Finalizing Construction of a New Data Center at BNL

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

[1]Brookhaven National Laboratory, Upton, NY 11973-5000, 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) at CERN (Switzerland) 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 in the near term and enable the Lab to address the needs of the future experiments at the High-Luminosity LHC at CERN and the Electron-Ion Collider (EIC) at BNL in the long term.

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 an international collaboration of over 3000 physicists [1, 2]. 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 [3–5]. The period of migration of IT equipment and services to the new data center is going to start with the installation of new core network equipment and the first new tape library for BNL ATLAS Tier-1 site in the new data center in 2021Q3 (within U.S. fiscal year (FY) 2021, covering Oct 2020 to Sep 2021 period), and is expected to extend until the end of FY2023. In this paper we highlight the key mechanical, electrical, and plumbing (MEP) components of the new data center, as well as high level overview of its central network systems. In this paper we describe the

* Corresponding author: alezayt@bnl.gov

plans to migrate a subset of IT equipment between the old and the new data centers in the calendar year (CY) 2021, conduct operations of both data centers in parallel starting from 2021Q3, and perform a gradual replacement of IT equipment currently deployed in the old data center in CY2021-2024 period. We also show the expected state of occupancy and infrastructure utilization for both data centers up to FY2026.

2 Existing Data Center

The existing 1,940 m^2 SDCC data center dates from the 1960's, with some additions made in 2009. It is a Tier I or "non-redundant" data center as defined by the Uptime Institute's Tier classification system [6]. 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 the 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 [7]. 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 powered. 2.3 MW of the UPS power is backed by diesel generator. Distribution of power across the data center areas is non-uniform, with a mix of 120V/208V single and 208V 3-phase circuits of various amperages (10A, 20A, 30A, and 50A rated). Combined, these characteristics make the existing data center ill-equipped to meet the reliability, availability, and serviceability (RAS) requirements of the SDCC Facility. These requirements are primarily 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 III 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 BNL campus chilled water for prolonged periods of time (provided that the diesel generator group is getting refuelled). The data center design targets a PUE of 1.2, and the seasonal variation of the monthly averaged PUE is expected to be in the 1.17–1.28 range given the climate conditions at BNL geographical location, 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 [7].

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 478 standard 42U, 19 in. wide, up to 1200 mm deep equipment racks. Total IT floor space is roughly 1,600 m^2 (17,000 ft^2).

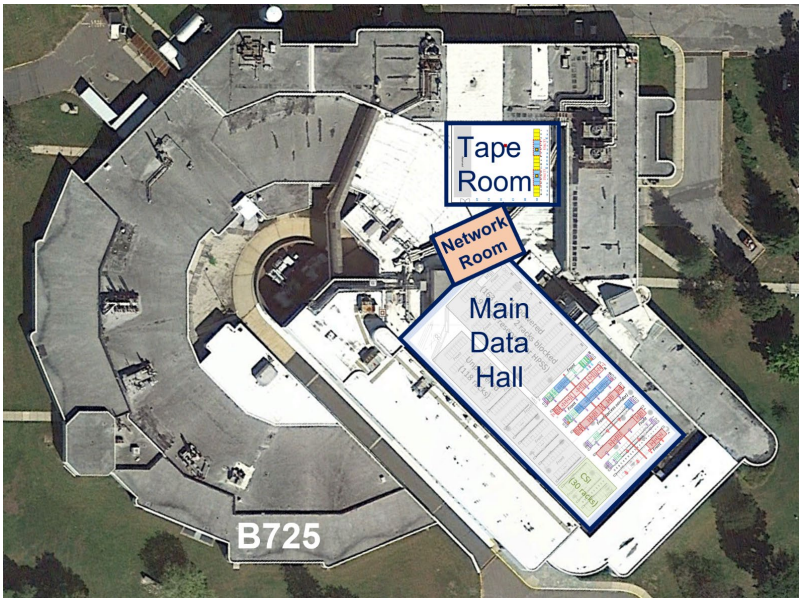

Fig. 1. New Data Center – Layout within former NSLS building.

