Framework - Package
Jul	6	Reconstruction Framework	
	13	Data and Analysis preservation	Data Preservation - What is preserved (simulated, DSTs, ...) - Location(s) - Access (S3, xrootd, rucio, ...)
	20	Documentation	Documentation: - Location of User documentation (wiki, repository,...) - Who will set up skeleton with list of topics (e.g. "Getting s
	27	EICUG Meeting	
Aug	3	Continuous Integration	Continuous Integration
	10	Containerization Official buids	Containerization - platform (Singularity, Docker, multi, ...) - Supported OSes - Location of images (e.g. cvmfs) Official builds - Location (e.g. cvmfs, container image, ...)
	17	Calibration DB Conditions DB	Calibration / Conditions DBs - Package - Server/Host - Access
	24	Calibration DB Conditions DB	
	31	Distributed Campaign Workflow	Distributed Campaign Workflow - Package (DIRAC, PanDA, STAR(?), ...)
Sep	7		
	14		

# Data and Analysis Preservation

Maxim Potekhin, Kolja Kauder

## Drawing on the CERN/community software/services stack

- HEPData, REANA, CERN Open Data Portal, Zenodo, ...

## PHENIX

- Direct photons analysis was successfully ported to REANA (using the BNL SDCC instance)
- HEPData packages published on the CERN portal at a healthy rate of ~20 per year
- Migration of PHENIX documentation to Zenodo reached a milestone – more than 600 items committed
  - All PhD theses and majority of all conference presentations for the past 6 years

## FAIR

- A new Research Coordination Network is being established with NSF support – FAIROS-HEP
- Will develop shared policies and cohesive infrastructure around data and publications
- Central concept is a “living publication” to preserve and extend physics results
- Will participate in organizational workshop at CERN next month

## sPHENIX

- Mandated by reviewers to give early attention to DAP
- Their attention is understandably occupied at the moment

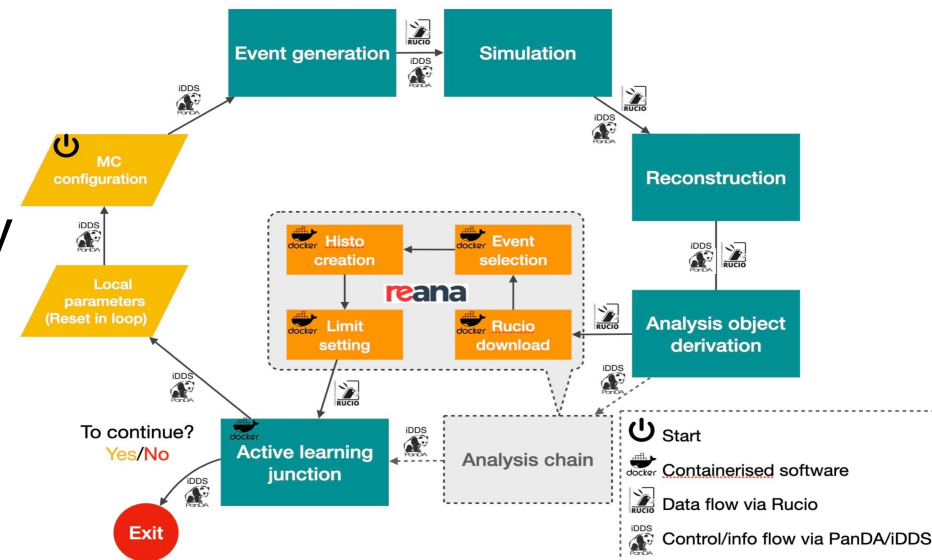
## EIC

- DAP one of the areas addressed in the (now) ePIC software planning process
- A determination to give it early attention

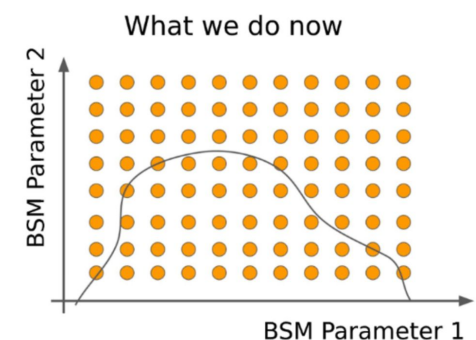
# Large Scale Complex Workflows

Wen Guan, Alexei Klimentov, Tadashi Maeno, Paul Nilsson, Christian Weber (Omega group), Torre Wenaus

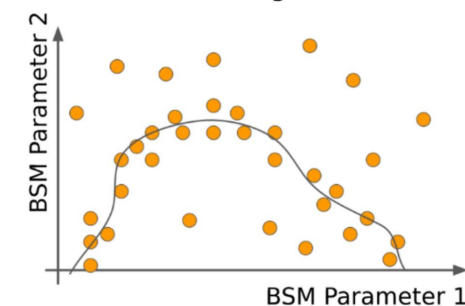
- Ready access to diverse large scale resources can greatly accelerate developing processing-intensive applications
  - Shorten turnaround times by orders of magnitude
  - **Expanded scope for scientific creativity in developing applications**
- We are developing such services with PanDA/iDDS
  - **Hyperparameter optimization service** in production use for ATLAS fast simulation (GAN based calorimeter simulation FastCaloGAN)
  - An adaptation of this service with the same iterative refinement structure uses iterative regression to efficiently calculate a limit surface by rapidly ‘learning’ where the surface is, a.k.a. ‘**active learning**’
  - NPPS and Omega groups at BNL are applying the active learning service in the  $H \rightarrow ZZ_d \rightarrow 4\ell$  dark sector analysis, pub notes in progress
    - Greater efficiency, scalability, automation enables a wider parameter search (instead of 1D, 2D or even 4D on large scale resources) and improved physics
- Working to generalize the services
  - **Use ATLAS work as a springboard for developing tools useful to the broader community**
- Exploring a new EIC use case: AI assisted detector design using Bayesian Optimization
- Plan collaboration with CSI on using funcX send work to LCFs
  - Such workflows can effectively leverage these GPU based machines
  - A previous Round Table on workflow management helped seed this



