Online Expansion Using `mmvdisk`

Is There Nothing New Under the Sun?

Raymond L. Paden, PhD
Storage Systems Technical Lead

17 Apr 19

Version 2a

rpaden@lenovo.com
512-858-4261

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

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)

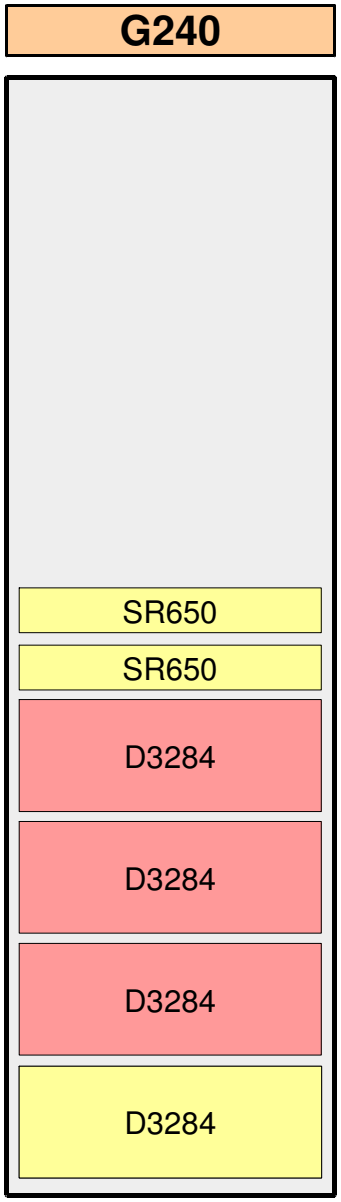
  Therefore it does not support hybrids... yet?
  Consider G220: cannot expand to G222 (i.e., 2 x 5U84 + 2 x 2U24)

- 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

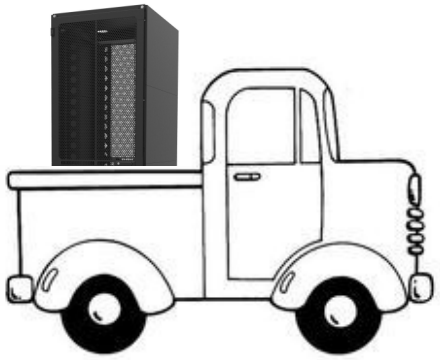
new and improved

G280

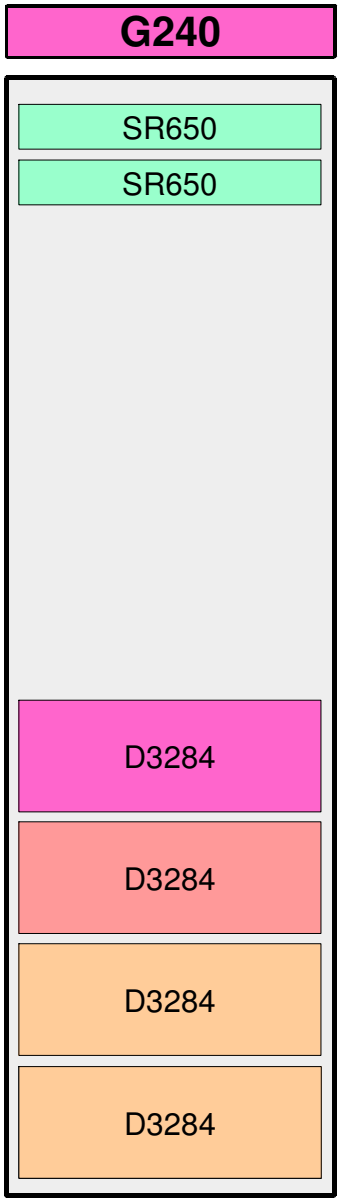
- SR650
- SR650
- D3284
- D3284
- D3284
- D3284
- D3284
- D3284
- D3284

Experiment:

- Modify all of my CST stanzas...
- 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!

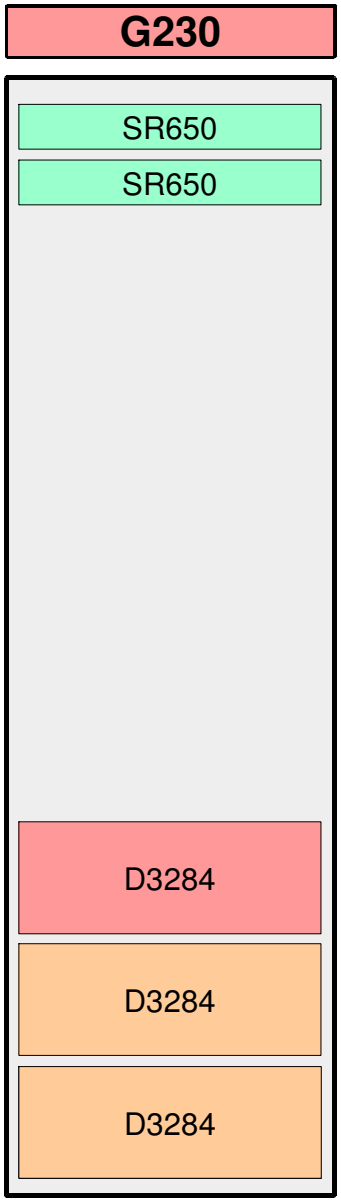
- Do it again...
  Expanding to G240 now gives 33% more capacity.

  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.

Conclusions

- `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.
- Online expansion is now available. Careful planning will allow customers to more effectively use it.

Questions and Answers

Backup Slides