NPP S&C Highlights at BNL in 2020 and Plans Ahead

Torre Wenaus, NPPS Group Leader, BNL Physics Dept
with contributions from many

JLab/BNL S&C Round Table
December 15 2020

BROOKHAVEN

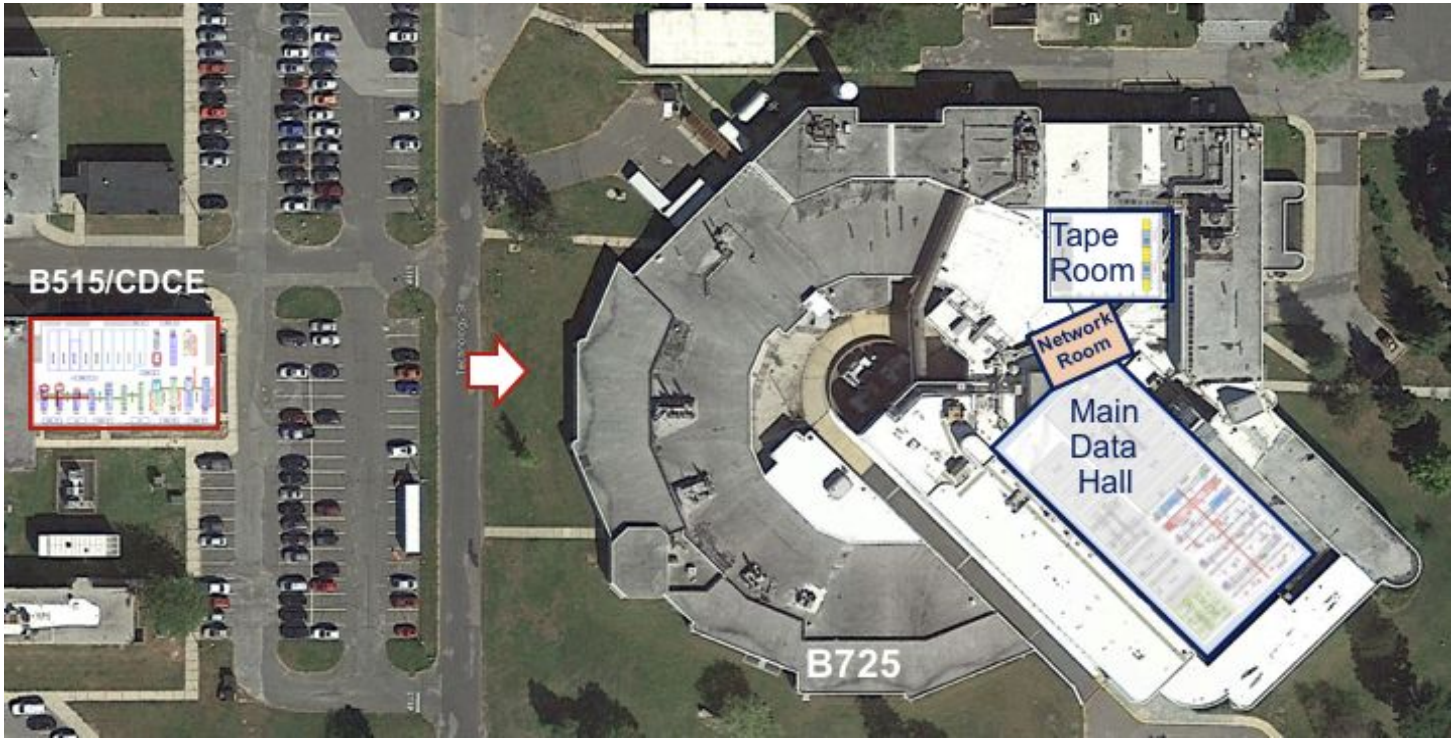
U.S. DEPARTMENT OF ENERGY

This talk

- Caveat: selective and incomplete, with more of an emphasis on software
- Within software, drawing mainly (but not entirely) on the Nuclear and Particle Physics Software (NPPS) group activities
  - Many software activities also in the STAR, (s)PHENIX, EDG (neutrino physics), Omega (ATLAS), NP and HEP theory and other groups
- No time for background on the experiments, going to assume they are familiar
- Outline… *with apologies for the terrifying slide count… I will skim quickly*
  - Computing facilities: SDCC
  - Software: NPPS
  - NP experiments
  - HEP experiments
  - Common software
  - Plans and priorities
- **Thank you to JLab for welcoming BNL into co-organizing the round table and inviting us to this annual tradition!**

Scientific Data and Computing Center (SDCC)

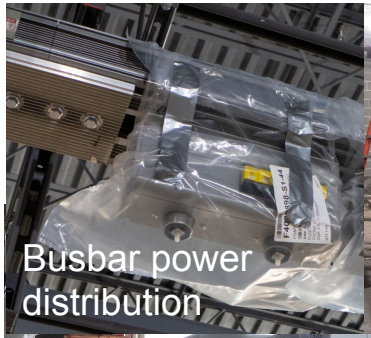
Transitioning to a new Data Center in BNL's 1st-generation light source building



Ofer Rind, BNL

SDCC: Transition to new Data Center

Construction delayed 3 months due to COVID-19 shutdown, now back within 1 month of schedule.



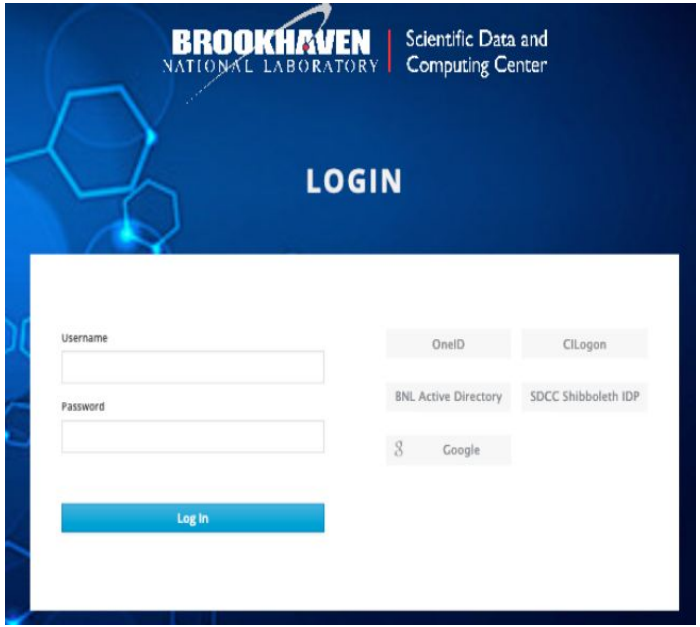
SDCC: Transition to new Data Center

Occupancy expected in Jun 2021.

Ofer Rind, BNL

SDCC: Secure Access

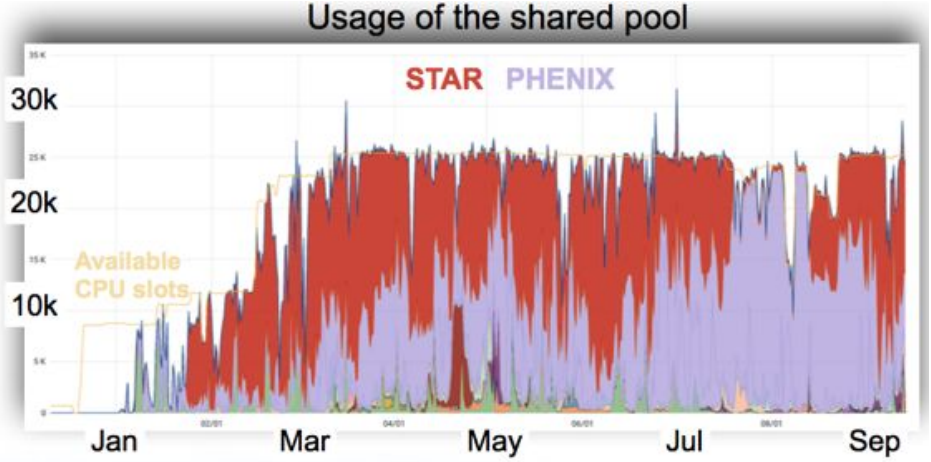
- SDCC becomes its own InCommon IdP
- Internal migration from shibboleth to keycloak, setting the stage for improved Single Sign-On and added security with Multi-Factor Authentication
- Updated password policy improves security and eliminates expiration cycle
- Deployment of improved NX service



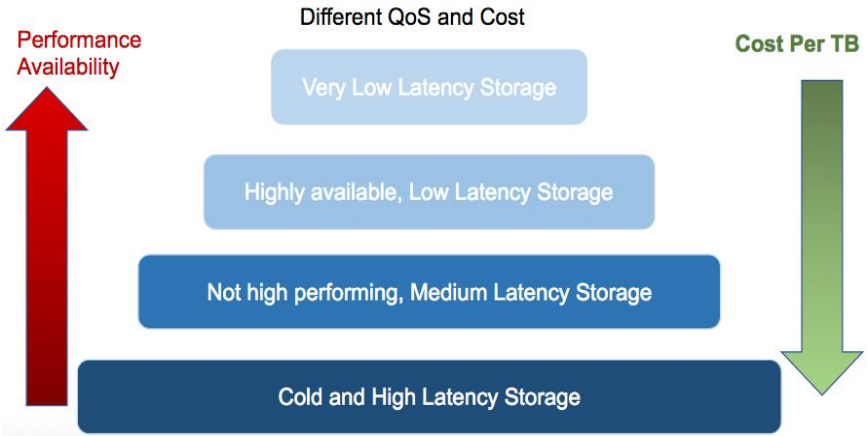
SDCC: Computing and storage

Migration to “shared pool” compute farm resource model

- Single batch farm for better utilization of available CPU cycles



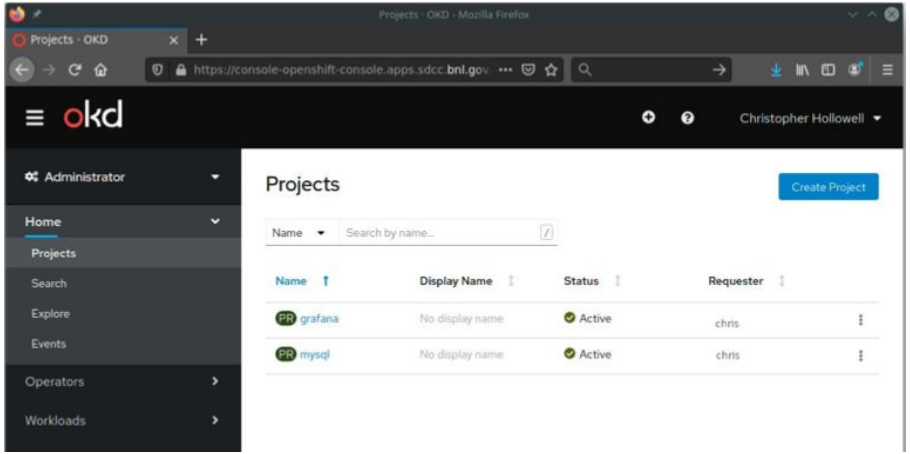
- ATLAS dCache upgrade (6.2)
  - Move to HTTP-TPC as primary data transfer protocol and enable new QoS capability
- Improved utilization of storage resources
  - Data Carousel and Multi-layer Automatic Storage (MAS)

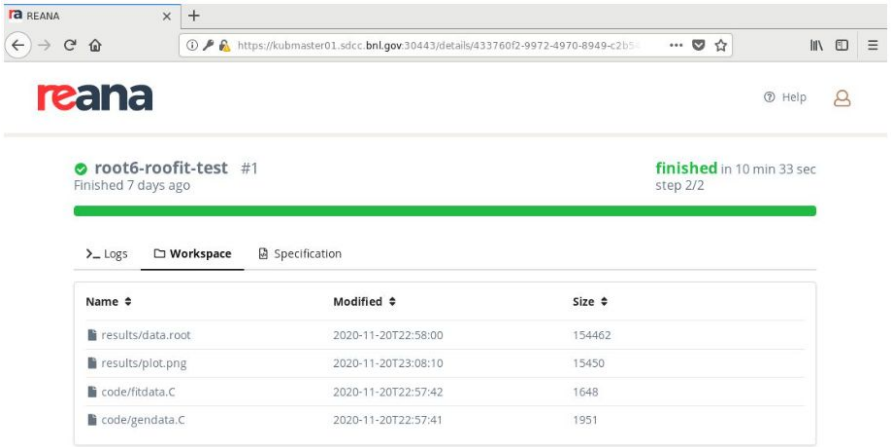


Ofer Rind, BNL

SDCC: Computing and storage

- Facility-wide transition to Lustre storage
  - NSLS-II: New Lustre-based S3 object store
  - Plus Globus endpoint leveraging new BNL Globus site license
- Deployment of new containerized analysis services
  - Openshift/OKD test instance for user services
  - REANA test instance deployed at scale



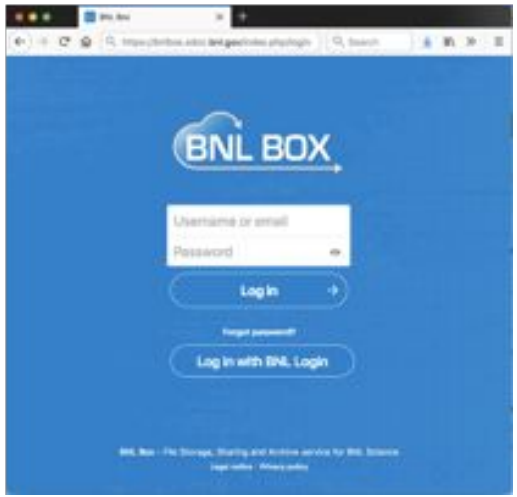


Ofer Rind, BNL

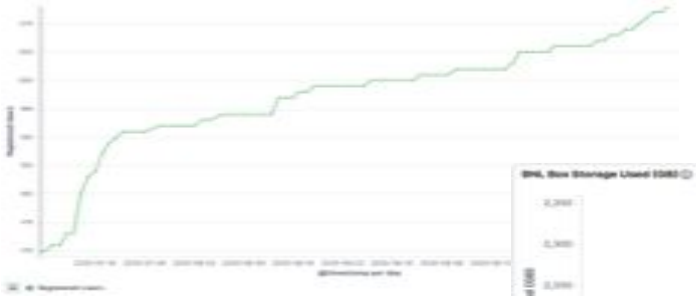
SDCC: User services

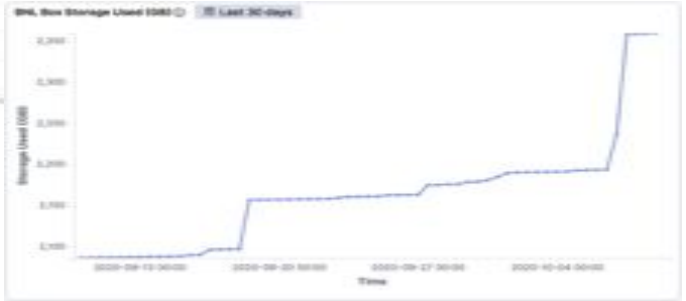
- Full deployment of Enterprise level File Sync-and-Share service (BNL Box)
  - Providing flexible, easy-to-use, unified cloud storage for all BNL scientific users
  - Nextcloud service on Lustre storage with HPSS tape interface for data archiving
  - Currently 232 users from across the BNL scientific landscape and growing

https://bnlbox.sdcc.bnl.gov



- New ELK stack monitoring and analytics - expanding to multiple SDCC services (e.g. Globus transfers)