Active learning with PanDA/iDDS + REANA



What active learning can do for us



Active Learning via iterative regression on a limit surface

# Diversity, Equity, Inclusion, Community

## Diversity, Equity and Inclusion

- Incrementally improved our still poor gender diversity, a long way to go in improving diversity in general
- Important to growing diversity: growing the job candidate pool
- We resolved (with the support of HR) to be more accommodating of working history gaps (time away from workforce), which has already paid off very nicely
- BNL's covid-influenced policy of allowing fully remote work (besides CERN, US only) expands our recruiting pool, as does our strong presence at CERN

## Mentoring and training

- S&C speaker contributions to the Physics Dept Summer Lecture Series
- We will participate in two university-led projects from the recent DOE Computational HEP Training call

## Community

- Many and ongoing contributions to the HEP Software Foundation (HSF)
- This [Software and Computing Round Table](#)
  - Encouraging knowledge transfer and common projects
- [Future Trends in Nuclear Physics Computing](#) workshop series
  - HEP content as well as NP
- CHEP organization contributions: Program Committee co-chair, International Advisory Committee members, session conveners



# 2023 Outlook

- **sPHENIX datataking** from ~April
  - Tracking, conditions database, production system in real-data production
- Intensive development program for **HL-LHC readiness** is ramping up (even as HL-LHC start date recedes!)
  - R&D demonstrator program towards technology decisions for HL-LHC Computing TDR
- **EIC** has entered an exciting new phase with **ePIC** established
  - BNL can't afford much software effort (bosses tell us) but we will make the most of what we have
  - Drawing on our other activities and expertise
  - Importantly, drawing on sPHENIX experience as we accrue it
- New opportunities? DESC and LuSEE emerged this year...
- **Analysis** tools focus: fast, (pseudo)interactive, leveraging community tools
- Building more **AI/ML** and complex workflows experience and activity
  - Services for large scale distributed complex workflows, AI/ML and related
  - HEP centric to date; hope to extend to AI assisted EIC detector design
- Building **data & analysis preservation** as an integral part of the experiment life cycle
  - Hard to get experiments' attention when data is flowing in, even though it's a direct benefit to analysis today
  - The powers above us continue to press its early importance

# Many thanks

Many thanks to those in the BNL Physics Department and elsewhere who contributed to this talk, and many more who have contributed to the work described

An inevitably partial list of contributors:

David Adams, Fernando Barreiro Megino, Doug Benjamin, Kaushik De, John De Stefano, Johannes Elmsheuser, Vincent Garonne, Wen Guan, Chris Hollowell, Jin Huang, Eddie Karavakis, Kolja Kauder, Alexei Klimentov, Eric Lancon, Paul Laycock, Meifeng Lin, Tadashi Maeno, Ruslan Mashinistov, Paul Nilsson, Marcin Nowak, Joe Osborn, Louis Pelosi, Victor Perevoztchikov, Chris Pinkenburg, Maxim Potekhin, Ofer Rind, Cedric Serfon, Dmitri Smirnov, Jason Smith, Alex Undrus, Gene Van Buren, Brett Viren, Jason Webb, Tony Wong, Shuwei Ye, Zhaoyu Yang, Xin Zhao

# Supplementary slides

# OKD Cluster Details

- Each cluster running OKD 4.10, and is provisioned with
  - 7 Dell R640 Servers
    - 3 HA control plane nodes, 4 worker nodes
      - Running Fedora CoreOS (FCOS) 35 deployed via OKD Installer-Provisioned Infrastructure (IPI)
        - CRI-O used as container runtime
    - Specs:
      - 2x Xeon Silver 4210 CPU @ 2.20 GHz
      - 128 GB RAM
      - 4x 25 Gbps NICs
      - 3x 480 GB SSDs
  - NetApp A250 Storage Appliance
    - 14 x 1.92 TB NVME drives (~26 TB raw)
    - ONTAP NetApp OS allows dynamic PV provisioning via Trident



*ATLAS OKD Cluster Hardware*

# Website Developments at SDCC

## New US ATLAS website

- <https://www.usatlas.org/>
- Drupal CMS with SDCC Keycloak, BNL and CERN based authentication
- Public and private areas for documentation

## sPHENIX MediaWiki updated

- SDCC Keycloak login with automatic role placement
- Public and private areas in conjunction with roles

## Additional Drupal sites

- <https://www.cosmo.bnl.gov/>
- <https://www.quantastro.bnl.gov/>
- <https://www.sphenix.bnl.gov/>

## Future developments

- Internal SDCC documentation site based on Jekyll and Git
- rsnapshot deployment to collect multiple VM backups to a single location for tape backup



About Get Involved Latest Activities Collaboration Resources Log in



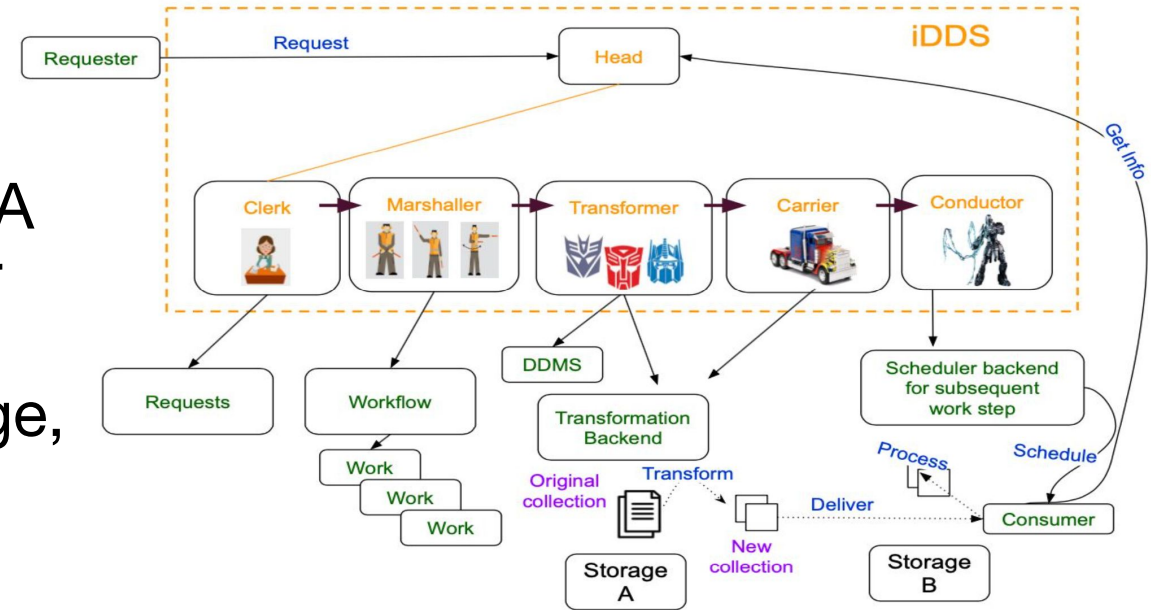
# Intelligent Data Delivery Service (iDDS)

iDDS is an experiment-agnostic add-on to PanDA (or other workload manager) supporting granular data delivery and **orchestration of complex workflows** that are efficient in their use of storage, network and processing resources

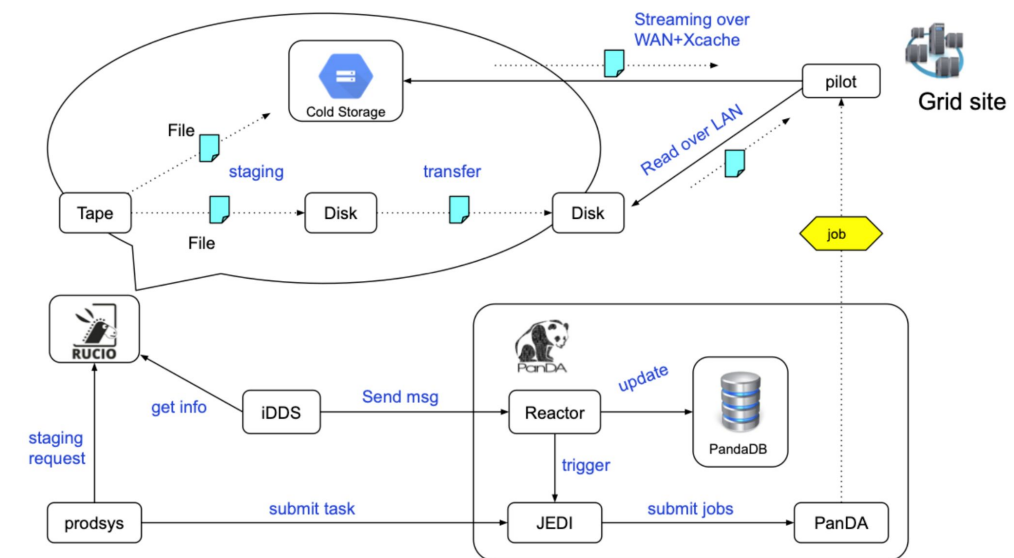
- A joint project with IRIS-HEP (NSF), project hosted by HSF
- Used by ATLAS, Rubin, sPHENIX

Used in a growing list of applications important for HL-LHC readiness and serving/scaling analysis

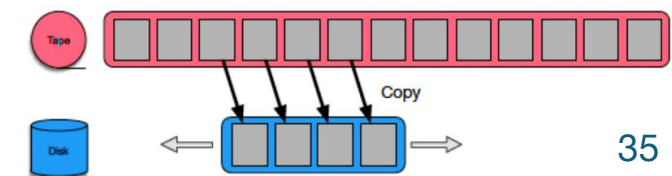
- **ATLAS Data Carousel** processes tape-resident data using a small disk storage footprint via a sliding window orchestrated by PanDA, iDDS and Rucio
  - In production for almost 2 years, reducing the storage needs of analysis object data, the dominant storage load for HL-LHC
  - Ongoing R&D to reduce the footprint and improve performance
- **Highly scalable ML services**
  - Enable analysts to run processing-intensive AI/ML applications on large scale, geographically distributed, heterogeneous resources
  - Shorten optimization and training latencies by orders of magnitude
- **Active learning services** drawing on the ML work



Intelligent Data Delivery Service (iDDS)



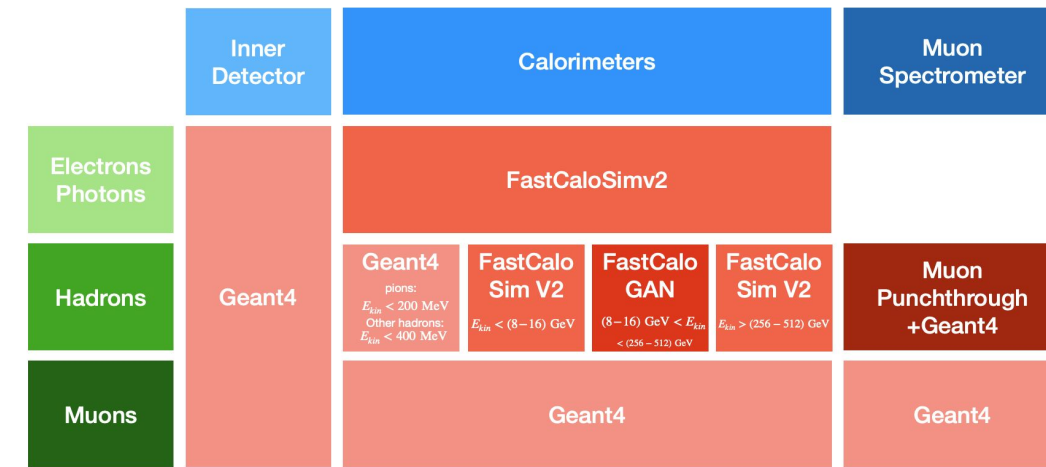
ATLAS Data Carousel using iDDS



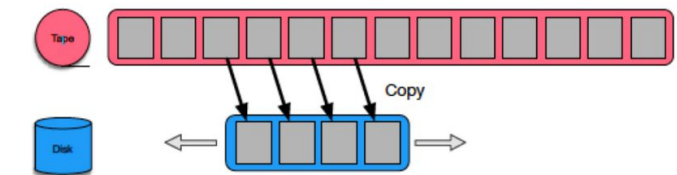
# Future R&D Towards HL-LHC and Beyond

We are leaders in the US ATLAS HL-LHC R&D program with much progress made, but many remaining challenges

