BNL Perspective : High Throughput and High Performance Computing at BNL

Alexei Klimentov

Scientific Computing and Data Facilities Division Director (Interim)

HEP/HPC Strategy Meeting

CERN, January 30, 2025

Outline

- BNL
  - Multi-purpose Lab
  - Nuclear and Particle Physics Experiments
    - From RHIC and ATLAS to DUNE and ePIC
- Projects at BNL*
  - REDWOOD Near Real Time Track (data streaming)
  - BNL CDS Advanced Computing Laboratory
  - Hierarchical, AI-Enabled Modeling of Future Supercomputers
  - HEP-CCE projects at BNL
  - BNL and UMass experience and AI/ML needs for GPUs
- Summary and conclusion

Glossary and Abbreviations (link)

*relevant to today's meeting

Brookhaven National Laboratory

Brookhaven Supports Data-rich Experimental and Computational Facilities and Programs

PHYSICS

Relativistic Heavy Ion Collider (**RHIC**): Supports nearly 2000 scientists worldwide

Electron-Ion Collider (**EIC**): Echelon0 and Echelon1 data storage (&archiving) and processing center, a new paradigm in physics and frontier for data science

Large Hadron Collider (**LHC**): Largest ATLAS Tier-1 center

Super KEKB: Tier-1 and RAW data storage center for high energy physics Belle II experiment

Quantum chromodynamics (**QCD**): computing facilities for Brookhaven Lab, RIKEN, and U.S. QCD communities

Basic Energy Sciences (BES) and Biological and Environmental Research (BER)

National Synchrotron Light Source II (**NSLS-II**): Newest and brightest synchrotron in the world; supports a multitude of scientific research in academia, industry, and national security

Center for Functional Nanomaterials (**CFN**): Combines theory and experiment to probe materials

Accelerator Test Facility (**ATF**): User facility for for advanced accelerator and laser research

Atmospheric Radiation Measurement (**ARM**) program: Partner in multi-site facility, operating its external data center

RHIC → EIC

LHC

QCD

NSLS-II

CFN

Brookhaven National Laboratory

Slide : N.D'Imperio et al

BNL Nuclear and Particle Physics Experiments

- RHIC : STAR : 1995 – 2025; sPHENIX : 2023 – 2025 (DAQ → Computing Center Data Streaming with data rate to tape higher than (per experiment) for LHC Run3)
  - **Experience builds know-how to support similar high-rate programs in the future**
    - sPHENIX data streamed from the detector (DAQ) to the computing center
    - RHIC experiments managed tape storage volume ~220 PB and it will be ~0.5EB by the end of 2025
  - Experiments at RHIC startup : BRAHMS, PHOBOS, PHENIX, STAR
- LHC : ATLAS : 2008 – 2026 / HL-LHC : 2030 – 2041
  - ATLAS is about 40% of BNL Computing Resources
- Super KEKB : Belle II : 2018 – 2030
- EIC : ePIC : Data Streaming Model and concept of Echelons (under development)
- DUNE : Computing model under development

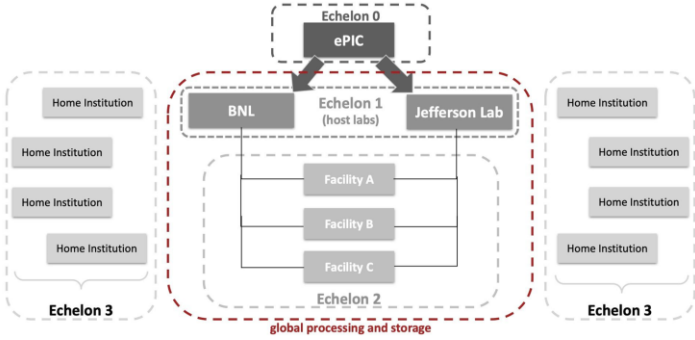
A unique situation, running (for more than 25 years) experiments in particle and nuclear physics, and new experiments with new ideas and new approaches. At the same time, research in other fields (ASCR, BER and BES)

