Construction of a New Data Center at Brookhaven National Laboratory

CHEP 2019 04-09 Nov 2019 Adelaide, AU

Presented by Shigeki Misawa (BNL)
Co-authors : Imran Latif (BNL), Alexandr Zaytsev (BNL)

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Brookhaven National Laboratory (BNL)

U.S. Department of Energy National Laboratory

Home of

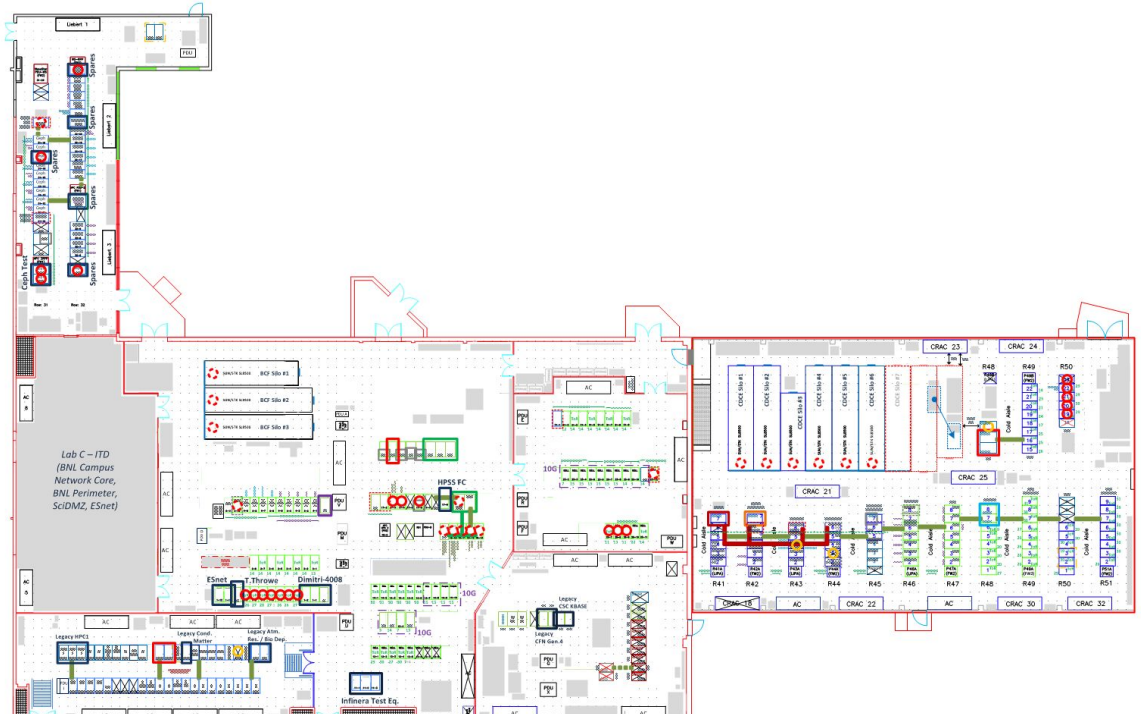
- Relativistic Heavy Ion Collider (RHIC)
- National Synchrotron Light Source II (NSLS-II)

Scientific Data and Computing Center (SDCC)

- Center for Scientific Compute at BNL
  - RHIC “Tier 0” facility
  - US ATLAS Tier 1 facility
  - Belle II data center outside of Japan
- Primary Resources
  - High Throughput Compute farm
  - High Performance Computing clusters
  - HPSS Mass Storage System

