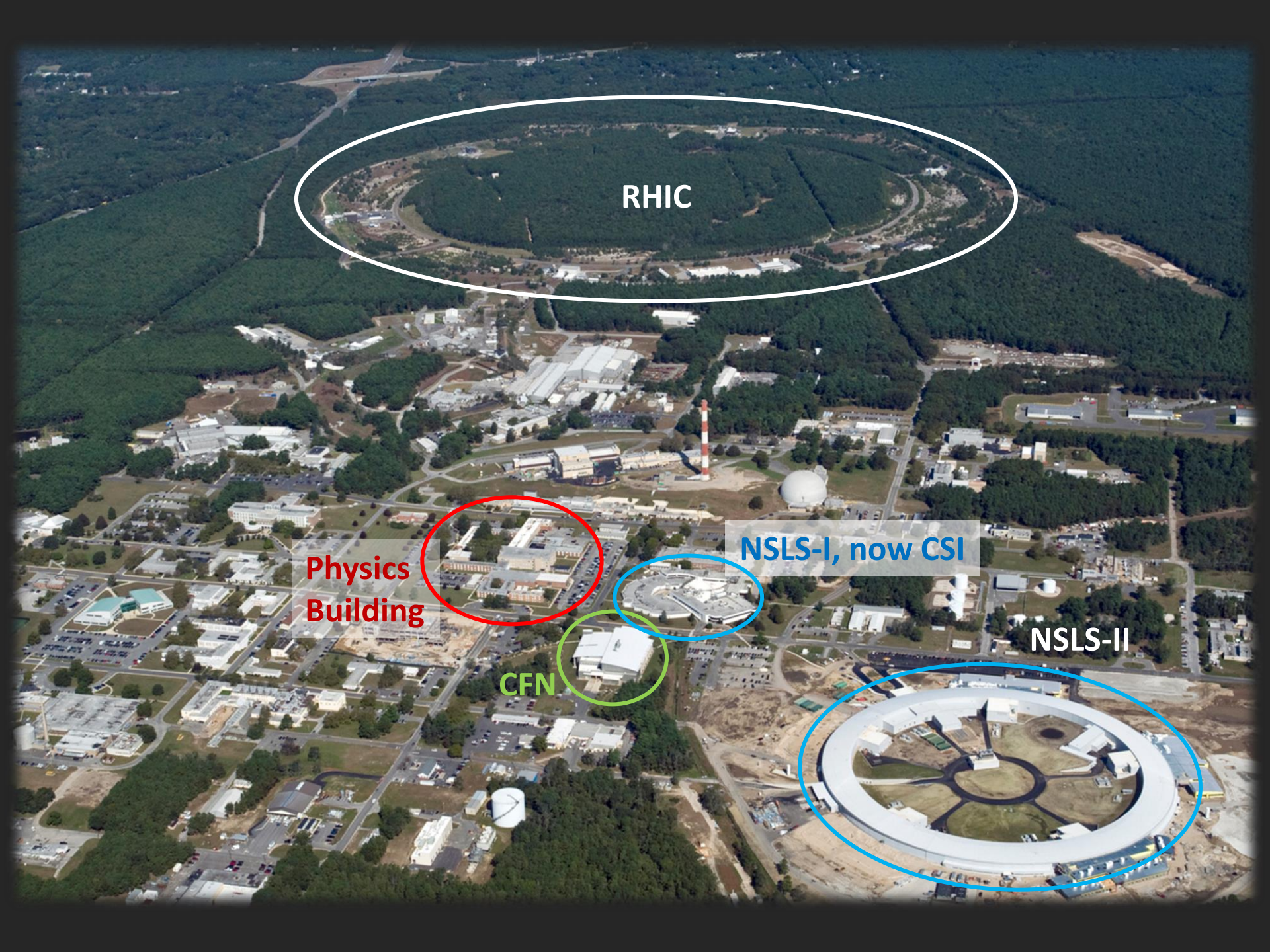
BNL SDCC/RACF Facility Overview

Alexandr Zaytsev

USQCD All Hands' Meeting

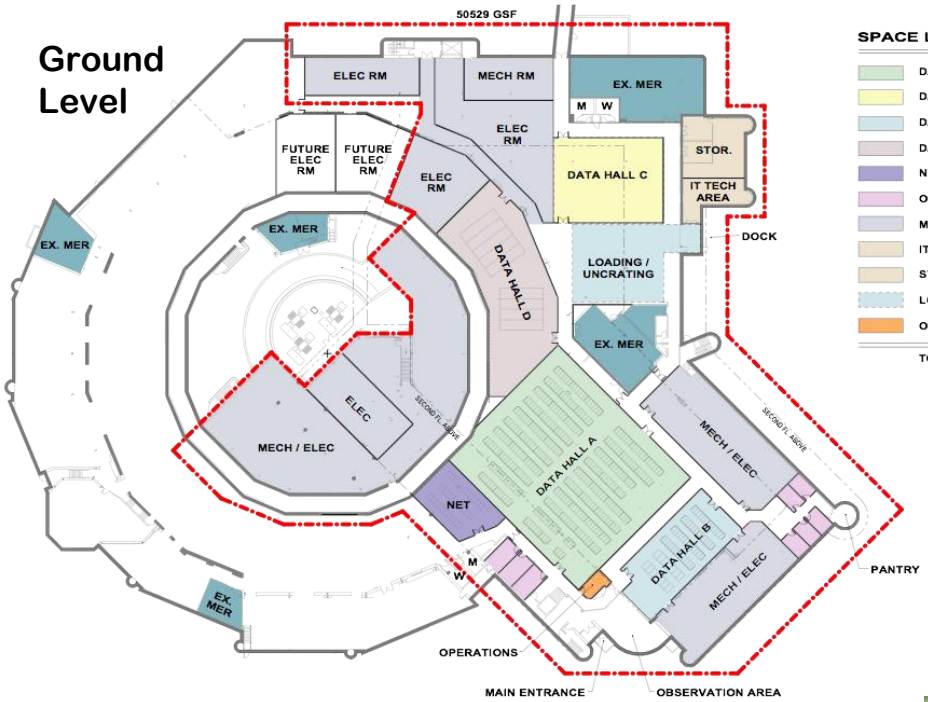
Jefferson Lab, April 28-29, 2017


RHIC
Physics Building
NSLS-I, now CSI
NSLS-II
CFN

SDCC/RACF at a Glance (1)