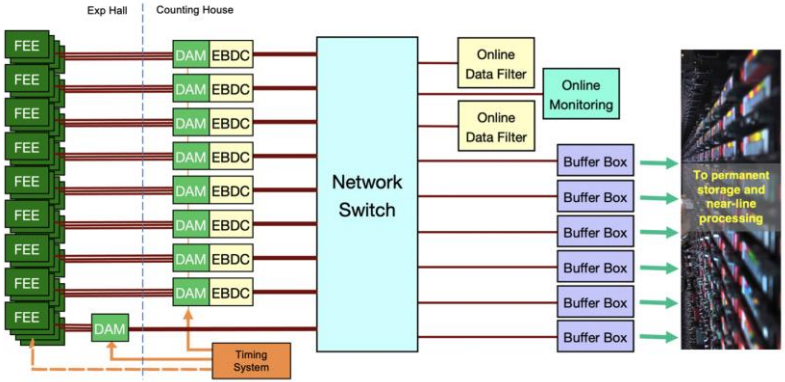
But the same supercomputing infrastructure in use by ALL

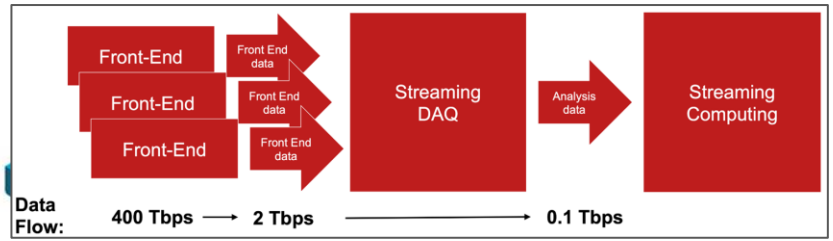
Brookhaven National Laboratory

ePIC Streaming Readout and Computing Model : Maximizing Physics Reach

- EIC luminosity is high, the cross section is not
- **Tractable to read out ~100% of the events**
  - Capture every collision event: they are all of physics interest
    - Backgrounds are substantial, they are in the data stream too
  - A **complete, unbiased event sample**
  - There is substantial noise reduction, compression in DAQ to reduce needed storag
- Data is streamed to prompt reconstruction for **early holistic view of the data**
  - Reconstruct, monitor/diagnose, calibrate, analyze as quickly as possible, O(1min)







The two-host-lab organization motivates the 'butterfly' model: BNL and JLab are symmetric peers

Echelon 0: ePIC experiment, DAQ system

Echelon 1: Two host labs, two primary ePIC computing facilities

Echelon 2: Global contributions leveraging commitments to ePIC computing from universities and labs domestically and internationally

Echelon 3: Supporting the analysis community where they are at their home institutes, primarily via services hosted at E1s/E2s

Slide : T.Wenaus et al

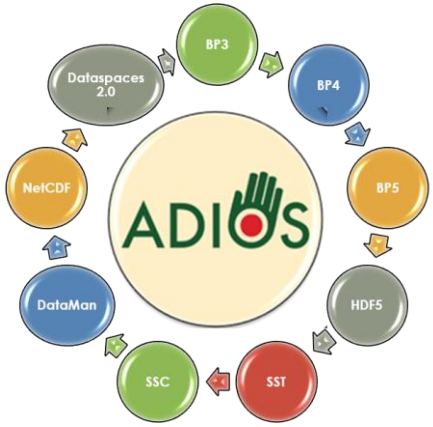
DOE ASCR REDWOOD Project 2024-2028 : Developing an Innovative Computing Ecosystem to Impact Science

