Smarter technology for all

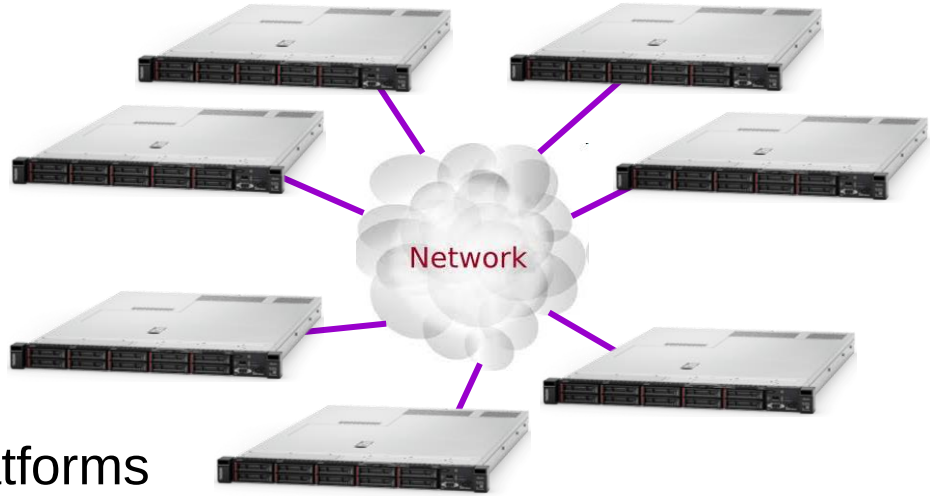
Architecting an ECE Solution

Sigrun Eggerling | 23.03.2023

seggerling@lenovo.com

IBM Storage Scale Erasure Code Edition Overview

- Delivers Storage Scale RAID over the network rather than within an enclosure
  - Declustered RAID
  - Disk Hospital
  - Highly fault-tolerant erasure codes
  - Dropped/corrupted write detection
  - Hardware neutrality

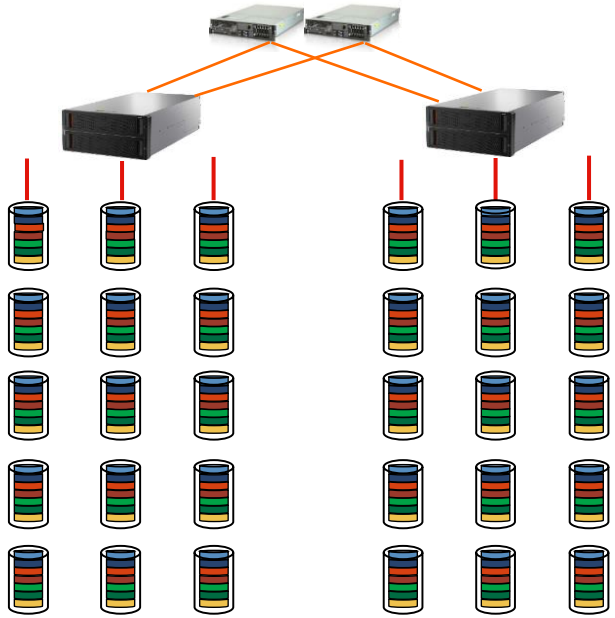


- Runs on commodity storage-rich servers
  - Leverage cost advantages of high-volume server platforms
  - Survive concurrent loss of nodes and disks

Lenovo Solutions for IBM Storage Scale RAID

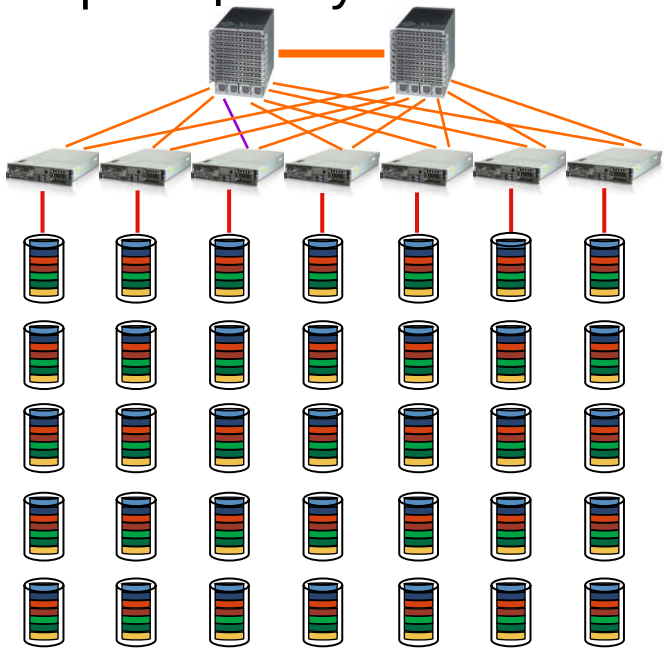
Lenovo DSS-G 2##

- File System RAID within building block
- SSD / HDD support
- ≤10 enclosures per recovery group
  - (192 solid state / 838 hard disk drives)



Lenovo DSS-G 100

- File System RAID across the network
- NVMe support
- ≤32 servers per recovery group
  - (320 NVMe drives)
- 3x HDR adapters per system

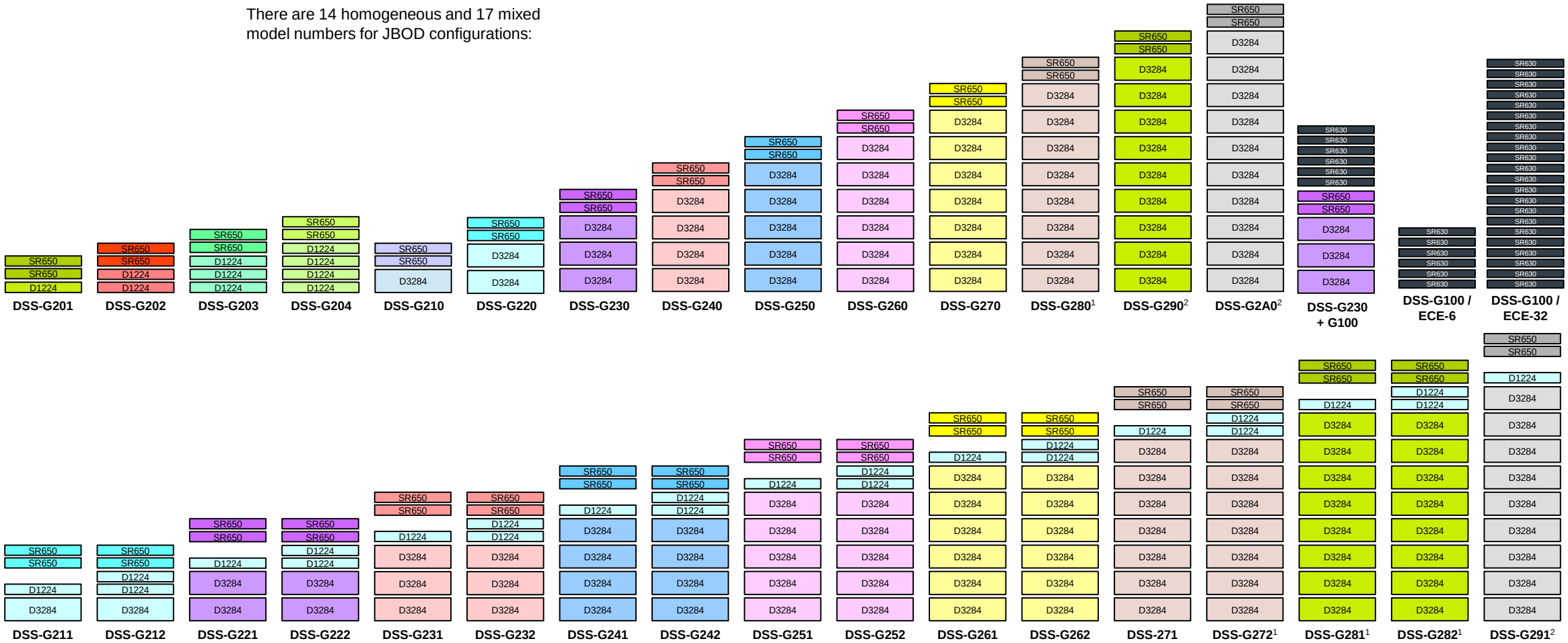


Lenovo DSS-G 4.4 Configurations

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.

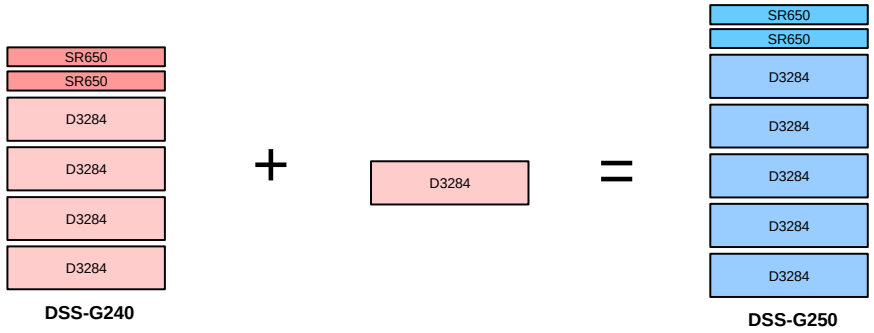
There are 14 homogeneous and 17 mixed model numbers for JBOD configurations:



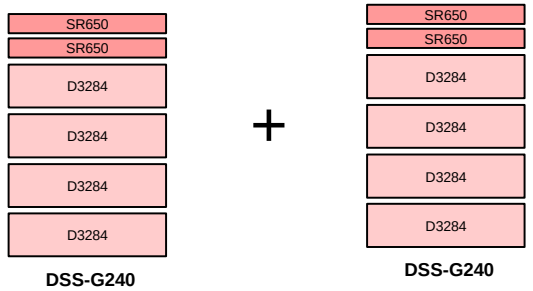
Footnotes:
1. Fits in 48U rack.
2. Requires 2 racks.

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 Recommended RG Sizes for Each Erasure Code

Recommended Recovery Group Size for each Erasure Code

Number of Nodes	4+2P	4+3P	8+2P	8+3P
4	**Not recommended** 1 Node	1 Node + 1 Device	**Not recommended** 2 Devices	**Not recommended** 1 Node
5	**Not recommended** 1 Node	1 Node + 1 Device	**Not recommended** 1 Node	**Not recommended** 1 Node
6 – 8	2 Nodes	2 Nodes [1]	**Not recommended** 1 Node	1 Node + 1 Device
9	2 Nodes	3 Nodes	**Not recommended** 1 Node	1 Node + 1 Device
10	2 Nodes	3 Nodes	2 Nodes	2 Nodes
11+	2 Nodes	3 Nodes	2 Nodes	3 Nodes

Note: For 7 or 8 nodes, 4+3P is limited to two nodes by recovery group descriptors rather than by the erasure code.

Table from IBM ECE Redbook. Color coding added by Lenovo.

Lenovo DSS-G Support

Requirement for ECE:

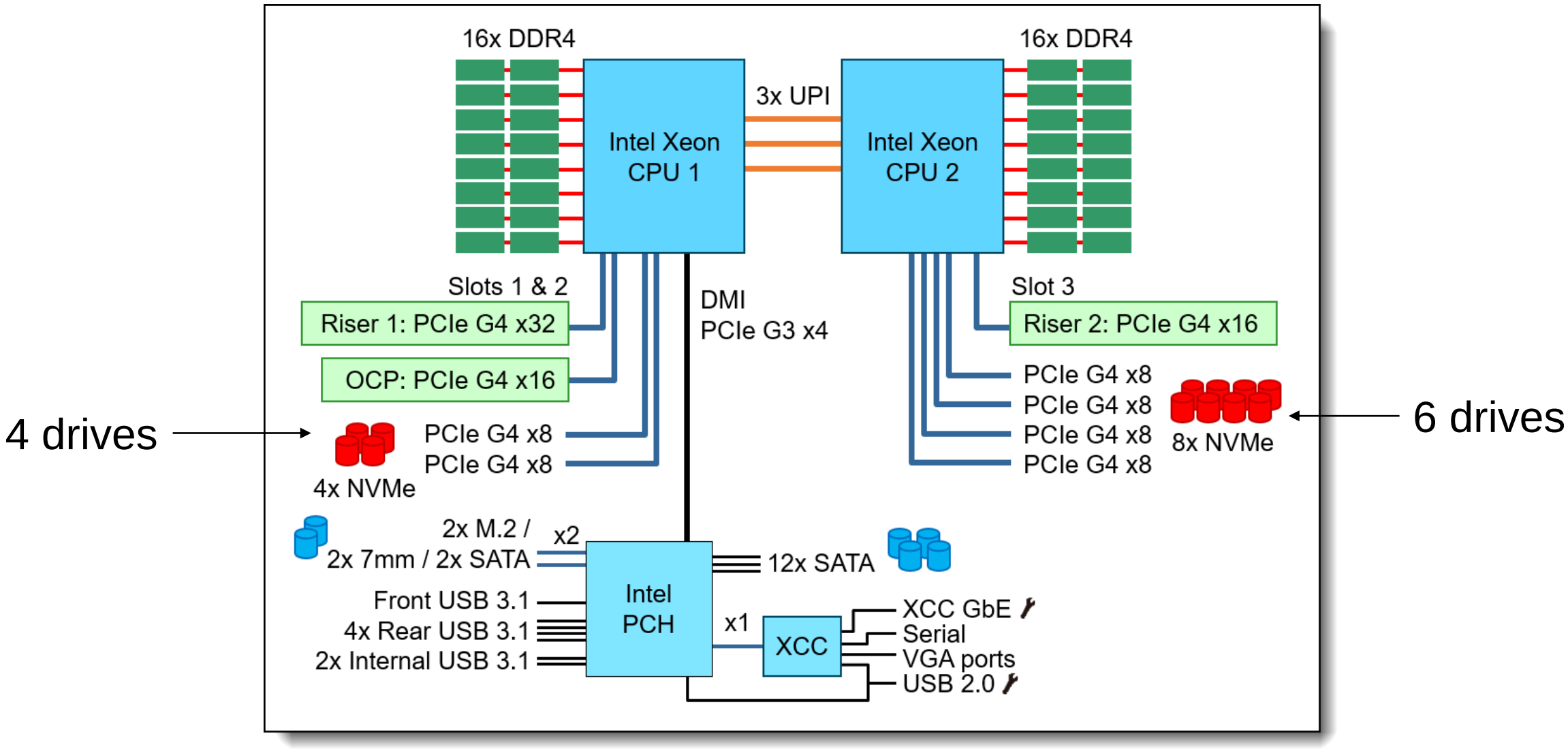
- Minimum **6** Servers for 4+2P
- Minimum **9** Servers for 4+3P [1]
- Minimum **10** Servers for 8+2P
- Minimum **11** Servers for 8+3P

Recommendation:

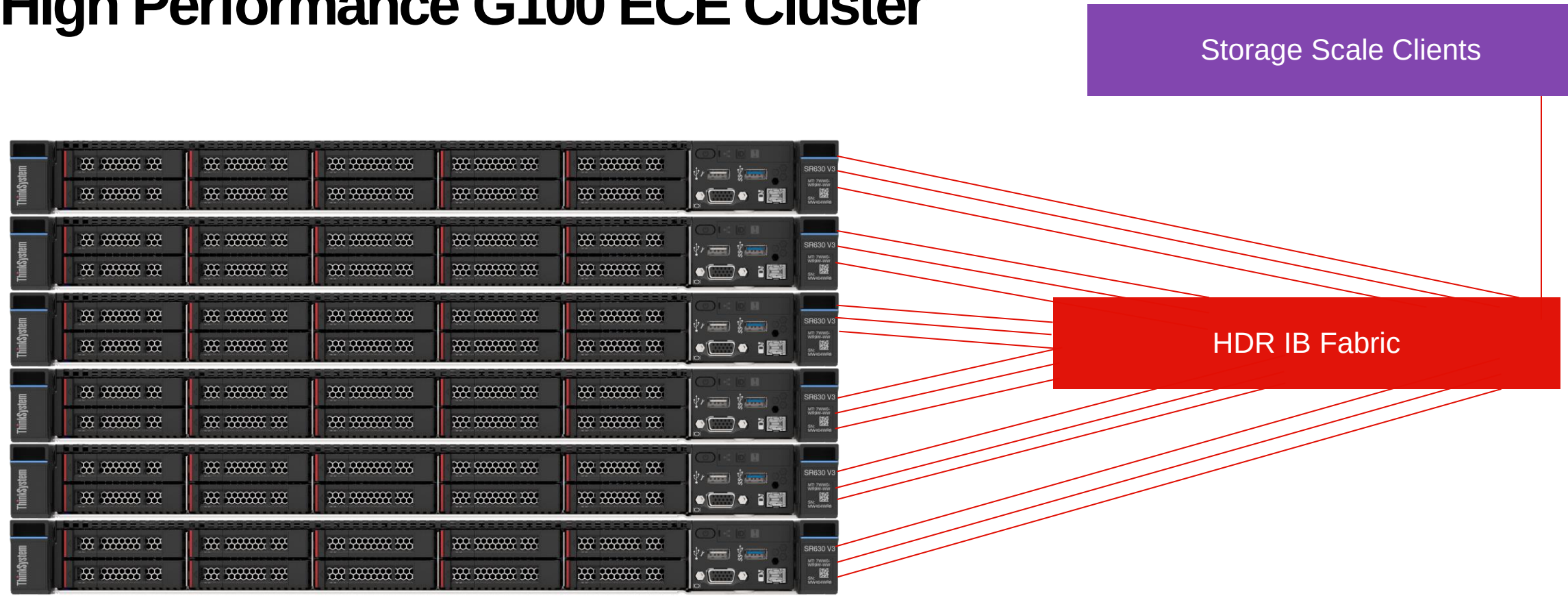
Add 2 or 3 more nodes
for rebuild scenarios...

- 8+ Servers for 4+2P (+2P)
- 10+ Servers for 4+3P (+3P)
- 12+ Servers for 8+2P (+2P)
- 14+ Servers for 8+3P (+3P)

SR630 V2 architecture



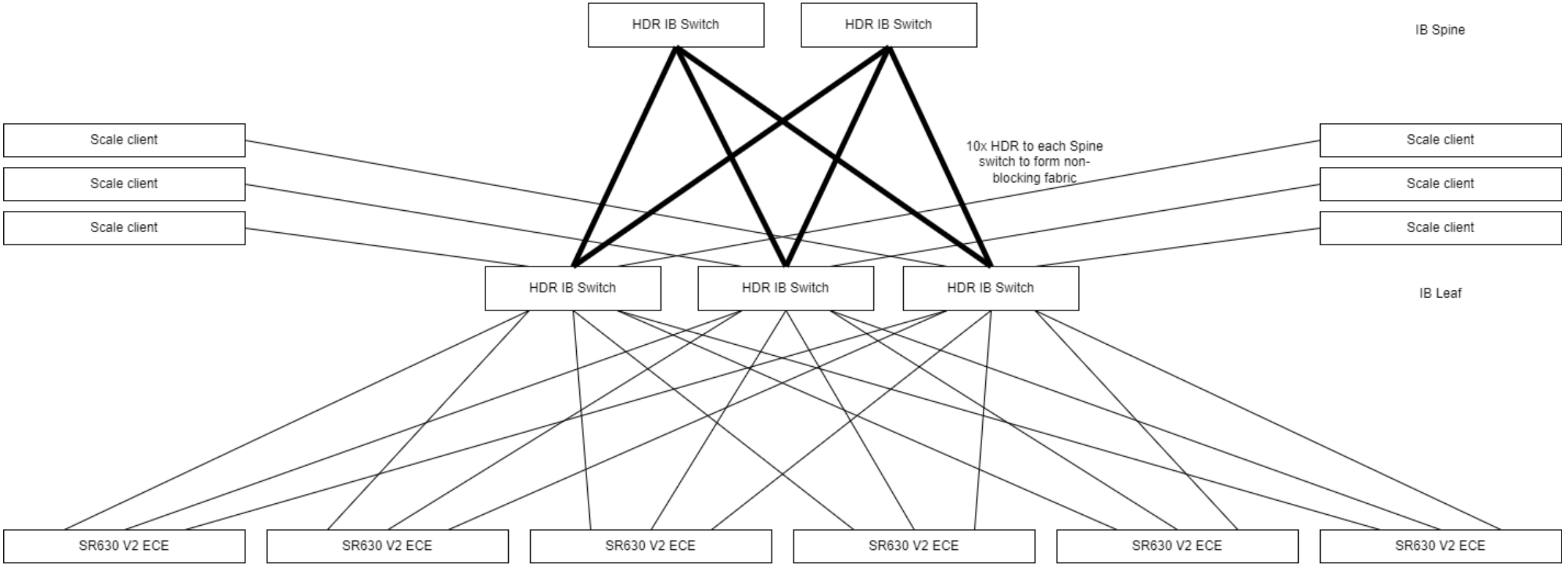
High Performance G100 ECE Cluster



ECE replication network is shared with Storage Scale Clients

Requires non-blocking HDR fabric between ECE nodes

High Performance G100 ECE Cluster Example topology



ECE replication network is shared with Storage Scale Clients

Requires non-blocking HDR fabric between ECE nodes

Split network G100 ECE Cluster

Storage Scale Clients

100Gb Ethernet

2x 100GbE ports per server (LACP bond)

verbsPorts configured between G100 servers to force ECE replication traffic over IB fabric backend network

2x HDR ports per server

HDR IB Fabric

Special config requires DSS-G approval, 2x ConnectX-6 HDR adapters + 1x ConnectX-6 HDR-100 (dual port) adapter

Requires non-blocking HDR fabric between ECE nodes

Lenovo DSS-G 100 configuration capacities

Configuration	SR630 servers	SR650 servers	D3284 drive enclosures	D1224 drive enclosures	Number of drives (min total capacity)	Number of drives (max total capacity)
DSS G100 ECE	*1*	*0*	*0*	*0*	*4x 2.5" (3TB)**	*10x 2.5" (61TB)**
DSS G100 ECE6	6	0	0	0	24x 2.5" (**19TB**)*	60x 2.5'' (921TB)*
DSS G100 ECE10	10	0	0	0	40x 2.5'' (32TB)*	100x 2.5'' (1536TB)*
DSS G100 ECE11	11	0	0	0	44x 2.5" (35TB)*	110x 2.5'' (1689TB)*
DSS G100 ECE32	32	0	0	0	128x 2.5" (102TB)*	320x 2.5'' (**4915TB**)*

Erasure Code Edition is limited to 32 servers in a single recovery group. Multiple recovery groups withing the same file-system are possible.

* Capacity is based on using 800GB (min) or 15.36TB (max) 2.5-inch NVMe SSDs.