- Located at Brookhaven National Laboratory on Long Island, NY
- Until recently the RHIC & ATLAS Computing Facility (RACF) was predominantly HTC oriented and the HPC component including the IBM BlueGene/{L,P,Q} machines was outside of the scope of RACF activities
- Nowadays the RACF is the main component of the Scientific Data & Computing Center (SDCC) within BNL Computational Science Initiative (CSI)
  - Provides full service computing for two RHIC experiments (STAR, PHENIX), and for ATLAS (US Tier-1 Site), along with some smaller groups: LSST, Daya Bay, DUNE, EIC, etc.
  - Hosts the BlueGene/Q machine (end of service for USQCD Sep. 2017)
  - Hosts and operates several HPC clusters supporting research in LQCD, NSLS-2, Biology, Center for Functional Nanomaterial (CFN), etc. – including the new Institutional Cluster (IC) and KNL cluster
- New systems expected before the end of 2017:
  - Extension of the IC cluster & the new USQCD cluster

SDCC/RACF at a Glance (2)

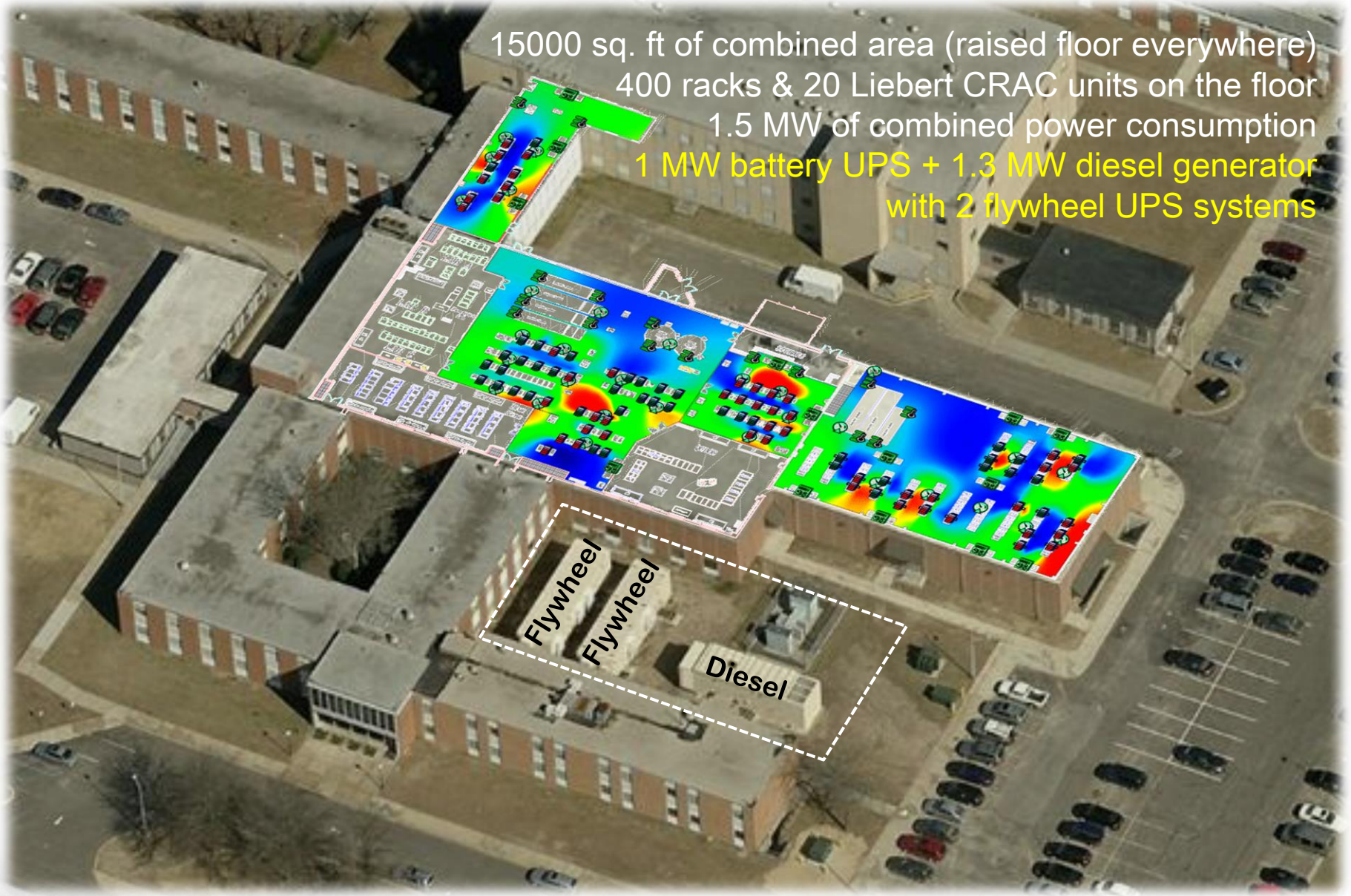


SPACE LEGEND	PROGRAM NET AREA (NSF)	DELINEATED AREA (NSF)
DATA HALL A	10800	9271
DATA HALL B	3450	2868
DATA HALL C	1950	2238
DATA HALL D	3600	4159
NETWORK EQUIPMENT	1000	1156
OFFICES	1680	1156
MECH/ELEC	22200	25017
IT TECH AREA	500	643
STORAGE	1200	1200
LOADING / UNCRATING	560	2579
OPERATIONS	150	133
TOTAL NET AREAS	48090	50529