- ML's new presence in the **fast simulation** chain, FastCaloGAN in AtlFast3, is a first step
  - ML is already processing intensive: 100 GPU days to train
  - With 90% of HL-LHC era simulation to use the fast chain, further investment will be rewarded (and help attract the physics groups)
- The success of the **Data Carousel** in economizing disk storage by processing a sliding window of tape-resident data is seeding new R&D to make the workflow still more storage efficient
- The growing importance of **streaming DAQ** often incorporating **AI/ML early in the processing** pipeline will drive R&D activity
- Supporting **AI/ML applications on large scale facilities** offers the most effective means of using DOE's largest **supercomputers** at a substantial scale
- The **software reengineering** program is yielding better, more efficient as well as GPU-capable software
- BNL-led joint R&D projects with Google and Amazon are advancing the capabilities of **analysis in the cloud**



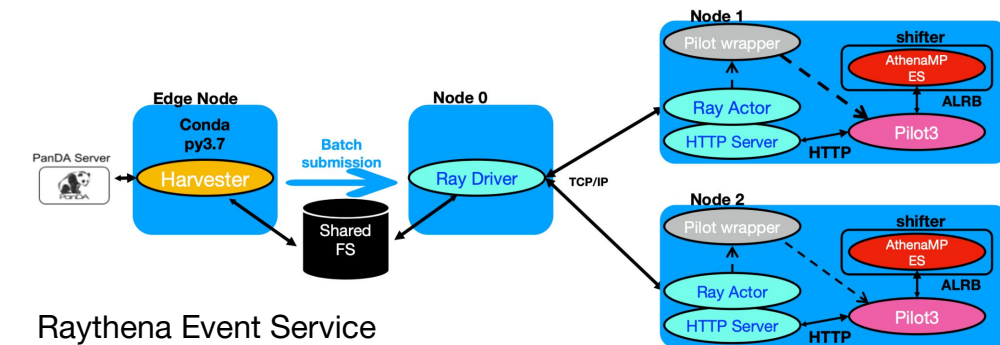
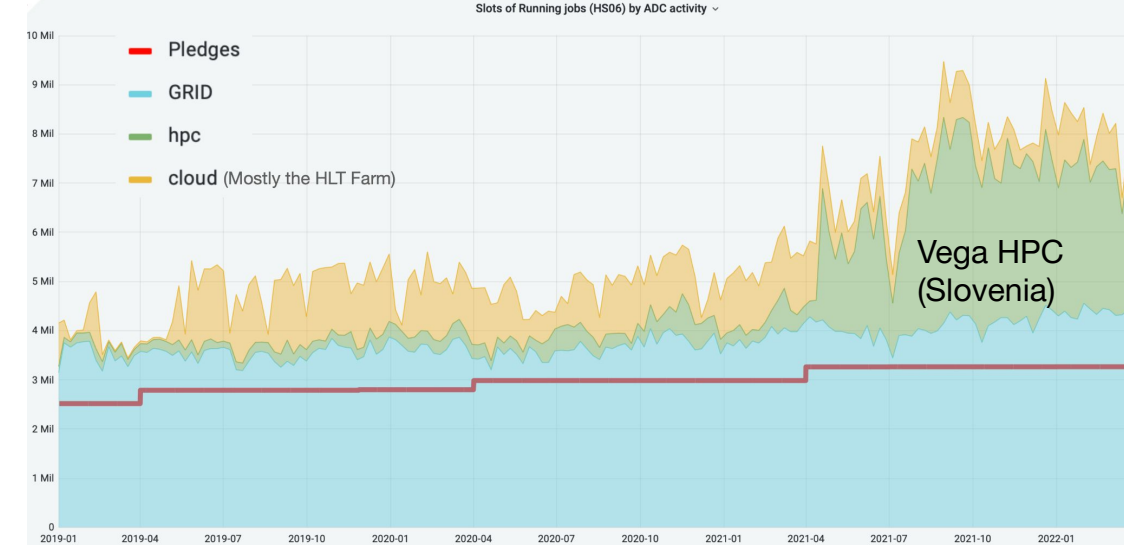
AtlFast3 components



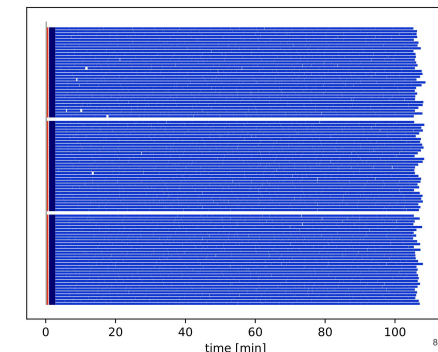
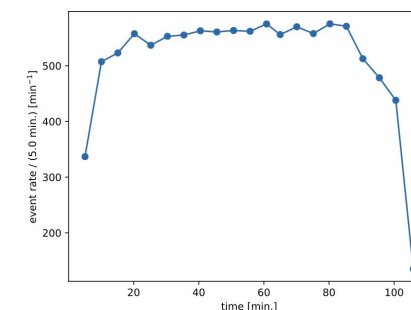
Data Carousel concept

# HPCs in ATLAS

- A decade of ongoing effort to use HPCs well
- BNL-led BigPanDA ASCR project 2012-2018
  - experiment-agnostic capability targeting HPCs, including DOE Leadership Computing Facilities (LCFs)
  - Oak Ridge's Titan in ATLAS PanDA production
  - Extended PanDA Titan usage to other science domains
- Event Service, PanDA's innovative mechanism to use HPCs with full node efficiency
  - BNL-led development, first demonstrated at CHEP 2013
  - Latest HPC version being commissioned at LBNL NERSC
- ATLAS uses HPCs extensively, drawing on PanDA's strong support and scalability
  - Dramatic recent example: Vega in Slovenia
- Current GPU-dominated LCF generation not used with ATLAS workflows
  - We don't have GPU-capable production workflows
- Rather, we focus on what these machines are increasingly optimized for: AI/ML
  - They are an important target of our large scale AI/ML services work
- And, we have ported the ATLAS software to ARM
  - Ready for ARM based HPCs
  - Already machines in Asia, coming soon in Europe and potentially US



Event-level processing management ensures each core is used fully






# ATLAS Google R&D Phase 1

- Initially 10 projects reviewed and approved at different scale
  - Resources, timeline and personpower estimation – done / in progress
  - Technical set up (including RSE and PanDA queues, CVMFS, Jupyter notebooks and Dask)