Ofer Rind, BNL

SDCC: User services

- New deployments of Invenio-based Digital Repositories
  - **SET**: NNSA repository with DOE OneID federated auth
  - **Covid-19-archive**: Zenodo-based for DOE National Virtual Biotechnology Lab (NVBL)
  - **Electron Ion Collider (EIC)**: Zenodo + CILogon/COmanage federated auth
  - **sPhenix**: Beta code release on production infrastructure
  - **InvenioRDM** Alpha release testbed

SDCC: User services

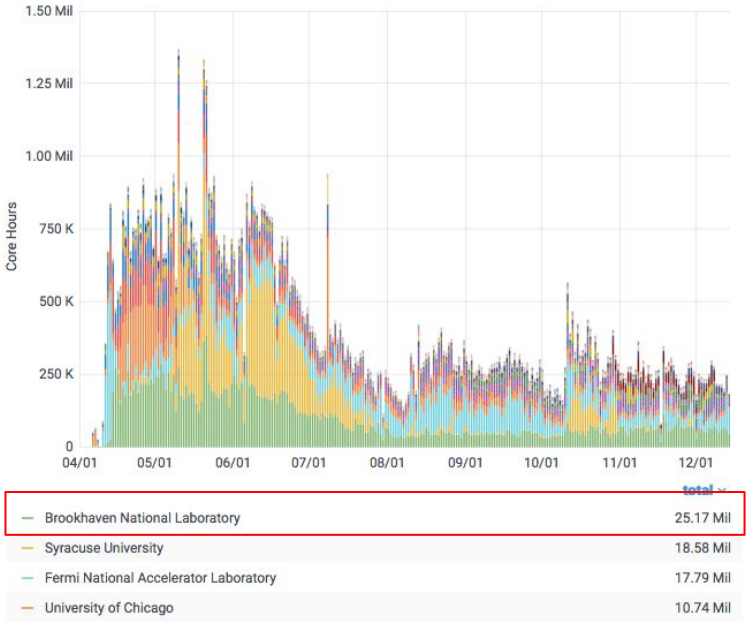
- Formed a new Experimental Support Group
  - “Embedded” personnel supporting experiment computing efforts (STAR, EIC, etc.)
- Deployed new Drupal web instances (SDCC internal, USATLAS, sPHENIX)

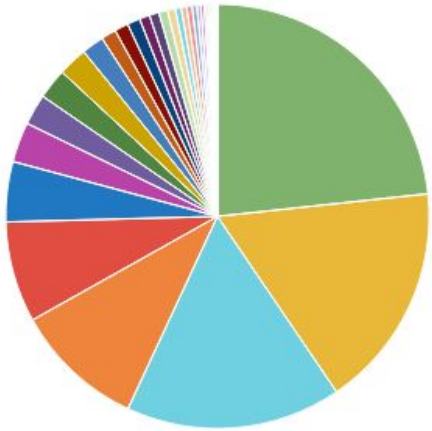
USATLAS Drupal instance

SPHENIX Drupal instance

SDCC: Covid-19 contribution

- Provided millions of core-hours via OSG for Covid-19 research
  - https://www.bnl.gov/newsroom/news.php?a=217177





	total
Brookhaven National Laboratory	25.17 Mil
Syracuse University	18.58 Mil
Fermi National Accelerator Laboratory	17.78 Mil
University of Chicago	10.74 Mil
Massachusetts Institute of Technology	8.31 Mil
University of Nebraska	4.96 Mil
University of Texas Arlington	3.30 Mil
University of Michigan	2.54 Mil
American Museum of Natural History	2.51 Mil
University of California San Diego	2.42 Mil
Vanderbilt University	1.84 Mil
Rutgers University	1.24 Mil

Ofer Rind, BNL

Nuclear and Particle Physics Software (NPPS) Group

The Nuclear and Particle Physics Software (NPPS) Group, created May 2019, consolidates much (not all) of the NPP software development in the Physics Department

More information at https://npps.bnl.gov

- Currently 14 HEP and 8 NP members, including two non-BNL members (ORNL, Stony Brook)
- Up by 4 members since it was formed, one current posting (HEP)
- Working on ATLAS, Belle II, DUNE, EIC, PHENIX, sPHENIX, STAR, Rubin Observatory

The group cultivates and leverages cross-experiment common efforts across HEP and NP

- Synergies in shared personnel, expertise, software
- 12 group members are applying their expertise to >1 experiment, up from 5 before NPPS

We also collaborate with several cross-cutting organizations and projects

- WLCG (LHC), HSF (international), IRIS-HEP (NSF), HEP-CCE (DOE), CSI (BNL/ASCR)
- and many others...

Collaborations

Stony Brook University

Argonne NATIONAL LABORATORY

HSF
HEP Software Foundation

Jefferson Lab

Google Cloud

iris hep
Institute for Research and Innovation in Software for High Energy Physics

BERKELEY LAB

Fermilab

SCIENTIFIC DATA MANAGEMENT
RUCIO

OAK RIDGE National Laboratory

ROOT
Data Analysis Framework

XRootD

BROOKHAVEN NATIONAL LABORATORY *Computational Science Initiative*

DIRAC
THE INTERWARE

CMS

CERN

CERN openlab

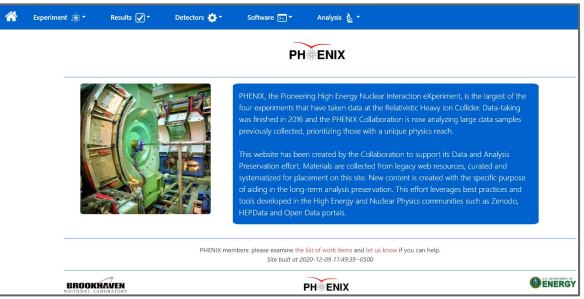
WLCG
Worldwide LHC Computing Grid

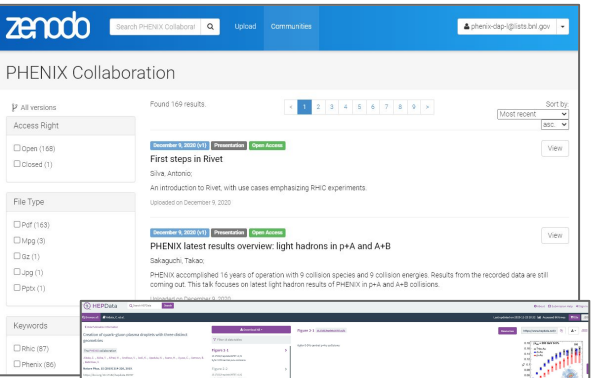
SLAC

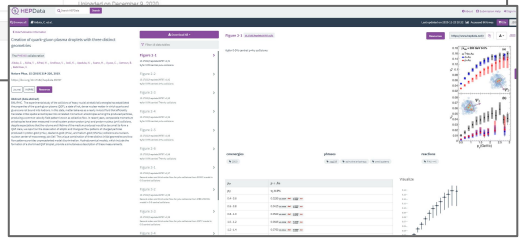
BROOKHAVEN

PHENIX: Data and analysis preservation (DAP)

c.f. Maxim's Trends workshop talk





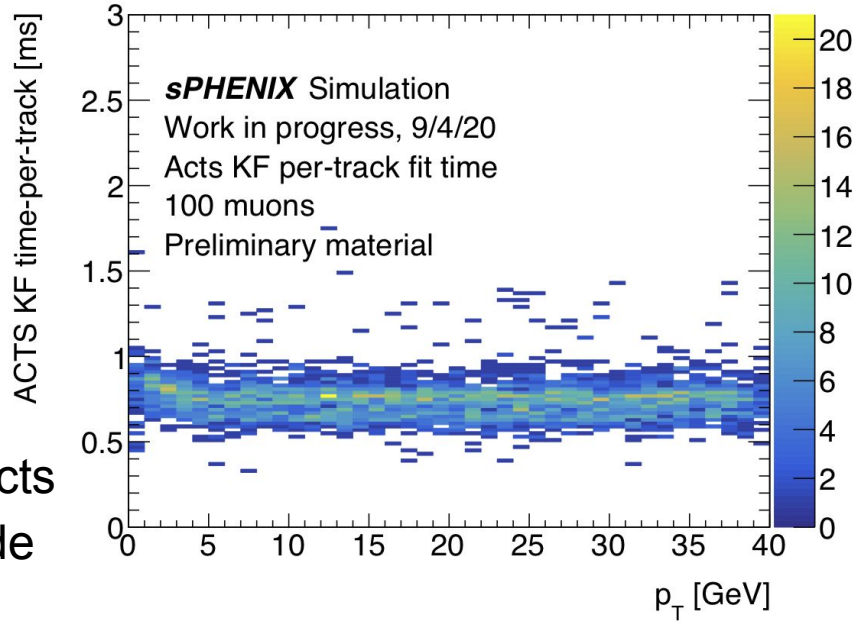


- Leveraging platforms and resources: reduce/eliminate in-house DAP solutions
- GitHub (web development, code preservation, documents etc)
  - GitHub/Jekyll site created from scratch replaced legacy php, aggregating info from many sources
- Zenodo (CERN) - world-class repo for preservation of documents etc
  - PHENIX materials from a variety of sources are collected, curated, tagged with metadata (including keywords) and committed Zenodo
- HEPData (CERN) - portal/repository for numerical data used in publications
  - Data is being collected, packaged and uploaded
- Rivet (*Robust Independent Validation of Experiment and Theory*) - exploring
- OpenData portal - capture and preserve analyses and samples of data
- REANA - a platform for reproducible analysis
- Collaborative opportunities with JLab?

Maxim Potekhin, BNL

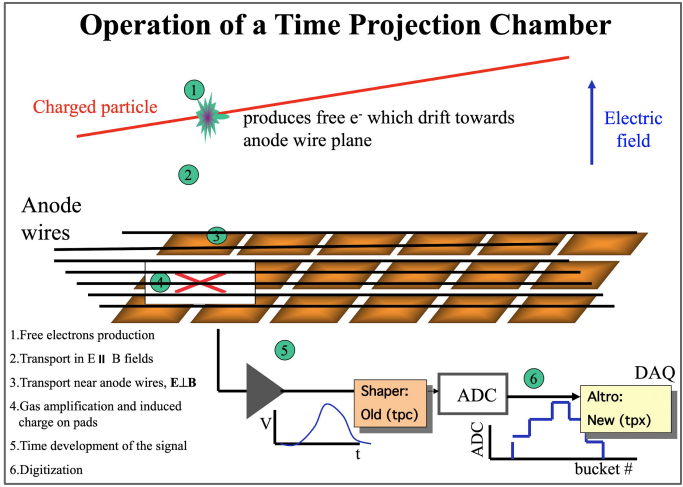
sPHENIX: Acts tracking

- Acts is now the default track fitter in sPHENIX software framework
- Have implemented a variety of tools:
  - Kalman Filter (KF) track fitting
  - Combinatorial KF track finding + fitting
  - Direct surface fitting (contributor to Acts!)
  - Initial and final vertex finding + fitting
  - Track seeding
  - Track projection to calorimeters
- Monitoring progress towards 3D fitting in Acts
- Advancing track fitting simulations to include TPC space charge distortions



sPHENIX: TPC full simulation

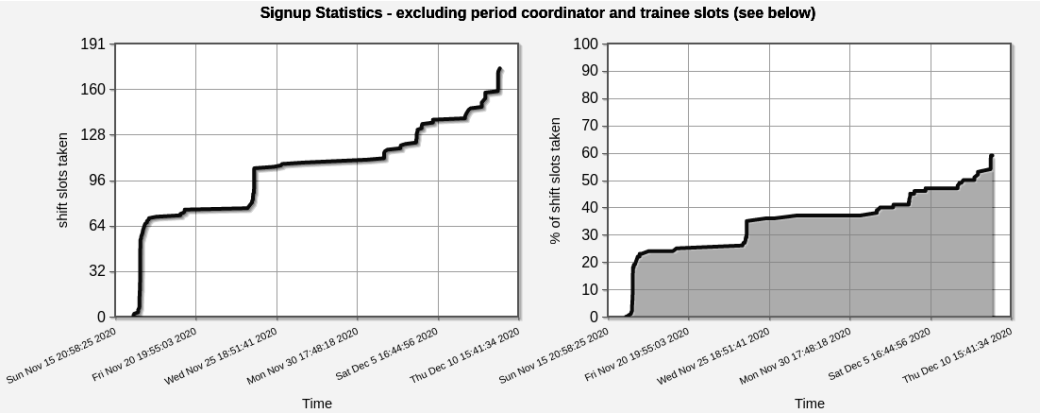
- Provide full simulation for sPHENIX TPC detector by adopting STAR's TPC Response Simulator TpcRS
- Accurate model is critical to confront TPC track reco with realistic data
- Support detailed modeling of space charge distortions, realistic characterization of hit quality
  - Electrons generated and transported to anode wires
  - Diffusion and distortions applied
  - Near anode wire ExB effects are accounted; E field not parallel to B field
- Cross checking distortion calculations with sPHENIX calibration group, in agreement
- Code refactored to be more modular and detector agnostic
  - https://github.com/bnlnpps/tpc-rs
- We plan to continue adapting the library to sPHENIX needs
- Also will pursue using STAR TPC data to test sPHENIX TPC track reconstruction



