Consolidating Scientific Computing Services at BNL

Tony Wong
Brookhaven National Laboratory
HEPIX @ Zeuthen (Spring 2016)

B.C.

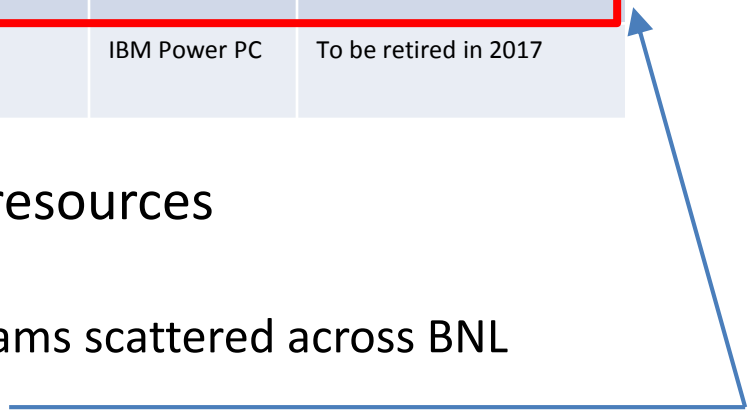
- BNL is a multi-disciplinary laboratory
- Legacy of field-specific funded research programs (HEP, NP, life sciences, energy sciences, etc)
- De-centralized computing services (except for WAN) is a natural result of this organizational structure
  - Dedicated computing services and staff is a plus for some programs, but…
  - Resource usage has built-in inefficiency
  - Duplication of services
  - Inability to leverage staff expertise due to program-specific constraints
- BNL is one of two (out of 10) DOE laboratories with de-centralized scientific computing services
- Recognized as a disadvantage when seeking to expand support to new initiatives such as NSLS-II and CFN at BNL

B.C.

- RACF was created in this de-centralized environment to support RHIC/ATLAS
- Expanded in ~2008 to support other activities in Physics Department
- RACF current portfolio of responsibilities
  - Nuclear and High Energy Physics
    - RHIC/eRHIC
    - ATLAS
  - Physics Intensity Frontier
    - DUNE
    - Daya Bay
  - Cosmic Frontier
    - LSST
    - Legacy projects

B.C.

Cluster Name	# of cores	Community	Type	Notes
RACF	~51,300	Nuclear & High Energy Physics	Intel Xeon	
Brookhaven Linux cluster	180	General	Intel Xeon	To be retired in 2016
CFN cluster	2,422	Center for Functional Nanomaterials	Intel Xeon	50% to be retired in 2016
HPC1 cluster	52	Computational Science Center	Intel Xeon Phi	GPU with IB inter-connect
Life S cluster	32	Biological, Environmental and Climate Sciences	Intel Xeon	
Biology cluster	10	Biology	Intel Xeon	To be retired in 2016
CMP cluster	80	Condensed Matter Physics and Material Sciences	Intel Xeon	
Blue Gene Q	60,928	Theory/LQCD	IBM Power PC	To be retired in 2017



- RACF has largest pool of HTC-centric resources
- Non-RACF managed equipment
  - Small clusters dedicated to other programs scattered across BNL
  - All maintained by a single IT person
  - Manpower intensive and not scalable

A.D.

- Computational Science Initiative (CSI) aims to change the organizational structure (see http://www.bnl.gov/compsci/)
- CSI plans to:
  - optimize resource utilization
  - re-organize staff responsibilities
  - support new BNL initiatives
  - leverage RACF expertise and experience
- RACF takes operational responsibility for scientific computing at BNL
- Acquire HPC-like cluster with enough computing power to serve the initial needs of the new initiatives
- Build new data center to accommodate expected increase in resources

A.D.

- RACF expanded portfolio of responsibilities
  - Nuclear and High Energy Physics
    - RHIC/eRHIC
    - ATLAS
    - LQCD
  - Physics Intensity Frontier
    - DUNE
    - Daya Bay
  - Cosmic Frontier
    - LSST
    - Legacy projects
  - Photon Physics (mainly NSLS-II)
  - Center for Functional Nanomaterials (CFN)
  - Future BNL activities?