Project	Responsible	Presentation	<extra> Resources
MC sample production for AL workflow	Z.Bhatti	<a href="#">Aug 24, 2022</a>	SW development was discussed at WFM SW <u>weekly</u> (Sep 15th)
Distributed ML	J.Sandesara	<a href="#">Aug 24, 2022</a> <a href="#">Sep 28 2022</a> <a href="#">Nov 2, 2022</a>	Sign off Sep 14 (postponed for 2 weeks) Additional big memory node has been added to <u>jupyter</u> cluster
S3 gateway	A.Hanushevsky	<a href="#">Aug 31, 2022</a>	500 GB + CPUs
Compact data formats, DASK	N. Hartmann	<a href="#">7 Sep. 2022</a>	O(100 TB) storage, O(1-10k) CPUs
BQ for <u>PanDA</u> WMS and Analytics	M.Grigorieva	<a href="#">14 Sep 2022</a>	Technologies should be demonstrated at small scale first

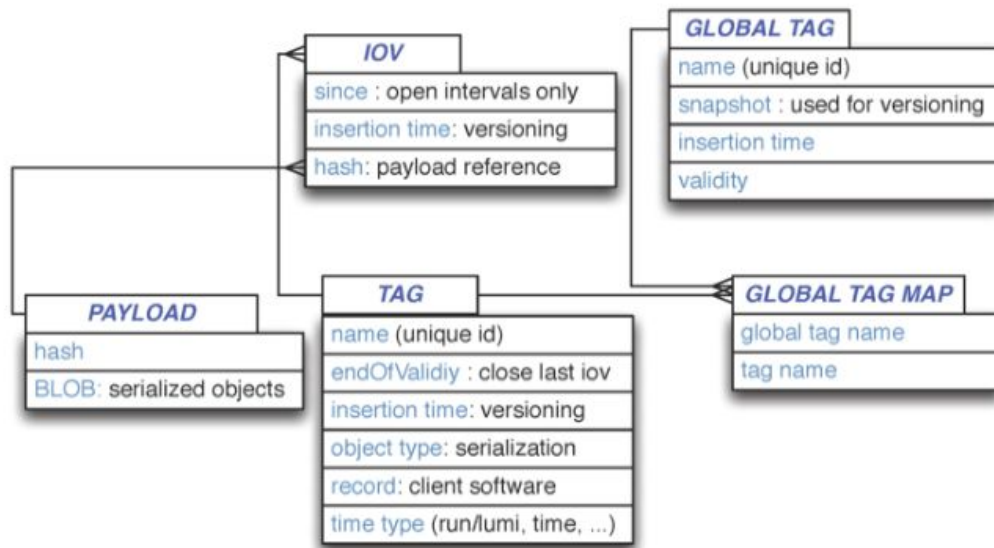
Project	Responsible	Presentation	<extra> Resources
aCTS CPU x86/ARM	A.Salzburger	<a href="#">Sep 14, 2022</a>	
Big Nodes and ARM	<u>J.Elmsheuser</u>	<a href="#">Sep 14, 2022</a>	Big Node set up @GCP : 48 CPUs, 250 GB disk, 192 GB memory
Analysis Facilities in the Cloud	L Heinrich	<a href="#">Sep 21 2022</a>	
<u>PanDA</u> /Desk Integration	P Nilsson	<a href="#">Sep 29 2022</a>	
Active Learning with <u>pchain</u> and <u>iDDS</u>	<u>R.Zhang</u>	<a href="#">Nov 9, 2022</a>	<u>PanDA</u> analysis queue

# ATLAS Software Infrastructure & User Support

- ❖ **asetup** environment configurator used by all interactively and in grid batch jobs
  - Running reliably with no interruption
  - Added regression tests in different containers on CVMFS;
  - Improved to be consistent and proper release version sorting between python2/3
- ❖  **code browser**
  - Transparent migration from RHEL6 to RHEL8
  - Implementation of **Universal Ctags** (replacing out-of-development Exuberant Ctags).
- ❖ **ML container images**  
  - Developed and optimized different types of ML Docker images (including tensorflow-gpu) with a new package/env manager **micromamba**.
  - Test deployment onto CVMFS in Singularity sandbox at BNL/CERN
- ❖ **dCache & Xcache** support at BNL
  - Improved **pnfs\_ls.py**, **Xcache\_ls.py** scripts to adapt to infrastructure evolution

# HSF Recommendations for Conditions Data access

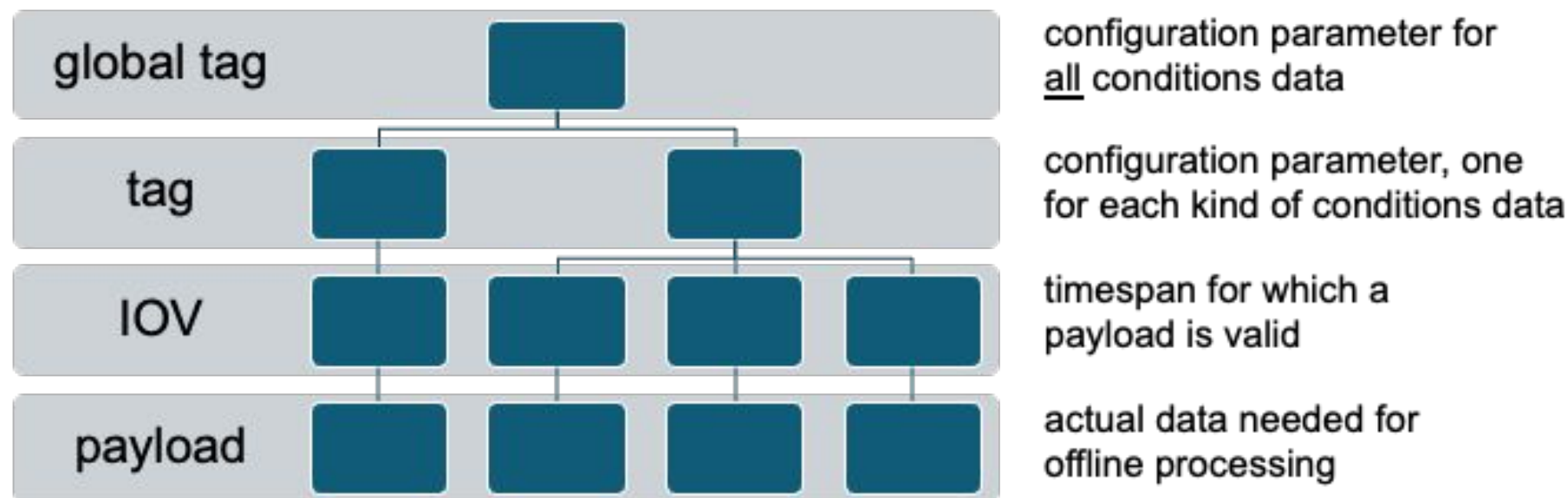
- From the HSF Conditions Data activity:  
<https://hepsoftwarefoundation.org/activities/conditionsdb.html>
- Key recommendations for conditions data handling
  - *Separation of payload queries from metadata queries*
  - *Schema below to enable appropriate configuration*