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 37.5% of this ultimate design capacity. The mechanical, electrical, and plumbing systems are implemented as standard sized power and cooling systems, allowing for increases in capacity beyond the initial buildout by increments of 1.2 MW of IT payload capacity.

3.2 Layout

IT equipment in the new data center is split between three separate rooms: a dedicated Tape Library Room, a Network Room, and the Main Data Hall (MDH) for compute and storage equipment. Separate rooms are used as the power, cooling and fire suppression systems are different for 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 (MDH), 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 base (37.5%) 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 base buildout area is delineated by the dash-dot lines in Figure 2.

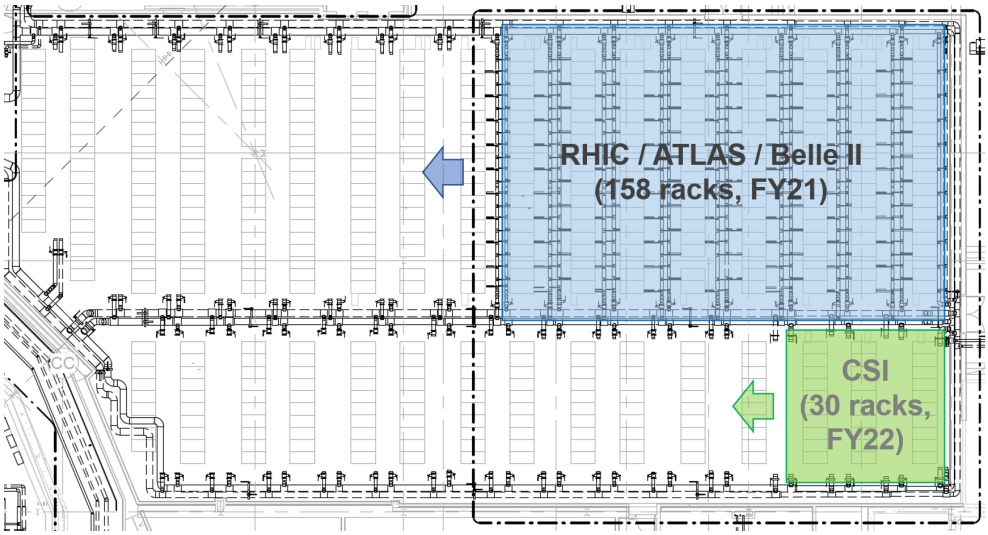

Fig. 2. New Data Center – Main Data Hall (MDH) floor plan.

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 such as IBM TS45000 series modular SCSI controlled tape libraries. Alternate library configurations are also possible to accommodate a different library form factor (frame composition and dimensions, within the limits of usable space of the Tape Library Room).

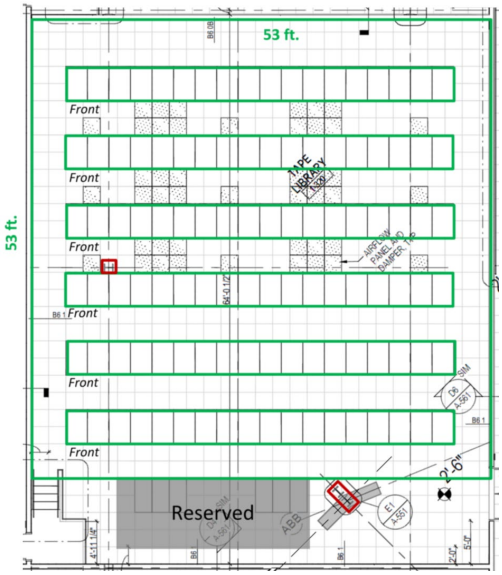

Fig. 3. New Data Center – Tape Library Room floor plan.