A.D.

Cluster Name	# of cores	Community	Type	Notes
RACF	~51,300	Nuclear & High Energy Physics	Intel Xeon	
CFN cluster	1500	Center for Functional Nanomaterials	Intel Xeon	To be retired in 2017(?)
Institutional cluster (phase 1)	~4,000	General	Intel Xeon	GPU with IB inter-connect
Blue Gene Q	60,928	Theory/LQCD	IBM Power PC	To be retired in 2017
LQCD cluster	TBD	LQCD	TBD	Acquisition expected in late 2017 or early 2018

- Institutional cluster (phase 1) to be deployed in July 2016
  - Replaces previously small, program-specific clusters scattered at BNL
  - Intended for HPC-like applications
- Phase 2 likely to double the size of the institutional cluster
  - Purchase timeline not decided yet
  - Xeon Phi Knights Landing a possibility
- LQCD cluster will replace Blue Gene Q
  - Xeon Phi Knights Landing a possibility
  - Some degree of integration and coordination with institutional cluster likely

Institutional Cluster

- Hardware configuration
  - 108 servers with dual Broadwell-based cpu's
  - Two Nvidia K80 GPU's per server
  - 128 GB ECC RAM
  - 2 TB SAS drives for data and swap space
  - 240 GB SSD for OS/swap
  - Non-blocking Infiniband EDR fabric
  - Estimated 20 kW/rack
  - 1 PB of GPFS storage (up to 24 GB/s bandwidth via EDR)
- Integration into RACF
  - Network
  - User account management
  - Ticket system
  - Environmental and activity monitoring
  - Other

Enhanced Responsibilities

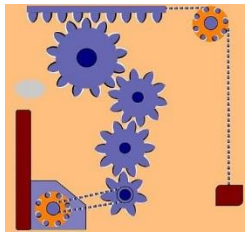
- Existing RACF operational responsibilities remains unchanged
- Different support model for institutional cluster
  - Separate MOU
  - Usage allocation committee (TBD)
  - Evaluate Slurm as workload manager (HPC cluster only)
  - Software management and provisioning (Openstack/Docker?)
- Resource sharing
  - Can HTCondor and Slurm co-exist on institutional cluster?
  - Possible back-fill by RHIC, USATLAS, OSG and others
  - Study to be conducted this summer
- Additional staff
  - One already hired (starting April 25th)
  - Two additional openings (one already posted, another soon)

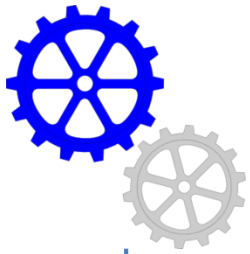
HTC and HPC

RACF

HTC

HPC





HTCondor

Slurm

Resource Sharing?

RHIC/ATLAS

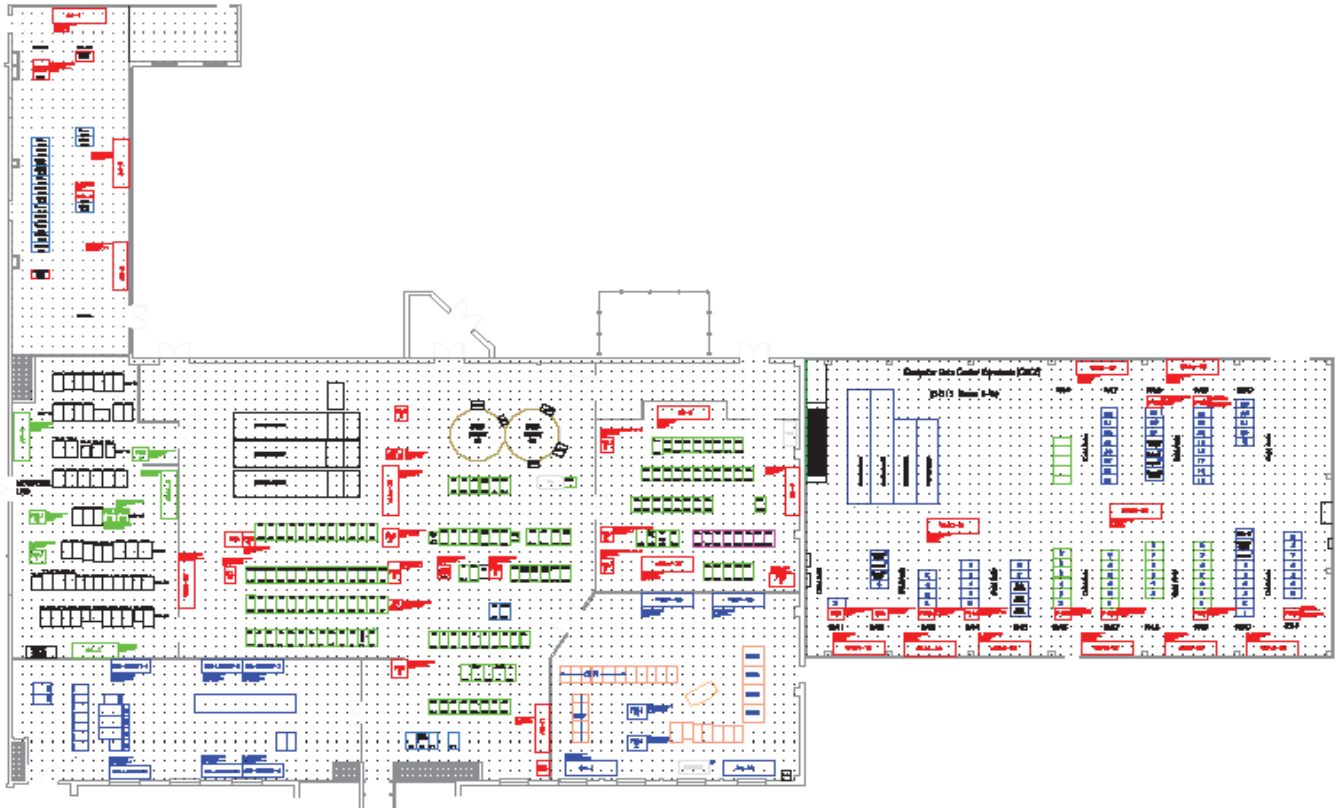
Institutional Cluster

New Data Center @ BNL

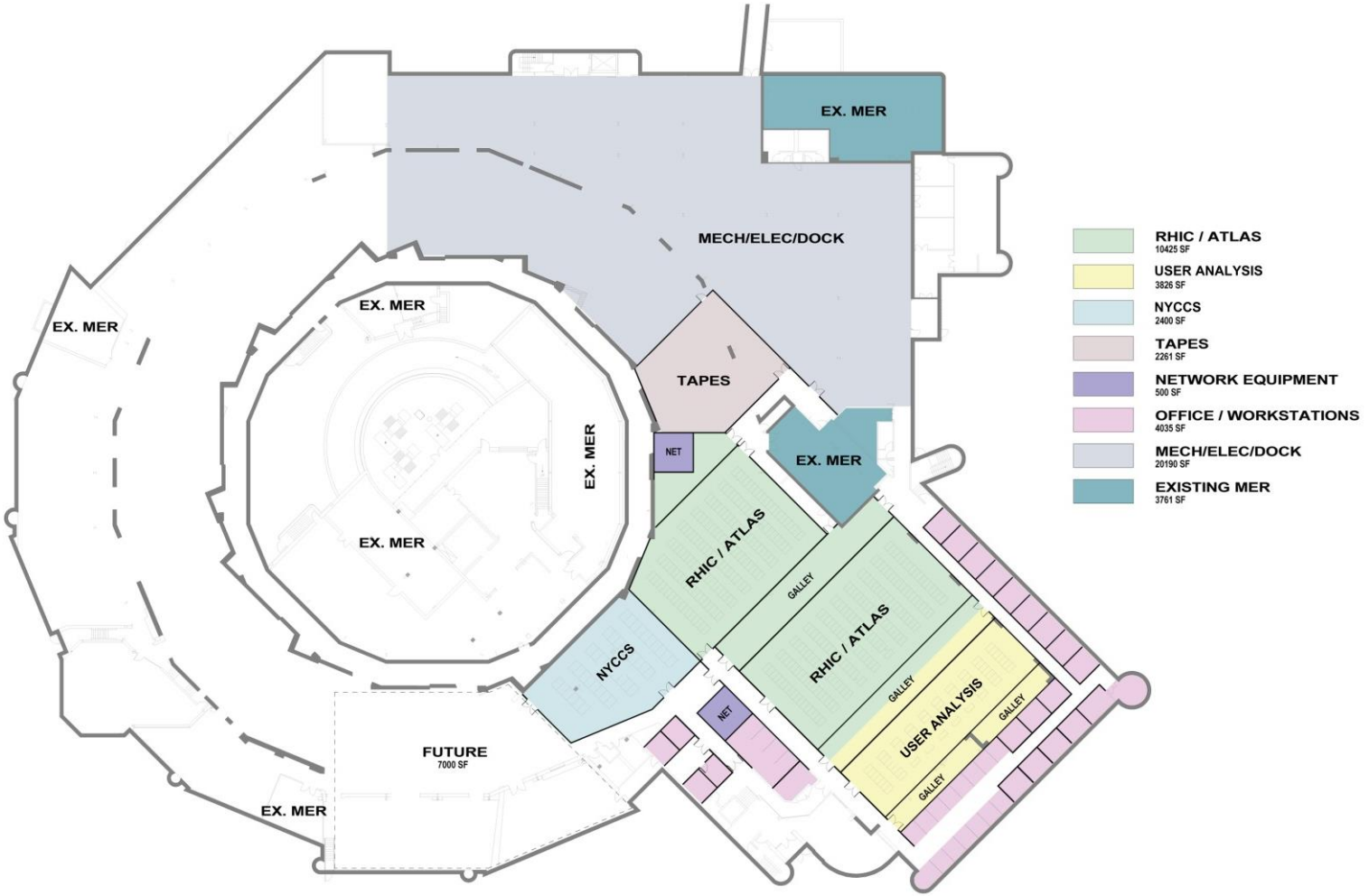
- DOE-funded Core Facility Revitalization (CFR) program
  - Identify and renovate key installations at BNL
  - Venue for HEPIX Fall 2015 meeting was renovated with CFR funds
- New Data Center identified as a key complement to CSI
  - CD-0 (Mission need) review successful – granted in September 2015
  - Preparing for CD-1 (Preliminary baseline range) review now – expect decision in Fall 2016
  - Estimate initial (design) funding in 2017 and construction in 2018-2020
  - Migration to new data center begins in late 2020 – near end of LHC LS2 and before Run 3 starts

Current Data Center @ BNL

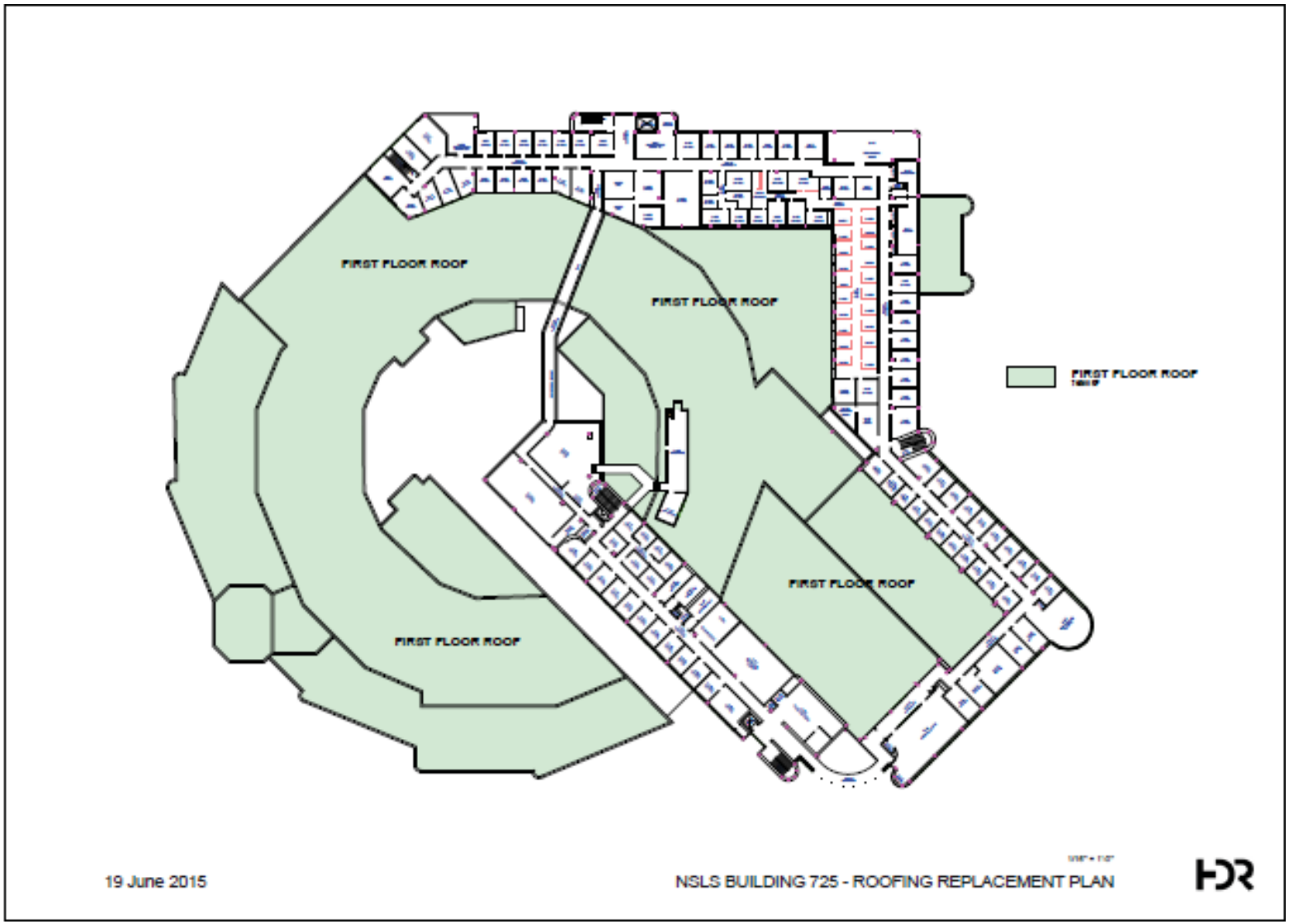
- Limited power, cooling and space in existing data center
- Piecemeal addition of two spaces in 2008 and 2009 to cope with new requirements
- Not sufficient to accommodate the needs of new initiatives



New Data Center (option 1)



Proposed New RACF Office Space



Key Facts of Proposed New Data Center

- 10 MW power feed (6 MW with UPS back-up)
- 32,000 ft^2 new facility (~2.5x larger than current data center) on 1^{st} floor
- 5,000 ft^2 for office space (2^{nd} floor)
- To achieve PUE of 1.2 - 1.4 (mandatory DOE requirement), consider:
  - Assisted external air cooling (~75-80% of the year)
  - Dedicated chillers as back-up for assisted external air cooling
  - Hot-aisle containment
  - Operate at higher intake temperature
- Designed to insure 99% uptime (3.5 days of downtime/year)
- Multi-user facility
  - USATLAS
  - RHIC/eRHIC
  - CFN
  - NSLS-II

Summary

- Higher visibility in scientific computing has been identified as a priority by BNL management
- The RACF is undergoing a transformational change
  - New responsibilities
  - Additional staff positions – contact me if need be
  - Management changes
- Yet another new data center at BNL (last one was in 2009) by 2020