Both **CMS** and **Belle II** use similar (though more complex) schema today

# HSF Recommendations – Key Concepts

How to get from global tag to actual conditions data:



**Possible user API for conditions data service:**

```
CondSvc.setGlobalTag(<GlobalTagName>);
```

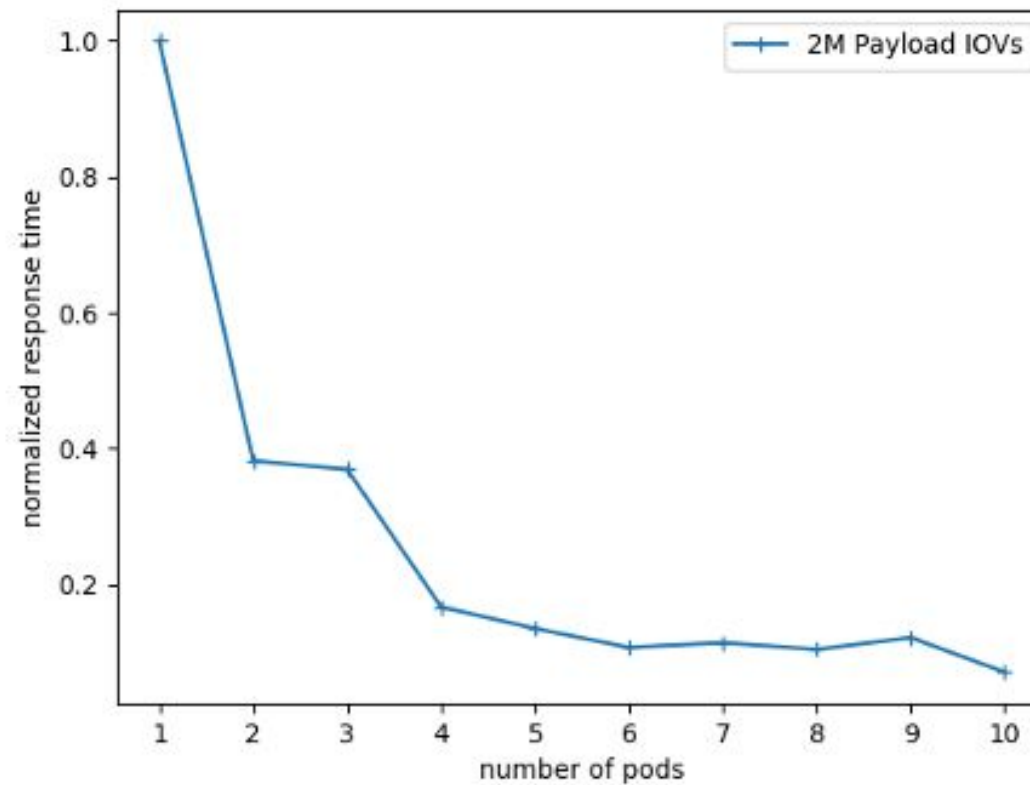
Once per job

```
CondSvc.get(<MyConditionsType>, <RunNumber>);
```

In each module

# Scaling Tests for NoPayloadDB conditions database

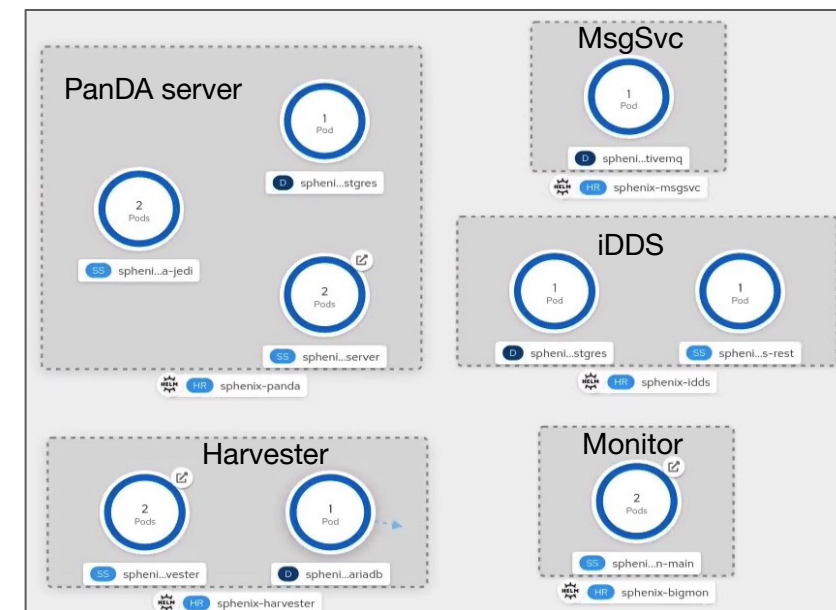
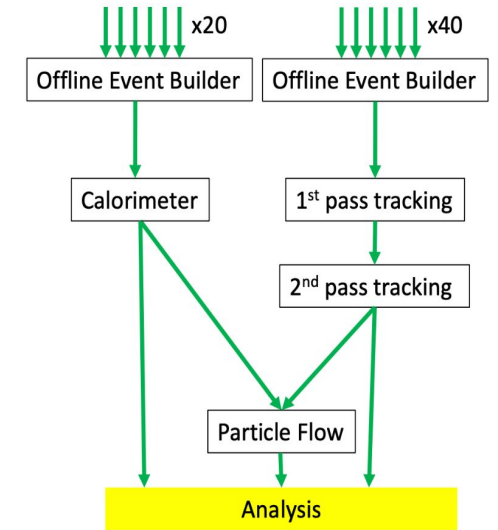
- Simulate expected access patterns in HTC environment
  - Submit 100 HTCondor jobs making 100 requests each
  - Investigate scaling of response times w.r.t. different parameters



# sPHENIX Distributed Computing

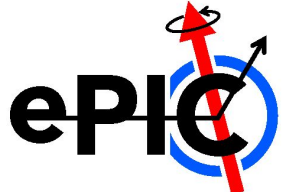
The Simply Handy Remote Execution Koordinator (SHREK) will orchestrate the near real-time production of the sPHENIX data.