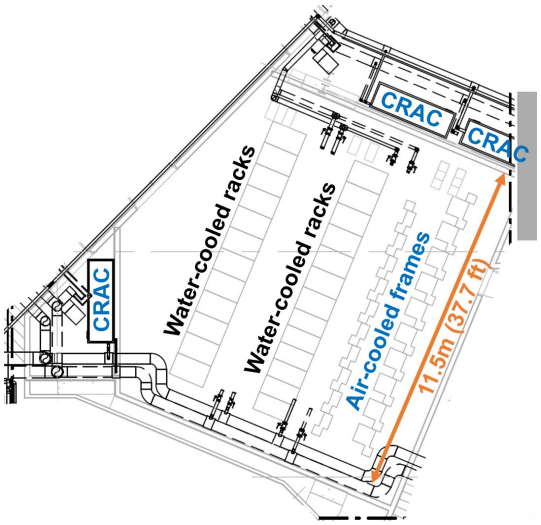

Fig. 4. New Data Center – Network Room floor plan.

The Network Room, shown in Figure 4, is approximately 120 m^2 (1,300 ft^2) and can host up to 24 racks for central network equipment and up to 20 two-post air-cooled frames for Ethernet copper and optical fiber distribution systems.

3.3 Electrical Power

Primary electrical power for the data center is implemented with multiple (up to 8 in the full buildout), 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 of IT load (PUE part of the data center electrical load) on top of the 1.2 MW of 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, 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. While 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 besides a limited scale 0.3 MW battery-based UPS added specifically to provide additional protection for central network systems of the new data center. The bypass and three primary power system modules with two shared diesel generators are in the base build of the new data center, allowing for 3.6 MW of IT load out of which 2.4 MW is protected by UPS and diesel backup in the initial buildout – as to be delivered to SDCC by the CFR project in 2021Q3. The fourth primary power system, required to reach the 4.8 MW (50% maximum capacity) buildout is expected to be added to the data center in FY2024-2025. Up to 3 additional 1.75 MW generator

modules are expected to be added to the shared diesel generator group in FY2022-2025 period, as we scale the IT load beyond the 2.4 MW profile, thus bringing the diesel generator group to 5 units out of 9 of maximum design capacity (serving 8 power system modules and the bypass system of the data center).

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 1.2 MW UPS system powers the network, tape, and HPC equipment.

3.4.1 HTC UPS Systems

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 200 kW 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 & Infrastructure UPS System

The single UPS system feeds five 300 kW PDUs, of which three are allocated for HPC computing in a non-redundant configuration. The HPC systems are to be physically placed in a dedicated area of the Main Data Hall as shown in Figure 2. The remaining two PDU's are configured in an A/B pair to feed network and tape library equipment in their respective rooms. At this point in time, the exact distribution to the HPC racks is undefined, as input power requirements are only expected to be finalized for the HPC storage and computing systems to be deployed in this area only by the end of FY2021. 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 an additional 300 kW battery-based UPS before reaching the PDUs. This additional protection is needed to make sure that under no scenario is the most critical part of the data center infrastructure which is the central network systems placed in the Network Room ever fed from the unprotected power source on both sides at the same time. 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 exchangers (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 exchangers. The Tape Library Room is cooled by two CRAC units in a 1+1 redundant configuration at the base (37.5%) 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 (MDH) utilizes active RDHx (i.e., RDHx with fans) for cooling HTC equipment. HPC cooling remains undefined at this moment, although infrastructure is in place for RDHx cooling. Since most of the heat dissipation of the IT load of the data center is going to be removed using water cooling under normal operational conditions, the ambient air temperature distribution in the Main Data Hall is expected to be near uniform, with a set point of 23.9° C (75° F).

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 and RDHx units is 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 Control & 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 infrastructure management (DCIM) will also be used to monitor and control systems within the data center. This includes, among other things, rack PDUs, rear door heat exchangers, and ambient temperatures. At this point in time, the production deployment of DCIM is in progress in the existing data center environment which is to be expanded into the new data center in 2021Q3.