Yuri Fisyak (BNL), Dmitri Smirnov (BNL)

STAR: Database & collaboration tools

- STAR Offline (Calibrations) DBs: (a) DB cluster scales pretty well with the ever-increasing load, (b) DB cluster has been moved to the virtualized SDCC-hosted service - 10 nodes in total, DB Load Balancer updated to take new nodes into account automatically
- RHIC Run services: STAR ShiftSignup, RunLog, Event Display, Experiment Status Plots are ready for the Run 2021, ShiftSignup now features a “merged” dues for 2020 and 2021 to accomodate for 2020 difficulties of getting shifts done; sign up is going well - about 60% of slots already taken
- EIC, SPHENIX, STAR phonebooks are now centrally maintained at https://phonebook.sdcc.bnl.gov
- Detector data collection framework (MIRA) and SSH Key Management system are updated with new features
- Two factor authentication implemented on various web services, tied into SDCC's kerberos service

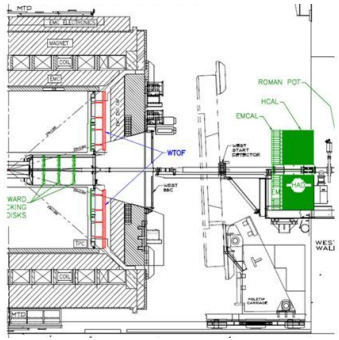


Dmitry Arkhipkin, BNL

STAR: geant4star

The STAR experiment has utilized starsim, a GEANT3 application, integrated with its C++ analysis framework, for simulation productions since STAR's inception

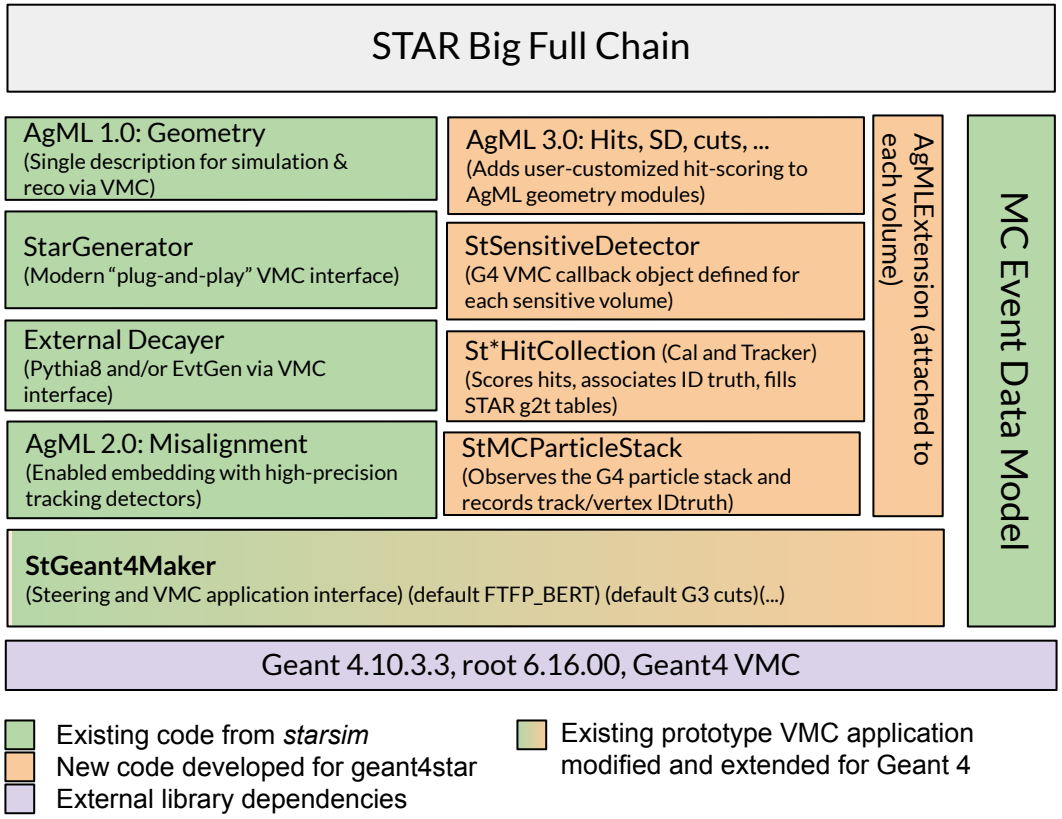
- Forward upgrade adds hadronic calorimetry starting in 2021
- Motivates a move to a Geant4 in order to leverage better hadronic physics modeling
- Would also alleviate issues with handling old Fortran codes



The geant4star project

- Leverages existing STAR simulations software stack
  - Utilizes the existing Monte Carlo Event Data Model (EDM)
  - Seamless integration with the STAR embedding & simulation chains
  - No changes required for existing slow simulators (digitizers)
- Test-driven development demonstrates correctness of EDM, consistency with *starsim,* and provides comprehensive test suite for continuous integration

STAR: geant4star



- Existing STAR software stack used for event steering, integration with digitization and reconstruction code, standard I/O
- StGeant4Maker initializes geometry and sensitive detectors, and provides the bridge to the Geant4 and Virtual MC (VMC) toolkits
- Generic sensitive detectors for trackers and calorimeters are implemented along the lines from early EIC discussions (A. Dotti, **Round table 2017**)
- AgML automatically creates sensitive detectors based on geometry definitions

geant4star is being developed alongside of a comprehensive test suite for validation and cross-comparison with *starsim*. See supplementary slides

STAR: Production

- Leveraging available HPC resource (Cori @ NERSC) for real data production + embedding
- Developed "blinded" productions for physics analyses of isobar collisions
- BirdView tool
  - Monitoring system for nightly jobs and production
  - Nightly jobs monitoring
    - STAR runs 255 jobs every night to track the library changes
    - Log extraction module records 25 parameters for a single job execution
    - Tracking changes was like finding a needle in a haystack
  - Production
    - Current production status
    - Prediction for completion of current dataset production
    - Disk space monitoring

BirdView

Amol Jaikar, Jerome Lauret, Gene Van Buren, BNL

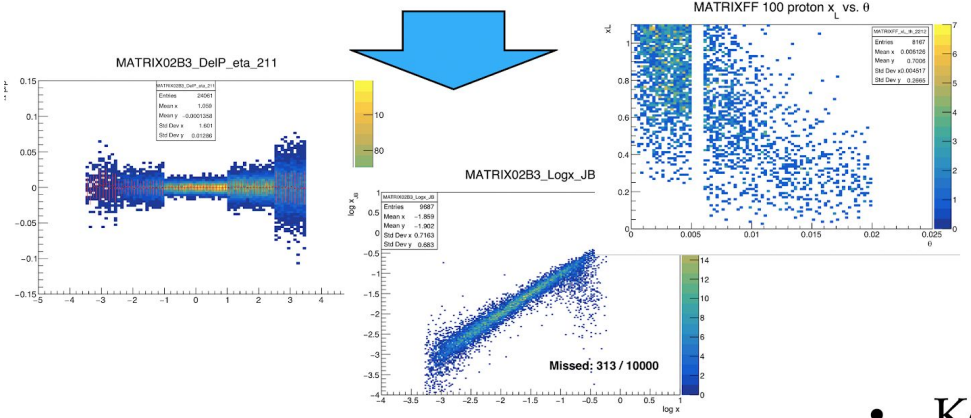
STAR: Software infrastructure modernization

- 32 to 64 bit migration completed
  - Presence of Geant3 in the active STAR software stack presented migration issues extending back to Zebra era
- Migration to ROOT6 underway
- Migration from CVS to git and GitHub underway
  - Pending detailed tests by the STAR production team
- Once the GitHub migration is done, migrate build system from Cons to CMake
  - Studies show significant reduction in built time with CMake
- Planning migration to Spack
  - Spack providing new packages for STAR such as ROOT6, Geant4, Geant4-vmc, KiTrack, GenFit, KFParticle; and previously used packages such as Vc, log4cxx, boost
  - Spack should make it easier to access and evaluate more modern compiler versions for production; take advantage of recent C++ features and better optimizations

EIC: Fast simulation

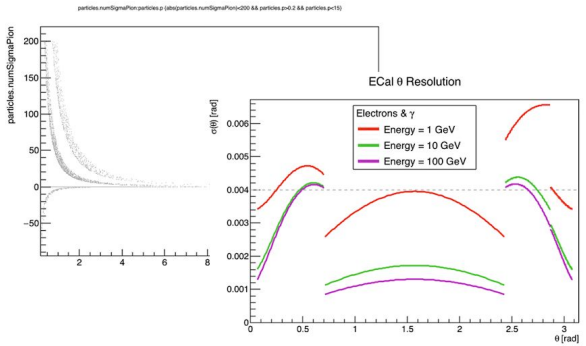
- EIC-smear is growing with the project

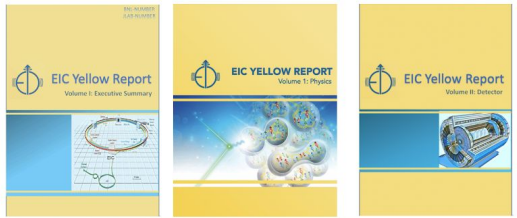
Nomenclature			Tracking						Electrons and Photons			π/K/p		HCAL	
			Resolution	Relative Momentum	Allowed X/X_0	Minimum-pT	Transverse Pointing Res.	Longitudinal Pointing Res.	Resolution σ_E/E	PID	Min E Photon	p-Range (GeV/c)	Separation	Resolution σ_E/E	Energy
↓ p/A	Far Backward Detectors	low-Q2 tagger	Not Accessible												
	Central Detector	Backward Detector	Reduced Performance	σ_p/p ~0.2%⊕p⊕5%	~5% or less X	70-150 MeV/c (B=1.5 T)	dca(xy) ~ 40/pT μm ⊕ 10 μm	dca(z) ~ 100/pT μm ⊕ 20 μm	1%/E ⊕ 2.5%/√E ⊕ 1%	π suppression up to 1/1E-4	20 MeV	≤ 10 GeV/c	≥ 3 σ	50%/√E⊕10%	~500MeV
				σ_p/p ~0.04%⊕p⊕2%					2%/E ⊕(4-8)%/√E ⊕ 2%	π suppression up to 1/(1E-3 - 1E-2)	50 MeV				
		Barrel		σ_p/p ~0.04%⊕p⊕1%		200 MeV/c	dca(xy) ~ 30/pT μm ⊕ 5 μm	dca(z) ~ 30/pT μm ⊕ 5 μm	2%/E ⊕(12-14)%/√E ⊕ (2-3)%	π suppression up to 1/1E-2	100 MeV	≤ 6 GeV/c		100%/√E~10%	
		Forward Detectors		σ_p/p ~0.04%⊕p⊕2%		70 - 150 MeV/c (B = 1.5 T)	dca(xy) ~ 40/pT μm ⊕ 10 μm	dca(z) ~ 100/pT μm ⊕ 20 μm	2%/E ⊕ (4*-12)%/√E ⊕ 2%	3σ e/π up to 15 GeV/c	50 MeV	≤ 50 GeV/c		50%/√E~10%	
				σ_p/p ~0.2%⊕p⊕5%											
↑ e		Instrumentation to separate charged particles from photons	Reduced Performance												
	Far Forward Detectors	Proton Spectrometer	Not Accessible												
		Zero Degree Neutral Detection													



In the pipeline:

- Validation, Tests
- Granularity, PID, …

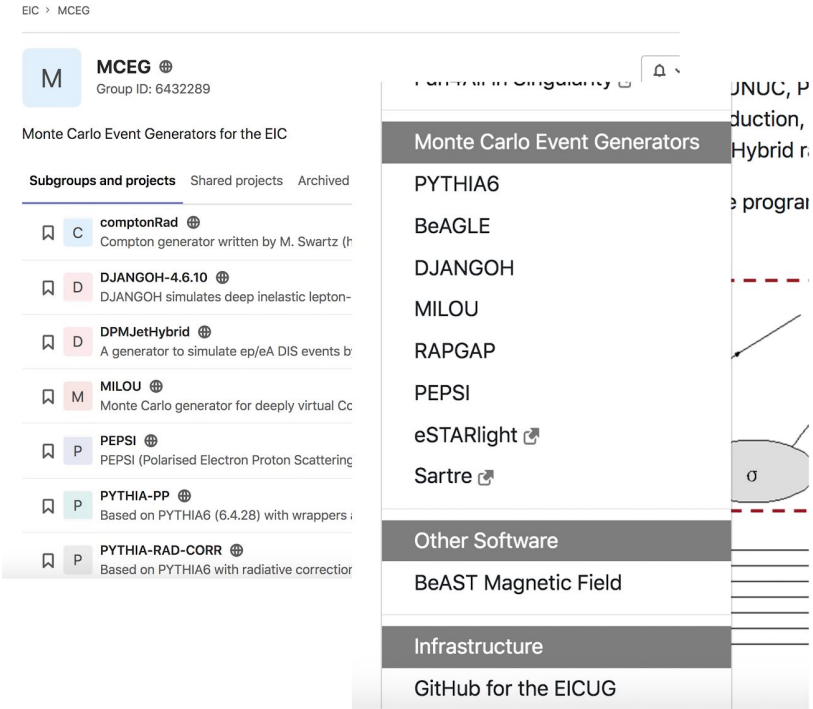


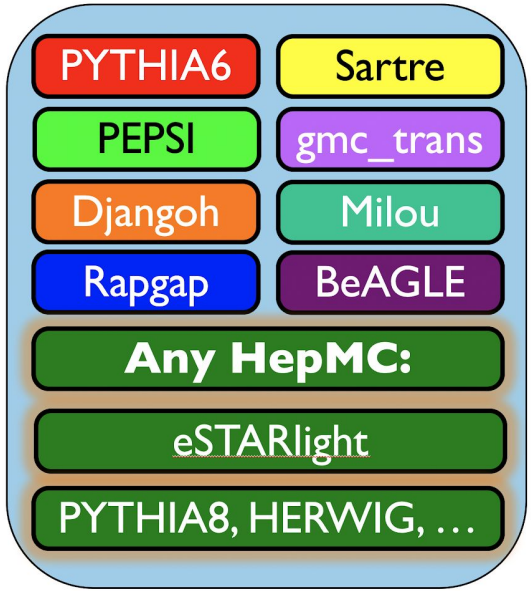


- Key ingredient for Yellow Reports!