- **donkey**: Identifies newly cataloged datasets (runs) as they become available in Rucio and dispatches reconstruction workflows.
- **shrek**: Maps sPHENIX reconstruction tasks onto PanDA workflows, assembling them from simple YAML files specifying the input data sets, the codes to be run, and the execution environment.
- Simulation-based workflows (evgen to reconstruction) demonstrated
- Data production monitoring to be implemented
- BNL PanDA instance installed in SDCC and under test
  - Aided by similarity to SLAC PanDA: k8s (OKD) and PostgreSQL
  - Still awaits production grade PostgreSQL back end
  - Using CERN instance in the meantime (as for Rubin)
- Rucio instance also in place, integration in progress



PanDA OKD components

# HepMC – the Glue that Binds All MC



Kolja Kauder

## Generate:

- For many, even recent, EGs, transformation via eic-smear remains the only way

## Store:

- HepMC as ROOT now supported in eic-smear

## Fine-tune:

- Secondary decays are supported but need careful handling – in creation and GEANT treatment

Studies for proposal	ATHENA	ECCE
DVCS in ep	EpIC	MILOU3D
DVCS (incoherent) in ed	EpIC	
DVCS in He-4		TOPEG
TCS in ep	EpIC	EpIC
$J/\psi$ in ep		eSTARlight, IAgar
$J/\psi$ in eA		eSTARlight, IAgar
$\phi$ in eAu/Pb	SARTRE, BeAGLE	SARTRE, BeAGLE
$\Upsilon(1S, 2S, 3S)$ in ep	eSTARlight, IAgar	
u-channel: $\omega, \rho$ in ep	eSTARlight	
$X, Y, \psi(2S)$ in $ep \rightarrow J/\psi \pi^+ \pi^- p$	elSpectro	elSpectro
Pion Form Factor		DEMPgen
Pion Structure Function		EIC_mesonMC
$A_1^n$ (He-3 double tagging)		DJANGO

Some of the generators used for EIC physics

## Read in:

- Treatment of multiple vertices (e.g. for background studies) in dd4hep needed fixes

kkauder commented on Nov 14, 2022 · edited by andresailer

Contributor

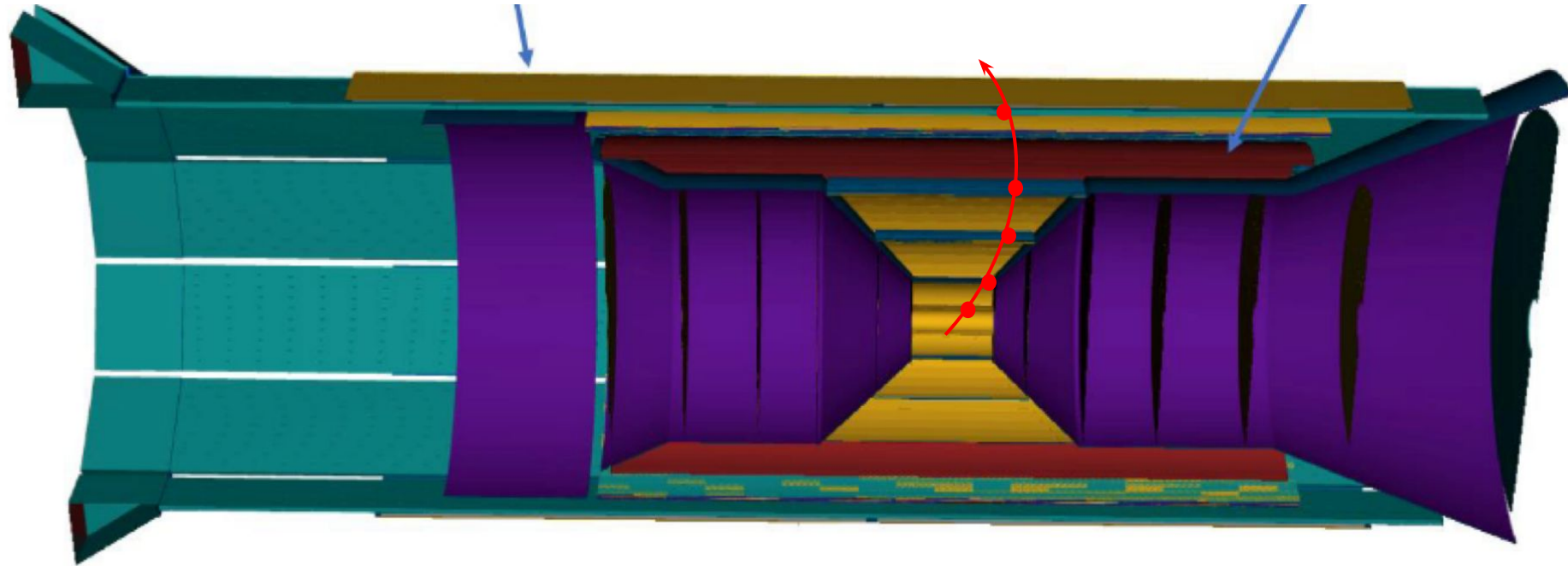
...

BEGINRELEASENOTES

- Final state HepMC particles were all attached to (0,0,0). Fixed by switching vertex creation for parentless particles to using their end-point instead, fixes [Final state particles in HepMC are forced to vertex 0,0,0 #1013](#)

ENDRELEASENOTES

# ePIC Track Reconstruction



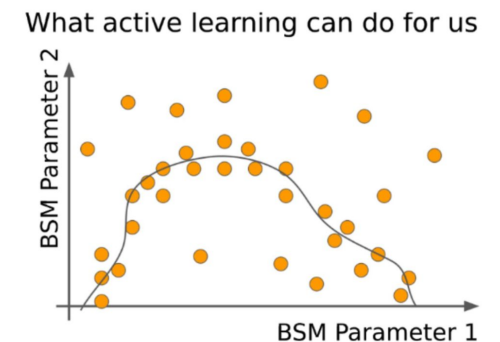
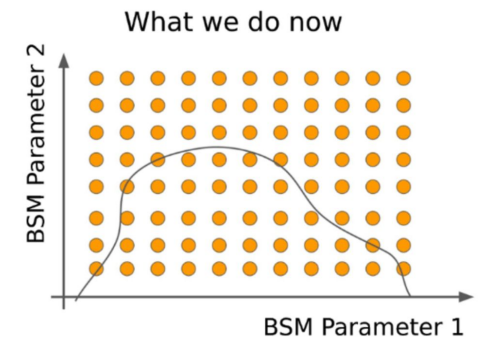
Major issue in ePIC track reconstruction is that we are still using truth seeded tracks for the `Acts::CombinatorialKalmanFilter`