3.7 Network

As mentioned previously, there is a dedicated room for all network equipment except for top of rack switches, distribution switches of the infrastructure racks, and low latency HPC interconnect switches that are expected to be deployed directly in the racks of the CSI segment of the Main Data Hall (MDH). Network cabling distance to all equipment racks in the Tape Library Room and the MDH from the Network Room falls within the 100 m

distance limit of certain types of Ethernet and Fibre Channel storage interconnect. The following types of network interfaces are expected to be most commonly used for connectivity distribution in the Main Data Hall using multi-mode fiber (MMF): 10 GbE, 25 GbE, 40 GbE, 100 GbE, (now) and 50 GbE, 200 GbE as well (in the next 5 years). The majority of the short- (SR4) and long-range (LR4) inter-switch links used by the central network systems of the data center are to be based on 100 GbE (now) and 400 GbE (in the next 5 years). The network core of the data center is fully IPv6- enabled and is capable of supporting the line cards with 400 GbE ports in the future in the existing modular central switched being deployed in it starting from FY2021. Thus, the transition to 400 GbE technology can happen in this environment before the next hardware refresh cycle for central network equipment which is anticipated in FY2025-2026 period.

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 to be deployed in the overhead position and which helps to maximize the utilization of the rack space by active equipment. 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.

The primary network link between the two data centers: the old one and the new one, is going to be commissioned in the initial configuration as 1.6 Tbps (unidirectional bandwidth) comprised of 16x 100 GbE long range (LR4) Ethernet uplinks which can be later upgraded to higher bandwidth based on 100 GbE LR4 or 400 GbE LR4 individual lines. The physical infrastructure of the interbuilding link consists of two 288-strand single-mode (SMF) fiber cables connecting central network equipment in the old and the new data center over redundant physical paths, thus theoretically allowing the deployment of up to 144 redundant high bandwidth Ethernet or Fibre Channel uplinks between the data centers.

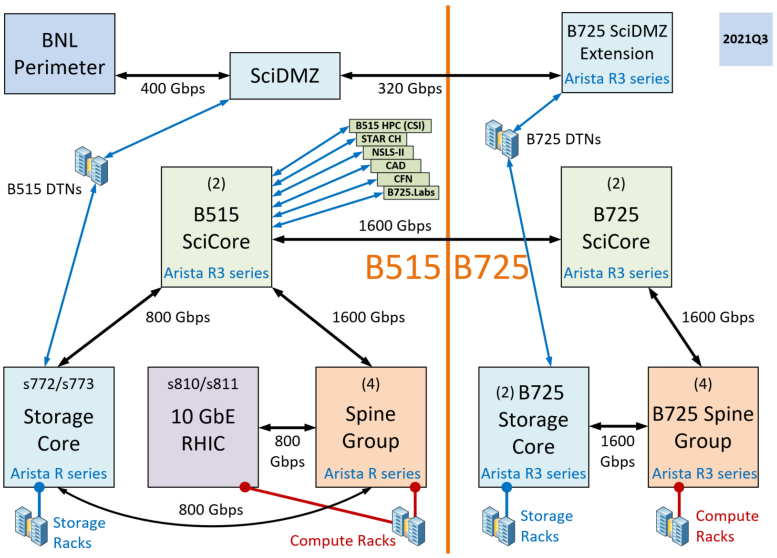

Fig. 5. High level diagram of network systems integration of the old data center (B515 on the left) and the new data center (B725 on the right), once the central network systems are extended into the new data center in 2021Q3.