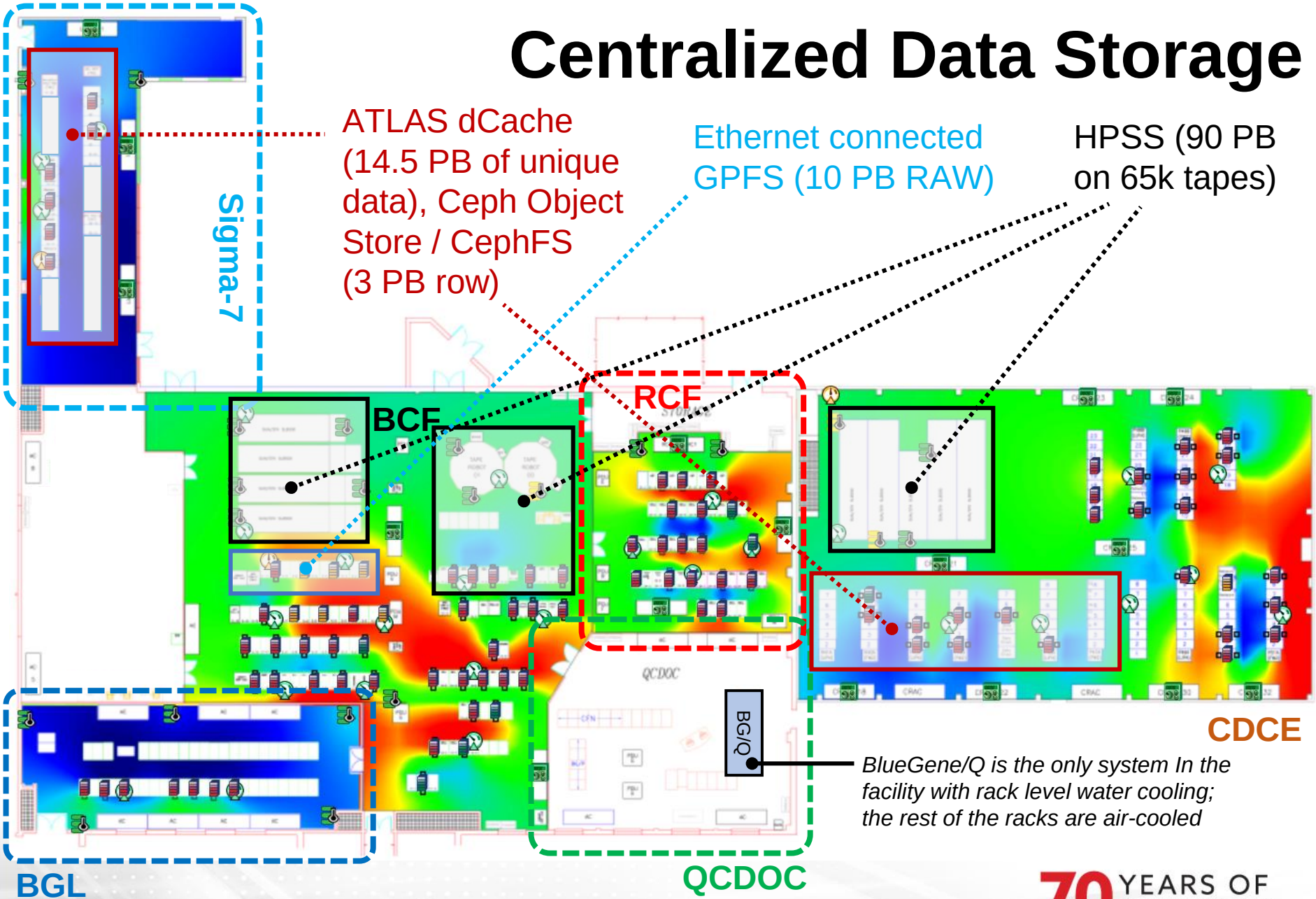
Phase I : 60% Space Plan, 3.6 MW

- Includes CSI office space already in use
- SDCC operational target is FY22

- CD-1 milestone approval received on Apr 17, 2017


15000 sq. ft of combined area (raised floor everywhere)
400 racks & 20 Liebert CRAC units on the floor
1.5 MW of combined power consumption
1 MW battery UPS + 1.3 MW diesel generator
with 2 flywheel UPS systems
Flywheel
Flywheel
Diesel

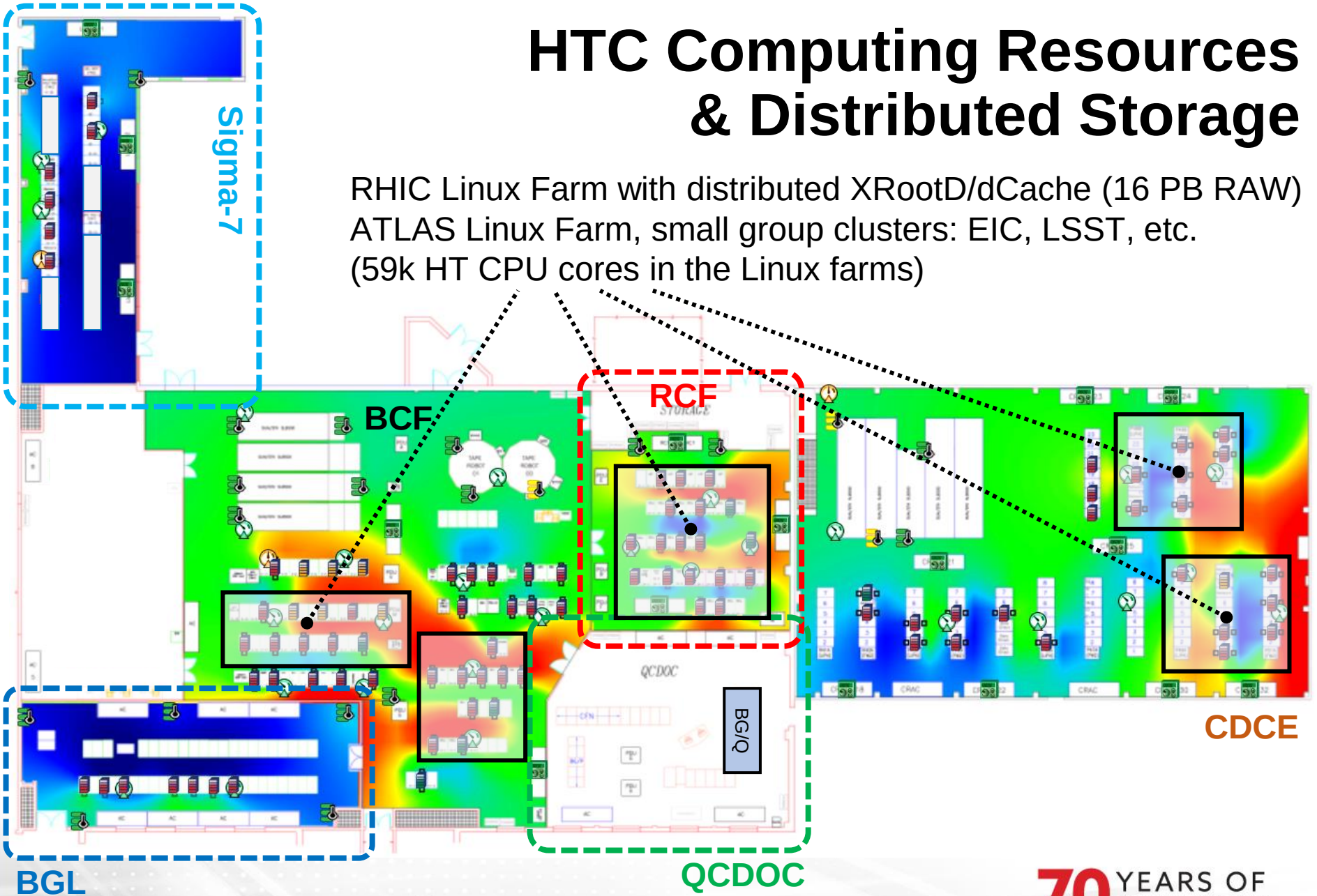
Centralized Data Storage



BlueGene/Q is the only system In the facility with rack level water cooling; the rest of the racks are air-cooled

HTC Computing Resources & Distributed Storage

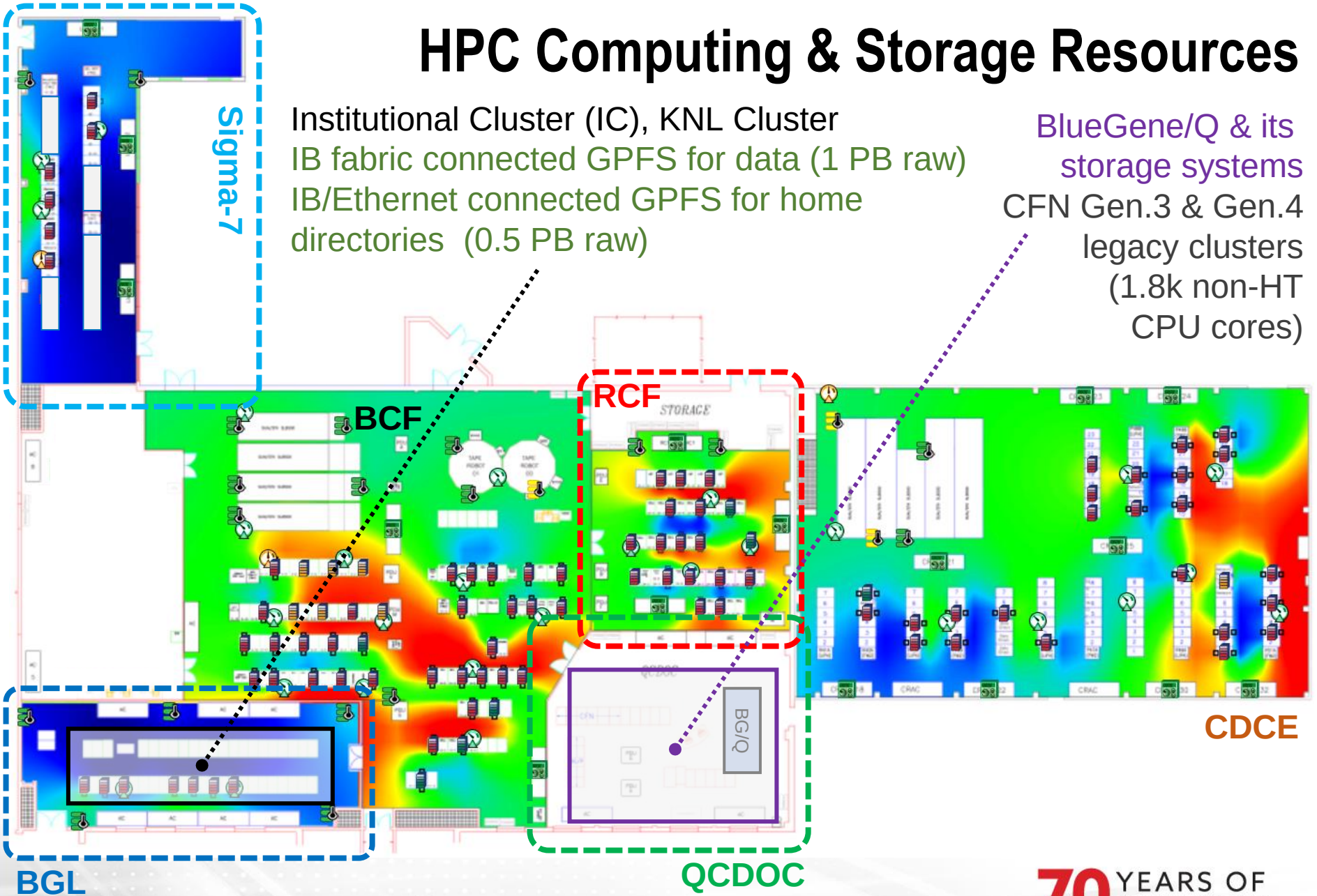
RHIC Linux Farm with distributed XRootD/dCache (16 PB RAW)
ATLAS Linux Farm, small group clusters: EIC, LSST, etc.
(59k HT CPU cores in the Linux farms)



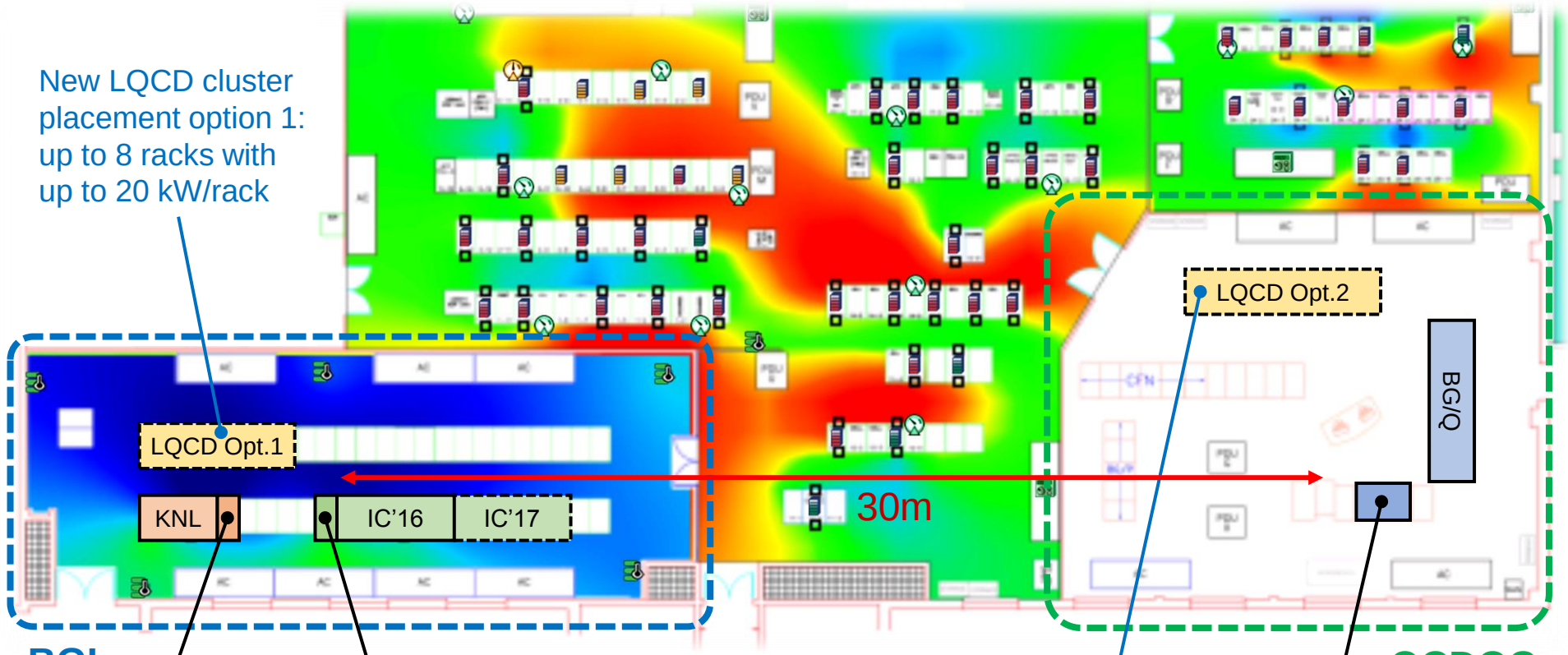
HPC Computing & Storage Resources

Institutional Cluster (IC), KNL Cluster
IB fabric connected GPFS for data (1 PB raw)
IB/Ethernet connected GPFS for home directories (0.5 PB raw)

