SDCC Operations During Transition to the New Data Center

Alexandr ZAYTSEV

alezayt@bnl.gov

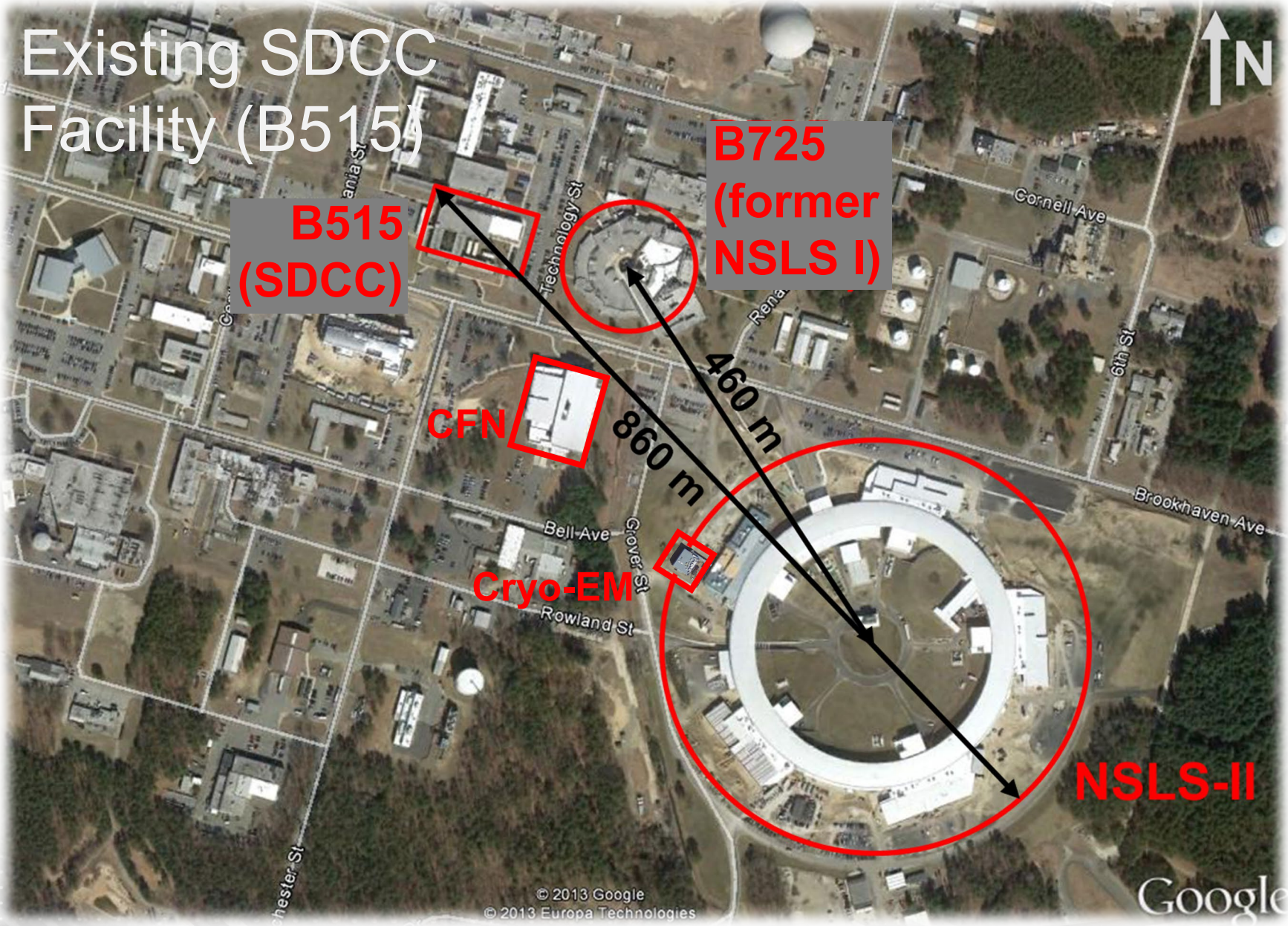
BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

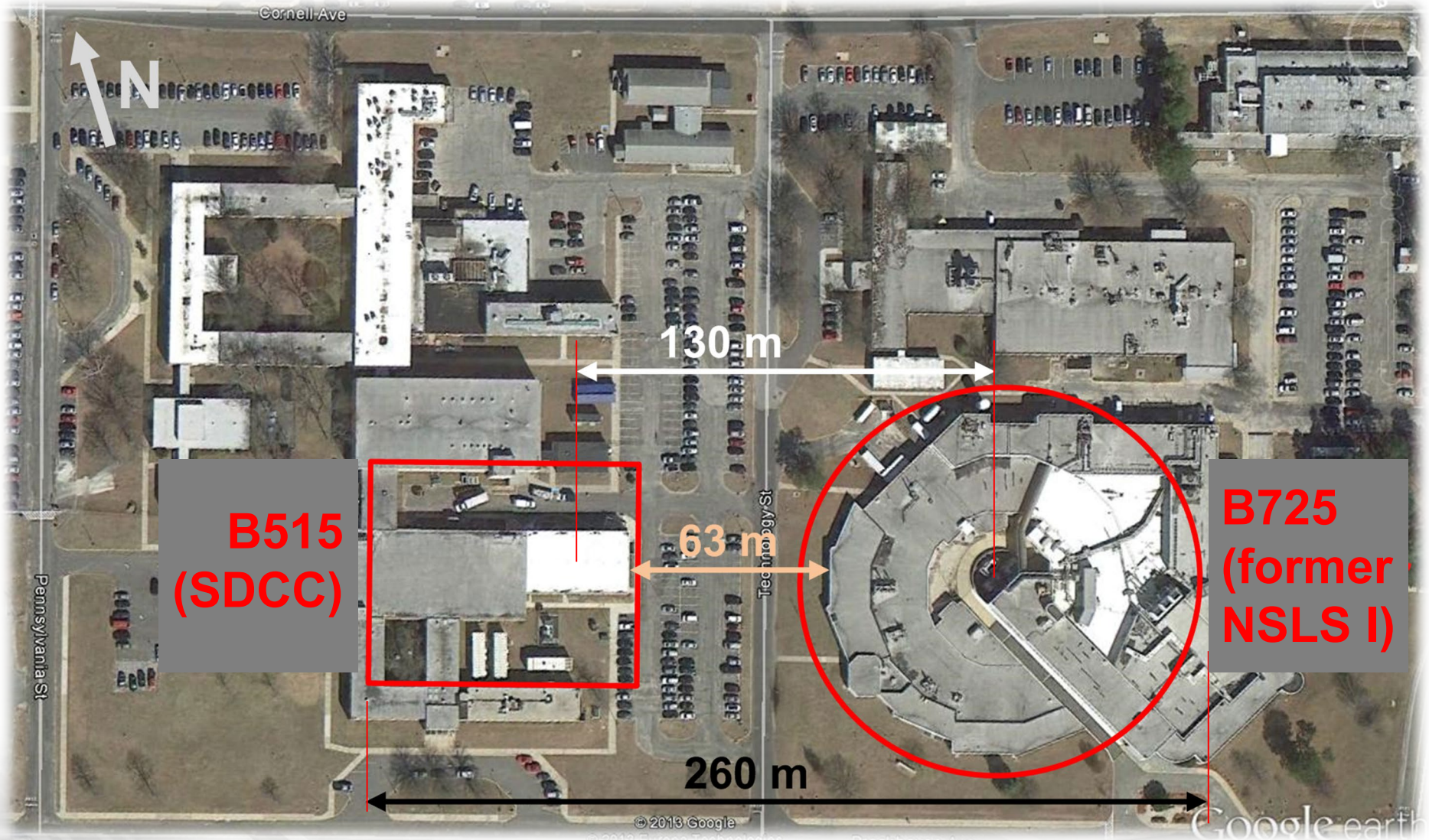
B515 Data Center

- Scientific Data and Computing Center (SDCC) Facility is currently operating a single B515 based data center which is a 1.5 MW scale air-cooled general purpose facility hosting 300 racks of equipment and 10 tape silos as of 2021Q1:
  - HTC and HPC computing systems (on the scale of 2500 nodes)
  - DISK storage (on the scale of 90 PB)
  - Robotic tape storage (on the scale of 200 PB)
  - *More details on the B515 data center configuration are given in the "SDCC Datacenter Transformation within the Scope of BNL CFR Project and Beyond" presentation delivered to HEPiX Spring 2019*
- Serving multiple international collaborations, research communities based at BNL, and also NY State based research organizations:
  - STAR, PHENIX and sPHENIX at RHIC (BNL) with two on site counting houses: one active (STAR CH) and another one expected to be reconnected to SDCC in 2021Q3 (sPHENIX CH)
  - ATLAS Experiment at the LHC (ATLAS Tier-1 Site)
  - Belle II Experiment at KEK (Belle II Tier-1 Site)
  - National Synchrotron Light Source (NSLS) II at BNL, currently with 29 active beamlines (expected to scale up to 60+ beamlines in the future) – *more details on SDCC support of NSLS II Facility to be given in the talk "Supporting a new Light Source at Brookhaven" immediately after this one in the session*
  - Center for Functional Nanomaterials (CFN) at BNL
  - BNL Computational Science Initiative (CSI) research groups and test labs
  - Cryo-EM Research Facility at BNL (currently in the process of joining in)
  - Simons Foundation (SF)

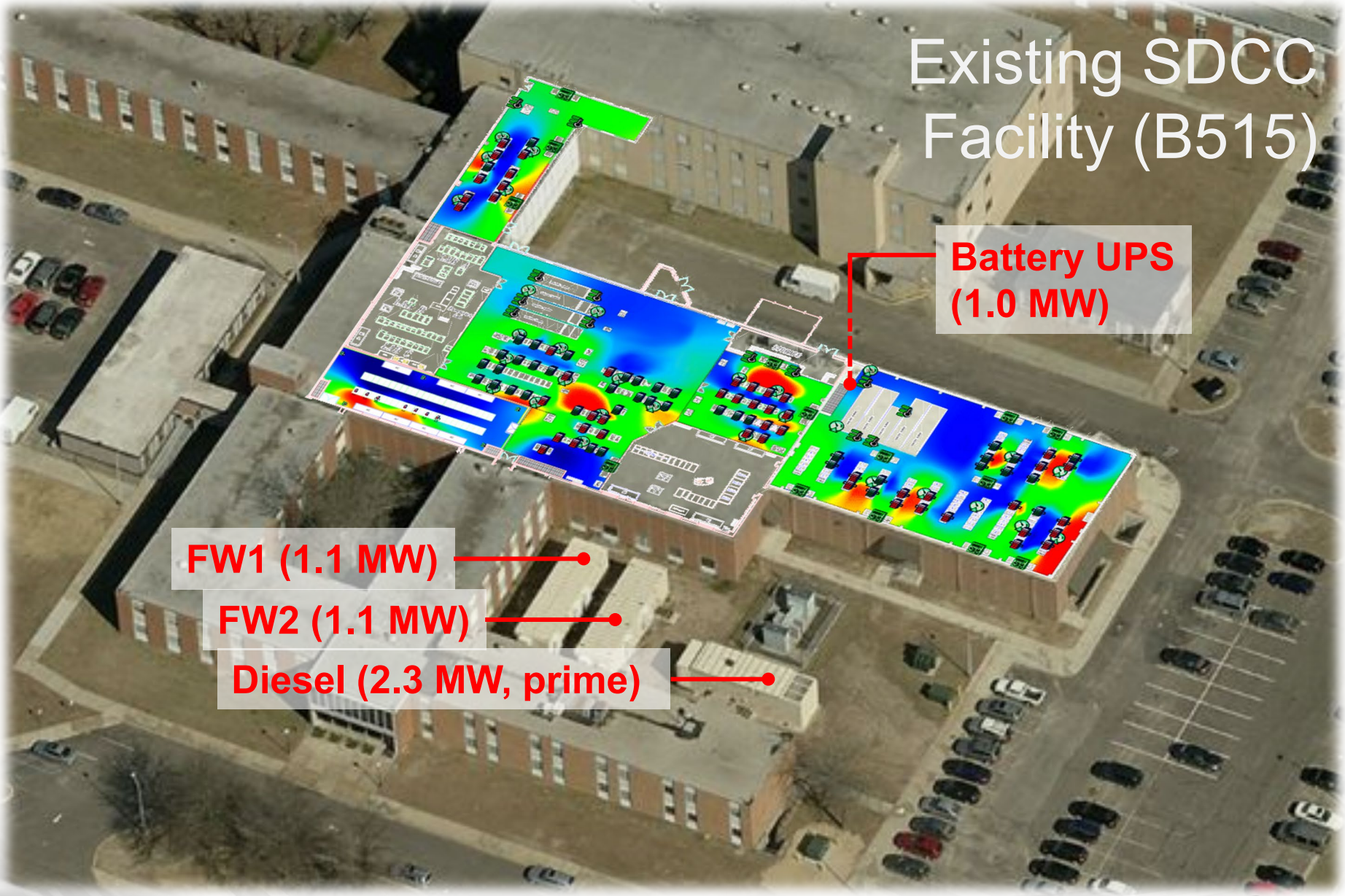
Existing SDCC Facility (B515)



Outlook: B515 & B725



Existing SDCC Facility (B515)

