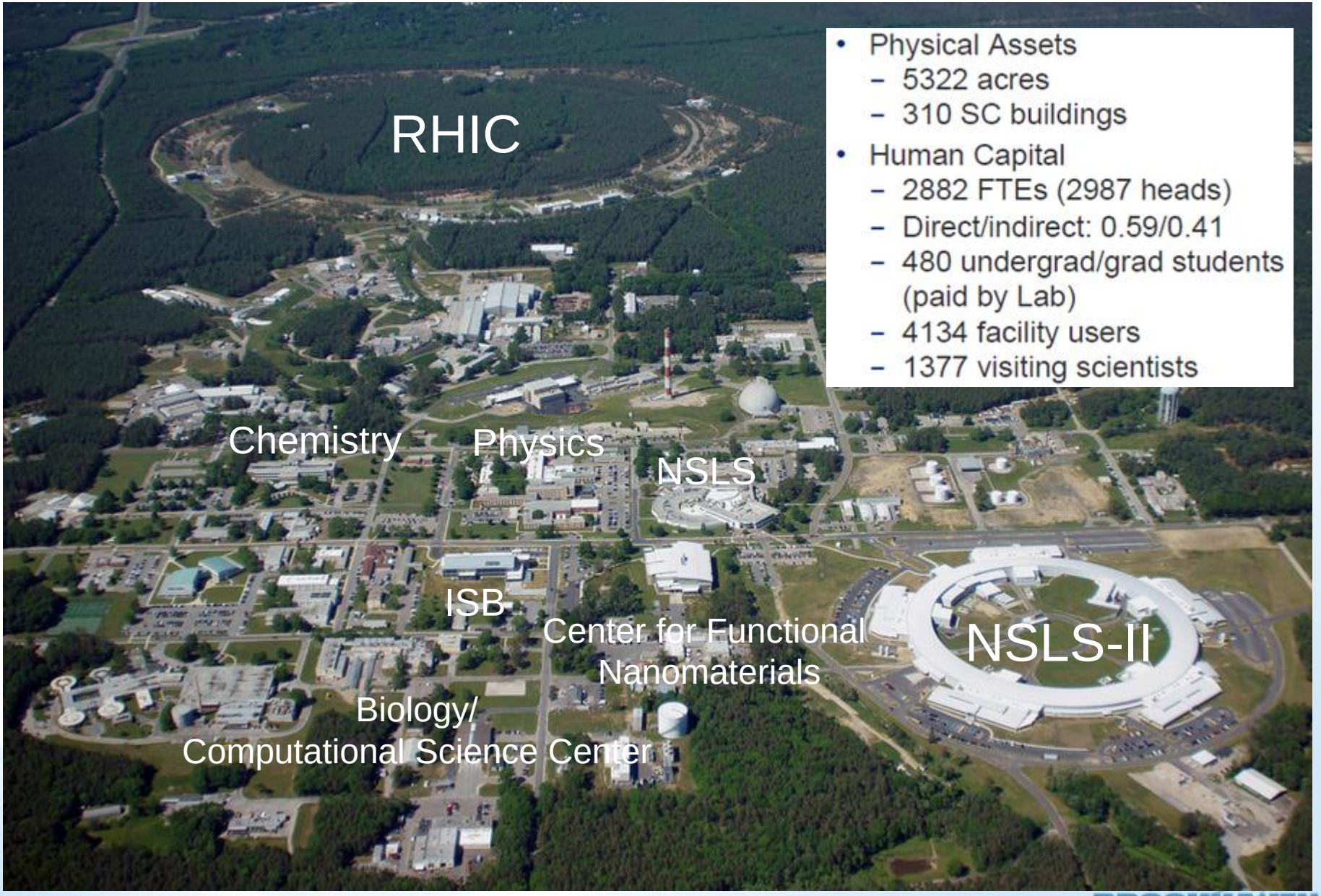
Driving ‘Big Data’ Science to Expand BNL’s Leadership in Discovery and Innovation

