

SDCC Network Operations Status Report (Aug 4, 2022)

Alexandr ZAYTSEV
alezayt@bnl.gov

The diagram illustrates the network architecture for the BNL SciDMZ and B725 SciDMZ Extension, divided into two main sections by a vertical orange line.

Left Section (BNL SciDMZ):

- BNL Perimeter:** Connected to the SciDMZ via a 400 Gbps link.
- SciDMZ:** Connected to the B515 SciCore via a 40 Gbps link.
- B515 SciCore (Arista R3 series):**
 - Connected to B515 DTNs via a 40 Gbps link.
 - Connected to B515 HPC (CSI), STAR CH, NSLS-II, CAD, CFN, B725 ACL, and sPHENIX CH via a 40 Gbps link.
 - Connected to (2) B515 Storage Core via an 800 Gbps link.
 - Connected to (4) B515 Spine Group via a 1600 Gbps link.
- (2) B515 Storage Core (Arista R series):**
 - Connected to B515 DTNs via a 40 Gbps link.
 - Connected to (2) B515 Storage Core via an 800 Gbps link.
 - Connected to (4) B515 Spine Group via an 800 Gbps link.
- (2) 10 GbE RHIC:** Connected to (4) B515 Spine Group via an 800 Gbps link.
- (4) B515 Spine Group (Arista R series):** Connected to Compute Racks via an 800 Gbps link.

Right Section (B725 SciDMZ Extension):

- B725 SciDMZ Extension (Arista R3 series):** Connected to the SciDMZ via a 320 Gbps link.
- B725 DTNs:** Connected to the B725 SciCore via a 40 Gbps link.
- (2) B725 SciCore (Arista R3 series):**
 - Connected to B725 DTNs via a 40 Gbps link.
 - Connected to (2) B725 Storage Core via a 1600 Gbps link.
 - Connected to (4) B725 Spine Group via a 1600 Gbps link.
- (2) B725 Storage Core (Arista R3 series):**
 - Connected to B725 DTNs via a 40 Gbps link.
 - Connected to (2) B725 Storage Core via a 1600 Gbps link.
 - Connected to (4) B725 Spine Group via a 1600 Gbps link.
- (4) B725 Spine Group (Arista R3 series):** Connected to Compute Racks via an 800 Gbps link.

Central Labels: B515 B725

Legend: Storage Racks (blue), Compute Racks (red).

RHIC 10G Arista 7500E Farm distribution switch (s810) incident of Jul 27, 2022

- One of the two 10G modular distribution switches (s810) serving the PHENIX half of the nodes in 18x direct 10G connected RHIC racks in B515 BCF and RCF areas experienced a complete chassis failure at 13:00 BNL LT on July 27, 2022
- The ITD Network Engineering detected an issue withing 10 min of failure and recovered the chassis withing 25 min of the failure. Since the chassis was recovered so quickly the incident had only limited impact on the running jobs
- The preliminary analysis shows that the incident was likely caused by a failure of one of the redundant fabric modules in the chassis, likely compounded by some other hardware issue
 - Normally a failure of a module shouldn't cause the complete failure of the chassis, and further investigation is underway to understand why it cascaded in such a way.
 - This switch model (Arista 7500E series 8-slot switch w/ MXP line cards) is phased out of production as of 2020Q4 everywhere in SDCC, but these 2 units serving 10G connected RHIC Farm nodes in B515 BCF and RCF areas are planned to remain until the retirement of all these CPU racks in FY23 (by Sep 30, 2023)
 - The vendor is unlikely to make any fixes to the system since this type of chassis is several generations behind the latest generations of Arista equipment: E->R->R2->R3 (currently used at scale in SDCC in both B515 and B725), but the system stayed in production without any incidents since Oct 2017 (for 4 years and 9 months+) so the chances of surviving on it for 14 more months without disruptions are high
 - The quantity of cold spare parts for this system that we are still in possession of should allow us to survive for one more year and hopefully avoid a costly replacement of these switches at the last year of life of the CPU racks they are serving
 - Multiple possible mitigation strategies are available to us in case we see another failure on s810 in the upcoming few months (merging s810/s811 line card set into a single s811 chassis, replacing the s810 chassis completely and so on, although each one of them is going to cause a downtime for s810 switch clients on the scale of 1-2 hours)
 - Replacing these central switches with 10 GbE ToRs would be costly, require much effort and likely cause a downtime for the clients on the timescale of 1-2 days, so at this stage it is considered as an overkill scenario

B725/MDH Floor Occupancy (2022Q3-2023Q1)

- 105 rack frames are already deployed in B725 MDH so far, out of which 25 are CPU racks, 15 are active storage and infrastructure racks, and 65 rack frames are provisioned for the future growth of the storage and infrastructure rack set
- 53 more rack spaces are available for placing more CPU racks in the Low Density Area



1st fully populated CPU row (rack frames are delivered pre-configured and retired with equipment in them)

Storage/infrastructure rows
(all rack frames are pre-deployed)

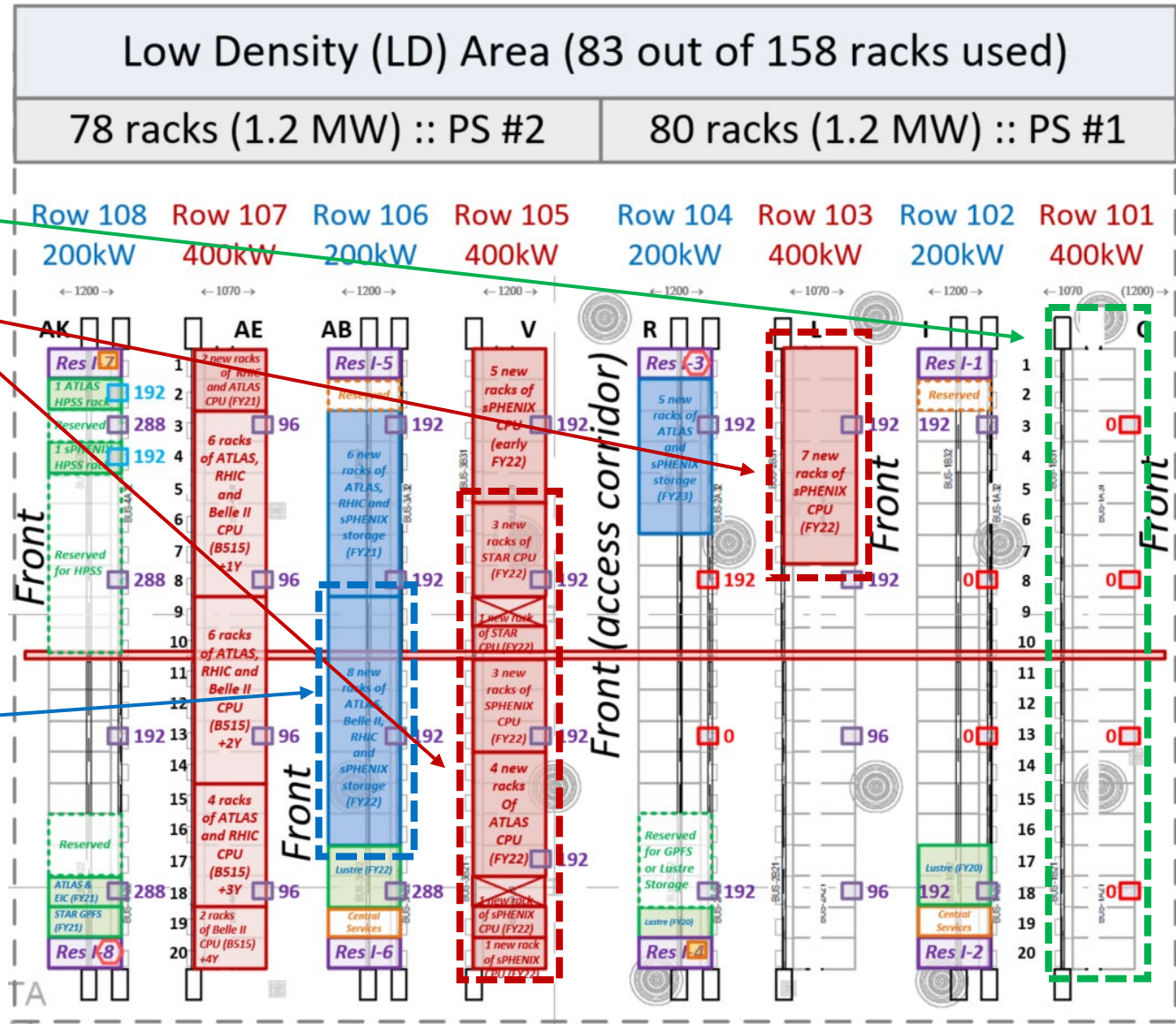
Aug 4, 2022

B725/MDH Floor Occupancy (2022Q3-2023Q1)

Copper connectivity provisioning is in progress for the row 101 for DCIM (the last row to be activated for Power Systems #1 and #2 areas of the MDH)

20 new CPU racks are to be deployed 2022Q4-2023Q1 as a result of FY22 ATLAS & RHIC CPU purchases)

8 more storage racks are being configured (40x JBOD+headnode pairs of combined FY22 DISK purchase)



Questions & Comments