Flexible Configurations in the Data Center

Is There Nothing New Under the Sun?

Raymond L. Paden, PhD
Storage Systems Technical Lead

22 Sep 19

Version 3a

rpaden@lenovo.com
512-858-4261

Some Ancient GPFS History

Good ole days...

- GPFS 1.*, 2.* required rebuilding FS to upgrade
- GPFS 3.* fixed this…
  It was no longer necessary to rebuild the FS to upgrade.
- mmadddisk has been available since GPFS 1.*
  You could expand FS anyway the HW would let you.
- By the time we got to GPFS 2.*, customers also learned how to mix servers and storage in the same cluster and FS
- Begining with GPFS 2.* and 3.* different OS'es were supported.
- Challenge:
  Development could not test every combination

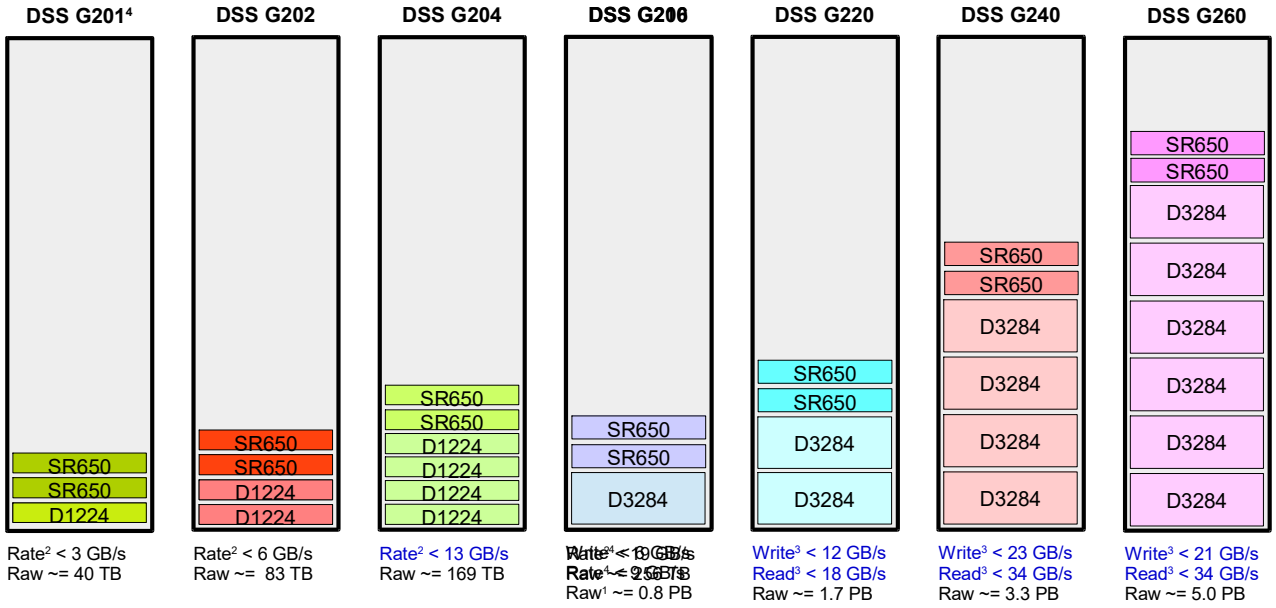
Ancient History

Some More Recent History

Then late in GPFS 3.* came a new beast...

- A *solution* called GSS using SW layer called GNR
  GSS is/was not an appliance!
  It was SW defined storage solution
  It was an integrated solution
  It was tested a single system with all of its components
  Freedom was gone!
- GSS supported/tested 7 configurations at using RHEL[1]
  GSS21s, GSS22s, GSS24s, GSS26s, GSS22, GSS24, GSS26
- Lenovo followed same pattern for its new DSS product.

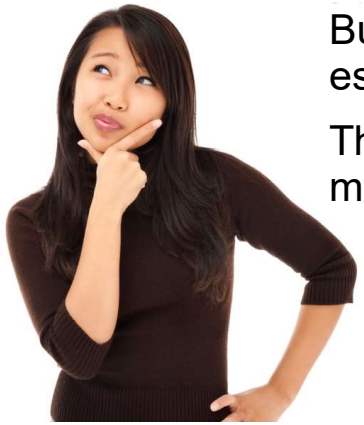
Recent History



Footnote:

1. GSS/DSS is network agnostic; in practical terms we do not test every network combination.

But This Wasn't Good Enough



But this isn't good enough, especially for new users.

They want more flexibility, more options.

Yes, experienced users usually find workarounds.