Michael Ernst
Brookhaven National Laboratory

BROOKHAVEN NATIONAL LABORATORY

a passion for discovery

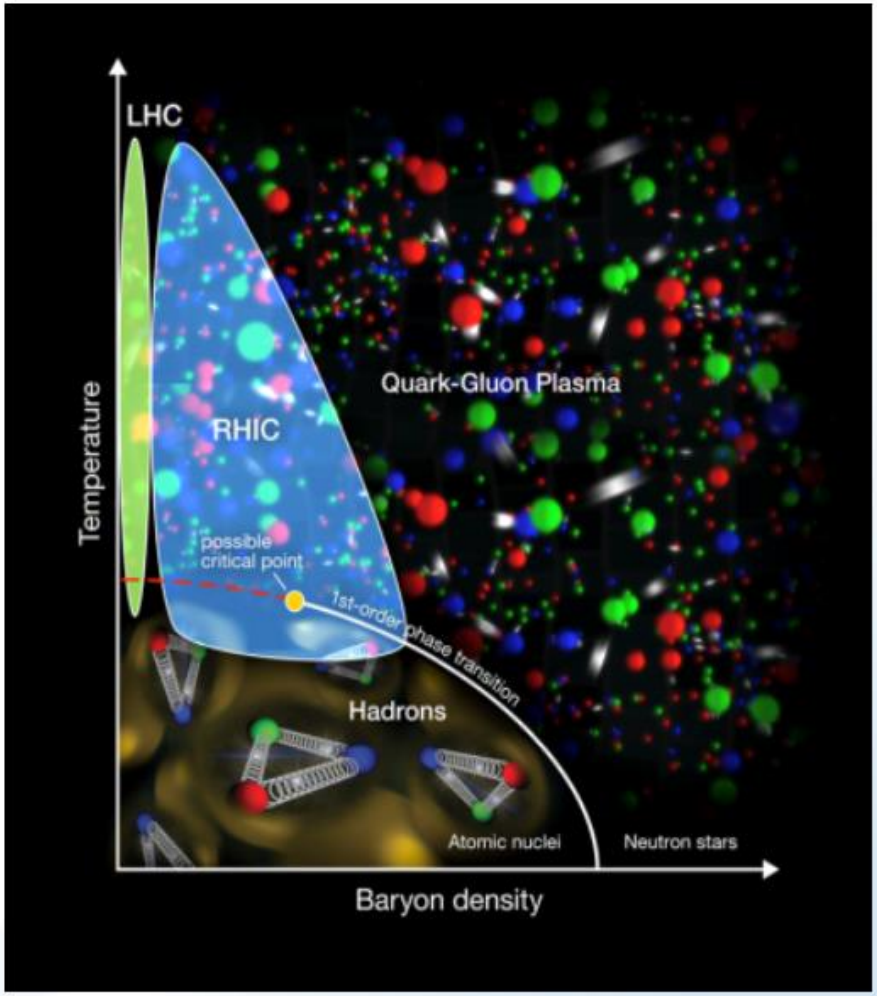
U.S. DEPARTMENT OF ENERGY | Office of Science



- Physical Assets
  - 5322 acres
  - 310 SC buildings
- Human Capital
  - 2882 FTEs (2987 heads)
  - Direct/indirect: 0.59/0.41
  - 480 undergrad/grad students (paid by Lab)
  - 4134 facility users
  - 1377 visiting scientists

RHIC is the Perfect Facility to Explore the Phases of Nuclear Matter

- QCD matter turns from a nucleon superfluid into a nucleon/hadron gas at approximately 100 billion degrees
- When heated to 2 trillion degrees at RHIC, nuclear matter suddenly turns into a liquid again
  - The most perfect liquid ever observed
- **Only RHIC has the energy range to observe where the transition occurs**



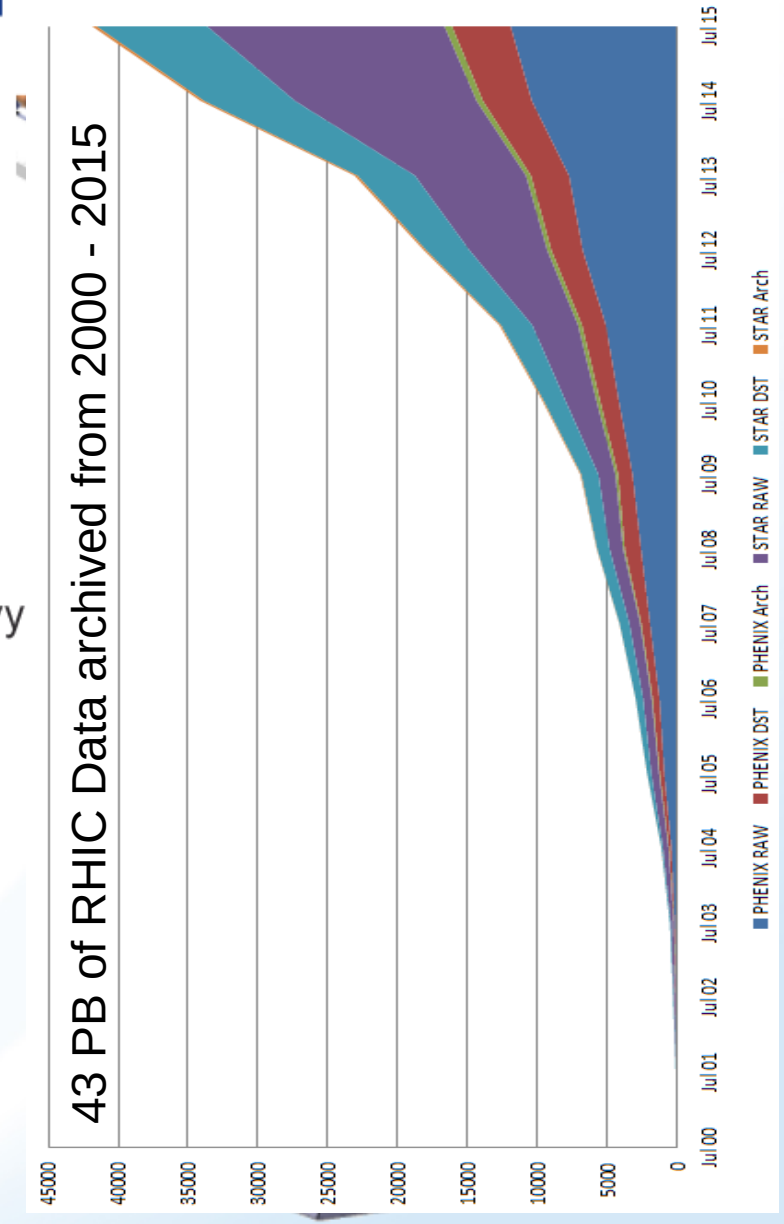
Completing the RHIC Mission

Status:

- RHIC-II configuration is now complete
  - 3D stochastic cooling
  - Vertex detectors in STAR (HFT) and PHENIX
- RHIC Run 15 – first proton-gold collisions

Plan: Complete the RHIC Mission in 3 campaigns:

- **2014 - 17:** Understand the properties of the strongly correlated quark gluon plasma using heavy quark spectroscopy
- 2018: Install low energy e-cooling
- **2019/20:** High precision scan of the QCD phase diagram
- 2020: Install sPHENIX upgrade
- **2021/22:** What makes the QGP a perfect fluid: Elucidate the fluid correlations using jets
- 2023/2024: RHIC shutdown and transition to eRHIC (assuming NSAC/DOE/Congressional alignment)



RHIC remains a unique discovery facility:
~3,000 citations/year, ~30 PhDs per year

Jerome will provide a lot more information on Computing in Nuclear Physics at BNL

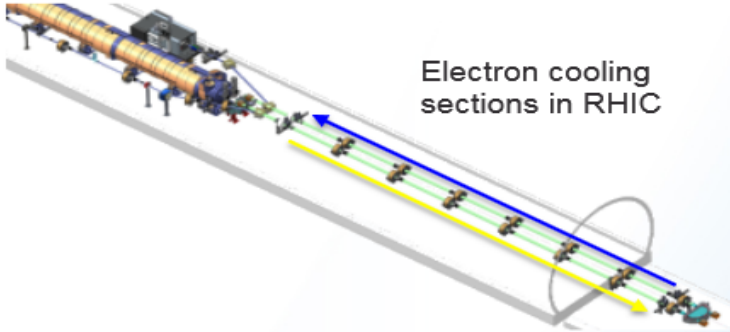
Completing the RHIC Mission

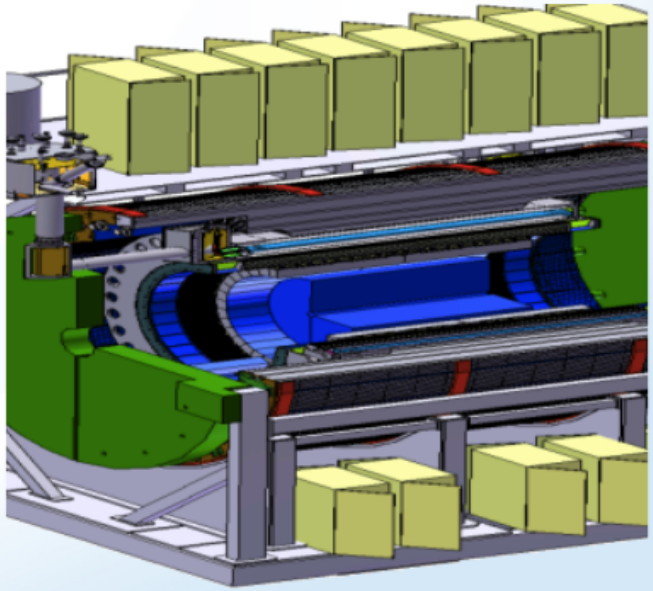
Status:

- RHIC-II configuration is now complete
  - 3D stochastic cooling
  - Vertex detectors in STAR (HFT) and PHENIX
- RHIC Run 15 – first proton-gold collisions

Plan: Complete the RHIC Mission in 3 campaigns:

- **2014 - 17:** Understand the properties of the strongly correlated quark gluon plasma using heavy quark spectroscopy
- 2018: Install low energy e-cooling
- **2019/20:** High precision scan of the QCD phase diagram
- 2020: Install sPHENIX upgrade
- **2021/22:** What makes the QGP a perfect fluid: Elucidate the fluid correlations using jets
- 2023/2024: RHIC shutdown and transition to eRHIC (assuming NSAC/DOE/Congressional alignment)

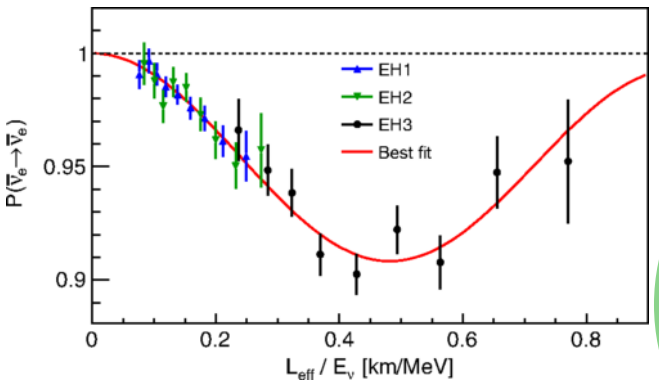




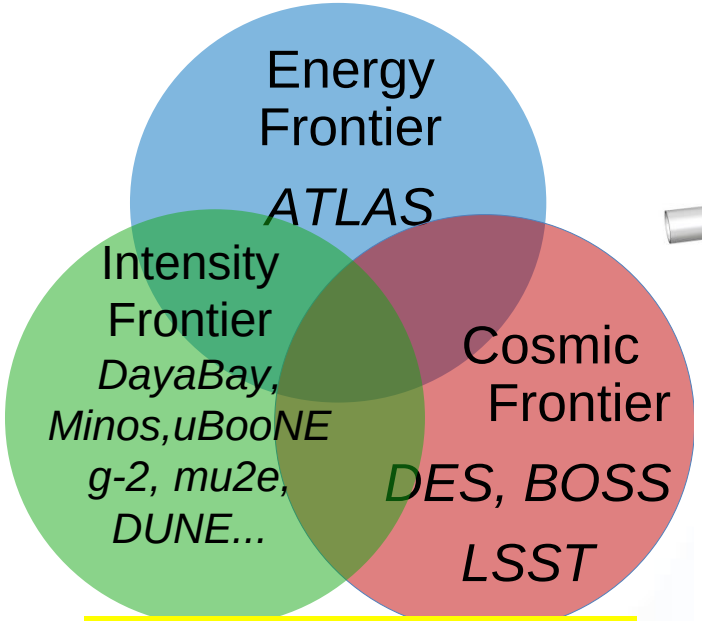
RHIC remains a unique discovery facility:
~3,000 citations/year, ~30 PhDs per year

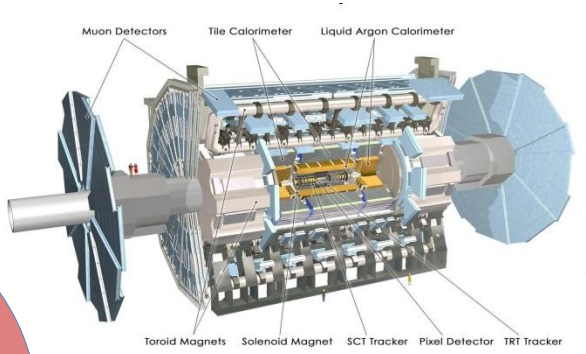
High Energy Physics at BNL

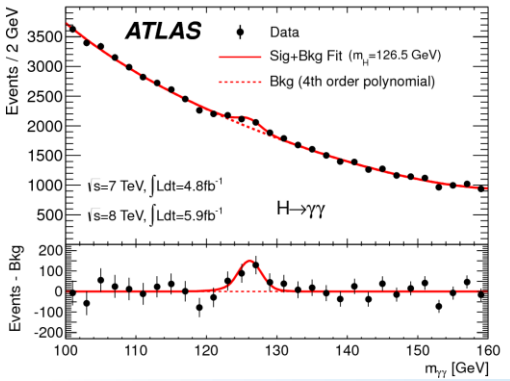
Daya Bay - $Sin^2 2\theta_{13}$



PRL 112, 061801 (2014)



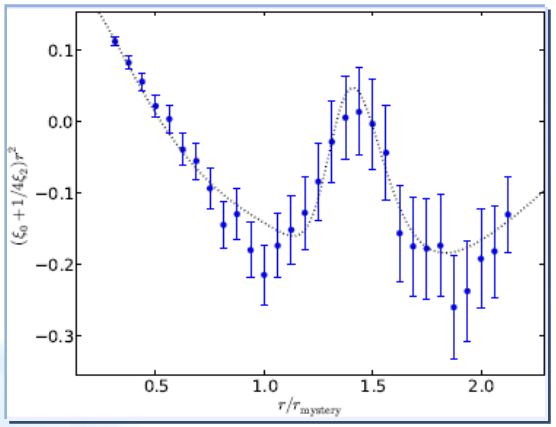




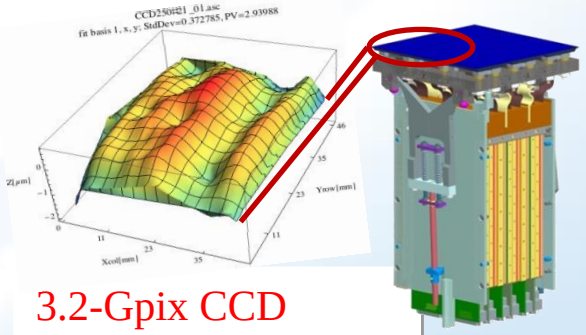
M(γ,γ) GeV

Phys. Lett. B 716 (2012)

