EPJ Web of Conferences **245**, 07042 (2020) https://doi.org/10.1051/epjconf/202024507042
CHEP 2019

Construction of a New Data Center at BNL

Imran Latif[1,*], *Shigeki* Misawa[1,**], and *Alexandr* Zaytsev[1,***]

[1]Brookhaven National Laboratory, Upton, N.Y. U.S.A

Abstract. Computational science, data management and analysis have been key factors in the success of Brookhaven National Laboratory's scientific programs at the Relativistic Heavy Ion Collider (RHIC), the National Synchrotron Light Source II (NSLS-II), the Center for Functional Nanomaterials (CFN), and in biological, atmospheric, and energy systems science, Lattice Quantum Chromodynamics (LQCD) and Materials Science, as well as our participation in international research collaborations, such as the ATLAS experiment at Europe's Large Hadron Collider (LHC) and the Belle II experiment at KEK (Japan). The construction of a new data center is an acknowledgement of the increasing demand for computing and storage services at BNL.

1 Introduction

The Computing Facility Revitalization (CFR) project at Brookhaven National Laboratory (BNL) is aimed at repurposing the former National Synchrotron Light Source (NSLS) building as a new data center for the Scientific Data and Computing Center (SDCC). The new data center will be operational in early 2021 to host compute, disk storage and tape storage equipment for the ATLAS experiment. ATLAS is one of the four main detectors at the LHC accelerator at CERN (European Center for Nuclear Research) and is operated by a international collaboration of over 3000 physicists [3, 4]. The data center will be available later in the year for all other groups supported by the SDCC including the STAR, PHENIX and sPHENIX experiments at the RHIC collider at BNL, the Belle II experiment at the High Energy Accelerator Research Organization (KEK) in Japan, and the Computational Science Initiative at BNL [5–7]. Migration of services to the new data center is expected to begin with the installation of new core network equipment and the first of several new tape libraries for ATLAS in early fiscal year 2021 (FY2021, from October 2020 to September 2021), and is expected to extend through 2024. This paper will highlight the key mechanical, electrical, and plumbing (MEP) components of the new data center. Also, we will describe our plans to migrate IT equipment between the current and new data centers, the inter-operational period in FY2021, gradual IT equipment replacement in FY2021-2024, and show the expected state of occupancy and infrastructure utilization for both.

2 Existing Data Center

The existing 1,940 m^2 SDCC data center dates from the 1960's, with some additions from 2009. It is a Tier 1 or "non-redundant" data center as defined by the Uptime Institute's Tier

*e-mail: ilatif@bnl.gov
**e-mail: misawa@bnl.gov
***e-mail: alezayt@bnl.gov

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

classification system [1]. The data center possesses a raised floor of varying heights (30 cm or 75 cm) and load capacities (750 to 1,500 kg/m^2) depending on location. All equipment is air cooled using cold air delivered by the under floor cold air plenum and generated by computer room air handling (CRAH) units distributed throughout the data center. Cooling capability is non-uniform and hot/cold aisle containment is not used. The data center can support racks dissipating up to 10 kilowatts (10 KW) but only in selected locations. Most areas can only support racks dissipating up to 8 KW. As configured, the old data center cannot meet the power usage effectiveness (PUE) of less than 1.5 for existing data centers mandated by the U.S. Government executive order in effect at the start of the project [2]. CRAH power is not generator backed, resulting in loss of cooling in the event of utility power failure. Chilled water for the CRAH units is sourced from the central BNL campus chillers. The available power in the existing data center is in excess of 4 megawatts (4 MW), of which 3 MW are either flywheel or battery UPS power. 2.3 MW of the UPS power is backed by diesel generator. Distribution of power is geographically non-uniform and is heterogeneous, with a mix of 120V/208V single and 208V three phase circuits of various amperages (10A, 20A, and 30A). In totality, these characteristics make the existing data center ill-equipped to meet the reliability, availability, and serviceability (RAS) requirements of the SDCC. These requirements are driven by the service level agreements set in the U.S. commitments to the Worldwide LHC Computing Grid, a global collaboration of computing centers supporting computing for the LHC [8, 9].

