Network Architecture & Operations of the RACF/SDCC Facility

Alexandr ZAYTSEV

alezayt@bnl.gov

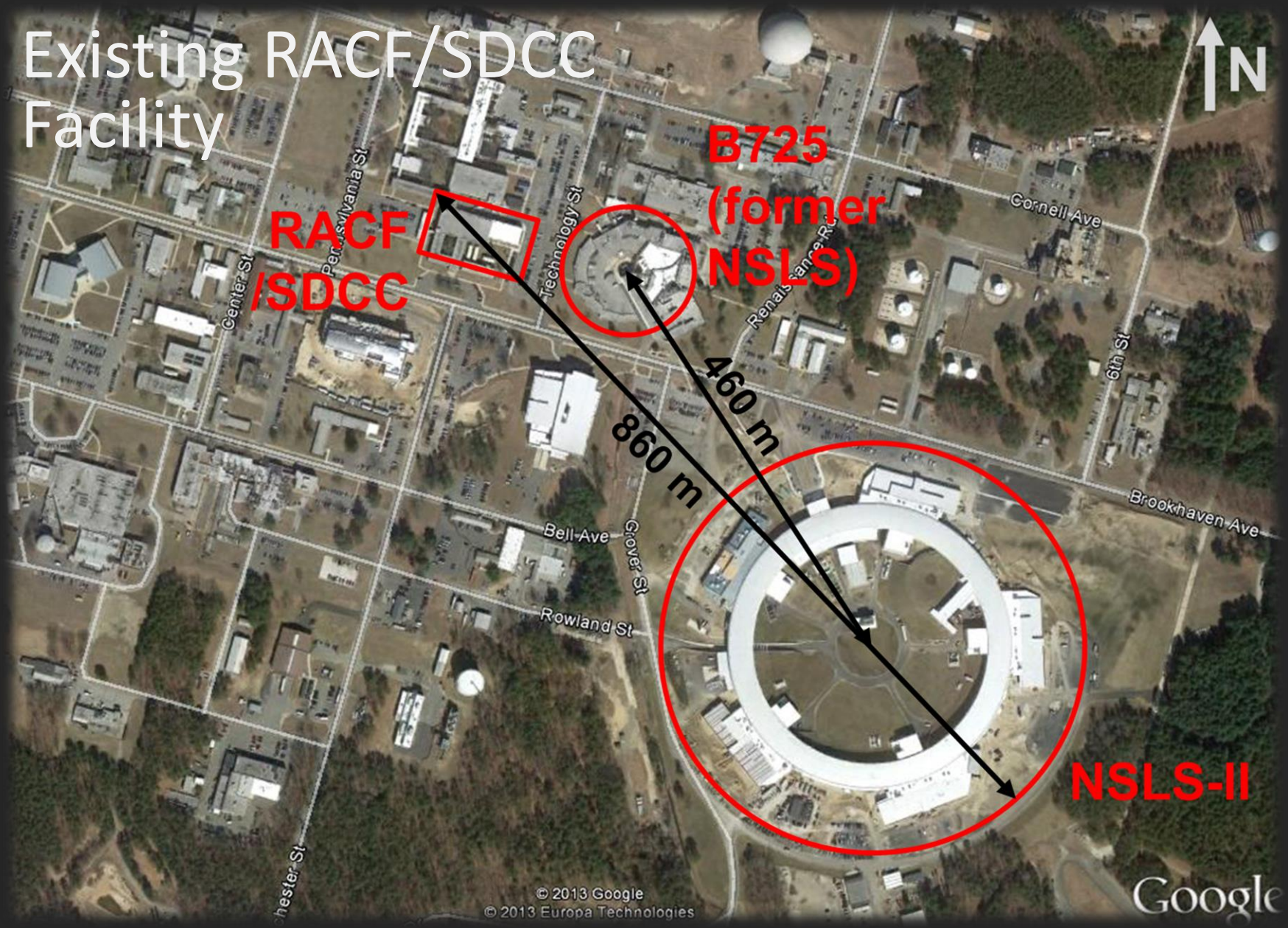
BROOKHAVEN NATIONAL LABORATORY

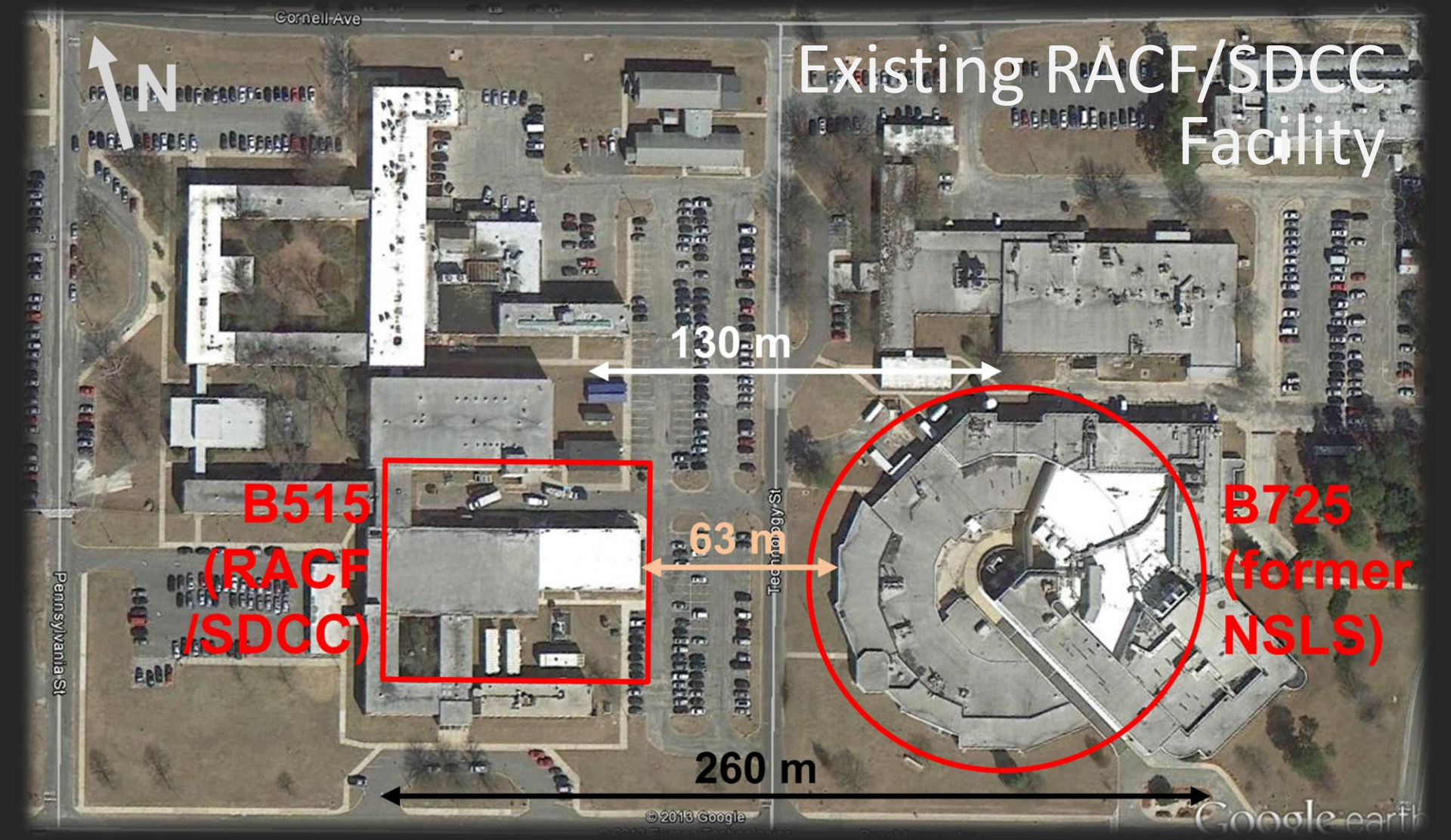
U.S. DEPARTMENT OF ENERGY

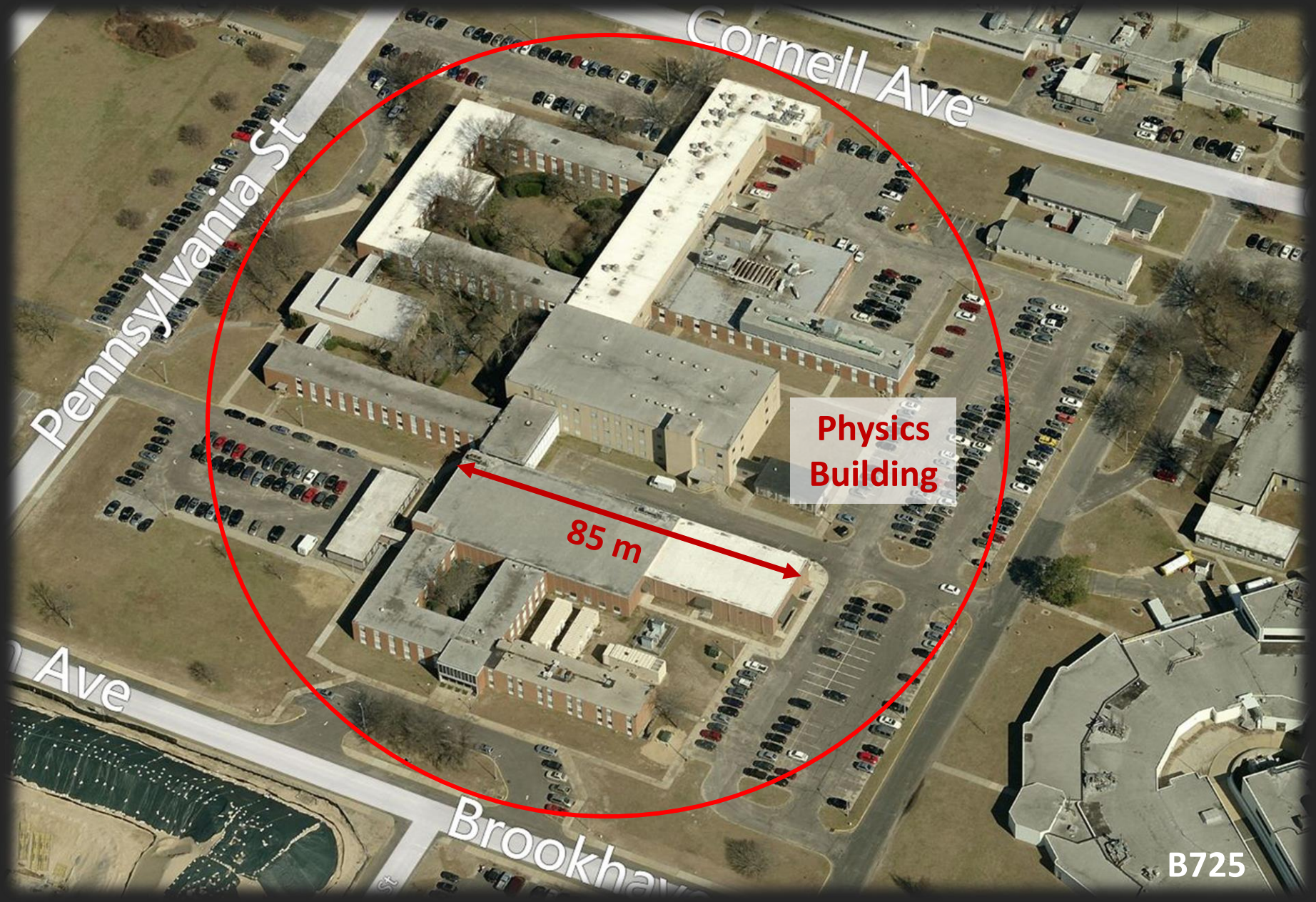
Outlook

- RACF/SDCC facility and its subsystems
- BNL and RACF/SDCC logical network structure
- Organizational structures involved
- Network core and major distribution systems
- Ongoing technology transitions
- Longer term outlook
  - Addition of the new data center (B725) in FY21
  - Finalizing B515 data center into the CDCE area in FY23
- Summary & Conclusion

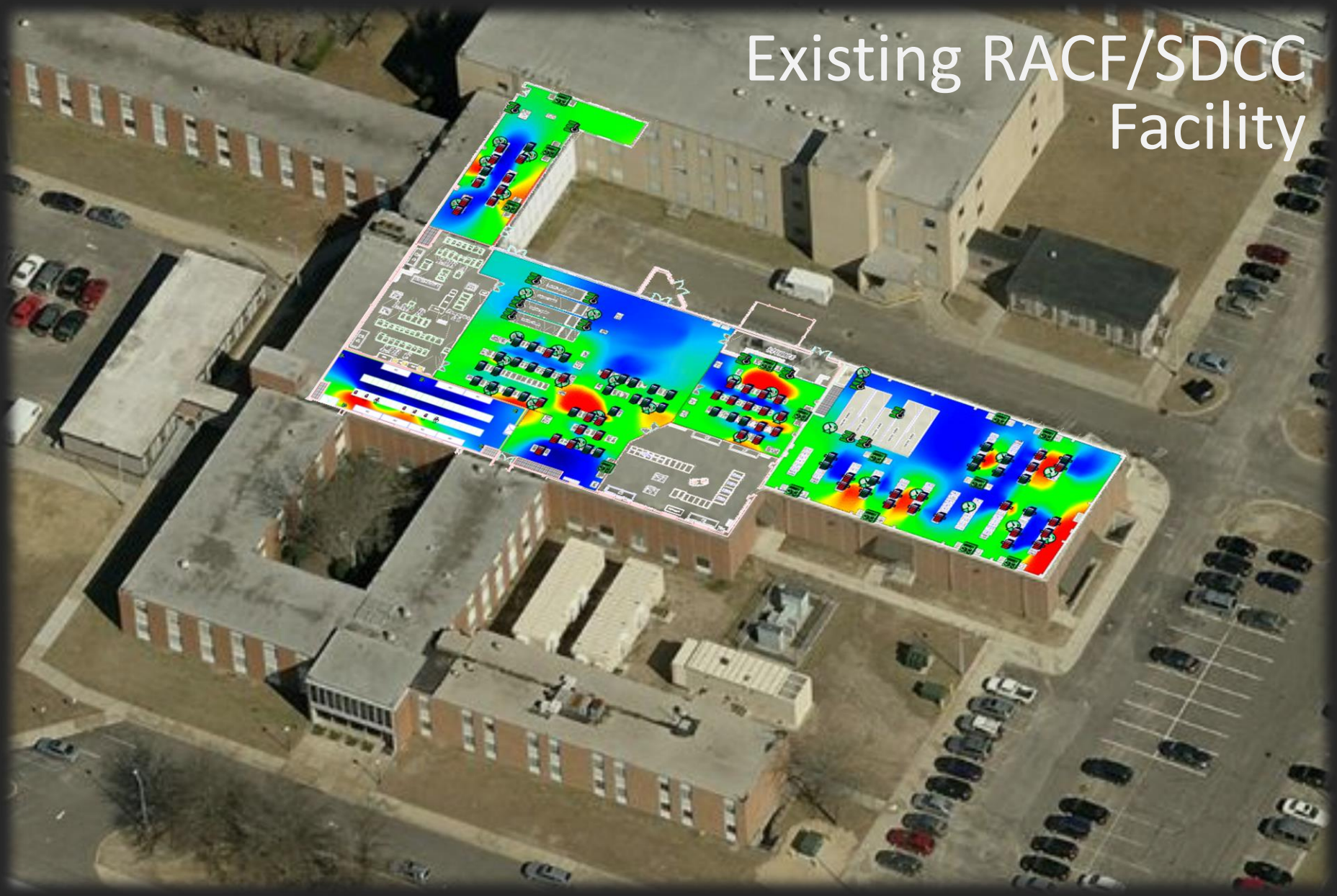
Existing RACF/SDCC Facility



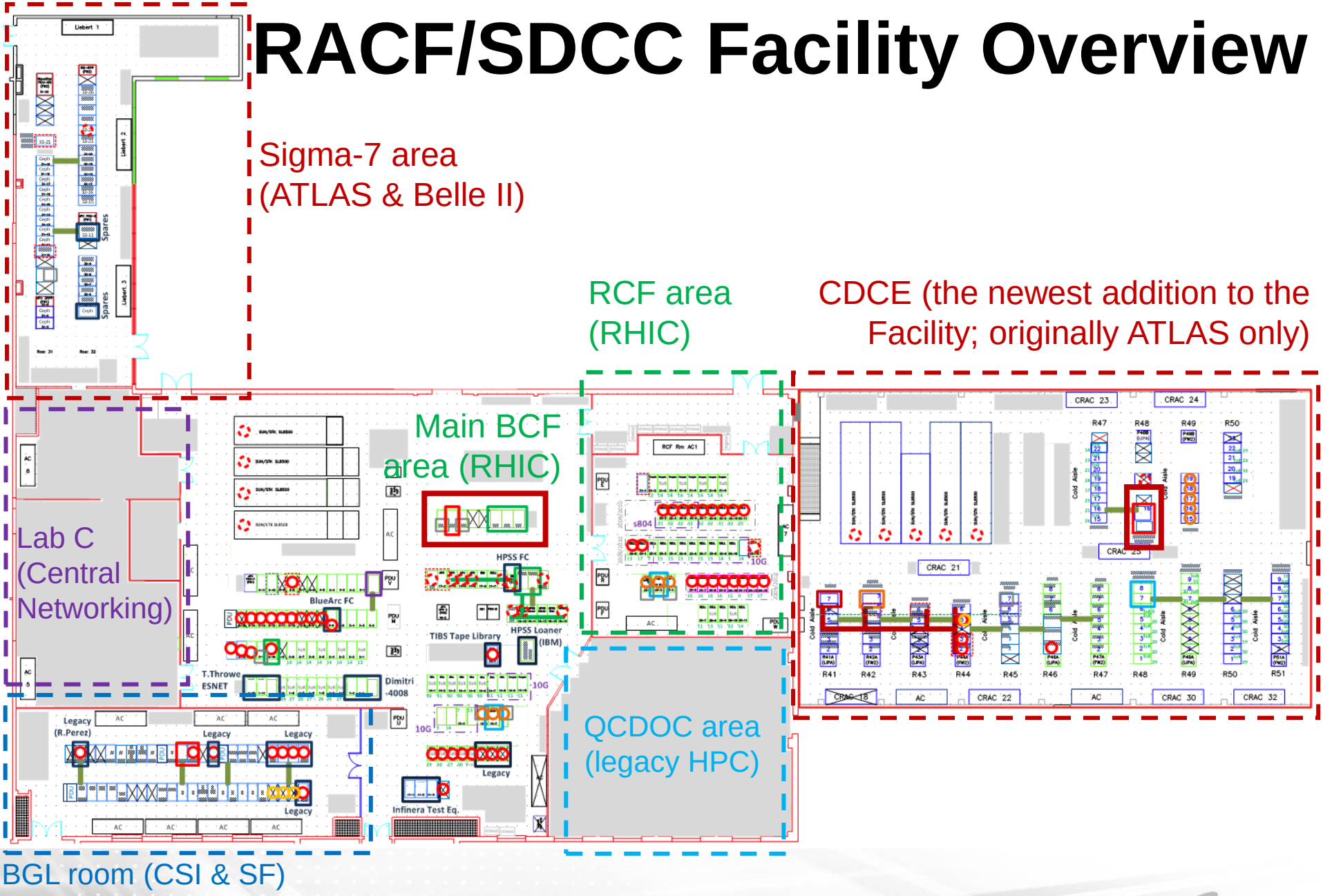

Existing RACF/SDCC Facility
N
Cornell Ave
Pennsylvania St
Technology St
130 m
63 m
260 m
B515 (RACF /SDCC)
B725 (former NSLS)
© 2013 Google
Google earth


Cornell Ave
Pennsylvania St
Physics Building
85 m
Ave
Brookhav
B725

Existing RACF/SDCC Facility



RACF/SDCC Facility Overview

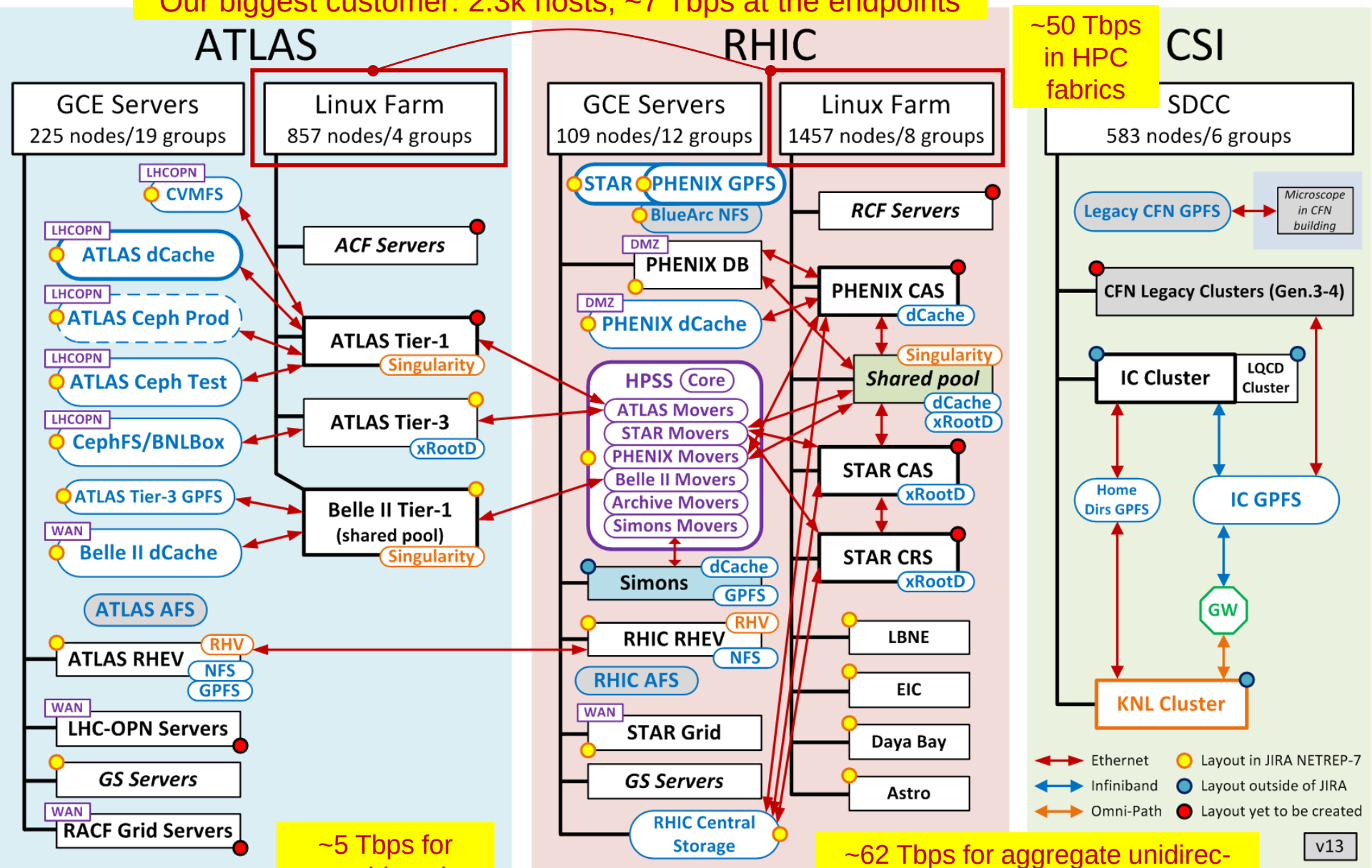


RACF/SDCC Facility Networks: Status and the Ongoing Transitions

- Merge of several client-specific network systems into a single network core (Science Core) based on Arista 7500E & 7500R series equipment
- Transition to a unified compute and storage architecture based on Arista ToR leaf switches connected to a unified spine group using ECMP routing / L3 isolation on the rack level (publicly routable /27 or /26 subnets for IPv4, /64 – for IPv6) for all 1 GbE connected compute nodes
- Migrating the majority of the storage headnodes requiring more than 2x 10 GbE of network connectivity to 25 GbE based network uplinks to efficiently utilize the 100GBASE-SR4 technology
- Phasing out Fibre Channel (FC) switch based storage interconnects systems and replacing those with SAS attached and direct FC attached storage
- Preparation for adding B725 based data center to the Facility in early FY21 (CFR Project), deploying the new B725 network infrastructure and handling the period of B515 and B725 data center inter-operation
- Physical consolidation of the B515 data center into the CDCE area by FY23

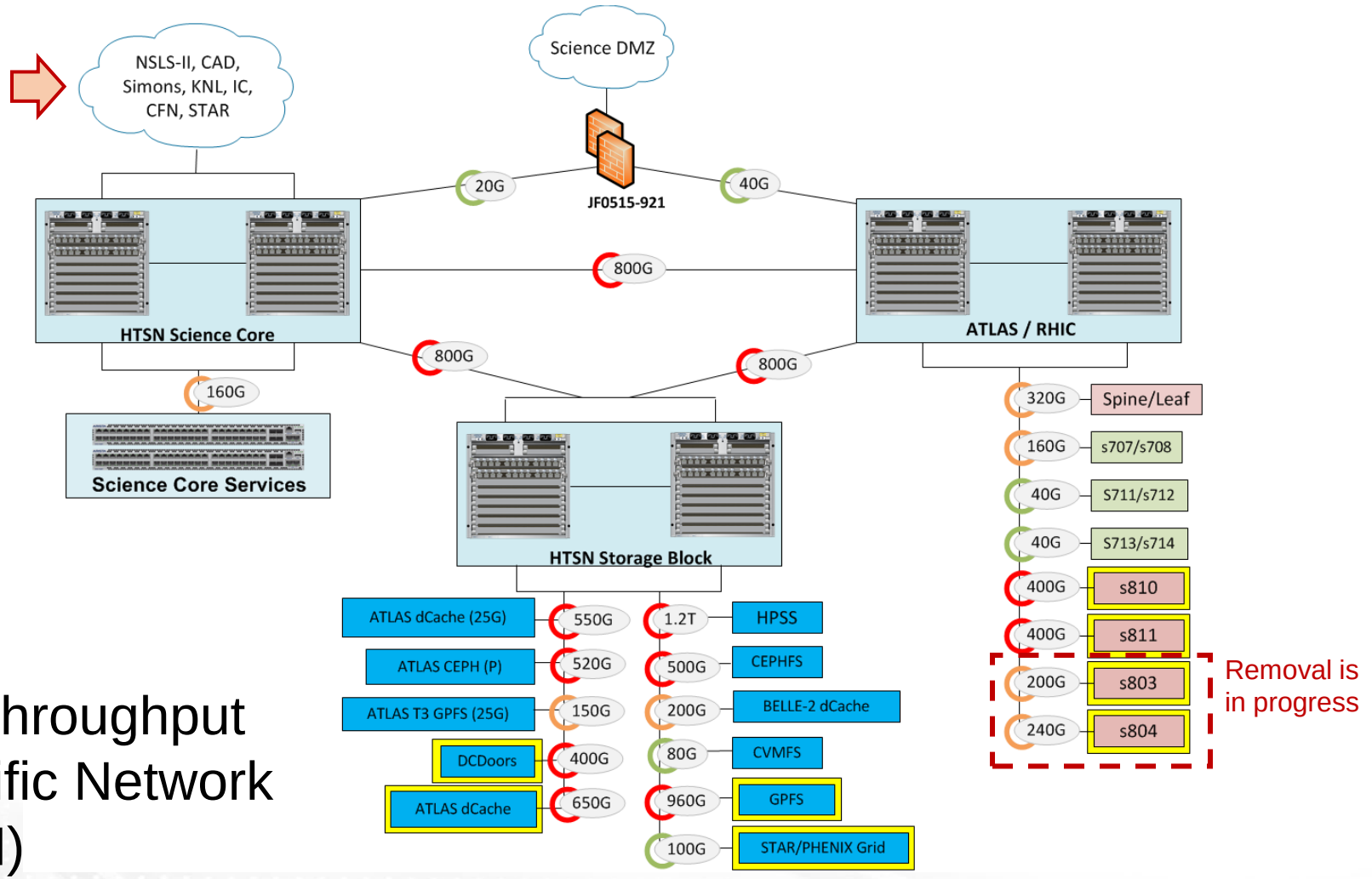
RACF/SDCC Subsystems (3231 nodes/49 groups on Sep 23, 2018)

Our biggest customer: 2.3k hosts, ~7 Tbps at the endpoints



Overall Architecture: Science Core

User groups residing on BNL Campus are getting connected here

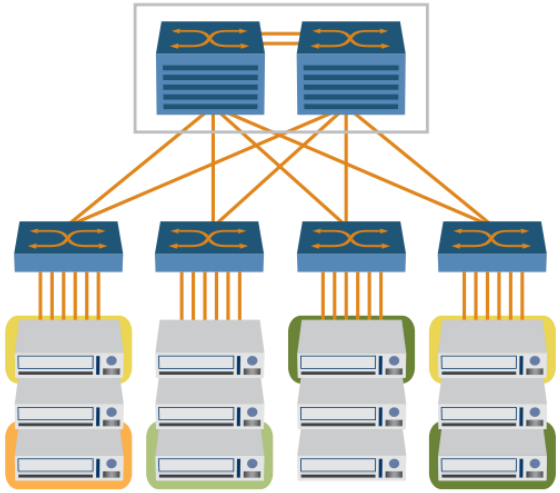


Removal is in progress

High Throughput Scientific Network (HTSN)

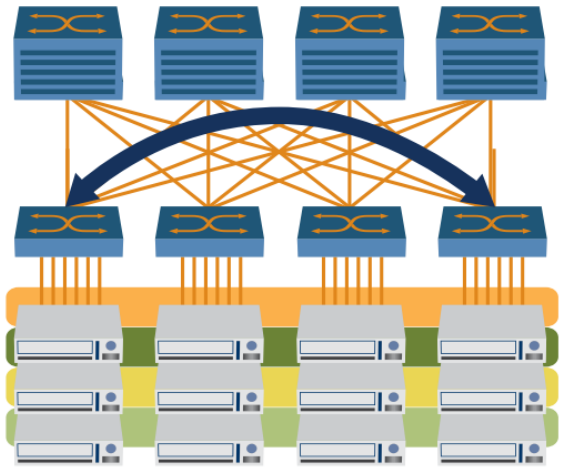
Overall Architecture: Arista “R” Series Portfolio

Multiple isolated L3 domains. Routing advertisement to the leafs via eBGP.



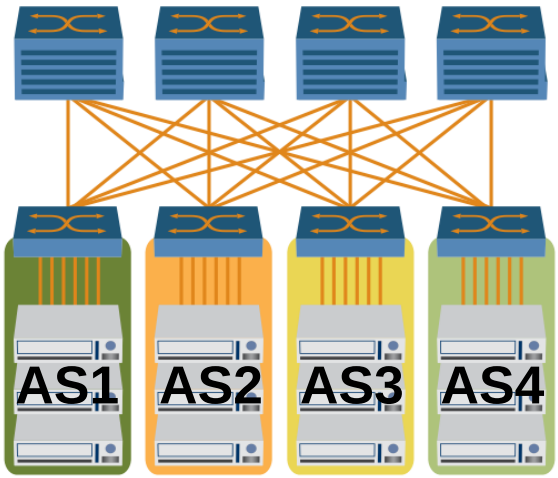
Layer 2 / MLAG

Scalability / reliability limits of L2 domains



L2 over Layer 3 VXLAN

Could be useful for extending relatively low / bandwidth & less critical traffic across multiple L3 domains



Layer 3 / ECMP

ECMP provides load balancing and adds resiliency across entire fabric

Overall Architecture Before 2017

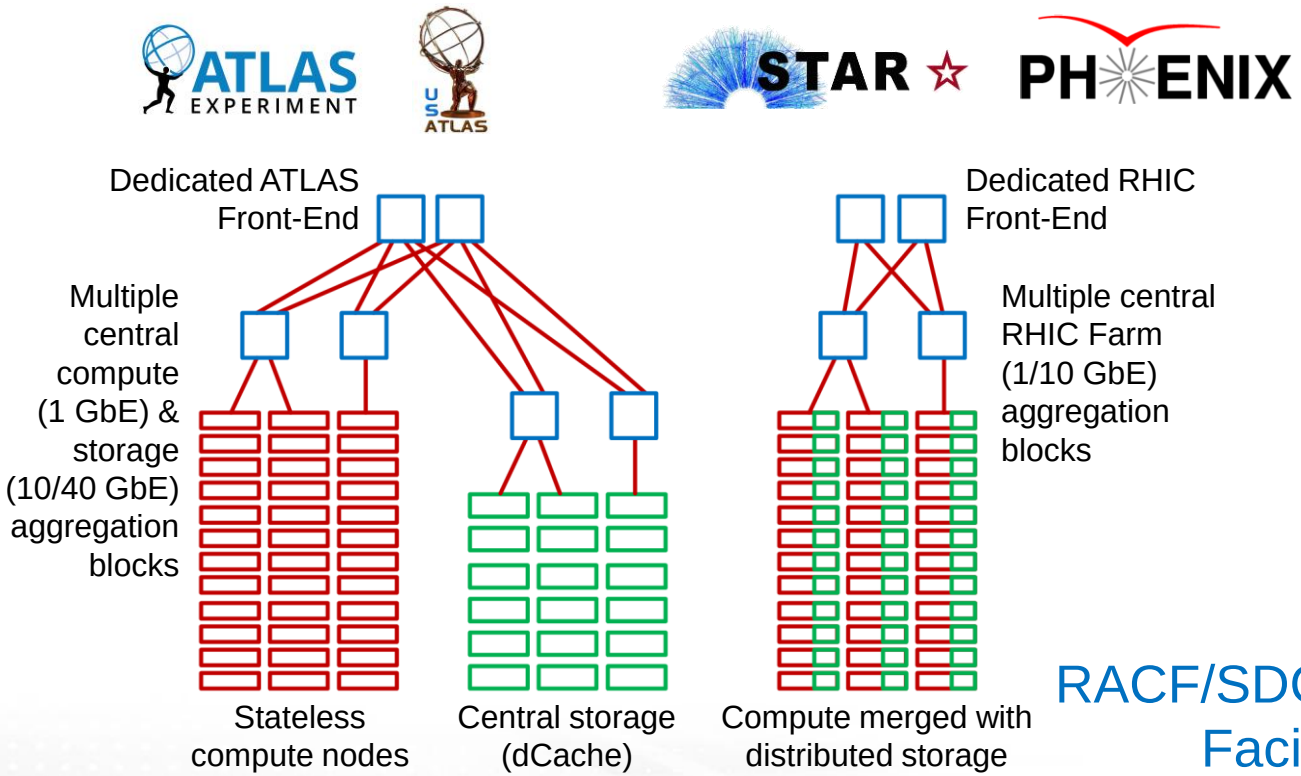
The World (IPv4/IPv6)

BNL Campus (IPv4 only)

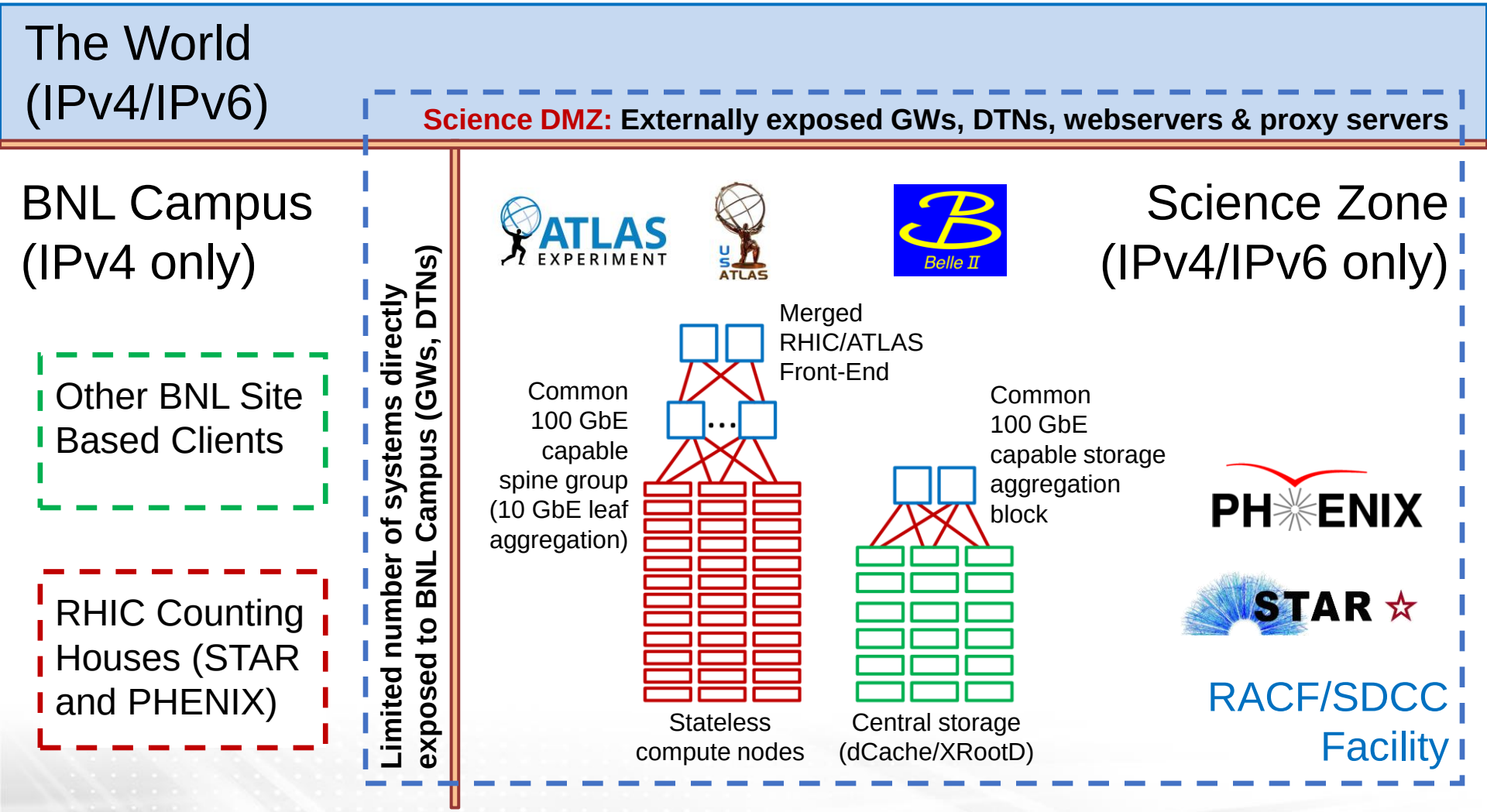
Other BNL Site Based Clients

RHIC Counting Houses (STAR and PHENIX)

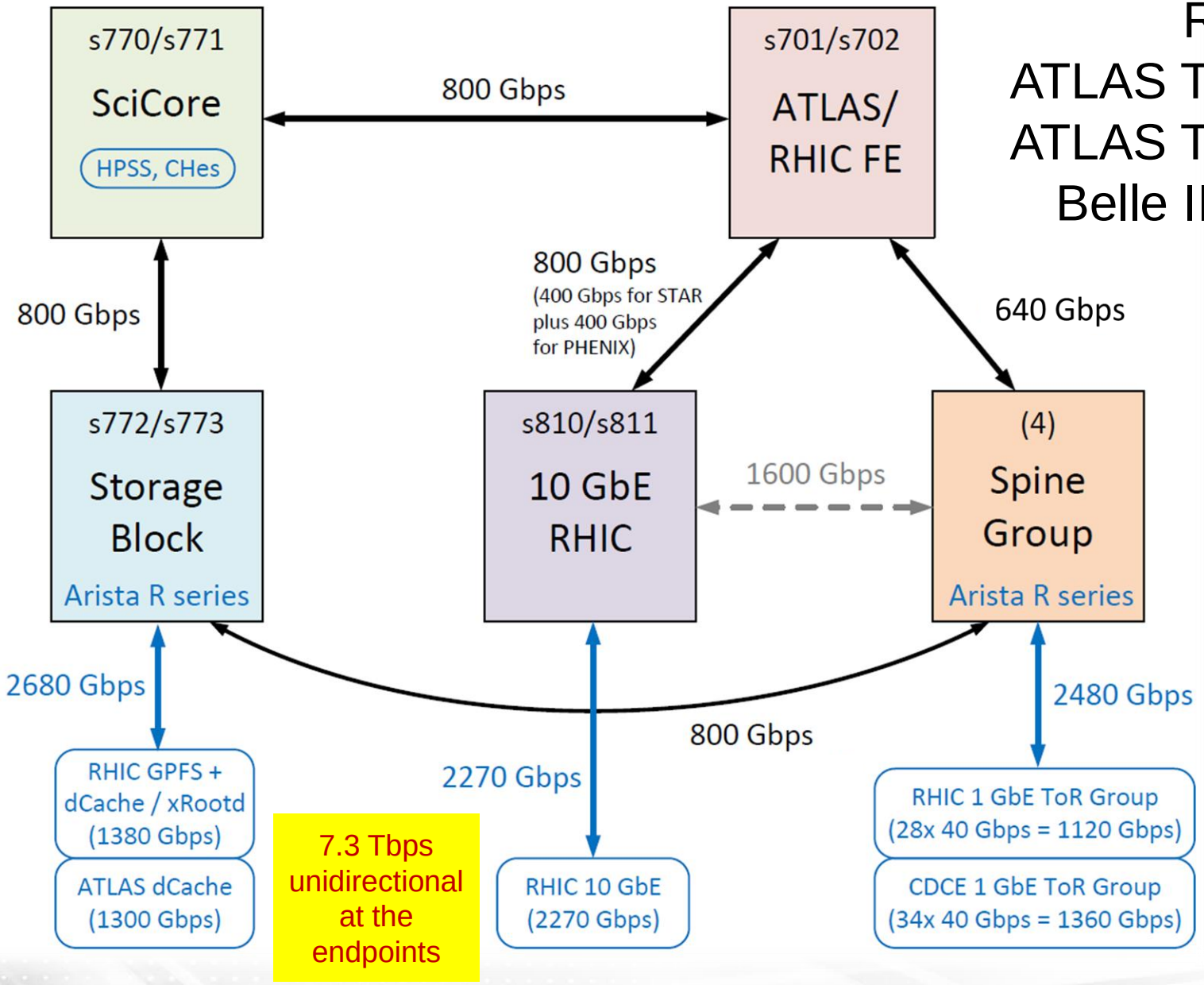
Science DMZ: **Externally exposed GWs, DTNs, webservers & proxy servers**



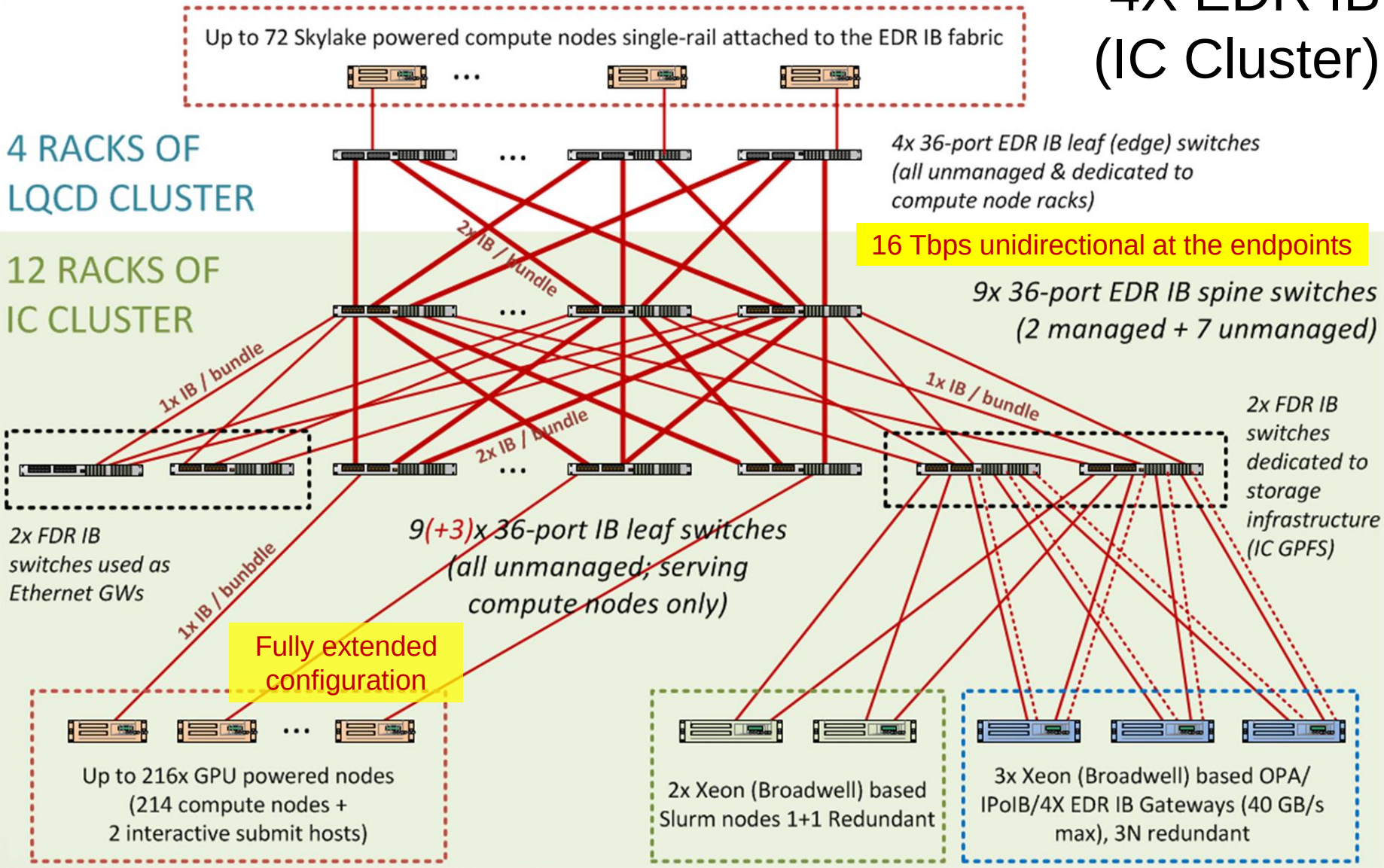
Overall Architecture Since 2018Q2



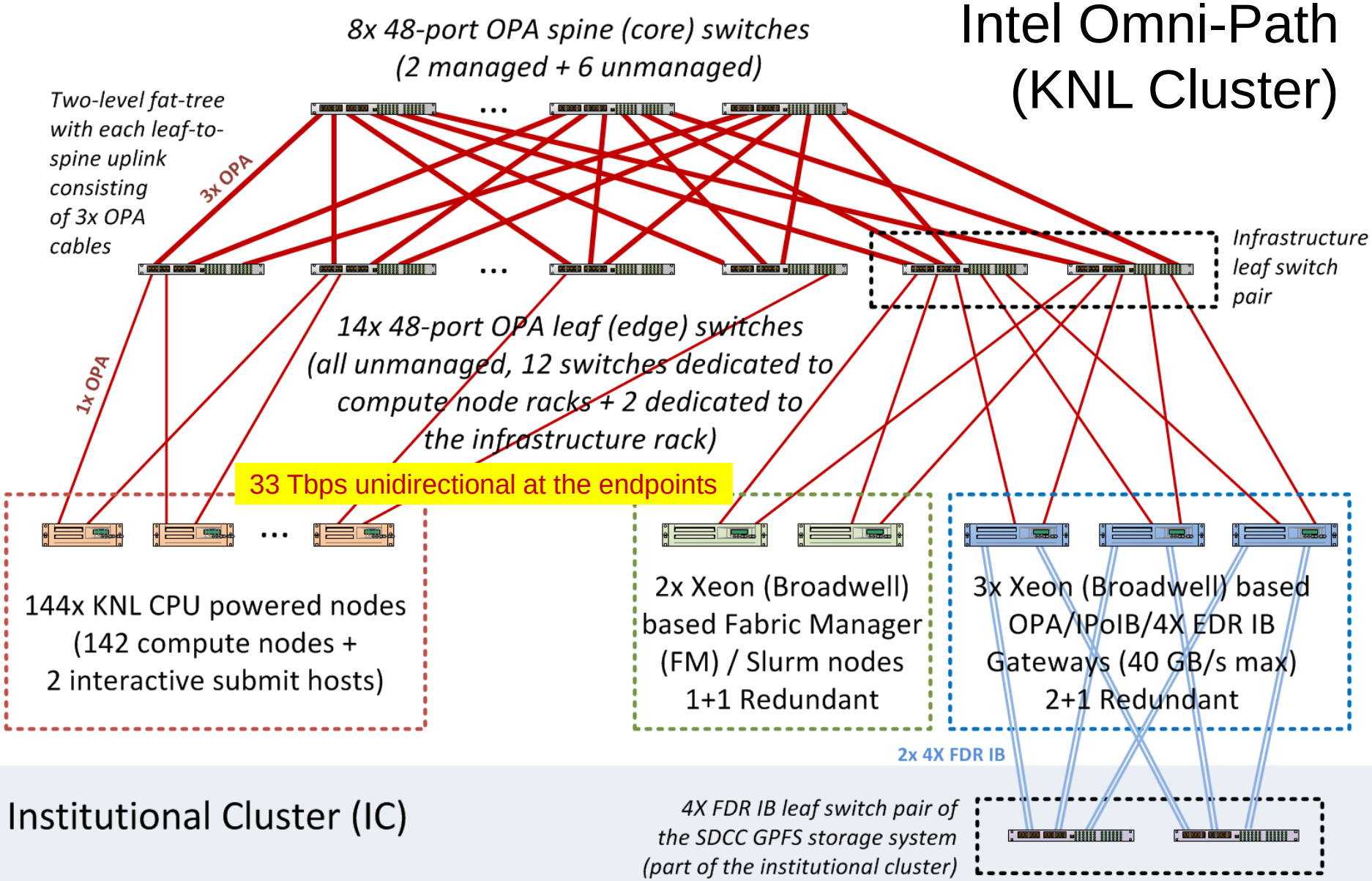
RHIC & ATLAS Tier-1 + ATLAS Tier-3 + Belle II Tier-1



4X EDR IB (IC Cluster)

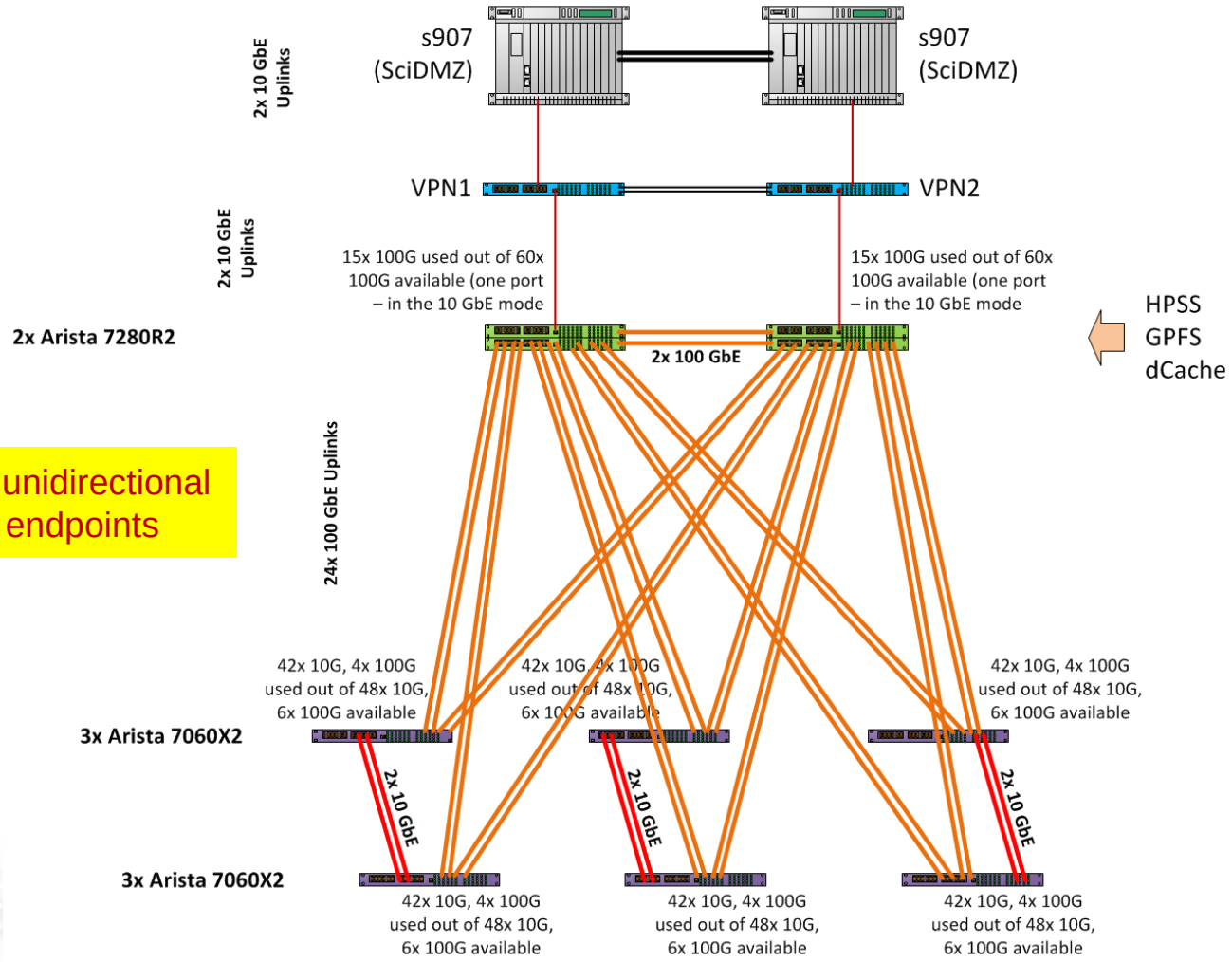


Intel Omni-Path (KNL Cluster)