HEP Theory Group Contributes to all Frontiers



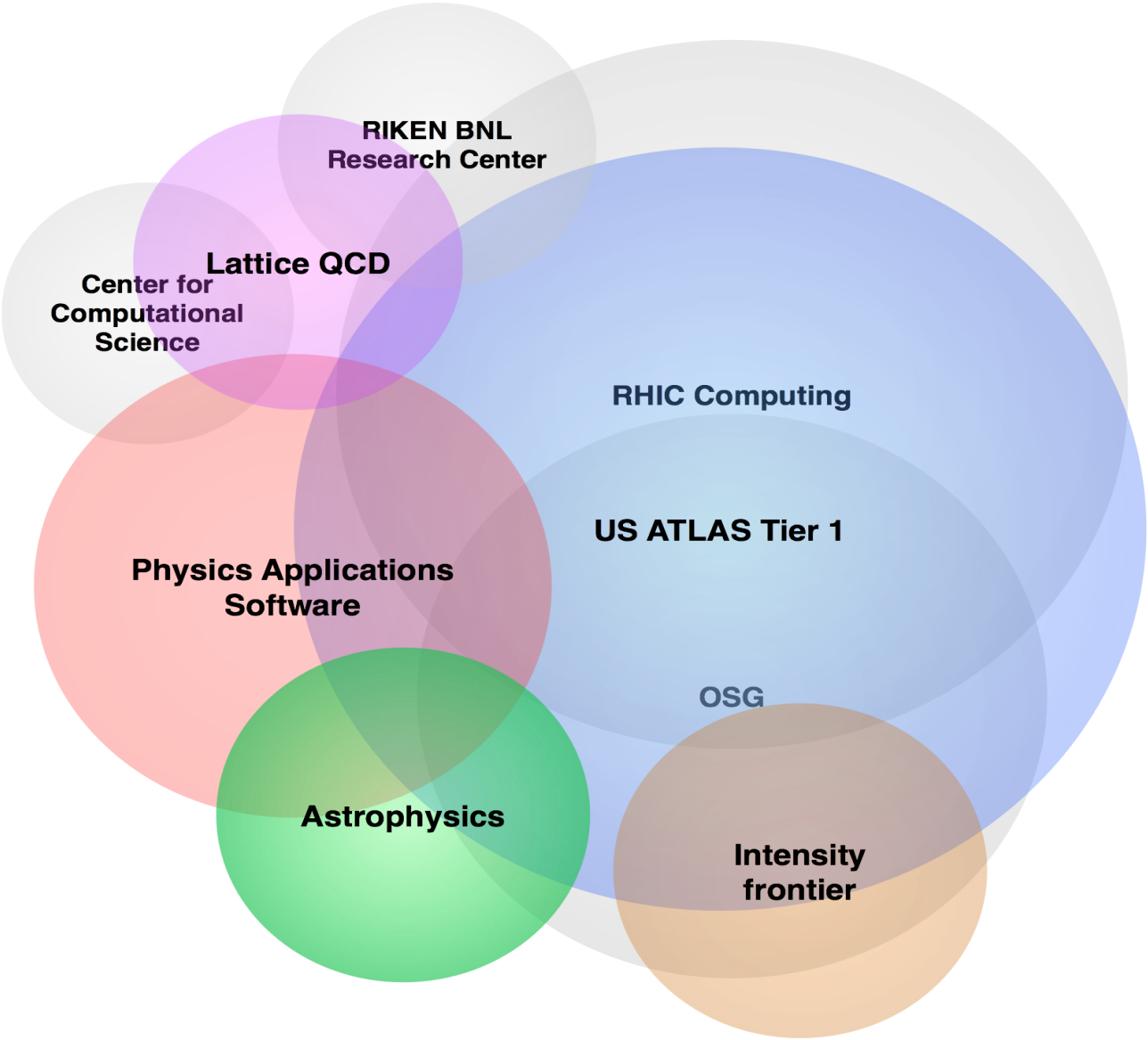
1st observation of high-z BAO peak via Lyman-α forest absorption



3.2-Gpix CCD camera for LSST

HEP Computing at BNL

Colors: HEP
Grey: non-HEP



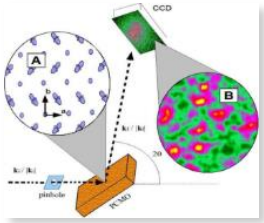
HPC in Experimental High Energy Physics

- Taking ATLAS as an example, in a year, ATLAS has grown from ε HPC use to 6% of ATLAS Grid resources.
- This could grow to 50% or more
  - One challenge is adapting the framework
  - Another is the data volume and mobility (it has to get back to Grid sites)
- In addition, with the new LHC data ATLAS' needs will grow
  - A factor of ~3 in 2-3 years
  - Another factor of ~3 in another 2-3 years
  - Ultimately a factor of ~100.
- Nevertheless, this is growing more slowly than the Leadership Class Facility (LCF) capacity – **HEP can fit**
- It is likely that HEP's needs will grow beyond this if the computing situation permits
- Other experiments are starting to show interest as well: perhaps another factor of three on top of this.

➢ ***Could OSG have a role in this?***

NSLS-II – 10,000 times brighter than NSLS

Coherent Soft X-ray Scattering (CSX-1)



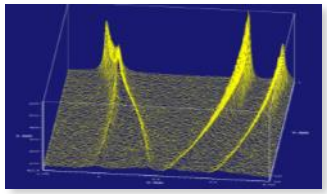
World-leading coherent flux XPCS, CDI

Imaging & dynamics in strongly correlated and magnetic materials

Fast Switching Polarization (CSX-2)

Resonant magnetic scattering, spectroscopy, XMCD

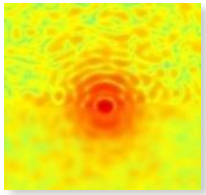
X-ray Powder Diffraction (XPD-1)



High energy Powder diffraction, scattering

Time-resolved in-situ in-operando extreme conditions
Understanding complex nanostructured materials