3 New Data Center

The new SDCC data center, being built in the shell of the former NSLS light source building (see Figure 1), is a "Tier 3" class data center that meets the RAS requirements of the SDCC. All critical data processing equipment will be supported by a fully redundant infrastructure (N+1) that is concurrently maintainable (i.e. without facility shutdown). The data center is also fully self sufficient, capable of operating without utility power or campus chilled water. Also, the data center design targets a PUE of 1.2, so the real world PUE should be well below the 1.4 maximum for new data centers mandated by the U.S. Federal Government executive order in effect at the time of project definition [2].

3.1 Capacities

If fully built out the new data center will be able to support 9.6 MW of information technology (IT) load, six 18 frame tape libraries, and $\approx$ 480 standard 42U, 19 in. equipment racks. Total IT floor space is roughly 1,600 m^2 (17,000 ft^2). The floor is a raised floor (height 76 cm (30 in)) and is rated for 2,440 kg/m^2 (500 lbs/ft^2). Water pipes for cooling are under the raised floor, while electrical and network services are provided from overhead busways and cable trays. The CFR project will only build out 50% of this ultimate design capacity. The mechanical, electrical, and plumbing systems are implemented as standard sized power and cooling systems, allowing for incremental increases in capacity beyond the initial 50% buildout.

3.2 Layout

IT equipment in the new data center are split between three separate rooms; a dedicated tape library room, a network room, and the main data hall for compute and storage equipment. Separate rooms are used as the power, cooling and fire suppression systems are different for

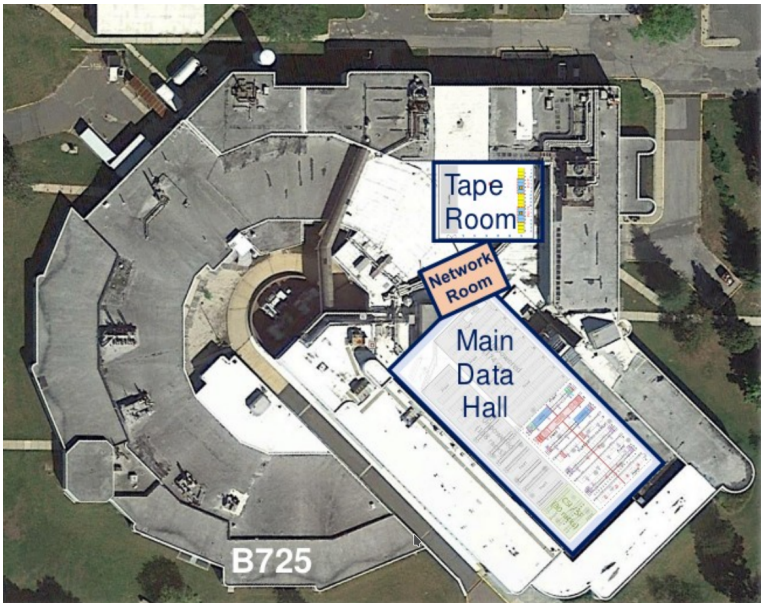

Figure 1. New Data Center - Layout within former NSLS building

the three rooms. Figure 1 shows the location of these rooms in the former NSLS building. All support equipment, except cooling towers, high voltage switch gear and diesel generators, are located within the building shell in the areas surrounding the IT rooms.

The main data hall, shown in Figure 2, is roughly 1,115 m^2 (12,000 ft^2) and is split into two areas; a High Throughput Computing (HTC) area and a High Performance Computing (HPC) area. The HTC area can host 16 rows of equipment, with 20 standard equipment racks per row. However, only 8 rows will be provided with power and cooling pipes in the 50% buildout. The HTC rows alternate between rows of 10 KW racks and rows of 20 KW racks. The 10 KW racks are for redundantly powered equipment, mostly consisting of servers for critical services and storage equipment. The 20 KW racks are for non-critical "stateless" compute servers. The HPC area consists of space for 15 rows of up to 10 standard equipment racks, of which 3 rows are energized in the base buildout. The 50% buildout area is delineated by the dash-dot lines in Figure 2.