EIC: Generators

- EIC MC collection is consolidated



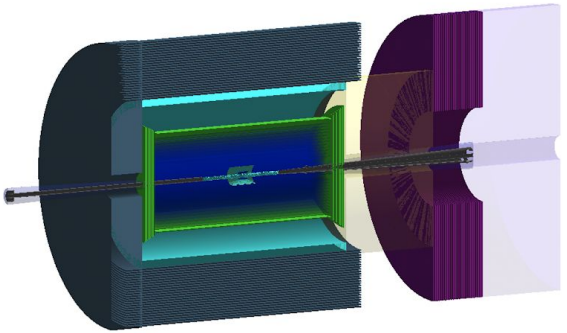


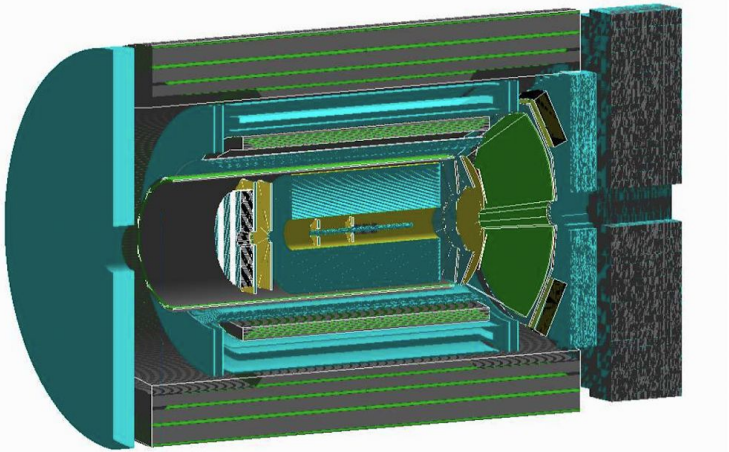
- … and growing!

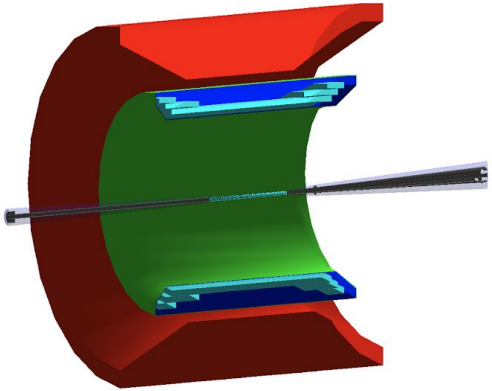
Kolja Kauder, BNL

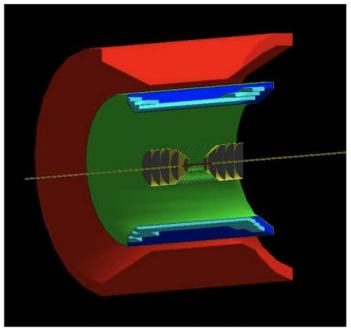
EIC: Detector Concepts in Fun4All

- One detailed detector concept based on the Babar magnet implemented
  - Jleic template + fieldmap
  - Beast template + fieldmap









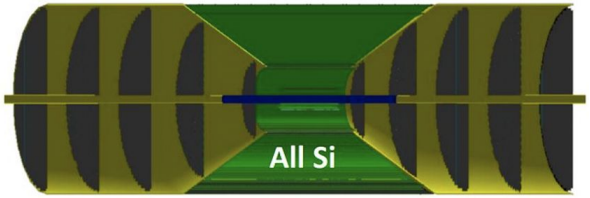
Detectors can be placed into templates:
All Silicon Tracker inside Beast magnet

Chris Pinkenburg, BNL

EIC: Integrated hybrid detector: It's all coming together

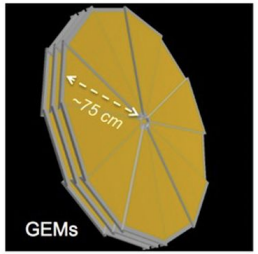
- **Integrated hybrid detector configurations are added to common simulation framework (Fun4All)**
  - TPC + Si disks + Si vertex
    - Including end cap material
  - MPGD barrel + Si vertex
  - Triple-GEM disks
    - Based on SBS GEMs
  - All Si detector
  - Ability to study various detector integrated configurations

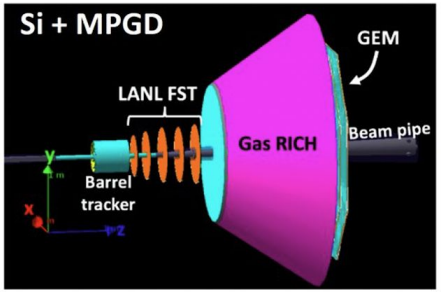
Rey Cruz-Torres et al: UCB



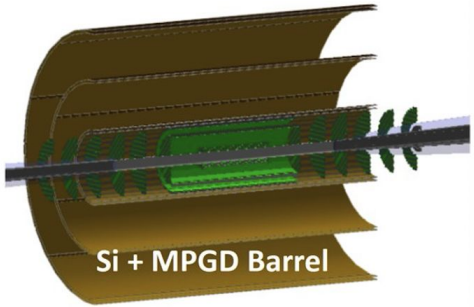
Cheuk-Ping Wong, et al: LANL

SBS GEMs

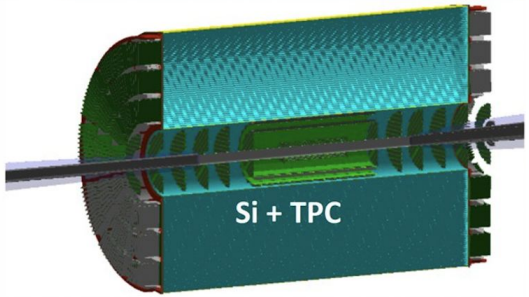




Qinhua Huang, et al , CEA Saclay



Håkan Wennlöf et al: UoB

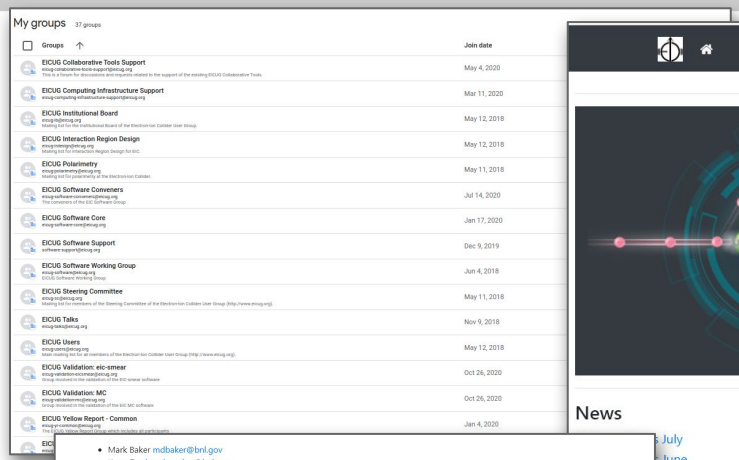


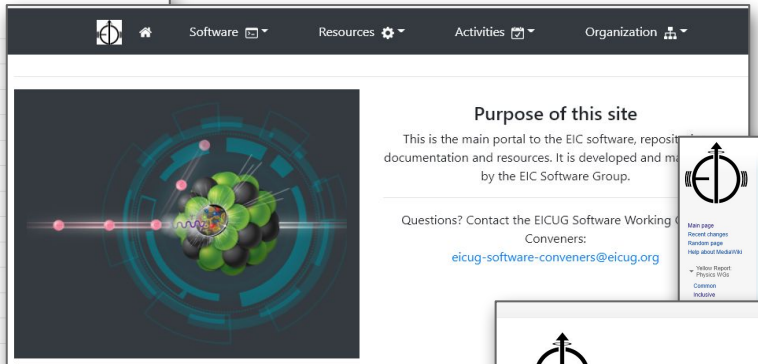
Matt Posik, Temple U

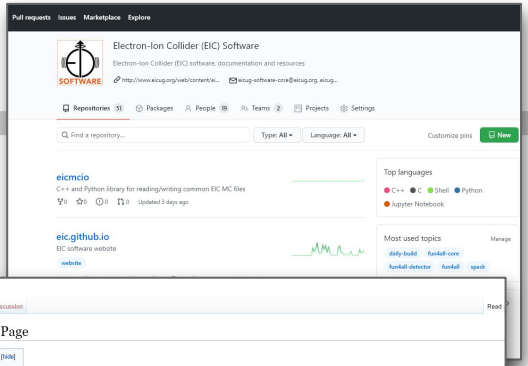
EIC: Infrastructure, communication, collaboration

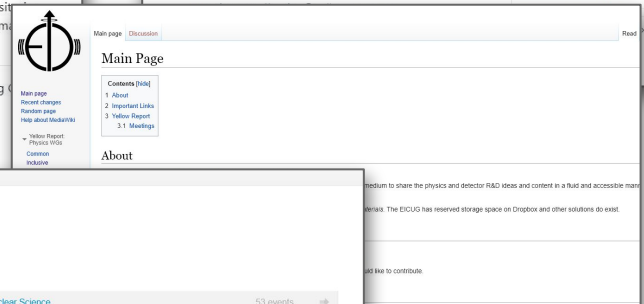
- EIC-wide collaborative tools (maintained mainly by NPPS):
  - The main EICUG website (Drupal-based, hosted by BlueHost)
  - Google services: 37 groups for distribution of e-mail, group calendars
  - BNL-hosted: Indico, Wiki, the EICUG Phone Book
- Data and code distribution
  - BNLBox cloud storage, containerized and cvmfs-hosted software
- https://github.com/eic managed by the EICUG Software Group
  - Covers 51 repositories, managed according to simple agreed upon guidelines
  - Efficient collaboration in core software, documentation, web development
  - Tools supporting developers: dOxygen, LXR code browser
- The EICUG Software Group Website https://eic.github.io/
  - Hosted on GitHub pages, secure, fast and efficient
  - Becoming a true hub for all EIC-related software, fostering teamwork and collaboration

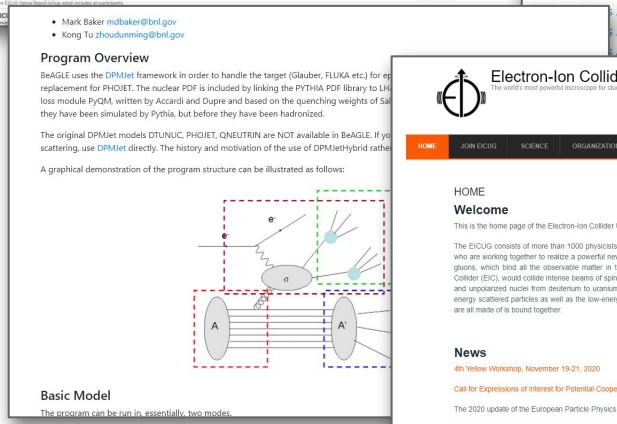
EIC Collaborative Infrastructure in pictures

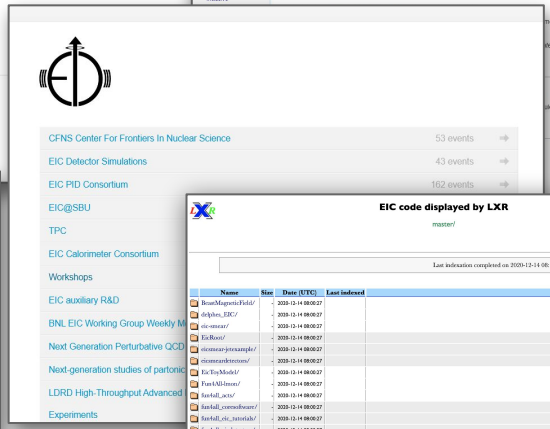


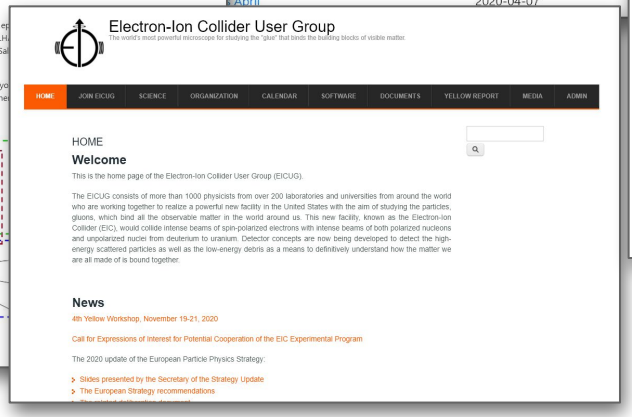


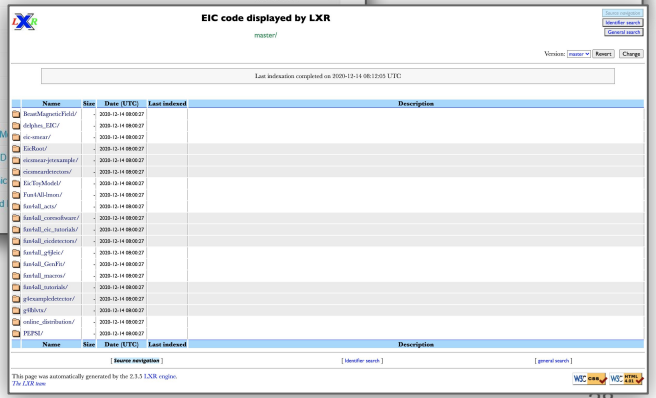












ATLAS: NPPS in ATLAS S&C

- Total ATLAS NPPS ~7.5+ FTE
- Core expertise in offline software and databases
  - Athena framework core expertise including its multiprocessing and multithreading variants
  - BNL develops ROOT I/O for ATLAS and works with the ROOT team on I/O issues
- Leading roles in ATLAS distributed software and computing since its inception
  - PanDA workload management system manages all ATLAS distributed production and analysis
  - Prodsys production system translates physicist requests into PanDA production
  - Many innovations to grow the resources available to ATLAS (HPCs, clouds, fine grained processing)
  - Pilot system (for Grid, Clouds, HPCs)
  - Data Management System (Rucio)
  - Monitoring
- Leading Role in ATLAS Core Software and Computing Model Project, also in DOMA and IRIS-HEP
- US ATLAS and ATLAS Software infrastructure support

  Long term support of ATLAS release build/test tools (~ 20000 Nightlies, CI, stable releases annually). Transitioned to modern open-source tools. Extending to new architectures
- Proposed and lead several successful R&D projects and Task Forces
  - Analysis Model Study Group – Analysis Model for the LHC Run3 and beyond
  - Containerization – use containers to run Monte-Carlo Production on a uniform way
  - Data Carousel
  - intelligent Data Delivery Service (iDDS)
  - HEP-Google
  - New workflows integration (ML, HPO, …) in ProdSys2/PanDA
  - PanDA beyond ATLAS
  - Data popularity studies and a new way to place/handle data replicas
  - Operational Intelligence

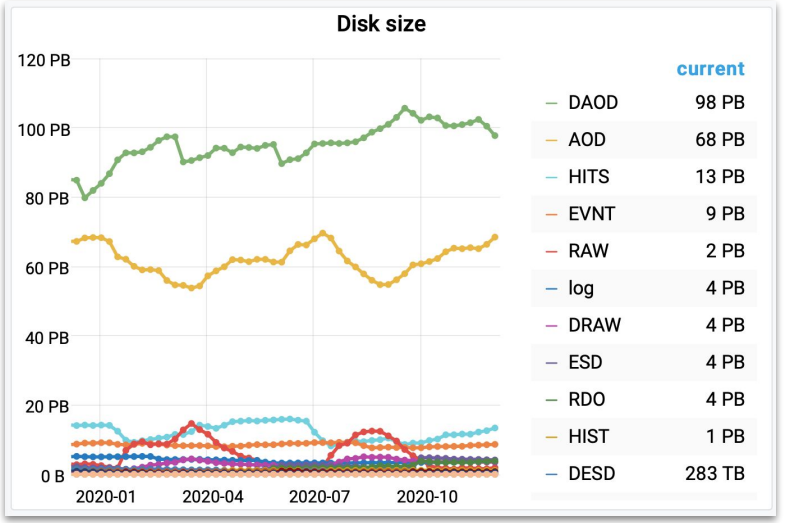
Alexei Klimentov, BNL

Our largest software activity

Skipping the fine print to give you the highlights...

ATLAS: Analysis Model for Run-3

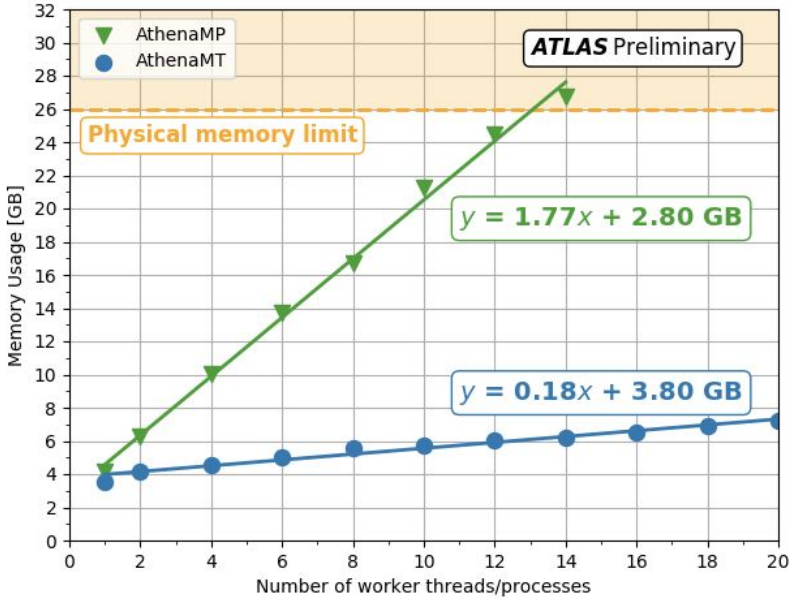
- ATLAS grid disk space is dominated by many different analysis products
- For Run3 introducing two main formats:
  - DAOD_PHYS suitable for most analysis, 40-50 kB/event
  - DAOD_PHYSLITE with calibrated physics objects, 10 kB/event
  - Introduce lossy float compression
- Keep only popular AOD on disk and use data carousel in production of DAODs
- Status:
  - DAOD_PHYS available for data/MC
  - DAOD_PHYSLITE available for data
- Leading BNL role in developing and implementing the model



Code is being ported to ATLAS's newly multi-threaded framework AthenaMT

ATLAS: Multithreaded Athena framework

- The Athena framework has been ported to support multi-thread processing with the goal to significantly save in memory usage
  - Previously ATLAS has relied on multiprocessing (AthenaMP)
- The migration is finished and will have first large scale production usage in Spring 2021 reprocessing all Run-2 data
- Memory savings vs. number of threads in reconstruction looks very promising (right)
- Large scale tests are currently conducted to identify the remaining thread related problems
- Software effort readying for Run-3 co-led by BNL



Memory usage vs. the number of threads in reconstruction of 500 simulated ttbar events including pile-up with an average number of interactions-per-bunch-crossing, <μ>, of 20.

ATLAS: Rucio distributed data management system

c.f. Mario Lassnig's Trends talk

Scientific data management system for exascale, manages 460+PB of ATLAS data and O(1PB/day) transfer, adopted by O(10) HEP experiments

- A busy development year
  - List of Rucio pull requests during this year: PR list (shared ATLAS/Belle II)
  - List of components under BNL responsibility: https://github.com/rucio/rucio/wiki
- Conveyor file transfer daemon
  - Finalization of multihop mode
    - Enable staging from CERN's new CTA tape manager to any site
  - Changes made to be able to use Google Cloud Storage in Rucio
    - cf. Google project on later slides
- Reaper2 file deletion daemon
  - New generation of the reaper with various improvements and bug fixes
  - Migration from previous generation completed

ATLAS Workflow and Workload Management on One Slide

Global ATLAS operations
Up to ~1.2M concurrent jobs
25-30M jobs/month at >250 sites
~1400 ATLAS users

Orchestrate all ATLAS Workflows :

- *MC Production*
- *Physics Groups WF*
- *Data reprocessing*
- *T0 spill-over*
- *HLT processing*
- *SW validation*
- *User's analysis*
- *ART*

Physics Group production requests

Meta-data handling

Distributed Data Management

AMI

Rucio

pyAMI

WFM

WMS

ProdSys2/ DEFT

Requests, Production Tasks

DEFT DB

Tasks

JEDI

Tasks Jobs

PanDA DB

PanDA server

Analysis tasks

ATLAS production requests

Jobs

harvester

harvester

harvester

harvester

pilot

pilot

pilot

pilot

Pilot scheduler

condor-g

ARC interface

ARC interface

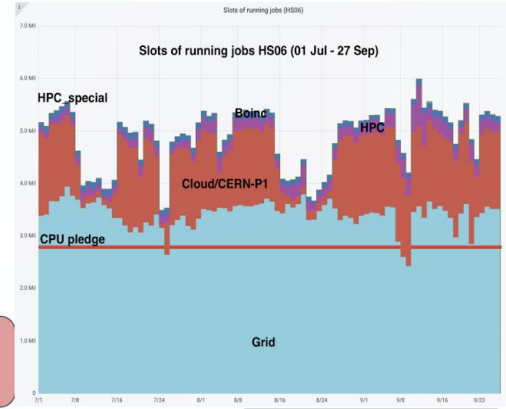
EGEE/EGI

OSG

NDGF

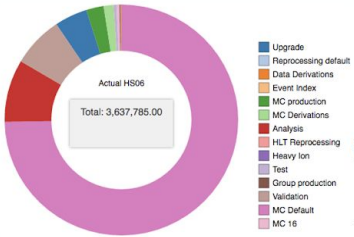
HPCs

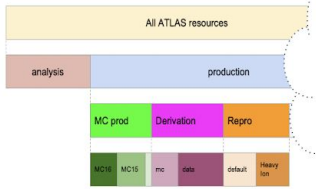
Worker nodes



First exascale workload manager in HENP
1.4+ Exabytes processed yearly in 2014/18
Exascale scientific data processing today

Shares/priorities





Support rich harvest of heterogeneous resources. Integrate WF and data flow

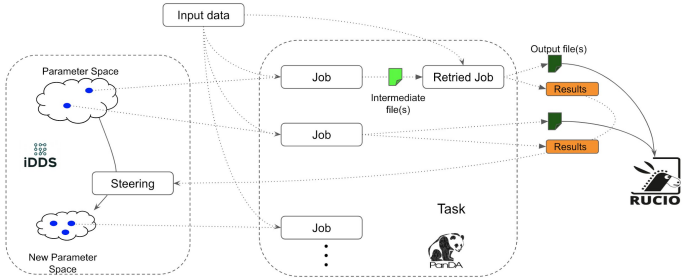
ATLAS: PanDA workload management system

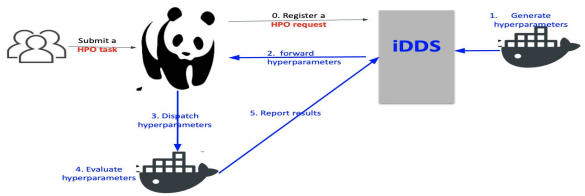
Exascale scientific Workload Management System. 1.3+ EB/year are processed and analyzed via PanDA, 1500+ users/year, ~400M jobs/year. Supports all workflows in ATLAS

- Migration of the entire PanDA system to **python 3**
- Next-gen **authentication/authorization**: OIDC/OAuth2.0 compliant AuthN/Z capable of ID federation among scientific and academic ID providers using Indigo IAM and CILogon
  - Leveraged also for sPHENIX (and soon EIC?) use of PanDA
- New CLI tool for **ML hyperparameter optimization**
- Refactoring of the user analysis infrastructure to support **user level containers**
- Migration to new cross-experiment information system CRIC (derived from ATLAS's AGIS)
- Optimization: cron consolidation, memory use reduction
- Performance study of **user analysis** to improve user experience
- Production Harvester instance on **Kubernetes**, with Helm chart for automated deployment
- PanDA client **support for JupyterLab**
- Communication between iDDS and PanDA/JEDI via ActiveMQ

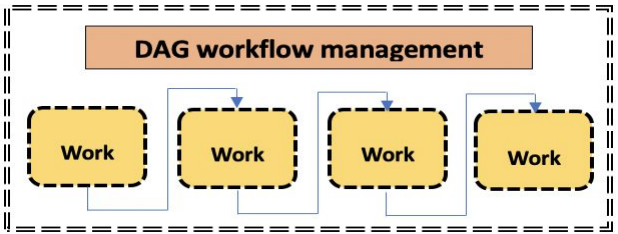
ATLAS: PanDA/iDDS (intelligent Data Delivery Service)

- Collaboration with IRIS-HEP on fine grained intelligent data-driven workflows
  - A further step in fine grained processing that began with the Event Service
- Thus far applied to ATLAS but not specific to ATLAS
  - Uses experiment-agnostic PanDA/iDDS R&D instance at CERN
- Successfully applied to a growing list of novel workflows
  - Data Carousel
    - Prompt start of processing when data appears from tape; avoids tails
  - HPO (Hyper Parameter Optimization)
    - Automated platform for scaled-up hyper-parameter optimization using distributed GPU resources on the grid, HPC, and clouds
    - Serving the only near-term payload we have for GPU-heavy HPCs: ML training/optimization
  - DAG based workflow management
    - High-level workflows specified by DAGs driving sophisticated workload scheduling
    - Interfaces nicely to DAG based experiments and tools: Rubin, Pegasus, ...
- Near-term plans
  - Improvements of user experience
    - Client and CLI development and improvements
    - Documentation and monitoring
  - Endeavour to support more use cases in multiple experiments
    - DAG based Active Learning is ongoing
    - Dynamic transformation and placement on demand, for example Derivation on Demand
    - Fine-grained data transformation and delivery, such as Event Streaming Service





iDDS HPO



iDDS DAG

ATLAS: PanDA Pilot

Responsible for executing and monitoring user and production system payloads on grid and HPC worker nodes and volunteered resources. Highlights from 2020 include:

- **Python 3** compatibility
- Major improvements to payload **containerization**
- **HPO** support with support for pre-, co- and post-processes
- Remote file open verification (for direct access jobs to prevent late failures)
- Improved support for running on **HPCs** (e.g. Summit)
  - with support for the second generation event service for HPCs '**Raythena**'
- Continued improvements in error handling
- Migration from AGIS to CRIC
- Caching of pilot proxy validation to prevent time consuming proxy cmd execution
- Payload proxy validation in long development now in final testing

ATLAS: PanDA/iDDS plans for 2021

- Continuous effort to improve **user experience**
  - Client API and CLI tools
  - Documentation
- Integration of iDDS with other IRIS-HEP products such as Service-X (data format transformations) and **Function-as-a-Service**, use cases directed at analysis
- panda-client on **JupyterLab**
- Support of more use cases in multiple experiments
  - **On-demand data processing** and placement
  - Fine-grained data transformation and delivery, e.g. **Event Streaming Service for analysis**
  - DAG-based **active learning**
- Increase of usage of the **HPO** service and its integration with more resources
  - Elastic distributed training with Horovod and Amazon EKS or Google GKE
  - Parallel HPO of fine grained models
    - ATLAS FastCaloGAN with 300 networks (100 η slices x 3 particle IDs) takes **100 GPU-days** to train

ATLAS: Data Carousel R&D project

An orchestration between workflow/workload management (WFMS), data management (DDM), data transfer (FTS) and data archiving services whereby a production campaign with its inputs resident on tape, is executed by staging and promptly processing a sliding window of X% (5%?, 10%?) of inputs onto buffer disk, such that only ~ X% of inputs are pinned on disk at any one time.

Data Carousel is one of R&D projects to address High Luminosity LHC distributed data processing challenge (scope, context and scale are different from RHIC), we are working very closely with CERN, 9 Tier-1s, FTS and dCache teams

Ultimate goal : use tape more actively and efficiently

Cycle through tape data, processing all queued jobs requiring currently staged data

We focus on ***efficiently*** using the ***available*** tape capacities

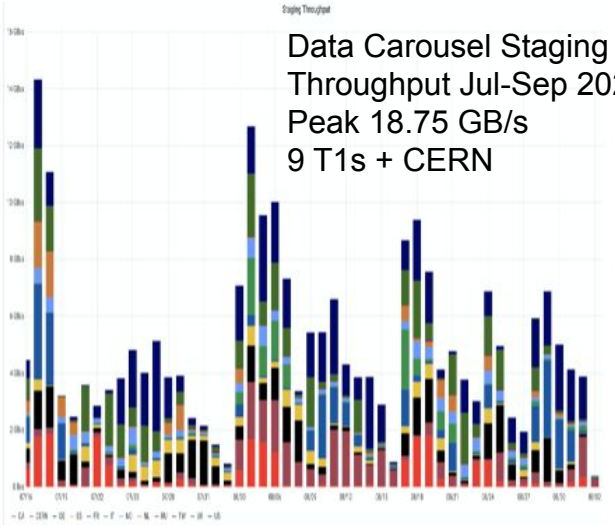
- Introduce little or no performance penalty to tape throughput, after integrating tapes into workflow
- Improve efficiency and throughput of tape systems by orchestrating the various components in the whole system stack, starting from better organization of writing to tapes
- Solutions should scale proportionally with future growth of tape capacity

'Data Carousel' LHC R&D was started in the second half of 2018 → to study the feasibility to use tape as the input to various I/O intensive workflows, such as derivation production and RAW data re-processing

…and "tape" could be any "cold" storage

ATLAS: Data Carousel project phases

- Phase I : Tape Sites Evaluation (2018)
  - Conduct tape staging tests, understand tape system performance at sites and define primary metrics
- Phase II : ProdSys2/Rucio/Facilities integration (2019-2020)
  - Address issues found in Phase I
  - Deeper integration between workflow, workload and data management systems (ProdSys2/PanDA/Rucio), plus facilities
  - Identify missing software components
- Phase III : Run **production at scale** for selected workflows (2020)
  - Run2 data reprocessing in Data Carousel mode : process ~22PB with 3.5 PB disk buffer respecting ATLAS CPU share policy and physics priorities. Run derivation production at all Tier-1s
  - Use iDDS in Data Carousel to decrease time between staging and job submission
  - Provide complete monitoring for Data Carousel and new components
  - Increase file size to improve tape I/O performance
  - Implement new policy to place data on TAPE and DISK



Data Carousel Staging Throughput Jul-Sep 2020
Peak 18.75 GB/s
9 T1s + CERN

'run production at scale' for data reprocessing and derivation production was ***demonstrated in 2020****, and our initial goal to have data carousel in* ***full production for LHC Run-3*** *is reached. We need to address Data Carousel in global (multi-VO) context and to work on smart tape writing*

Alexei Klimentov, BNL

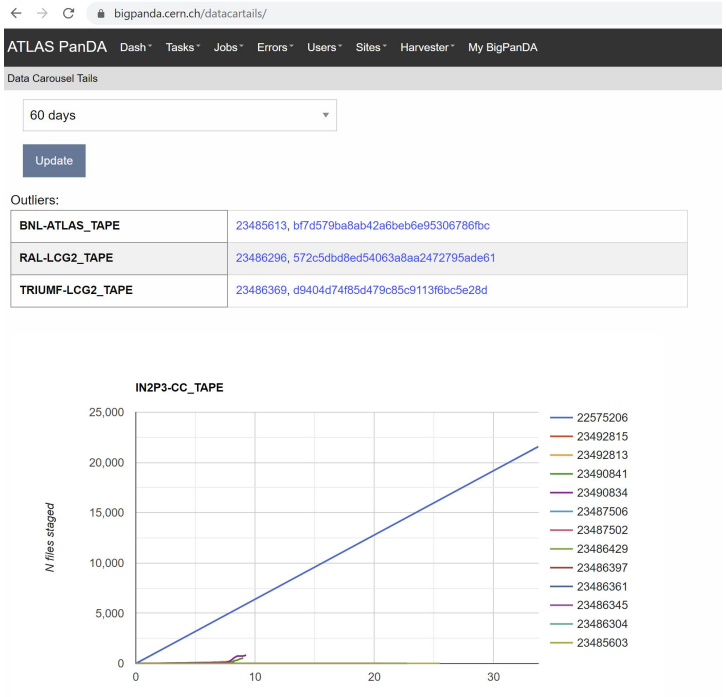
ATLAS: PanDA (and related) monitoring

2020 Milestones

- **Jobs buster.** A tool for root-cause analysis of large losses of walltime or massive fails in jobs, part of an '**Operational Intelligence**' project in ATLAS computing ops
- AGIS -> CRIC migration
- **Data Carousel tails dashboard.** Provides automatic detection of outliers in data transfers
- **iDDS monitoring**. A new set of views for the recently developed data delivery system
- **GPU monitoring.** A common activity with HSF prmon to deliver GPU utilization metrics for jobs running in PanDA

2021 Plans

- **MLFlow monitoring.** Finalize implementing container based setup for MLFlow monitoring of HPO jobs
- **Operational intelligence.** Perform R&D for raising up the precision of Jobs Buster results
- **Documentation.** Improve the BigPanDA documentation to bootstrap users on using PanDA

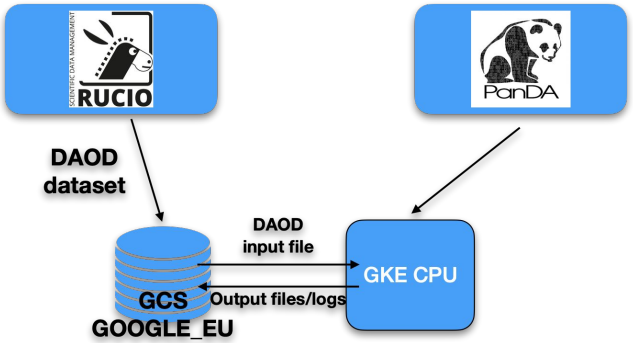


Data Carousel tails dashboard

Sergey Padolski, BNL

ATLAS: R&D project with Google

- New phase of a collaboration extending back to 2013
  - Google cloud for production, “data ocean” global analysis with Google/PanDA/Rucio
- US ATLAS, Google, DOE agreed on a 3-track project in June 2020
  - Automated **data handling between Hot/Cold storage**, data analysis and virtualization
  - New **data formats and I/O** performance studies
  - **Machine learning**, new architectures and algorithms
- Thus far
  - Successful GKE/GCS integration for the first time with full Rucio/PanDA workflow
  - 103 slides of Year-1 retrospective and Year-2 planning :-) Sep 2020
  - Several valuable studies on data formats and ML
  - Important benefit to Rubin (LSST) work, later slides
- Year 2 planning underway
  - Focus on analysis facilities and interactive analysis using compact data formats

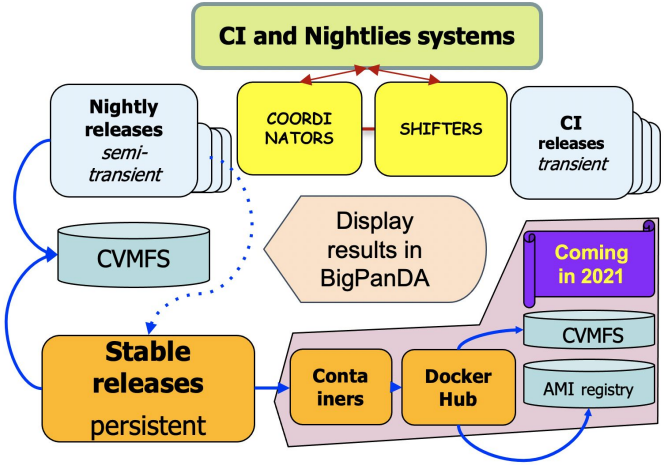


ATLAS: Software release building

- CI and Nighty systems under NPPS responsibility
- Technology-rich software development workflow
  - GitLab, Jenkins, BigPanDA, CVMFS
  - 3000 participants
- **30 nightly branches** with local & Grid testing
- Up to **100 CI builds daily** for all MR updates
- In 2020
  - 26000 build jobs completed
  - 50 million packages compiled
  - Intelligent workflow and job scheduling implemented
    - CPU use efficiency increase by a factor 1.7

2021 plans

- Workflow streamlining, transition to GitLab CI
- Containerization of all releases
- Solutions for container storage, registration, validation



Alex Undrus, BNL

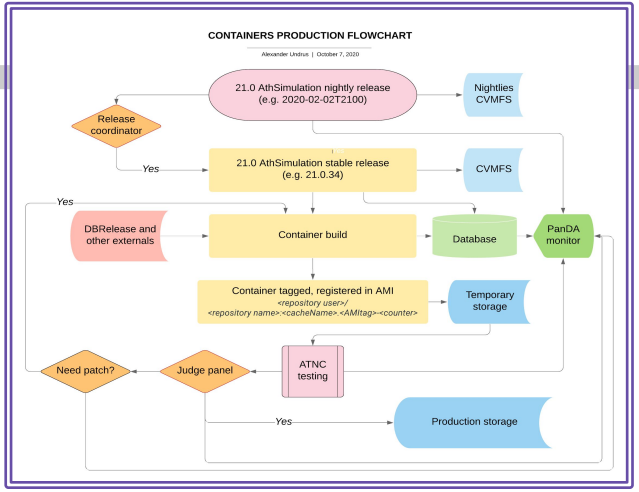
ATLAS: Release containerization

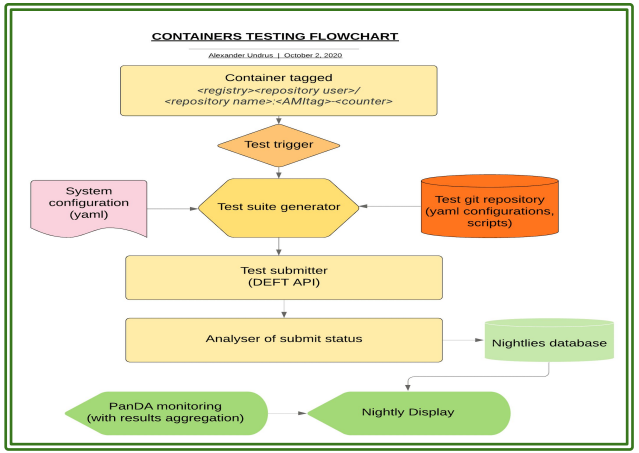
ATLAS Release Containerization Task Force

- Formed in Feb. 2020, Alex Undrus co-coordinating
- Goal: develop procedure for ATLAS releases containerization
  - Automated, integrated into release building workflow
  - “Fat” containers suitable for use on HPCs
- Containers for stable and nightly releases are built
  - Already used for production at Munich, Prague, NERSC
- Containers are catalogued, stored on DockerHub and CVMFS
- ATNC (ATLAS Testing Nightly Containers) framework is created

2021 plans

- Develop sustainable container storage, retention policies
- Work with BigPanDA monitoring team on ATNC dashboard
- Make use of the asetup tool for generation of the container runtime environment





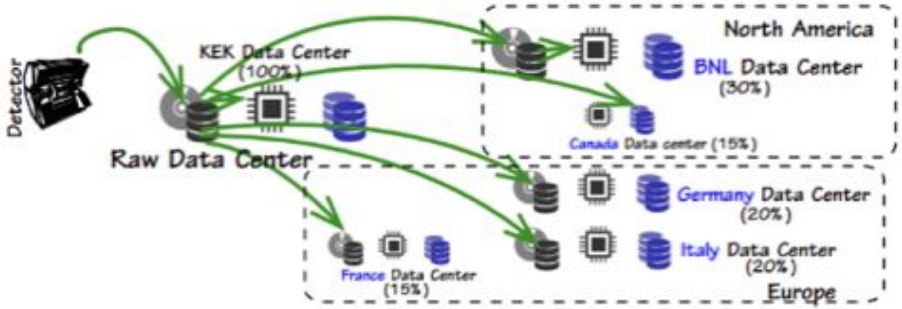
Alex Undrus, BNL

ATLAS: Challenges ahead

- **Analysis** facilities and analysis workflows
  - User analysis beyond the Grid
  - New analysis workflows
  - Run analysis payloads in containers
  - Run end-user analysis in clouds
    - Demonstrate features and capabilities not available (or limited) on WLCG sites and university resources
    - For example, use of ML tools, analytics and other features of Google Cloud Platform for physics analysis
- **Dynamic Data Placement** and Dynamic Data Replicas Handling
- Deeper **Rucio, file transfer, ESnet** integration (and global monitoring)
- ATLAS core SW releases to support **multithreaded** processing
- **Data Carousel** as a base model for production and analysis workflows
  - Data Carousel in multi-VO context
  - Automation in tails handling
  - Data Carousel for end-user analysis tasks

- **Decrease operational effort** to support distributed computing
- EuroHPC/Exascale initiatives and WLCG. **How efficiently can we use new generation of HPCs**?

Belle II: Data and calibration



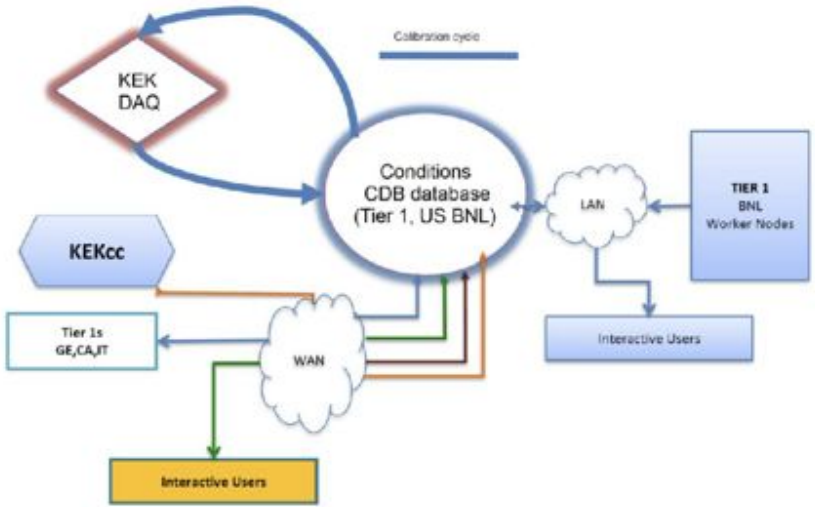
- SDCC@BNL is currently the only raw data center outside of Japan, that changes next year (see figure)
- Prompt calibration moved from KEK to SDCC in 2020 (with EDG/NPPS influence and help)

- Belle II is a very successful example of EDG/NPPS/SDCC collaboration

- Belle II **US PI** is Jaffe (EDG)
- Belle II US **L2 manager for computing** is Laycock (NPPS)
- Belle II **US Tier1** is in SDCC
- **Distributed data management** and **conditions database** software development effort is in NPPS

Paul Laycock, BNL

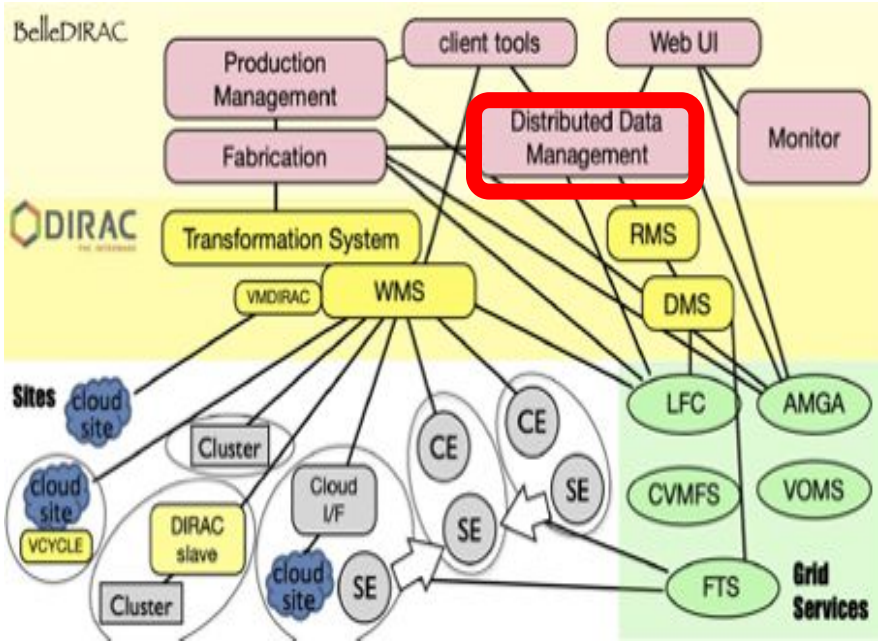
Belle II: Conditions database



- Belle II conditions database operated from SDCC
- Software development in NPPS (Ruslan)
  - Added running global tags in 2020
  - Adding finer granularity authorisation in 2021
- In 2021, Laycock will take responsibility for the CDB client side tools
- Potential to leverage this software to support other experiments like sPhenix and DUNE
  - Plenty of details to work out in 2021

Paul Laycock, BNL

Belle II: Distributed data management



- Custom DDM software upgraded by NPPS (Sergey, Ruslan) to cope with early Belle II data
- Upgraded to use Rucio thanks to a new Dirac plugin (Cedric) and updated Belle II Dirac APIs (Ruslan)
- LOTs of testing!
  - Belle II is a running experiment
- Transparent migration planned for January 2021, enabling
- Many new developments for Belle II that may be useful for other experiments, already merged into main repository
  - Chained subscriptions, lightweight monitoring...

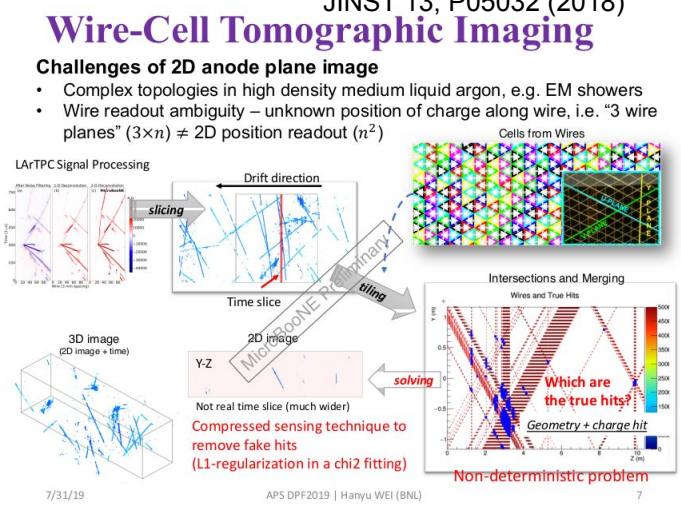
Paul Laycock, BNL

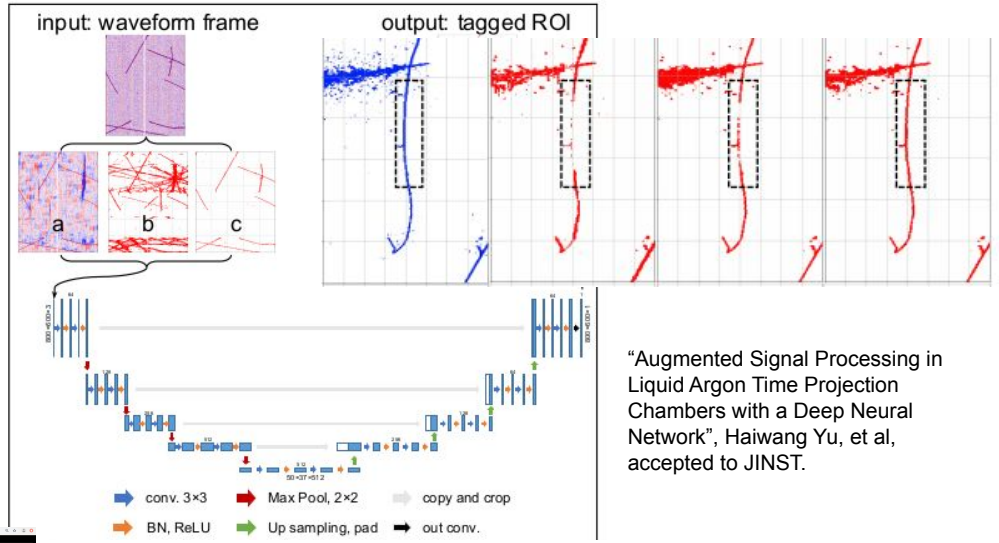
DUNE: Wire-Cell Toolkit, Prototype and R&D Software

- Toolkit: general, portable, ST/MT data-flow programming component based
  - Main users: ProtoDUNE/DUNE, MicroBooNE, support for ICARUS + R&D detectors
  - Batteries: best available wire-LArTPC simulation and signal processing, includes noise sim and filter, 3D tomography, 3D clustering, file I/O, logging, configuration, etc
- Prototype: basis for one of MicroBooNE's leading analyses
  - WC team generalizes and optimizes prototype s/w and "ports" it to toolkit for all LArTPC
  - Includes high-level 3D reco, ionization/scintilation matching, pattern recognition, dE/dx
  - High-efficiency signal recovery via AI/ML, AI/ML vertex identification.
- R&D with Wire-Cell
  - Add support for PCB-anode R&D detector, new Garfield++ field calculator,
  - HEP-CCE/PPS investigations into Kokkos using WCT as test mule.
  - BNL LDRD award to EDG/CSI for precision GAN-based sim/data systematics
    - GAN to "Augment" sim to be more data like, drive sim improvements
    - Quantify residual systematic differences, propagate them through reco
    - Initially target neutrino/LArTPC then sPHENIX jet reco / TPC sim

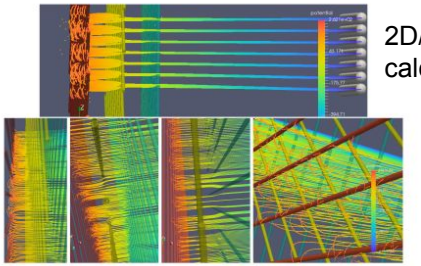
DUNE: Various Wire-Cell products

JINST 13, P05032 (2018)

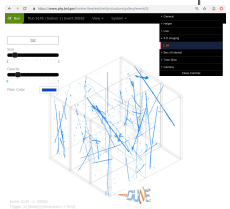




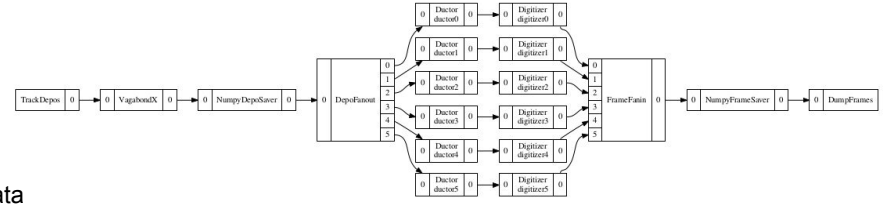
"Augmented Signal Processing in Liquid Argon Time Projection Chambers with a Deep Neural Network", Haiwang Yu, et al, accepted to JINST.



2D/3D field calculations



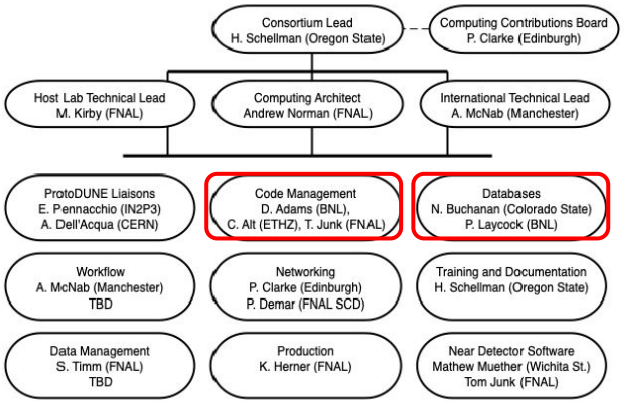
"Bee" 3D WebGL based data visualization, Chao Zhang



Example Wire-Cell Toolkit multi-threaded data flow processing graph.

Brett Viren, BNL

DUNE: S&C design and development



CDR Chapter 4

Frameworks (Andrew Norman, Paul Laycock, Mathew Muether)

Chapter 5

Databases - (Paul Laycock and Norm Buchanan)

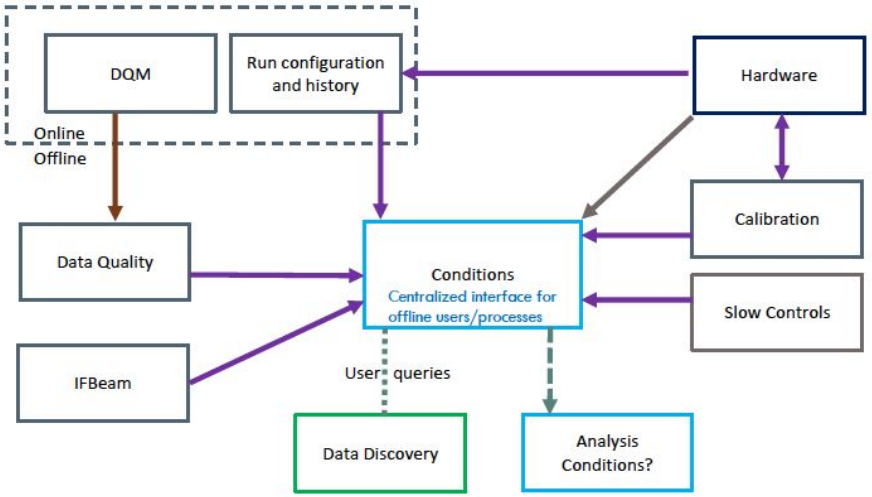
Chapter 18

Code Management - (Tom Junk, Mathew Muether, David Adams..)

- Wire-Cell as described
- NPPS in DUNE S&C org chart
  - Code Management: David Adams
  - Databases: Paul Laycock
- Task forces figuring out S&C design
  - Software framework: Adams, Brett Viren and Laycock (co-chair)
  - Metadata: Laycock
- Data prep and releases: Adams
- Participating in full DUNE CDR planned for early 2021

DUNE: Databases

DUNE offline database sketch



More study and discussion needed to converge on final design.