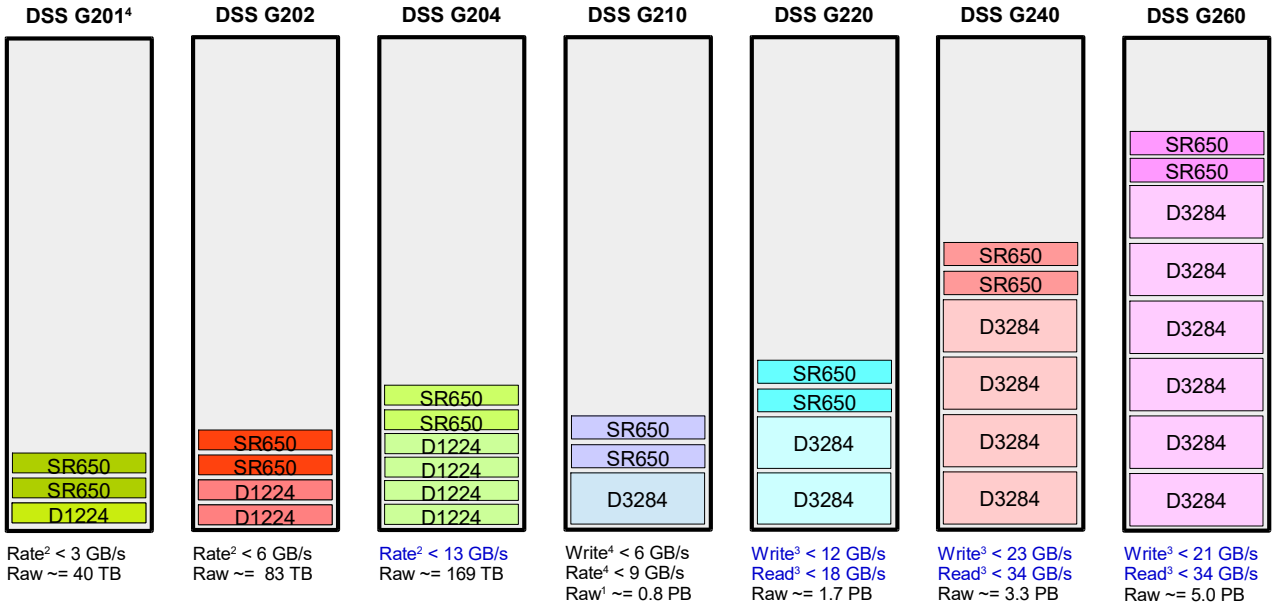
Example:
Using 3xG220 rather than 1xG260 so you can expand in smaller increments.

But its still a hassle!

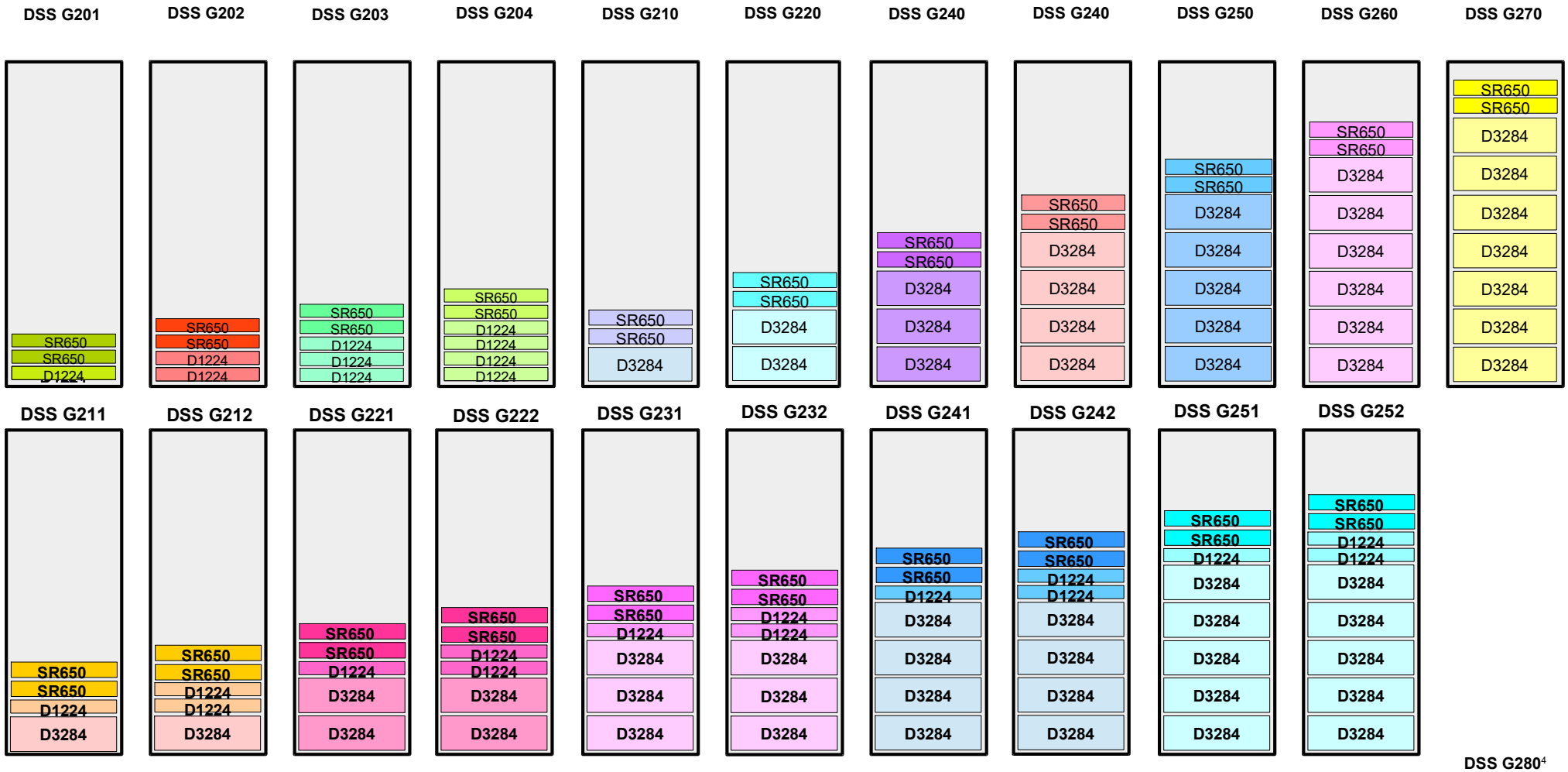
So we added new features:

1. Online expansion using mmvdisk
2. Support odd number of enclosures
3. Hybrid configurations

With these new features, we are going from this...

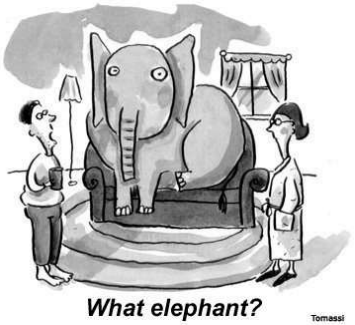


Lenovo



Entire DSS-G Product Line on One Page

- From 7 to 25 basic configurations
- Also there is rackless, G100, Mestor, HDR, OPA, 100GbE!
- Remember, because this is a solution, every one of these configurations are tested! This limits what can be done.



mmvdisk
Major New Feature Starting with GPFS 5.0.2.*

Goals:

- Provide unified conceptual framework that simplifies GNR administration
- Enforce/encourage GNR best practices for the following tasks:
  - GNR server configuration (`mmvdisk server`)
  - Recovery group configuration (`mmvdisk recoverygroup`)
  - Configuring vdisk NSDs (`mmvdisk vdiskset`)
  - Configuring vdisk based FS (`mmvdisk filesystem`)
- Eliminate manual stanza file editing

Command structure:

```
mmvdisk <noun> <verb> <parameters>
```

Command short cuts:

```
mmvdisk rg <verb> <parameters>
mmvdisk vs <verb> <parameters>
mmvdisk fs <verb> <parameters>
```

Central Concept: vdiskset

- A collection of uniform vdisk NSDs from one or more RGs is called a *vdiskset* (VS).
- A vdisk based FS is configured using one or more vdisksets.

Legacy vs. mmvdisk Command Structure:

- Compatibility between the legacy and mmvdisk command structures is strictly limited.
- The `mmvdisk rg convert` converts all components of a cluster to use mmvdisk command structure.

Using `mmvdisk rg convert` is a one way street.
Once converted there is no going back!

Find general overview of mmvdisk command structure at following URL:
https://www.ibm.com/support/knowledgecenter/SSYSP8_5.3.1/com.ibm.spectrum.scale.raid.v5r01.adm.doc/bl1adv_mmvdiskmanage.htm

The following URL provides a good example of how to create mmvdisk RG/FS:
https://www.ibm.com/support/knowledgecenter/SSYSP8_5.3.1/com.ibm.spectrum.scale.raid.v5r01.adm.doc/bl1adv_mmvdiskoutlineusecase.htm

`mmvdisk`: Helpful Links

The following Spectrum Scale website provides many helpful links:

https://www.ibm.com/support/knowledgecenter/SSYSP8_5.3.1/com.ibm.spectrum.scale.raid.v5r01.adm.doc/bl1adv_mmvdiskmanage.htm

Here are several links of particular interest from previous link:

- Outline of an mmvdisk use case
  Lists 7 steps needed to create a FS using `mmvdisk` command structure
- Use case for mmvdisk
  Screen scrapes for building ESS system following the outline from the previous link.
- Converting existing recovery groups to mmvdisk management
  Example using the `mmvdisk recoverygroup convert` command
- Replacing a pdisk using mmvdisk
- Link to *IBM Spectrum Scale RAIDVersion5.0.2: Administration*
  https://www.ibm.com/support/knowledgecenter/SSYSP8_5.3.2/raid_adm.pdf

mmvdisk: Seven Steps for Creating FS

Step 1: Create mmvdisk Node Class

```
mmvdisk nc create –node-class <NC Name>
```

Step 2: Validate topology

This is an alternative to mmgetpdisktopology ; topsummary

```
mmvdisk server list --node-class <NC Name> --disk-topology
```

Step 3: Configure recovery group servers for each nodeclass.

```
mmvdisk server configure --node-class <NC Name> --recycle all
```

Step 4: Create recoverygroups (RG)

```
mmvdisk rg create --recovery-group <left, right RG names> --node-class <NC Name>
```

Step 5: Define vdisksets (VS)

It is necessary to define the VS before creating them.

```
mmvdisk vs define --vdisk-set vs1 --recovery-group <left, right RG names>
<parameters>
```

Step 6: Create the defined VS

```
mmvdisk vs create --vdisk-set <VS name(s)>
```

Step 7: Create a file system (FS) spanning 1 or more VS

This creates the NSD and the VS

```
mmvdisk fs create --file-system <FS Device Name> --vdisk-set <VS name(s)>
<parameters>
```

Comments:

Lenovo provides following commands (see <xCAT Server>:/install/dssg/bin)

- dssgmkstorage.mmvdisk (performs steps 1 – 3)
- dssgmkfs.mmvdisk (performs steps 4 – 7 creating FS following Lenovo best practices)

  This command cannot support all possible combinations, but the --dryrun parameter to provide an example to follow.

Online Expansion
Major New Feature Starting with GPFS 5.0.2.*

Goal: Add new enclosures to existing GNR building blocks.

- Start small, grow larger
- This can be done an active system without a maintenance window

Command: mmvdisk rg resize

- e.g., `mmvdisk rg resize --rg dss17,dss18 -v no`

Restrictions:

- Enclosures must be homogeneous; for example...

  Consider G210 (8TB disk): cannot expand to G220 adding a 5U84 (10TB disks)

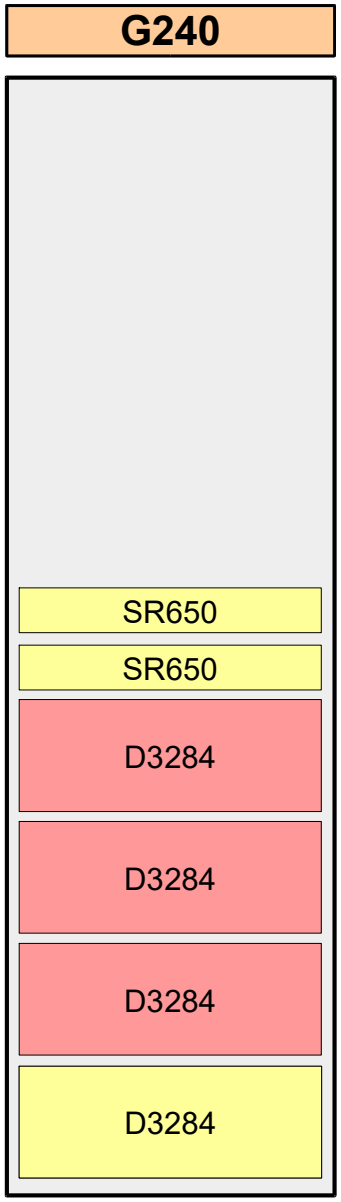
  Therefore it does not support hybrids... yet?
  Consider G220: cannot expand to G222 (i.e., 2 x 5U84 + 2 x 2U24)

- Add one increment at a time

Comment:
`mmvdisk` works well with hybrids, but online expansion does not.

Following slides present set of *experiments* to illustrate how online expansion works that motivate some recommended best practices.

Online Expansion
Expand One Increment at a Time?



Experiment:

- Start with G210
- Expand to G240 in one step...

```
[root@dss23 config_ray]# mmvdisk rg resize --rg dss23,dss24

mmvdisk: Obtaining pdisk information for recovery group 'dss23'.
mmvdisk: Obtaining pdisk information for recovery group 'dss24'.
mmvdisk: Analyzing disk topology for node 'dss23-ib0.cluster'.
mmvdisk: Analyzing disk topology for node 'dss24-ib0.cluster'.
mmvdisk: Validating existing pdisk locations for recovery group 'dss23'.
mmvdisk: Validating existing pdisk locations for recovery group 'dss24'.
mmvdisk: The resized server disk topology is 'DSS-G240 7X06CTO1WW LSI1BUS
PCI 1,2,3,4'.
mmvdisk: Server disk topology 'DSS-G240 7X06CTO1WW LSI1BUS PCI 1,2,3,4'
does not support resizing.
mmvdisk: Command failed. Examine previous error messages to determine
cause.
```

- What went wrong?
- The G240 stanza in the CST (Comp Spec Topology) file needs following:

```
SourceSignature="1[84,1-1,2-42]"
```

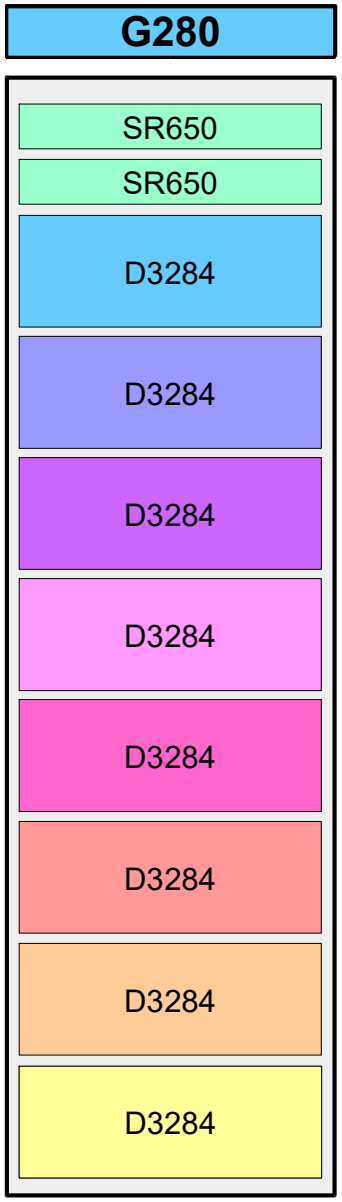
This tells GNR that a G240 can be expanded from a G210

There can only be one source signature in a stanza

This is ***NOT*** customer tunable!

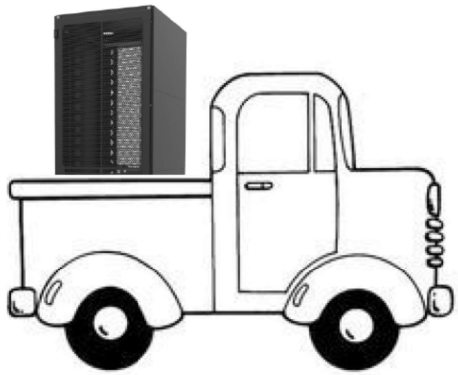
Online Expansion
Expand One Increment at a Time

new and improved

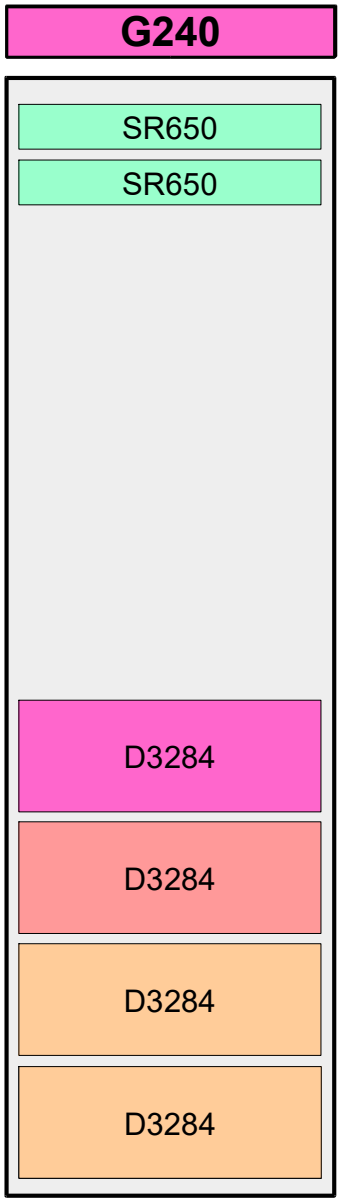


Experiment:

- Start with G210
- Expand to G220, wait for re-balance to complete...
- Expand to G230, wait for re-balance to complete...
- Expand to G240, wait for re-balance to complete...
- Expand to G250, wait for re-balance to complete..
- Expand to G260, wait for re-balance to complete..
- Expand to G270, wait for re-balance to complete..
- Go to Best Buy and get a 44U rack
- Expand to G280, wait for re-balance to complete..



Online Expansion
Configuring the Base System



Question:

- So what do we do with all of this new space?
- e.g., start with G220, expand to G230 and later G240

Another Experiment:

Call it vs1

- Start with G220 creating vdiskset using all capacity; i.e., 100%
- Expanding to G230 gives us 50% more capacity.
  Can we add it to the existing FS?

  Create new vdiskset for the new capacity with *same parameters* as existing FS. But notice the set-size parameter...

  mmvdisk vs define --vdisk-set vs2 --recovery-group dss17,dss18 --code 8+2p --block-size 16m --set-size 33% --nsd-usage dataAndMetadata

Set-size specifies percent of available capacity to use.
But its only ½ the size of vs1!

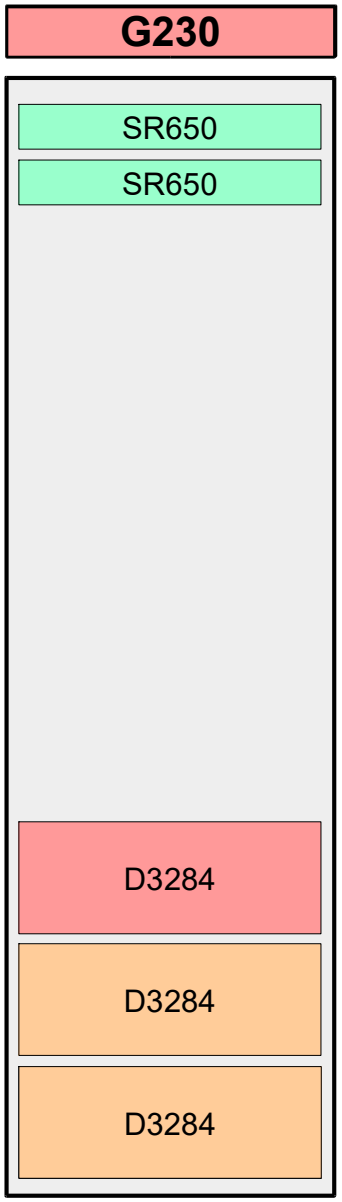
- Do it again...
  Expanding to G240 now gives 33% more capacity.

  mmvdisk vs define --vdisk-set vs3 --recovery-group dss17,dss18 --code 8+2p --block-size 16m --set-size 25% --nsd-usage dataAndMetadata

  Following best practice, vdisksets vs2 and vs3 cannot be added to original FS; instead create two new FS (e.g., /fs2 and/fs3)

- So what can be done about this?

Online Expansion
Configuring the Base System



Yet Another Experiment:

- Starting with a G220, create first vdiskset using 50% of the space.

  mmvdisk vdiskset define --vdisk-set vs1 --recovery-group dss17,dss18 --code 8+2p --block-size 16m --set-size 50% --nsd-usage dataAndMetadata

- Next create an exact copy of the first vdiskset;
  This will result in 2 vdisksets, each using 50% of the capacity, each one spanning both enclosures in both RGs.

  mmvdisk vdiskset define --vdisk-set vs2 --copy vs1 --recovery-group dss17,dss18 --force-incompatible

- Then when expanding to G230, make another copy of vs1.

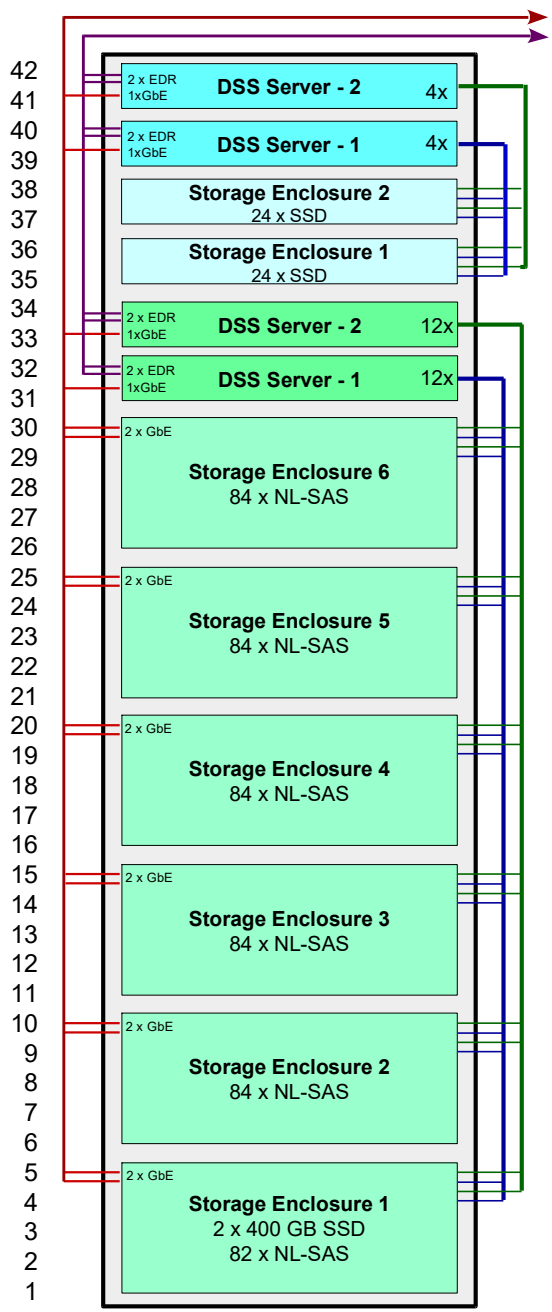
  mmvdisk vdiskset define --vdisk-set vs3 --copy vs1 --recovery-group dss17,dss18 –force-incompatible

  This can then be added to the existing FS since the new vdiskset is the same size as the others.

  mmvdisk filesystem add --file-system fs_16m --vdisk-set vs3

- As best practice, when installing **first** GNR BBs (with more than 1 enclosure) configure multiple uniform vdisksets to allow for expansion.

DSS-G
Multi-tiered G260/G202 vs. Hybrid G262



G260+G202: More fast capacity & more performance
- DSS-G 2.4, GPFS 5.0.2-3.0.1
- RHEL 7.6

This is 2 building blocks

- **2** x DSS Servers
- 2 x D1224 Storage Enclosures (aka, 2U24)
  Capacity using 1.2TB SSD with 8+2P
  • Raw = 57.6 TB, Usable < 39 TB
  16 MiB Streaming rate
  • Write < 15 GB/s, Read < 24 GB/s

- **2** x DSS Servers
- 6 x D3284 Storage Enclosures (aka, 5U84)
  Capacity using 10 TB NL-SAS with 8+2P
  • Raw = 5.02 PB, Usable < 4.0 PB
  16 MiB Streaming rate
  • Write < 34 GB/s, Read < 42 GB/s

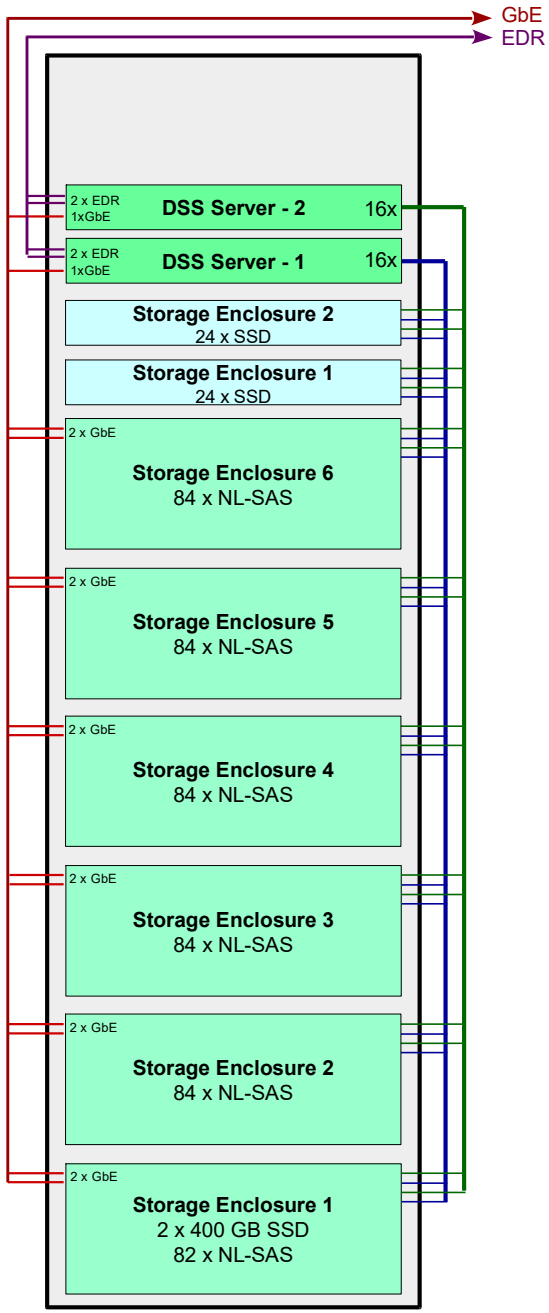
- Aggregate Performance
  16 MiB Streaming rate
  • Write < 49 GB/s, Read < 66 GB/s

G262: Only more fast capacity
- DSS-G 2.4, GPFS 5.0.2-3.0.1
- RHEL 7.6

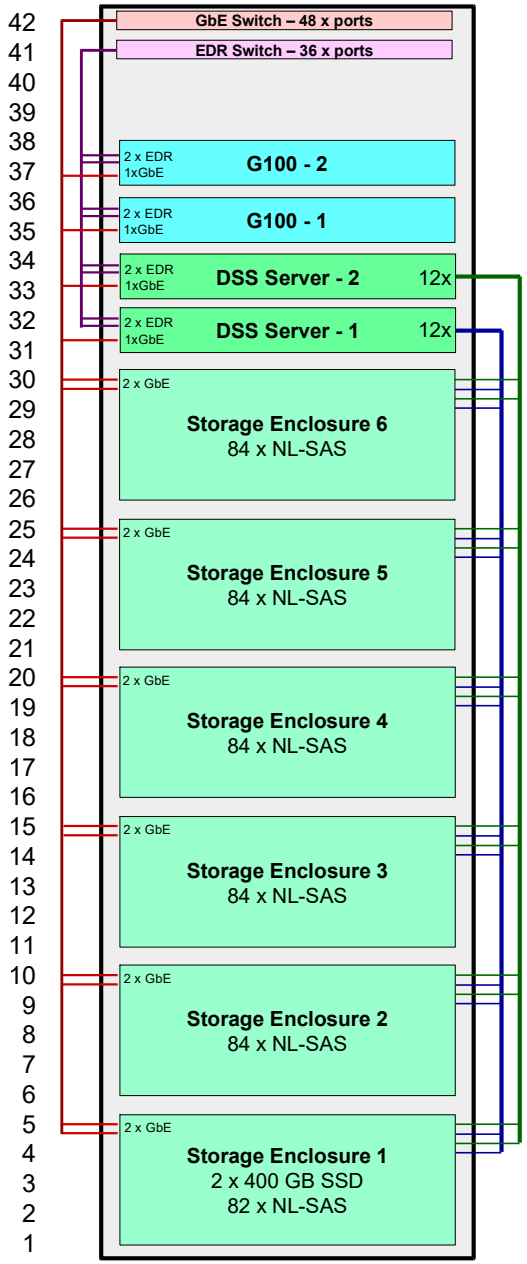
This is *only* 1 building block

- **2** x DSS Servers
- 2 x D1224 (2U24) + 6 x D3284 (5U84) Storage Enclosures
  Capacity using 1.2 TB SSD and 10 TB NL-SAS with 8+2P
  • Raw = 5.08 PB, Usable < 4.02 PB
  Aggregate 16 MiB Streaming rate
  • Write < 34 GB/s, Read < 42 GB/s

Comments:
1. mmvdisk does not yet support online expansion for hybrids
2. Hybrids are not intended primarily as a dedicated MD store.



DSS-G
Case Study: Solution Using Multi-tiered G260/G100



2 x Building Block
- DSS-G 2.2, GPFS 5.0.2-1
- RHEL 7.5

Building Block #1 (system pool)
- 2 x NSD Servers (SR650 with Sky Lake where *each* server has following:
  2 x sockets (Gold 6142; aka, Sky Lake), 18 x cores per socket
  DRAM: 192G (16 GB/DIMM)[6]
  2 x 2xEDR adapters (GPFS)[1]
  1 x GbE adapter (general admin)[2]
  8 x 800G NVMe (n.b., mirroring is *not* enabled)[3]

Building Block #2 (data pool)
- 2 x DSS Servers (SR650 with Gold 6142 (Sky Lake)) where each server has following:
  2 x sockets, 18 x cores per socket
  DRAM: 192G (16 GB/DIMM)[6]
  2 x 2xEDR adapters (GPFS)[1]
  1 x GbE adapter (general admin)[2]
- 6 x Storage Enclosures (D3284... aka, 5U84)
  2 x 400 GB SAS SSD[4]
  502 x NL-SAS using 8+2P[5] with 1 "hot spare" per enclosure per RG
- Capacity: 4TB NL-SAS -> raw ~= 2.00 PB, usable < 1.4 PB

Aggregate Performance (preliminary results, more testing needed)
- Streaming Data Rate: write < 38 GB/s, read < 45 GB/s (up to as high as 51 GB/s)
- mdtest data rates: write ~= 169,000, read ~= 134,000
  Command line used: ./mdtest -d /fs_16m/mdt/ -i 3 -n 16384 -F -w 3072 -C -E -r -T -p 15 -u -v

Footnotes
- Common Practice: use only 1 port per adapter; 2nd port used for HA as needed.
- GPFS can use either the GbE network or IPoIB for GPFS administration.
- As configured, there is no redundancy; to get same performance with failover redundancy requires 4 x G100.
- SSD is used for logtip backup (n.b., GNR MD). The primary copy is stored in NVRAM.
- Enclosure protection is guaranteed using 8+2P.
- Servers can also be configured using 384G (24 x 16G DIMMs), 768G (24 x 32G DIMMs) or 1536G (24 x 64G DIMMs) But largest supported pagepool size < 1024G.

Conclusions

- Since DSS-G is a “solution”, all supported configurations go through integration testing. This limits the number of configurations that can be supported, but provides a more robust solution to customers.
- Within the framework of a solution, DSS-G now offers greater flexibility in terms of initial deployment and future growth.
  - Online expansion
  - Support for up to 8 enclosures
  - Support for odd numbers of enclosures
  - Support for hybrid configurations
- Each configuration comes with trade-offs of cost and performance.
- `mmvdisk` is the future for GNR (aka, Spectrum Scale RAID)
  - Use `mmvdisk` on new installs where it makes sense.
  - Convert existing legacy systems where feasible.

  Note: Cluster must be running at GPFS 5.0.2.* or later.

Questions and Answers