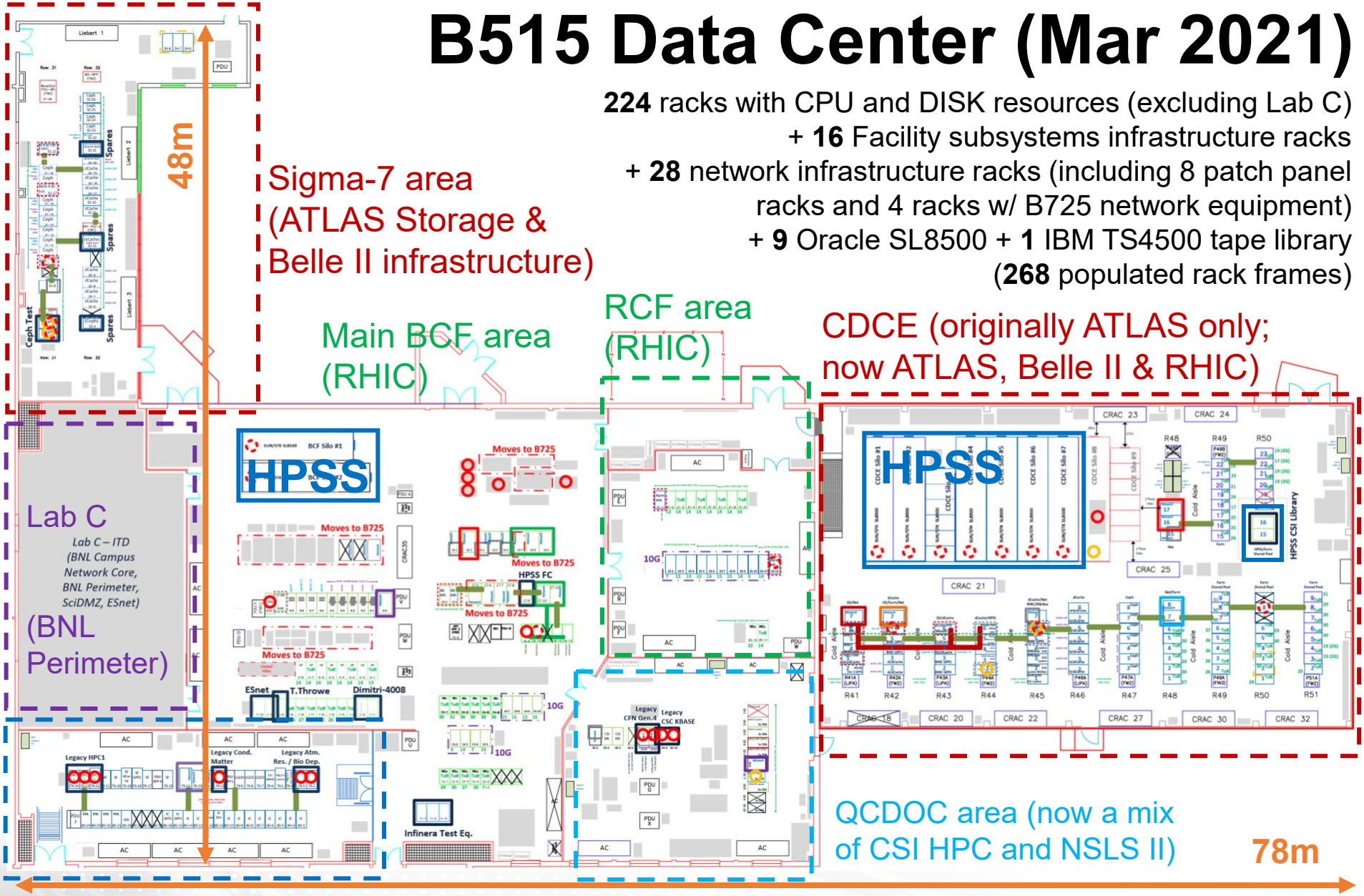


B515 Data Center Areas

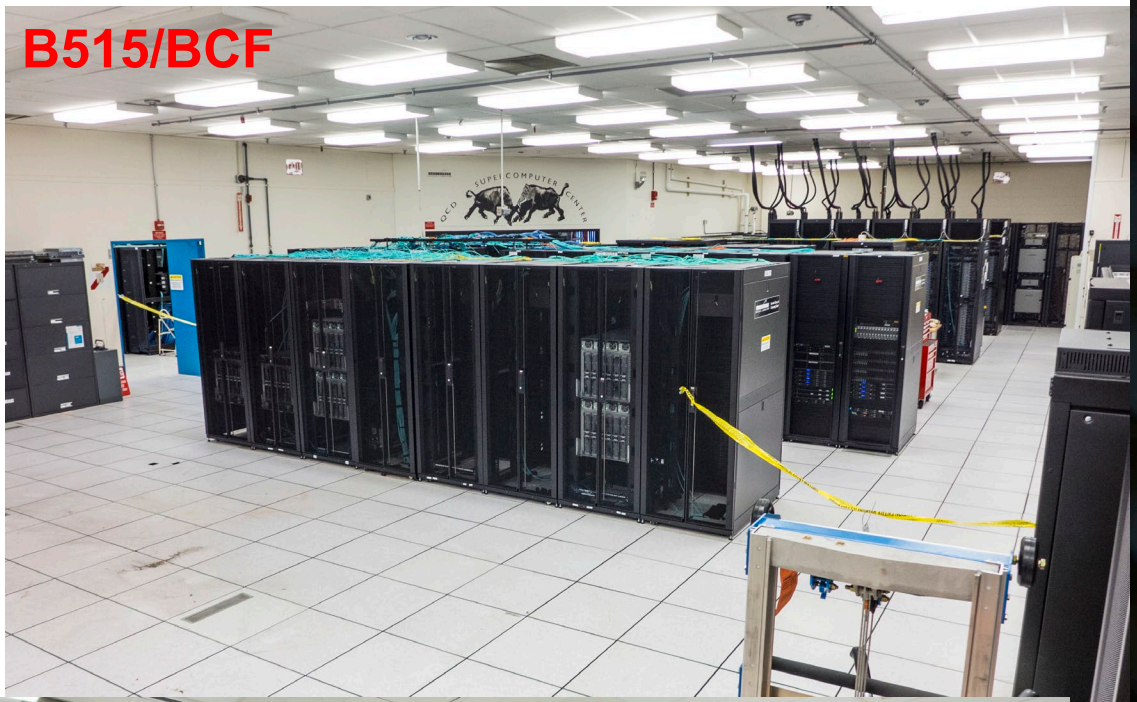
- SDCC B515 data center incorporates three main IT areas located in the same building each provided with its own cooling solution yet sharing the power distribution and power backup infrastructure
  - **BCF area** (the oldest part of the data center with operational history pre-dating the SDCC Facility and going back to the 1960s), now subdivided into several physically isolated areas:
    - **RCF** (RHIC Computing, Inergen based fire suppression)
    - **QCDOC** (former BG/Q supercomputer area, now converted to host Simons Foundation resources)
    - **BGL** (former BG/L supercomputer area, now converted to host CSI and part of SF resources)
    - **Lab C** (IT department managed central network resources, BNL perimeter equipment, ESnet equipment, fiber cross for the entire BNL site, Inergen based fire suppression)
    - **Main BCF** (the rest of space left)
  - **Sigma-7 area** (highest power redundancy and cooling infrastructure reliability area added to B515 data center in 2009; no Inergen based fire suppression here though) – primarily hosting ATLAS and Belle II equipment
  - **CDCE** (the newest expansion of B515 data center added in 2009) – the only area (besides Lab C) in the B515 data center that is expected to remain operational after FY23

B515 Data Center (Mar 2021)

224 racks with CPU and DISK resources (excluding Lab C)
+ **16** Facility subsystems infrastructure racks
+ **28** network infrastructure racks (including 8 patch panel racks and 4 racks w/ B725 network equipment)
+ **9** Oracle SL8500 + **1** IBM TS4500 tape library
(**268** populated rack frames)



B515/BCF

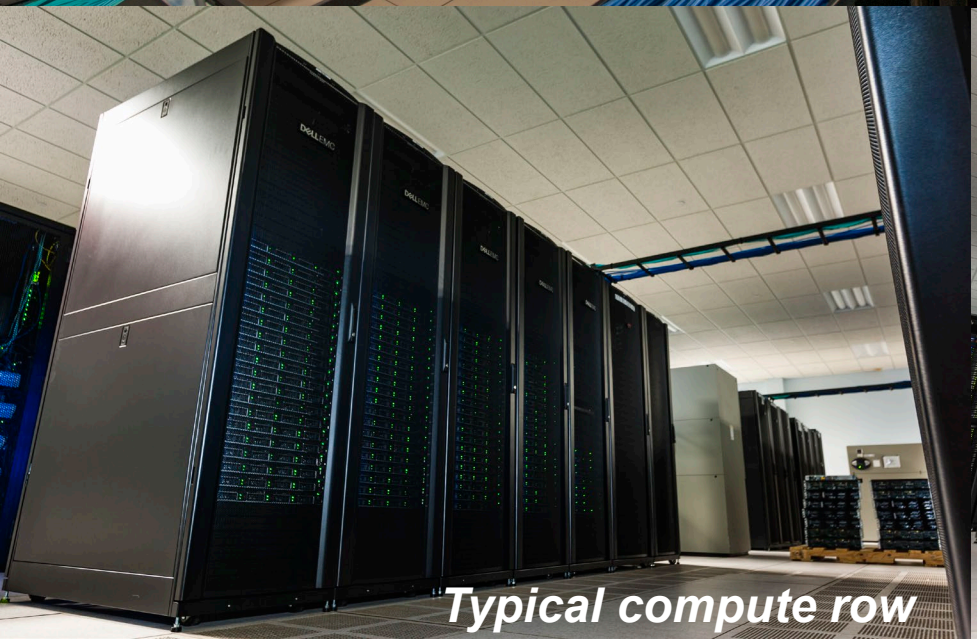


Oracle SL8500 silos being moved to CDCE


B515/BGL
HPC row


B515/Sigma-7
APC
CAUTION
NOISE HAZARD
HEARING PROTECTION REQUIRED
CAUTIO
NOISE
HAZA
HEARING PROTECT
REQUIRED IN THIS

B515/CDCE



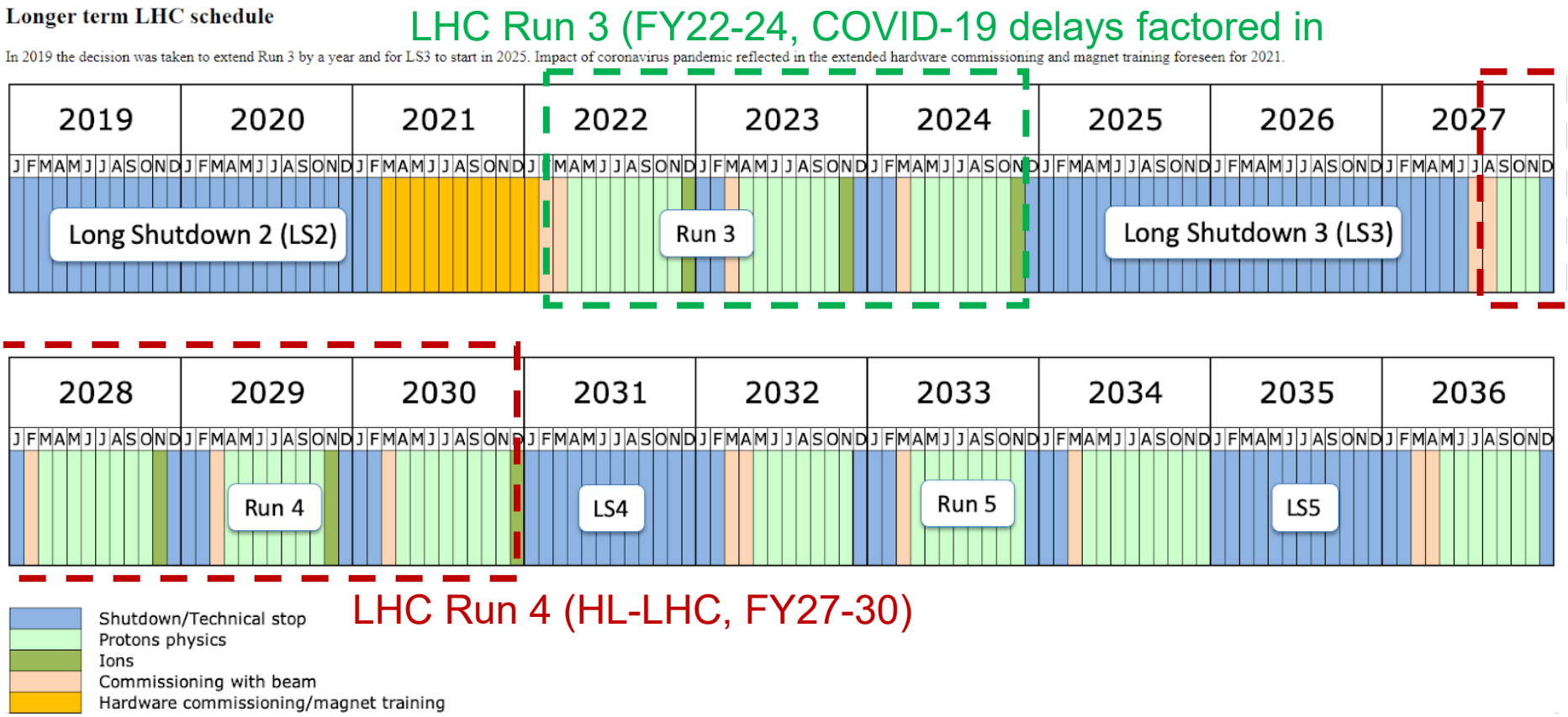
Typical compute row

Typical HW RAID storage row

The Path Forward

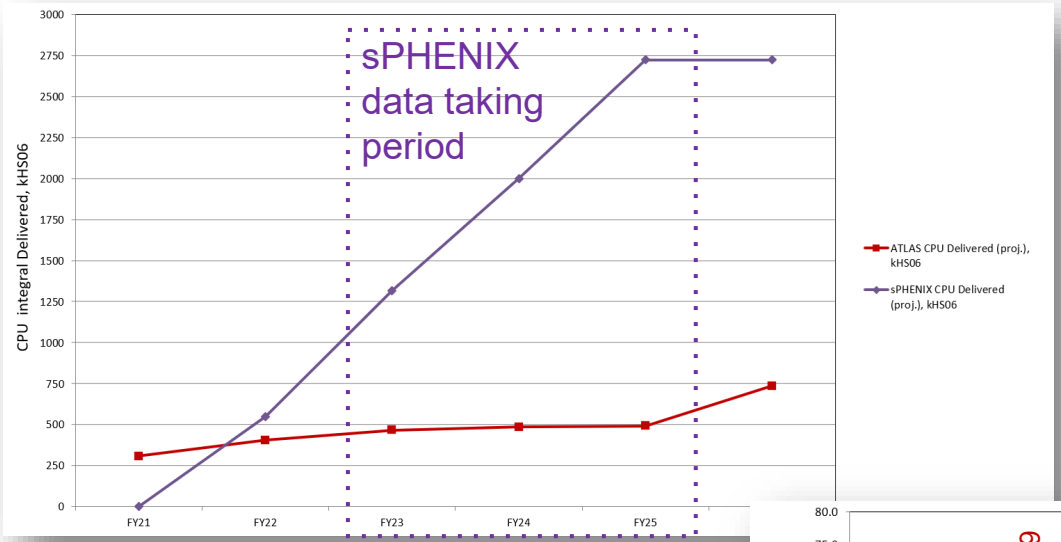
- The existing SDCC B515 data center is a highly non-uniform facility aggregating the history of several decades of infrastructure solutions, with none of its areas providing the feature set needed for addressing the future points of growth associated with:
  - sPHENIX Experiment at RHIC beginning the data taking in FY23 and scaling up in FY23-25
  - STAR Experiment at RHIC continuing taking data in FY23-25
  - ATLAS / High Luminosity LHC (HL-LHC) starting from FY27
  - Scaling of the NSLS-II Facility to 60 active beamlines
- Furthermore, the following challenges to address driven by the need to increase IT equipment power density and efficiency of operations:
  - Increasing the power density of HTC CPU racks up to 20 kW/rack (current limit is about 12 kW/rack)
  - Scaling up the CSI HPC systems and increasing the power density of HPC racks up to 30 kW/rack (current limit is about 15 kW/rack)
  - Scaling the combined Facility power profile beyond 2 MW of IT load while protecting all its IT payload from the site-wide power outages and allowing the payload to remain operational during the prolonged site-wide power outages
  - Improve the overall PUE of the SDCC Facility below 1.4

The Path Forward (Cont.)



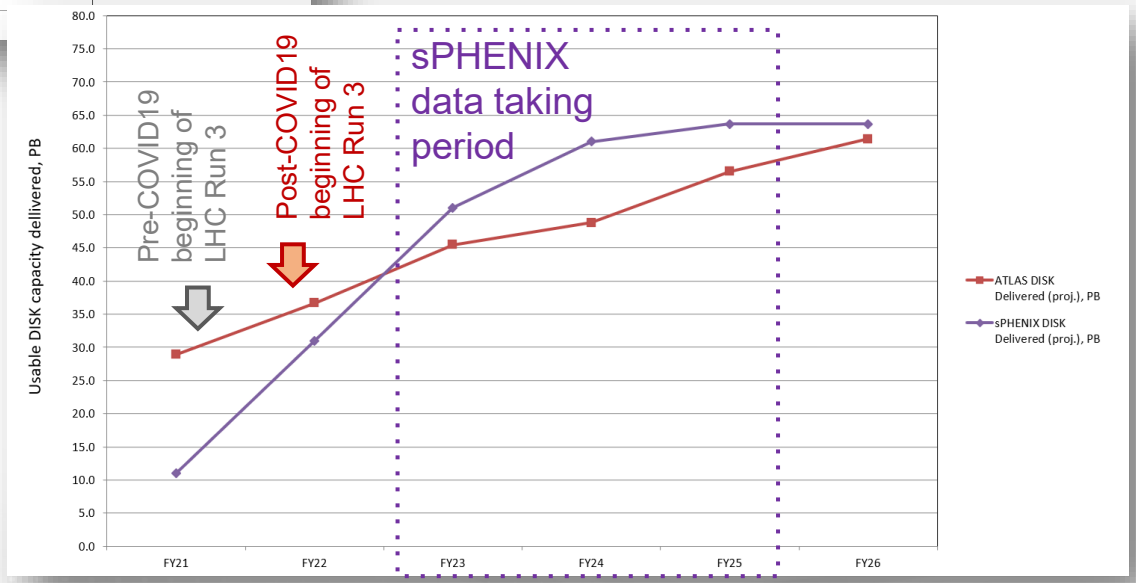
https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm

B725: Main Capacity Drivers of FY21-26 Period

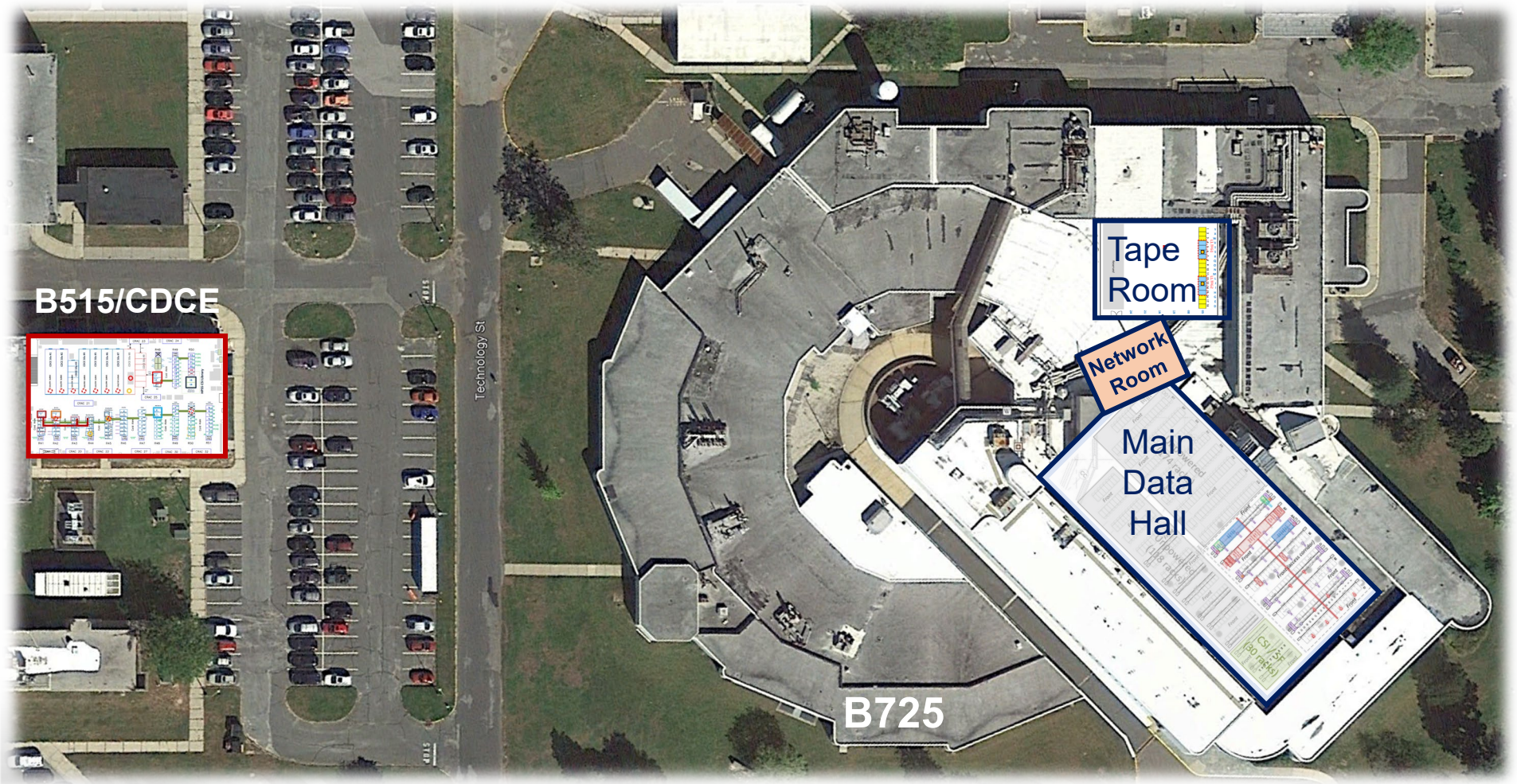


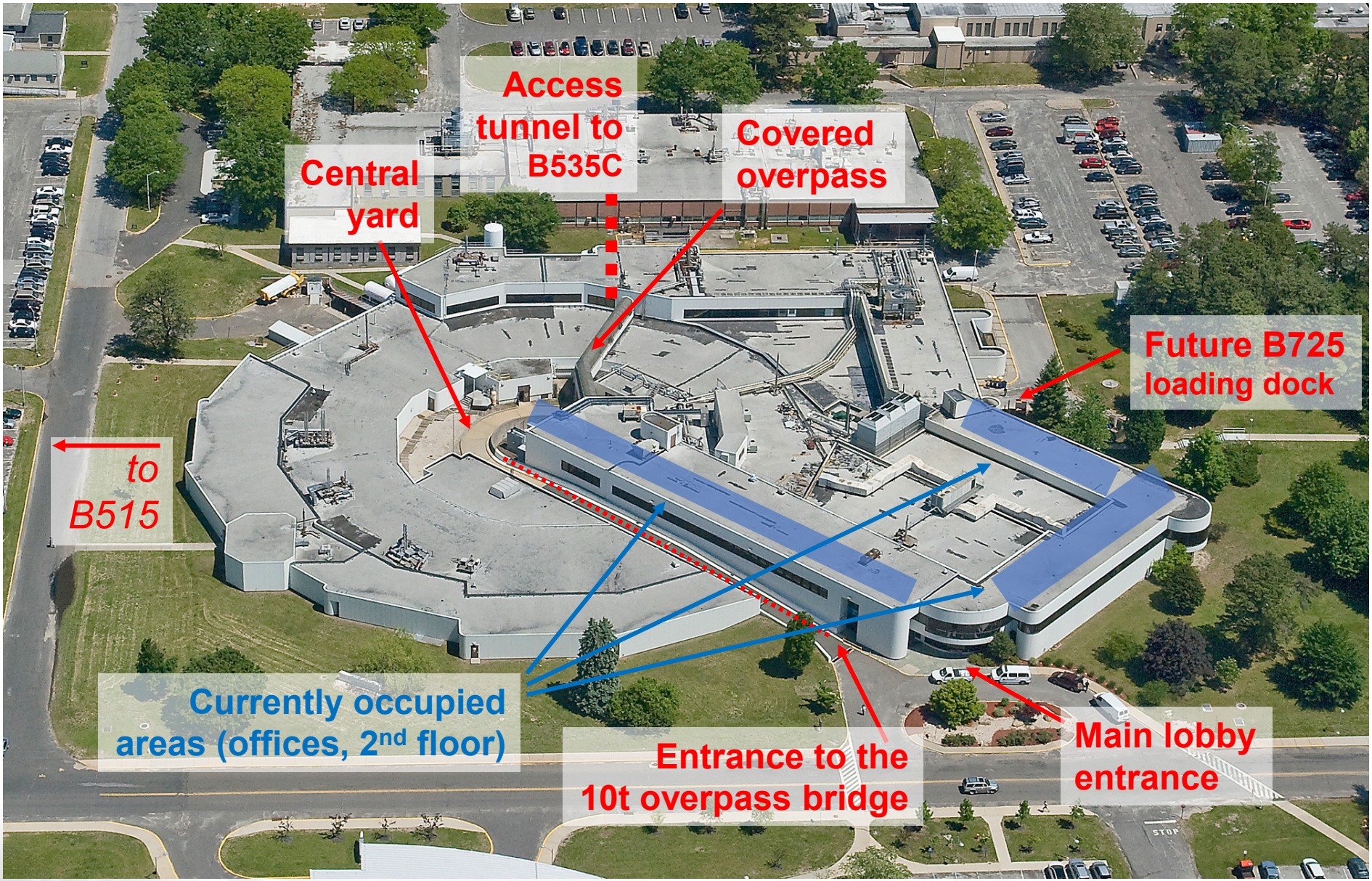
- ATLAS Tier-1 ramping up for the LHC Run 4 in FY25-26
- sPHENIX ramping up for data taking runs of FY23-25 starting from FY22
- STAR is also expected to take data in FY23-25

- sPHENIX SDCC equipment deployment stages:
  - FY20-21: 11 PB of usable DISK storage capacity is to be delivered for the MC campaign
  - CPU purchases begin in FY22 and the CPU integral reaches the 2.75 MHS06 plateau in FY25; RAW and DST disk buffer capacity is purchased in the same period (60 PB usable by FY25)
  - The HPSS tape libraries and movers are purchased in FY22 and FY24 (total capacity of ~40k tape slots)



New data center is being designed & constructed for the SDCC Facility in B725 (former NSLS I building) in FY19-21 period, with migration of most of the DISK and all the CPU resources (primarily via gradual HW refresh process) to the new data center from the existing B515 data center to happen in FY21-23

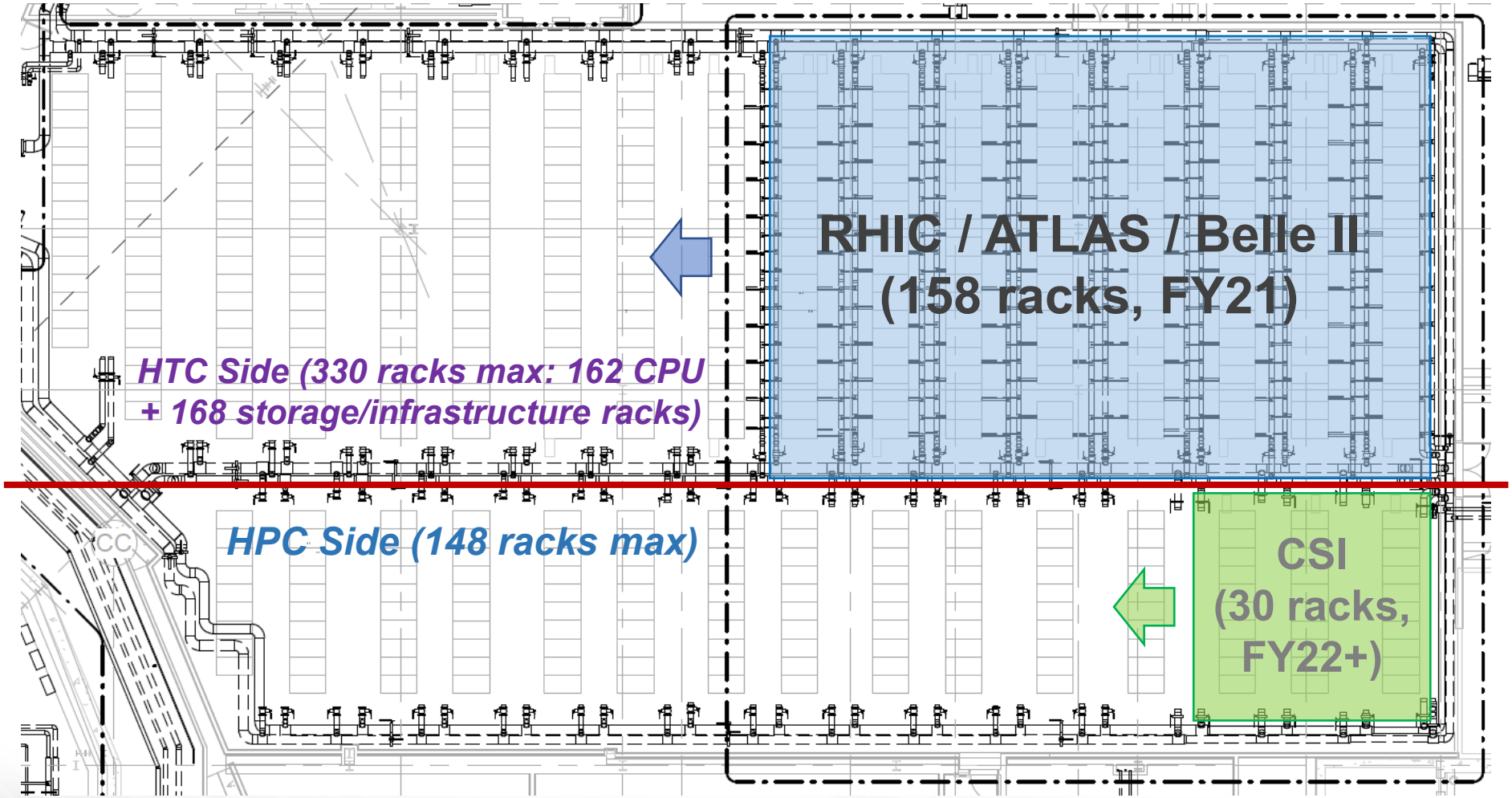



Access tunnel to B535C
Central yard
Covered overpass
Future B725 loading dock
to B515
Currently occupied areas (offices, 2nd floor)
Entrance to the 10t overpass bridge
Main lobby entrance

BNL Core Facility Revitalization (CFR) Project: Design & Construction of B725 Data Center

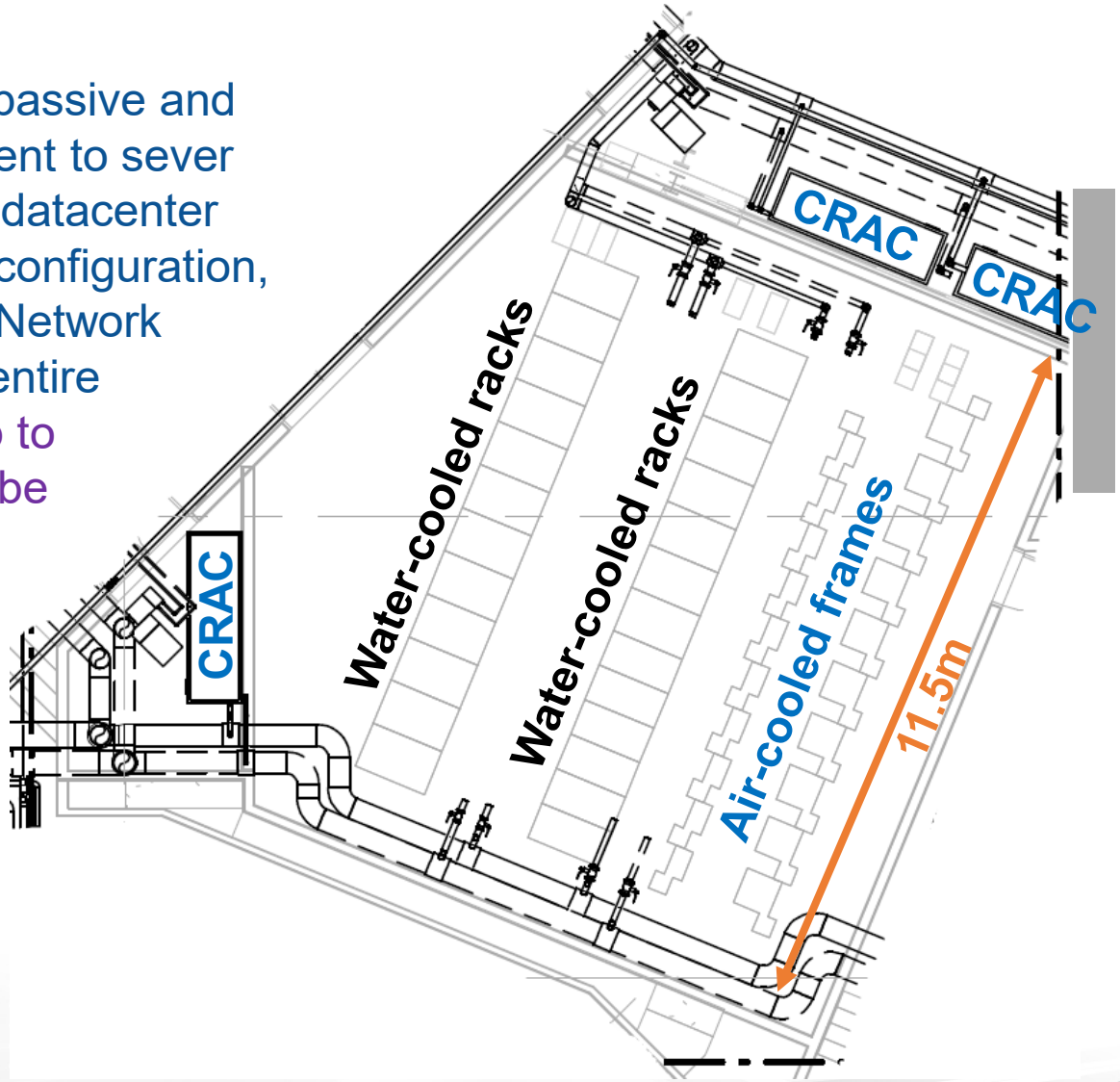
- Main design features of the SDCC B725 data center:
  - A single large data hall for CPU and DISK resources **(Main Data Hall, MDH)** divided into two aisles:
    - HTC (RHIC/ATLAS/Belle II) aisle: 16 rows of ~20 racks each + one row of 16 racks
    - HPC (CSI): 14 rows of 10 racks each plus one row of 8 racks)
    - 478 rack positions in total with up to 9.6 MW of IT load combined in a fully built out configuration (2030 timeframe)
  - 188 rack positions to become available starting from 2021Q3
    - 158 rack positions for RHIC/ATLAS/Belle II with 2.4 MW of power/cooling available
    - 30 rack positions for CSI with 900 kW of power/cooling available
    - 2.4 MW of diesel generator backup power (IT load) available
    - Unlocking 290 remaining rack positions will require construction of additional electrical rooms, installation of additional power distribution and UPS equipment, chillers, cooling towers and diesel generators.
  - Dry-pipe/pre-action double interlock sprinkler system for fire suppression in the MDH
  - Inergen based fire suppression in the Tape Room and the Network Room
  - Standard APC 42U racks (600mm wide, 1070 mm deep (HTC compute) or 1200 mm deep (high capacity JBOD storage)) are to be used across the entire floor of B725 MDH:
    - All equipped with watercooled rear-door heat exchangers with chilled water supplier from under the raised floor
    - Isolation valves on the row-level and individual rack level on the water pipes
    - Zoned drainage system in the concrete floor
    - Nothing but water is distributed under the raised floor in the B725 MDH
  - 3 level of overhead cable/power distribution: power/busbar, fiber tray with mini-racks attached, copper tray with RJ-45 patch panels attached
    - No patch panels to be placed inside the racks with equipment

B725 Data Center: Main Data Hall



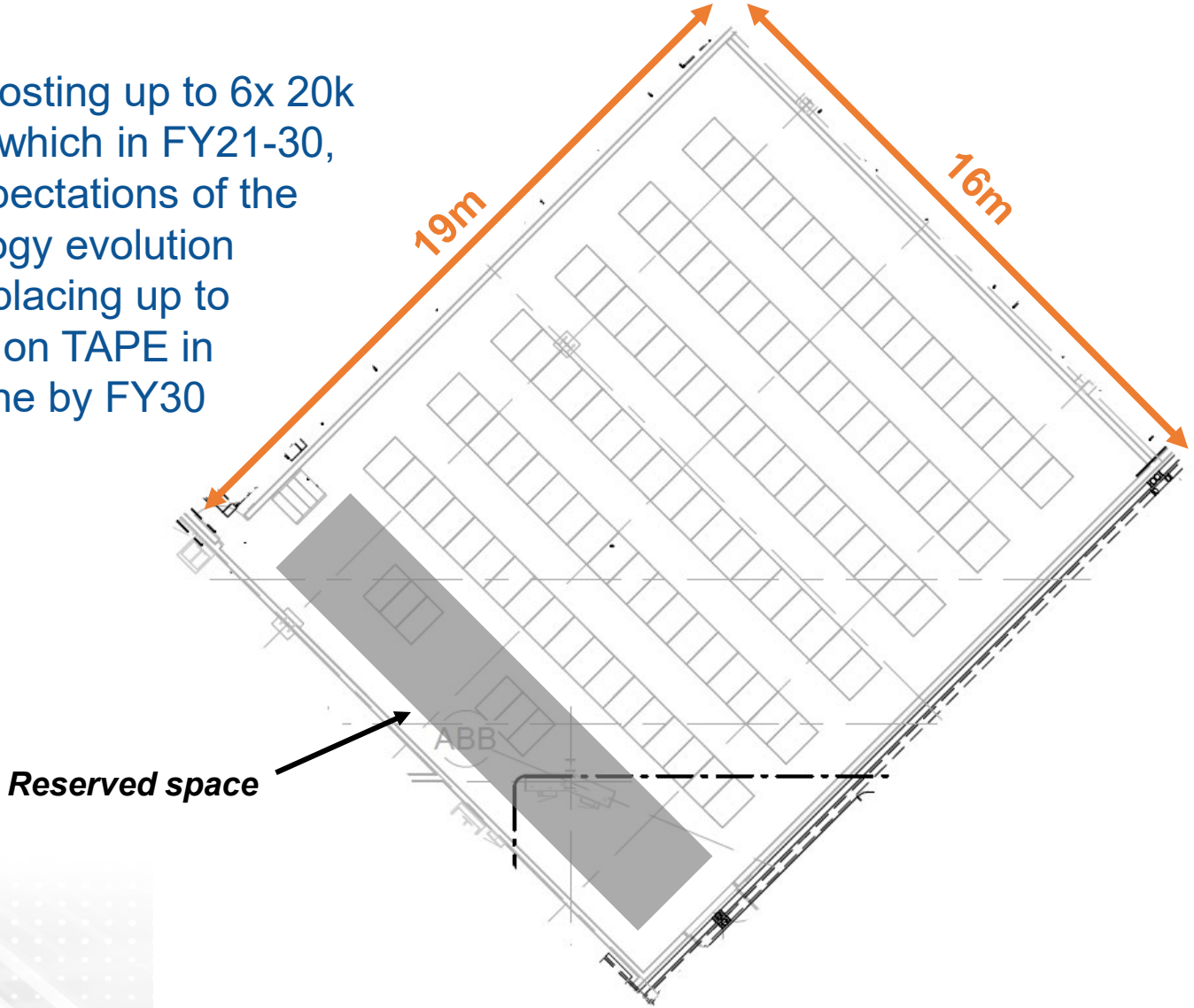
B725 Data Center: Network Room

Capable of hosting all passive and active network equipment to sever the fully built out B725 datacenter in 9.6 MW / 478 racks configuration, including ESnet / BNL Network Perimeter serving the entire BNL site if required (up to 0.5 MW of IT load can be deployed in this area)



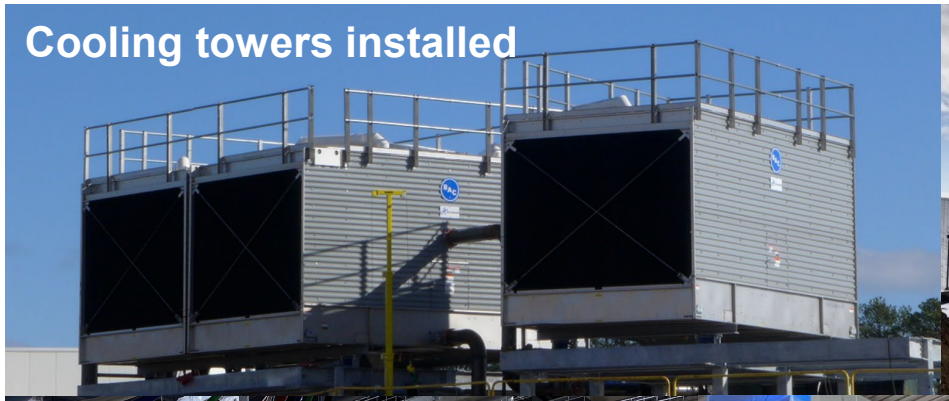
B725 Data Center: Tape Room

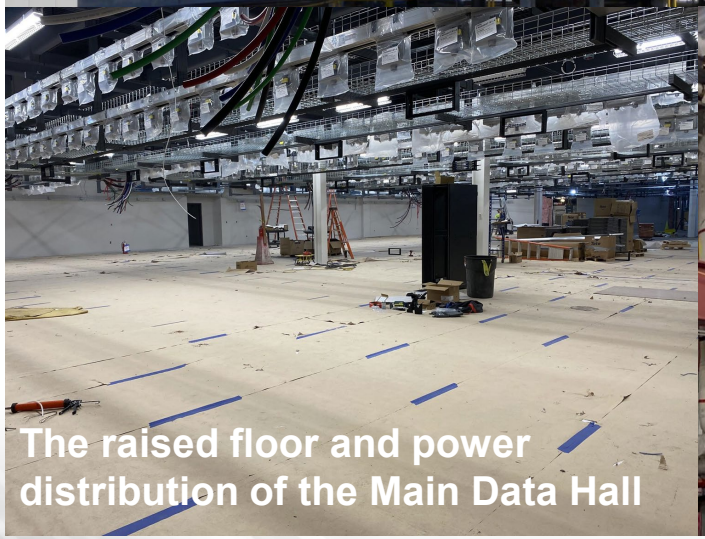
Capable of hosting up to 6x 20k slot libraries which in FY21-30, given the expectations of the LTO technology evolution could mean placing up to 3 EB of data on TAPE in this area alone by FY30

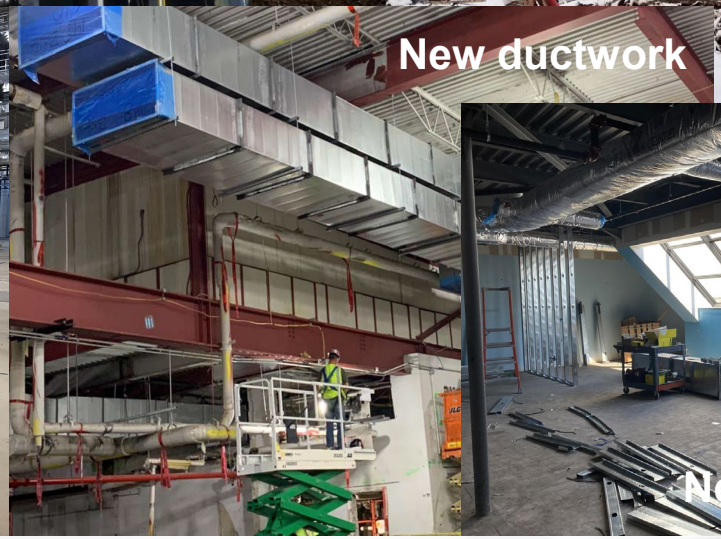


Status of B725 Datacenter Construction

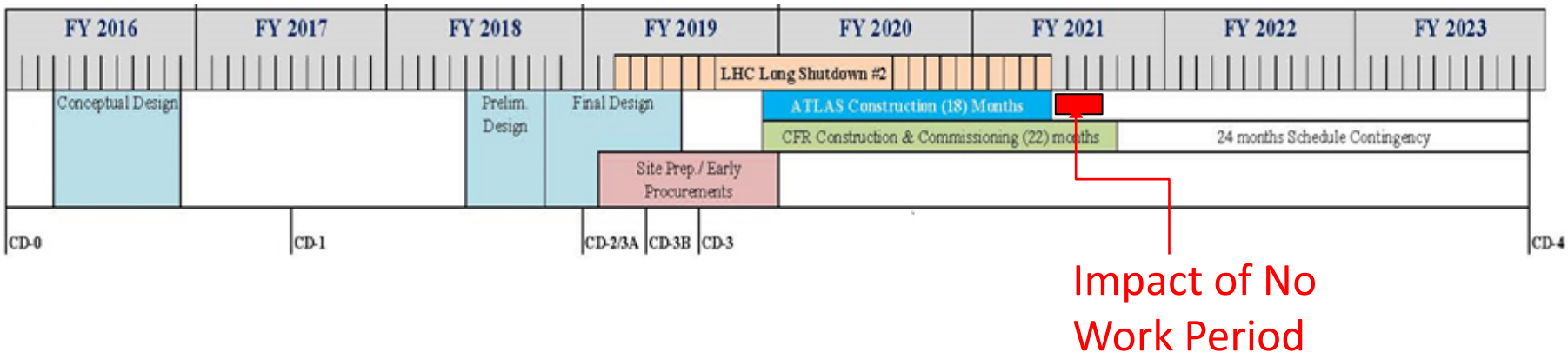
July 2020 – Mar 2021: construction is going ahead after 3 months of delay in 2020Q2 due to COVID-19. The early occupancy of B725 datacenter is expected to begin for ATLAS in June 2021 (network equipment deployment is expected to start in Aprl-May 2021; occupancy for all tenants is expected to start in July 2022)







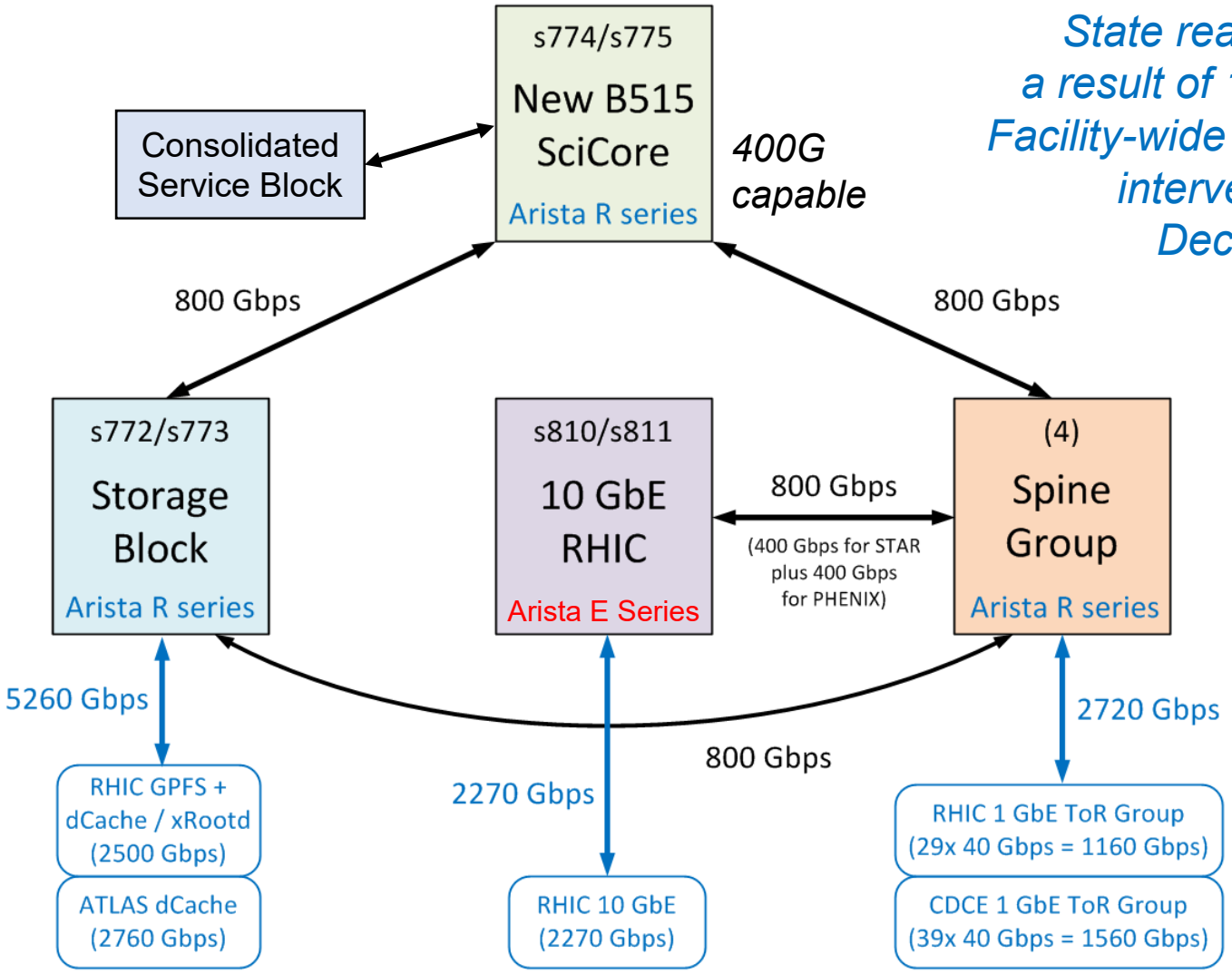
CFR Project and Impact of COVID-19



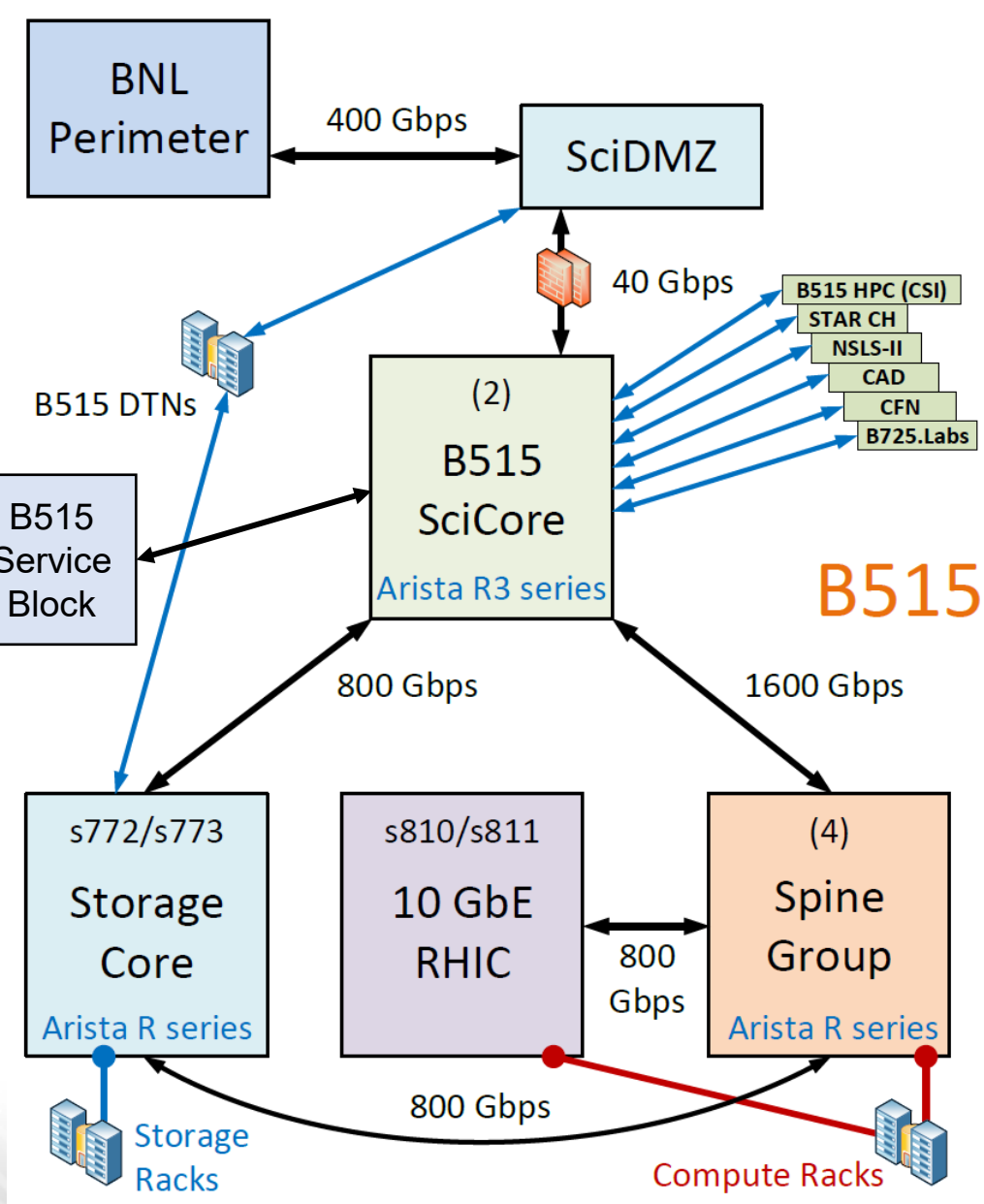
COVID-19 Schedule Impacts:

- 3 Months “No work” period (April, May, June 2020)
- Delays realized for equipment procurements (6-8 weeks)
- Delays realized for de-mobilization and re-mobilization as well as phased re-start of construction activity
- The early occupancy shifted from 2021Q1 to 2021Q3 as a result

B515 Data Center: Network Consolidation

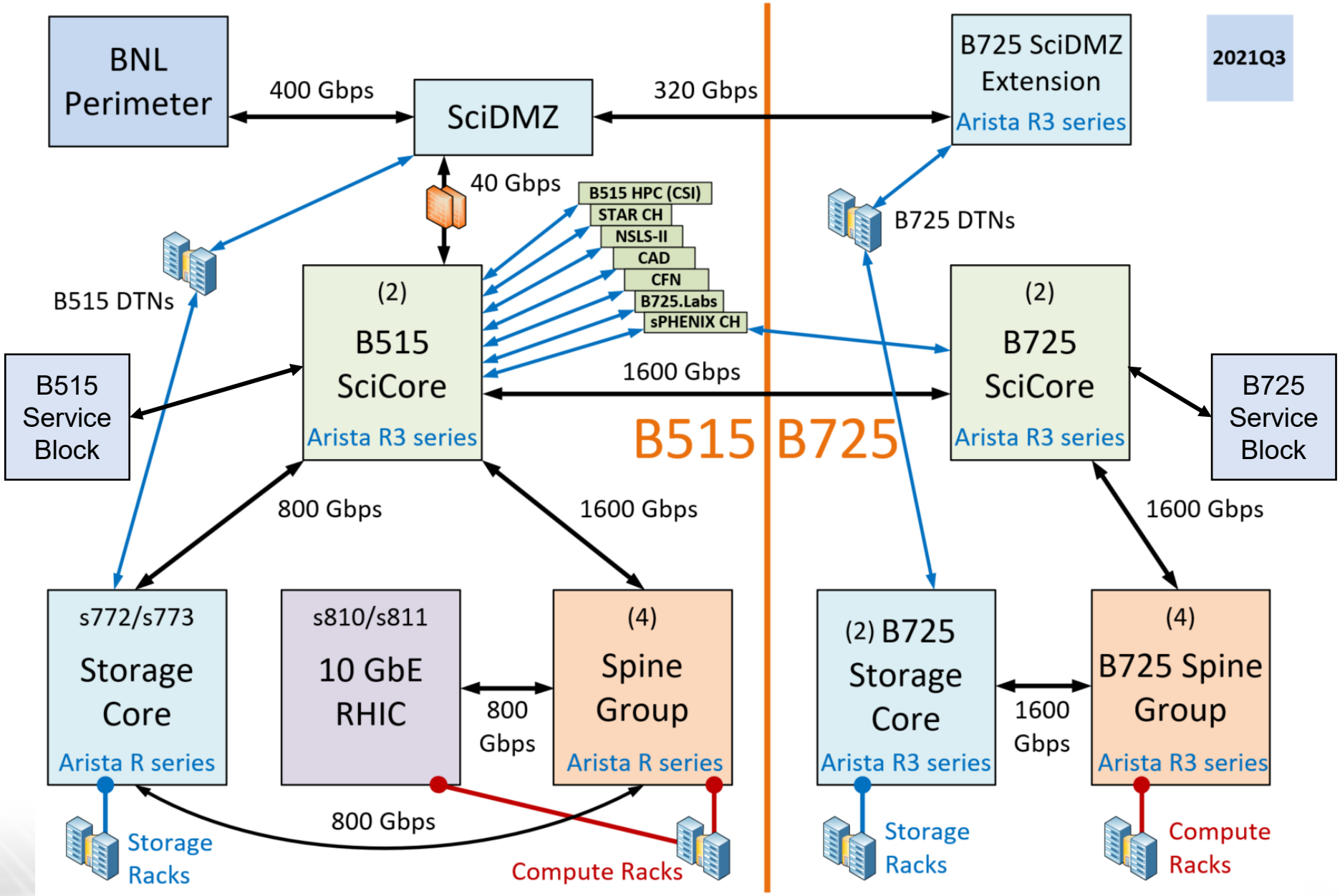


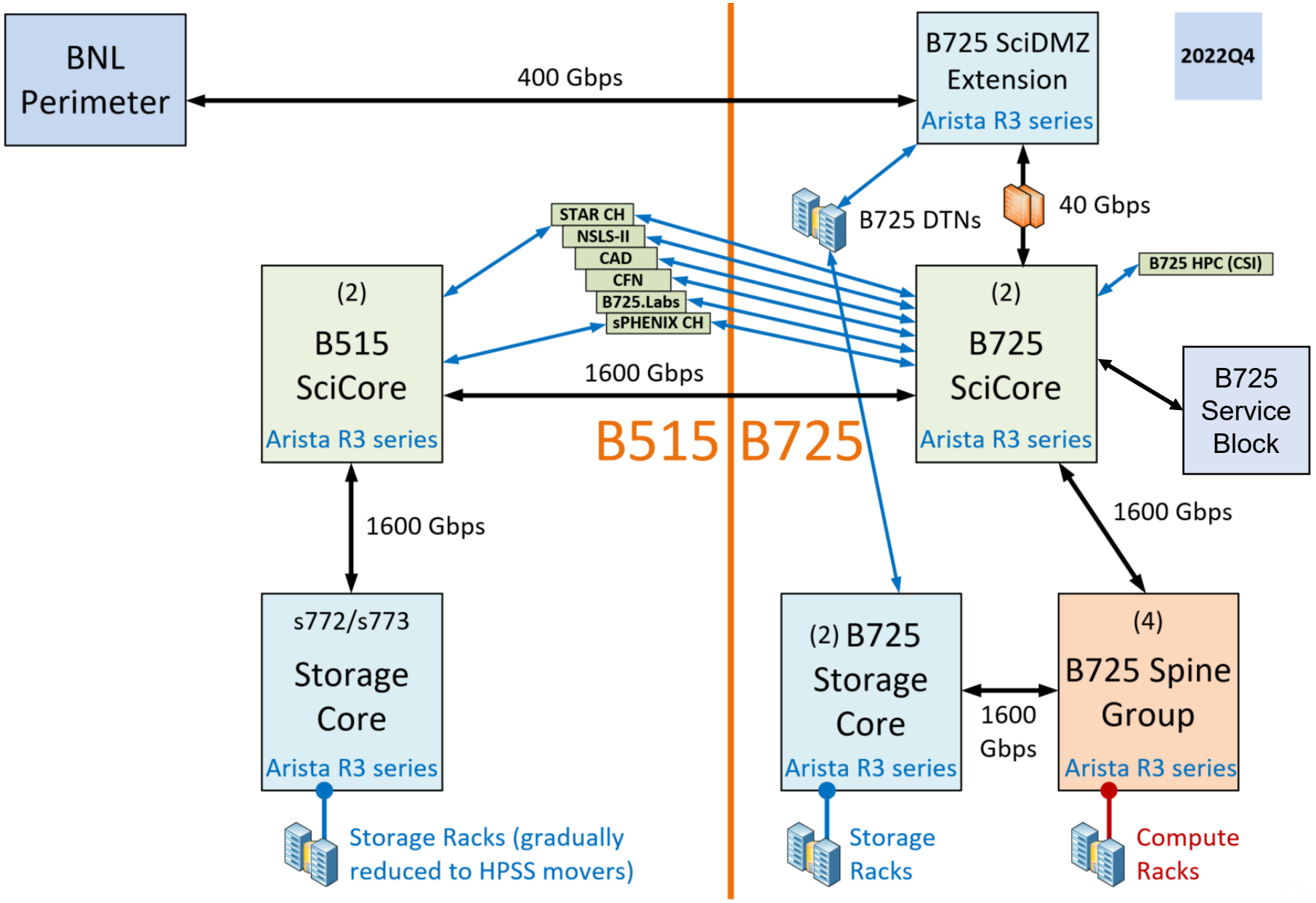
State reached as a result of 12h long Facility-wide network intervention of Dec 1, 2020



2020Q4

The expansion of SDCC central network into B725 is to expected start in 2021Q2 in order to have them ready for deploying IT payload on the floor of B725 in 2021Q3


BNL Perimeter
400 Gbps
SciDMZ
320 Gbps
B725 SciDMZ Extension
Arista R3 series
2021Q3
40 Gbps
B515 HPC (CSI)
STAR CH
NSLS-II
CAD
CFN
B725.Labs
sPHENIX CH
B725 DTNs
B515 DTNs
(2)
B515 SciCore
Arista R3 series
(2)
B725 SciCore
Arista R3 series
1600 Gbps
B515 Service Block
B725 Service Block
B515
B725
800 Gbps
1600 Gbps
1600 Gbps
s772/s773
Storage Core
Arista R series
s810/s811
10 GbE RHIC
800 Gbps
(4)
Spine Group
Arista R series
(2) B725 Storage Core
Arista R3 series
1600 Gbps
(4)
B725 Spine Group
Arista R3 series
800 Gbps
Storage Racks
Compute Racks
Storage Racks
Compute Racks


BNL Perimeter
400 Gbps
B725 SciDMZ Extension
Arista R3 series
2022Q4
B725 DTNs
40 Gbps
STAR CH
NSLS-II
CAD
CFN
B725.Labs
sPHENIX CH
(2) B515 SciCore
Arista R3 series
1600 Gbps
B515 B725
(2) B725 SciCore
Arista R3 series
B725 HPC (CSI)
B725 Service Block
1600 Gbps
1600 Gbps
s772/s773 Storage Core
Arista R3 series
(2) B725 Storage Core
Arista R3 series
1600 Gbps
(4) B725 Spine Group
Arista R3 series
Storage Racks (gradually reduced to HPSS movers)
Storage Racks
Compute Racks

B515 / B725 Data Center Transition in FY21-23

- Compute resources (HTC):
  - All of the 1 GbE connected Farm racks in B515 are ToR switch based as of now.
  - All new HTC compute node racks are placed in CDCE since FY18.
  - The compute node racks that still have at least 2 year of vendor supported life as of FY21 are to be moved to B515 and connected to B725 Spine Group in Jul-Sep 2021 timeframe (18 such racks exist in CDCE serving ATLAS, RHIC and Belle II).
  - The configuration of the 10 GbE connected part of the RHIC CPU Farm is frozen since FY18 and not going to be extended anymore. This part of the RHIC Farm is to be retired as it goes off support in B515 along with its Arista 7508E switch pair serving it and replaced by the new ToR switch and 2x 1 GbE / 10 GbE attached compute nodes deployed in B725 in FY22-23.
  - 1 GbE connected parts of RHIC and ATLAS CPU Farms are to remain in B515 and to be gradually replaced with new compute node racks deployed in B725 in FY22-23 period.
  - No compute node racks are expected to be hosted in B515 after FY23.
- Compute resources (HPC)
  - CSI HPC clusters (Institutional Cluster (IC), KNL, Skylake/LQCD) located in the BGL room and the legacy HPC clusters located in the QCDOC room are expected to be retired by the end of FY23 and replaced with the new Institutional Cluster to be deployed in the CSI segment of the B725 MDH floor in FY22 *– more information on the subject is in the "New institutional resources at BNL" presentation delivered to this HEPiX on Mar 16, 2021.*
  - CSI Machine Learning (ML) cluster deployed in the QCDOC area can either be retired and replaced by a new system in B725 or moved to B725 MDH (CSI segment) by the end of FY23.
  - The NSLS II HPC racks recently deployed in QCDOC area are to be migrated to B725 MDH (CSI segment) before the end of FY23.

B515 / B725 Data Center Transition in FY21-23 (Cont.)

- Storage (dCache/xRootD/Ceph):
  - The bulk purchase of RHIC and ATLAS dCache storage equipment of FY21 is going to target B725 MDH as the deployment area, as well all purchased of ATLAS, Belle II and RHIC dCache storage henceforth. The first storage racks are expected to be operational on the floor of B725 MDH in 2021Q3.
  - No physical migration of storage racks between B515 and B725 at all, as gradual replacement of storage backends by new equipment deployed directly in B725 is much safer and less disruptive way to migrate DISK resource capacity.
  - This gradual replacement is expected to be completed by the end of FY23, at which time the storage backends of all dCache instances deployed in the Facility are expected to be based on JBODs.
- Storage (GPFS/Lustre/Central NFS)
  - No physical migration of Lustre/GPFS storage racks between B515 and B725 as well
  - New RHIC and ATLAS related Lustre installations are expected to be deployed in B725 in 2021Q3.
  - The ATLAS Tier-3 GPFS system currently deployed in CDCE is expected to be replaced by a new Lustre installation in B725 MDH in FY22
  - New central NFS appliance is to be deployed in B725 in FY23, and the existing one in CDCE is to be retired shortly after that.
  - The existing RHIC GPFS systems located in B515 are expected to be replaced by (likely) Lustre installations in B725 MDH in the FY24-25 timeframe.

B515 / B725 Data Center Transition in FY21-23 (Cont.)

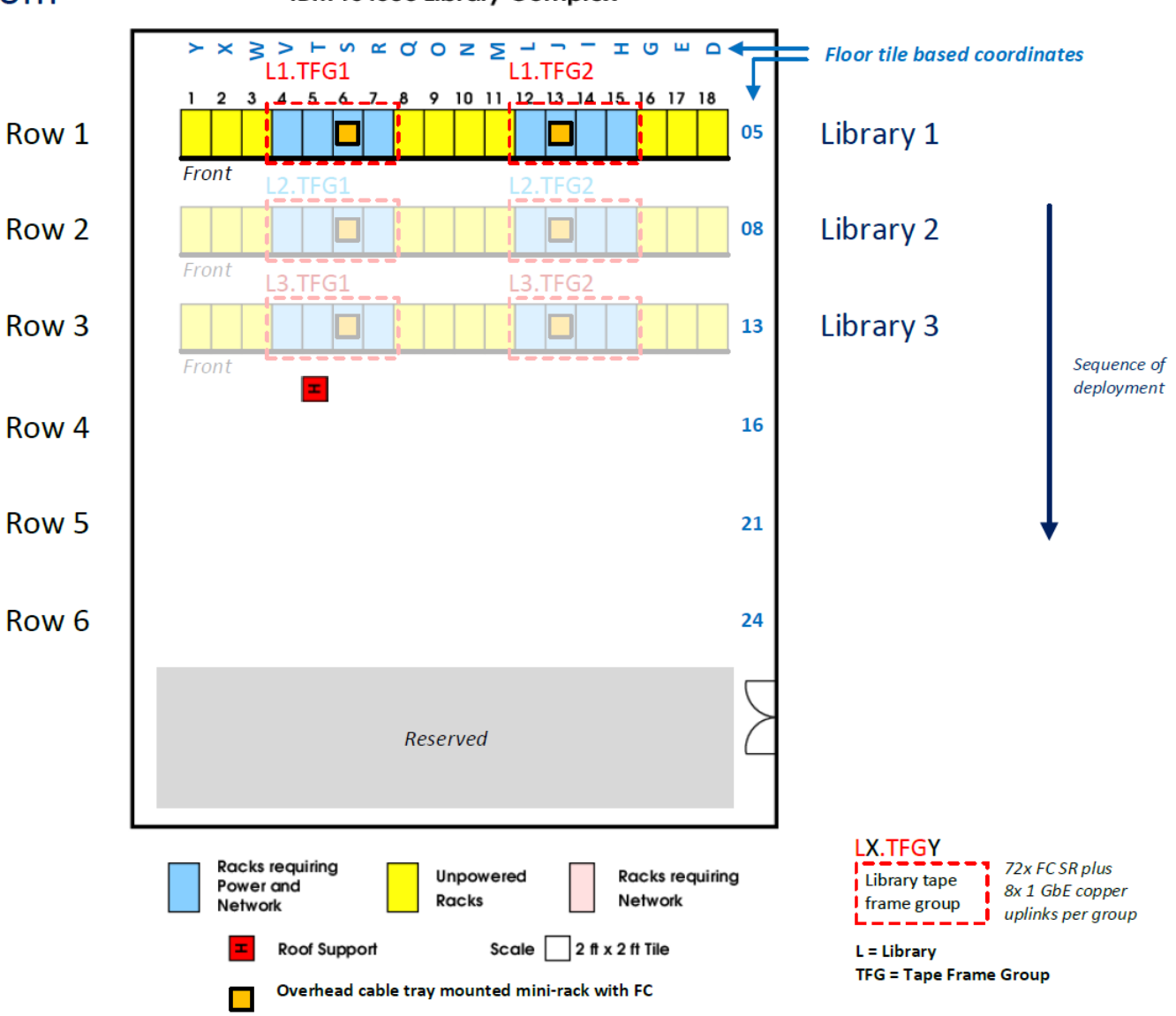
- Virtualization platforms (RHEV & Farm VMMS system):
  - The SDCC RHEV instance is to be replaced with the new installation in B725 in FY21.
  - Farm VMMS system and other Farm infrastructure components currently deployed in B515 infrastructure racks are expected to be replaced by the new servers deployed in the infrastructure racks on the floor of B725 MDH by FY24.
- HPSS
  - The first B725 ATLAS tape library (IBM TS4500 with ~20k tape slots) and the first HPSS mover rack are expected to be deployed in B725 in 2021Q3.
  - The HPSS Core Servers are to be moved to B725 MDH infrastructure racks in Jul-Aug 2021.
  - The exiting 9x Oracle SL8500 tape silos deployed om SDCC Facility are to be consolidated to CDCE on the B515 side by the end of FY21 (schedule affected by COVID-19 countermeasures since this on-site work can only be done by a specific group of contractors).
  - The consolidation of B515 HPSS complex to B515/CDCE area is expected to be completed by the end of FY21, to be followed by the retirement of all legacy FC infrastructure in main BCF.
  - The HPSS installation in B725 is expected to grow up to 6 libraries (up to 576 tape drives and 120k tape slots) and 9 mover racks on the floor of MDH by FY30 if necessary.
  - More tape libraries can be added to B515/CDCE area (both Oracle SL8500 and IBM TS4500) past FY23, if necessary.
- Infrastructure racks
  - 4 infrastructure racks distributed across PS #2 area are to be available in B725 starting from 2021Q3 – to be increased to 8 infrastructure racks distributed across PS #2 and PS #1 areas in early FY22.
  - All B515 infrastructure racks are expected to be retired by FY25.

B725 Tape Room

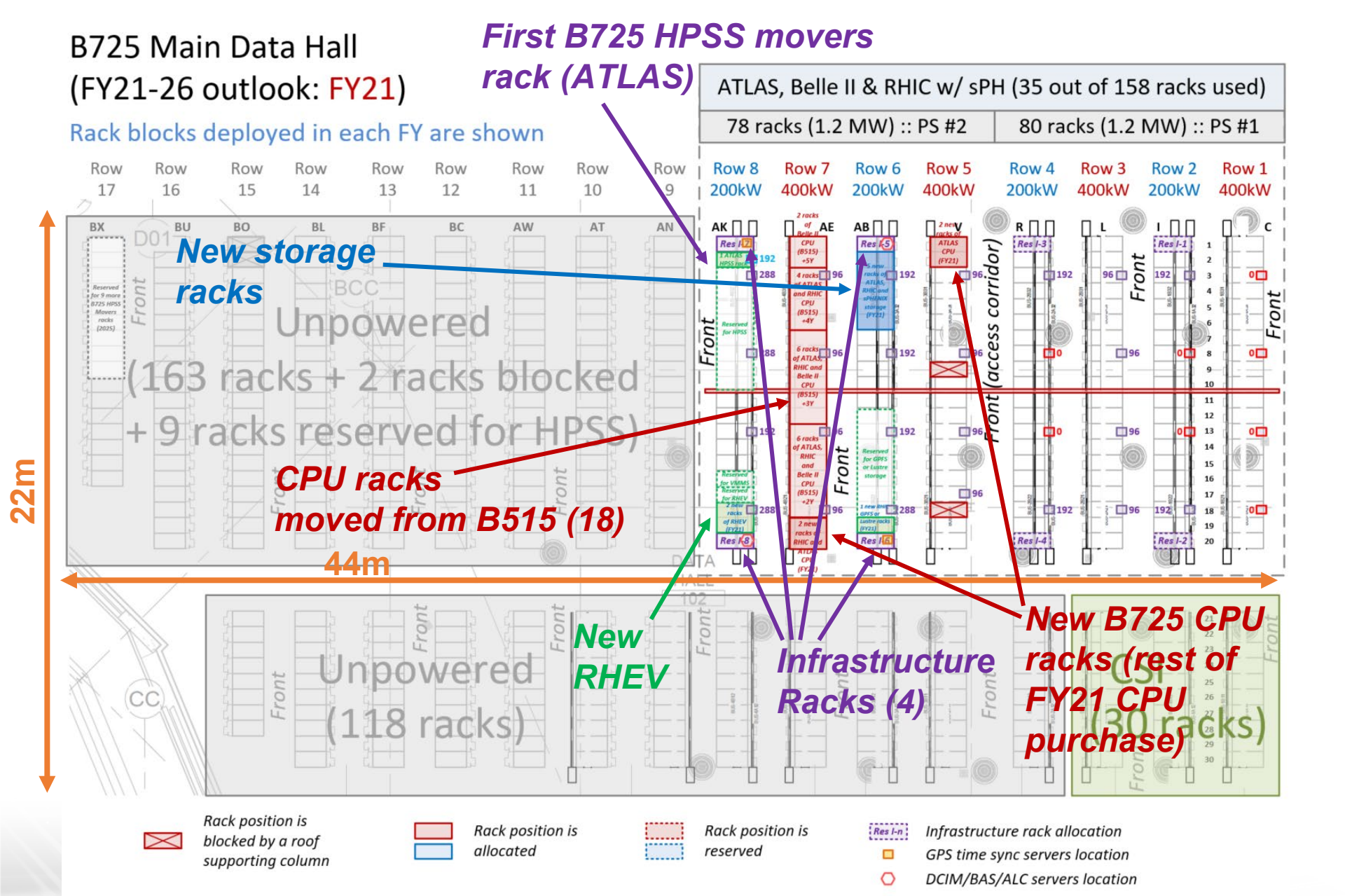
ATLAS (FY21)

sPHENIX (FY22)

sPHENIX (FY24)

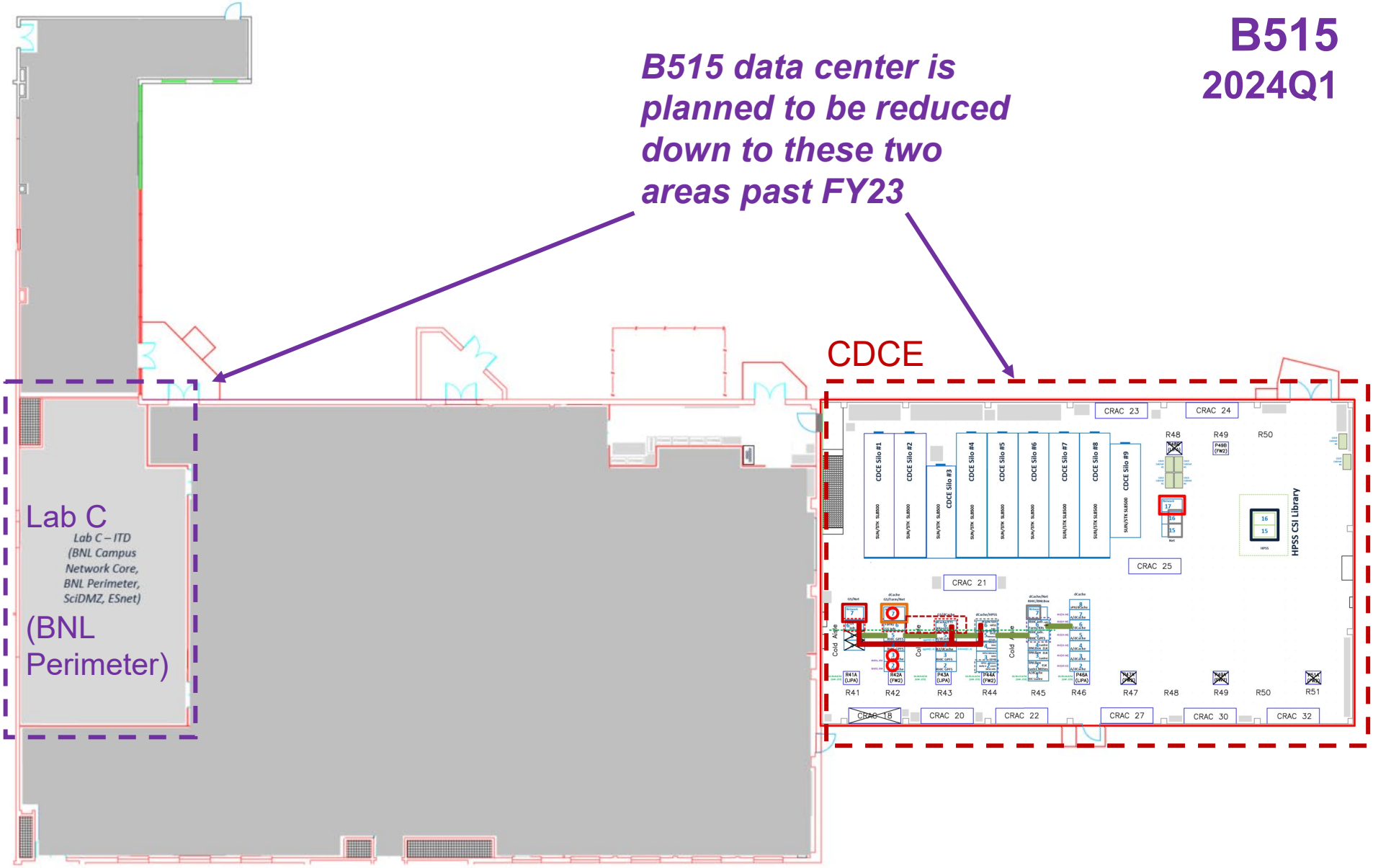


STEP 4
FY24

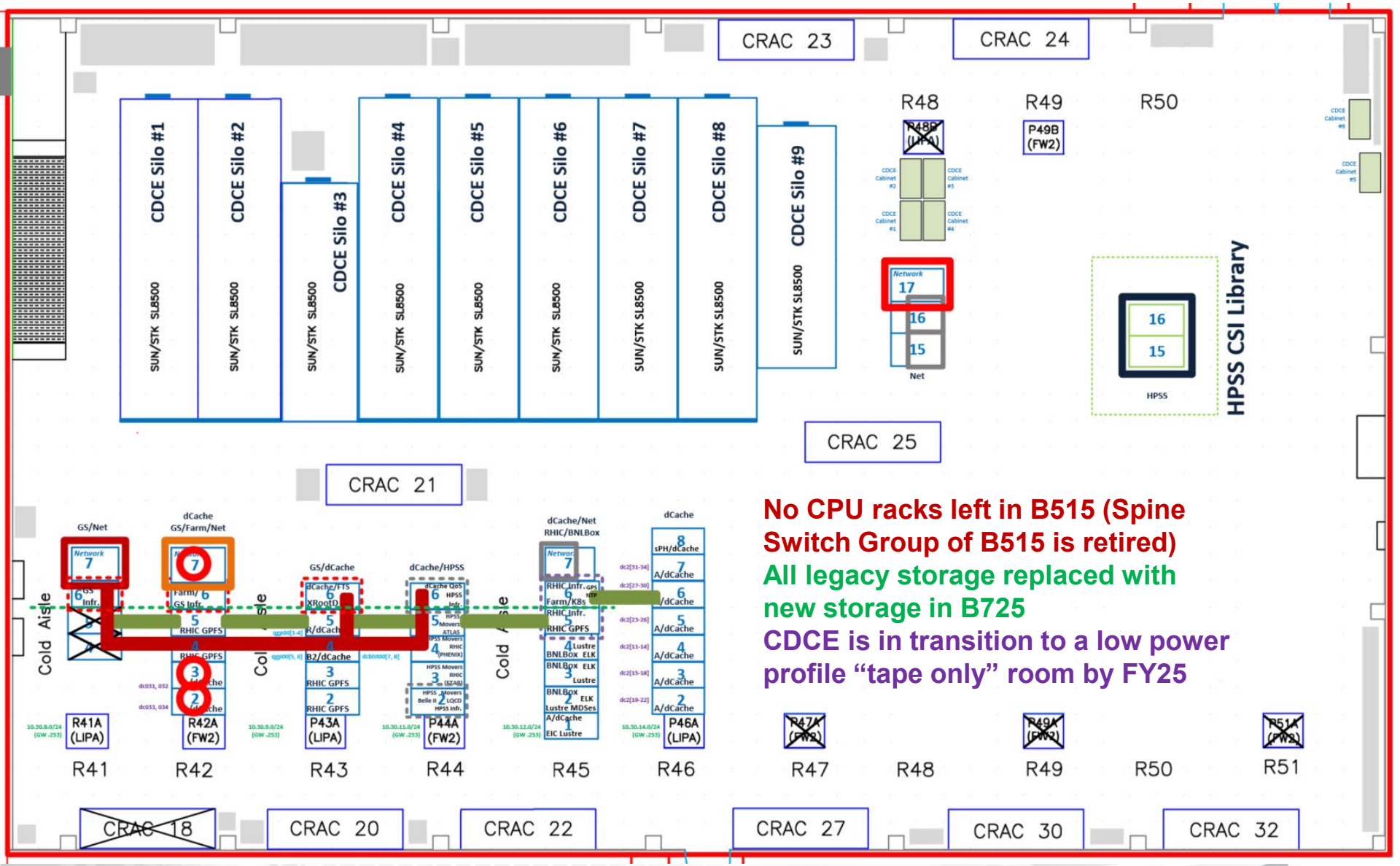

B725 Main Data Hall
(FY21-26 outlook: FY21)
Rack blocks deployed in each FY are shown
First B725 HPSS movers rack (ATLAS)
ATLAS, Belle II & RHIC w/ sPH (35 out of 158 racks used)
78 racks (1.2 MW) :: PS #2
80 racks (1.2 MW) :: PS #1
Row 17
Row 16
Row 15
Row 14
Row 13
Row 12
Row 11
Row 10
Row 9
Row 8 200kW
Row 7 400kW
Row 6 200kW
Row 5 400kW
Row 4 200kW
Row 3 400kW
Row 2 200kW
Row 1 400kW
New storage racks
Unpowered
(163 racks + 2 racks blocked
+ 9 racks reserved for HPSS)
CPU racks moved from B515 (18)
22m
44m
Front (access corridor)
Unpowered
(118 racks)
New RHEV
Infrastructure Racks (4)
New B725 CPU racks (rest of FY21 CPU purchase)
CSI
(30 racks)
Rack position is blocked by a roof supporting column
Rack position is allocated
Rack position is reserved
Infrastructure rack allocation
GPS time sync servers location
DCIM/BAS/ALC servers location

B515
2024Q1

B515 data center is planned to be reduced down to these two areas past FY23



Expected SDCC Datacenter Layout by the end of FY23 on B515 side

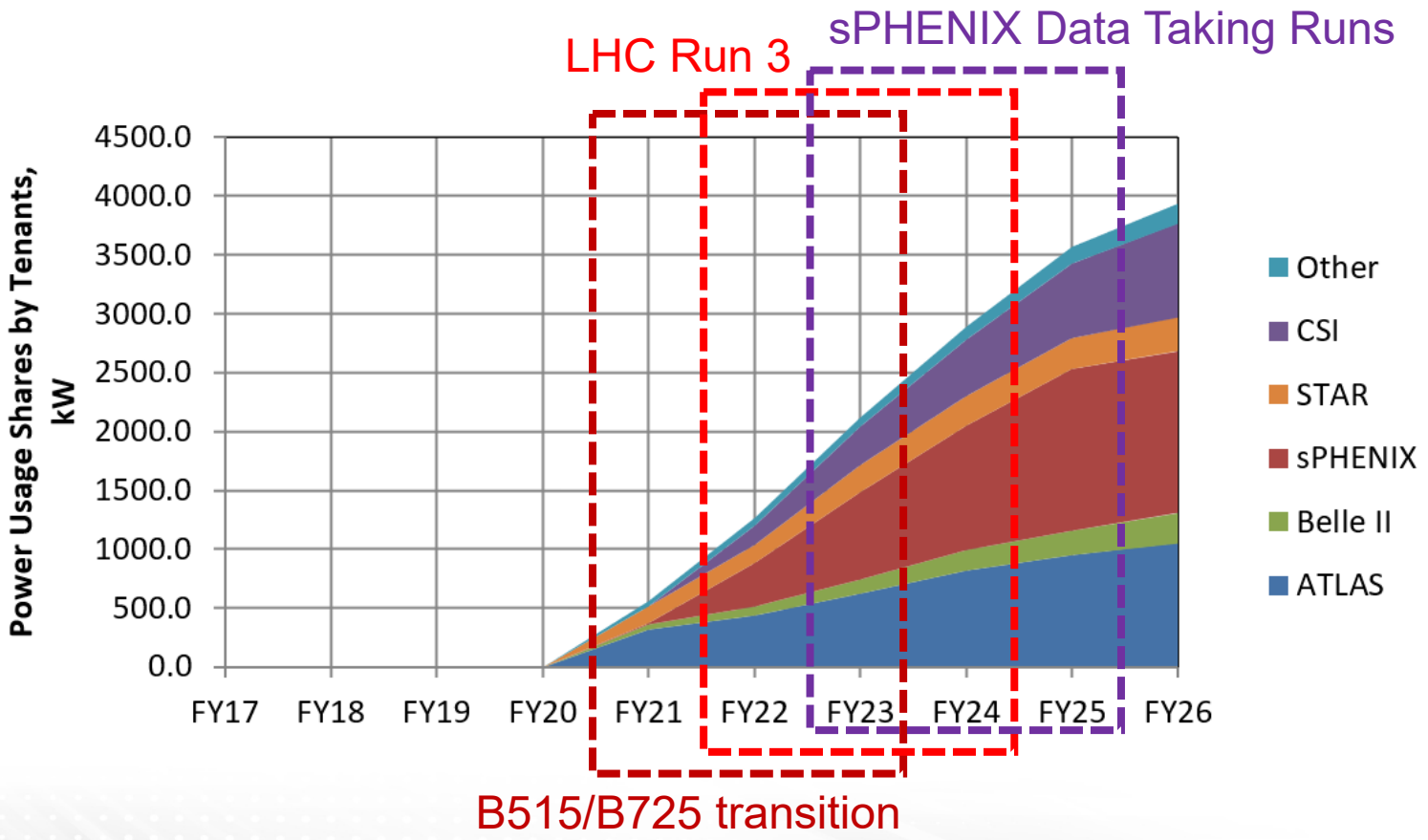


No CPU racks left in B515 (Spine Switch Group of B515 is retired)

All legacy storage replaced with new storage in B725

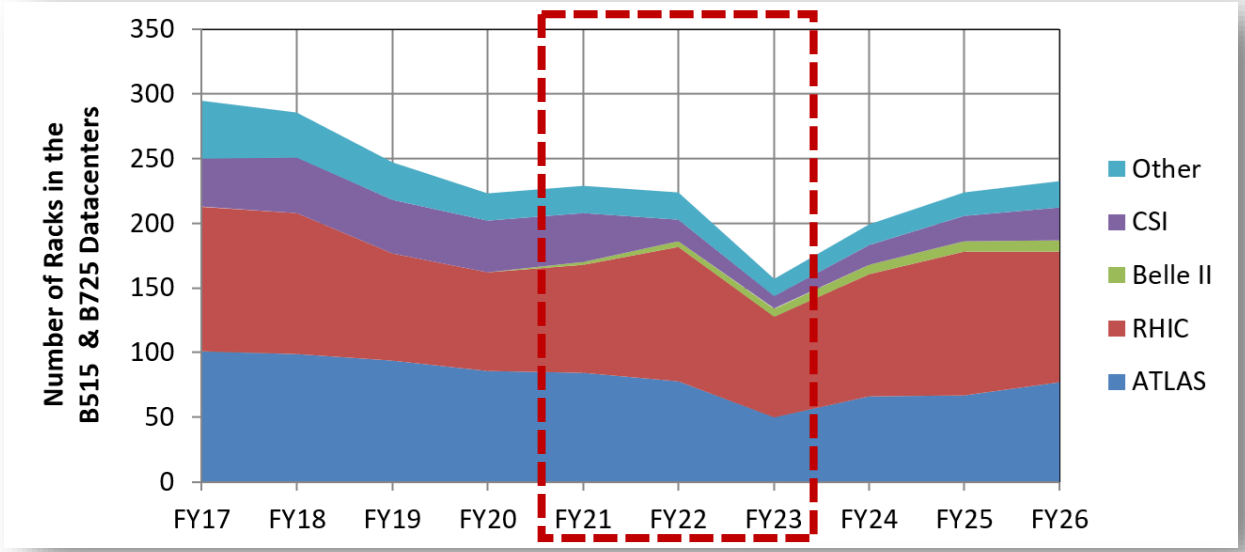
CDCE is in transition to a low power profile "tape only" room by FY25

Expected B725 Data Center Scale-Up in FY21-26 (IT Power Profile, kW)

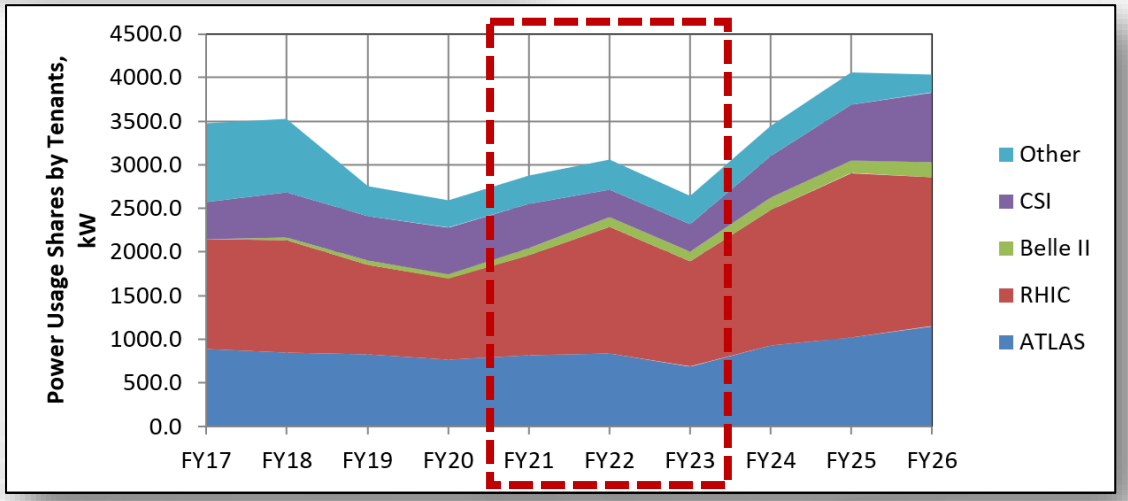


B515 / B725: Floor Occupancy Projection

B515/B725 transition



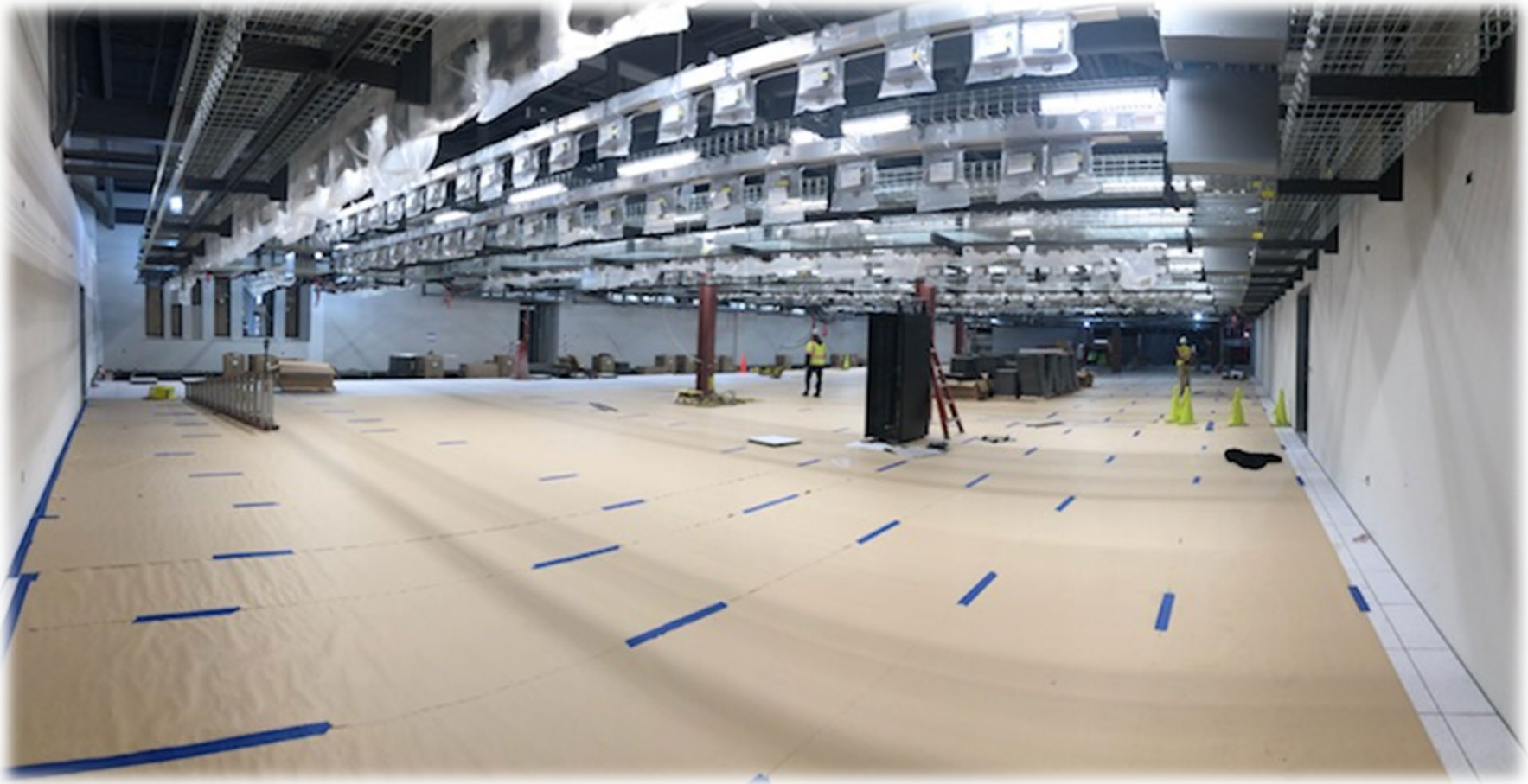
B515/B725 transition



Summary

- The CFR project has finished the design phase in the first half of 2019 and then entered the construction phase in the second half of 2019 which is currently projected to be finished in May-June 2021 timeframe.
- The existing B515 data center is already on track for consolidation into the CDCE and Lab C areas since FY18
- The occupancy of the B725 data center ATLAS is expected to start in June 2021, and occupancy for all tenants – in July 2021.
- A gradual migration of compute and storage capabilities between the B515 and B725 data centers is expected to occur in FY21-23 period with only a subset of existing CPU racks physically moving between two locations in 2021Q3
- B515 data center reduction to the CDCE and Lab C areas is expected to be completed by the end of FY23 which would mark the end of the transition period

Questions & Comments



https://www.sdcc.bnl.gov