BlueGene/Q & its storage systems
CFN Gen.3 & Gen.4 legacy clusters (1.8k non-HT CPU cores)



HPC Computing & Storage Resources

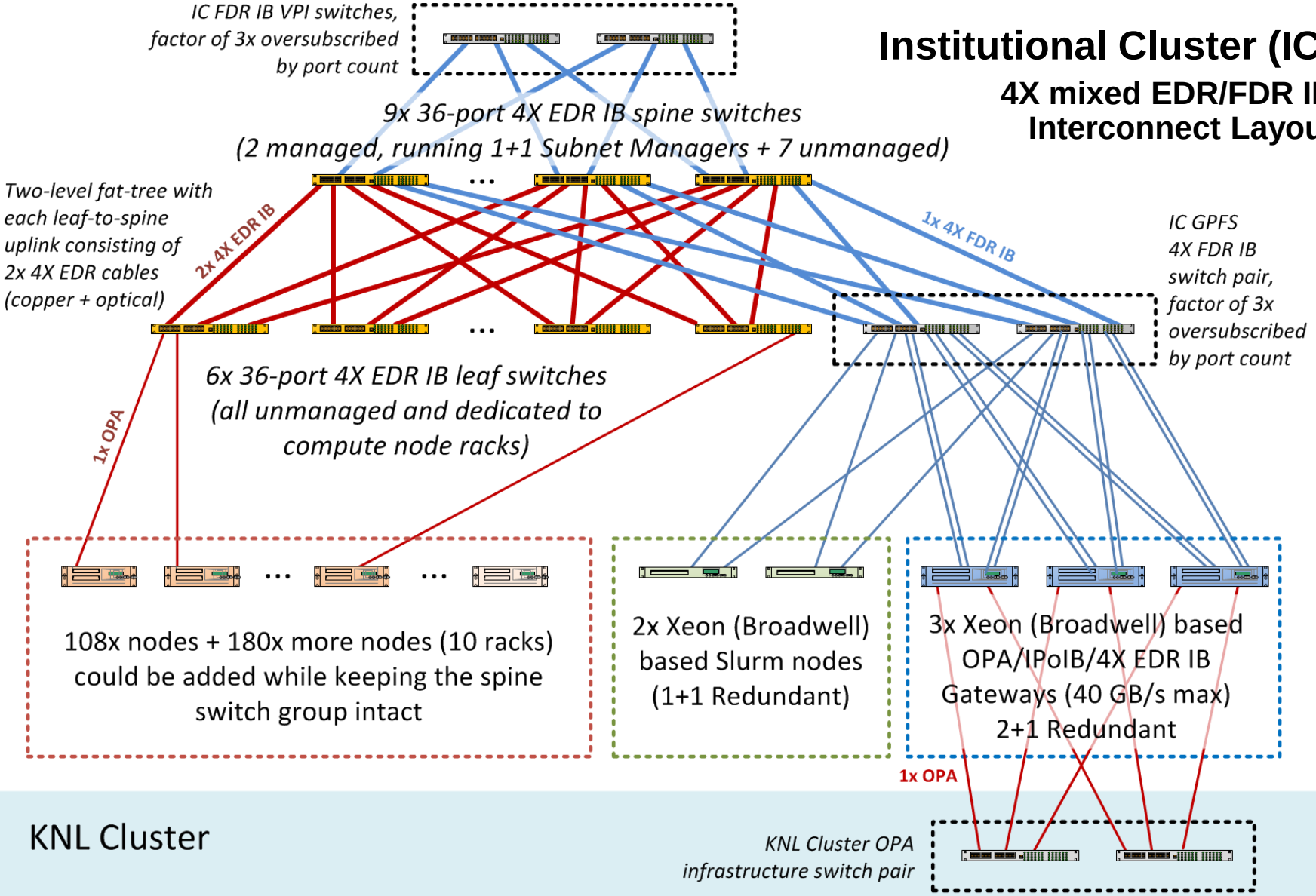


Institutional Cluster (IC)

- Our first HPC cluster available to the entire BNL community
  - Operational since January 4, 2017
- 108 compute nodes
  - Dual Xeon Broadwell (E5-2695 v4) CPU's with 36 physical cores on each
  - Two NVidia K80 GPU's
  - 1.8 TB SAS drive + 180 GB SSD for temporary local storage
  - 3.8k non-HT CPU cores / 256 GB RAM
- Two level fat-tree Mellanox 4X EDR IB interconnect
- 1 PB of GPFS storage (raw) with up to 24 GB/s I/O bandwidth capability (shared with KNL cluster as well)

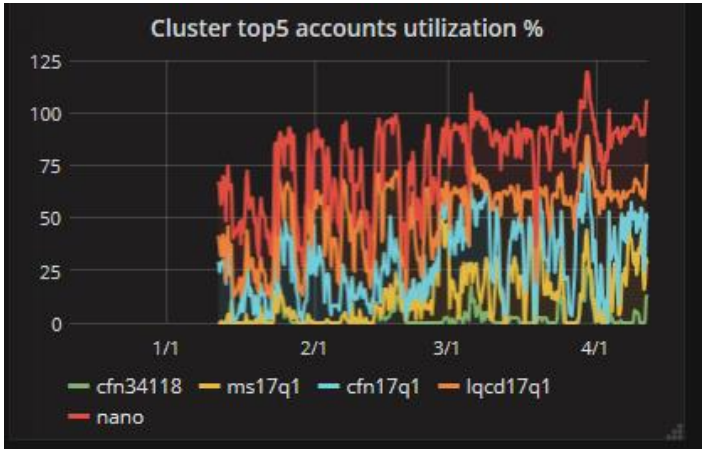
Institutional Cluster (IC)