Current SDCC Data Center

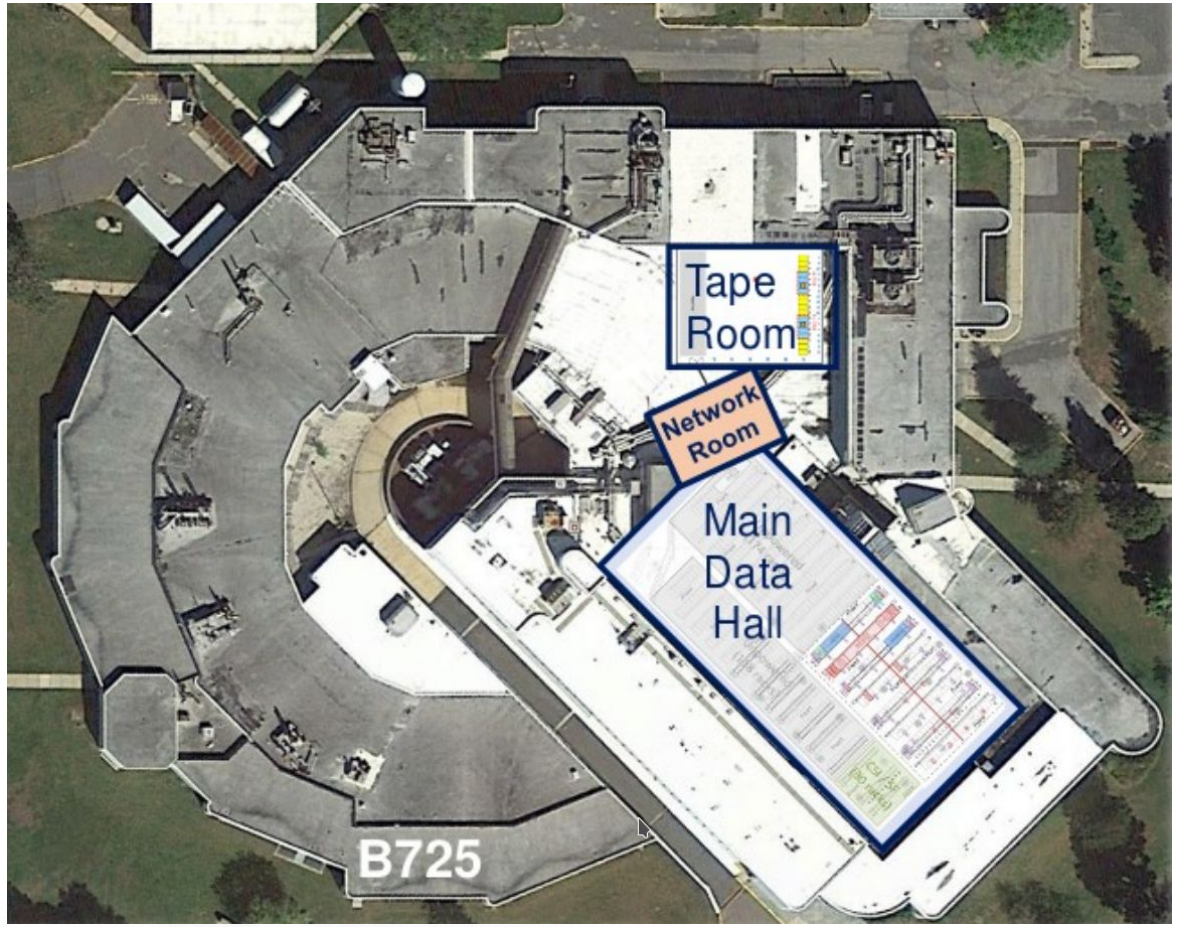
- Built in 1960's with additions in 2009
- 1940 m^2 (~20,000 ft^2) area
  - 30 cm/75 cm raised floor
  - 750 - 1500 kg/m^2 load rating
- "Tier I" non-redundant data center
  - (using the "Tier" classification defined by the Uptime Institute)
- CRAH based cooling
  - Under floor supply plenum
  - Dependent on campus Chiller
  - Not on generator power
  - 10 KW/rack max cooling (non HPC)
  - No hot/cold aisle containment
  - Non-uniform cooling capability
  - Unable to meet mandated PUE < 1.4



- > 4 MW electrical power
  - 3 MW UPS power (battery/flywheels)
  - 2.3 MW diesel generator
  - Irregular physical distribution
  - 30A single/three phase circuits (non HPC)

New Data Center

- Repurposed NSLS Light Source building
- “Tier III” Class data center*
  - Redundant infrastructure
  - Concurrently maintainable
  - Completely self sufficient in emergencies
- “Ultimate” capacity (not deployed)
  - 9.6MW electrical power
  - Six 18 frame tape libraries
  - ~480 equipment racks
- Outfitting ~50% of ultimate capacity
- PUE < 1.4, with 1.2 as design target



* using the “Tier” classification defined by the Uptime Institute

Modular Power / Cooling

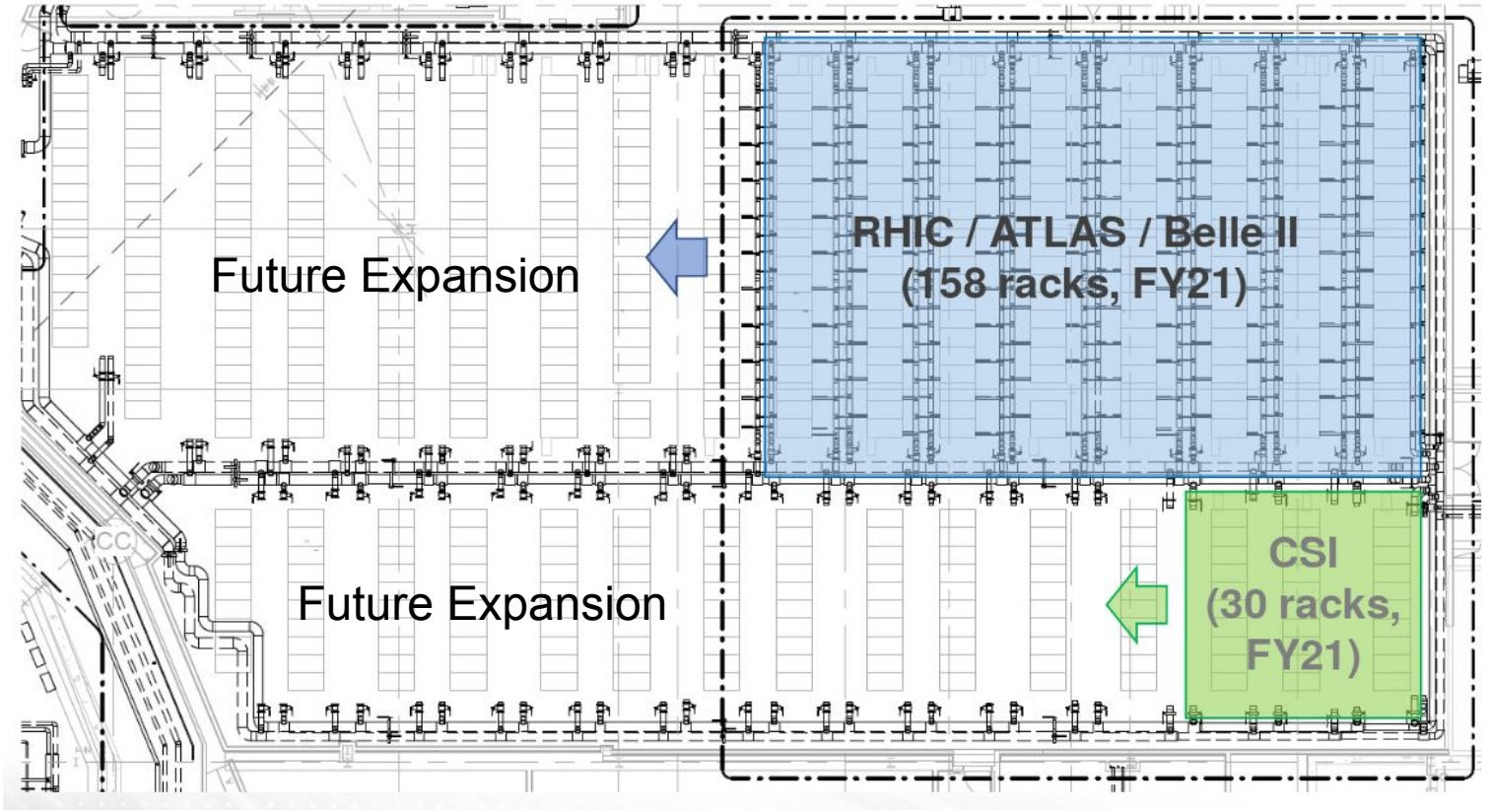
- Ultimate 9.6 MW buildout
  - 8 power+cooling “modules”
  - 1 maintenance bypass power/cooling “module”
- 1.2 MW IT load per “module”
  - 1.75 MW Diesel generator
    - 24 hour fuel tank
    - Externally sited
  - 1.2 MW VRLA based UPS system
    - 5 minute run time @ max load
    - Sited indoors
  - 445 ton (1.57 MW) chiller
    - Chiller on generator
    - Critical water pumps on UPS

- Current “base” buildout (3.6 MW)
  - Three 1.2 MW modules (almost)
    - except 3rd generator is an option
  - Fourth 1.2 MW module is an option
  - Maintenance Bypass module
- Maintenance bypass module
  - Power bypass - 1.2 MW utility power
    - UPS is an option
  - Cooling bypass - One chiller equivalent using BNL campus chilled water.
    - Fourth chiller is an option
  - Diesel generator - In design, but no option to purchase.

