Smarter technology for all

Lenovo storage update & Lessons from a DSS-G Upgrade

Sigrun Eggerling, HPC Storage Engineer
Jens Werner, HPC Consultant

| 2022-10-20

2022 Lenovo Internal. All rights reserved.

Lenovo Distributed Storage Solution for IBM Spectrum Scale™

Software defined storage with IBM Spectrum Scale

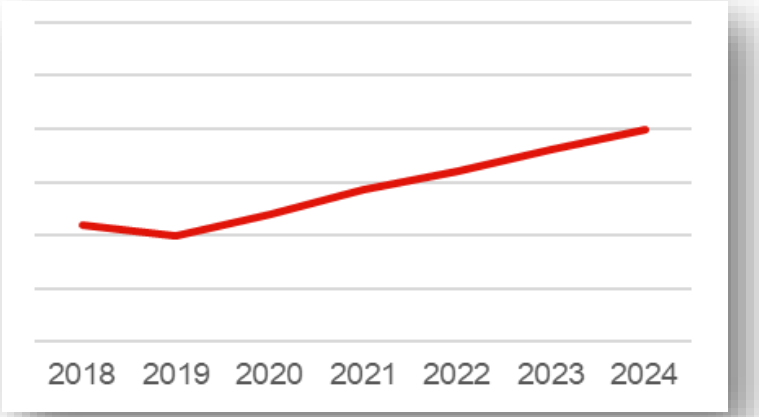
- Longterm partnership for growth in HPC storage
- Strategic OEM contract

Independence and compete-tiveness in HPC storage

- Broad parallel FS coverage (HDD, SSD, NVMe)
- Winner of multiple awards from prominent HPC/AI orgs

Key differentiation from Exascale to Everyscale

- From Wall Street to LRZ and KMA, DSS-G adds value
- **300+** PB 2018-2020 installed



Lenovo 2022 Lenovo Internal. All rights reserved.

Lenovo Solutions for IBM Spectrum Scale™

Lenovo DE Series

- Spectrum Scale DAE/DME
- Midrange SAN block storage
- Entry level NVMe AFA

- Focus: midrange solutions

Lenovo DSS-G 2##

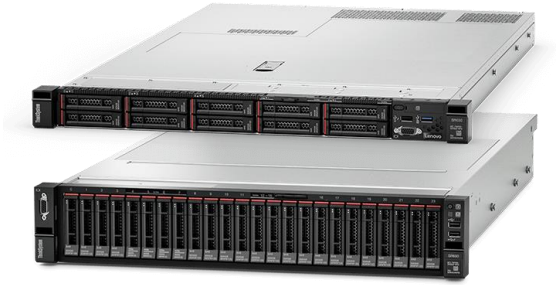
- Spectrum Scale DAE/DME
- direct-attached block storage

- Focus: high end solutions

Lenovo DSS-G 100

- Spectrum Scale ECE
- server-based block storage

- Focus: NVMe solutions



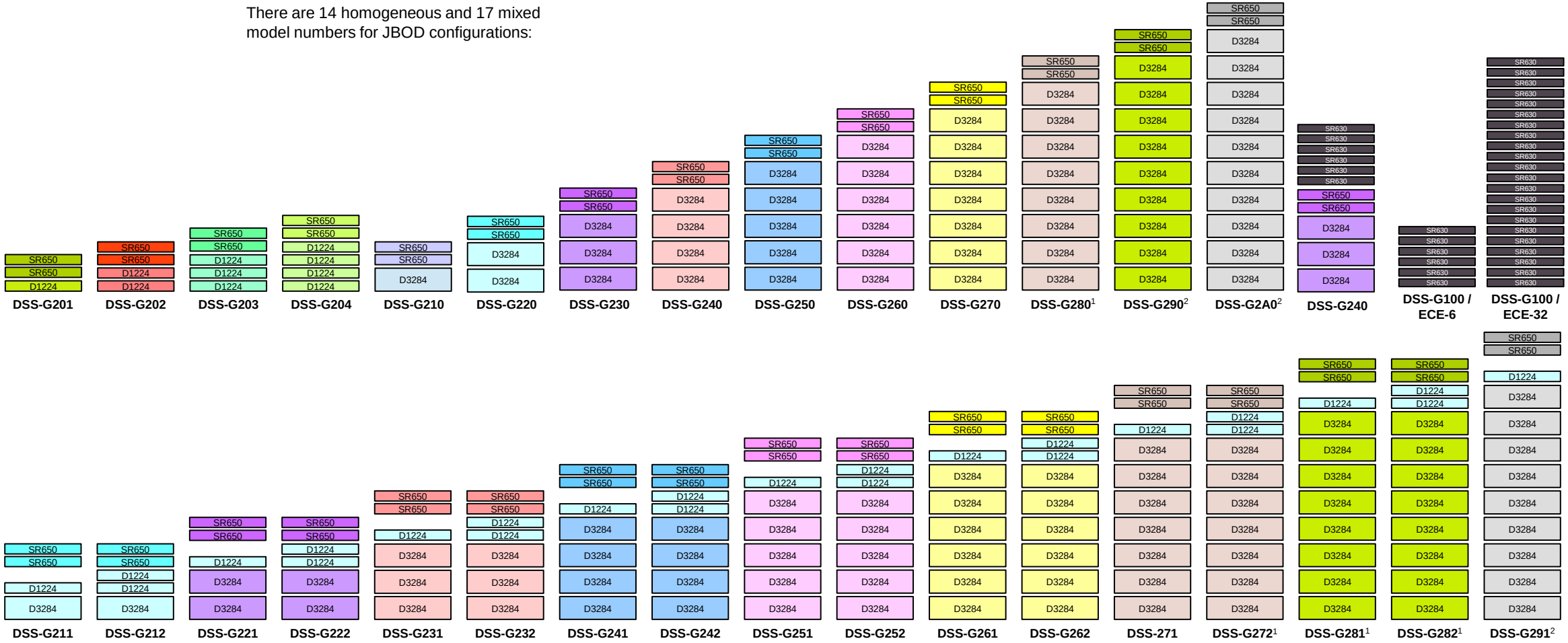
Lenovo 2022 Lenovo Internal. All rights reserved.