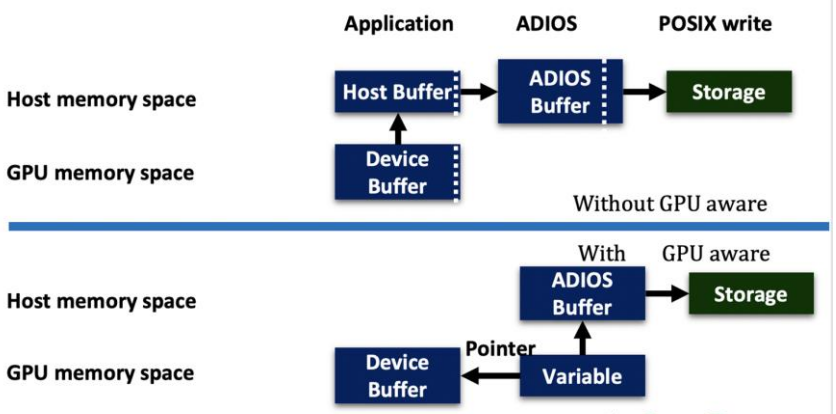
- Three National Labs, Brookhaven, Oak Ridge and SLAC, and three universities, Carnegie Mellon, University of Massachusetts Amherst, and University of Pittsburgh
- Lead PI A. Klimentov (BNL) with co-PIs A. Hoisie, T. Maeno and S. Yoo (BNL); S. Klasky (ORNL); and W. Yang (SLAC); as well as a BNL team of computing scientists, IT engineers, and physicists
  - Four Tracks : Near Real Time WF; HighThroughput WF, Monitoring and Integration; System (WMS and Distributed Computing Modelling
- Ecosystem of research platforms connected to address scientific challenges involving complex workflows and exabyte data volume in heterogeneous computing environments.
- Data placement and complex workflows created by brokering and partitioning algorithms and associated runtime to help science applications better exploit architectural features found in DOE's computing infrastructure.
- Incorporate timely new algorithms to provide near-term high impact on science (with domain scientists).
- Software and algorithms will be demonstrated at scale for several scientific domains, e.g., particle physics, astronomy, and nuclear fusion.
- Make software and algorithms easily available to the research community for broader, long-term impact.

Brookhaven National Laboratory · Oak Ridge National Laboratory · SLAC · University of Massachusetts Amherst · Carnegie Mellon University · University of Pittsburgh

REDWOOD Near Real Time WF : ADIOS: high-performance publisher/subscriber I/O framework:

- Easy-to-use, high performance I/O abstraction to allow for on-line/off-line data subscription service
- Sustainable production software for self-describing data-streams
- Declarative, publish/subscribe API separated from the I/O strategy
- Multiple implementations (engines) provide functionality and performance
- Rigorous testing ensures portability
- GPU-aware to reduce data movement
- Scalable I/O from laptop to exaflop computers
- https://github.com/ornladios/ADIOS2

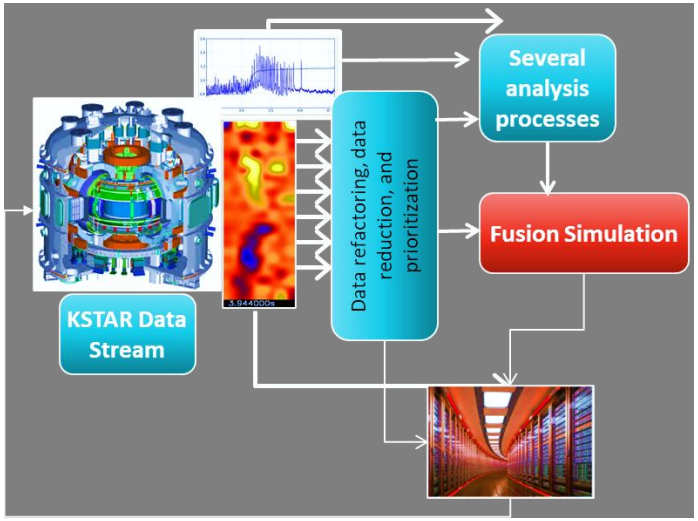




Godoy, W. F., Klasky, S. ,et al. (2020). ADIOS-2: The adaptable input output system. a framework for high-performance data management. SoftwareX, 12, 100561