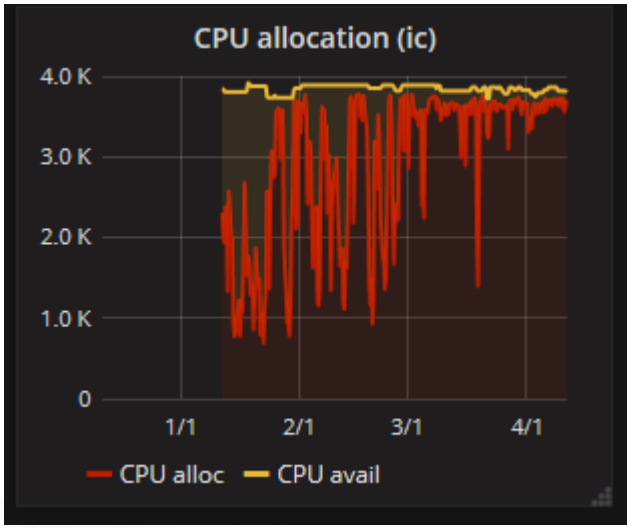
4X mixed EDR/FDR IB Interconnect Layout



Institutional Cluster (IC)

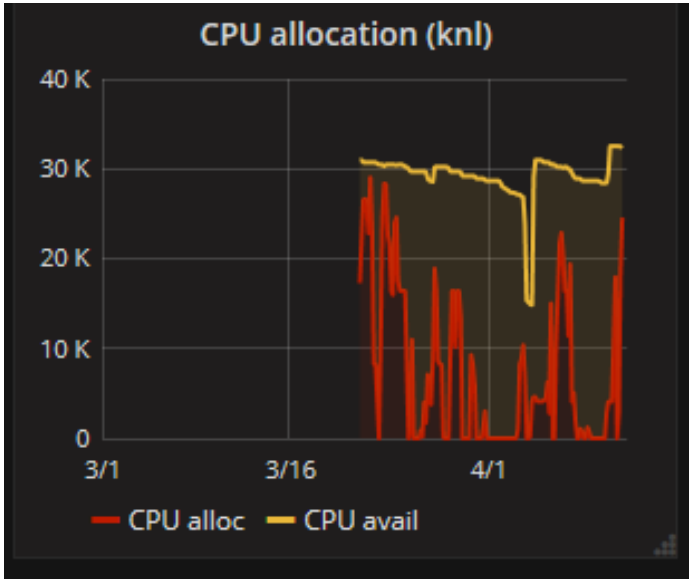
- Have fixed initial problems encountered with hardware failures, GPU and HCA cards performance issues, and support
- Currently ~120 registered users
- Cluster utilization approaching 95%
- Uptime nearly 100% over past three months
- Expansion under active discussion
  - Extent to be determined by expected demand (maximum 108 nodes on this extension)
  - P100 instead of 2x K80 GPU's





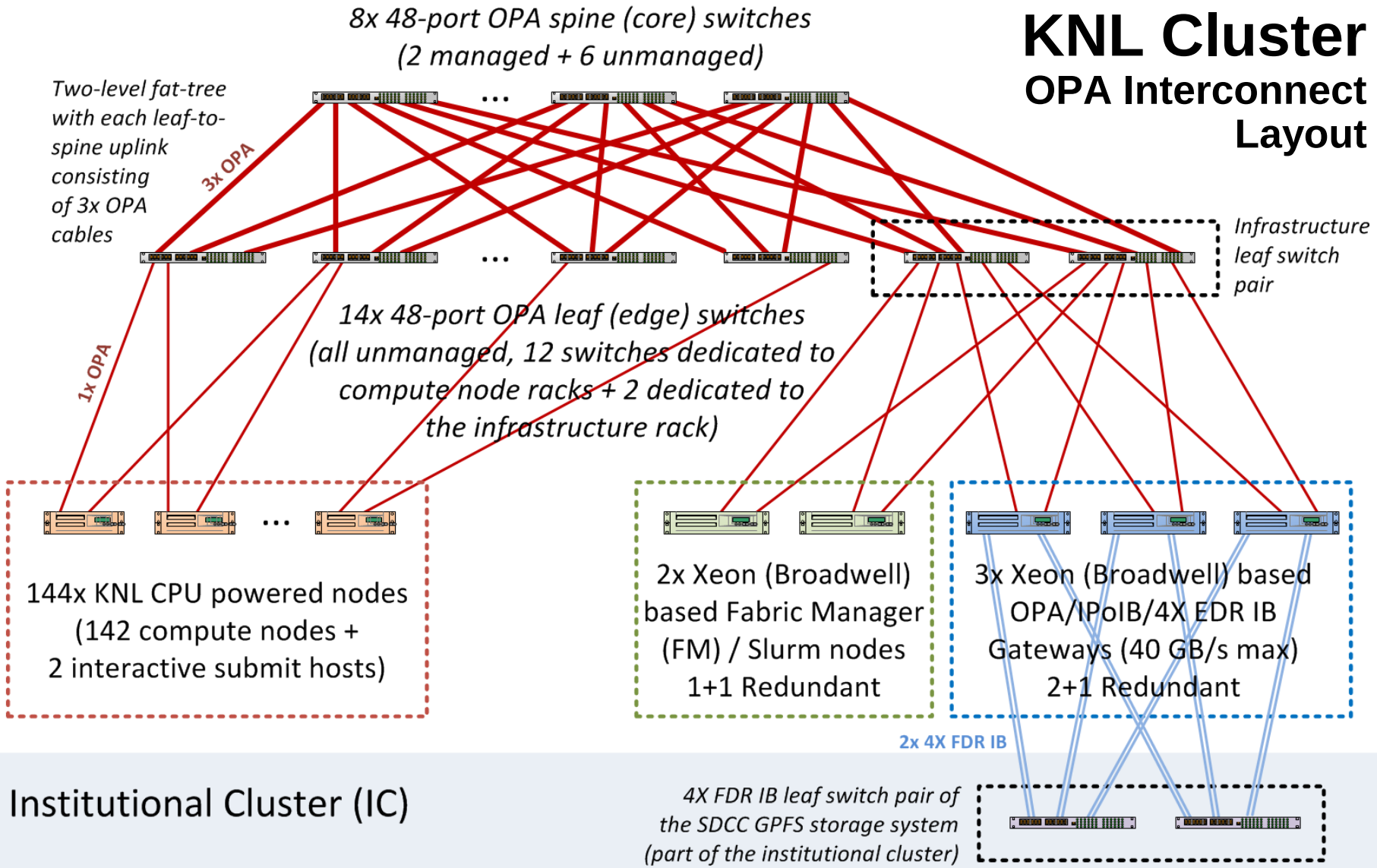
KNL Cluster

- 142 compute nodes + 2 submit nodes
  - Single Intel Xeon Phi CPU 7230:
    64 physical / 256 logical cores @ 1.3 GHz
    (1.5 GHz maximum turbo mode)
    per node
  - Dual Intel Omni-Path (OPA) PCIe x16
    HFI cards (non-blocking 200 Gbps =
    25 GB/s unidirectional)
  - 36k HT CPU cores / 36 TB of RAM in total



- Two level fat-tree single fabric Intel OPA interconnect system built out of 8x spine (core) plus 14x leaf (edge) 48-port Intel Omni-Path switches – 630 OPA uplinks (all copper) in total
  - Bisection bandwidth for compute nodes alone is about
    14 Tbps = 1.7 TB/s
- Unlike IC, achieving stability and performance with KNL cluster has required significant dedicated effort

KNL Cluster
OPA Interconnect Layout



Institutional Cluster (IC)

KNL Cluster: Early Issues

- Intel KNL CPU recalls: ~10 CPUs (out of 144) replaced by KOI so far, more replacements are on the way [Low]
- OPA cables early failures: about ~4 passive copper OPA cables (out of ~600) failed partially or completely shortly after installation – all replaced by KOI; no more dead cables for at least 4 months now [Low]
- Node/chassis HW stability issues: 2 problematic chassis (out of 36): PSU subsystem suspected, replacement by KOI is on the way [Moderate]
- Dual-rail OPA performance limitations with MPI / low rank jobs: at least 4 ranks per node are needed for the KNL compute nodes to fill the 25 GB/s (unidirectional) pipe: a problem unique to dual-rail OPA with KNL combination (it doesn't affect KNL with dual-rail 4X EDR IB), Intel notified back to Nov'16, the solution is still pending [High]
- Maintenance problems with high density OPA cable placement (both intra-rack and inter-rack): complete rewiring of OPA interconnect by KOI was needed back to Dec'16 in order to tackle the problem, the solution seems to be adequate [Low]
- HW driven performance degradation issues:
  - Severe clock-down to 40-80 MHz: only observed once but on the entire machine, power distribution glitch suspected, BIOS/BMC upgrade to S72C610.86B.01.01.0231.101420161754 / Op Code 0.28.10202, Boot Code 00.07 back to Dec'16 seem to have slved the problem [Low]
  - Subsequent issue with syscfg not working correctly with this new test BIOS, now solved by Intel [Low]

KNL: Issues Encountered While Trying to Enter Production

- KNL CPU performance degradation:
  - MCDRAM fragmentation leads to significant performance degradation over time (up to factor of 3x), node reboot is needed to solve the problem: latest XPPSL releases seems to alleviate the problem [Moderate]
  - Stepping out from RHEL7.2 kernel v3.10.0-327 to any other RHEL or vanilla custom built kernel resulted in systematic ~20% performance drop: the problem seems to be solved in the latest RHEL7.3 kernel v3.10.0-514.16.1 [Moderate]
- KNL CPU performance variation from node-to-node and from run-to-run:
  - Normally arising from the intrinsic systematic performance variation between physical cores on the KNL CPU (up to 13%), but was aggravated up to ~30% by stepping out of RHEL7.2 kernel v3.10.0-327: the problem seems to be solved in the latest RHEL7.3 kernel v3.10.0-514.16.1 [Moderate, needs confirmation]
- Using the non-default subnet prefix with Intel OPA Fabric Managers in out setup makes the OPA fabric unstable: the problem is reported to Intel, but not being actively investigated since an easy workaround exists (using the default prefix) [Low]
- KNL CPU instability after attempted MCDRAM/NUMA mode change via syscfg: complete power cycle of the node is needed after the mode change, otherwise the node ends up in an unusable state; the issue prevents full integrated of the KNL Cluster with Slurm – pending the BIOS update by Intel that is expected to solve the problem [High]

❖ More details are available in William Strecker-Kellogg's talk at HEPiX Spring 2017: https://indico.cern.ch/event/595396/contributions/2532420/

KNL Cluster: Performance Variation Example

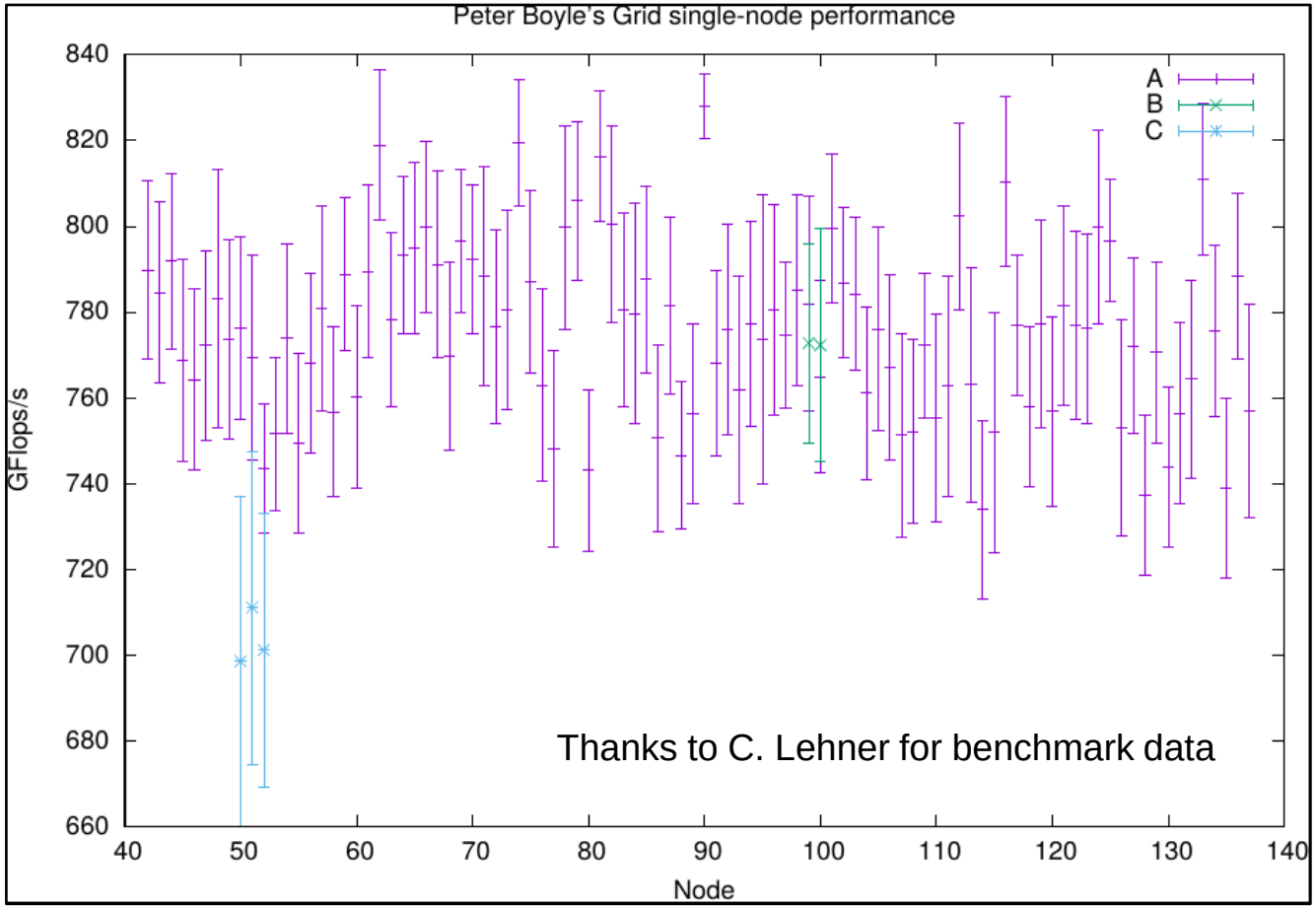
RHEL kernel versions involved:

A = 3.10.0-327 (original RHEL 7.2 Deployed by KOI)

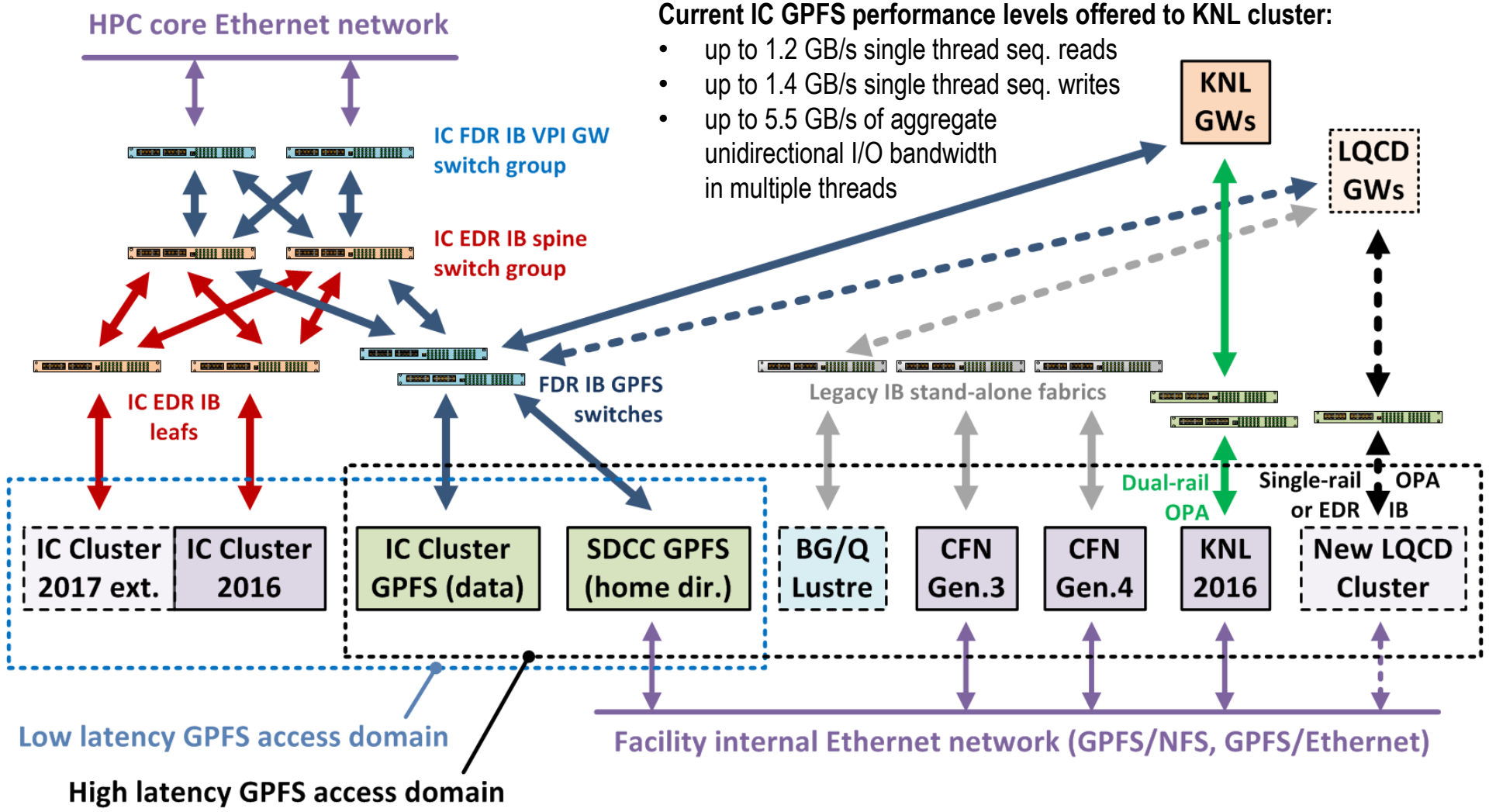
C = 3.10.0-327.36

B = 3.10.0-514.16.1 (latest RHEL 7.3)

Vanila kernel builds from v4.x showed even higher performance variation



SDCC HPC Storage Interconnect Layout



18

Summary

- SDCC/RACF facility has significantly expanded its HPC component in 2016
- Two new HPC clusters (IC & KNL) are deployed:
  - IC cluster by the HPE is now in full scale production with a diverse user base with near 100% resource allocation. All the cluster configuration and hardware issues encountered during the deployment process we resolved by the HPE in ~3 months.
  - KNL cluster by KOI Computing (offering the Intel platform). A significant number of hardware and software stack issues related to KNL CPUs and Intel OPA technology we encountered, out of which at least two remain unsolved until now (6 months after deployment). The expectations are that these remaining issues to be solved by Intel before the end of 2017, yet it's not guaranteed.
- USQCD community is already utilising the resources of BNL IC cluster on regular basis and we are hoping to achieve the same state of affairs with BNL KNL cluster in the near future
- Extension of the IC cluster and addition of the new USQCD cluster are expected in 2017Q3-4:
  - The IC cluster is to be extended with dual Xeon Broadwell, dual NVidia P100 GPU based compute nodes provided with a single-rail fat-tree 4X EDR IB interconnect
  - Multiple architectures are still being considered for the new USQCD cluster, including another KNL based system with single-rail OPA or 4X EDR IB fat-tree interconnect
  - Since several hardware and software stack issues related to the Intel KNL CPUs and the OPA technologies are not yet solved by Intel, relying on this particular architecture may result in extra manpower needed to maintain and operate such a system at the same level of resource availability and hardware stability as demonstrated by the BNL IC cluster
  - The process estimating the scale and costs of the necessary power and cooling infrastructure upgrades needed to accommodate the extended IC cluster and the new LQCD cluster is underway

Questions & Comments?