Lenovo DSS-G 4.1 Configurations

There is only one model number for ECE: DSS-G100.

But a Lenovo ECE solution will consist of multiple ECE servers where each one will be configured with at least 6 but not more than 32 servers. Storage solutions can be a hybrid mix of DSS-G and ECE Here are 3 examples.

There are 14 homogeneous and 17 mixed model numbers for JBOD configurations:

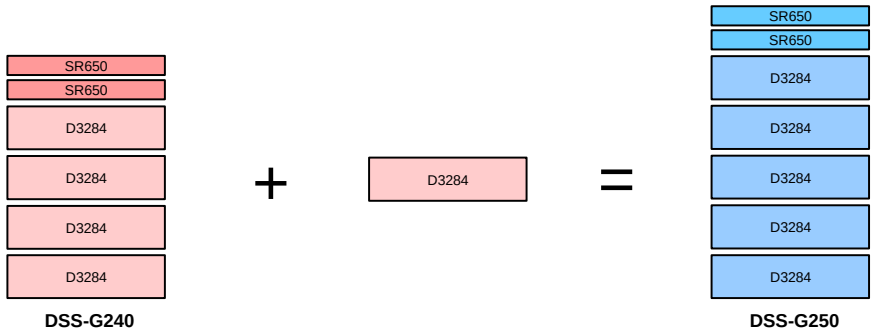


Footnotes:
1. Fits in 48U rack.
2. Requires 2 racks.

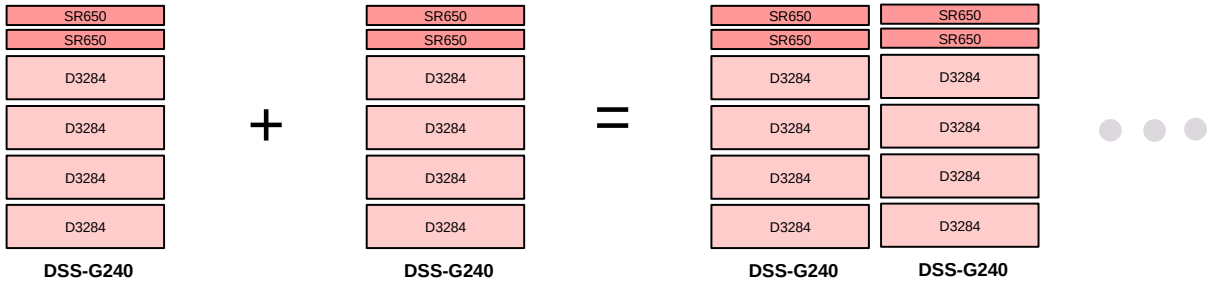
Lenovo 2022 Lenovo Internal. All rights reserved.