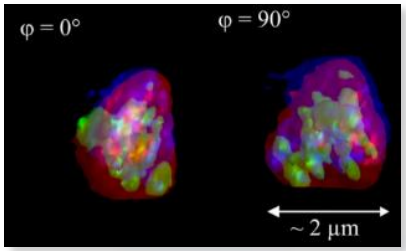
Coherent Hard X-ray Scattering (CHX)



Non-equilibrium and heterogeneous dynamics in soft matter, at buried interfaces, biomaterials, glasses, driven systems

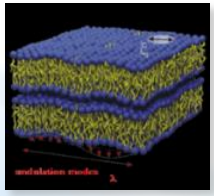
100x greater time resolution in XPCS studies of dynamics

Sub-um Resolution X-ray Spectroscopy (SRX)



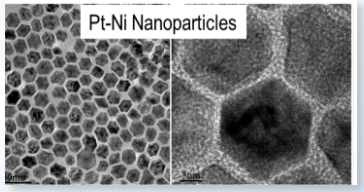
World-leading spectroscopy in sub-100 nm spot

Inelastic X-ray Scattering (IXS)



THz dynamics in liquid, glassy, and crystalline materials with nanoscale inhomogeneities

Hard X-ray Nanoprobe (HXN)



Nanoscale imaging with fluorescence and diffraction

100m long beamline
~10 nm baseline

NSLS-II – 10,000 times brighter than NSLS

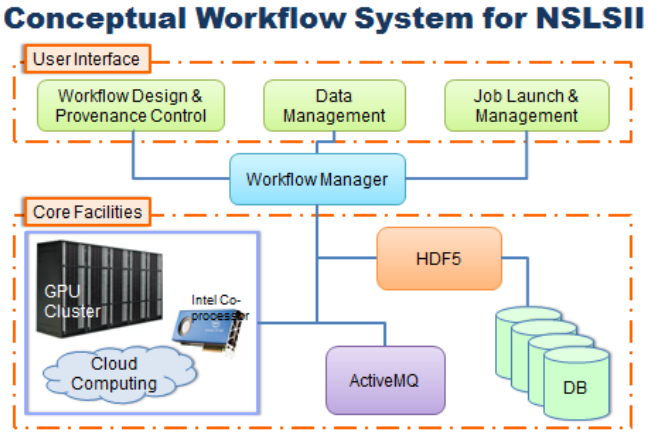
- Data rates at NSLS-II will eventually be > **90 TB/day** driven by use of high speed 2D detectors and the brightness of NSLS-II allowing for faster data collection.
- At the same time the synchrotron user community has transitioned from “expert” users to scientists who use the synchrotron as a “tool”.
- Currently most users do not:
- … have the hardware at home to analyze complex data.
- … have the analysis software to properly analyze the data.
- Full utilization of NSLS-II (28 beamlines, ~2500 users in 2018) will require the facility to store and manage the data and provide analysis and visualization tools for local and remote users. – ***The community needs the computing resources and a fabric of services that enable the knowledge discovery process***
- ***This is not just a local facility issue ! A role for OSG here?***
- Security for access to these data from 1000s remote locations with challenged network. (Cable / DSL / University)
- Wide variation in data types out of the lab (e.g. 1000s small files and some large datasets) to a very diverse set of institutes.

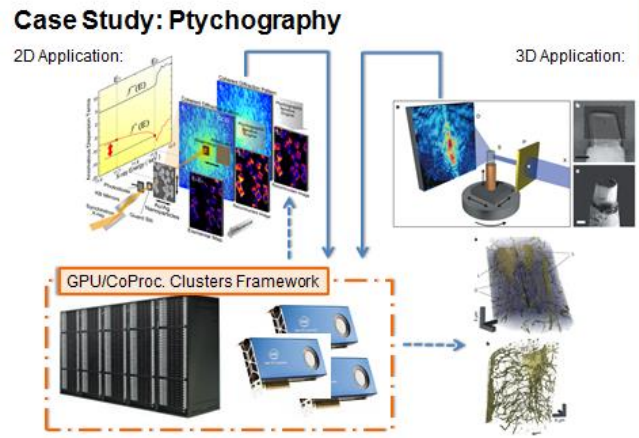
NSLS-II – 10,000 times brighter than NSLS

- The challenge is this community does not just need a repository for data and processing. **Data Science** is changing how we approach experiments and use facilities.

- Could OSG become a focus for experimenters at Lightsource Facilities to work with data scientists in a collaborative environment?

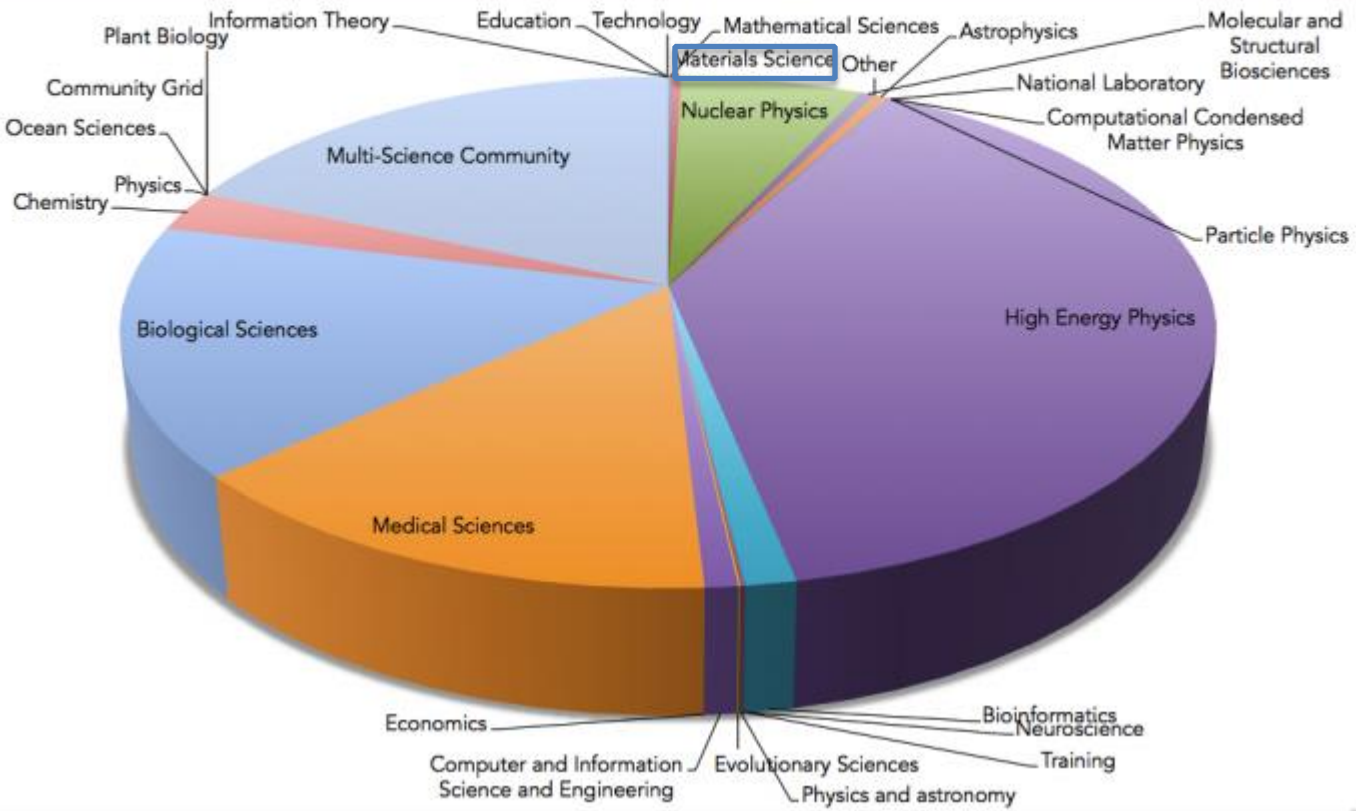
- With the shift in the synchrotron user community, training and collaborating with user groups is extremely important. OSG could become a nexus for such collaborations.