The tape library room, shown in Figure 3, is approximately 310 m^2 (3,300 ft^2) and can accommodate up to six 18 frame linear tape libraries. Alternate library configurations are also possible to accommodate a different library form factor. The network room, shown in Figure 4, is approximately 120 m^2 (1,300 ft^2).

3.3 Electrical Power

Primary electrical power for the data center is implemented with multiple, equal sized power modules, each supporting 1.2 MW of IT load. Each module consists of one 1.75 MW diesel generator with a 24 hour fuel tank and one 1.2 MW valve regulated lead acid (VRLA) battery based uninterruptible power systems (UPS) with a 5 minute run time at maximum load. The diesel generator is sized to power the chiller and water pumps required to cool the 1.2 MW IT load in addition to the IT load. Critical water pumps in the cooling system are on UPS to maintain water circulation during the transition from utility power to diesel generator. Also,

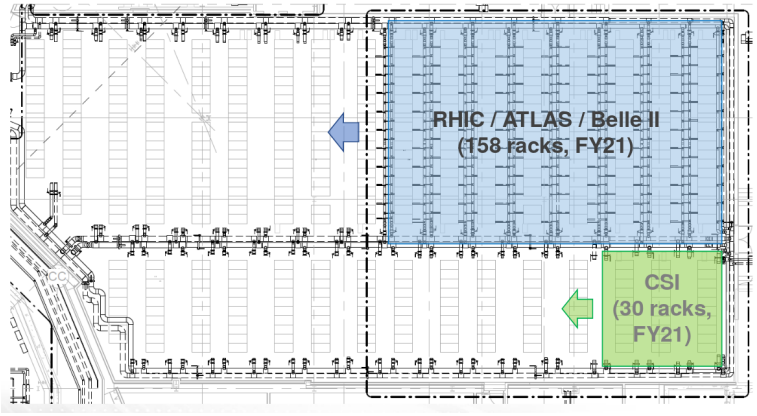

Figure 2. New Data Center - Main data hall floor plan

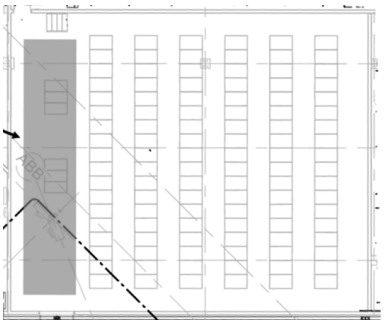

Figure 3. New Data Center - Tape library room floor plan

the generators are connected in parallel, allowing any generator to feed any of the 1.2 MW UPS systems. There is a single 1.2 MW maintenance bypass module to allow for maintenance of any one of the 1.2 MW UPS systems. Load downstream of each 1.2 MW UPS system is connected to the UPS and the bypass system by a static transfer switch. In this configuration, the bypass module can replace one of the primary UPS systems in the event of a failure or a maintenance shutdown of a primary UPS. The input power for the 1.2 MW bypass system is either from the utility feed or from the diesel generators, there is no UPS in the bypass system. In project management lingo, the bypass and three primary modules are in the "base" build, allowing for 3.6 MW of IT load. The fourth primary, required to reach the 4.8 MW 50% buildout, is an "add alternate" to the base build.

3.4 Electrical Distribution

Each 1.2 MW UPS is assigned to one of two possible loads. Two UPS systems power the HTC compute farm, storage, and infrastructure server equipment. One UPS system powers the network, tape, and HPC equipment.