Scaling Lenovo DSS-G

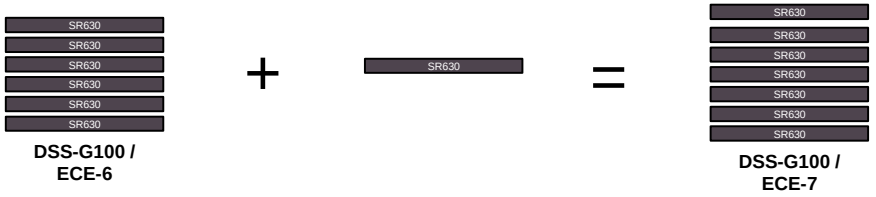
Spinning disk solutions can have enclosures added



Multiple DSS-G systems can be deployed in the same cluster/file-system



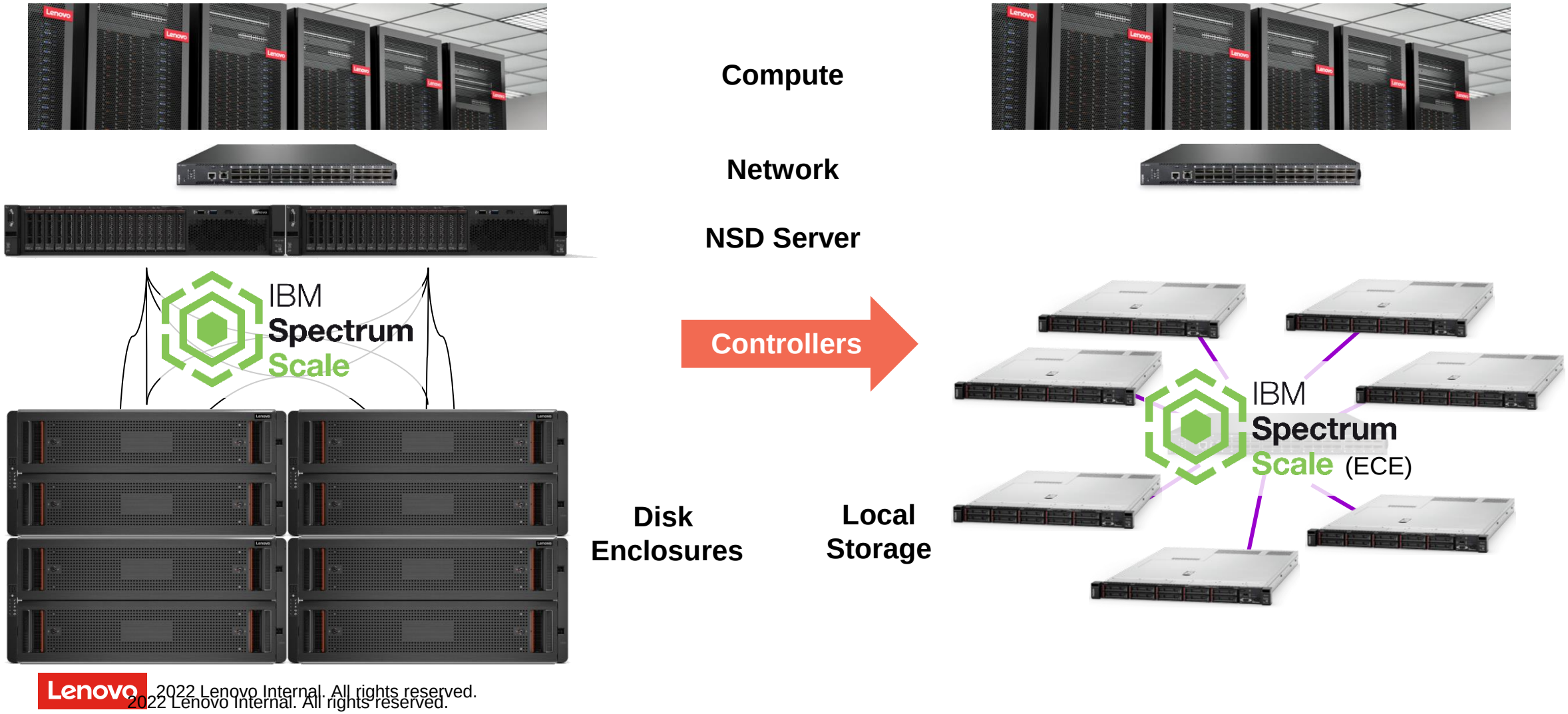
DSS-G100 ECE solutions can grow by server



Lenovo 2022 Lenovo Internal. All rights reserved.

Lenovo Distributed Storage Solution

Traditional (enclosure-based) versus G100 with ECE



Lenovo 2022 Lenovo Internal. All rights reserved.
2022 Lenovo Internal. All rights reserved.

DSS-G Current Releases

Product	Release date	Description	Edition	Supported configurations	Core component	Network stack	Remarks	Software download	LeSI Best Recipe
DSS-G 4.2b	2022-09-01	Full installation / RHEL 8.4	standard or advanced	DSS-G2xy with SR650 / SR650 V2	Spectrum Scale DAE or DME 5.1.3-1.27	Mellanox OFED 5.6-1.0.3.3 or Cornelis OPXS 10.11.1.3.1	• Release notes (text file) • DSSGversion (text file) • Documentation (ZIP archive)	Lenovo Service Connect	LeSI 22B.2
			erasure	DSS-G100 ECE with SR630 / SR630 V2	Spectrum Scale ECE 5.1.3-1.27				
DSS-G 2.10b		Full installation / RHEL 7.9	standard or advanced	DSS-G2xy with x3650 M5 / SR650	Spectrum Scale DAE or DME 5.1.3-1.27	Mellanox OFED 4.9-5.1.0.0 or Cornelis OPXS 10.11.1.3.1	• Release notes (text file) • DSSGversion (text file) • Documentation (ZIP archive)		

https://support.lenovo.com/us/en/solutions/ht510974-lenovo-distributed-storage-solution-for-ibm-spectrum-scale-dss-g#DSS-G-4.2b

4.3 will likely be the last to support xCAT – move to Confluent. Lenovo will supply a migration guide!

Lenovo 2022 Lenovo Internal. All rights reserved.

DSS-G HW migration & Upgrade

September 2022

Lenovo 2022 Lenovo Internal. All rights reserved.

Original Setup

- 44 x DSS-G220 (10 TB HDD) for Data (4x SAS HBA config)
- 2x DSS-G202 (7,68 TB SSD) for Metadata (4x SAS HBA config)
- 10x SR630 as Manager-Quorum nodes
- 52 PB usable space (8+2p)
- Intel Omnipath OPA Fabric (2x HCA per NSD)
- Separated in 2 clusters with one FS each (2/3 – 1/3)
- DSS-G release 2.5a:
  - RHEL 7.6 Kernel 3.10.0-957.41.1
  - Spectrum Scale RAID 5.0.4-PTF-1-efix6 (5.0.4-1.6)
  - Intel OPA IFS 10.10.0.0.445

Lenovo 2022 Lenovo Internal. All rights reserved.

Hardware rework

- Goal
  - Place a Mellanox ConnectX-4 VPI adapter in slot4 (connection to a new non-OPA cluster via Ethernet)
- Challenge
  - No free PCI slot left for this Mellanox adapter
- Solution
  - Remove SAS HBA from slot 4 and re-cable SAS connections for 3x SAS HBA solution

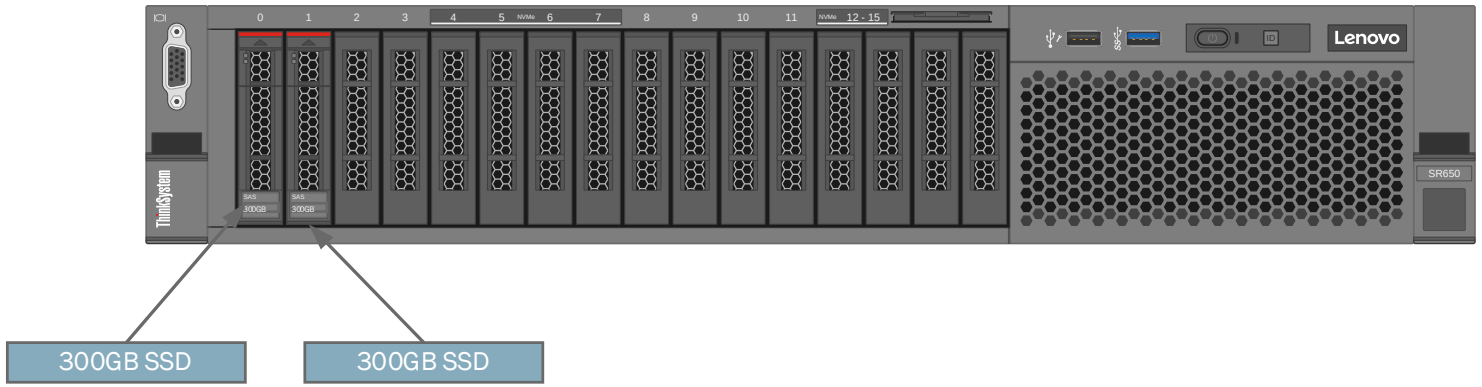
This can only be done **offline**. Required **3 days per cluster** with 2 technicians and 2 admins.

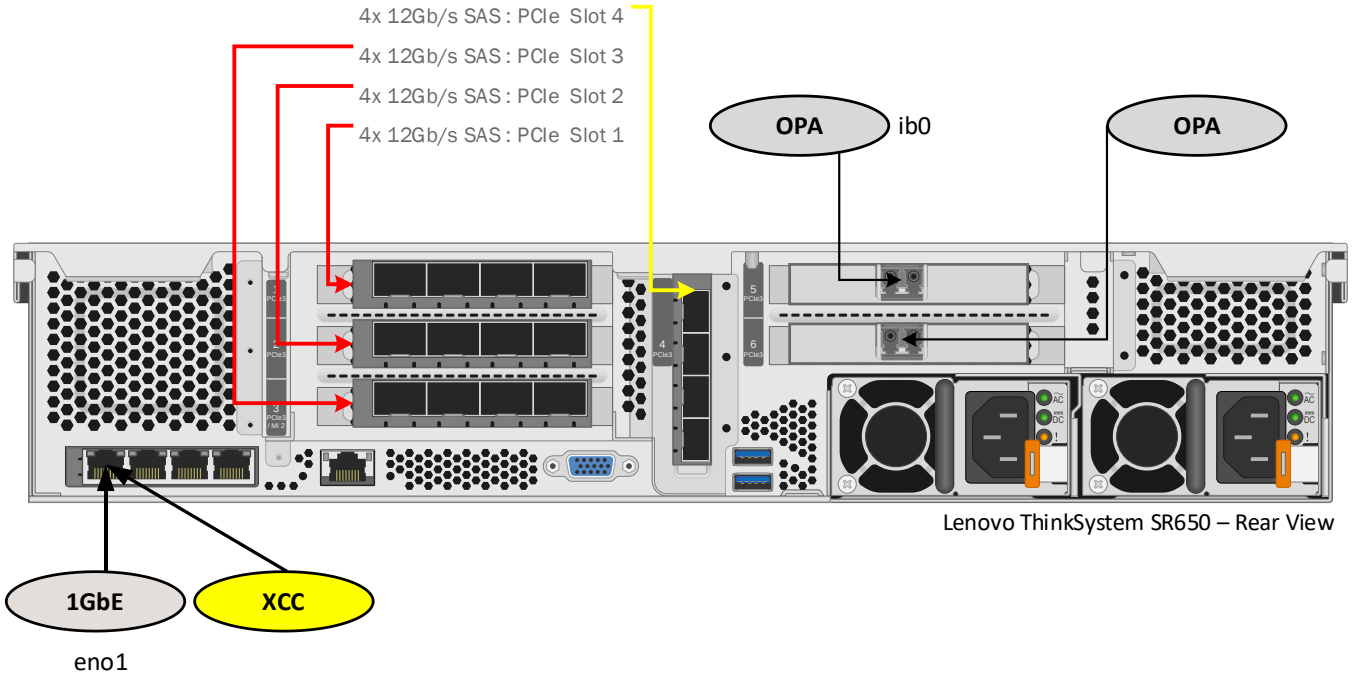
Requires Lenovo validation of change!

Lenovo 2022 Lenovo Internal. All rights reserved.

DSS-G Server Node configuration – Initial Configuration

Lenovo ThinkSystem SR650 – Front View





Lenovo ThinkSystem SR650 – Rear View

2022 Lenovo Internal. All rights reserved.

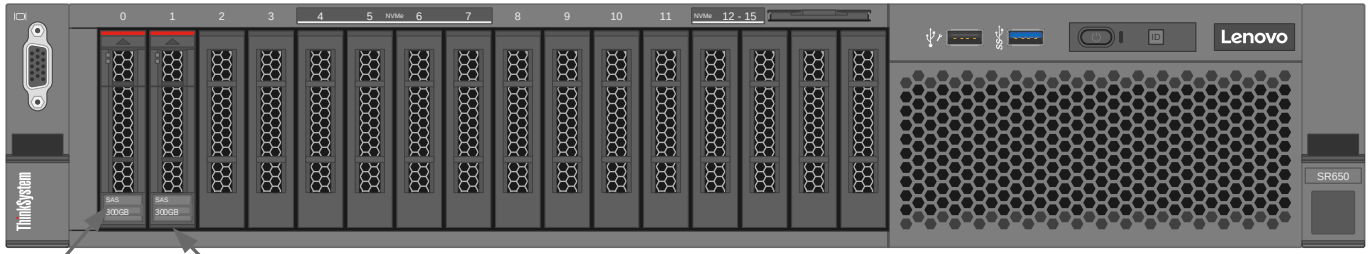
DSS-G Server Node configuration

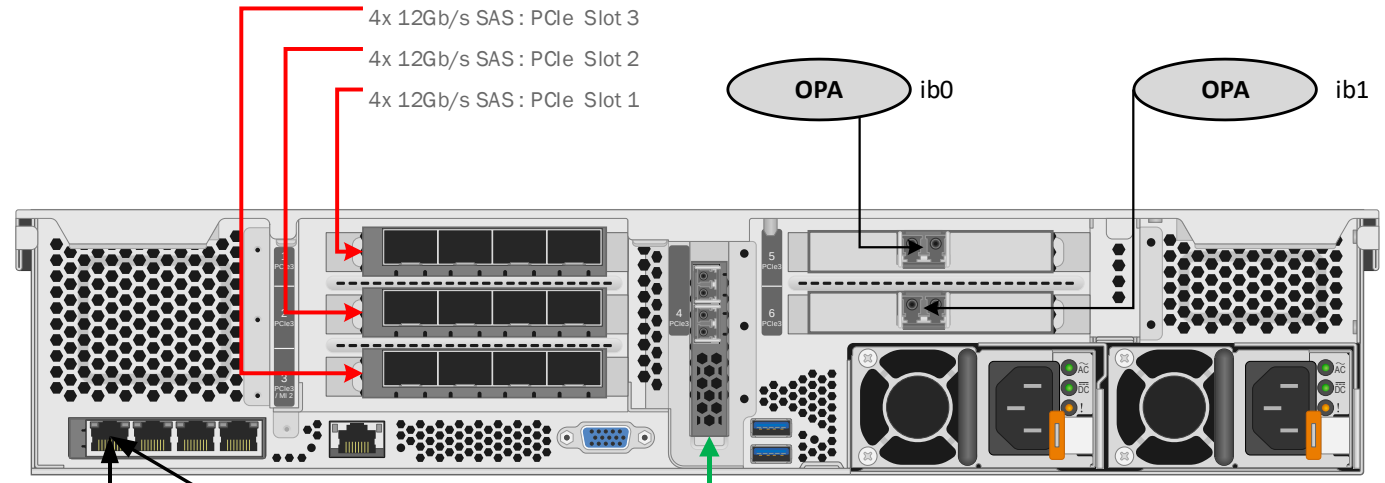
Component	Option	Quantity	Notes
Processor	Intel Xeon Gold 6142 16Core, 150W 2.6GHz Processor	2	Dual Socket 32 Cores total
Memory	ThinkSystem 16GB TruDDR4 2666 MHz (2Rx8 1.2V) RDIMM	12	192 GB total memory
Intel OPA Single port 100Gb/s adapter	Intel OPA 100 Series HFA Single port x16 PCIe 3.0 HFA	2	Configure both ports with IPoOPA and RDMA for GPFS
Quad port 10GbE LOM	ThinkSystem 10Gb 4-port Base-T LOM	1	4x 1/10GbE Ethernet ports LOM option
Internal Disks	ThinkSystem 2.5" 300GB 10K SAS 12Gb Hot Swap SFF HDD	2	Internal disks configured with RAID-1
RAID adapter	ThinkSystem 930-8i 2GB Flash PCIe 12Gb RAID Adapter	1	Internal SAS RAID adapter connecting to 2 Internal SAS disks Also provides NVRAM partitions for Scale GNR
Quad Port SAS adapter	ThinkSystem 430-16e SAS/SATA 12Gb HBA	3	Connect to 6x Lenovo D3284 High density storage 12Gb/S SAS links (PCI Slots 1,2,3,4)
XCC port	ThinkSystem XClarity Controller Standard to Advanced Upgrade	1	1GbE Ethernet port to XCC
OPA Network	Intel OPA 100 Series Single-port PCIe 3.0 x16 HFA	2	To be used for Spectrum Scale connectivity to Compute cluster (RDMA)
40GbE Ethernet	Mellanox ConnectX-4 2x100GbE/EDR IB QSFP28 VPI Adapter	1	Spectrum Scale cluster communications : Daemon and Administration interface

Lenovo 2022 Lenovo Internal. All rights reserved.

DSS-G Server Node configuration – Target Configuration

Lenovo ThinkSystem SR650 – Front View





Lenovo ThinkSystem SR650 – Rear View

1GbE

eno1

XCC

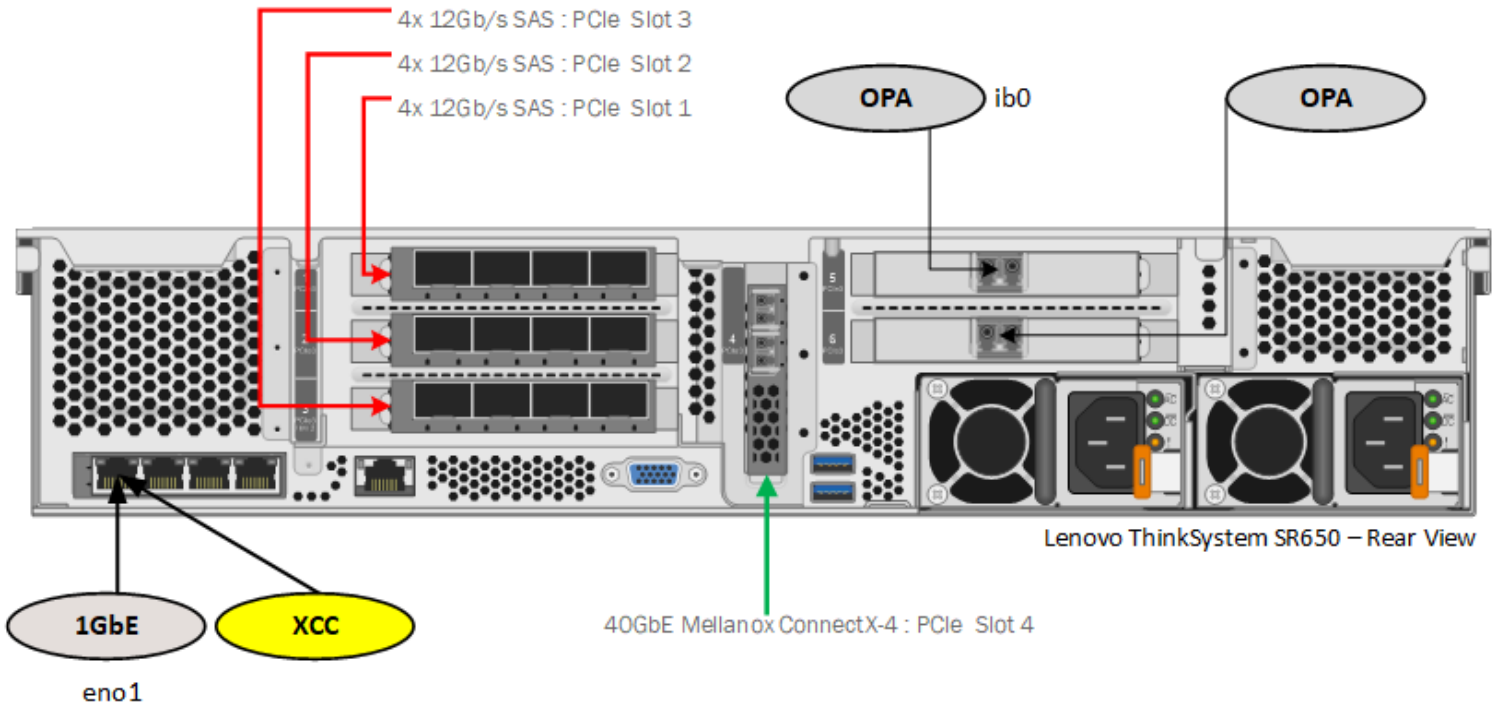
40GbE Mellanox ConnectX-4 : PCIe Slot 4

Slot 4 changed

Lenovo 2022 Lenovo Internal. All rights reserved.

DSS-G220/DSS-G202 Building Block : Target SAS Port Mapping

Server,slot,port	Enclosure,IOM, port	SR650 location	D3284 location
1,1,1	1,1,0	\S1\P1	\ESM B\SAS A
1,2,1	2,0,0	\S2\P1	\ESM A\SAS A
1,3,1	1,0,0	\S3\P1	\ESM A\SAS A
1,3,2	2,1,0	\S3\P2	\ESM B\SAS A
2,1,1	1,1,1	\S1\P1	\ESM B\SAS B
2,2,1	2,0,1	\S2\P1	\ESM A\SAS B
2,3,1	1,0,1	\S3\P1	\ESM A\SAS B
2,3,2	2,1,1	\S3\P2	\ESM B\SAS B

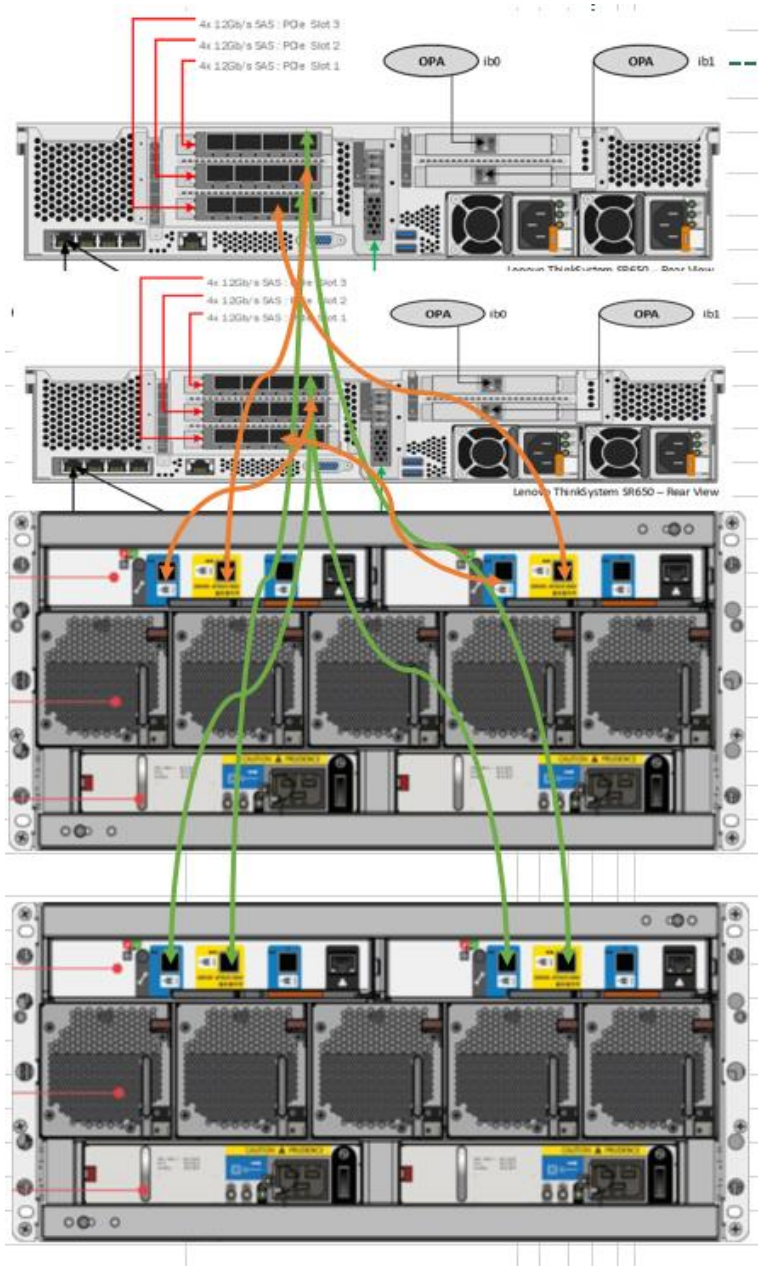


Lenovo 2022 Lenovo Internal. All rights reserved.

Re-cabling

- After recabling check the topology with mmgetpdisktopology / topsummary
- Set the Mellanox HBA to Ethernet only with mstconfig (part of mstflint package)

```
mstconfig -d 58:00.0 set LINK_TYPE_P1=2 LINK_TYPE_P2=2
```



Cabling changed

Cabling same as before

Lenovo 2022 Lenovo Internal. All rights reserved.

Software Upgrade 1/2

- Target Release DSS-G 4.1b:
  - RHEL 8.4 Kernel 4.18.0-305.40.2
  - Spectrum Scale RAID 5.1.2-PTF2.efix9
  - OPA IFS 10.11.1.0.2
  - Released 05.05.2022 (SS sub date)
- Goal:
  - Reinstall all NSD Server + Manager Nodes with DSS-G release 4.1b or RHEL 8.4 + GPFS
  - Update all FW, also on Manager Nodes to matching LESI 22A release.
  - Ensure that only OPA RDMA (OFED) stack will be installed

Lenovo 2022 Lenovo Internal. All rights reserved.

Software Upgrade 2/2

- Challenges:
  - DSS-G setup scripts are not able to deal with Omnipath and Mellanox (VPI) adapters at the same time. Manual fix needed for this.
  - If Mellanox + Onmipath HCAs installed, some unwanted UEFI changes happen.
  - Adapt all customer specific post-scripts from RHEL 7 to RHEL 8 (eg. Network settings)
- Solutions:
  - Change the (M)OFED install script that only OPA SW stack will be installed.
  - Change the (M)OFED FW update script to perform upgrades on both HCA types.
  - Adapt the UEFI settings batch files (processed by onecli)
  - Create a new postscript that changes the type of Mellanox HCA to ethernet (again)
  - Recreate scripts to adapt RHEL 8 specific config settings to all customer post(boot) scripts. This was the most time consuming part.

Lenovo 2022 Lenovo Internal. All rights reserved.

Observations

- Very limited HW failures
  - One SAS HBA
  - Two SAS cables (may have been defect before)
  - Two Enclosure PSUs (both on same enclosure – of course ☺)
  - 4 HDDs were not able to receive the FW upgrade
- One HDD prevents an entire Building Block from starting (core dumping NSD nodes).
  - Means that the entire FS is not accessible.
- Never seen such a “bad” disk before. Causes some type of SAS bus flapping events.
- After two weeks of operation 8 slow pdisks were sorted out by disk-hospital (> 7k HDDs active)

- Read performance increase round about 20 percent!

Lenovo 2022 Lenovo Internal. All rights reserved.

Post upgrade tasks

- Upgrade the min. Release Version to latest
- Upgrade the Recovery Groups to new version
- Convert Recovery Groups from legacy to mmvdisk style

Planned tasks:

Upgrade Filesystem version to latest (dependencies to client clusters)

Lenovo 2022 Lenovo Internal. All rights reserved.

Smarter technology for all

Lenovo

thanks.