Material Science is small on OSG today

Science Throughput



Rob Gardner

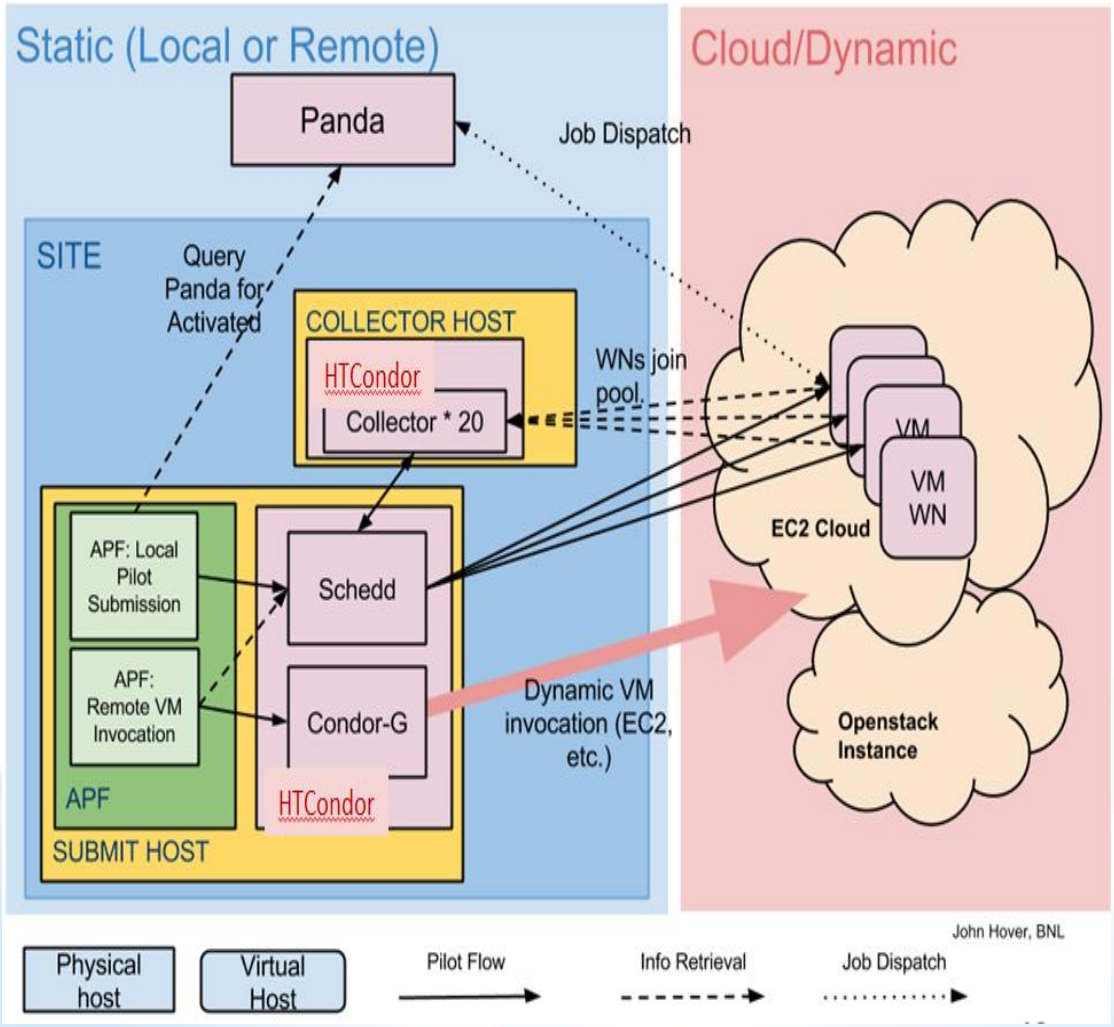
flexible and nimble

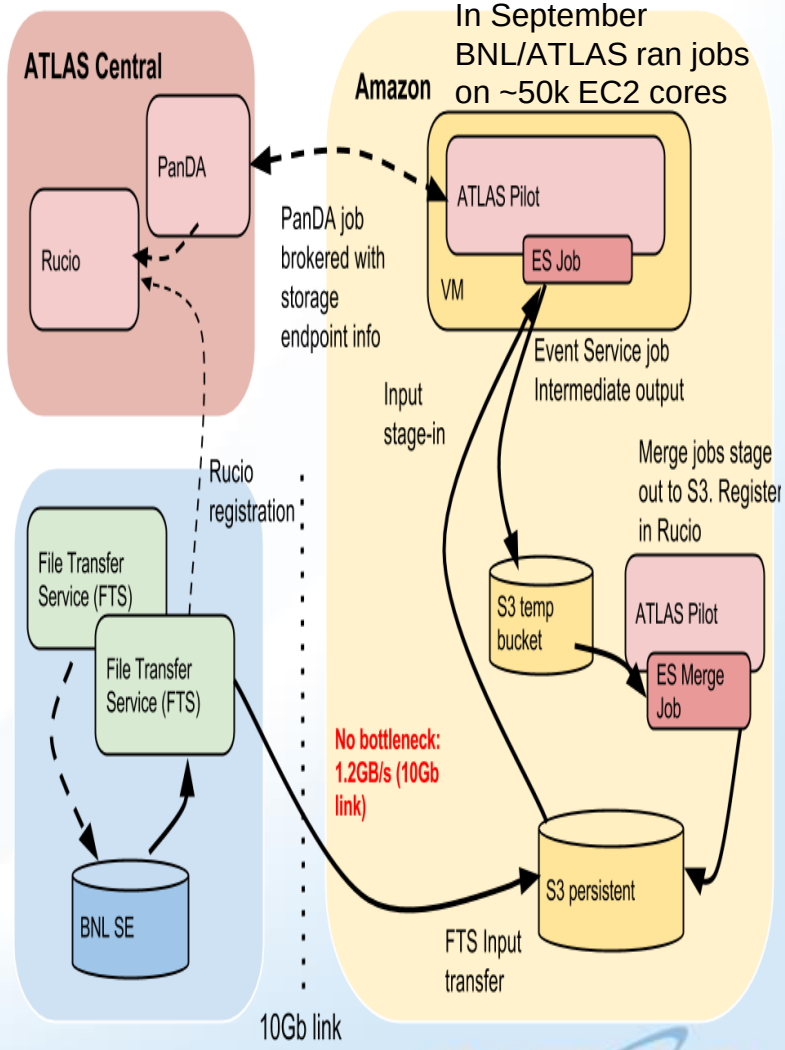
- Grid
- Cloud
- HPC

Provisioning and Using Cloud Resources

Programmatically instantiates
Compute resources in the Cloud on demand

Fine-grained data processing making
"Volatile"/AWS SPOT Resources valuable





Strengthen Leadership in Data Driven Discovery

- Lab Program Development has enabled creation of a Center for Data Driven Discovery (C3D)
- C3D has four core missions:
  - Provide a focal point for leading data science research
  - Educate the next generations of expert data scientists
  - Translate research advances into tools and expertise that lead to measurable scientific progress and improved industrial competitiveness
  - Demonstrate its capabilities through joint scientific projects (e.g. HEX NIH Beamline at NSLS-II and EIC (Electron Ion Collider)) and industrial collaborations (starting with e.g. GE, Pfizer that worked already with the NSLS)

C3D Approach

- The ability to make sense of data at large volumes and faster speeds is foundational to science, energy, health, national security and industry.
- C3D will develop and provide:
  - The expertise to analyze and unravel societies' most challenging data science needs
  - The software tools to enable C3D and others to address these challenges
  - The physical computational environment to solve these challenges in a timely and effective manner
    - Local and offsite resources
- Potential for Collaboration with OSG