3.8 Transition to Operations

Migration to the new data center is going to be performed in a manner of a phased migration, with newly purchased equipment installed in the new data center starting from 2021Q3 (which permits a significant portion of FY2021 purchases of compute and data storage equipment to target the deployment directly in the new data center), and concurrent staged retirement of existing equipment in the old data center (the process that is expected to be completed by the end of CY2024). This schedule is already taking into account all the delays inflicted on data center construction process due to the COVID-19 pandemic countermeasures implemented in U.S.A., State of NY, U.S. Department of Energy (DOE), and Brookhaven Laboratory in 2020Q2-3 period.

Except for selected items, notably the newest compute node racks in the existing data center (currently 18 racks with 2 or more years of projected lifetime on the data center floor as of FY2021, which corresponds to about 7% of the number of racks with IT equipment deployed in the old data center as of Feb 2021), 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 in 2021Q2 for which all preparations are made as of Feb 2021 to make sure that this can be done completely transparently to the old data center operations.

Timing of the migration is critical as Run 3 of the LHC is expected to start in early CY2022 due to the additional delays related to COVID-19 pandemic, and no critical services can be taken offline once the run starts [11]. Due to 1 year delay in the Run 3 schedule the construction of the new data center is expected to be finished more than half a year before the beginning of the run which would provide ample opportunity to ramp up its operations for ATLAS, RHIC and Belle II collaborations by the time the ATLAS Run 3 starts. The first round of deployment of HPC compute and storage systems for the Computational Science Initiative at BNL on the floor of the Main Data Hall of the new data center is expected to be carried over during the 2021Q4 period. The process of gradual transition to the new data center expected to be completed by the end of CY2024. At the end of the transition, only the tape libraries (mostly the legacy Oracle SL8500 silos, with recent addition of limited scale IBM TS4500 2-frame library) and associated servers will be operating in the old data center.

Figure 6 shows the projected evolution of the total number of racks with IT load to be deployed in the new data center in FY2021-2026 period with a high level breakdown between the main different data center clients as anticipated in the plans made as of Feb 2021.

4 Summary

The new data center at BNL will substantially enhance the capabilities of the SDCC Facility. 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, energy consumption and simplify the installation of new equipment.

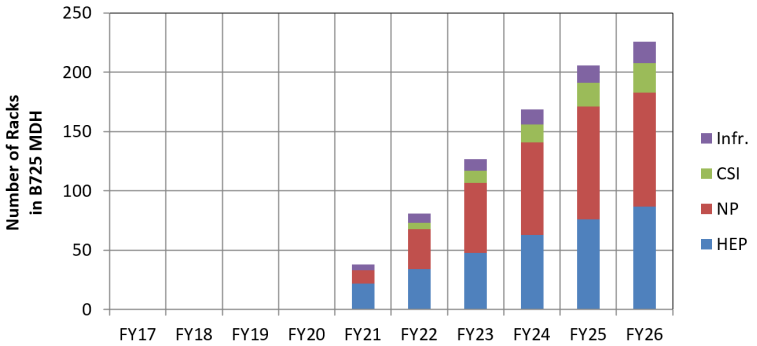

Fig. 6. Projected evolution and composition of the number of racks to be deployed in the new data center in FY2021-2026 period. The racks are grouped into four broad categories: Infrastructure (Network Room and Main Data Hall (MDH)), Computational Science Initiative (CSI; primarily HPC in the MDH which also includes 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) resources), Nuclear Physics (STAR, PHENIX and sPHENIX detectors at RHIC collider at BNL) and High Energy Physics (ATLAS detector at the LHC and Belle II detector at KEK).

References

[1] The ATLAS Collaboration *et al.,* The ATLAS Experiment at the CERN Large Hadron Collider, 2008 JINST 3 S08003: https://iopscience.iop.org/article/10.1088/1748-0221/3/08/S08003/meta

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

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

[4] The Belle II Collaboration, "The Belle II Experiment at KEK (Japan)": https://www.belle2.org

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

[6] Uptime Institute, Tier Classification System: https://uptimeinstitute.com/tiers

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

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

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

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

[11] European Organization for Nuclear Research (CERN), "LHC Long Term Schedule" (2019-2036): https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm