Direct Data Streaming at BNL Present and the Future: Network Centric View

Alexandr ZAYTSEV

alezayt@bnl.gov

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

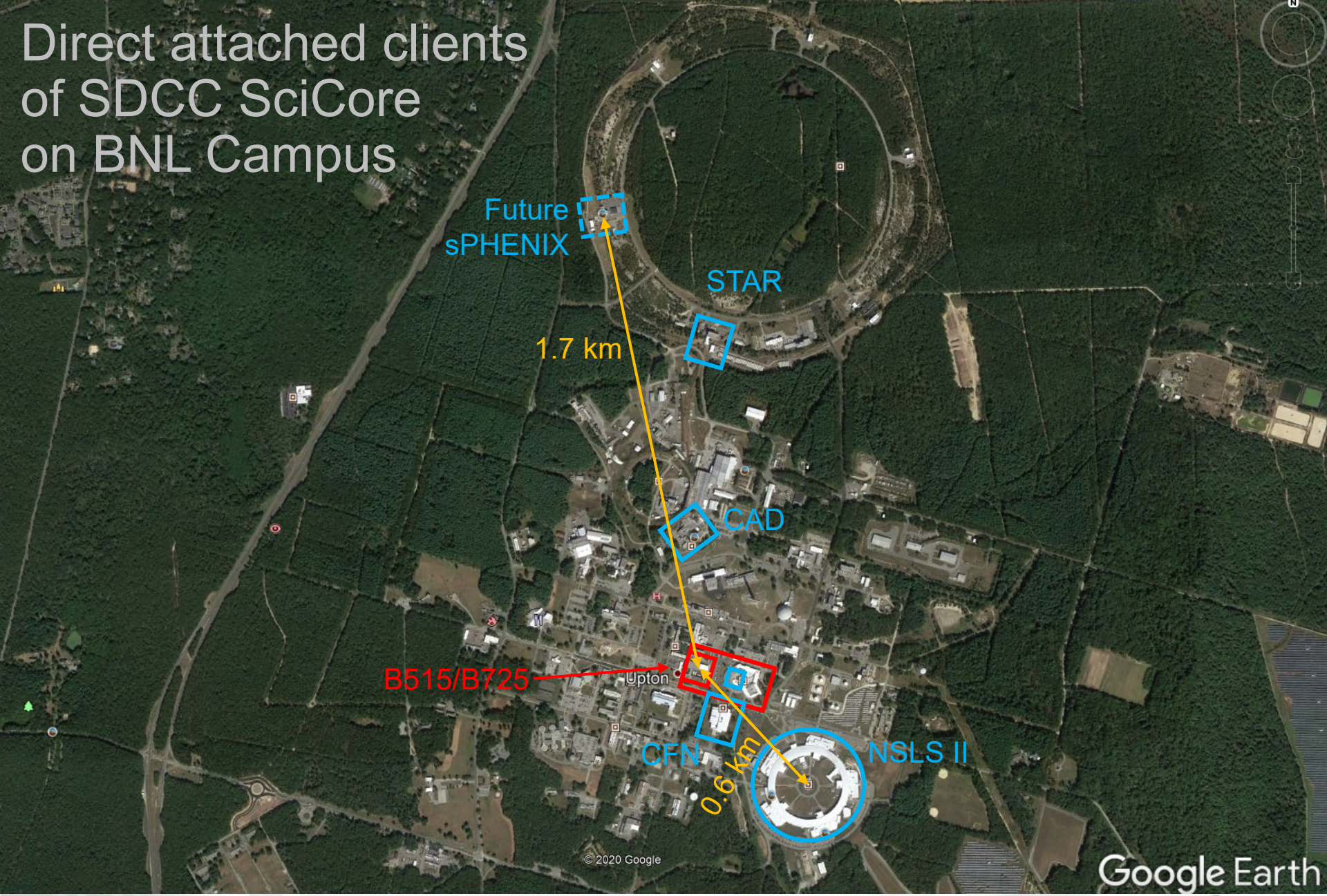
Introduction

- As the cost of hight bandwidth SR uplinks (up to 100m range) and LR uplinks (up to 10km range over the SMF G.652 9/125μ fiber) for Ethernet dropped significantly in 2019-2020, deploying multi-100 Gbps network channels between experimental installations (Lab spaces and Counting Houses and central data storage and data processing Facilities has become a viable option
- At the same time the data throughput capabilities of the top-of-the-line detectors are reaching into the 0.5-1 Tbps range (peak unidirectional bandwidth), e.g. the latest generation of CCD detectors being developed for 4D-STEM microscopy are designed with 480 Gbps capability.
- This drives the need to place a significant amount of IT infrastructure to be placed in the immediate vicinity of the detectors – in the Lab spaces in a traditional approach – to capture, sore and process the data stream, which is not always desirable or even possible due to the power distribution / cooling and airflow / noise and vibration requirements

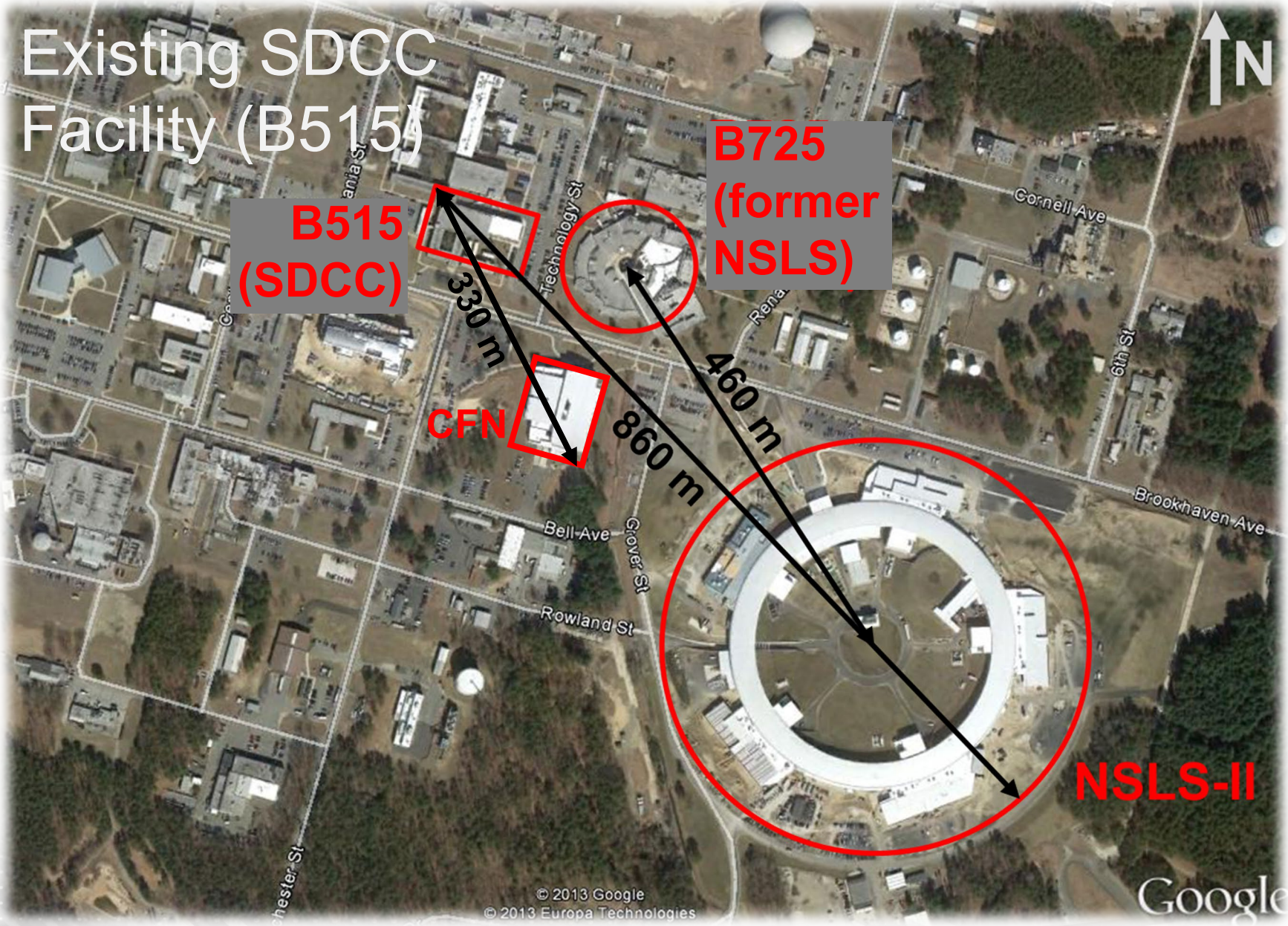
Introduction

- All of these factors combined are driving the development of direct data streaming systems allowing the offload of the experimental data produced by the detector readout systems to the remote IT facilities (datacenters) over the Campus (1-10 km range over the fiber path) and WAN (10 km range and above over the fiber path) networks
  - As a result, the physical boundary between the traditional online (Trigger/DAQ) and offline parts of data processing pipelines for many mid-scale detector experiments
- This trend is visible across many National Labs, and the ESnet has multiple developments in this area in particular:
  - Please refer to the summary report by Inder Monga (ESnet) in Mar 2019: https://www.sc-asia.org/2019/wp-content/uploads/2019/03/1_1140_Inder-Monga.pdf
- Here at BNL we have a long established history of using the on-site SMF fiber infrastructure for bringing the data collected by the DAQ systems of RHIC experiments to the B515 based RACF/SDCC datacenter, and over the last few years a set of directly connected systems was expanded significantly

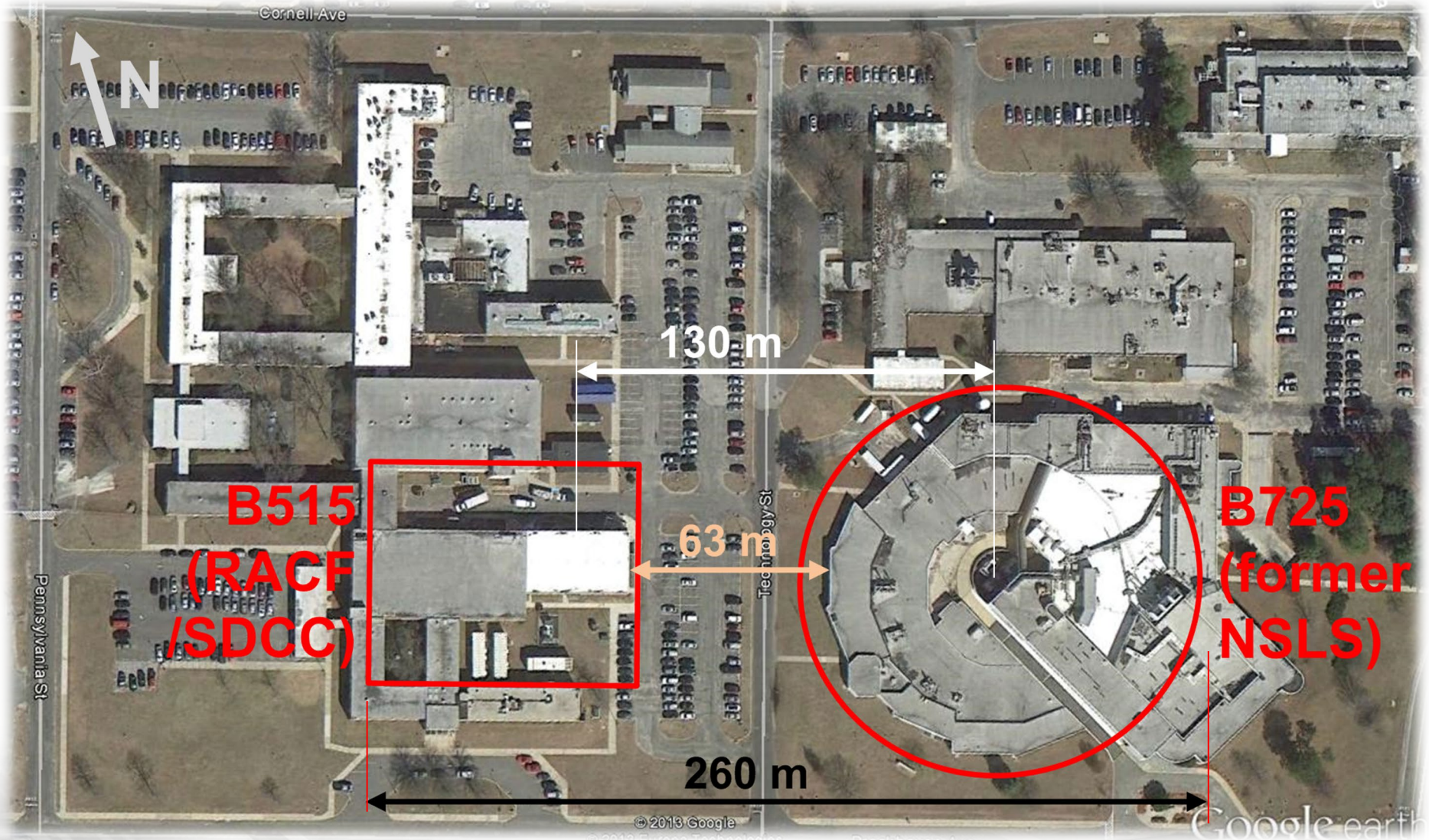
Direct attached clients of SDCC SciCore on BNL Campus



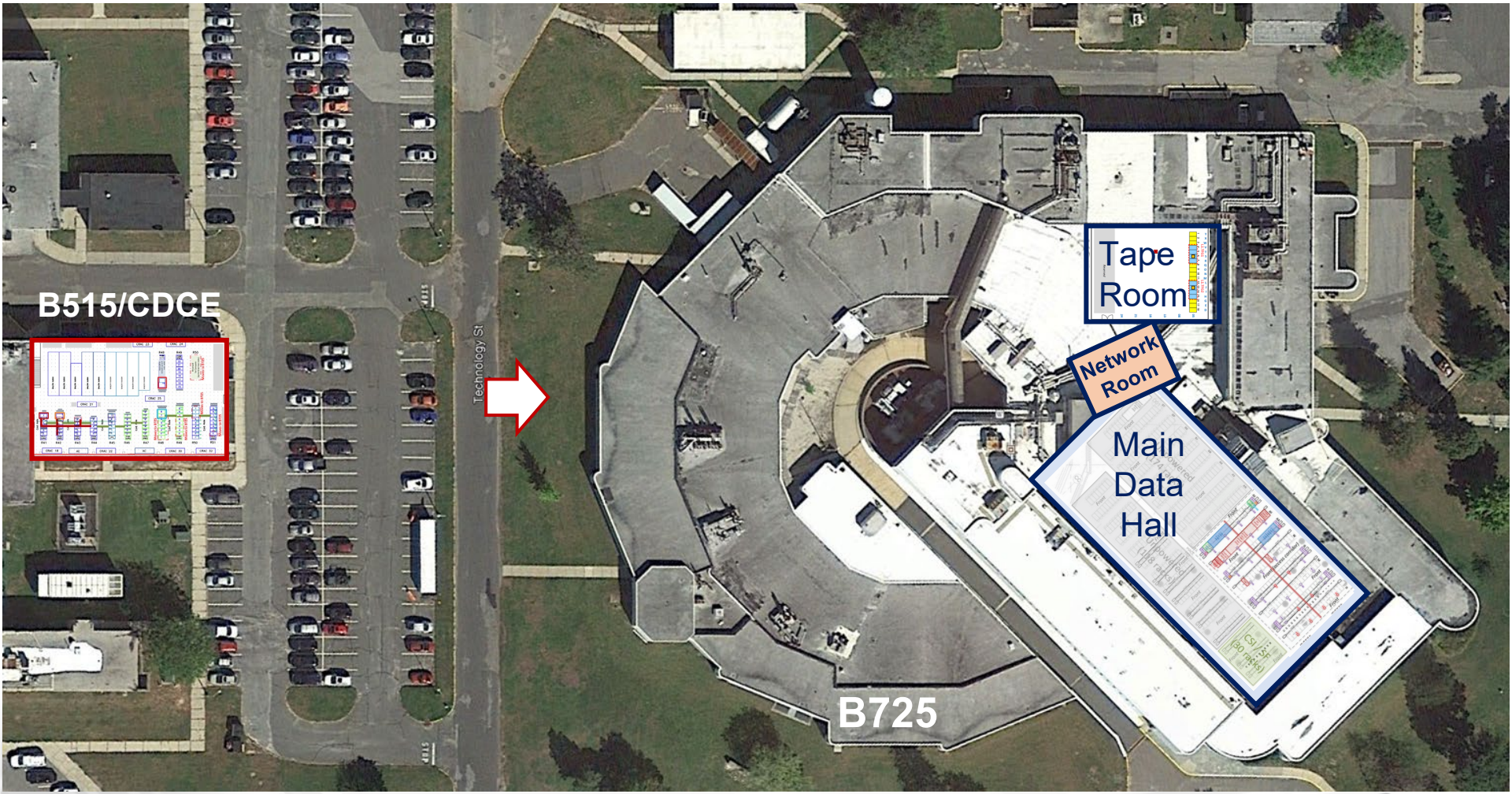
Existing SDCC Facility (B515)



Outlook: B515 & B725

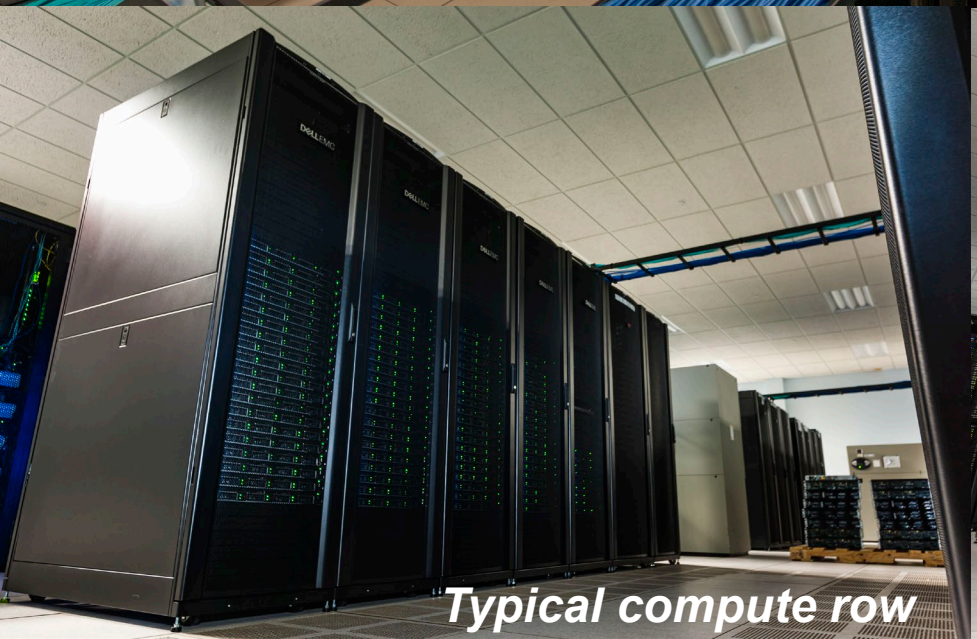


New datacenter is being designed & constructed for the SDCC Facility in B725 in FY19-21 period, with migration of most of the spinning disk storage and all of the compute capability to it from the existing B515 based datacenter to happen in FY21-23




B515/BGL
HPC row

B515/CDCE

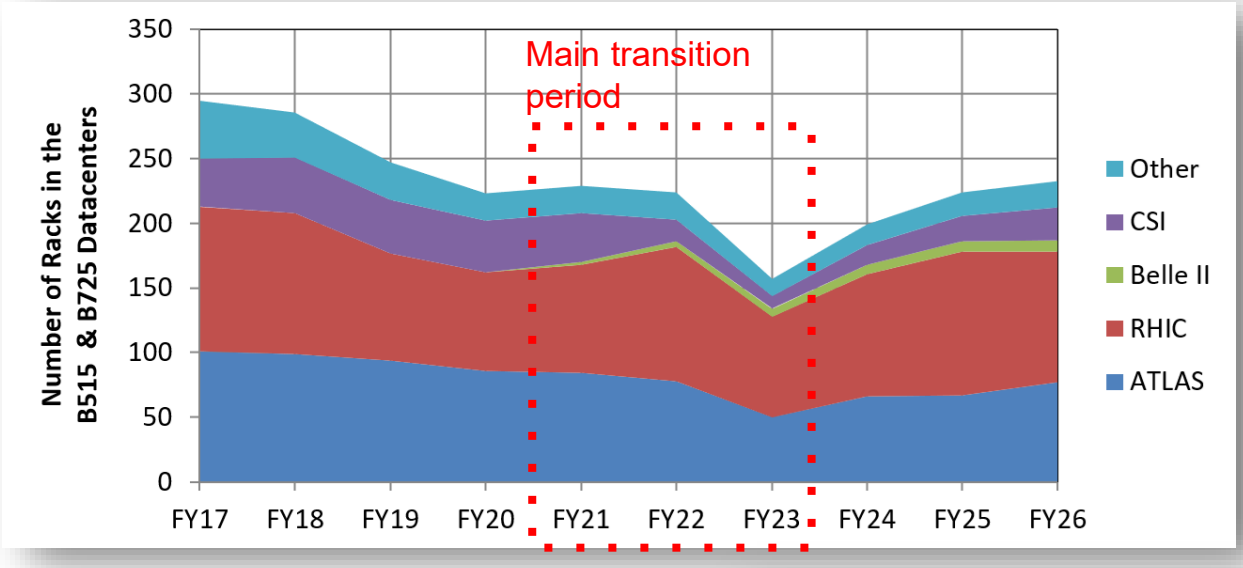


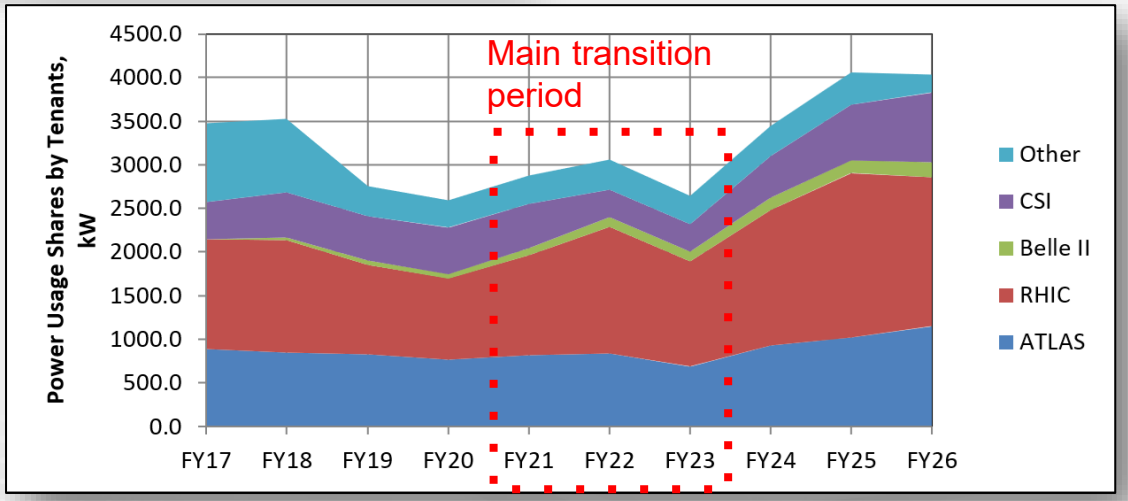
Typical compute row

Typical HW RAID storage row

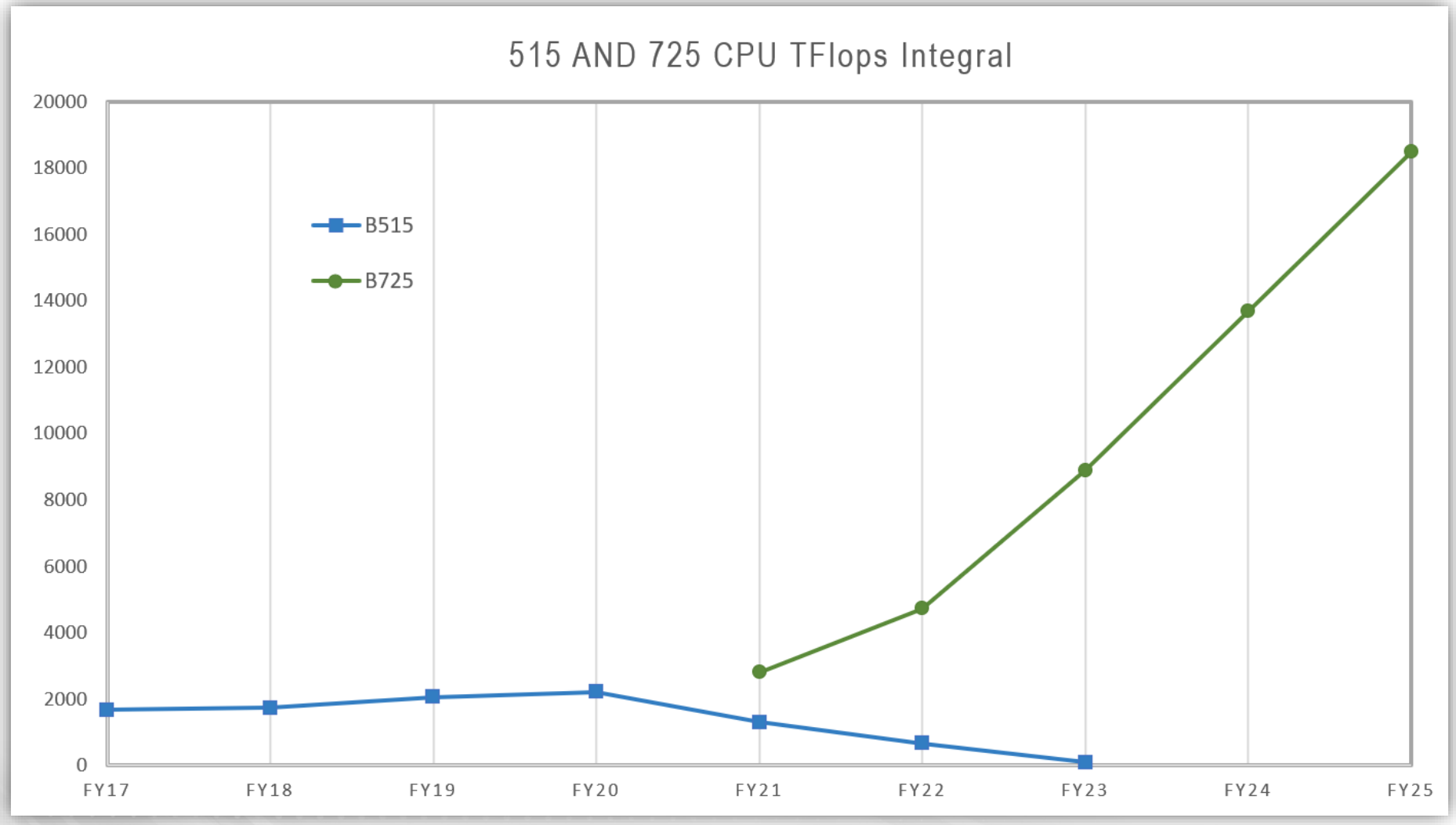
B725 Central Network Equipment Pre-deployment in B515

B515 / B725: Floor Occupancy Projection

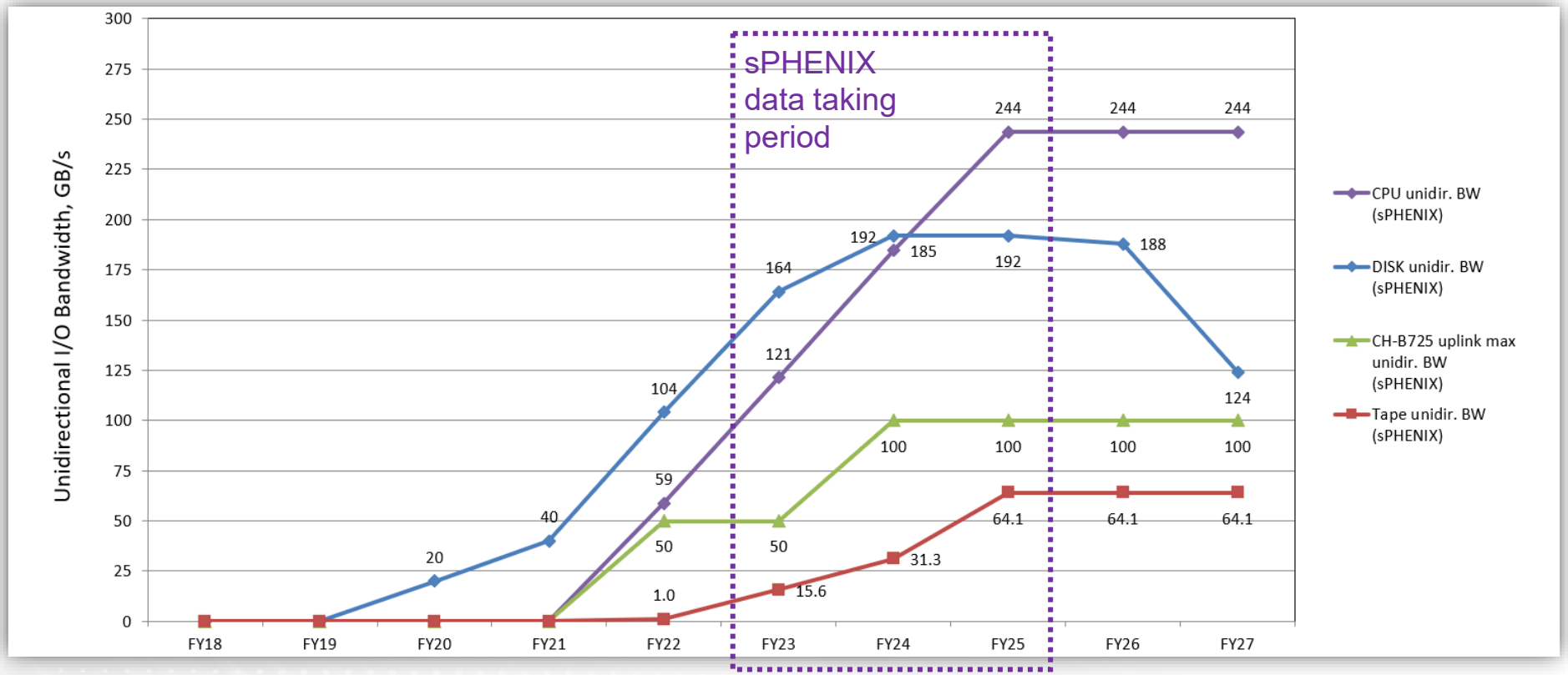




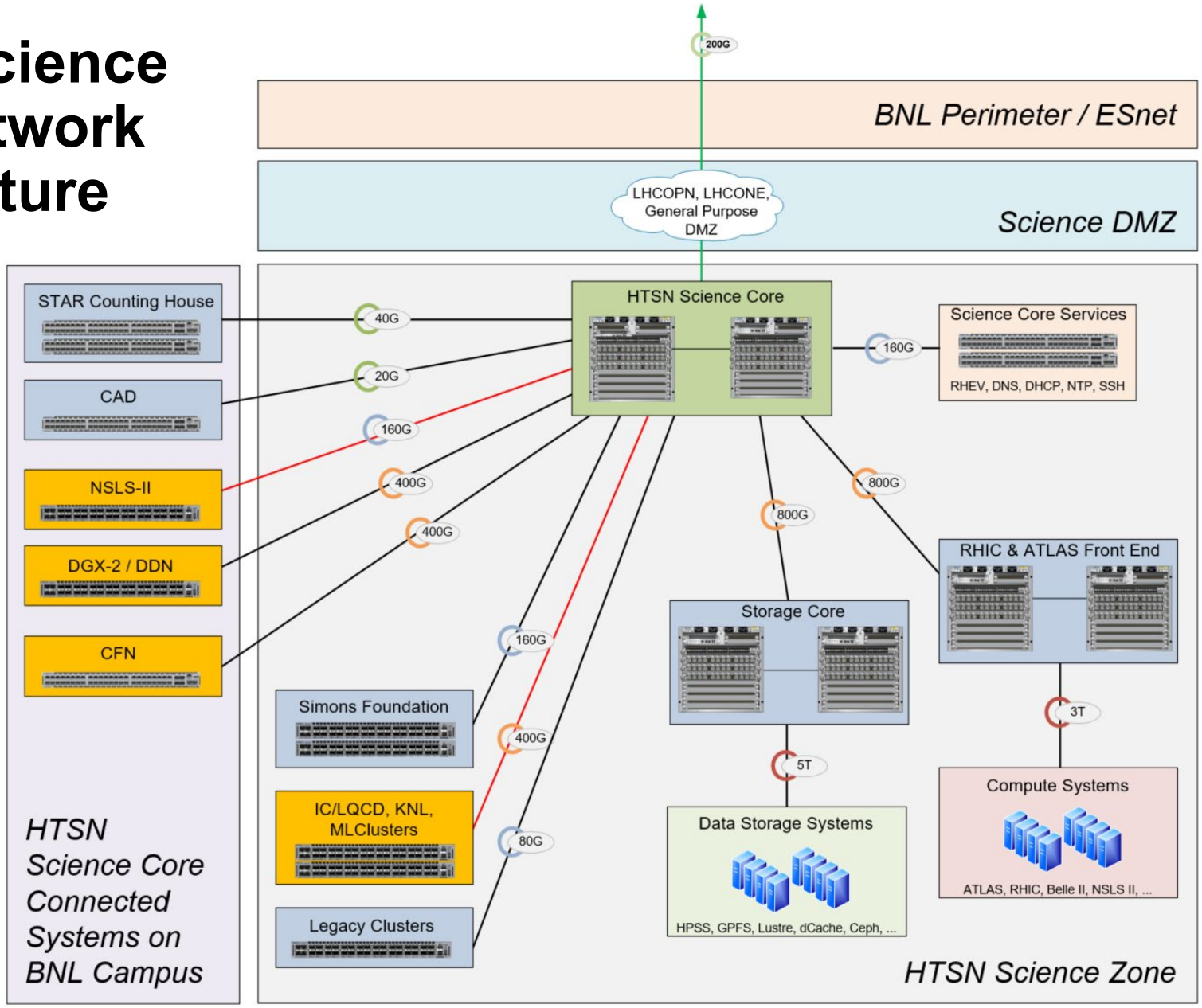
B515 / B725: CPU Capacity Projection



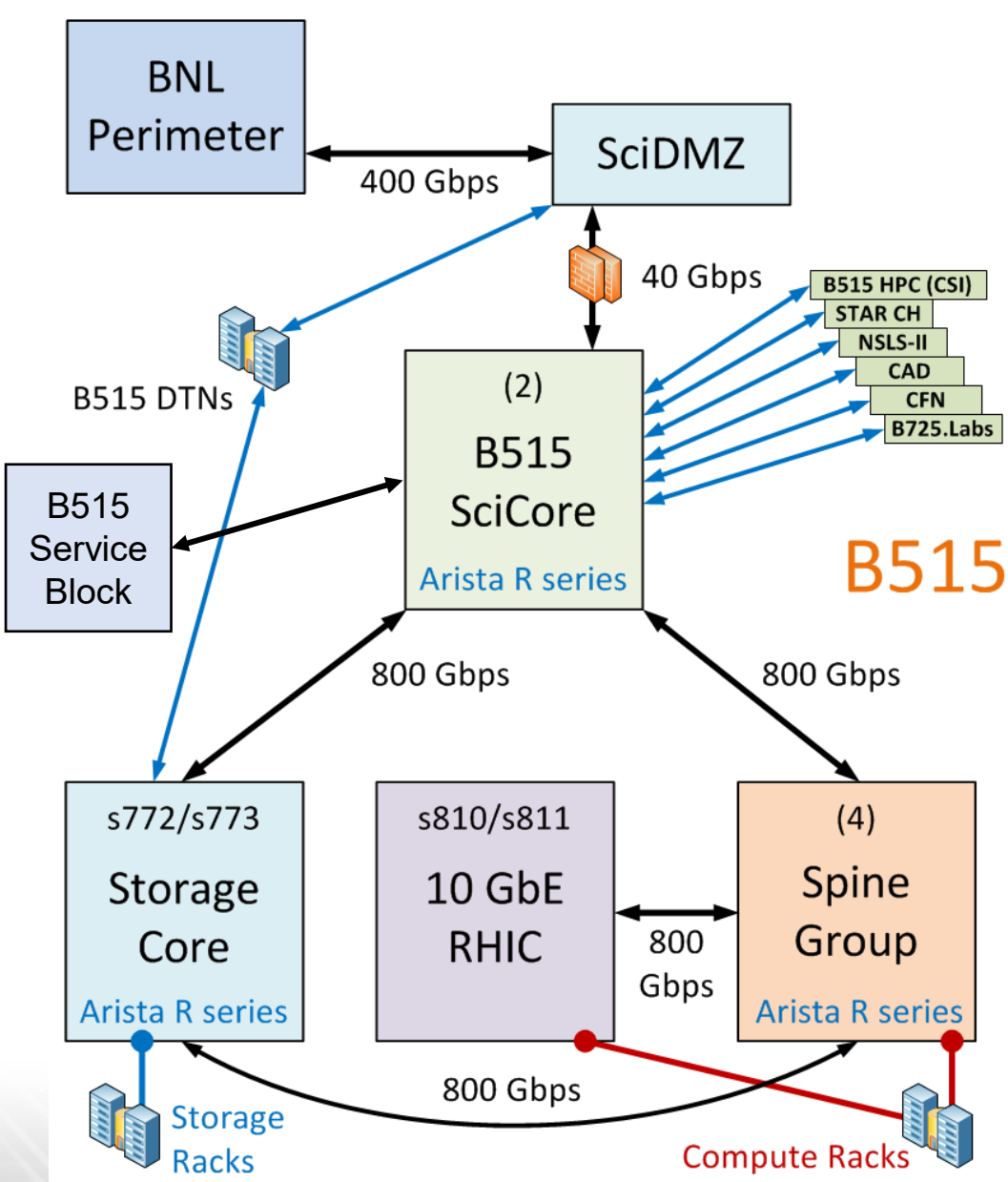
New Network I/O Capacity Drivers for SDCC: sPHENIX Experiment at RHIC

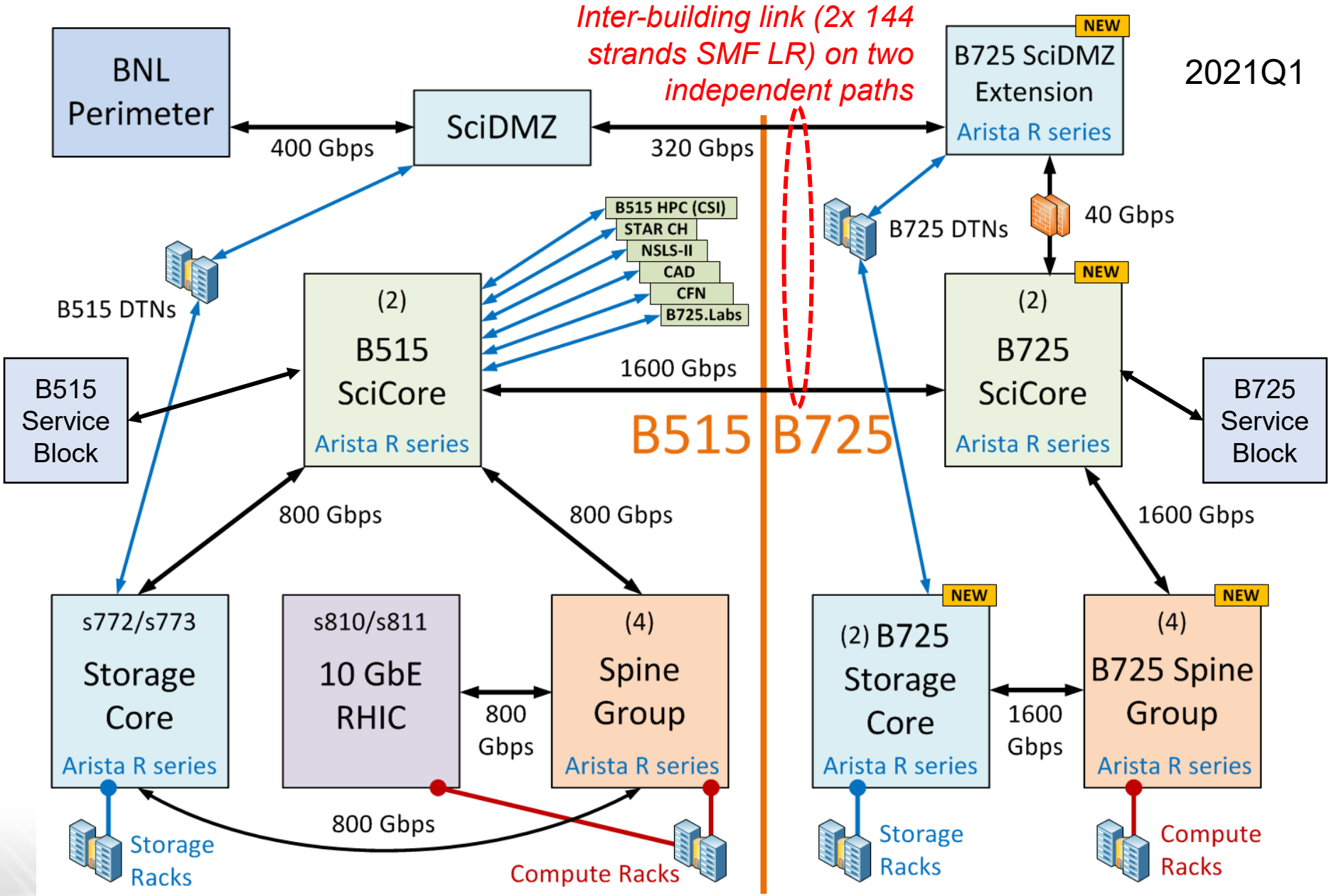


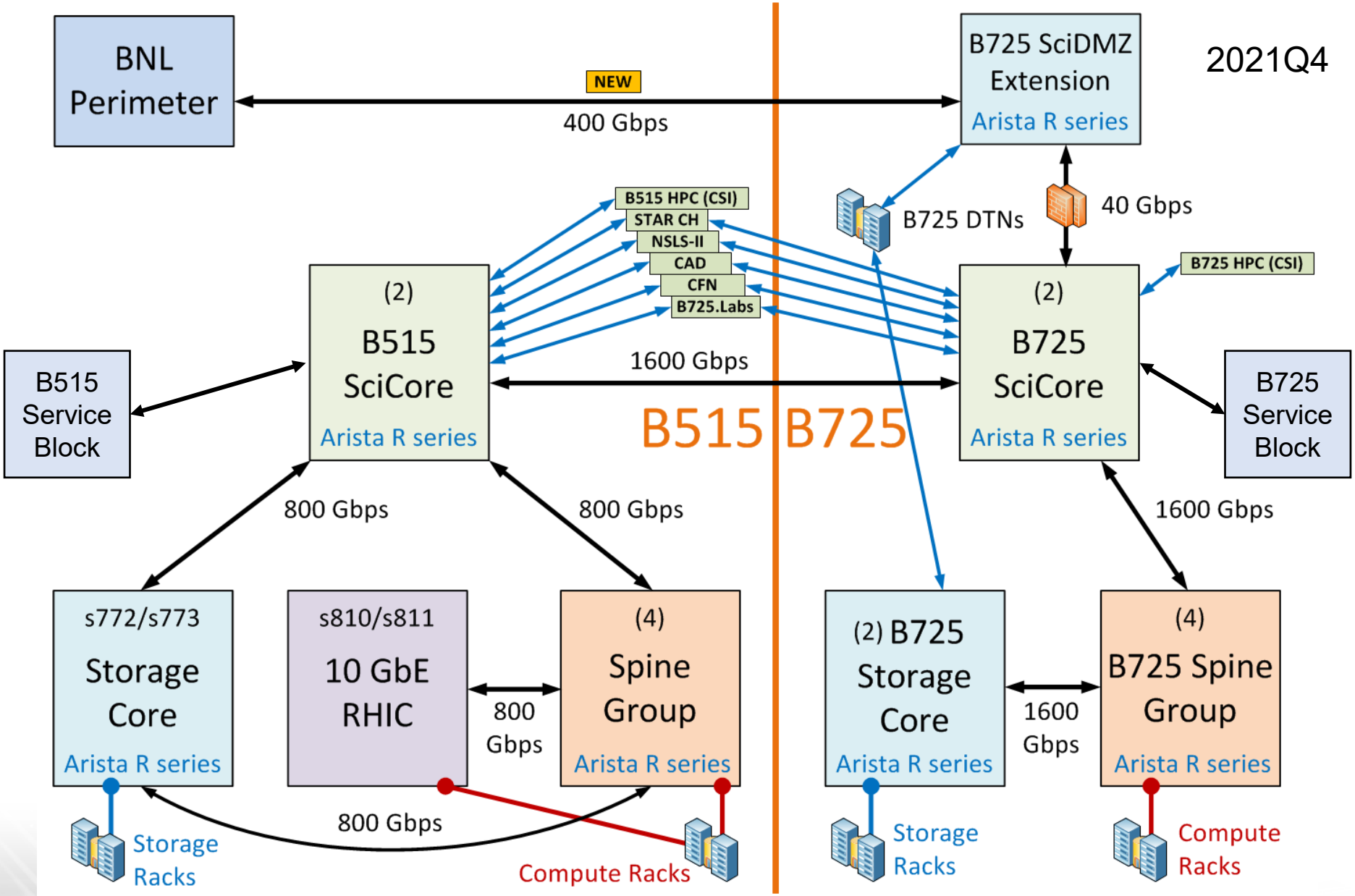
SDCC Science Core Network Architecture

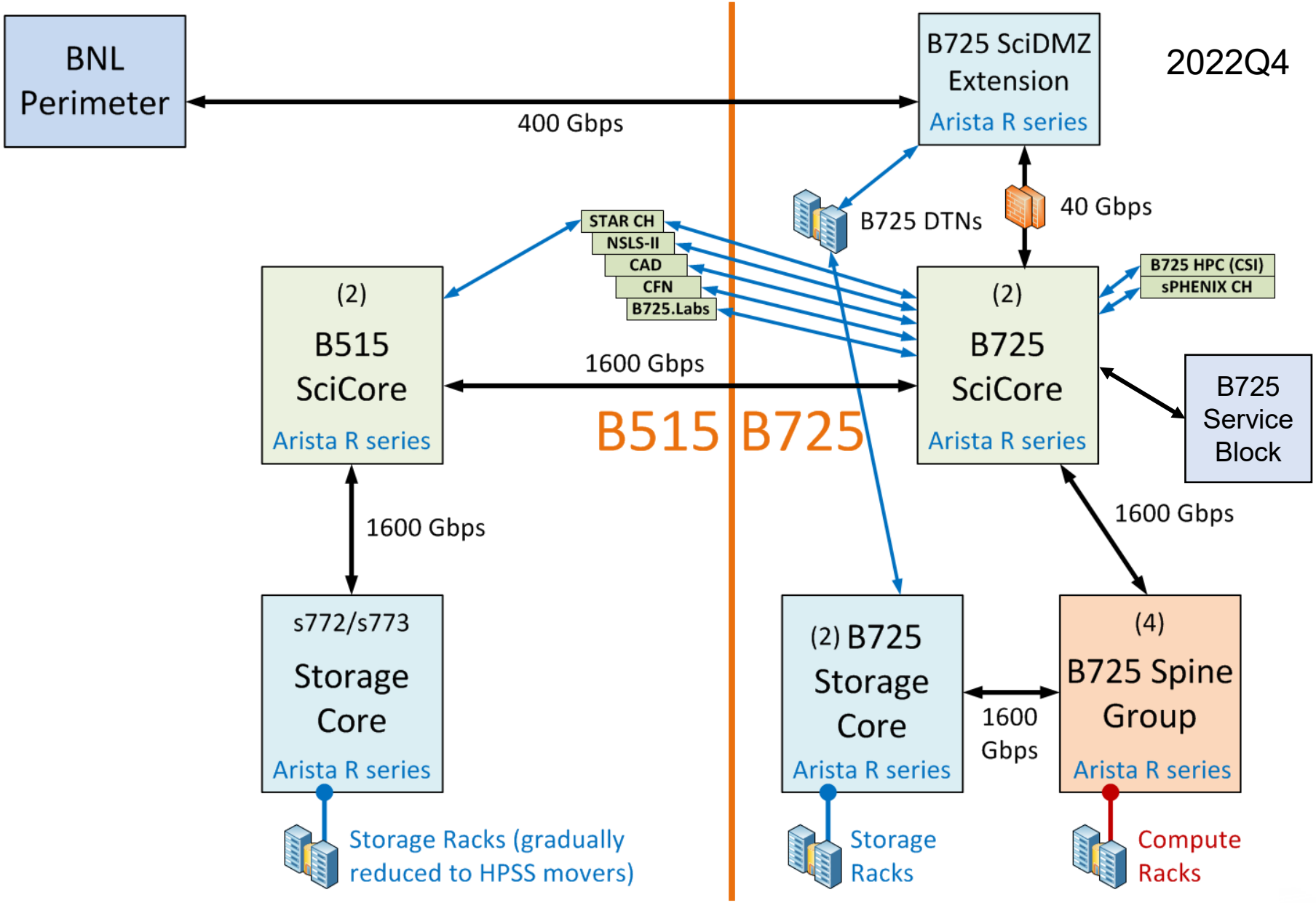


2020Q2


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

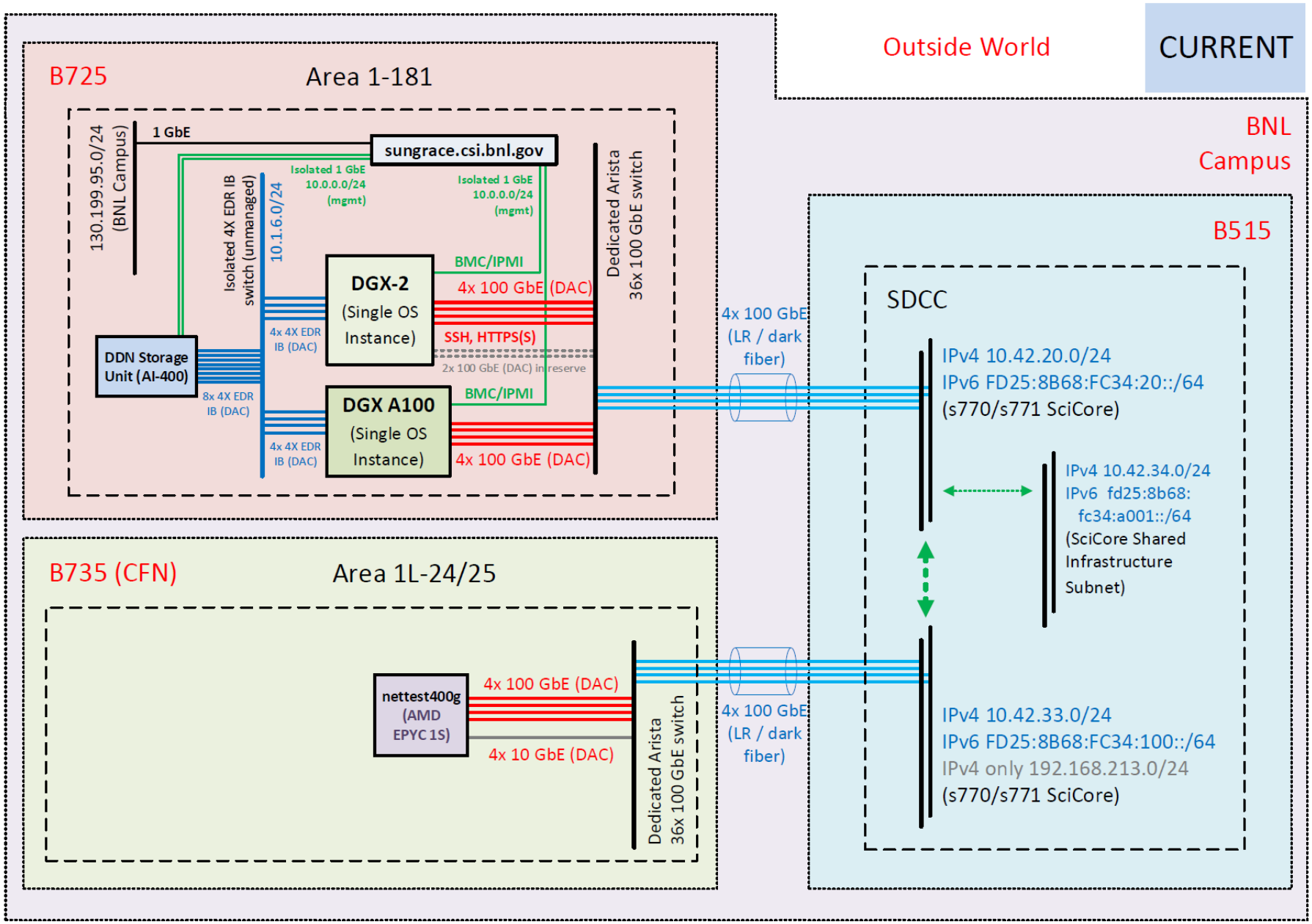

Inter-building link (2x 144 strands SMF LR) on two independent paths
2021Q1
BNL Perimeter
400 Gbps
SciDMZ
320 Gbps
NEW
B725 SciDMZ Extension
Arista R series
B515 HPC (CSI)
STAR CH
NSLS-II
CAD
CFN
B725.Labs
B515 DTNs
B725 DTNs
40 Gbps
(2) B515 SciCore
Arista R series
1600 Gbps
NEW
(2) B725 SciCore
Arista R series
B515 Service Block
B725 Service Block
B515 B725
800 Gbps
800 Gbps
1600 Gbps
s772/s773 Storage Core
Arista R series
s810/s811 10 GbE RHIC
800 Gbps
(4) Spine Group
Arista R series
NEW
(2) B725 Storage Core
Arista R series
1600 Gbps
NEW
(4) B725 Spine Group
Arista R series
800 Gbps
Storage Racks
Compute Racks
Storage Racks
Compute Racks

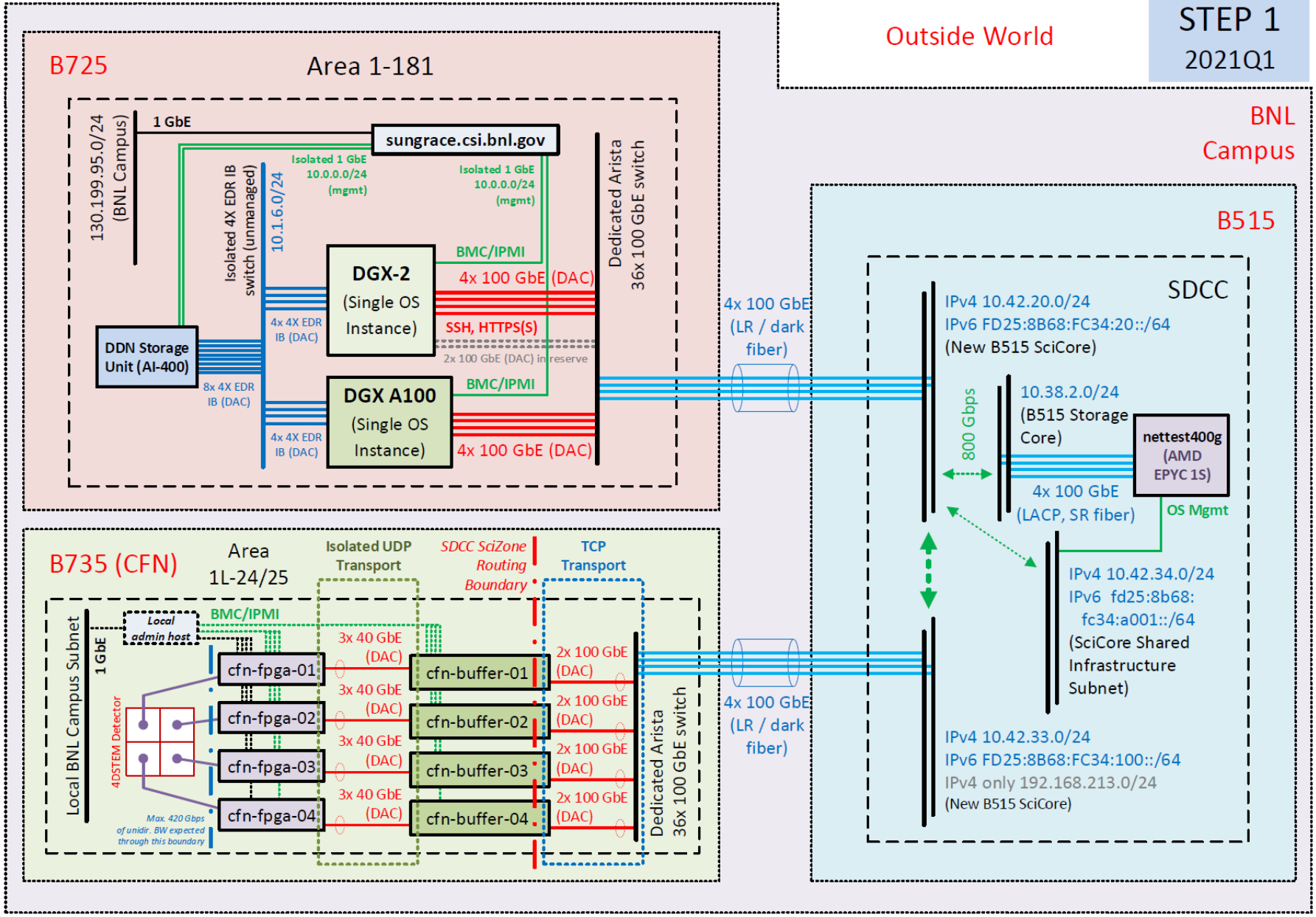

BNL Perimeter
NEW
400 Gbps
B725 SciDMZ Extension
Arista R series
2021Q4
B515 HPC (CSI)
STAR CH
NSLS-II
CAD
CFN
B725.Labs
B725 DTNs
40 Gbps
B725 HPC (CSI)
(2)
B515 SciCore
Arista R series
1600 Gbps
(2)
B725 SciCore
Arista R series
B515 Service Block
B725 Service Block
B515 B725
800 Gbps
800 Gbps
1600 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 R series
1600 Gbps
(4)
B725 Spine Group
Arista R series
800 Gbps
Storage Racks
Compute Racks
Storage Racks
Compute Racks


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

Data Transfer Mechanisms Involved

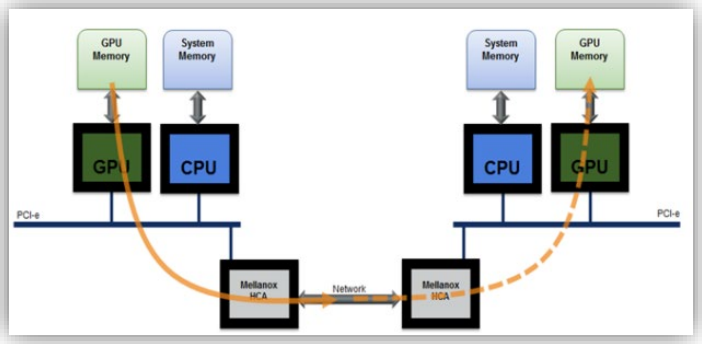
- Reaching up to full utilization of available capacity of network infrastructure built out of multiple lanes of 100 GbE is a considerable challenge, especially if the number of systems talking across such a link is limited
- Multiple options for the underlying network protocols are available:
  - Standard TCP/IP over 1500 MTU and Jumbo-frame Layer-2 uplinks, making sure that LACP redundancy isn't a limiting factor: manual balancing TCP connections across multiple lanes is needed – this usually implies that some buffering at the data source is still needed
  - Pure UDP/IP driven high bandwidth transfers are theoretically possible given the expected high reliability of the links, yet it is taking a considerable effort on the switch configuration along the path to ensure high efficiency of the UDP transfers
  - Mounting GPFS or Lustre filesystems over across the uplinks using native Ethernet (depending on the physical length of the fiber path, dealing with the RTT in 10s of millisecond range may become an issue)
  - RDMA over Converged Ethernet (RoCE): IB encapsulation into Ethernet which could be particularly useful for the RAM to RAM transfers eliminating the need for data storage buffers on DAQ side completely (RoCE implementation in Mellanox ConnectX-5 and ConnectX-6 product lines is of particular interest in case of BNL)

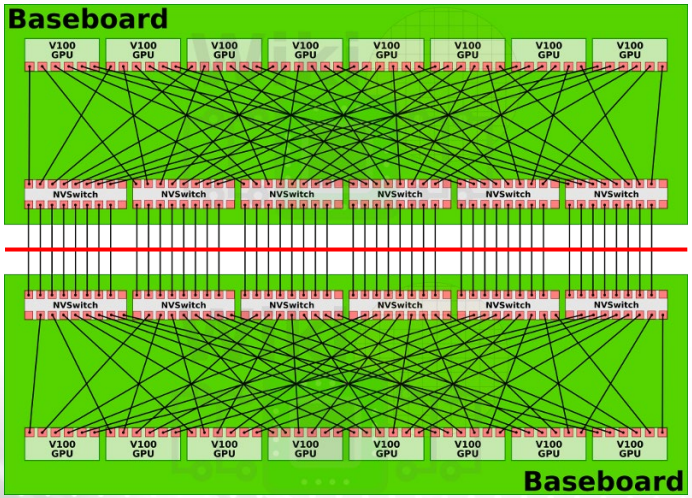

Outside World
CURRENT
BNL Campus
B725
Area 1-181
130.199.95.0/24 (BNL Campus)
1 GbE
sungrace.csi.bnl.gov
Isolated 1 GbE 10.0.0.0/24 (mgmt)
Isolated 1 GbE 10.0.0.0/24 (mgmt)
Isolated 4X EDR IB switch (unmanaged)
10.1.6.0/24
Dedicated Arista 36x 100 GbE switch
BMC/IPMI
DGX-2 (Single OS Instance)
4x 100 GbE (DAC)
SSH, HTTPS(S)
2x 100 GbE (DAC) in reserve
4x 4X EDR IB (DAC)
DDN Storage Unit (AI-400)
8x 4X EDR IB (DAC)
BMC/IPMI
DGX A100 (Single OS Instance)
4x 4X EDR IB (DAC)
4x 100 GbE (DAC)
4x 100 GbE (LR / dark fiber)
B515
SDCC
IPv4 10.42.20.0/24
IPv6 FD25:8B68:FC34:20::/64
(s770/s771 SciCore)
IPv4 10.42.34.0/24
IPv6 fd25:8b68:fc34:a001::/64
(SciCore Shared Infrastructure Subnet)
B735 (CFN)
Area 1L-24/25
nettest400g (AMD EPYC 1S)
4x 100 GbE (DAC)
4x 10 GbE (DAC)
Dedicated Arista 36x 100 GbE switch
4x 100 GbE (LR / dark fiber)
IPv4 10.42.33.0/24
IPv6 FD25:8B68:FC34:100::/64
IPv4 only 192.168.213.0/24
(s770/s771 SciCore)


Outside World
STEP 1
2021Q1
BNL Campus
B725
Area 1-181
130.199.95.0/24 (BNL Campus)
1 GbE
sungrace.csi.bnl.gov
Isolated 1 GbE 10.0.0.0/24 (mgmt)
Isolated 1 GbE 10.0.0.0/24 (mgmt)
Isolated 4X EDR IB switch (unmanaged)
10.1.6.0/24
Dedicated Arista 36x 100 GbE switch
BMC/IPMI
DGX-2 (Single OS Instance)
4x 100 GbE (DAC)
SSH, HTTPS(S)
2x 100 GbE (DAC) in reserve
DDN Storage Unit (AI-400)
4x 4X EDR IB (DAC)
8x 4X EDR IB (DAC)
BMC/IPMI
DGX A100 (Single OS Instance)
4x 4X EDR IB (DAC)
4x 100 GbE (DAC)
4x 100 GbE (LR / dark fiber)
B515
SDCC
IPv4 10.42.20.0/24
IPv6 FD25:8B68:FC34:20::/64
(New B515 SciCore)
800 Gbps
10.38.2.0/24
(B515 Storage Core)
nettest400g (AMD EPYC 1S)
4x 100 GbE (LACP, SR fiber)
OS Mgmt
IPv4 10.42.34.0/24
IPv6 fd25:8b68:fc34:a001::/64
(SciCore Shared Infrastructure Subnet)
B735 (CFN)
Area 1L-24/25
Isolated UDP Transport
SDCC SciZone Routing Boundary
TCP Transport
Local BNL Campus Subnet
1 GbE
Local admin host
BMC/IPMI
4DSTEM Detector
cfn-fpga-01
cfn-fpga-02
cfn-fpga-03
cfn-fpga-04
3x 40 GbE (DAC)
cfn-buffer-01
cfn-buffer-02
cfn-buffer-03
cfn-buffer-04
2x 100 GbE (DAC)
Dedicated Arista 36x 100 GbE switch
Max. 420 Gbps of unidir. BW expected through this boundary
4x 100 GbE (LR / dark fiber)
IPv4 10.42.33.0/24
IPv6 FD25:8B68:FC34:100::/64
IPv4 only 192.168.213.0/24
(New B515 SciCore)

Data Transfer Mechanisms Involved

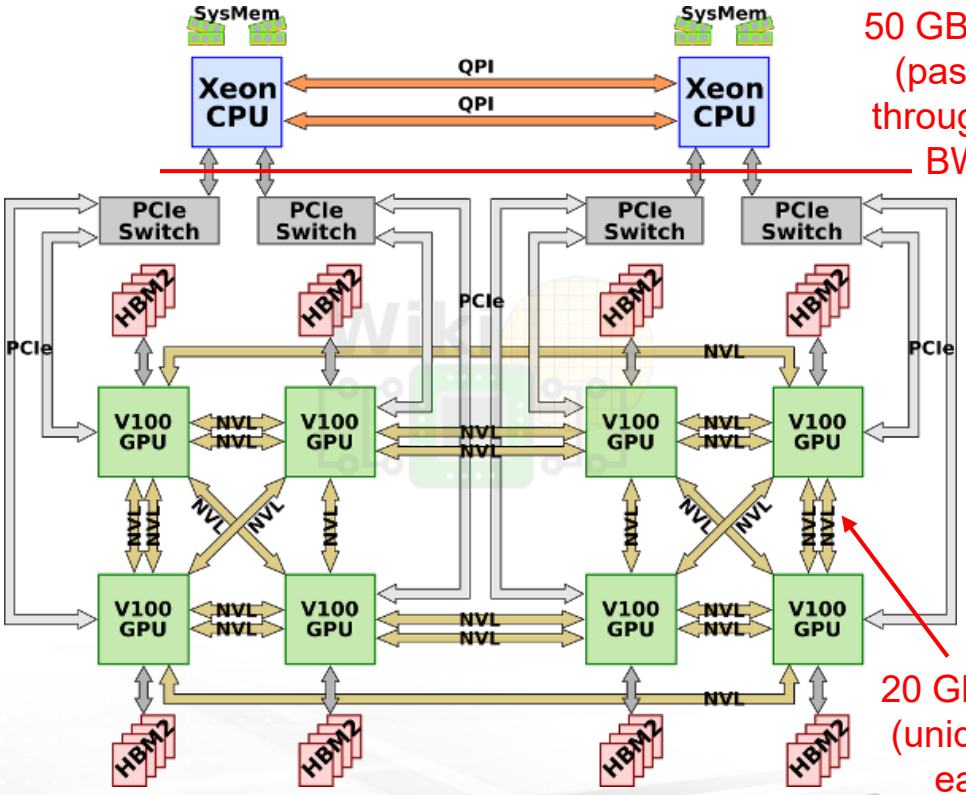
- Multiple options for the underlying network protocols are available:
  - In case of a GPU-enabled systems on one or both sides of a data transfer channel RoCE technology can be used to avoid CPU / system RAM bottlenecks and perform GPU RAM to GPU RAM transfers





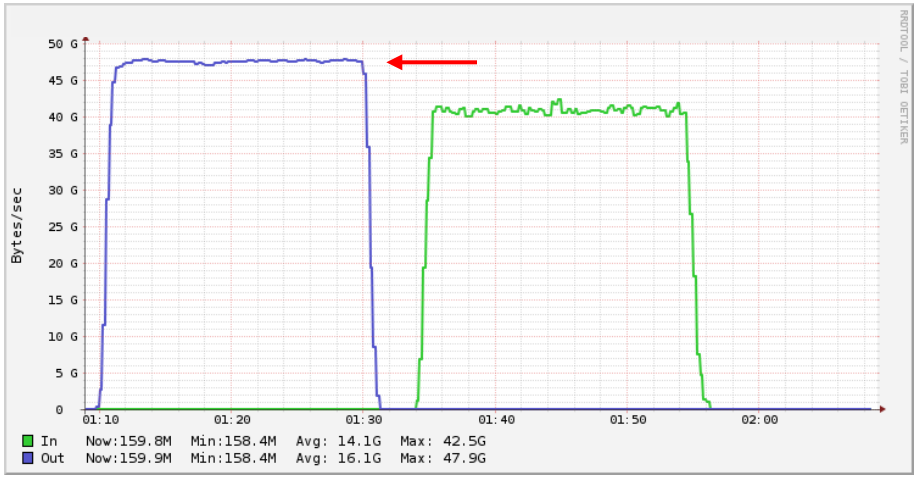
960 GB/s (unidir.)

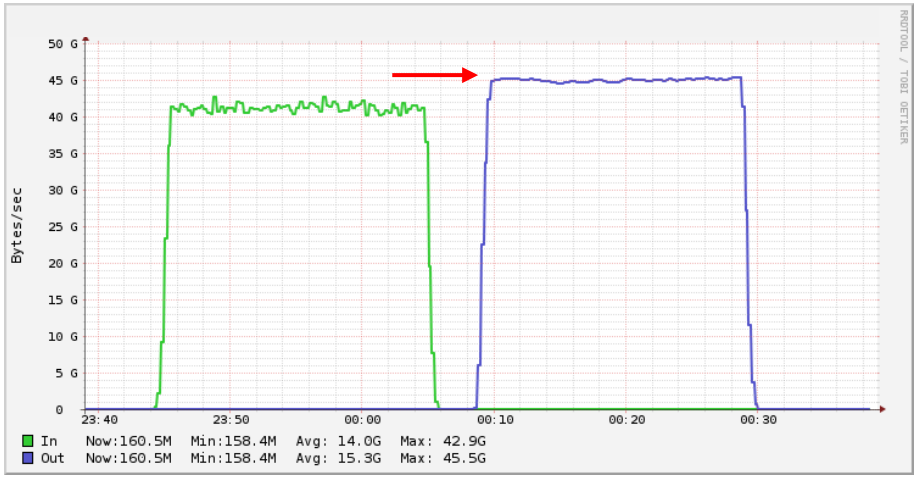
Nvidia DGX-2



Data Transfer Mechanisms Involved

- Evaluation of 400 Gbps channel capability between B515 SDCC and B725 CSI Advanced Computing Lab (ACL) in 2020Q3
  - 16 2x 25 GbE (LACP) attached systems on SDCC side (16x OS instances)
  - 4x 100 GbE (LACP) attached DGX-2 unit (single OS instance)
  - iperf3 with TCP/IP MTU 1500





- DGX-2 to SDCC: 16x 2 threads: max 42.9 GB/s = 86% of theoretical max
- SDCC to DGX-2: 16x 6 threads: max 47.9 GB/s = 96% of theoretical max

The Outlook for Future Developments

- The deployment of 4D-STEM detector in B735 CFN is expected to happen in 2021Q1 which would trigger the full scale build up of DAQ system for it. This should allow us to fully explore the potential of the direct streaming of the experimental data acquired by this detector to the systems deployed in the SDCC datacenter and B725 CSI ACL
- The NSLS II is significantly increasing their presence in the SDCC datacenter in 2020Q4 with potential for triggering upgrade of their uplink to the SDCC SciCore from 160 Gbps to 400 Gbps
  - The CryoEM facility recently added to the NSLS II complex while being currently satisfied with capabilities of 160 Gbps uplink of the NSLS II may push that boundary if they ever deploy detectors similar to 4D-STEM of CFN
- The sPHENIX Counting House is expected to get their 200-400 GbE uplink to SDCC datacenter based on 100 GbE lanes in FY21-22 in preparation for the beginning of sPHENIX datataking in FY23, which can be upgraded to 800 Gbps using 400 GbE lanes in FY24. With the current design the DAQ system of sPHENIX is retaining the data buffers on the CH side.
- The STAR Experiment at RHIC is expected to collect data throughout FY22-25 period and the investigation of the investigation of the options for direct offloading of experimental data from STAR DAQ system to the STAR CH buffer located in the SDCC datacenter is ongoing
- The upgrade of BNL Perimeter to 400 GbE capable equipment is likely to happen in FY23 (mostly driven by requirements of ATLAS Tier-1 site at BNL at the moment)

Questions & Comments