Slide : S.Klasky, N.Podhorszky et al

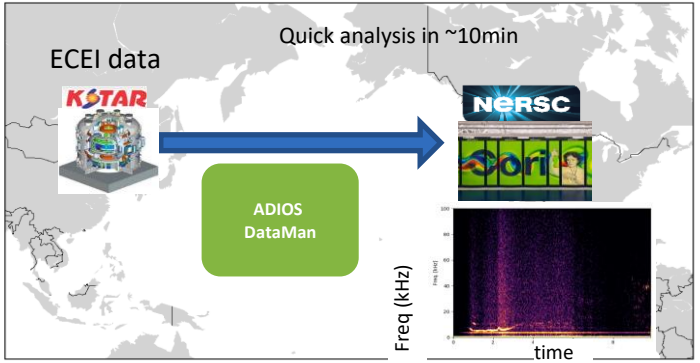
Near-real time networked analysis of fusion experimental data (KSTAR)

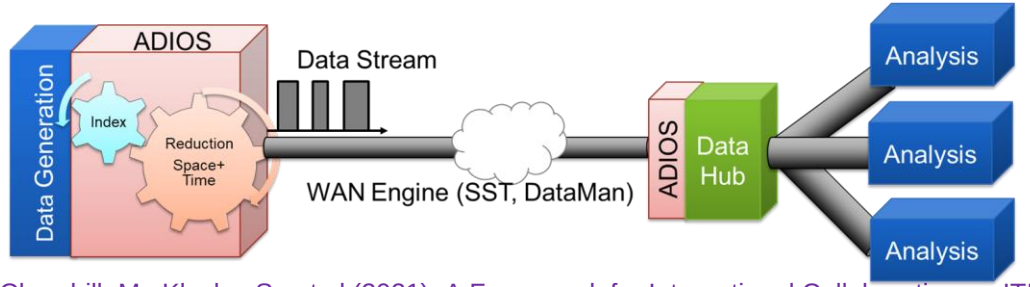


Near real-time streaming analysis of **KSTAR** data

– Through a **WAN** from South Korea to NERSC

Use of ADIOS **DataMan** Engine ⟶ **500 MB/s**





Reduced analysis time from 12h to 10 minutes

Slide : S.Klasky, N.Podhorszky et al

Churchill, M., Klasky, S., et al.(2021). A Framework for International Collaboration on ITER Using Large-Scale Data Transfer to Enable Near-float-Time Analysis. Fusion Science and Technology, 77(2), 98-108.

BNL CDS. Advanced Computing Laboratory (ACL): A Unique Capability

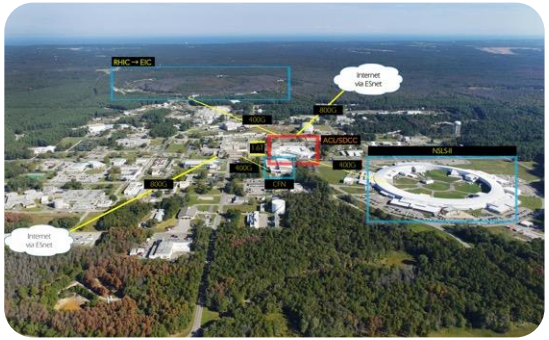
- ACL is ready to ingest data directly from experiments in RHIC/EIC facilities, in NSLS-II and CFN in a new computer lab space.
- CAT-AI Integrated in the HAI-FI initiative as a main thrust area.
- Collaborative space for codesign of new data-intensive technologies for experimental science with DOE facilities and industry, as well as funding from other federal agencies.

Brookhaven National Laboratory

Slide : A.Hoisie et al

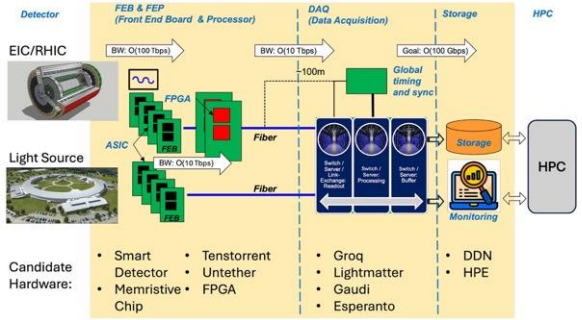
BNL Research Project. Codesign in Action for Experimental Science Computing: Architectures, Systems, and Testbeds

Advanced Computing Lab

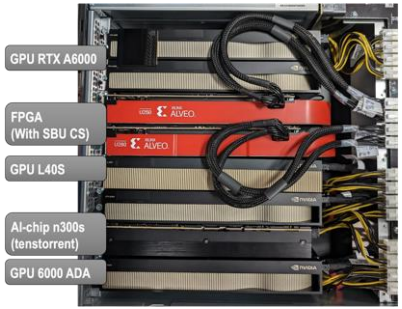


Unique collaborative testbed facility with access to live, actual data from diverse experiments, such as CFN (microscopy), NSLS-II, and RHIC/EIC, for codesign of architectures and experimental workflows.

Experimental Science Workflows

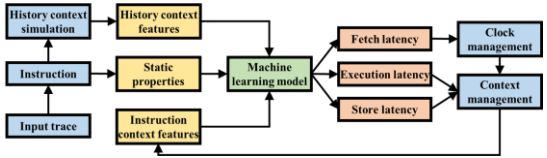


Extreme data challenges, heterogeneity, large spectrum of spatial and temporal computing scales from the edge to the extreme – real-time to long computational campaigns.

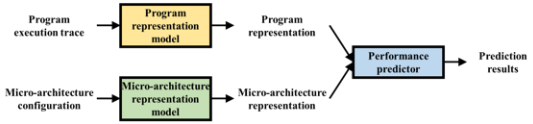


AI-based Modeling and Simulation

Leading the charge with SimNet and PerfVec
Accurate simulation faster by orders of magnitude compared with Discrete-Event Simulation



SimNet: AI-based architecture simulation https://github.com/lingda-li/simnet



PerfVec: AI-based Architecture modeling https://github.com/PerfVec/PerfVec

Li, L. S. Pandey, T. Flynn, H. Liu, N. Wheeler, and A. Hoisie. 2022. SimNet: Accurate and High-Performance Computer Architecture Simulation using Deep Learning. POMACS 6(2):Article 25. DOI: 10.1145/3530891.

Pandey, S., L. Li, T. Flynn, A. Hoisie, and H. Liu. 2022. Scalable Deep Learning-Based Microarchitecture Simulation on GPUs. SC22, pp. 1-15. DOI: 10.1109/SC41404.2022.00084.

Brookhaven National Laboratory

Staff: Lingda Li; Thomas Flynn; Yuhui Ren; Alexei Klimentov; Adolfy Hoisie

Slide : A.Hoisie et al

HEP-CCE Project : Making HPC More Accessible

- HPC systems are getting more diverse and complex, in terms of both the hardware and software environments
  - **Hardware:** Processing units: CPU, GPU, APU, etc, diverse memory hierarchies
  - **Software:** Different programming models, software modules, job manager, authentication/authorization
  - There is a need to enhance the current software ecosystem to make HPC more accessible to NPP users

- **Goal**: Develop cross-cutting solutions to meet HEP computing needs with advanced HPC capabilities, leveraging DOE ASCR and HEP technologies and incorporating latest AI/ML advancements.

- Part of the **HEP-CCE** project, BNL is investigating portable workflow management solutions that can enable seamless workflow executions across Grid and LCF sites (and cloud computing in the future)
  - ATLAS and DUNE workflows as initial test cases
  - For ATLAS workflows, leverage the **PanDA** workflow management system and **Globus Compute** (FuncX)
    - *This is in service of AID2E as well as ATLAS, to get AID2E workflows onto HPCs, in particular Perlmutter*
  - For DUNE workflows, investigate integrating **Superfacility API** with existing workflow management systems
  - Also closely monitor and coordinate with the IRI development

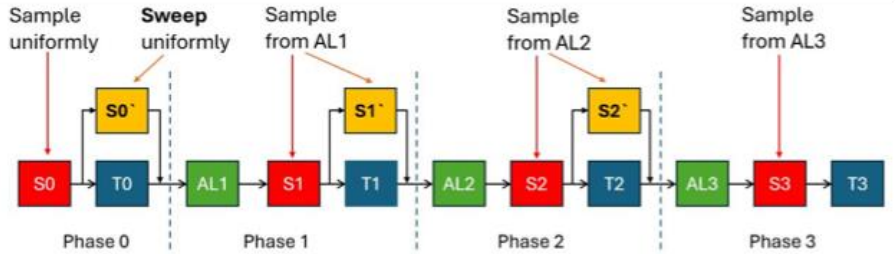
- We also need application software that can take advantage of the massive parallelism on HPC systems
  - Portable programming models are preferred, but also open to vendor APIs if proven advantageous.
  - **Manual porting is time consuming => Leveraging LLMs to automate part of the process**
  - Ongoing work evaluates using open-source/commercial LLMs with RAG to perform various coding tasks: **code documentation, summarization, explanation, porting and optimization.**

Atif, Mohammad, et al. "Evaluating portable parallelization strategies for heterogeneous architectures in high energy physics." *arXiv preprint arXiv:2306.15869* (2023).

Non-NPP HPC Workflows

- BNL's non-NPP modeling and simulation workflows mainly come from
  - Climate science (direct numerical simulations)
  - Materials science (density functional theory)
  - Biology (molecular dynamics simulations)
  - Machine learning training and inference
- Many of them increasingly couple machine-learning payloads with the high-fidelity simulations
  - Workflow orchestration/resource management to maximize performance and/or resource utilization is being developed

Atif, Mohammad, et al. "Fourier neural operators for spatiotemporal dynamics in two-dimensional turbulence." *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 2024.

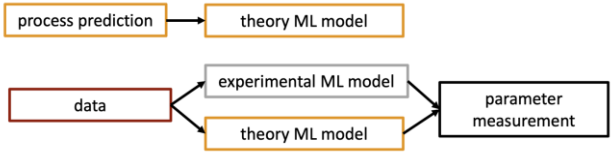


Wang, Tianle, et al. "An Active Learning-Based Streaming Pipeline for Reduced Data Training of Structure Finding Models in Neutron Diffractometry." *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024.

Brookhaven National Laboratory

Slide : M.Lin et al

AI/ML and HPC

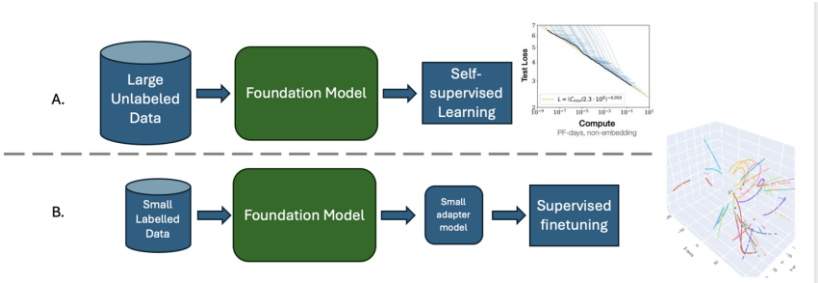


(1) Motivation and challenge :

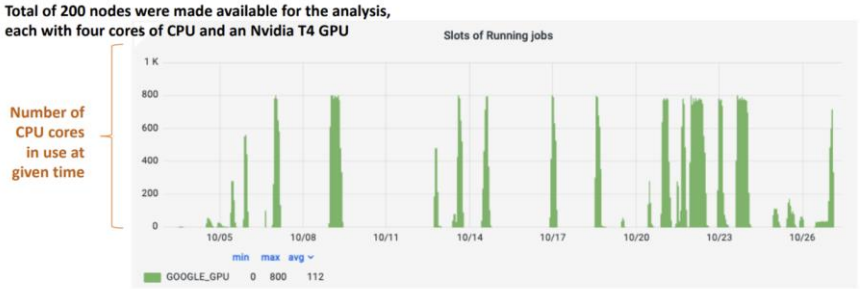
- Towards AI-ready data. Two user facilities (RHIC and NSLS II) and in multiple international science collaboration. There is a trend in applying more AI models on data.
- Large-scale AI model training. Building scientific foundation model. Fast prototyping at local HPC and full-scale training at OLCF, ALCF and NERSC.
- Efficient real-time AI inference. Edge-computing, FPGA, ASCI, high-throughput inference (EIC) and low-latency decision making (NSLS-II).

(2) Motivation and challenge :

- do novel type of analyses and find novel ways to explore HEP data with ML, use large GPU partitions available on HPC



A. Neural scaling behavior build credibility for large scale funding
B. Generalization to downstream task particle tracking application



J.Sandesara, R.Coelho Lopes de Sa, V.Martinez Outschorn, F.Barreiro Megino, J.Elmsheuser, A.Klimentov. ATLAS Data Analysis using a Parallel Workflow on Distributed Cloud-based Services with GPUs https://doi.org/10.1051/epjconf/202429504007

Summary and Conclusion

- High-Performance Computing Facilities represent the most substantial computational resources available to the scientific community in AI/ML era
- HPC systems are getting more diverse and complex, in terms of both the hardware and software environments
  - **Hardware:** Processing units: CPU, GPU, APU, etc, diverse memory hierarchies
  - **Software:** Different programming models, software modules, job manager, authentication/authorization
  - There is a need to enhance the current software ecosystem to make HPC more accessible to NPP users
- HPC integration and HPC interfaces is a challenge for NPP and other users
  - Interoperability of HPC resources
  - Payloads submission, data management and workflows orchestration
  - Cybersecurity requirements
- AI/ML workflows adaptation for HPC is another common topic
  - There is an effort at BNL to design AI/ML workflows to exploit the strengths of accelerator processors for payloads like data analysis and model training
- Data management
  - Data caching, ingestion and export, this also includes data transfer between LCF/HPC centers and data transfer between LCF/HPC and National Labs
- Workflows
  - AI/ML workflows (with high level parallelism) are still the main candidates for BNL NPP community

All of the above will be strongly influenced by the DOE's Integrated Research Infrastructure programme.

Acknowledgements

These slides drew on presentations, proposals, discussions, comments and input from many.
Thanks to all, including those I've missed.

Thanks to D.Benjamin, R.Coelho Lopes de Sa, N.D'Imperio, J.Elmsheuser, A.Hoisie, S.Klasky, M.Lin, C.Pinkenburg, N.Podhoszky, Y.Ren, C.Serfon, T.Wenaus, A.Wong, F.Wurthwein, S.Yoo…

Abbreviations and Glossary

- ACL - Advanced Computing Laboratory at CDS
- AF - Analysis Facilities
- AID2E - AI-Assisted Detector Design for EIC
- ASCR - DOE Office of Science. Advanced Scientific Computing Research
- Belle II – HEP experiment in Japan
- BER - Biological and Environmental Research
- BES - Basic Energy Sciences
- BNL - Brookhaven National Laboratory
- CAT-AI - Center for Advanced Technologies for Artificial Intelligence @ACL
- CFN – Center for Functional Nanomaterials at BNL
- CDS – Computational and Data Science Directorate at BNL
- DUNE – Deep Underground Neutrino Experiment at FNAL
- EIC – Electron Ion Collider at BNL
- ELK - Elastic search, Kibana and Logstash, aka ELK stack
- ePIC – EIC experiment
- HEP-CCE – DOE HEP funded project, Computing Center for Excellence
- Human-AI-Facilities Integration (HAI-FI)
- KSTAR - Korea Superconducting Tokamak Advanced Research;
- LDRD - Lab Directed Research and Development
- NSLS – National Synchrotron Light Source Facilities , Directorate at BNL
- NP – Nuclear Physics
- NPP – Nuclear and Particle Physics, Directorate at BNL
- PD - Program development funds
- REDWOOD – DOE ASCR funded project (2024-2028)
- QCD – Quantum Chromodynamics
- RHIC - Relativistic Heavy Ion Collider at BNL
- SDCC – Scientific Data and Computing Center
- SOW - Statement of Work
- sPHENIX – RHIC experiment
- STAR – RHIC experiment
- WF - workflow
- WMS - Workload Management System