



# Lenovo Solutions with ECE ( Spectrum Scale Erasure Code Edition )

Lenovo™

Spectrum Scale Strategy Days, 04-Mar-2021 | Michael Hennecke

# Lenovo Documentation for DSS-G, and ECE on DSS-G100



## Lenovo Distributed Storage Solution for IBM Spectrum Scale (DSS-G) (ThinkSystem based) Product Guide

Lenovo Distributed Storage Solution for IBM Spectrum Scale (DSS-G) is a software-defined storage (SDS) solution for dense scalable file and object storage suitable for high-performance and data-intensive environments. Enterprises or organizations running HPC, Big Data or cloud workloads will benefit the most from the DSS-G implementation.

DSS-G combines the performance of the Lenovo ThinkSystem SR650 servers, Lenovo D1224 and D3284 storage enclosures, and industry leading IBM Spectrum Scale software to offer a high performance, scalable building block approach to modern storage needs.

Lenovo DSS-G is delivered as a pre-integrated, easy-to-deploy rack-level engineered solution that dramatically reduces time-to-value and total cost of ownership (TCO). All DSS-G base offerings described in this product guide are built on Lenovo ThinkSystem SR650 servers, Lenovo Storage D1224 Drive Enclosures with high-performance 2.5-inch SAS solid-state drives, and Lenovo Storage D3284 High-Density Drive Enclosures with large capacity 3.5-inch NL SAS HDDs.

Combined with IBM Spectrum Scale (formerly IBM General Parallel File System, GPFS), an industry leader in high-performance clustered file system, you have an ideal solution for the ultimate file and object storage solution for HPC and Big Data.

**Did you know?**

The DSS-G solution gives you the choice of shipping fully integrated into the Lenovo 1410 rack cabinet, or with the Lenovo Client Site Integration Kit, 7X74, which allows you to have Lenovo install the solution in a rack of your own choosing. In either case, the solution is tested, configured, and ready to be plugged in and turned on; it is designed to integrate into an existing infrastructure effortlessly, to dramatically accelerate time to value and reduce infrastructure maintenance costs.

Lenovo DSS-G is licensed by the number of drives installed, rather than the number of processor cores or the number of connected clients, so there are no added licenses for other servers or clients that mount and work with the file system.


Lenovo provides a single point of entry for supporting the entire DSS-G solution, including the IBM Spectrum Scale software, for quicker problem determination and minimized downtime.



Figure 1. Lenovo DSS-G Model G260

[Click here to check for updates](#)

Lenovo Distributed Storage Solution for IBM Spectrum Scale (DSS-G) (ThinkSystem based) 1



## DSS-G Declustered RAID Technology and Rebuild Performance


Provides an overview of the Spectrum Scale RAID technology

Introduces the Spectrum Scale RAID terminology

Explains how to calculate the volume of critical and non-critical rebuilds

Demonstrates the rebuild performance of the DSS-G declustered RAID technology

Michael Hennecke



[Click here to check for updates](#)



## Lenovo ThinkSystem SR630 Server (Xeon SP Gen 2) Product Guide

Lenovo ThinkSystem SR630 is an ideal 2-socket 1U rack server for small businesses up to large enterprises that need industry-leading reliability, management, and security, as well as maximizing performance and flexibility for future growth. The SR630 server is designed to handle a wide range of workloads, such as databases, virtualization and cloud computing, virtual desktop infrastructure (VDI), infrastructure security, systems management, enterprise applications, collaboration/email, streaming media, web, and HPC.

Featuring the second generation of the Intel Xeon Processor Scalable Family (Xeon SP Gen 2), the SR630 server offers scalable performance and storage capacity. The SR630 server supports up to two processors, up to 3 TB of memory capacity with TruDDR4 DIMMs or up to 7.5 TB of memory capacity with a combination of TruDDR4 DIMMs and Intel DC persistent memory modules (DCPMMs), up to 12x 2.5-inch or 4x 3.5-inch drive bays with an extensive choice of NVMe PCIe SSDs, SAS/SATA SSDs, and SAS/SATA HDDs, and flexible I/O expansion options with a LOM slot, a dedicated storage controller slot, and up to 3x PCIe slots.

The following figure shows the Lenovo ThinkSystem SR630 with 2.5-inch hot-swap drives.




Figure 1 Lenovo ThinkSystem SR630 with 2.5-inch drive bays

**Did you know?**

The SR630 server features a unique AnyBay design that allows a choice of drive interface types in the same drive bay: SAS drives, SATA drives, or U.2 NVMe PCIe drives.

The SR630 server offers onboard NVMe PCIe ports that allow direct connections to the U.2 NVMe PCIe SSDs, which frees up I/O slots and helps lower NVMe solution acquisition costs.

The SR630 server delivers outstanding memory performance with Performance+ 2933 MHz DIMMs, which is achieved by supporting two-DIMMs-per-channel configurations at speeds up to 10% faster than the Intel specification defines, while still maintaining world-class reliability.

[Click here to check for updates](#)

Lenovo ThinkSystem SR630 Server (Xeon SP Gen 2) 1

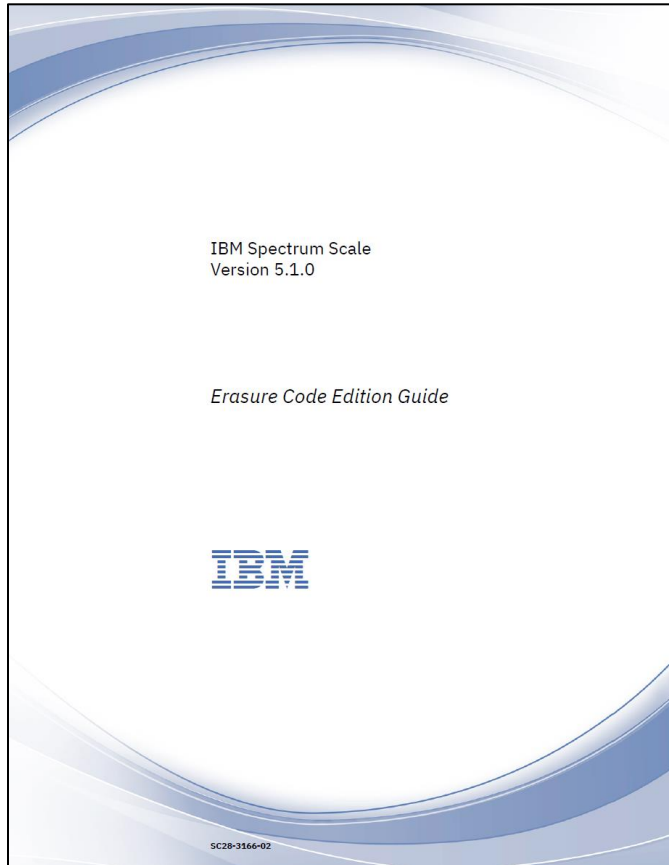
<https://lenovopress.com/lp0837>

<https://lenovopress.com/lp1227>

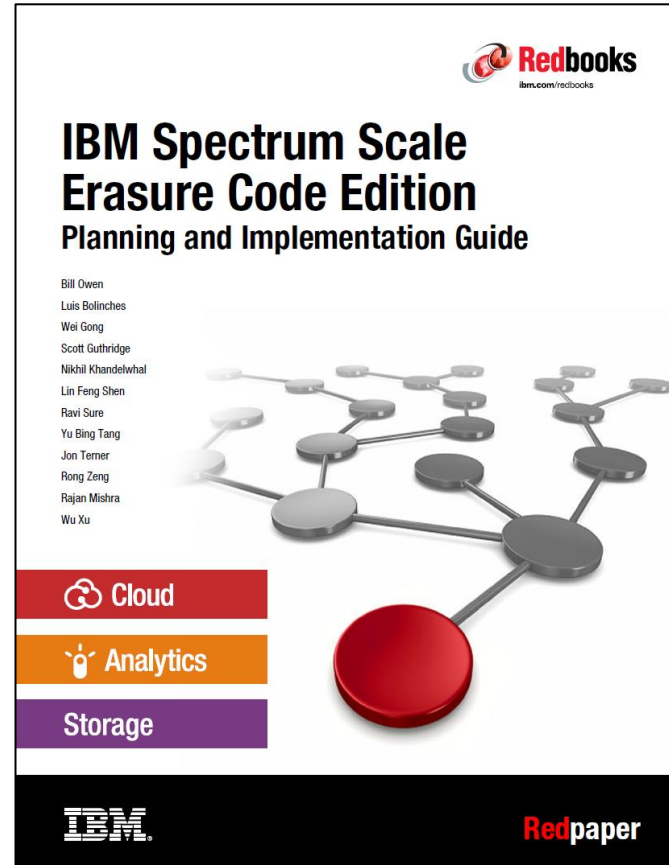
<https://lenovopress.com/lp1049>



# IBM Documentation for ECE



Scale 5.1.0: [scale\\_ece.pdf](#)  
Scale 5.1.0: [raid\\_adm.pdf](#)



<https://www.redbooks.ibm.com/abstracts/redp5557.html?Open>

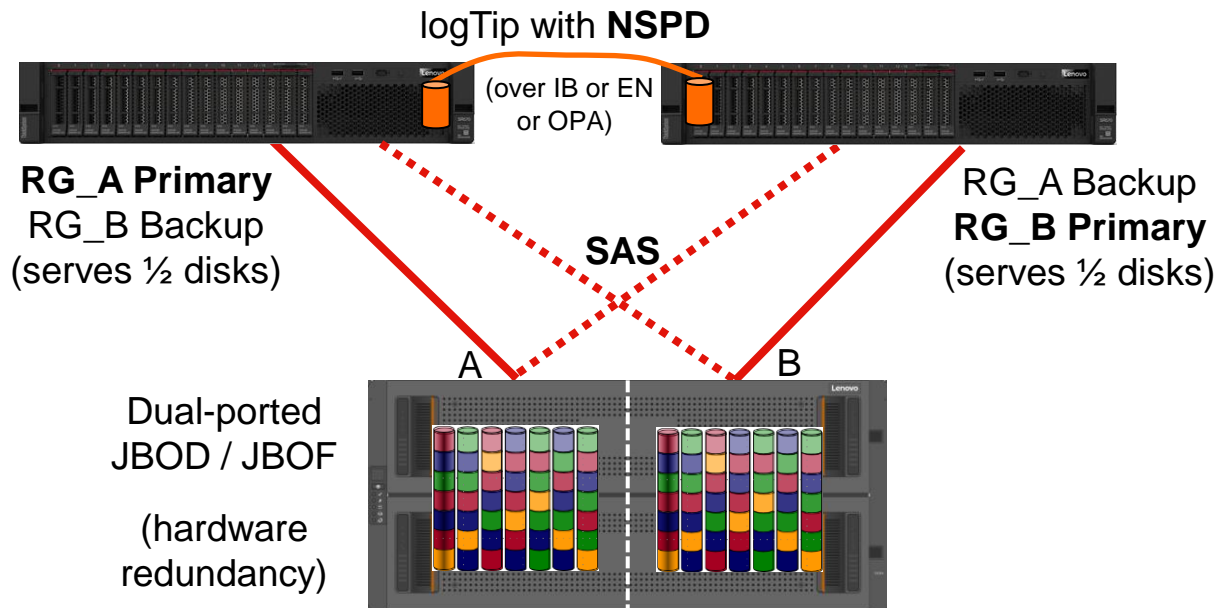


IBM Knowledge Center, e.g.  
[Outline of an mmvdisk use case](#)  
(Example is for a „paired RG“,  
but the same steps also apply for  
ECE's „scale-out RG“)

# Positioning IBM Spectrum Scale ECE (Erasure Code Edition)

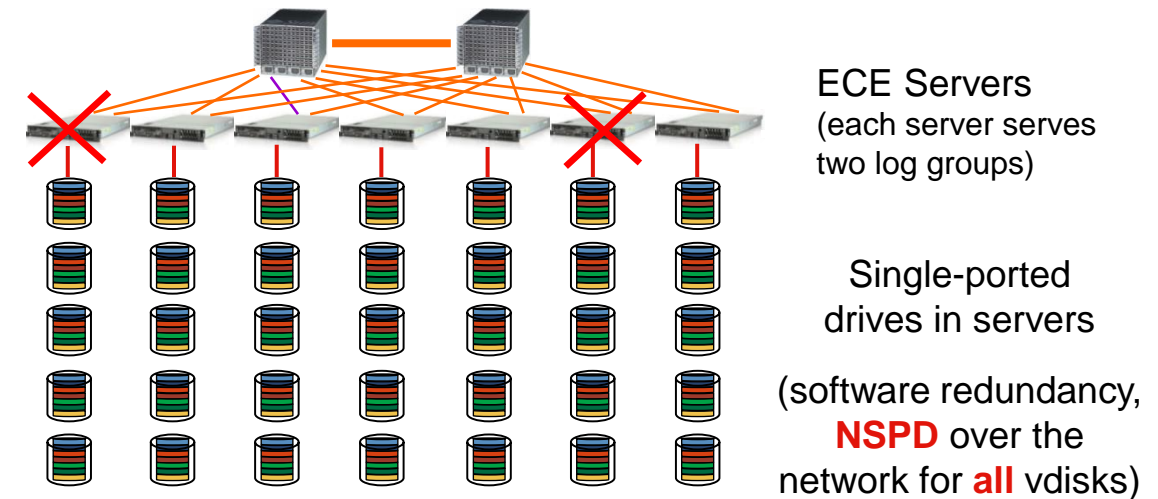
- **DSS-G2xy** is a hardware-redundant solution (dual-ported JBODs) with software RAID
- Fixed sizes; expansion @ full enclosures
- **NL-SAS** and **SSD** support (no NVMe)
- Disk- or capacity-based licensing

## DSS-G2xy with DAE / DME: Two servers + JBOD(s):



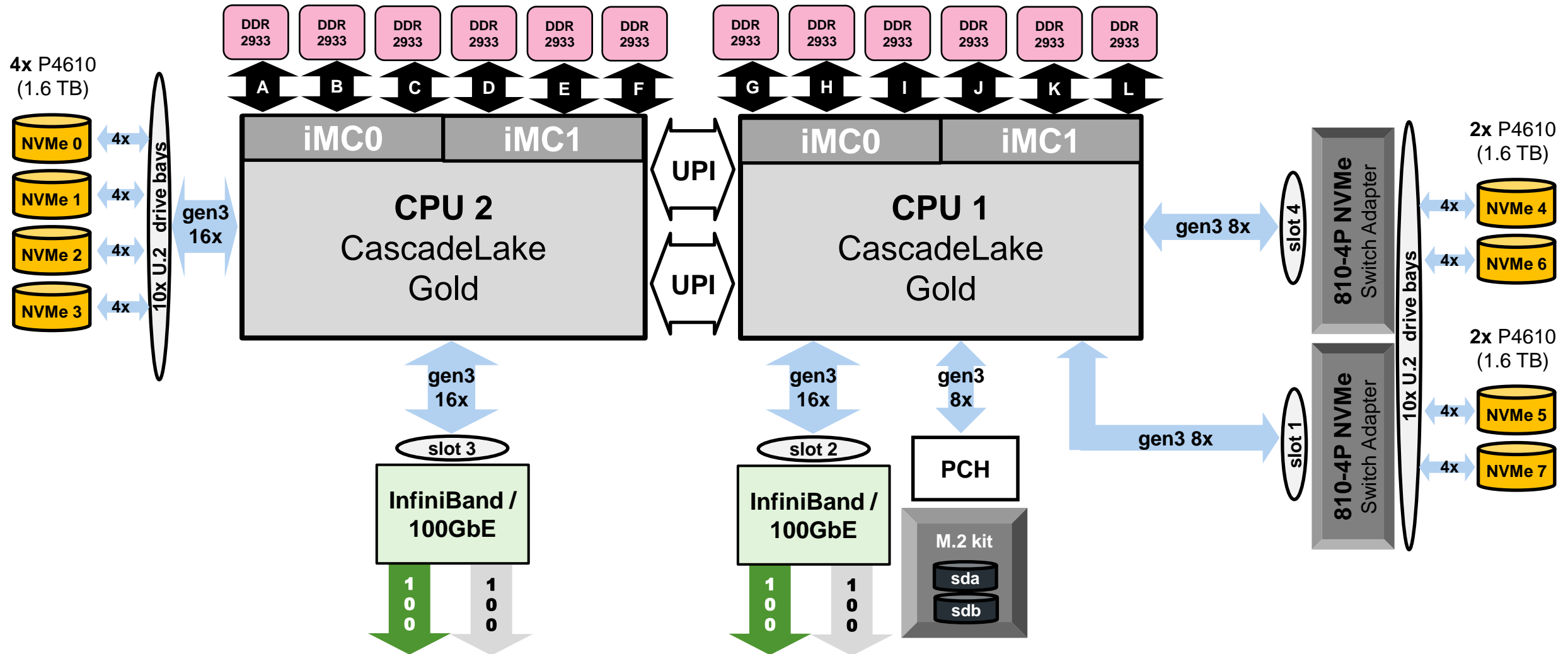
- **DSS-G100** is a server with internal **NVMe**
- **ECE** is a scale-out solution providing software redundancy (network RAID) for internal disks
- ECE cluster size: min. **6** nodes; max: 128 incremental expansion (+ 1 node); max. 32 nodes / ECE RG
- Capacity-based licensing (disk-based is WIP)

## DSS-G100 with ECE: Scale-out; ≤32 servers/RG:



ECE supports:  
4+2P, 4+3P, 8+2P, 8+3P; 3Way, 4Way repl.

# Lenovo DSS-G100 NVMe-rich Server: ThinkSystem SR630

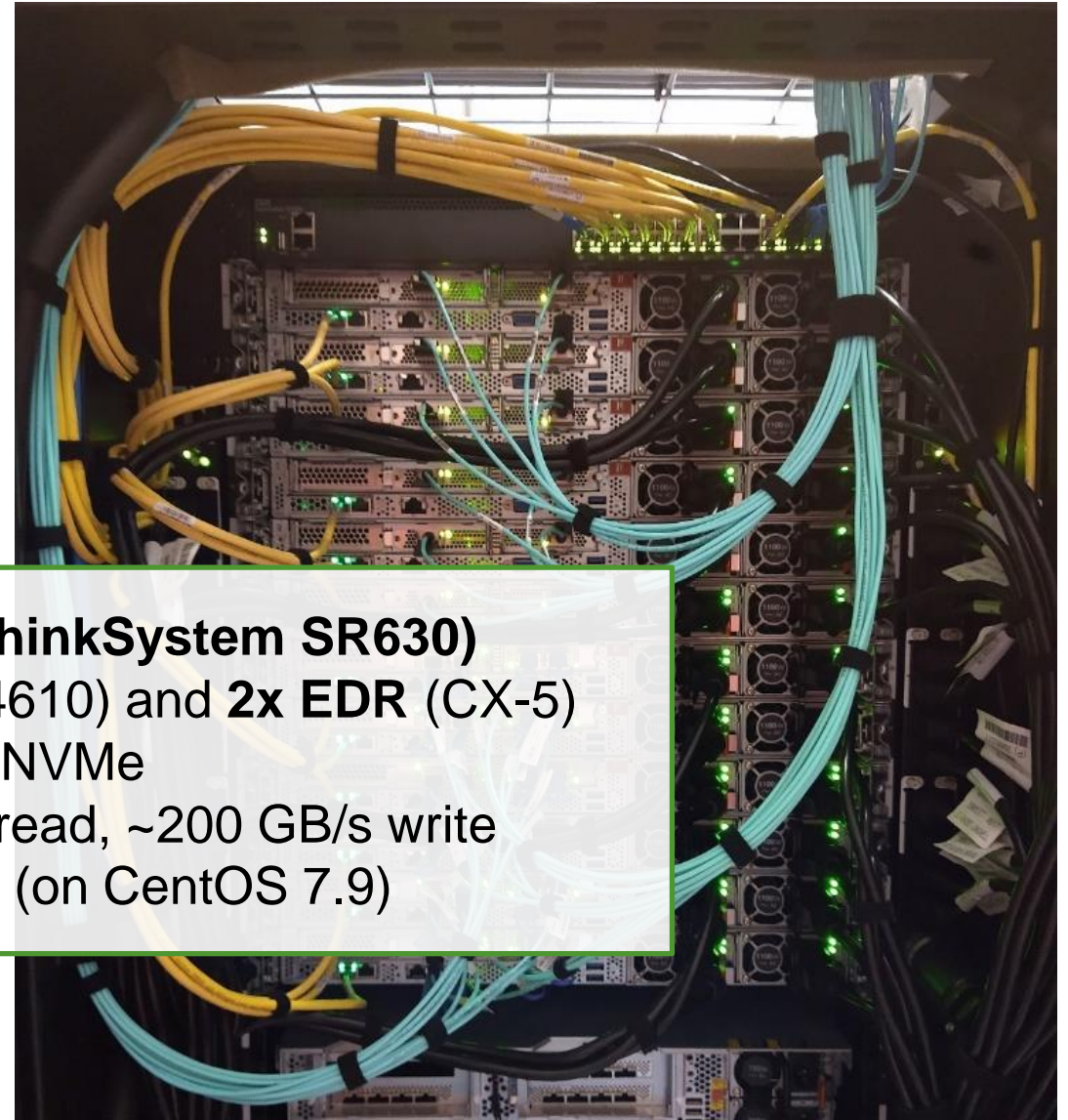
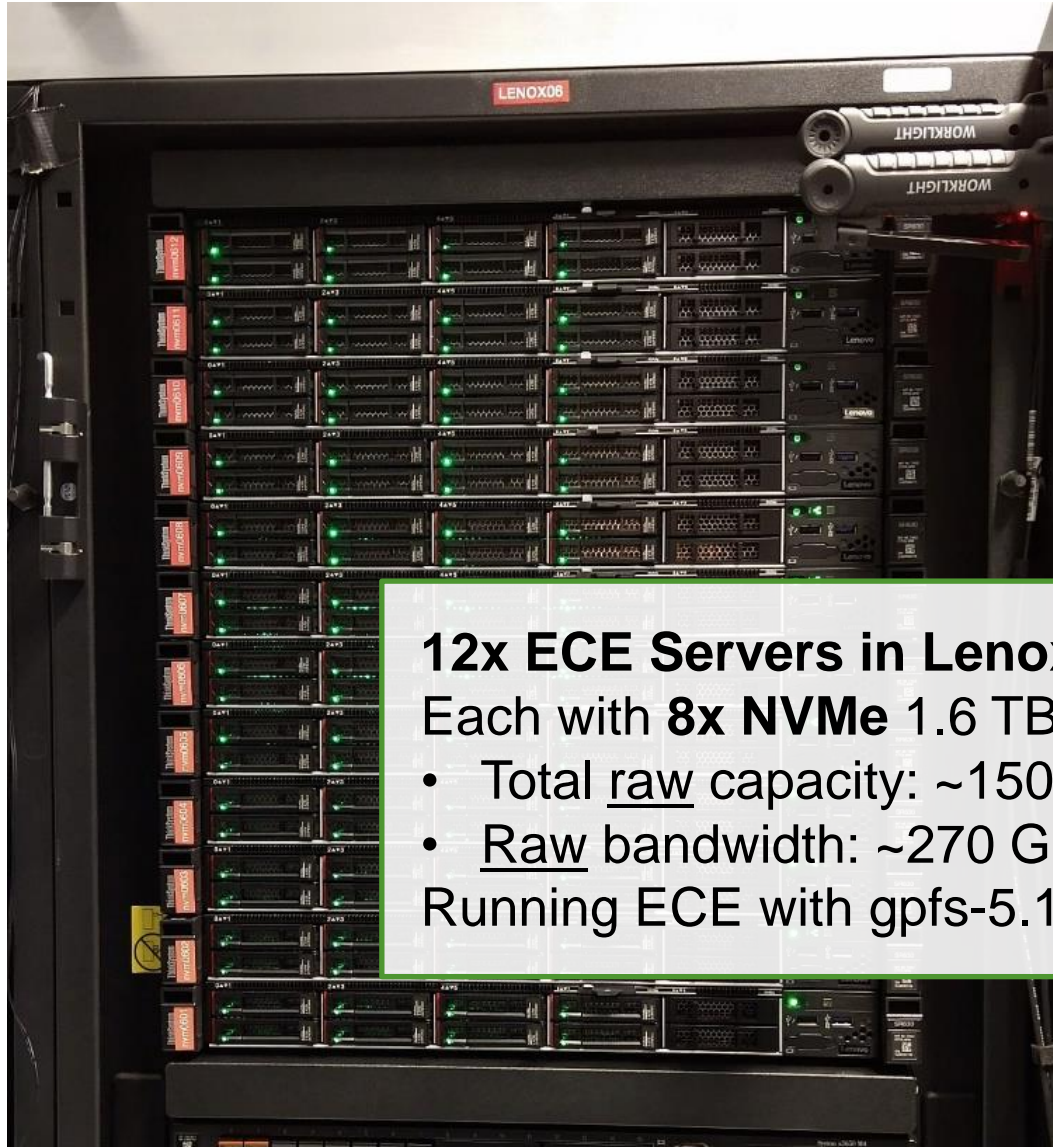








# ECE in Lenovo's HPC Innovation Center Stuttgart



**12x ECE Servers in Lenox (ThinkSystem SR630)**  
Each with **8x NVMe 1.6 TB (P4610)** and **2x EDR (CX-5)**

- Total raw capacity: ~150 TB NVMe
- Raw bandwidth: ~270 GB/s read, ~200 GB/s write

Running ECE with gpfs-5.1.0.2 (on CentOS 7.9)

# Lenovo Recommended RG Sizes for each Erasure Code

Table 4-3 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.

## 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

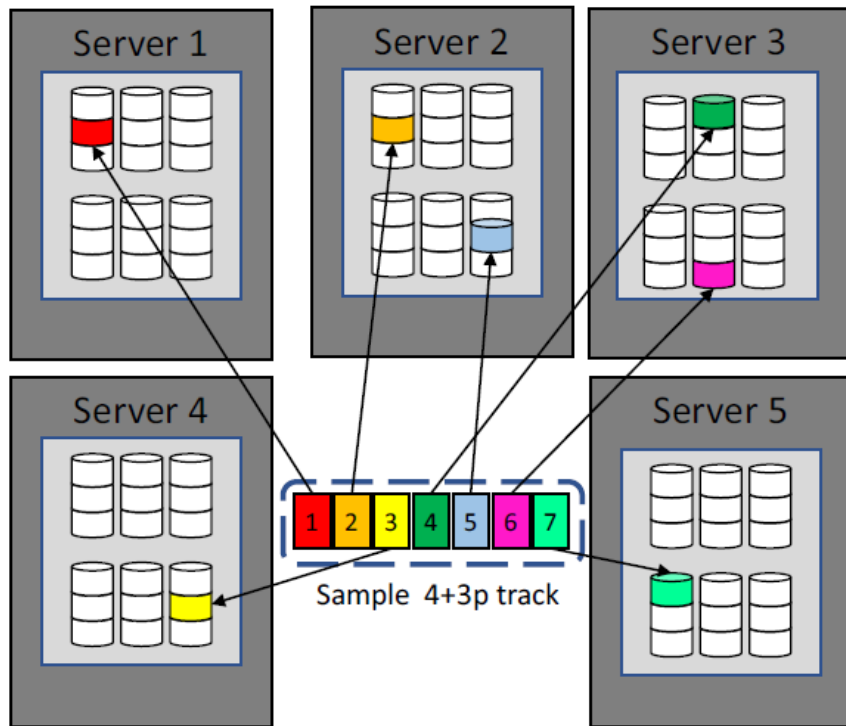
## Recommendation:

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

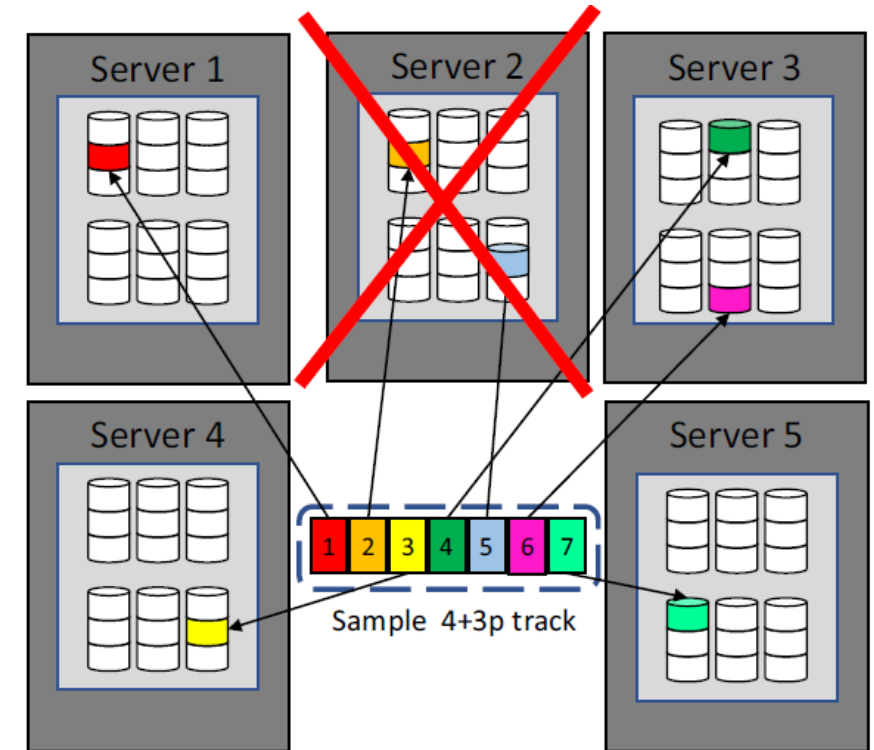
- 8+ Servers for 4+2P (+2P)
- 10+ Servers for 4+3P (+3P)
- 12+ Servers for 8+2P (+2P)
- 14+ Servers for 8+3P (+3P)



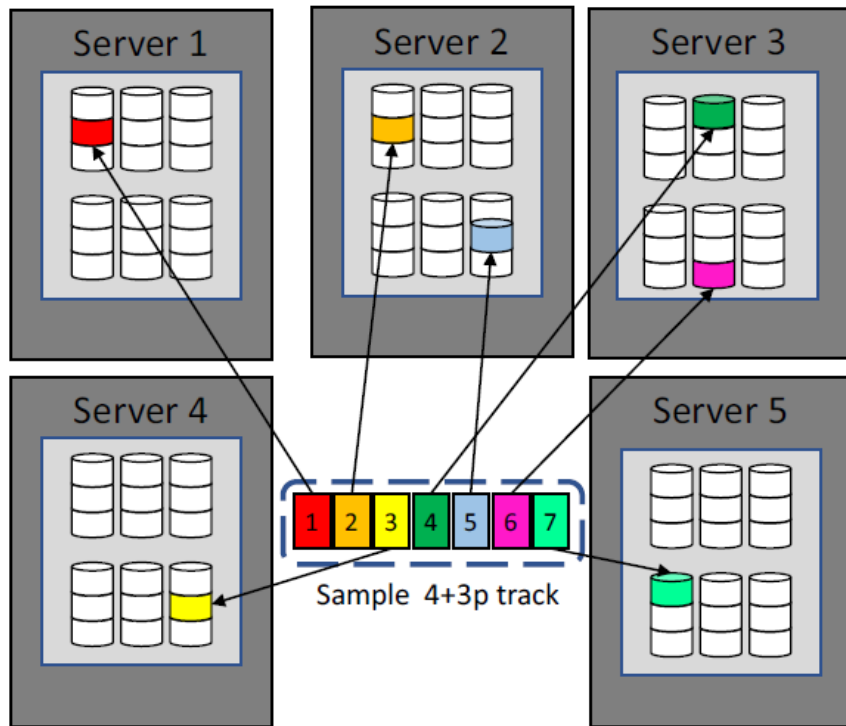
# ECE Fault Tolerance Example: 4+3P on 5 Servers



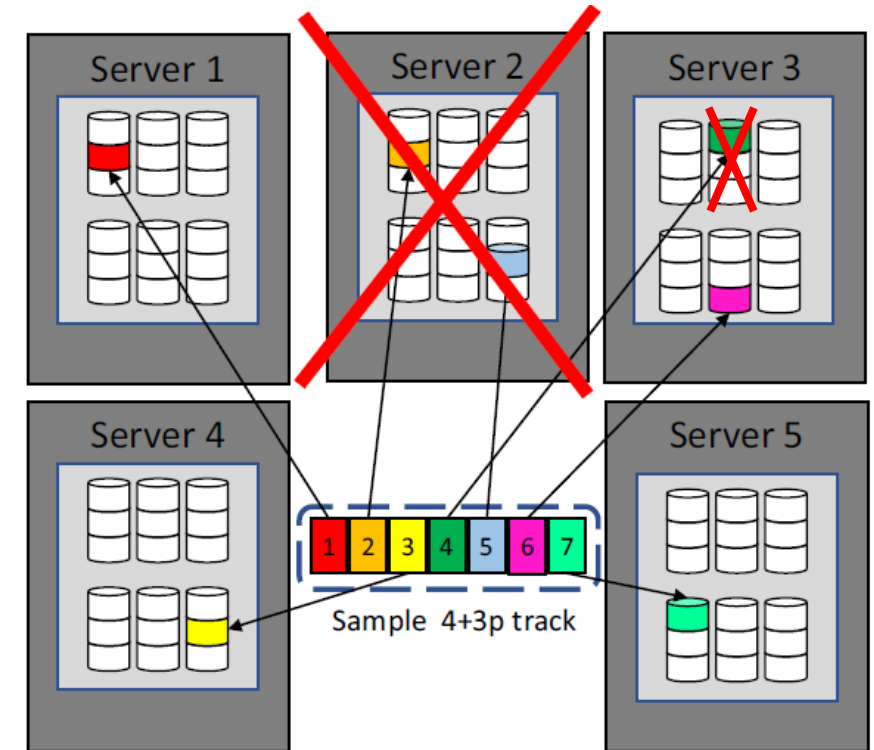
single server  
fault



# ECE Fault Tolerance Example: 4+3P on 5 Servers

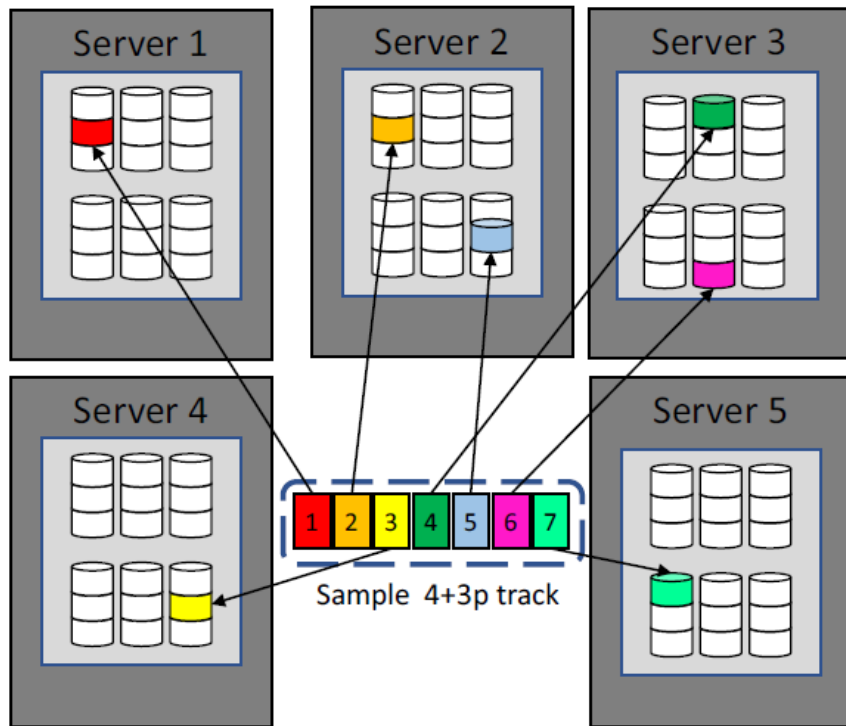


single server  
+ 1 disk fault



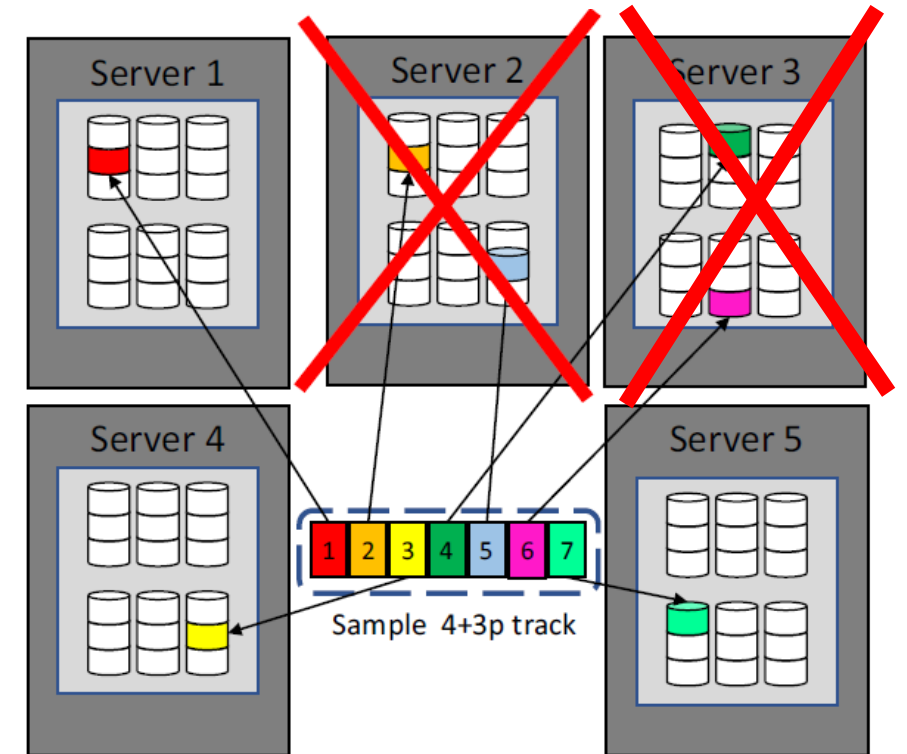


# ECE Fault Tolerance Example: 4+3P on 5 Servers



**data loss**

**double server fault**



# Intel NVMe Drive Options for DSS-G100

Other OEM vendors' NVMe drives that are supported in the SR630 should also work...

Drive Series	Storage Technology	Capacity [GB]	Sequential Read [MB/s]	Sequential Write [MB/s]	Random Read [k IOPS]	Random Write [k IOPS]	Read Latency [usec]	Write Latency [usec]	Active Power [W]	Idle Power [W]	Write Endurance [PBW]	Write Endurance [DWPD]
<b>Entry (~1 DWPD)</b>												
P4510	64layer 3D TLC NAND	1000	2850	1100	465,0	70,0	77	18	10,0	5,0	1,92	1,05
P4510	64layer 3D TLC NAND	2000	3200	2000	637,0	81,5	77	18	12,0	5,0	2,61	0,72
<b>P4510</b>	64layer 3D TLC NAND	<b>4000</b>	3000	<b>2900</b>	625,5	113,5	77	18	14,0	5,0	6,30	0,86
P4510	64layer 3D TLC NAND	8000	3200	3000	620,0	139,5	77	18	16,0	5,0	13,88	0,95
<b>Mainstream (~3-5 DWD)</b>												
P4610	64layer 3D TLC NAND	1600	3200	2100	620,0	200,0	77	18	13,3	5,0	12,25	4,20
<b>P4610</b>	64layer 3D TLC NAND	<b>3200</b>	3200	<b>3000</b>	640,0	200,0	77	18	13,8	5,0	21,85	3,74
P4610	64layer 3D TLC NAND	6400	3000	2900	640,0	220,0	77	18	14,6	5,0	36,54	3,13
<b>Performance (~30 DWPD)</b>												
P4800X	3D Xpoint	375	2400	2000	550,0	550,0	10	10	18,0	5,0	20,50	29,95
P4800X	3D Xpoint	750	2500	2000	550,0	550,0	10	10	18,0	6,0	41,00	29,95



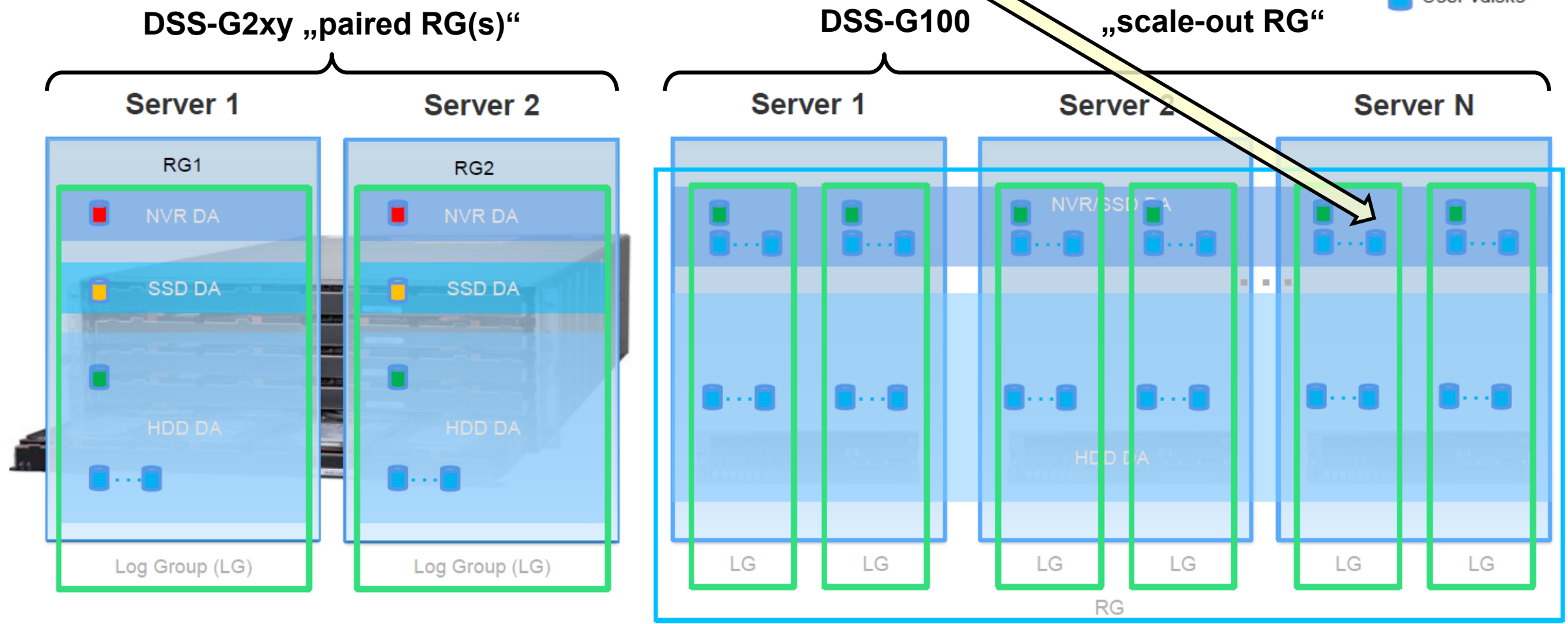
# File System Blocksizes Supported by ECE

Disk Media	4+2P, 4+3P	8+2P, 8+3P
<b>HDD</b> (NL-SAS)	1M, 2M, 4M (8kiB subblocks) 8M (16kiB subblocks)	1M, 2M, 4M (8kiB subblocks) 8M, <b>16M</b> (16kiB subblocks)
<b>Flash</b> (SAS-SSD, NVMe)	1M, 2M (8kiB subblocks)	1M, 2M, <b>4M</b> (8kiB subblocks)

# Hardware Resource Partitioning – Log Groups etc.

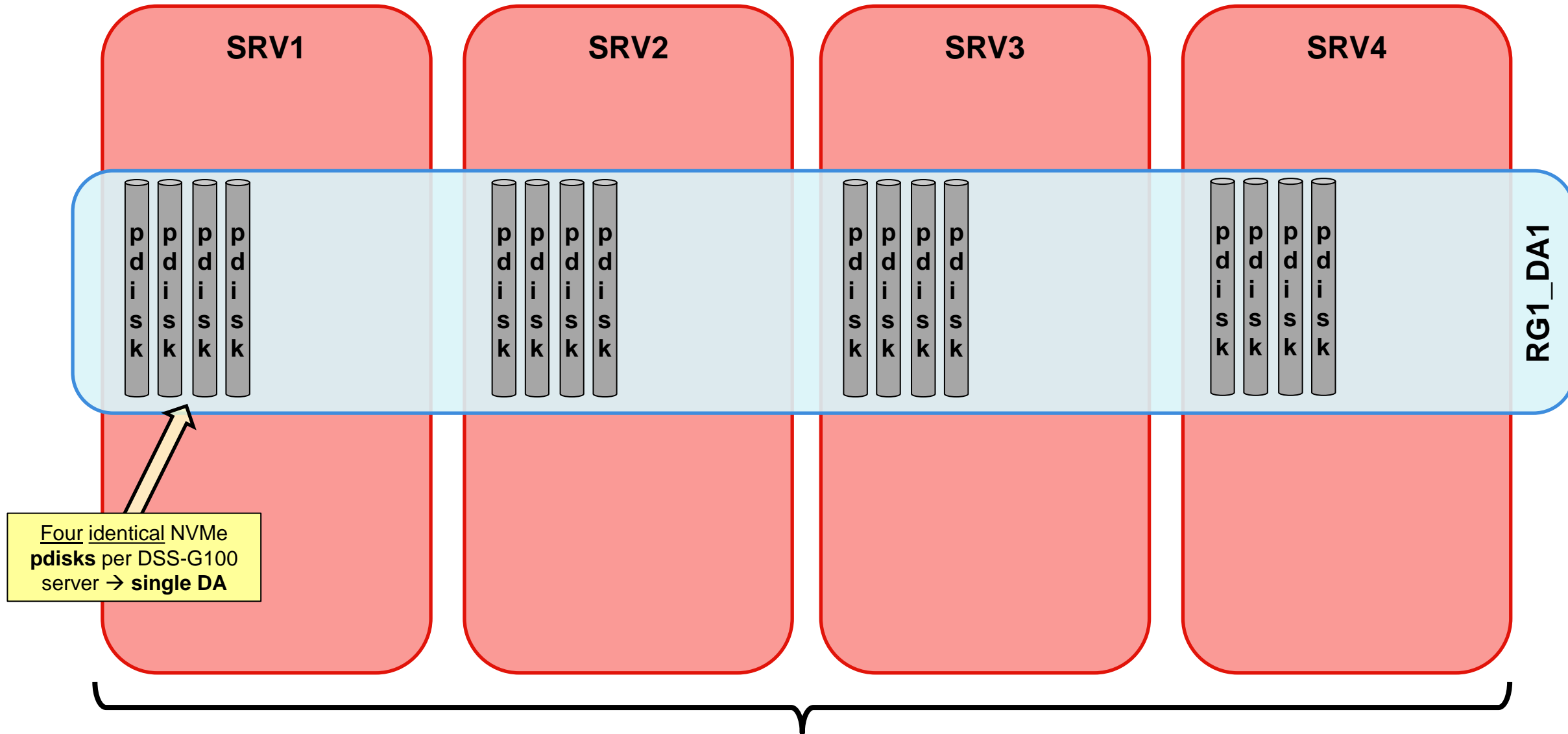
Disk icons highlight **vdisk ownership** by a Log Group (LG), **not pdisk location**.  
The actual ECE vdisks are distributed / declustered across **all** the servers in the **RG**.  
Transactions for the **user** vdisks are logged to the **logHome** vdisk owned by the same **LG**.

- Logtip vdisk
- Logtip backup vdisk
- Loghome vdisk
- User vdisks

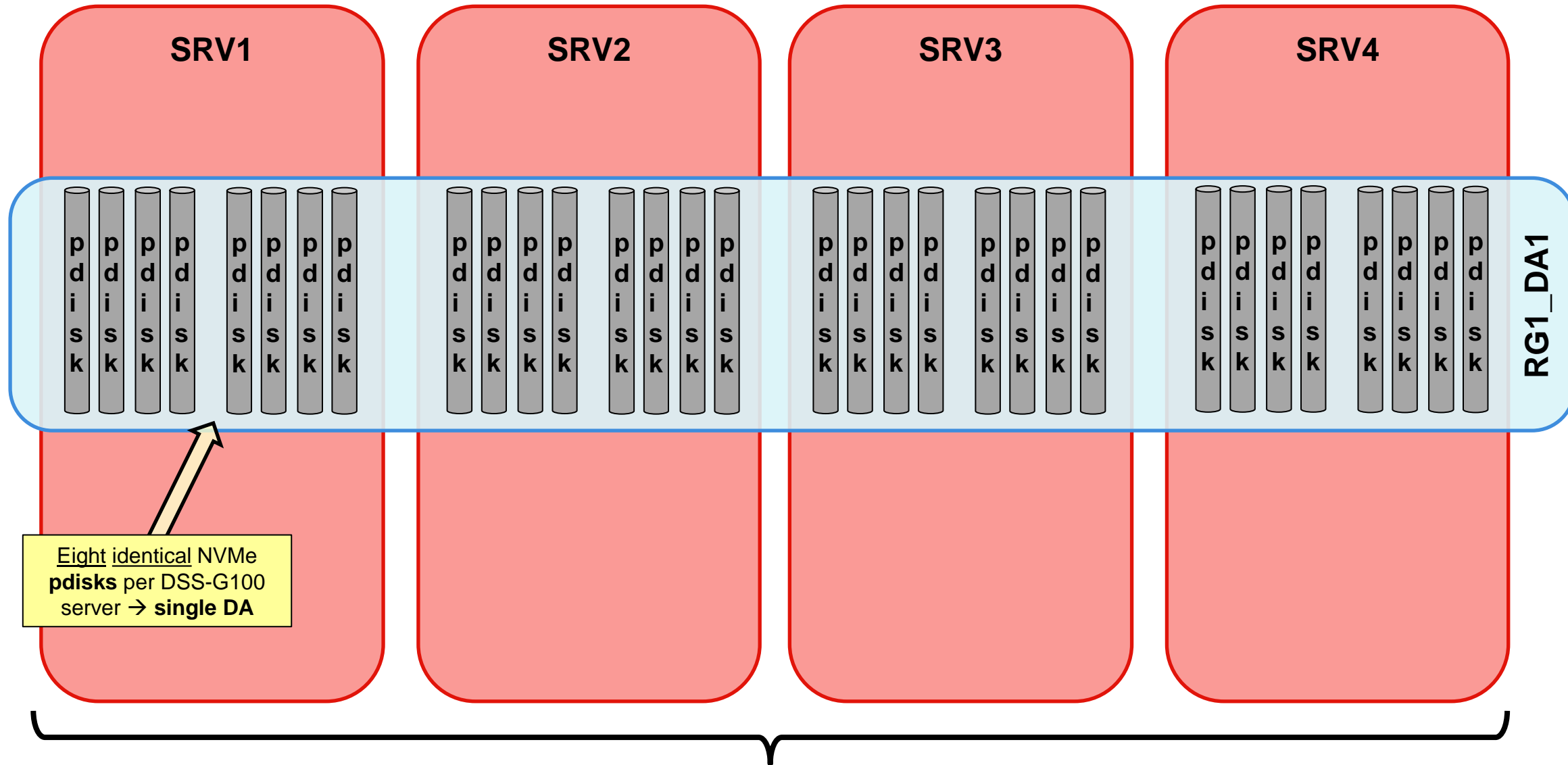




# Spectrum Scale RAID Components on ECE @ DSS-G100

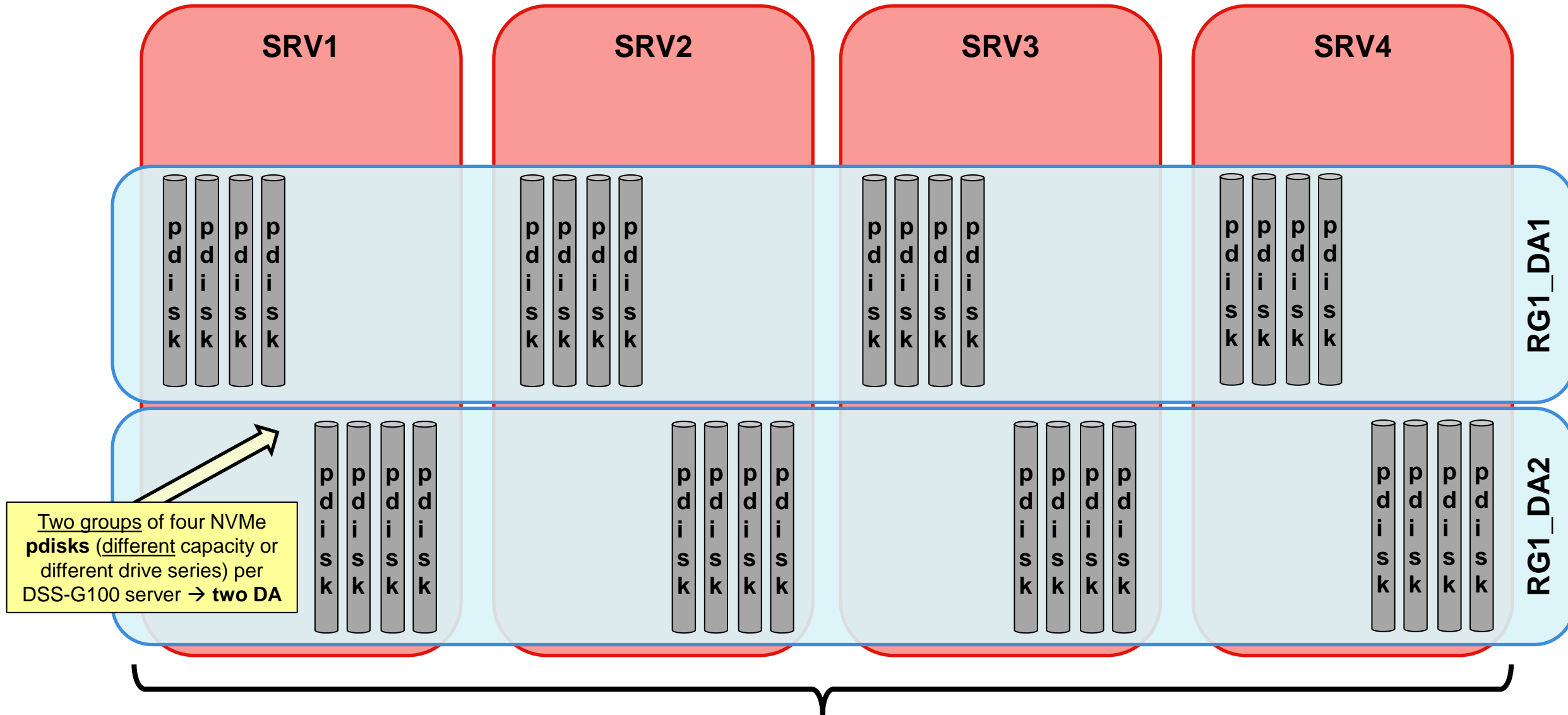


# Spectrum Scale RAID Components on ECE @ DSS-G100

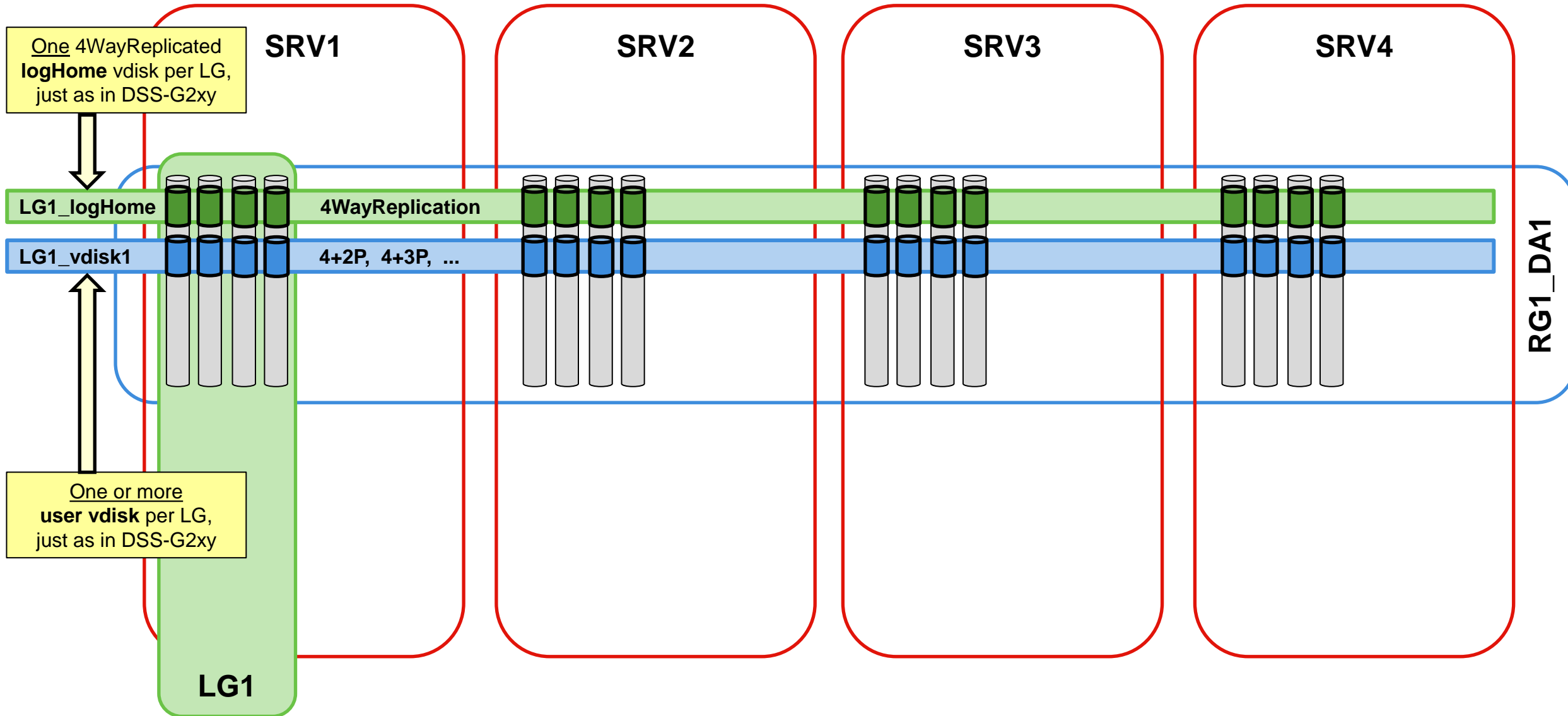




# Spectrum Scale RAID Components on ECE @ DSS-G100

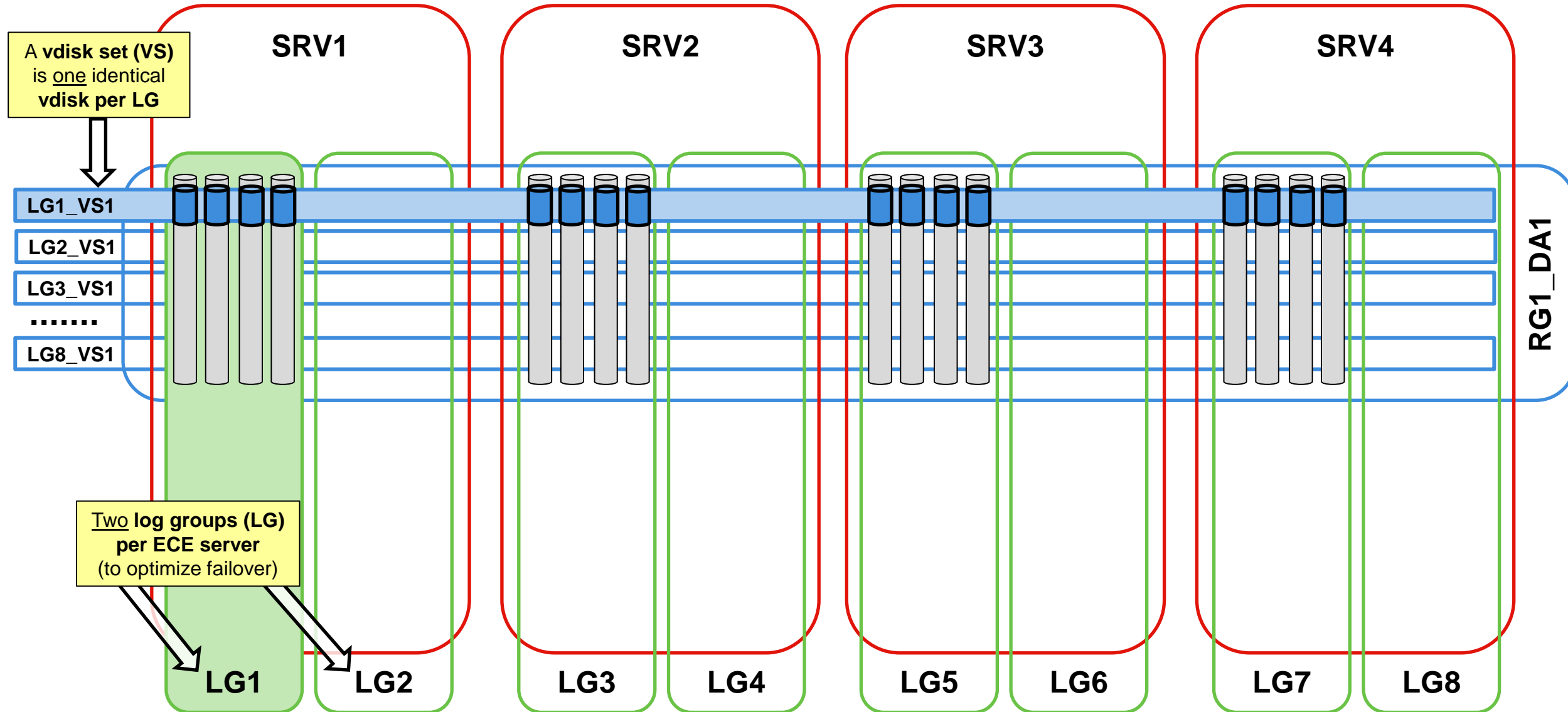


# Spectrum Scale RAID Components on ECE @ DSS-G100

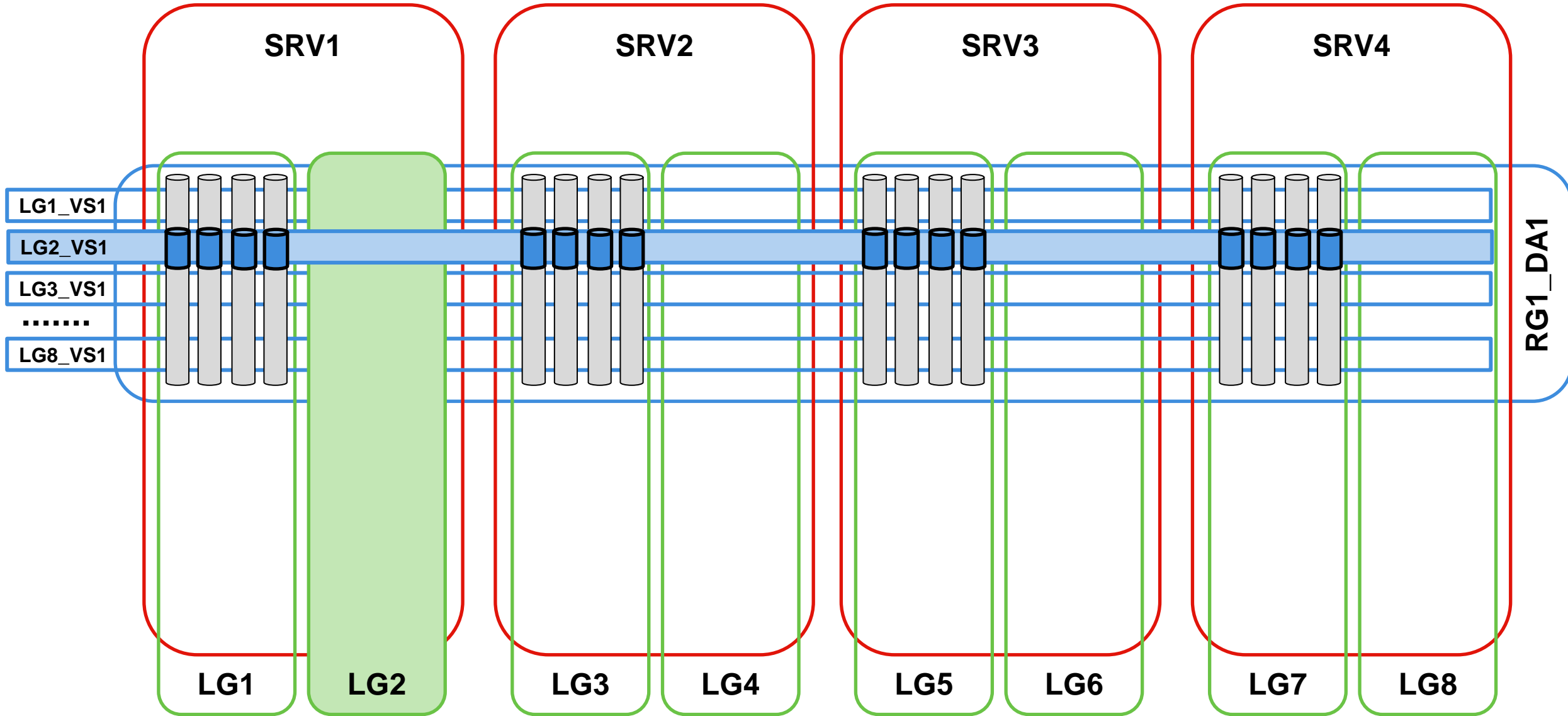




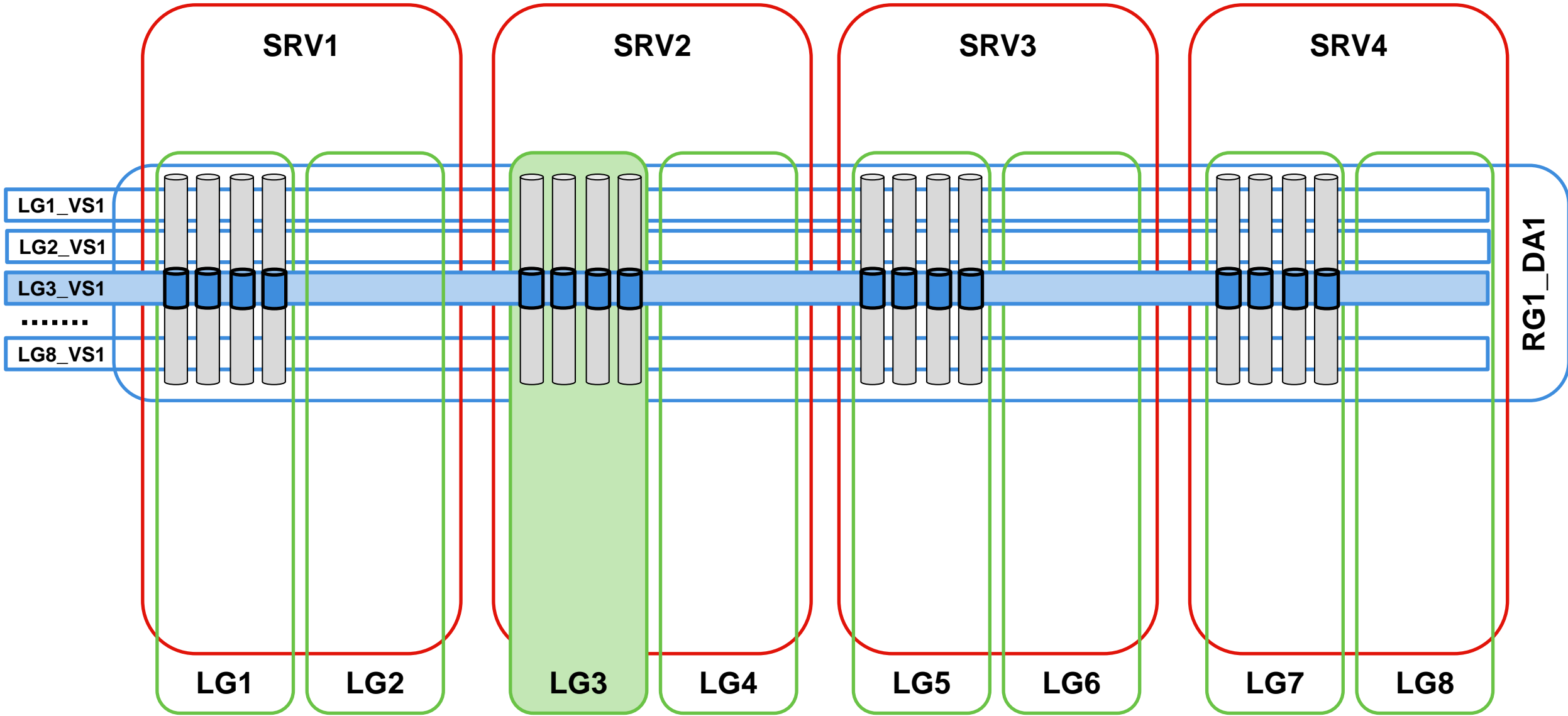
# Spectrum Scale RAID Components on ECE @ DSS-G100



# Spectrum Scale RAID Components on ECE @ DSS-G100

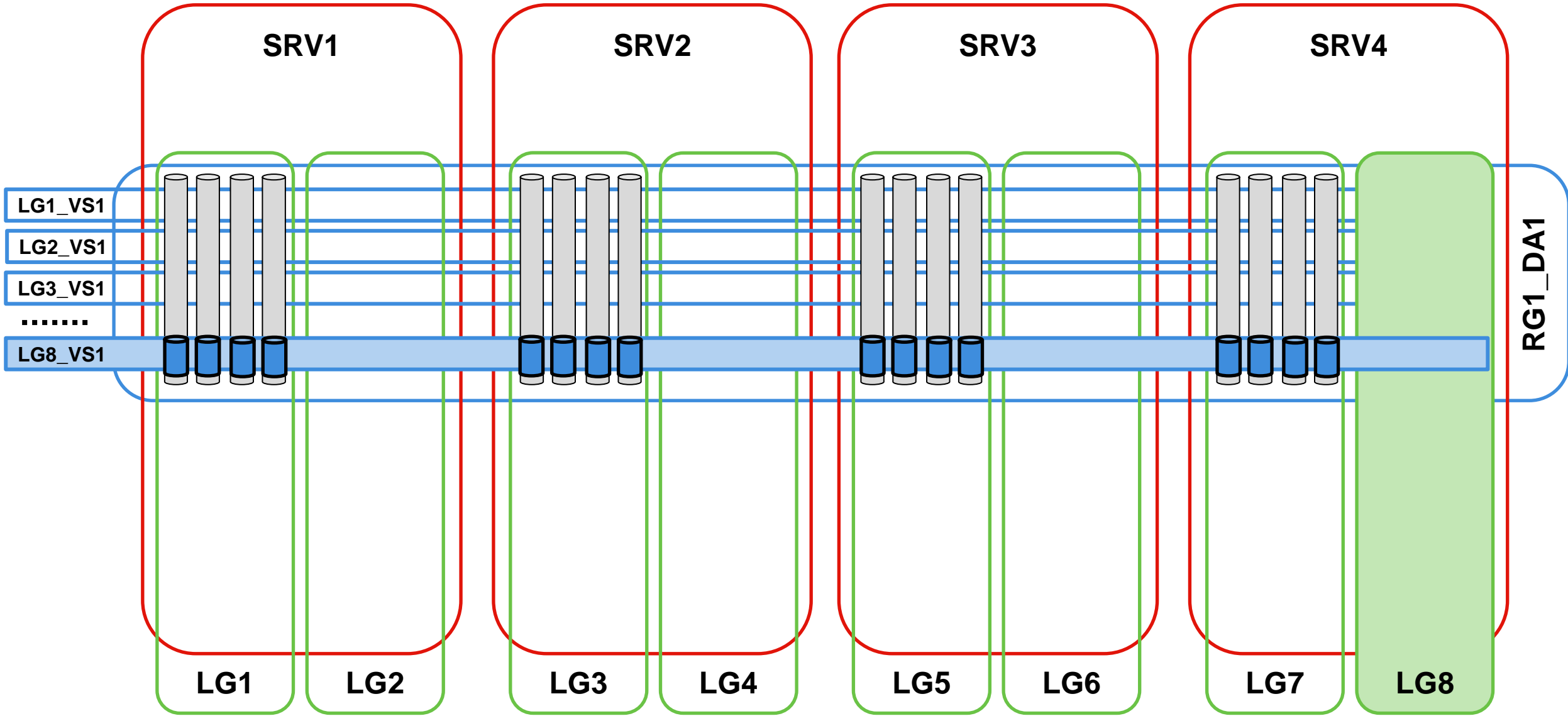


# Spectrum Scale RAID Components on ECE @ DSS-G100

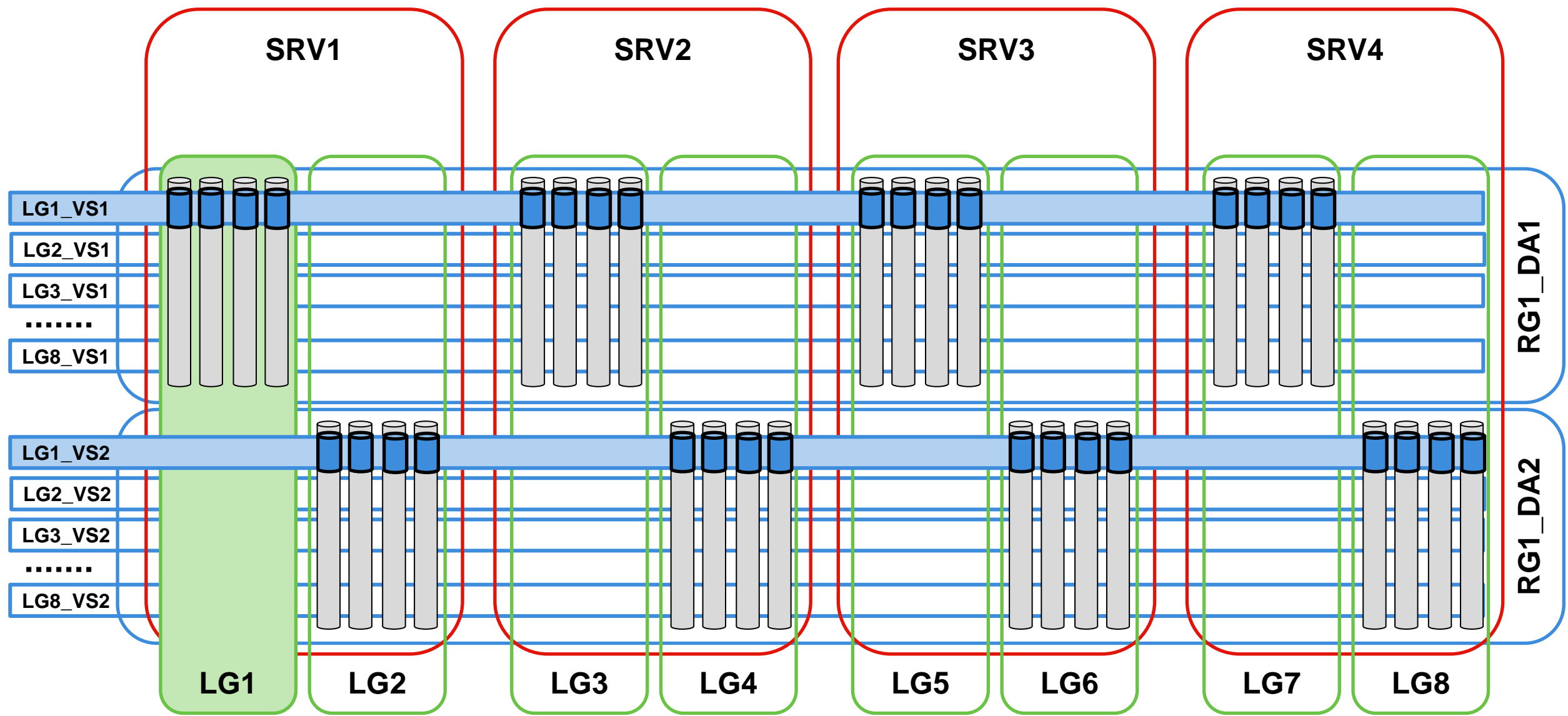




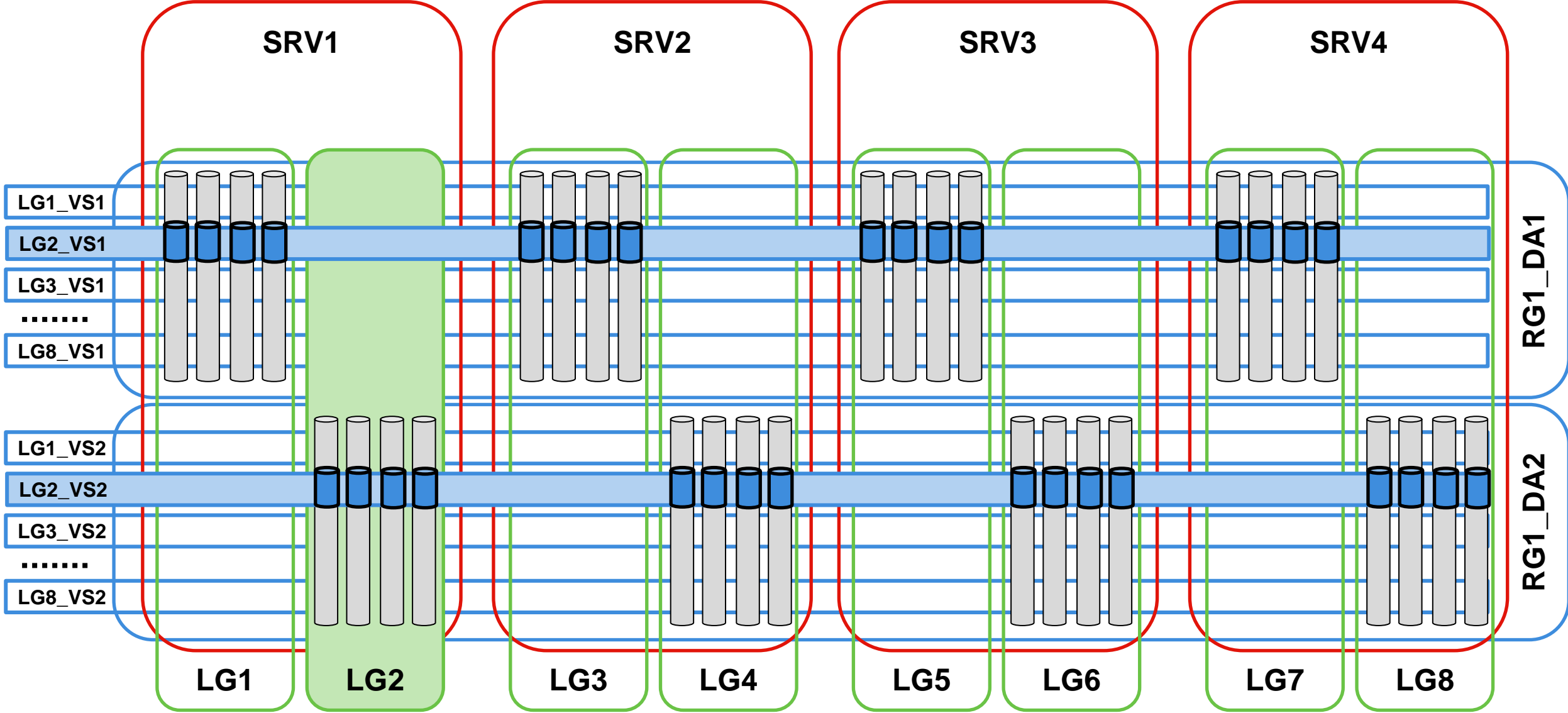
# Spectrum Scale RAID Components on ECE @ DSS-G100



# Spectrum Scale RAID Components on ECE @ DSS-G100

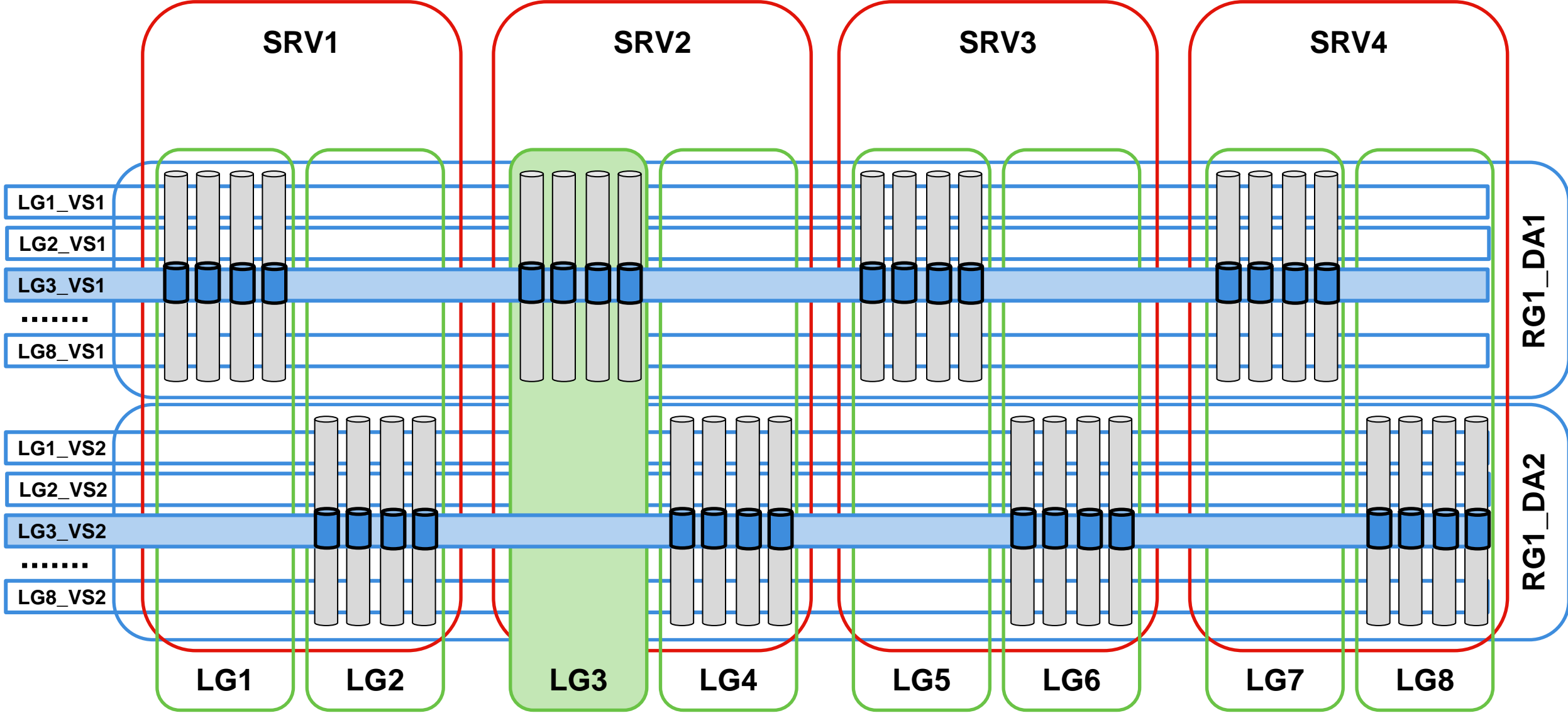


# Spectrum Scale RAID Components on ECE @ DSS-G100

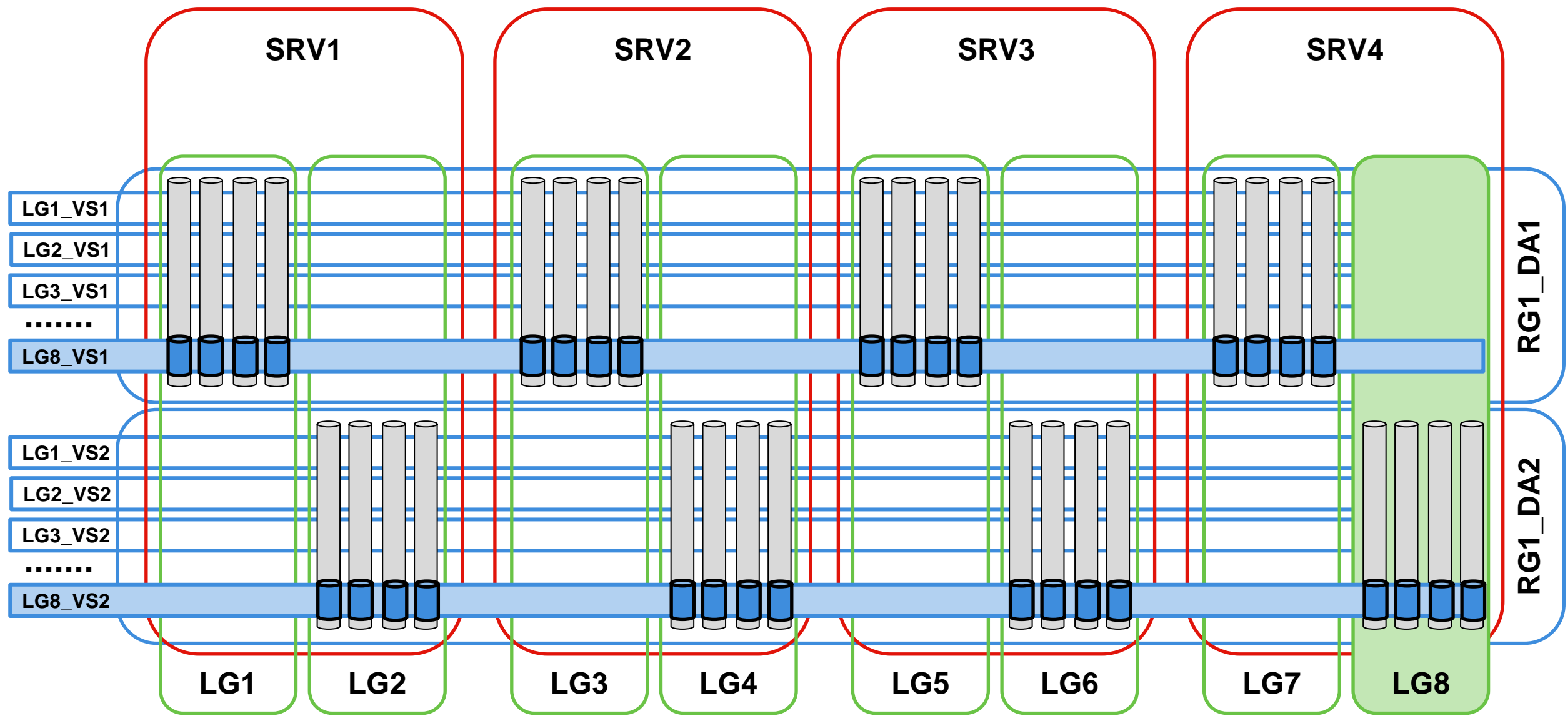




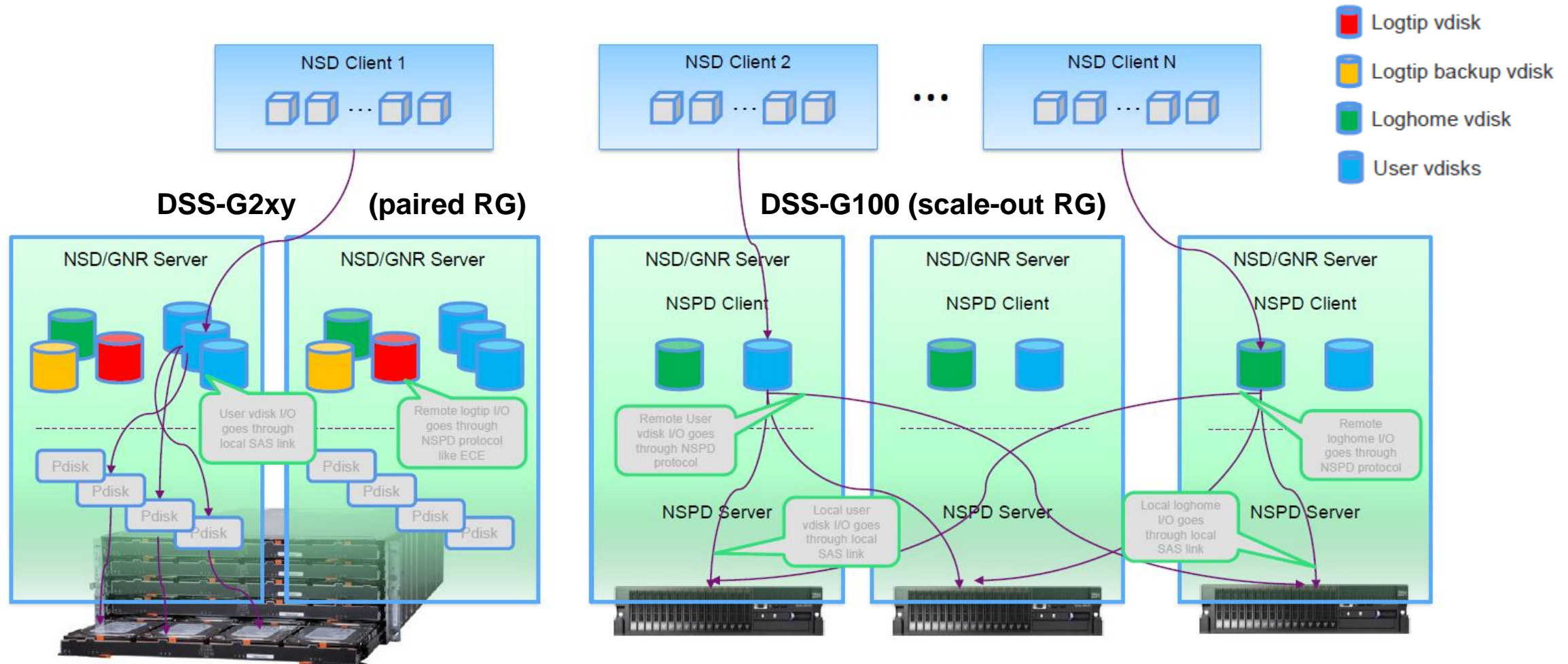
# Spectrum Scale RAID Components on ECE @ DSS-G100



# Spectrum Scale RAID Components on ECE @ DSS-G100



# NSD I/O Path – DSS-G2xy versus ECE@DSS-G100



# Summary

- Spectrum Scale Erasure Code Edition (ECE), as a superset of Scale DME, is now added to the IBM / Lenovo OEM agreement
  - Lenovo can sell fully integrated ECE solutions as of now (DSS-G100)
- DSS-G100 @ ECE is an excellent scale-out solution for Scale on NVMe
  - Pay attention to Lenovo's minimum and recommended ECE cluster sizes...
  - Software RAID setup is managed through the **mmvdisk** command set
- ECE read bandwidth is typically good
  - currently debugging an odd read performance issue with gpfs-5.1.0.2
- ECE write performance work is ongoing (target: DSS-G 3.2, ~June 2021):
  - Need to do more testing with GNR **TRIM** support enabled (available since gpfs-5.0.5)
  - Network performance analysis / tuning needed for **NSPD** traffic at wirespeed loads
    - „fat“ front-end network, versus „split“ front-end and back-end networks



# thanks.

**mhennecke @ lenovo.com**



# Bi-Directional Traffic on a Single ConnectX-6 EDR Card/Port

```
root@cmp2641:~  
0 0 100 0 0 0 938B 478B 0 0 357 534  
0 0 100 0 0 0 1086B 318B 0 0 481 739  
0 0 100 0 0 0 1653B 2342B 0 0 616 585  
0 0 100 0 0 0 1480B 566B 0 0 391 547  
0 0 100 0 0 0 1240B 598B 0 0 375 553  
0 0 100 0 0 0 1404B 476B 0 0 446 688  
0 0 100 0 0 0 786B 318B 0 0 333 523  
0 2 97 0 0 0 8667k 9365M 0 0 248k 34k  
0 6 93 0 0 0 1 2760M 11G 0 0 352k 73k  
0 15 79 0 0 0 6 12G 11G 0 0 448k 131k  
0 18 77 0 0 0 5 12G 11G 0 0 431k 118k  
0 18 78 0 0 0 5 12G 11G 0 0 427k 116k  
0 18 79 0 0 0 4 12G 11G 0 0 408k 122k  
0 17 79 0 0 0 4 12G 11G 0 0 391k 160k  
0 17 79 0 0 0 3 12G 11G 0 0 388k 154k  
0 17 79 0 0 0 4 12G 11G 0 0 405k 198k  
0 17 79 0 0 0 4 12G 11G 0 0 389k 186k  
0 17 80 0 0 0 3 12G 11G 0 0 336k 161k  
0 18 79 0 0 0 3 12G 11G 0 0 335k 150k  
0 17 80 0 0 0 3 12G 11G 0 0 338k 171k  
0 17 79 0 0 0 3 12G 11G 0 0 384k 184k  
0 18 79 0 0 0 3 12G 11G 0 0 360k 161k  
0 17 79 0 0 0 3 12G 11G 0 0 383k 167k  
0 18 79 0 0 0 3 12G 11G 0 0 380k 164k  
0 17 79 0 0 0 4 12G 11G 0 0 359k 164k  
0 17 79 0 0 0 4 12G 11G 0 0 343k 153k  
0 17 79 0 0 0 4 12G 11G 0 0 364k 170k  
0 17 79 0 0 0 4 12G 11G 0 0 355k 159k  
0 17 79 0 0 0 4 12G 11G 0 0 336k 149k  
0 17 79 0 0 0 4 12G 11G 0 0 321k 139k  
0 17 79 0 0 0 3 12G 11G 0 0 346k 151k  
0 17 79 0 0 0 3 12G 11G 0 0 344k 162k  
0 17 79 0 0 0 3 12G 11G 0 0 346k 151k  
0 17 79 0 0 0 4 12G 11G 0 0 353k 134k  
0 18 78 0 0 0 4 12G 11G 0 0 355k 124k  
0 18 78 0 0 0 4 12G 11G 0 0 343k 131k  
0 17 79 0 0 0 4 12G 11G 0 0 337k 136k  
0 14 83 0 0 0 3 12G 2292M 0 0 226k 141k  
0 10 88 0 0 0 2 9081M 4346k 0 0 148k 97k  
0 0 100 0 0 0 846B 318B 0 0 323 471  
0 0 100 0 0 0 1326B 318B 0 0 425 542  
---total-cpu-usage--- --net/total-- ---paging-- ---system--  
usr sys idl wai hiq siq|_recv_ send|_in_ out|_int_ csw  
0 0 100 0 0 0 1452B 420B 0 0 504 711  
0 0 100 0 0 0 1392B 828B 0 0 437 546  
0 0 100 0 0 0 1086B 318B 0 0 385 491
```

```
root@cmp2641:~  
[root@cmp2641 ~]# iperf -c cmp2642-ib0 -l 2m -P 16 -t 30 | grep SUM  
[SUM] 0.0-30.0 sec 332 GBytes 94.9 Gbits/sec  
[root@cmp2641 ~]# iperf -c cmp2642-ib0 -l 4m -P 16 -t 30 | grep SUM  
[SUM] 0.0-30.0 sec 331 GBytes 94.9 Gbits/sec  
[root@cmp2641 ~]# iperf -c cmp2642-ib0 -l 1m -P 16 -t 30 | grep SUM  
[SUM] 0.0-30.0 sec 332 GBytes 95.0 Gbits/sec  
[root@cmp2641 ~]# iperf -c cmp2642-ib0 -l 1m -P 8 -t 30 | grep SUM  
[SUM] 0.0-30.0 sec 331 GBytes 94.8 Gbits/sec  
[root@cmp2641 ~]# iperf -c cmp2642-ib0 -l 1m -P 16 -t 30 | grep SUM  
[SUM] 0.0-30.0 sec 332 GBytes 95.0 Gbits/sec  
[root@cmp2641 ~]#
```

```
root@cmp2670:~  
[root@cmp2670 ~]# iperf -c cmp2641-ib0 -l 2m -P 16 -t 30 | grep SUM  
[SUM] 0.0-30.0 sec 332 GBytes 95.1 Gbits/sec  
[root@cmp2670 ~]# iperf -c cmp2641-ib0 -l 4m -P 16 -t 30 | grep SUM  
[SUM] 0.0-30.0 sec 332 GBytes 95.0 Gbits/sec  
[root@cmp2670 ~]# iperf -c cmp2641-ib0 -l 1m -P 16 -t 30 | grep SUM  
[SUM] 0.0-30.0 sec 332 GBytes 95.0 Gbits/sec  
[root@cmp2670 ~]# iperf -c cmp2641-ib0 -l 1m -P 8 -t 30 | grep SUM  
[SUM] 0.0-30.0 sec 322 GBytes 92.2 Gbits/sec  
[root@cmp2670 ~]# iperf -c cmp2641-ib0 -l 1m -P 16 -t 30 | grep SUM  
[SUM] 0.0-30.0 sec 332 GBytes 95.0 Gbits/sec  
[root@cmp2670 ~]#
```