Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

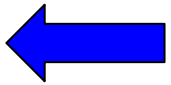
SDCC - Architecture

Shigeki Misawa
Scientific Data and Computing Center

March 7, 2024

@BrookhavenLab

SDCC - Architecture

- The first in a potential, multi-presentation series. If desired, content of future talks are as follows:
  - **Architecture**  **This Talk**
  - Overview of services
  - Stakeholder support
  - Storage services
  - Compute services
  - Groupware and software deployment services
  - Facility support (“Infrastructure”) services

Not Your Traditional Data Center

- Founded (1990's) to provide offline compute/storage for the experiments at RHIC
  - Direct access from experiment "DAQ" to tape library at the data center
  - Dedicated compute farm for event "reconstruction", as well as physics analysis
- Subsequently tasked with supporting ATLAS, an HEP experiment at the Large Hadron Collider (LHC)
  - Part of the ATLAS world wide computing grid
  - Provide "Grid" accessible compute and data storage resources
- SDCC - Scope expanded to offer services to other groups

<u>Data intensive computing is in SDCC's DNA</u>

Next Gen Data Intensive Research

- FY2013 - SDCC anticipated explosion of data intensive experiments at BNL
  - 10x increase in data volume/rate from next generation NP (sPHENIX) and HEP High Luminosity-LHC experiments (HL-ATLAS)
    - 100s of petabytes generated per year
  - Proliferation of data intensive experiments from new groups (e.g. NSLS-II, CFN)
    - Smaller and more numerous than NP/HEP collaborations
    - Geographically dispersed in the BNL campus
    - Limited infrastructure to support in-place compute/storage resources

SDCC Data Center

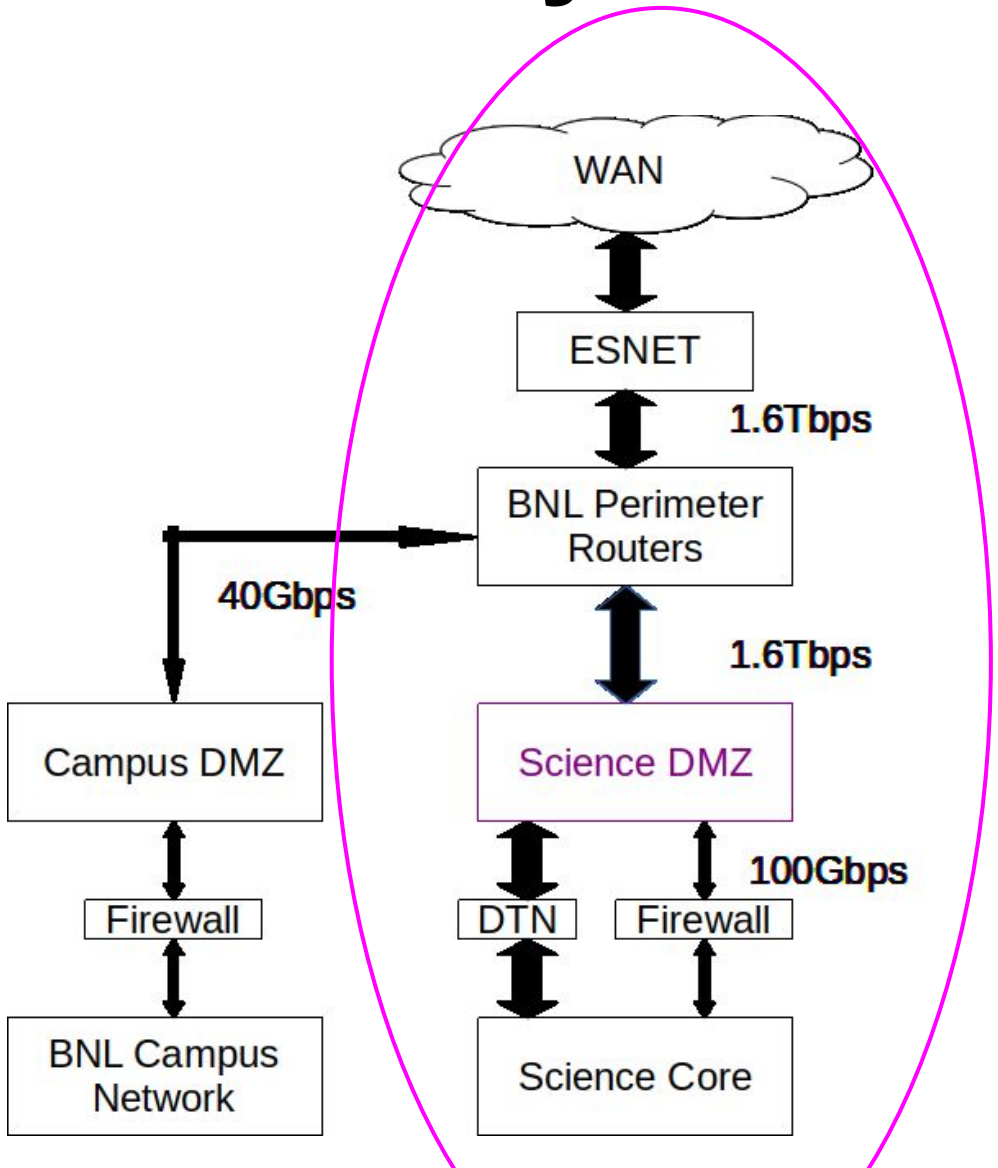
- New, highly available data center capable of supporting compute and data storage resources required by next generation data intensive research
  - HPC/HTC compute farms
  - Specialized compute systems
  - Scalable storage (performance and capacity)
  - High bandwidth WAN connectivity
- Architecture of the SDCC is specifically designed to allow data intensive experiments to directly access selected resources in the data center

Supporting Data Intensive Research

- FY13-FY15 High Throughput Science Network (HTSN) architecture, developed to support data intensive research.
  - Science DMZ
    - Termination point for high bandwidth WAN connectivity
  - Science Core
    - High bandwidth, “frictionless” network for scientific data within the BNL campus
    - Can connect scientific instruments directly to data center compute/storage resources
    - Also interconnects compute/storage resources within the data center

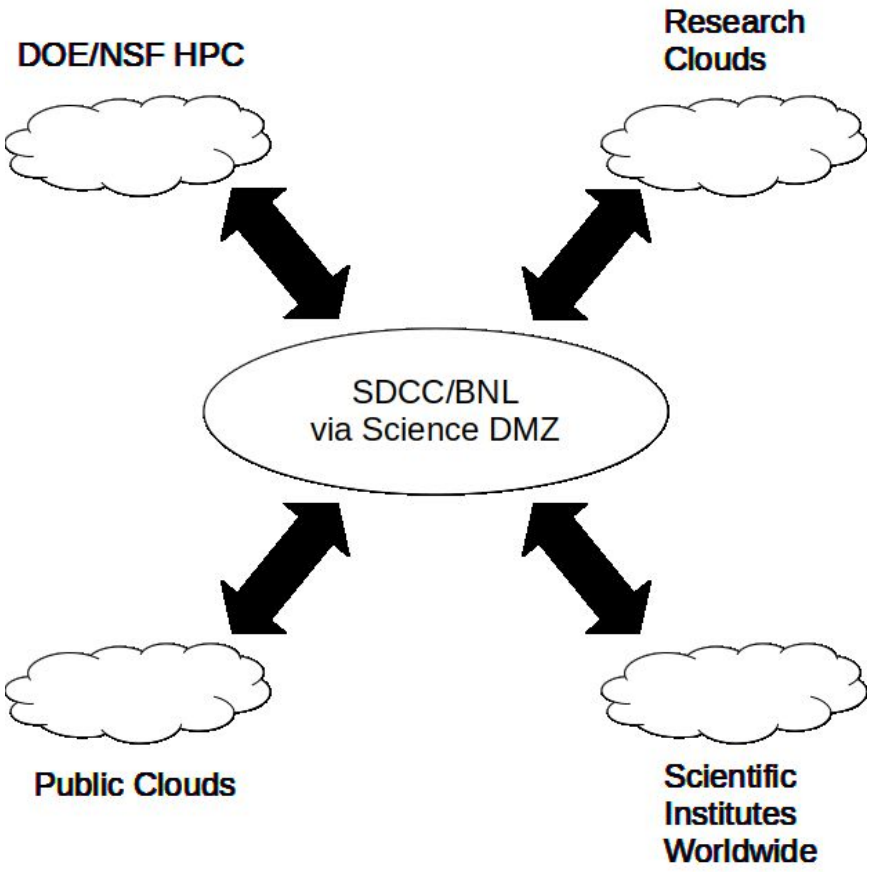
Science DMZ WAN Connectivity

- Science DMZ and Core are completely decoupled from the BNL campus network
  - Science and campus connect only at the BNL perimeter
- Science DMZ supports IPv4 and IPv6
  - Critical for international groups
- Dedicated 100 Gbps firewall protects Science Core network from WAN
- DTN's on the DMZ enable high bandwidth (100s of Gbps) data transfers to/from the WAN



Science DMZ Rationale

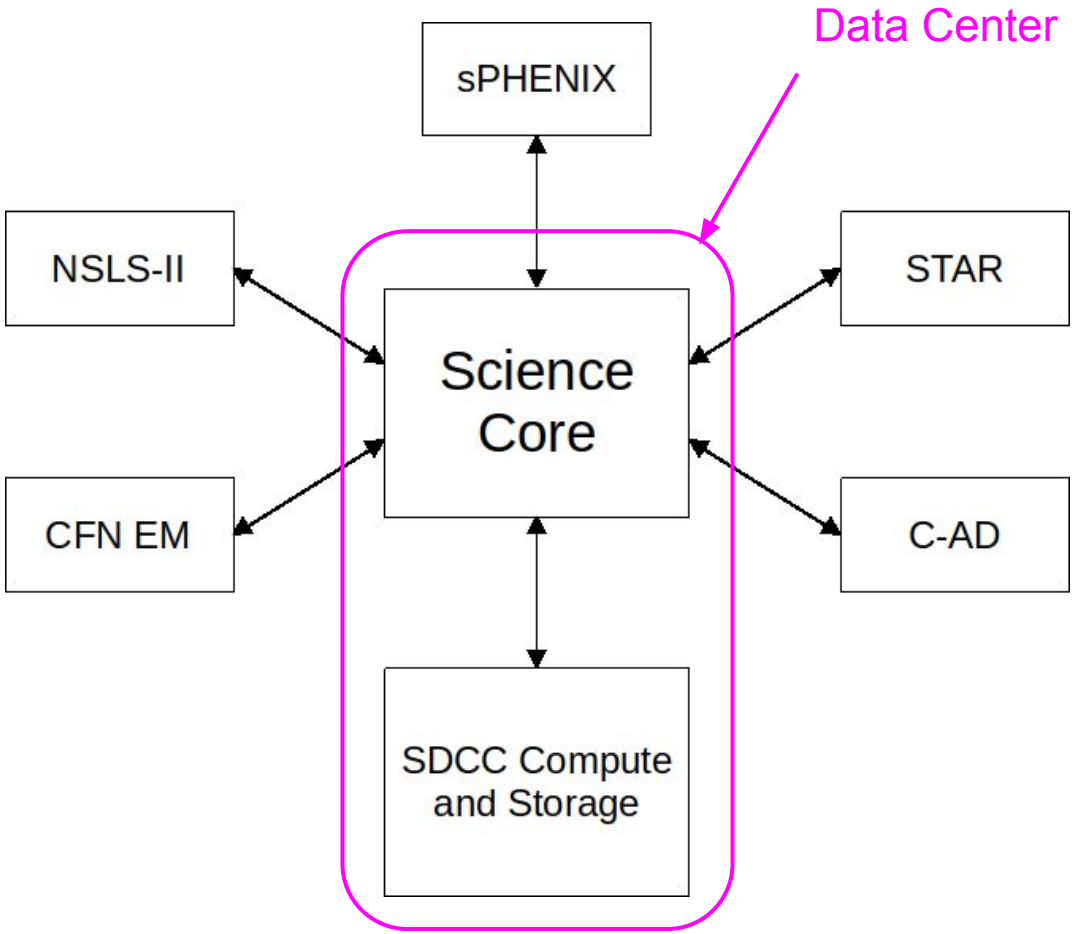
- High bandwidth WAN connectivity is needed to access resources outside of BNL
  - e.g. DOE/NSF HPC facilities, public cloud, other institutes
- Computing trends are making access to external resources more important
  - Proliferation of new services
  - Increasingly targeted hardware resources, e.g. GPU systems
- Science DMZ is heavily used by ATLAS and other HEP programs. Use by EIC experiments is expected in the future



Science Core (“External” Connectivity)

- Connects selected subnets in the SDCC to sites at BNL outside of the data center
  - sPHENIX
  - STAR
  - NSLS-2
  - CFN Electron Microscopy
  - C-AD
- Network bandwidth is limited by link speed, up to 100 Gbps [1], and # links connecting endpoints

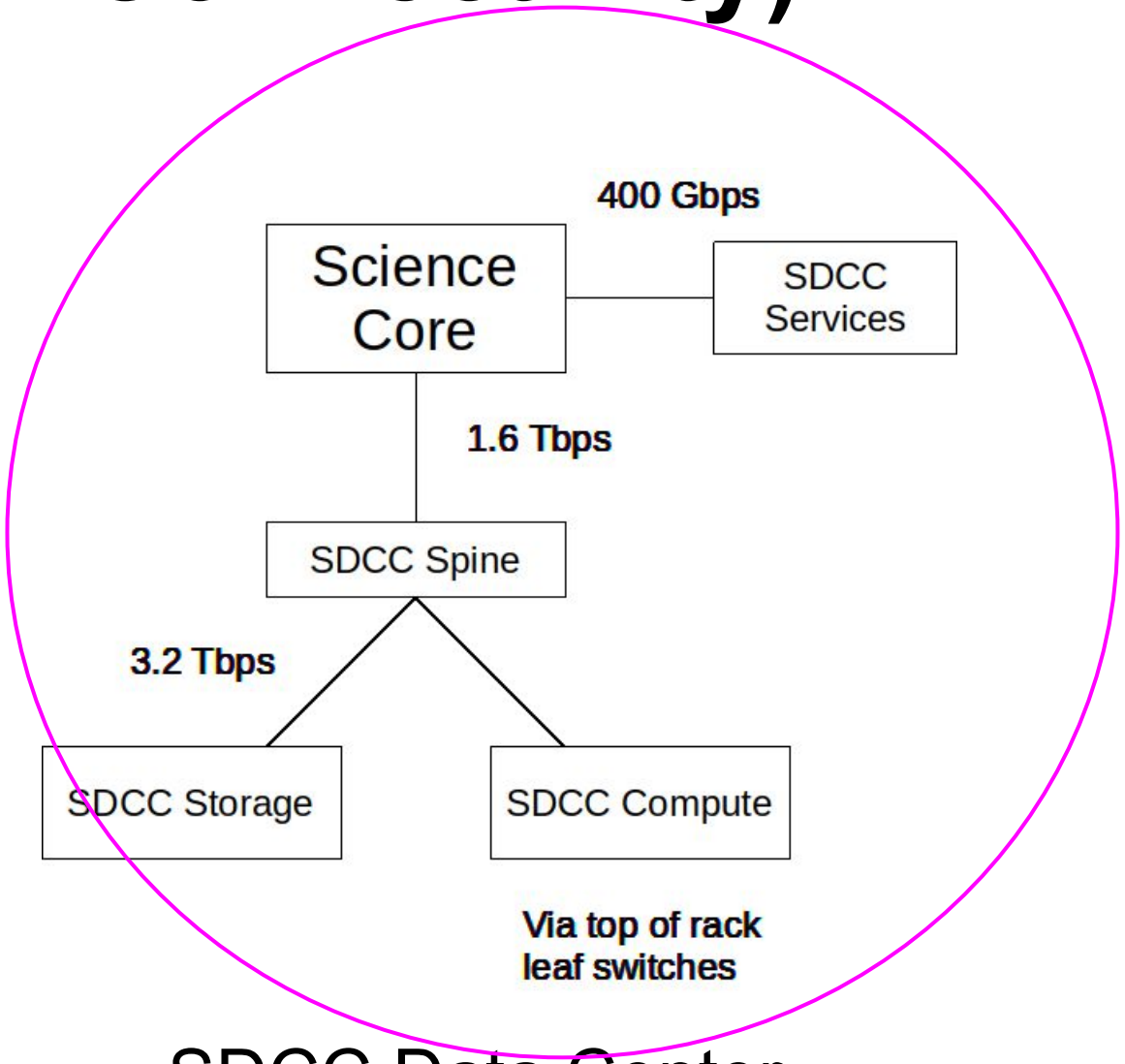
[1] Upgradable to 400 Gbps per fiber pair



External Connectivity

Science Core (Internal Connectivity)

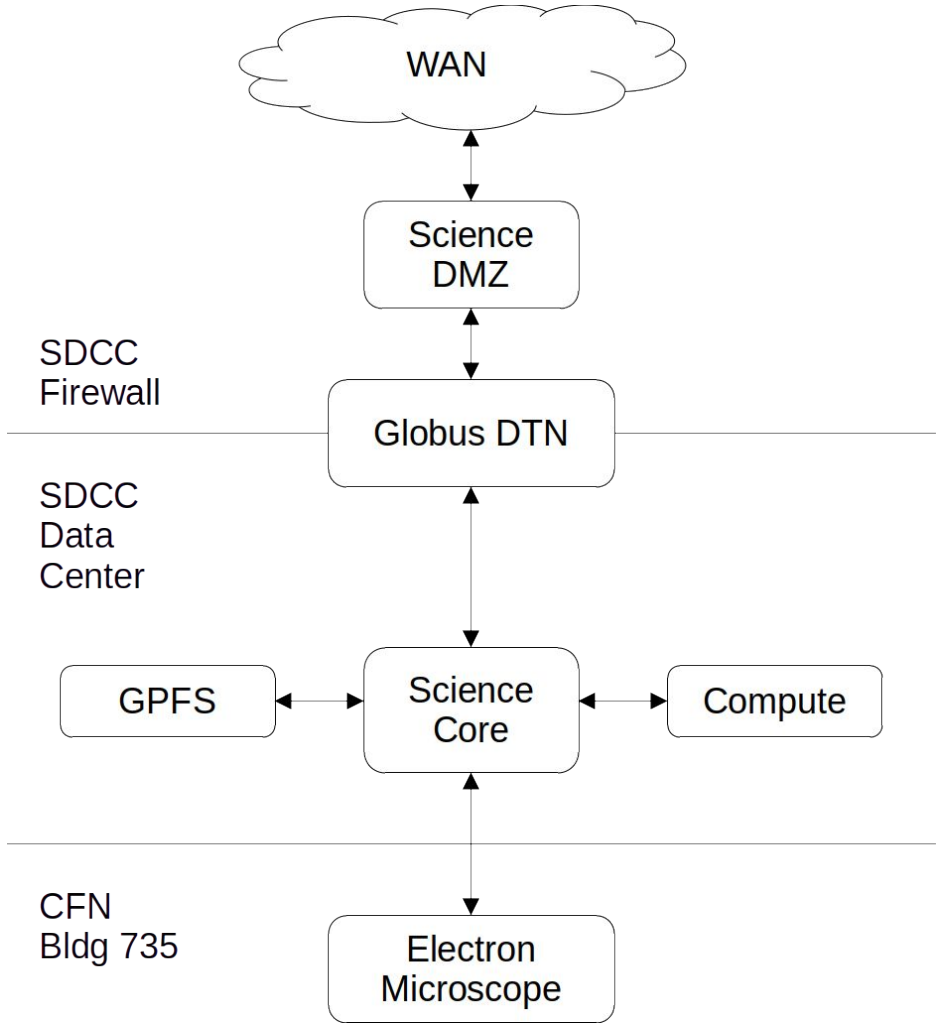
- Provides full connectivity between resources inside the data center.
  - Data flows between storage and compute are isolated within the SDCC spine and leaf network
  - Data flows to other, mostly lower bandwidth SDCC services routed through Science Core
- SDCC internal network is mostly IPv4 with gradual introduction of dual stack IPv4/IPv6 just starting (for ATLAS)



SDCC Data Center

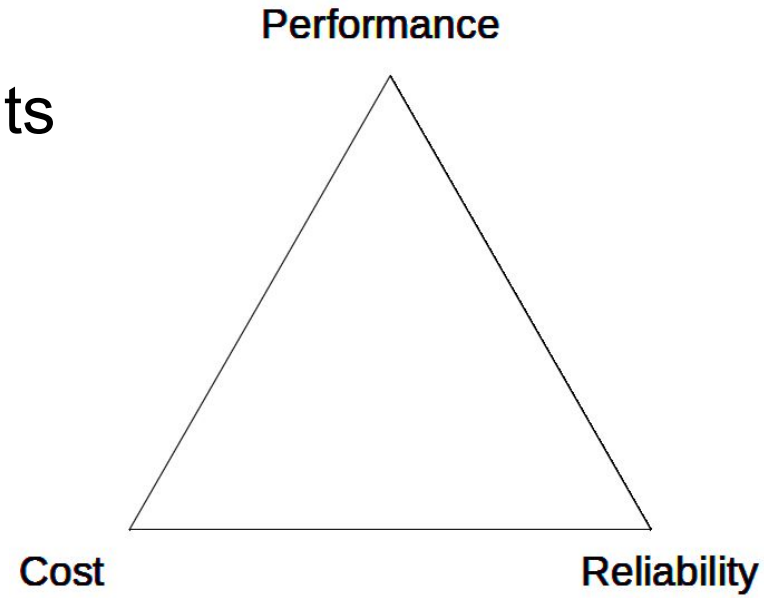
Science Core Proof of Concept (~ FY16)

- CFN E-TEM
  - Direct writes to GPFS at SDCC from CFN systems
  - Analysis of GPFS resident data on BNL Institutional Cluster
  - WAN transfers via sftp/Globus



Remote Access to Storage Now in Production

- Direct access to data center storage through Science Core from remote sites now widespread
- Storage provided varies depending on requirements
  - Industry standard file sharing (NFS/SMB)
    - NSLS-II, CFN EM
  - Scale out parallel file systems (disk)
    - sPHENIX, CFN EM, NSLS-II
  - Scale out bulk data storage (disk)
    - ATLAS (LAN/WAN)
  - Nearline and archival storage (tape)
    - sPHENIX, STAR, ATLAS



Requirements

CFN Model in production

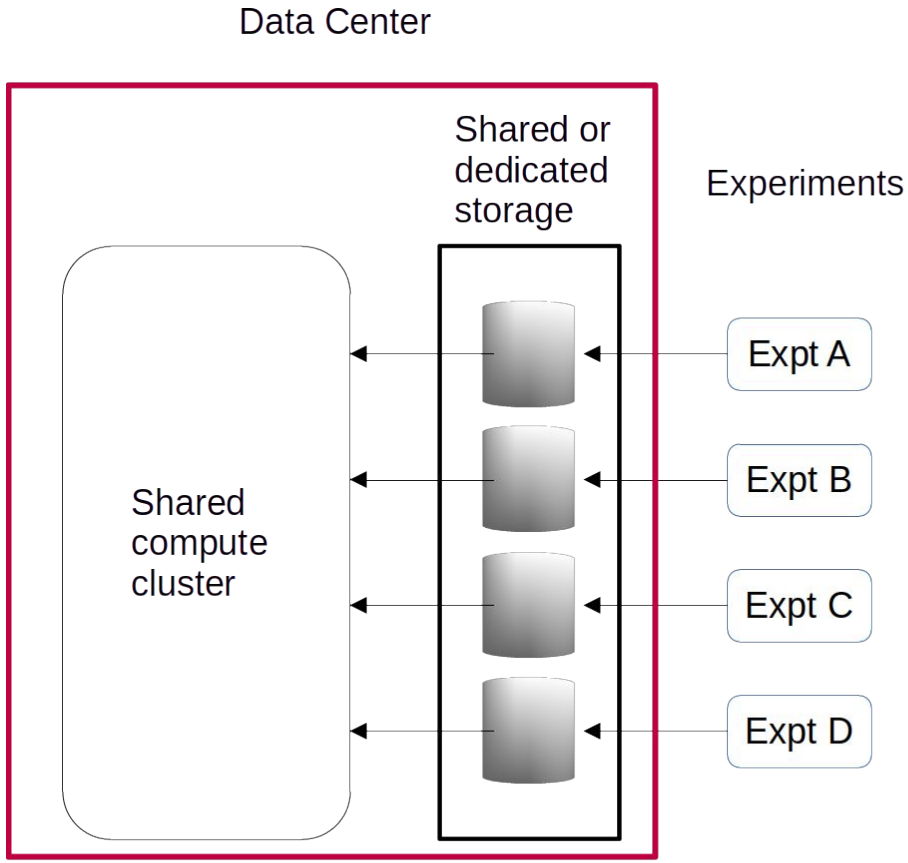
- sPHENIX uses dedicated compute resources to process Lustre resident data written directly from the experiment hall into Lustre
  - Close coupling of “DAQ”, Lustre, and data center compute critical for their streaming DAQ system.
- Work in progress at CFN to connect their latest high frame rate imaging system (~40GB/sec data rate) to SDCC based storage resources
  - Include possibility to streaming data directly to compute resources, bypassing storage

Other Uses of Science Core

- Following set of slides cover other potential uses of the capabilities of the Science Core
- Applicability of these configuration to BNL research is a question for the Technical Advisory Board
- Other configurations are possible, limited by the imagination
  - However, all imaginable configuration aren’t necessarily practical.

Shared Computing Resources

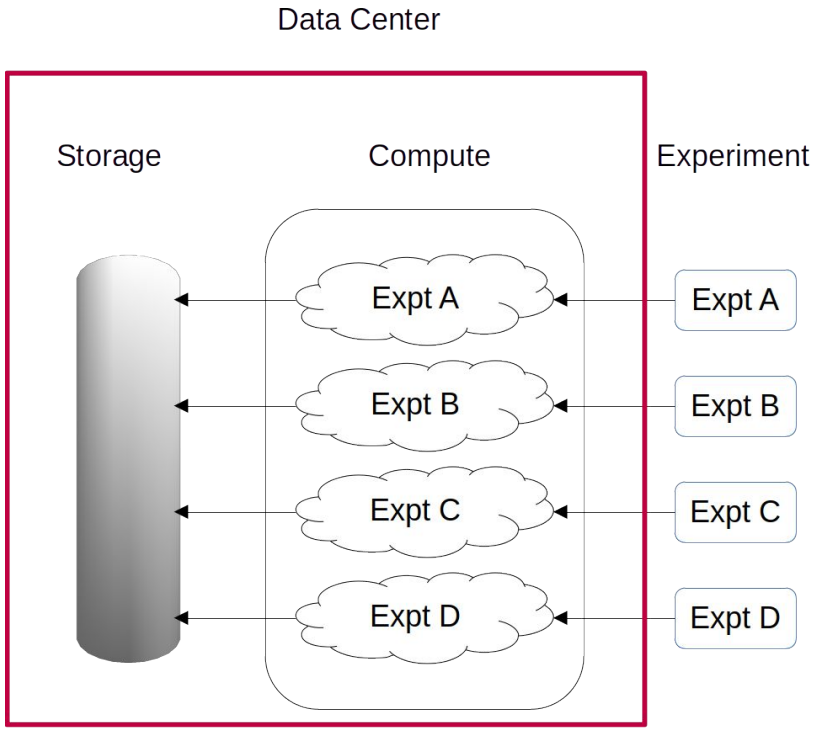
- Replication of CFN model
  - but with shared resources
- Potential reduction in costs
  - Particularly if individual experiment “duty cycles” are low
  - Significant commonality in compute requirements necessary
  - Simplified software infrastructure
  - With many experiments, can achieve “critical mass” making more diverse resources affordable
  - Can be refined/enhanced with R&D (ASCR or LDRD?)



Shared compute cluster in the data center
Relatively simple architecture. Details in configuration has an impact on the capabilities of the system

Configurable Private Clouds ?

- Guaranteed, dedicated resources
- Tailored to experiment specifics
- May spread costs over more experiments
  - Allows for access to more and more diverse computing equipment
- Enhanced isolation might be possible within VM or container frameworks
  - Requires R&D (ASCR?)
  - May be too costly, cumbersome or unnecessary

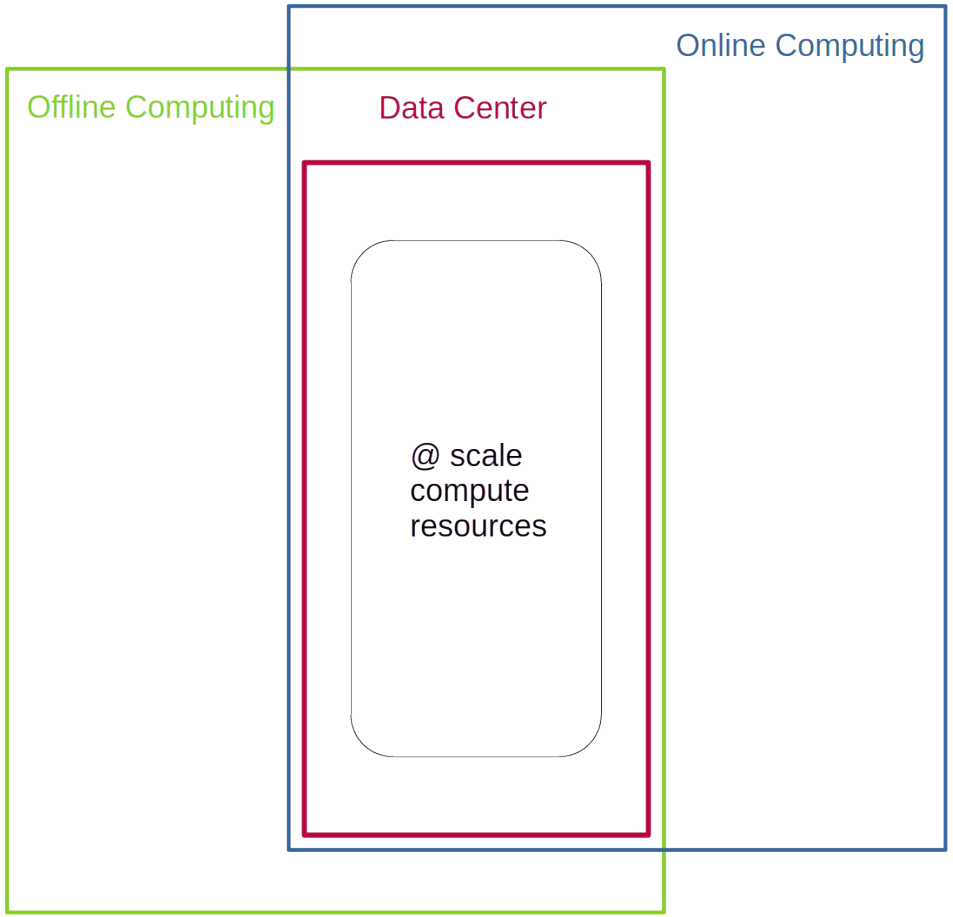


Multi-Tenant Enterprise Cloud in the data center

Resources for each private cloud drawn from a potentially heterogeneous mix of compute nodes, e.g, CPU only, GPU-lite, GPU-heavy, large cache, large memory, fast CPU cores, more cpu core

Retaskable Computing

- “Move” data center computing resources between offline and online computing depending on collider operations
- May be of use if significant online resources are needed during collider operations
- Limited # configuration changes per year (two per year is doable)
- Marginal cost to support maximum isolation decreases with scale
- Bandwidth requirements to remote “online” resources may impact viability



Configure data center compute resources for online computing during collider operation and offline compute during shutdown periods.

Access to Other SDCC Services

- Science Core enables groups to access other SDCC services beyond compute and storage
  - Data transfer and management
    - Globus Connect Server
    - sFTP DTN for simple, infrequent data transfers, albeit at low bandwidth
    - Rucio/FTS based data transfer and management
  - Long term and medium term archival storage
  - Web based data analysis (e.g. Jupyter notebook)
- An overview of the portfolio of SDCC services is the subject of a future presentation

Things to Consider

- Utilization of SDCC based resources from outside the data center requires single mode fiber from the experiment site to Bldg 725
- Science Core supports 100 Gbps/fiber pair and with line card upgrades can support 400 Gbps/fiber pair
- Public clouds are driving down costs and are actively migrating to 400GbE
  - Cost may be lower than one might imagine in the future
  - Data intensive science is likely to benefit from of this trend
- IEEE Projects for 800Tbps and 1.6Tbps Ethernet using 200Gbps signaling are active with completion expected between 2026 and 2030.

Concluding Remark

- Science Core and DMZ creates new opportunities
  - Allows computing resource to be brought closer to the experimental apparatus
  - May make resources available to research group that might not otherwise be able to afford if they were worked independently
  - Makes resources outside of BNL accessible to experiments
  - Potentially opens up avenues of research that would otherwise not be possible
- Mission of the SDCC is to partner with research group to enable data intensive research.