New Data Center Layout

Tape Room

- Main data hall for compute
- Dedicated tape room
- Separate network room (not shown)



Main Data Hall

- Blue 10 KW/20 KW per rack
- Green 30 KW per rack

Main Data Hall

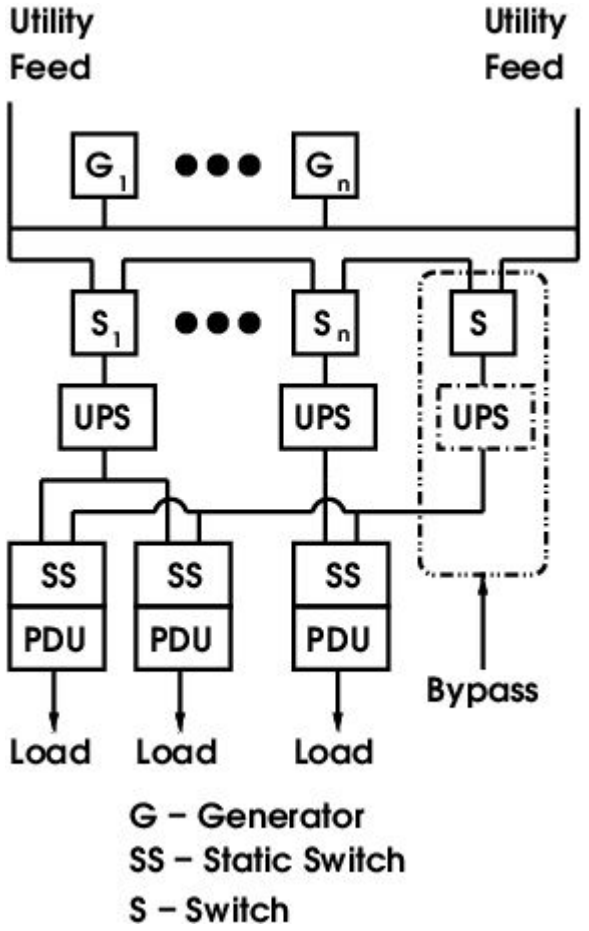
- Supports standard 42U rack
- ~1,115 m^2 (12,000 ft^2) total area
  - 75 cm raised floor
  - 2440 kg/m^2 load rating
  - ~50% of area “energized”
- Partitioned floor plan
  - ⅔ area for HTC
  - ⅓ area for HPC
- Under floor chilled water pipes
  - Four water supply/return loops
    - Two loops for all HTC areas
    - Two loops for all HPC areas
  - 60℉ (15.5℃) supply temperature
- HPC partition
  - 15 rows, 10 racks per row
    - 3 rows “energized”
  - Electrical distribution unspecified
  - Water cooling solution unspecified
- HTC partition
  - 16 rows, 20 racks per row
    - 8 rows “energized”
    - Alternating “compute”/”storage” rows
  - Cooling via active rear door heat exchange (RDHx)
  - Power via overhead busway
    - Three phase 50A tap boxes
    - Two 50A rack PDUs per rack
    - 14.4 KW usable per rack PDU

Power Distribution

- HTC “storage” rows
  - “A/B” power @ 10 KW per rack
  - Two 200 KW busway per row
- HTC “compute” rows
  - “A” side only power @ 20 KW per rack
  - One 200 KW busway per 10 racks
- Each HTC UPS system (2 total)
  - “A” side - 3 x 400 KW PDU
    - 2 x 200 KW “storage” busways
    - 4 x 200 KW “compute” busways
  - “B” side - 1 x 400 KW PDU
    - 2 x 200 KW “storage” busways
  - PDUs connected to UPS and bypass through static switch
- HPC rows (energized)
  - “A” side only power
  - 300 KW per 10 racks
- HPC UPS system
  - “A” side - 4 x 300 KW PDU
    - 3x300 KW for HPC
    - 300 KW to network/tape rooms
  - “B” Side - 300 KW PDU
    - 300 KW to network/tape rooms
  - PDUs connected to UPS and bypass system through static switch
  - 300 KW UPS on bypass feed for network/tape room PDU if no UPS on bypass system

Power Distribution

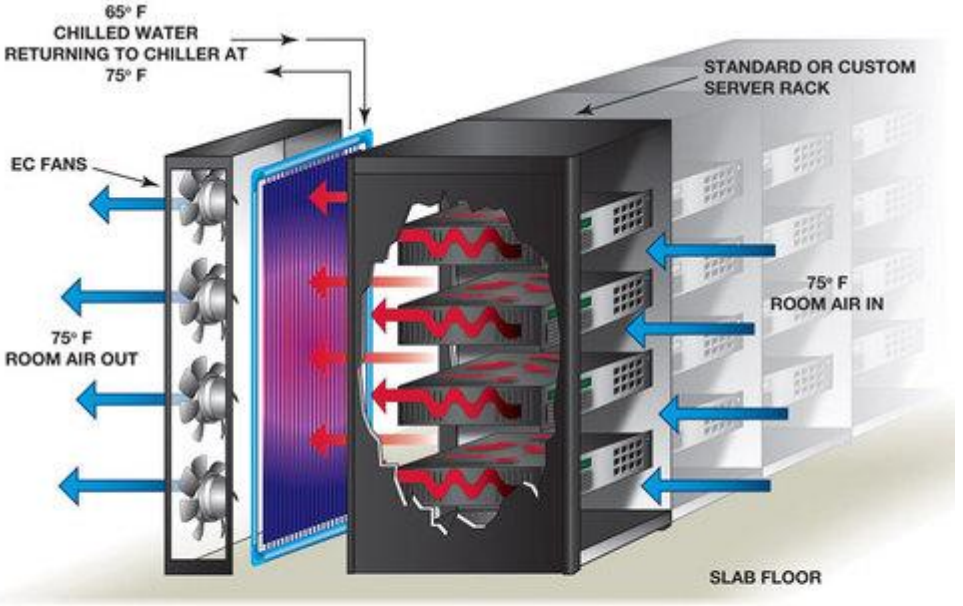
- All generators connect to ALL UPS systems
- Two paths from generator to UPS
- Bypass system connected to generators
- Two power connections to PDU
  - To one primary UPS
  - To bypass system
- Concurrent maintainability on all equipment except HTC compute node PDU and busways



Cooling System

- Variable speed, magnetic bearing chillers
- Cooling towers with variable speed fans
- Two water side economizers
- Cooling solutions
  - CRACs in tape, mechanical, and electrical rooms
  - CRACs and active RDHx in network room
  - Active RDHX in main data hall
- Water pipe size to racks

  (assumes active RDHx)
  - 15 KW per “storage” rack
  - 30 KW per “compute” rack
  - 300KW per HPC row (10 racks)



- Racks in row alternate back/front branch pipe for water source
  - 50% racks lose water in case of branch pipe shutdown

Networking/Monitoring

- Standalone network room
  - < 100 m to all racks
  - All network equipment except ToRs and low latency HPC switches.
- Overhead cable tray
  - Separate copper/fiber trays
  - Runs parallel to rows
- Overhead network patch panels
- Three separate networks
  - Production network
    - Direct optical or ToR+optical uplink
  - Local, row based IPMI - ToR based
  - BAS/DCIM monitoring (in data hall)
    - Direct 1Gbase-T to network room

- Infrastructure Monitoring
  - BNL campus building automation system (BAS) monitors
    - mechanical systems
    - electrical systems
    - plumbing systems
  - DCIM system monitors
    - Rack PDUs
    - Rear door heat exchanges
  - DCIM prototyping in progress in the old data center.

Data Center Migration

- LHC Run 3 starts CY2021
  - No disruptions to ATLAS Tier 1 operations
- New data center occupancy
  - ATLAS areas ready before CY2021
  - Other areas become ready for occupancy throughout CY2021
- Phased migration plan
  - Critical ATLAS systems “move” prior to CY2021
  - Other systems are “transitioned” to the new data center over time

- Migration Plan
  - Facility network extend to new data center in late CY2020
  - Starting CY2021, all new equipment installed in new data center
    - Includes new tape libraries in time for LHC Run 3
  - Only newest compute nodes in old data center physically relocated to new data center
  - Remaining equipment in old data center will retire in place.
  - Old data center vacated by CY2023-24, except nine legacy tape libraries and associated servers.

Current Status

- Contract for construction has been awarded.
  - Contracted review caused minor delay
  - No impact to data center availability for ATLAS

- General contractor is in post award, pre-construction phase with subcontractors (Buyout phase in construction lingo)

- Purchase of equipment with long lead times are “in flight”

- Pre-construction, non contract work completed
  - Abatement projects (removal of hazardous materials)
  - Pre demolition surveys

Additional Information

- CHEP 2016 presentation on new data center rationale
  - https://iopscience.iop.org/article/10.1088/1742-6596/898/8/082009/pdf

- HEPiX Spring 2019 presentation on data center migration plans
  - https://indico.cern.ch/event/765497/contributions/3351452/