- Bringing LHC experience to the offline database interface for DUNE, to support distributed computing access patterns
  - Existing solutions at FNAL for different use cases, many of them good
  - Likely improve FNAL solutions rather than adopt anything from BNL, but collaboration comes in many flavours
- Maxim Potekhin and Paul Laycock helped drive the hardware DB requirements for the FNAL upgrade, now in production

Paul Laycock, BNL

Rubin Observatory (former LSST)

Rubin has high demand computing requirements:

- Core hours yearly: 4.5^7 (in 2023) to 4.1^8 (in 2032)[1]
- Storage (Normal + High Latency + Latent) 130PB in 2023 to 2,376PB in 2032[1], cumulative processed data size is 500PB[2]

Few calls in 2020:

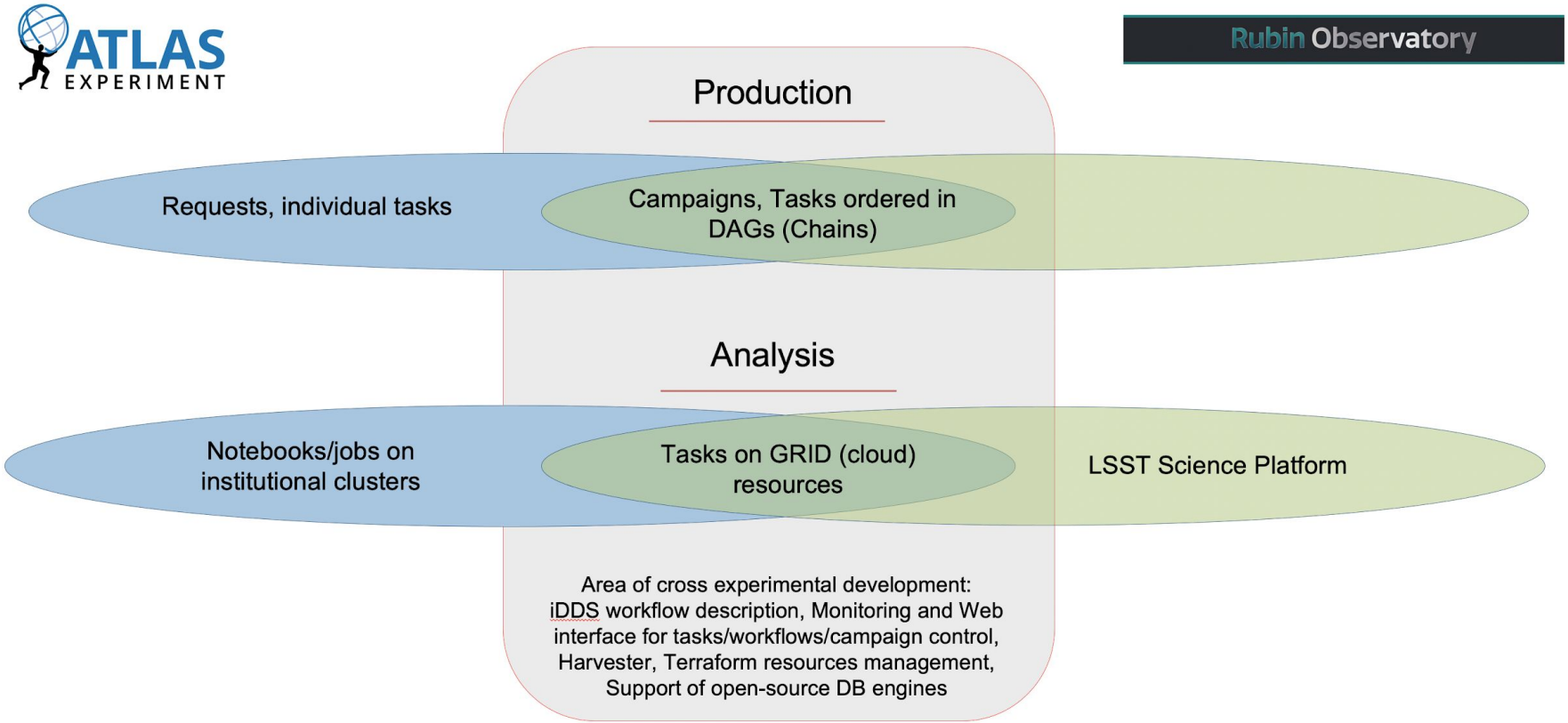
- Interim Data Facility (IDF). Selected to mitigate risks and correspondent delays of selection the US Data Facility. A 3 years contract was finally assigned to Google
- US Data Facility. A contract assigned to SLAC in October
- Workflow/Workload Management Platform. Under selection. Should support Rubin data processing pipeline and manage computation on few computing centers

BNL Participated in all calls, NPPS focused on evaluation PanDA as W(F)MS

[1] https://dmtn-135.lsst.io/

[2] https://arxiv.org/pdf/0805.2366.pdf

Rubin: Workflow management synergy with ATLAS



Sergey Padolski, BNL

Rubin: NPPS milestones

- DOMA PanDA instance was tuned up to work in testing mode (PanDA server/JEDI, PanDA monitoring, iDDS, Harvester, DOMA version of Pilot, CRIC) *(thanks to Tadashi and other people)*
- First tests were performed on BNL GRID resources
- Google + PanDA + Rubin proof of concept project:
  - Reused the ATLAS + Google R&D results in the K8S Harvester features
  - Container-based deployment of Rubin framework
  - Developed support of DAG in iDDS
  - Performed a load tests and optimization of a whole processing system
  - Performed processing on the HSC RC2 dataset (Subaru Hyper Suprime-Cam) on Google Cloud
- Development of the PanDA plugin for Rubin framework and deployment configuration (in process)

Rubin: NPPS plans for 2021

- Foremost objective, convince them to adopt PanDA officially :-) (circa June)
- Finalize initial **integration of PanDA and Rubin** framework
- Start working with Rubin Security Office to identify and fix potential misalignment between created setup and Rubin **security requirements**
- Start working with system integrator providing Terraform deployment for LSST to identify possible perspectives for dynamic resources provisioning and allocation with Rubin (DOMA) PanDA setup
- Follow to updates of the Rubin Framework Batch Processing subsystem to keep **workflow management** plugin up to date
- Develop **functional tests** setup to run periodically test of the whole system against new software releases and middleware modifications
- Identify the **functionality currently missing** in the iDDS + Rubin FW based workflow management (e.g. production **campaign management**), design and implement missing functionality in different components
- Depending on decision on the **DB Server** (Oracle, Postgres, MySQL) support migration of core components to the new backend
- Conduct **high load tests** to make sure that setup is ready for working in highly concurrent environment

Sergey Padolski, BNL

Common software

- We cultivate common software projects in-house and in our collaborations
- HSF, WLCG, IRIS-HEP, HEP-CCE etc.
- Bold at right are common software we're involved in

ACTS

Athena

ATLAS CI and nightly system

Conditions database

Data Carousel

DIRAC

DUNE prompt processing system and Data Quality Monitor

dunetpc

EicRoot

eic-smear

ATLAS event service

Fun4All

Harvester

Intelligent data delivery system (iDDS)

Invenio

Jekyll

LXR code browser

PanDA workload management system

PanDA monitoring system

PanDA pilot

Pegasus

ATLAS production system

ROOT

Rucio

Scikit-HEP

Wire-Cell

Zenodo

Plans and priorities (largely with my NPPS hat on)

- **Sustaining and improving software and operations** in our running experiments: ATLAS, Belle II, STAR
- Continuing to **grow cross-experiment contributions**, common activities, collaborations
- **sPHENIX readiness** as BNL's next major experiment, to start up in 2023
- The intensive development program remaining for **HL-LHC readiness**
- **Supporting the EIC** as it moves to detector collaborations and the detailed design phase
- Growing **new opportunities**: Rubin Observatory, DUNE, ...
- Building **DAP** as an integral part of the experiment life cycle, from PHENIX to others
- Building **ML** experience and activity, beyond the scaled-up ML services work we're doing
- Solving challenges where software meets the evolving **technology landscape**
- **Supporting and developing software as an integral part of the science programs of the BNL Physics Department**

Finally

- I’m out of time!

Many thanks

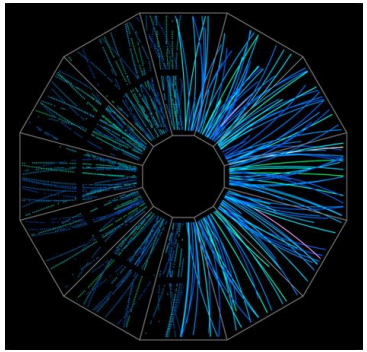
- Many thanks to those in the BNL Physics Department and elsewhere who contributed to this talk, and the many more who have contributed to the work described
- An inevitably partial list of contributors:
  - Jaroslav Adam, David Adams, Dmitry Arkhipkin, Fernando Barreiro Megino, Kaushik De, Johannes Elmsheuser, Yuri Fisyak, Jin Huang, Amol Jaikar, Kolja Kauder, Hongwei Ke, Alexander Kiselev, Alexei Klimentov, Eric Lancon, Paul Laycock, Tadashi Maeno, Ruslan Mashinistov, Paul Nilsson, Marcin Nowak, Joe Osborn, Sergey Padolski, Victor Perevoztchikov, Chris Pinkenburg, Matt Posik, Maxim Potekhin, Ofer Rind, Cedric Serfon, Dmitri Smirnov, Alex Undrus, Gene Van Buren, Brett Viren, Jason Webb, Shuwei Ye

Supplemental

STAR: geant4star example unit test: TPC hits

Unit tests focus on provable behavior from known inputs to the simulations.

A muon thrown down the center of each TPC sector will encounter 72 sensitive gas divisions, leave an energy deposition in each of them, and have a volume ID from which we can decode the struck element of the detector.



Test for a single hit illustrated below. Summary of all tests for TPC illustrated at right.

```
 ------------------------------------------------------------------
  - given - A hit on that track
 [The hit should have a nonzero volume_id] id=75 vid=772 de=0.000004 ds=2.198872  -passed-
 [The hit should have an energy deposit > 0]  -passed-
 [The hit should have a path length > 0]  -passed-
 [The hit should have a nonzero momentum]  -passed-
 [The hit should have a nonzero log10(gamma)]  lgam=3.727149 (needs to be filled)  -passed-
 [The hit should have a length > 0]  -passed-
 [The hit should have adc, pad and timebucket set to zero]  -passed-
 [Track's momentum at hit should be < initial value]  -passed-
 [Hit position should be w/in the fiducial volume of the sector] (-94.549805 -163.875229 98.589195 / in phi=1 r=1 -passed-
 [The padrow should be 1 <= pad <= 72]  -passed-
 [The sector should be 1 <= sector <= 24]  -passed-
 [The detector state is in (0,1,2)]  -passed-
 [The decoded sector number should be 7]  sector=7 -passed-
```

pass	fail	total	track tests
24	0	24	A muon must have been processed by geant
24	0	24	The track should have a start vertex
24	0	24	The start vertex should be in vertex table
24	0	24	There should not be a stop vertex in the TPC
24	0	24	The start vertex should be on the z-axis
24	0	24	The track should be primary
12	0	12	The track should have an eta=0.500000
12	0	12	The track should have an eta=-0.500000

pass	fail	total	hit tests
1824	0	1824	The hit should have a nonzero volume_id
1824	0	1824	The hit should have an energy deposit > 0
1824	0	1824	The hit should have a path length > 0
1824	0	1824	The hit should have a nonzero momentum
1824	0	1824	The hit should have a nonzero log10(gamma)
1824	0	1824	The hit should have a length > 0
1824	0	1824	The hit should have 0 adc, pad and timebkt
1824	0	1824	Track's momentum at hit should be < initial
1824	0	1824	Hit position should be w/in fiducial volume
1824	0	1824	The padrow should be 1 <= pad <= 72
1824	0	1824	The sector should be 1 <= sector <= 24
1824	0	1824	The detector state is in (0,1,2)

Typical unit tests execute w/in a few minutes, permitting rapid validation of code modifications.

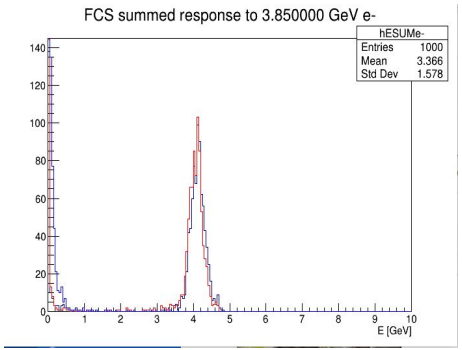
Jason Webb, BNL

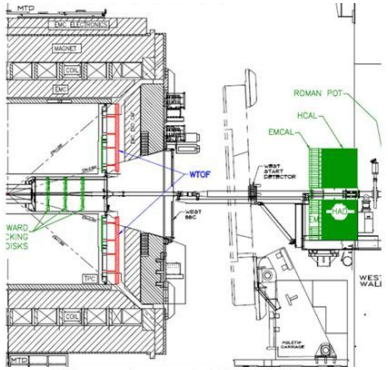
STAR: geant4star comparative test: Forward calorimetry

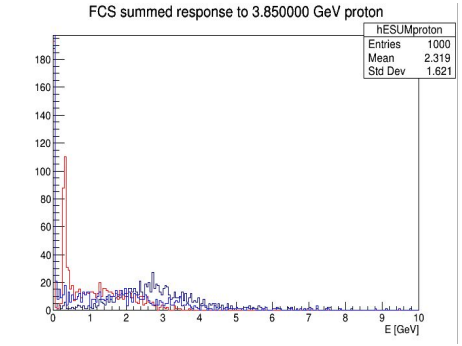
Comparative tests address the questions of whether the response of the detectors to incident particles agrees with data, existing simulations test beams and/or our expectations.

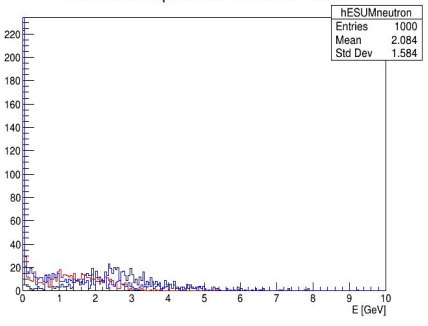
The first STAR calorimeter simulations using a fully negrated geant4 application are illustrated at right. Energy sums (using fixed, un-tuned sampling fractions) for single 3.85 GeV electrons, protons and neutrons are shown.

Results are reasonably consistent with expectations, and will improve as subsystem experts begin to tune the simulations.









EM calorimeter response (@ sampling fraction of 0.20)
Hadronic calorimeter response (@ sampling fraction of 0.0145)
Summed calorimeter response

Jason Webb, BNL

STAR: geant4star test coverage

Unit Tests

- Setup known configuration of particles
- Run through known detector setup
- Checks for *provably correct* results in data structures
  - e.g. known calorimeter cell registers energy

Comparative Tests

- Accumulate a set of plots for each subsystem
- Compare results between two different simulation packages
- With sufficient statistics, demonstrates consistency between packages, and correctness where answers are independently known

Current set of tests are reasonably complete, and are sufficient for supporting FCS studies in 2021. Additional testing and tuning to be done by subsystem experts as we reconcile the simulation to data.

subsystem	mcEDM unit tests	comparative
TPC[0]	passing	highly consistent
FTS (Si)	passing (simplified test)	TBD
FTS (sTGC)	passing (simplified test)	TBD
FCS	passing	reasonable (needs tuning)
BEMC	passing[3]	reasonable (stat limited)
EEMC	passing[2]	reasonable (stat limited)
MTD	passing[4,5]	skeleton
EPD	Volume ID scheme valid?	skeleton
BTOF	passing[4]	reasonable 1/b vs p
VPD (tof start)	TBD	TBD
ZDC	N/A Geometry model is the wrong detector design and technology, ZDC exists outside of the world volume, Beamline magnet *fields* are not implemented	

[0] Need to test the prompt-hits region
[1] Waiting on final geometry model(s)
[2] First eta-ring does not see hits in all 5 z-sections of the calorimeter
[3] Few hits in neighboring towers
[4] Energy loss=0 in some cells
[5] Not entirely sure of the expected volume IDs

Jason Webb, BNL