Institutional Cluster (IC)

The Highest Ethernet Bandwidth Per Compute Node
Deployed for Simons Foundation in 2018: 20 Gbps Per Node



2.3 Tbps unidirectional at the endpoints

Client Connectivity Bandwidth Range

- 100 Mpbs Ethernet (monitoring systems only) [<1 Gbps/node]
  - Copper only (RJ45)
- 1-4x 1 Gbps Ethernet [4 Gbps/node]
  - Mostly over copper (RJ45, HDE – being phased out)
- 1-4x 10 Gbps Ethernet (HPSS, most of the DTNs) [10 Gbps/node]
  - Mostly over MMF fiber (SFP+), SFP+ direct attach 10 GbE copper
- 1x 56 Gbps IPoIB/4X FDR IB (IC/LQCD cluster, IC Cluster GPFS) [56 Gbps/node]
  - Copper/fiber, QSFP
- 1x 100 Gbps IPoIB/4X FDR IB (IC/LQCD cluster) [100 Gbps/node]
  - Copper/fiber, QSFP
- 2-4x 25 Gbps Ethernet (ATLAS/Belle II/RHIC dCache & XRootD, new GPFS) [100 Gbps/node]
  - Fiber only (SFP28)
- 1-4x 40 Gbps Ethernet (previous generation GPFS, HPSS, special DTNs) [160 Gbps/node]
  - Fiber only (MTP)
- 2x 56 Gbps IPoIB/4X FDR IB + 2x 40 GbE (BNLBox) [192 Gbps/node]
  - Copper/fiber, QSFP
- 2x 100 Gbps IPoOPA/OPA Gen.1 (KNL cluster) [200 Gbps/node]
  - Copper only, QSFP
- 2x 100 Gbps IPoOPA/OPA Gen.1 + 4x 56 Gbps IPoIB/4X FDR IB (KNL/IC/LQCD cluster OPA/IB gateways) [424 Gbps/node]
  - Copper/fiber, QSFP

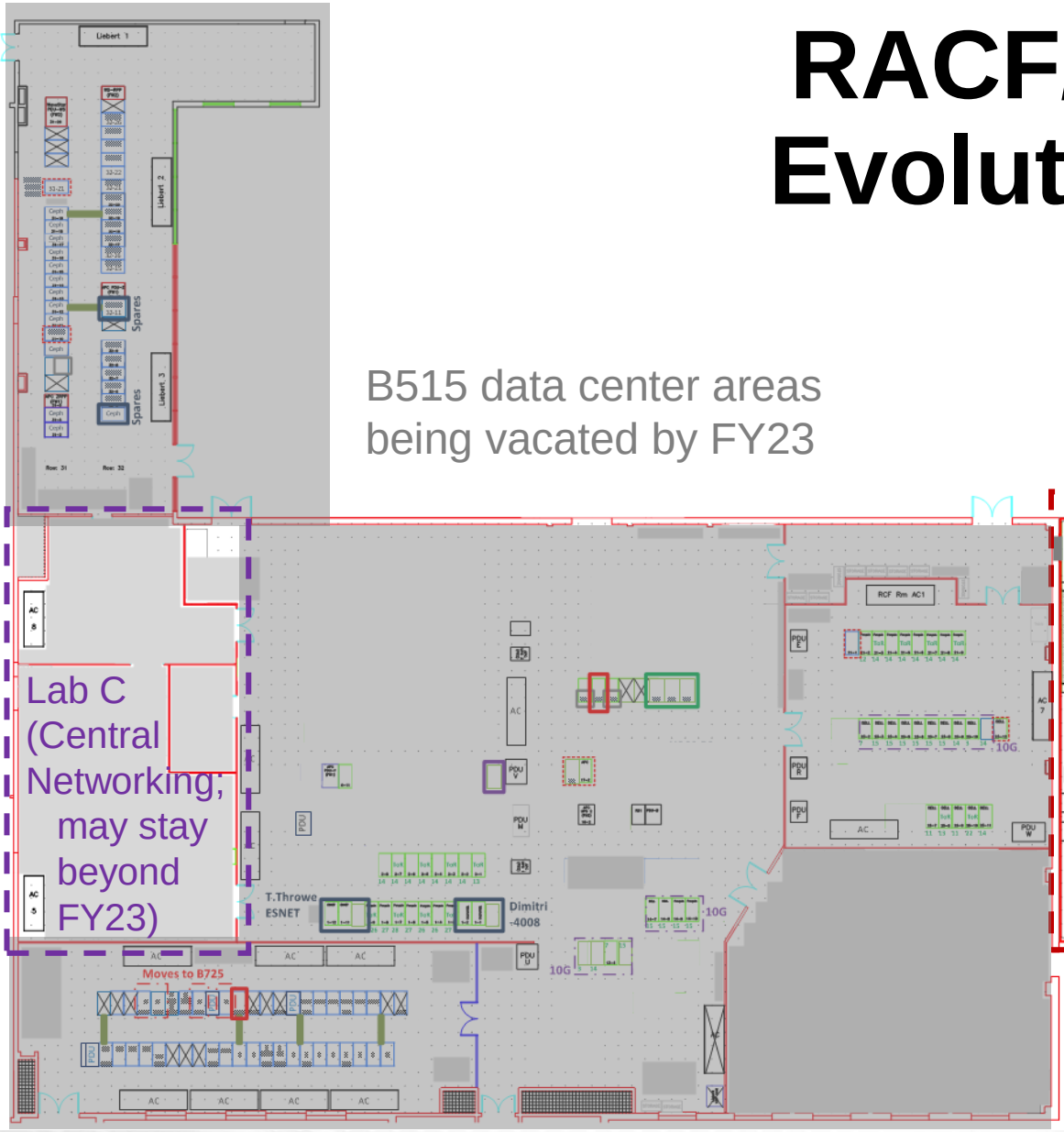
RACF/SDCC Facility Extension in FY20-21: Adding B725 Data Center

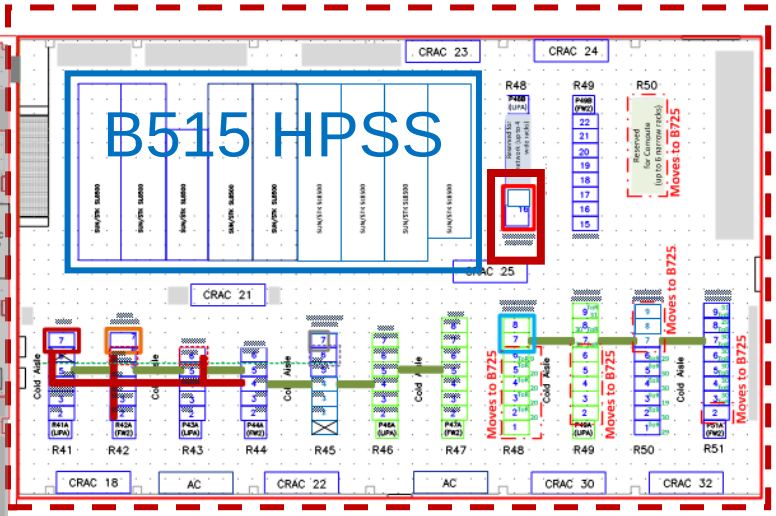
- New Arista 7500R equipment based Science Cores are to be established in both B515 (CDCE) and B725 (dedicated network room) data centers serving as merged ATLAS/RHIC/Belle II Front-Ends on each side and providing an inter-building link capacity of at least 1.2 Tbps (LR range)
  - A CWDM ring based solution for the inter-building link can also become viable should the inter-building link capacity grow up to 3-4 Tbps and or should some other BNL Campus facility joins the RACF/SDCC data center group
- SciDMZ connectivity is going to be extended from B515 to B725 over the inter-building link (400 Gbps initial capacity) in FY20-21
- Dedicated Arista 7500R equipment based spine group and storage block are to be established in B725 (dedicated network room)
- The initial deployment of equipment on the floor of B725 data center in FY21 is expected to be dominated by ATLAS, Belle II, RHIC compute node racks
- The central storage remaining in B515 will be replaced gradually by new storage in B725 over FY21-24 period
- The HPSS system is expected to stay split between B515 and B725 in FY21-23

RACF/SDCC Facility Evolution in FY21-23

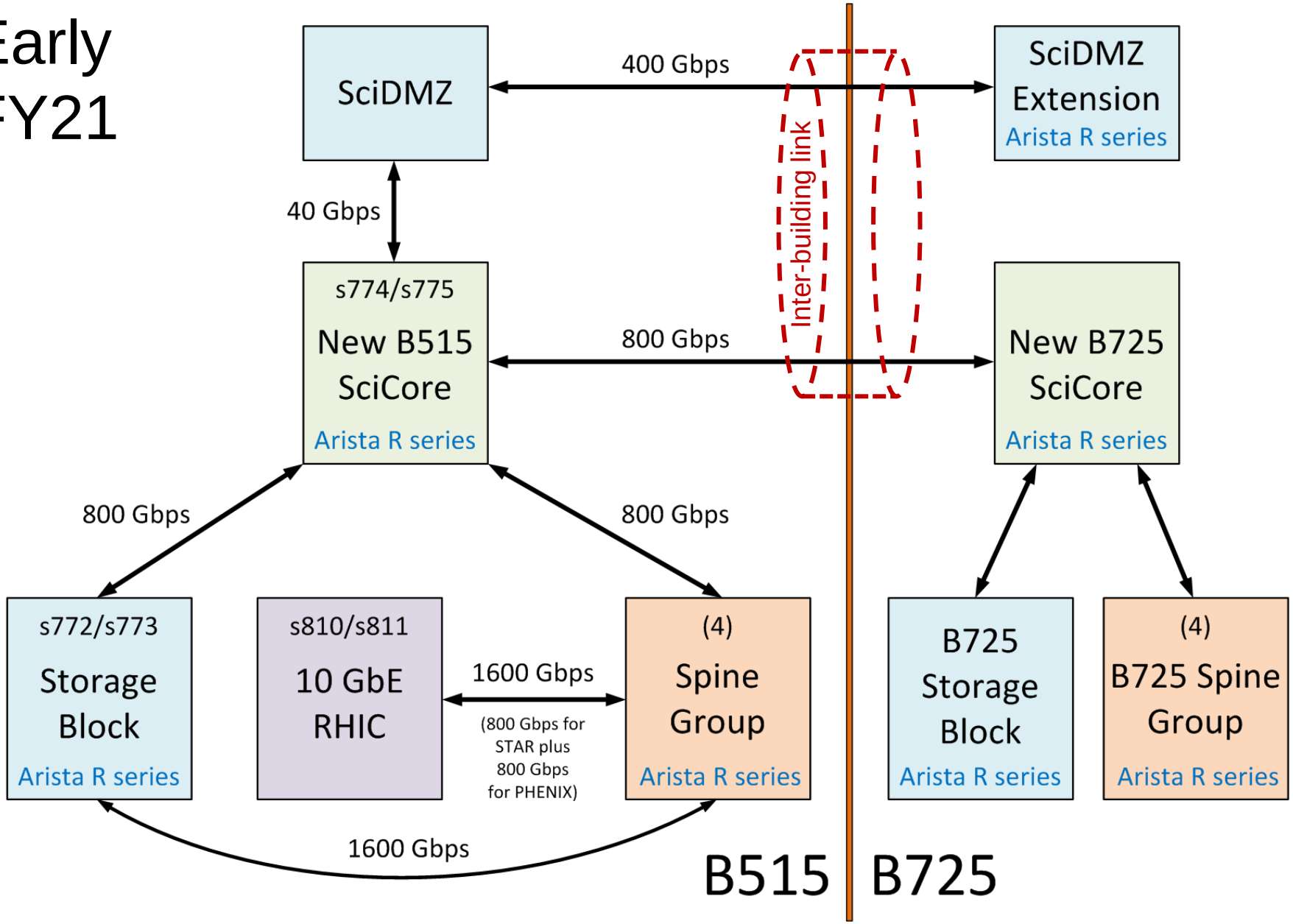
B515 data center areas being vacated by FY23

B515 data center consolidated to CDCE (FY23)

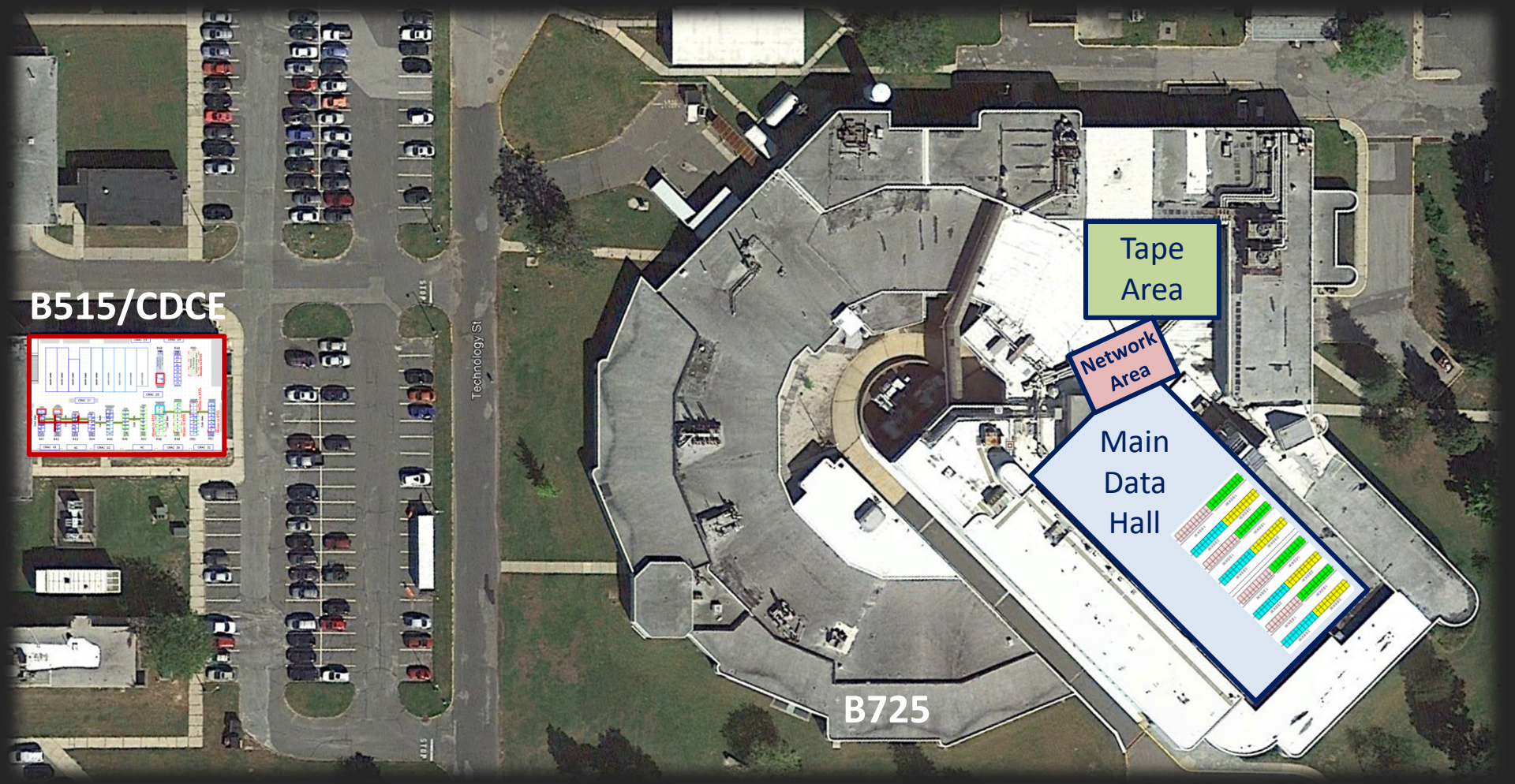




Early FY21


SciDMZ
400 Gbps
SciDMZ Extension
Arista R series
Inter-building link
40 Gbps
s774/s775
New B515 SciCore
Arista R series
800 Gbps
New B725 SciCore
Arista R series
800 Gbps
800 Gbps
s772/s773
Storage Block
Arista R series
s810/s811
10 GbE RHIC
1600 Gbps
(800 Gbps for STAR plus 800 Gbps for PHENIX)
(4)
Spine Group
Arista R series
B725 Storage Block
Arista R series
(4)
B725 Spine Group
Arista R series
1600 Gbps
B515
B725

Addition of B725 Data Center in FY21



Summary

- Summary of the current state and ongoing transitions related to the network systems of RACF/SDCC data center is given
- The architecture of the network systems of the Science Core and its current utilization is overviewed
- Expected evolution of the RACF/SDCC data center in FY21-23 period is highlighted

Questions & Comments