Ongoing work to implement `Acts::OrthogonalSeedFinder` into `ElCRecon`

Discussions recently started on seeding strategies and longer term track reconstruction plans

# Applying ML services via active learning in analysis

- The hyperparameter optimization (HPO) service we developed is in production use for FastCaloGAN, part of the production ATLAS fast simulation AtI Fast3
- In working with the analysis community to find the next large scale application of the HPO service, what emerged was active learning, algorithmically a close relative
- The active learning technique we're using was developed by our ATLAS NYU colleagues
  - [“Excursion Set Estimation using Sequential Entropy Reduction for Efficient Searches for New Physics at the LHC”](#), Kyle Cranmer et al, ACAT 2019
  - in calculating an iso-contour surface  $f(x)$ , conventional approach uses a grid with a sampling density not informed by the unknown  $f(x)$
  - instead, use an iterative approach, using information about  $f(x)$  from previous evaluation cycles to sample parameter space more efficiently
  - find an iterative algorithm that suggests points to evaluate that help the most in finding the contour
  - Kyle and colleagues found a computationally efficient one
- In response to interest from analyzers (in particular that team), we adapted our ML hyperparameter optimization service to serve this similar iterative refinement algorithm
- The entire workflow from event generation -> simulation -> reconstruction -> derivation -> limit setting analysis and its iterative refinement loop is implemented and automated using PanDA and iDDS
  - It employs grid and REANA (Reproducible Research Data Analysis Platform) processing resources
- Modular and containerized
  - Analysts provide components specific to their analysis
- Now exploring a new EIC use case: AI assisted detector design using Bayesian Optimization



Active Learning via iterative regression on a limit surface

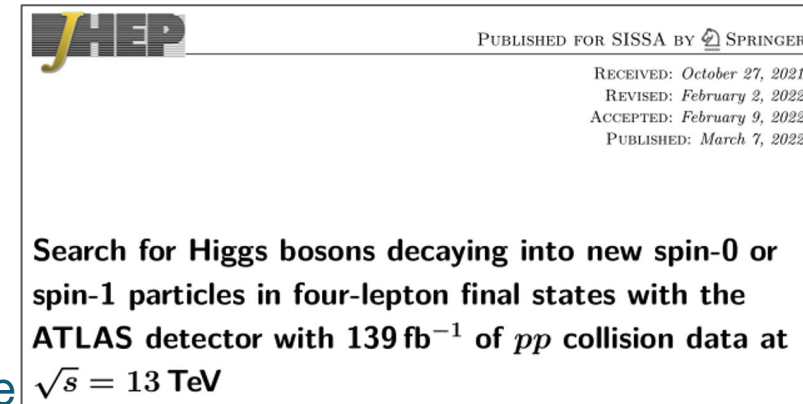
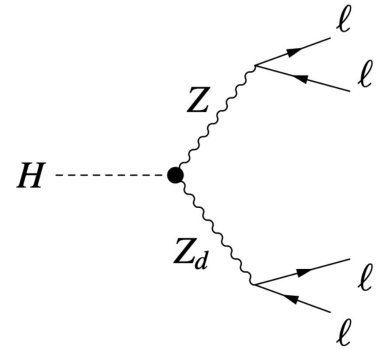
# Active learning service in the $H \rightarrow ZZ_d \rightarrow 4\ell$ dark sector analysis

Earlier this year ATLAS published a search for  $H \rightarrow Z_d Z_d \rightarrow 4\ell$  and  $H \rightarrow ZZ_d \rightarrow 4\ell$ , a Standard Model extension with a dark vector gauge boson

Due to limited resources the analysis allowed only one free parameter ( $Z_d$  mass)

The automation and processing scale offered by the active learning service can enable extracting limits with more free parameters

- Active learning in the ( $Z_d$  mass - kinetic mixing parameter) 2D plane is being explored now
- The  $ZZ_d$  progenitor mass is also a parameter of interest, may differ from the Higgs, may be incorporated
- The present study focuses on  $H \rightarrow ZZ_d$ ; with active learning we could explore joint  $H \rightarrow ZZ_d$   $H \rightarrow Z_d Z_d$  limit setting with the full 3D phase space of  $Z_d$  mass and mixing parameters



# ML in sPHENIX

Maxim Potekhin

The first ML application developed in sPHENIX was created for signal feature extraction (e.g. amplitude and time of the peak), which was traditionally performed using a fitting technique. A ML model was developed and tested with encouraging results, improving the speed by up to three orders of magnitude, with adequate precision. An interesting part of the project – while the model is created and trained in the Python environment using Keras, it is deployed as a part of the sPHENIX C++ software stack by leveraging ONNX – a cross platform library for integration of neural network models into a variety of software environments.

There is currently interest in the sPHENIX community in applying these techniques in other areas, such as clustering algorithms for the electromagnetic calorimeter with a view to enhance physics performance (e.g. efficiencies, rejection factors etc) compared to the existing analytical solutions.

One of the sPHENIX collaborators recently demonstrated a superior performance of the ML-based clustering for discrimination of various types of events in the electromagnetic calorimeter. This work is ongoing.

# EIC Infrastructure and Collaborative Tools

Maxim Potekhin

A busy year for collaborative tools in the EIC community

- From proto-collaborations to ePIC

Development, support of tools

- EIC Users Group website, redesigned in 2022
- EIC UG Software Working Group website
- Google Group and BNL hosted mailing lists, calendars
- Dropbox and Google Drive storage
- A few wiki instances
- YouTube channel

Steadily expanding use of our favorite website platform, Jekyll + git(hub)

- HSF, NPPS, PHENIX, EIC, ...

EIC UG has adopted Zenodo, drawing on expertise/experience from our PHENIX Zenodo implementation

# NPPS Collaborations