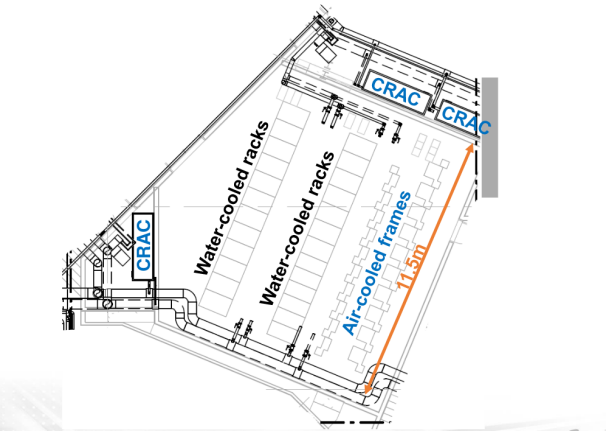

Figure 4. New Data Center - Network room floor plan

3.4.1 HTC UPSes

Each HTC UPS system feeds four 400 KW power distribution units (PDUs). As mentioned previously, each PDU is also connected to the 1.2 MW bypass system. Each PDU feeds two 200 KW overhead busway that are used to distribute power to individual racks. Two of the PDUs connected to a single HTC UPS, supporting a total of 4 overhead busways, supply power to forty 20 KW equipment racks. The remaining two 400 KW PDUs attached to a single HTC UPS are configured in an A/B pair, yielding two A/B pairs of 200KW busways. The two A/B pairs support forty 10 KW equipment racks. In total the two HTC UPS systems support eighty 20 KW racks and eighty 10 KW racks.

Contained in each 20 KW rack are two 3-phase 208V 50A in cabinet power distribution units (aka CDUs, "power strips" or in rack power distribution units). Both CDUs are connected to the same busway, and are used to power "stateless", single power supply, compute servers (one CDU per server) Although each CDU provides up to 14.4 KW of usable IT equipment (after 20% derating), by convention racks will be populated such that no more than 10 KW of power is drawn from each CDU. Each 10 KW rack also contains two of the same type of CDUs; however each CDU is connected to a different busway: one CDU to the "A" busway and the other to the "B" busway. The 10 KW racks are designed for storage systems, infrastructure servers and other critical systems that require redundant power. This design provides CDU commonality between 10 KW and 20 KW racks and allows for higher powered racks, up to 14.4 KW and 28.8 KW respectively, with no equipment changes. However, busway capacity limits the number of racks that can be supported at the higher power consumption.

3.4.2 HPC UPS

The single HPC UPS system feeds five 300 KW PDUs, of which three are allocated for HPC computing in a non-redundant configuration. At this point in time, distribution to the racks is undefined, as input power requirements have not been defined for the HPC system(s). The remaining two PDU's are configured in an A/B pair to feed network and tape library equipment in their respective rooms. Finally, like the HTC PDUs, the HPC PDUs are also connected to the 1.2 MW bypass system; however, for the A/B pair, the bypass power flows

through a 300 KW UPS before reaching the PDUs. Power distribution in the network room is through A/B pair overhead busways, while tape library power is hard wired to distribution panels.

3.5 Cooling

The new data center utilizes two cooling methods, computer room air conditioner (CRAC) based air cooling with under floor cold air plenums and active rear door heat exchanges (RDHx). Use of CRACs instead of CRAHs is partially driven by the higher chilled water temperature in the new data center. CRACs have compressors as they are air conditioners, while CRAHs are basically heat exchanges. The tape library room is cooled by two CRAC units in a 1+1 redundant configuration at 50% buildout. Full buildout requires the addition of a third CRAC unit, resulting in an N+1 redundant configuration. Network equipment is also air cooled, with redundant CRAC units. However, chilled water infrastructure will be installed in the room to support RDHx units if necessary. Mechanical and electrical rooms are also cooled with CRAC units.

The main data hall utilizes active RDHx (i.e. RDHx with fans) for cooling HTC equipment. HPC cooling remains undefined, although infrastructure is in place for RDHx cooling. There are four chilled water loops in the main data hall to support the cooling solution, two loops for the HTC areas (10 KW/20 KW racks, 20 racks per row) and two loops for the HPC areas (10 racks per row, 30 KW per rack). Water supply pipes are sized to allow for 30 KW power dissipation per rack in rows with 20 KW and 30 KW racks and 15 KW power dissipation per rack in rows with 10 KW racks assuming RDHx as the cooling mechanism. However, there is insufficient chilled water capacity to support running all racks at these higher levels. (Note also that the power distribution system also cannot support all racks running at these higher power levels.) For redundancy, RDHx units in any given row alternate between branch piping in front and behind each rack for chilled water.

Chilled water for the CRAC units and RDHx are supplied by multiple 445 ton (1.57 MW) chillers, with one chiller matched to each 1.2 MW power system. Three chillers are in the base project with the fourth chiller, like the fourth power system, a project "add alternate". Chiller maintenance/redundancy is accomplished via a "maintenance bypass" chiller that consists of a heat exchange driven by the BNL central chiller plant. Chilled water temperature is 15.5° C (60° F).

3.6 Monitoring

As part of the process to meet mandated energy efficiency regulations and meet RAS requirements, the mechanical, electrical and plumbing systems will be monitored by the BNL campus building automation system (BAS) [10]. A separate, data center monitoring system (DCIM) will also be used to monitor and control systems within the data center. This includes, among other things, rack PDUs, rear door heat exchanges, and ambient temperatures. At this point in time, a prototype of the DCIM is being deployed in the existing data center.

3.7 Network

As mentioned previously, there is a dedicated room for all network equipment except for top of rack and low latency HPC switches. Network cabling distance to all equipment racks in the tape room and main data hall from the network room falls within the 100 m distance limit of certain types of Ethernet. All network cabling between the network room and equipment racks are in overhead cable trays, with separate trays for copper and fiber cabling. Network

patch panels for rack connectivity to the network room are also overhead. There will be three independent networks in the new data center, the production data network, the BAS/DCIM monitoring network, and a local, row based out of band management network for IT equipment.

3.8 Migration

Migration to the new data center is a phased migration, with newly purchased equipment installed in the new data center, and concurrent staged retirement of existing equipment in the old data center. Except for selected items, notably the newest compute nodes in the existing data center, no equipment will be physically moved from the old data center to the new one. As would be expected, during the migration period, the SDCC will be operating equipment in both data centers. To make the phased migration possible, the SDCC network will be extended to the new data center.

Timing of the migration is critical as Run 3 of the LHC is expected to start in early CY2021 and no critical services can be taken off line once the run starts [11]. Although the new data center is not expected to be completed by early calendar year 2021 (CY2021), the construction schedule is designed to allow early occupancy of selected portions of the data center to allow ATLAS services at the SDCC to be migrated prior to the start of Run 3. Network connectivity to the new data center will be established in late CY2020 and critical ATLAS and SDCC support services will be moved to the new data center in late CY2020 or early CY2021. Migration of non critical ATLAS services and remaining SDCC services is expected to start later in CY2021 when allowed by the construction schedule. The migration process is expected to end at some point between CY2023 and CY2024. At the end of the migration, only the legacy tape libraries and associated servers will be operating in the old data center.

4 Summary

The new data center at BNL will substantially enhance the capabilities of the SDCC. From the perspective of SDCC customers, the new data center will allow the SDCC to support more and higher powered equipment and increase facility availability. From the perspective of the SDCC, it will reduce operational complexity, maintenance burdens, and energy consumption and simplify the installation of new equipment.

References

[1] Uptime Institute, *Tier Classification System* (2021 February). Retrieved from https://uptimeinstitute.com/tiers.

[2] U.S. Federal Government. (2015, March 19), "Executive Order (EO) 13693, Planning for Federal Sustainability in the Next Decade". Retrieved from https://www.fedcenter.gov/programs/eo13693/.

[3] The ATLAS Collaboration, Journal of Instrumentation **Volume 3**, (2008).

[4] The ATLAS Collaboration, "The ATLAS Experiment at CERN". Retrieved from https://atlas.cern.

[5] Brookhaven National Laboratory, "RHIC Relativistic Heavy Ion Collider". Retrieved from https://www.bnl.gov/rhic/.

[6] Belle II, "Belle II". Retrieved from https://www.belle2.org.

[7] Brookhaven National Laboratory, "Computational Science Initiative". Retrieved from https://www.bnl.gov/compsci.

[8] Worldwide LHC Computing Grid (WLCG). Retrieved from https://wlcg.web.cern.ch/.

[9] Worldwide LHC Computing Grid (WLCG), "Signed Memoranda of Understanding". Retrieved from https://wlcg.web.cern.ch/mou/signed.

[10] U.S. Federal Government Office of Management and Budget, (2019, June 25), "Memorandum M-19-19 Update to Data Center Optimizition Initiative (DCOI)". Retrieved from https://datacenters.cio.gov/policy/.

[11] European Organization for Nuclear Research (CERN), (2020, January 16), "LHC Long Term Schedule". Retrieved from https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm.