IBM Elastic Storage System Support Webinar

ESS Disk Drive Replacement

—

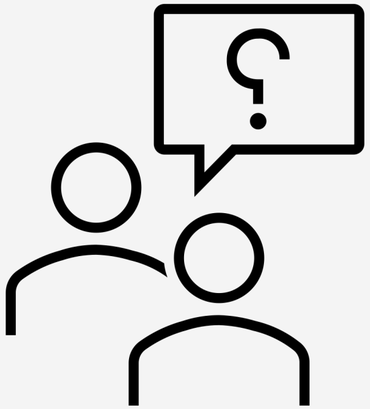
Tim Cook
Scott “Tex“ Nance

June 2022

Agenda

1. Drive replacement overview
2. GPFS Spectrum Scale Native Raid (GNR)
3. What is the disk hospital?
4. Identifying a failing or failed drive
5. Commands used during the removal and replacement process
6. What happens next?
7. Q&A

Questions?



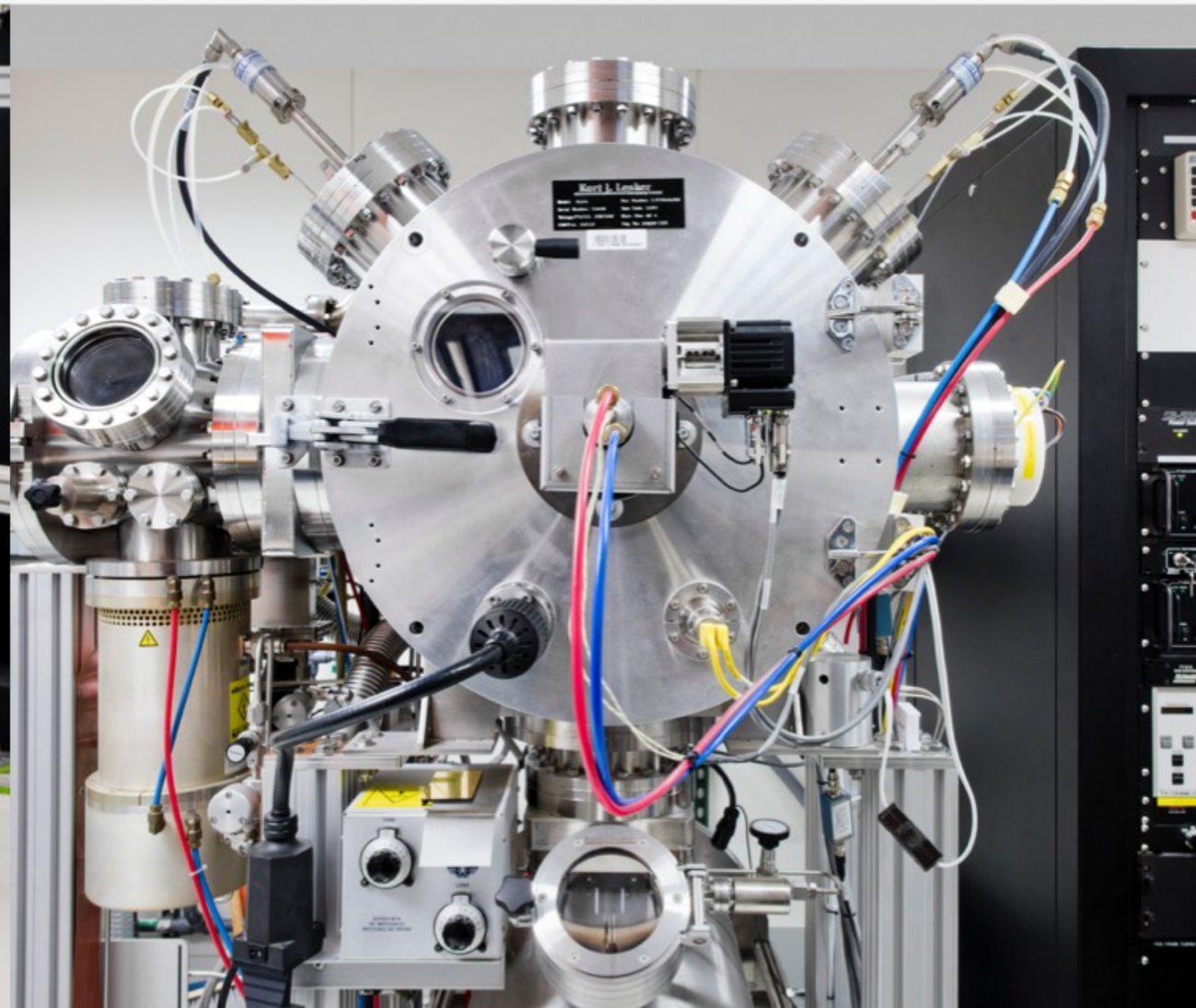
Post questions in the monitored Q&A box in Webex

We also have time at the end for live questions

Disk Replacement Overview

- Identify failing or failed drive
  - Call home automatic detection
  - Graphical User Interface (GUI) identification
  - Health checks using CLI commands including mmlspdisk, mmlsvdisk and mmhealth
- Open a case with the IBM ESS support team via a hardware support ticket
- Provide key information for the ESS support team
  - Specific disk drive or drives that have failed (Output from mmlspdisk all --not-ok)
  - Field Replaceable Unit (FRU) part number
  - How you would like the drive replaced (IBM SSR or Customer Replaceable Unit CRU)
  - Shipping address and contact information (email and phone number)
- Part is replaced and failed part is returned to IBM

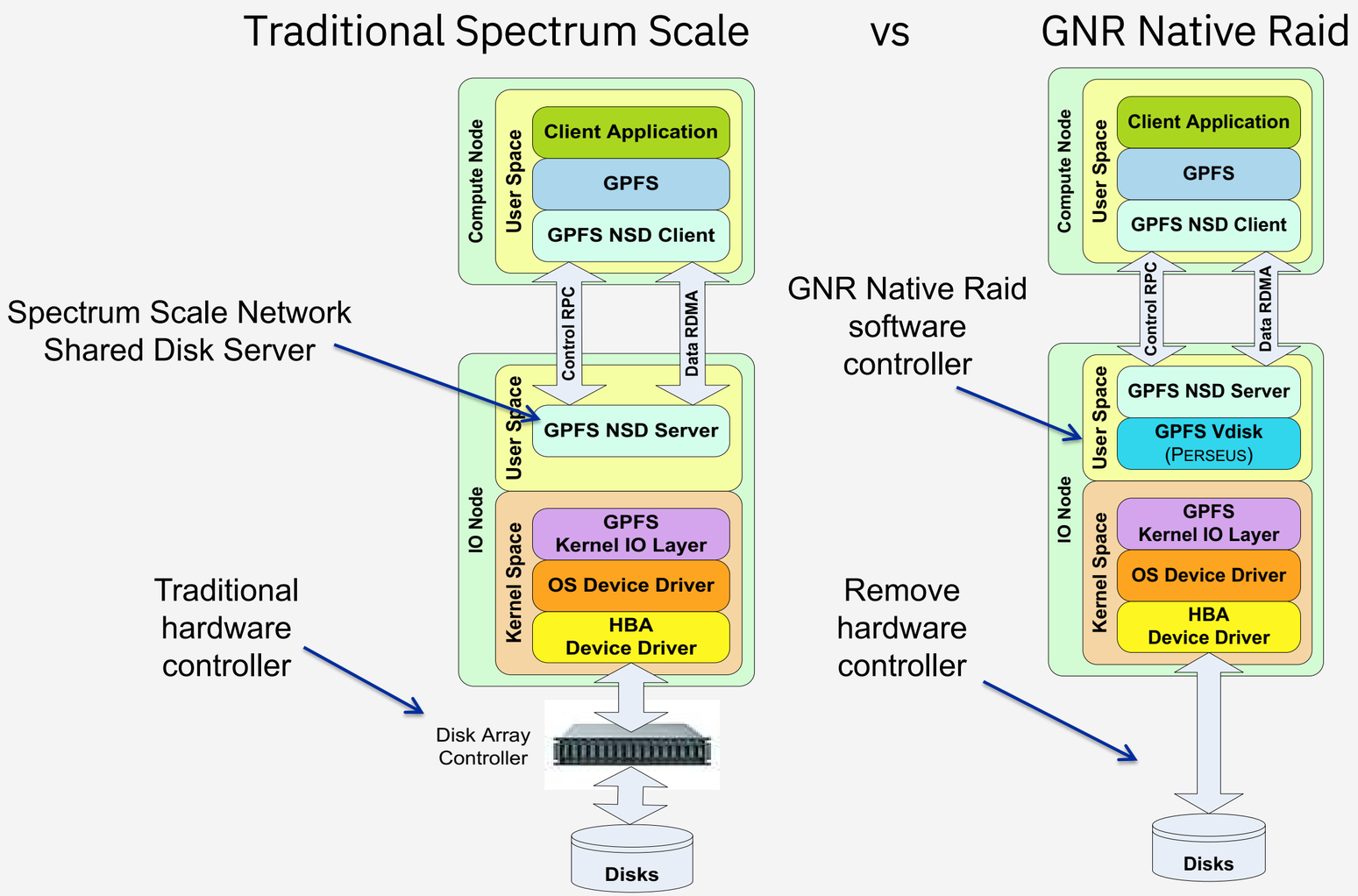
GPFS Spectrum Scale Native Raid (GNR) + Disk Hospital



What is GPFS Spectrum Scale Native Raid (GNR)?

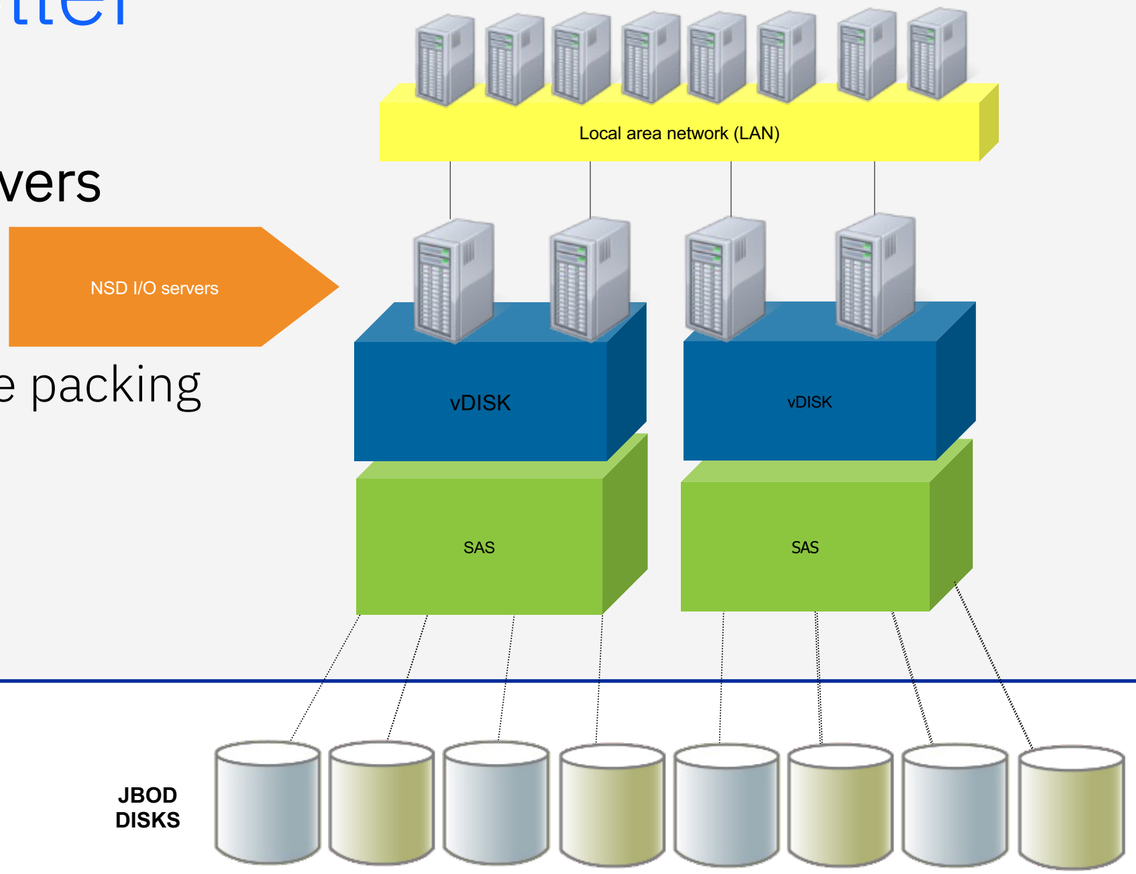
- GNR is a software implementation of storage RAID technologies within GPFS
  - Allows interfacing with standard Serial Attached SCSI (SAS) disks in a dual-ported JBOD array
  - No external RAID storage controllers required
- Distributed Raid: Distributes data, redundancy , and spare space uniformly across all of the disks in the JBOD
- Pdisk-group fault tolerance
  - Error correction codes ensure that missing data is fully recoverable
  - Checksum provides an end-to-end data integrity check to protect against data corruption and lost disk writes
  - Highly reliable 2-fault-tolerant and 3-fault-tolerant Reed-Soloman based parity codes coupled with 3-way and 4-way replication to protect against data loss
- Provides a familiar interface and flexible hardware configuration options
- Journaling
- Automatic recovery

GPFS Spectrum Scale vs GPFS Native RAID (GNR)



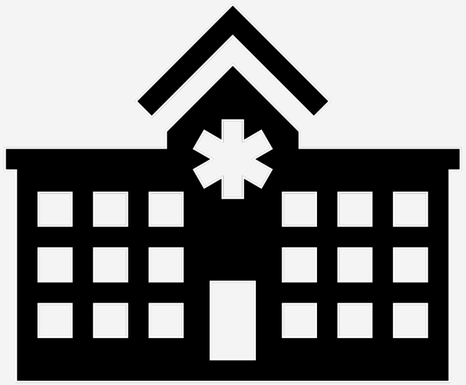
No Hardware Storage Controller

- GNR Software RAID operates on the I/O servers
  - SAS attached JBOD
  - Special JBOD storage drawer for very dense drive packing
  - Solid-state drives (SSDs) for metadata storage



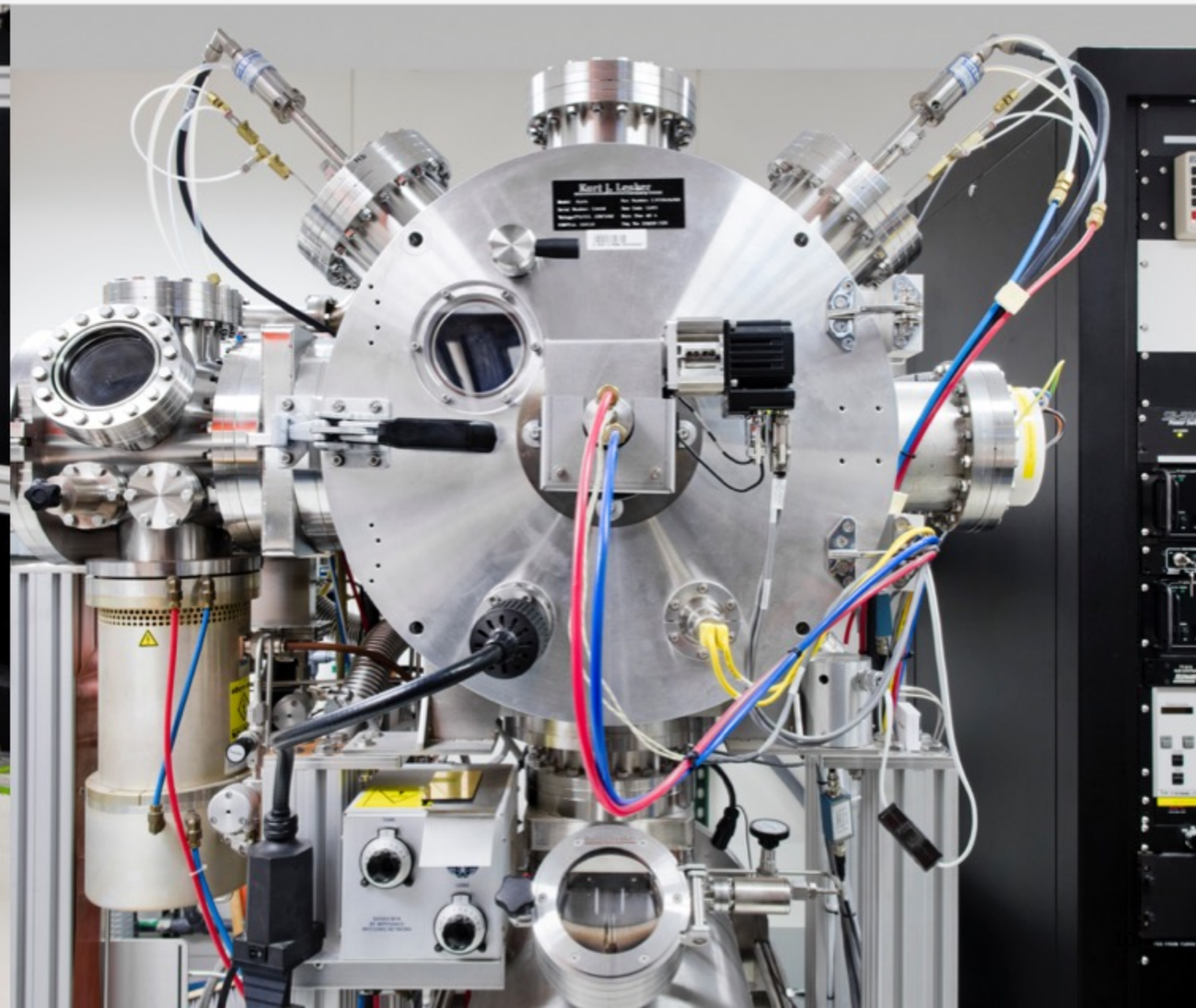
- Features
  - Auto rebalancing
  - Only 2% rebuild performance hit
  - End-to-end, disk-to-GPFS-client data checksum

Disk Hospital Overview



- The disk hospital is a key feature of GNR
- Asynchronously diagnoses errors and faults in the storage subsystem
- Limits the impact of a faulty pdisk to protect I/O operations
- What happens when there is a suspected problem with a pdisk
  - The pdisk is admitted to the disk hospital
  - GNR uses the vdisk redundancy codes to reconstruct any lost or erased data blocks

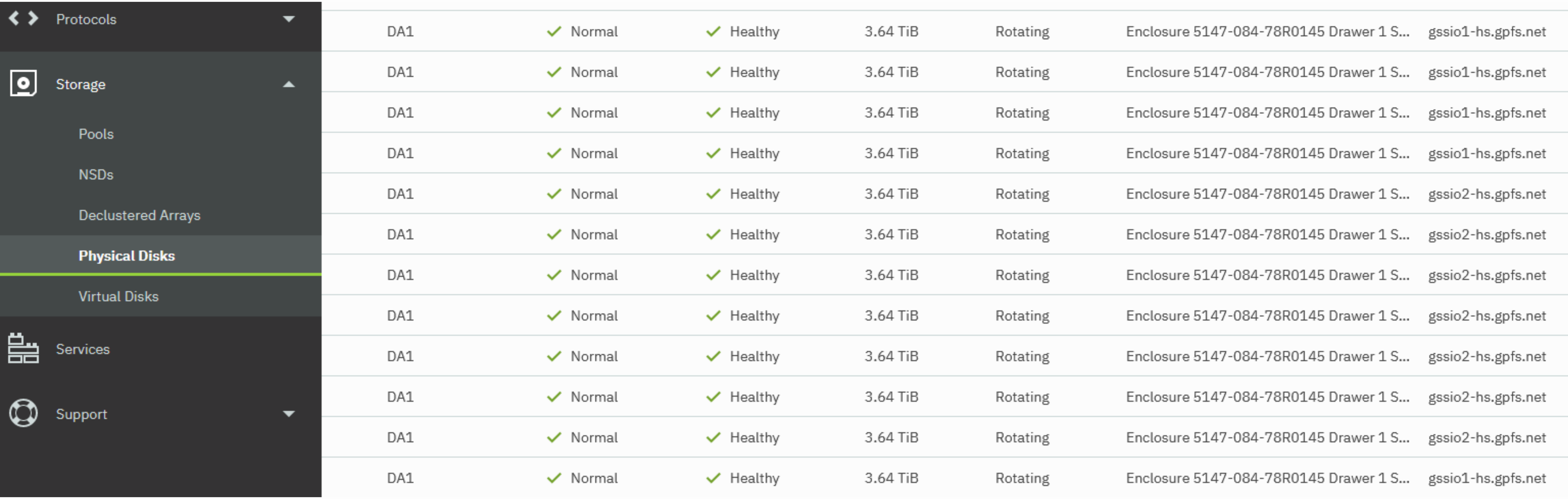
Identifying a Failing or Failed Drive



Using the Graphical User Interface (1 of 2)

Storage – Physical Disks

Shows all Healthy Disks

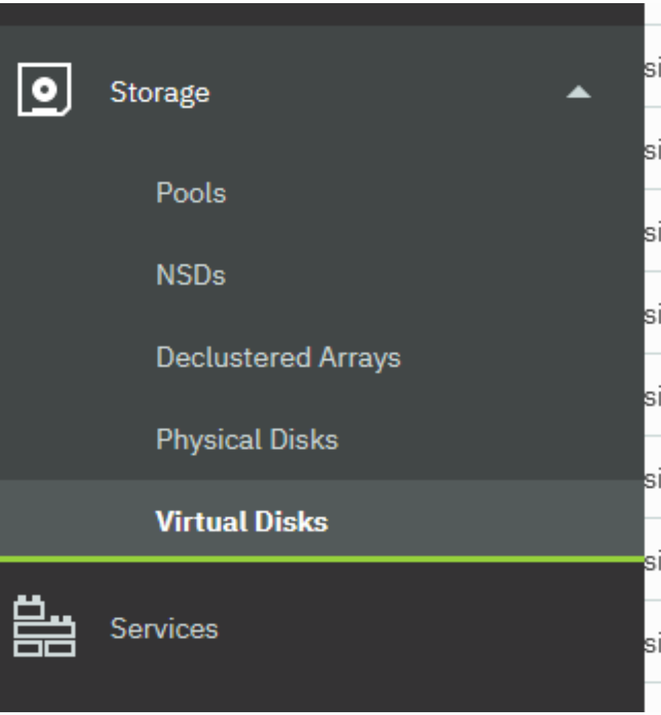


DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio1-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio1-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio1-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio1-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio2-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio2-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio2-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio2-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio2-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio2-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio2-hs.gpfs.net
DA1	✓ Normal	✓ Healthy	3.64 TiB	Rotating	Enclosure 5147-084-78R0145 Drawer 1 S...	gssio1-hs.gpfs.net

Using the Graphical User Interface (2 of 2)

Storage – Virtual Disks

Shows all Healthy Disks



sio1-hs	DA1	Healthy	111.3 GiB	Reed-Solomon 8+2p	gpfs0	8 MiB
sio1-hs	DA1	Healthy	111.3 GiB	Reed-Solomon 8+2p	gpfs0	8 MiB
sio1-hs	DA1	Healthy	111.3 GiB	Reed-Solomon 8+2p	gpfs1	8 MiB
sio1-hs	DA1	Healthy	5.9 GiB	Three-way replication	gpfs0	1 MiB
sio1-hs	DA1	Healthy	5.9 GiB	Three-way replication	gpfs1	1 MiB
sio1-hs	DA1	Healthy	253.7 GiB	Reed-Solomon 8+2p	polos	4 MiB
sio1-hs	DA1	Healthy	253.7 GiB	Reed-Solomon 8+2p	polos	4 MiB
sio2-hs	DA1	Healthy	2.5 TiB	Reed-Solomon 8+2p	coderepo	16 MiB

Example of Failed Drive from GUI

Hardware

Servers

0

Disk Enclosures

1

Physical Disks

1 / 28

e1s14	rg_gssio2	DA1	Normal	Healthy
e1s15	rg_gssio2	DA1	Normal	Healthy
e1s16	rg_gssio2	DA1	Normal	Healthy
e1s17	rg_gssio2	DA1	Draining	Degraded
e1s18	rg_gssio2	DA1	Normal	Healthy
e1s19	rg_gssio2	DA1	Normal	Healthy

e1s15	rg_gssio2	DA1	Normal	Healthy
e1s16	rg_gssio2	DA1	Normal	Healthy
e1s17	rg_gssio2	DA1	Replaceable	Failed
e1s18	rg_gssio2	DA1	Normal	Healthy
e1s19	rg_gssio2	DA1	Normal	Healthy

Using CLI Commands

- mmhealth cluster show --unhealthy
- mmhealth node show -N all --unhealthy
- mmlspdisk all --not-ok (This is the preferred command to supply when you open a case)
- mmlspdisk all --not-ok | egrep "name | state”
- mmvdisk pdisk list --rg all --not-ok
- mmlsrecoverygroup rgname -L --pdisk | grep -v “2, 4” (this will also show any missing paths)

Note: if you have more than one disk it is also a good idea to include the gpfs.snap data.

Example of a Failed Drive Using the mmvdisk Command

```
[root@ems1 ~]# mmvdisk pdisk list --rg rg_gssio2

                                declustered
recovery group  pdisk              array     paths  capacity  free space  FRU (type)       state
--------------  ------------  -----------  -----  --------  ----------  ---------------  -----
rg_gssio2       e1s13         DA1              2  3576 GiB    3342 GiB  01EJ599          ok
rg_gssio2       e1s14         DA1              2  3576 GiB    3344 GiB  01EJ599          ok
rg_gssio2       e1s15         DA1              2  3576 GiB    3342 GiB  01EJ599          ok
rg_gssio2       e1s16         DA1              2  3576 GiB    3344 GiB  01EJ599          ok
rg_gssio2       e1s17         DA1              2  3576 GiB    3344 GiB  01EJ599          simulatedDead/draining/replace
rg_gssio2       e1s18         DA1              2  3576 GiB    3344 GiB  01EJ599          ok
rg_gssio2       e1s19         DA1              2  3576 GiB    3344 GiB  01EJ599          ok
rg_gssio2       e1s20         DA1              2  3576 GiB    3342 GiB  01EJ599          ok
rg_gssio2       e1s21         DA1              2  3576 GiB    3344 GiB  01EJ599          ok
rg_gssio2       e1s22         DA1              2  3576 GiB    3344 GiB  01EJ599          ok
```

Data Needed by Support for Drive Replacement

- mmlspdisk all --not-ok (This is the preferred command to supply when you open a case)

```
[root@ems1 ~]# mmlspdisk all --not-ok
pdisk:
   replacementPriority = 1000
   name = "e1s17"
   recoveryGroup = "rg_gssio2"
   state = "simulatedFailing/draining"
   internalState = 02000.4c0
   fru = "01EJ599"
   server = "gssio2.gpfs.net"
   userLocation = "Rack RACK01 U02-03, Enclosure 5147-   024-G7BA005 Drive 17"
   hardwareType = SSD
   nPaths = 2 active 4 total
```

- Make sure the case is opened using the serial number of the enclosure not the EMS
- How you would like the drive replaced (IBM SSR or CRU)
- Address where the system is located or where the replacement part should be shipped (this is really important when there are multiple locations or sites)

Example of data needed from an Actual Case

Problem Description: Failed 14T sas drive. fru = "01LU841"

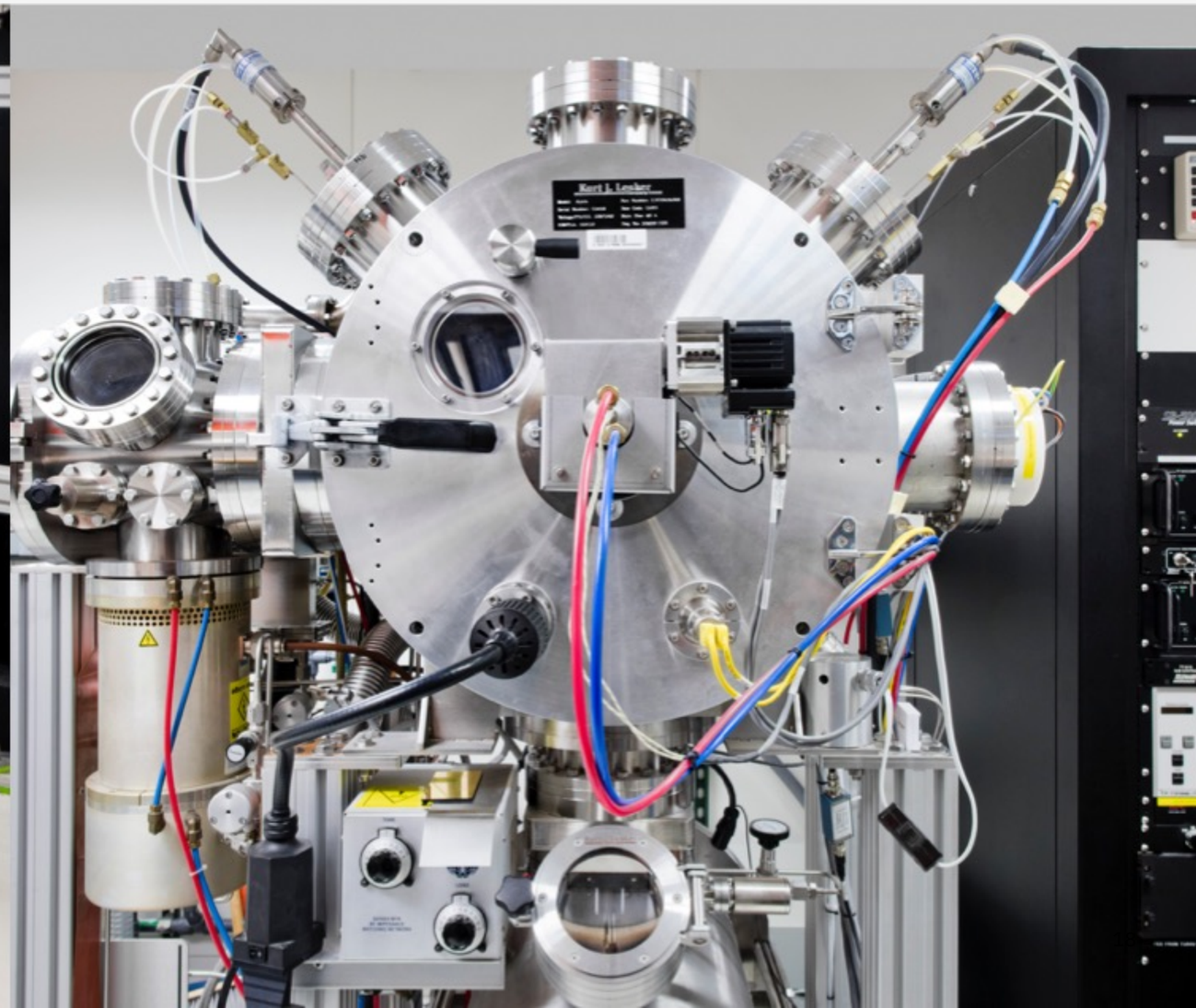
Please send a CRU with return mailer

Bubba Smith c/o Redneck Auto Parts International
713 SE 10th Ave
Amarillo, TX 79101
806-356-3850

```
pdisk:
  name = "e4s071"
  recoveryGroup = "rg_gssio4-hs"
  declusteredArray = "DA1"
  state = "failing/serviceDrain/replace"
  fru = "01LU841"
  location = "78T62TX-71"
  server = “rapgssio1.redneckauto.com"
  userLocation = “Enclosure 78T62TX Drive 71“
```

NOTE: the customer information has been changed and some of the data has been removed to save space.

Commands Used During the Removal and Replacement Process



Disk Replacement Overview

1. Prepare the pdisk for removal

2. Physically replace the pdisk (either an IBM SSR or customer)

3. Complete the replacement process

Step 1 – Prepare the pdisk for Removal

Example: Legacy - Non-MMVDISK with expected output:

```
# mmchcarrier rg_gssio2 --release --pdisk e1s17

[I] Suspending pdisk e1s17 of RG rg_gssio2 in location SV21314035-5-
1.
[I] Location G7BA005-17 is Rack RACK01 U02-03, Enclosure 5147-024-
G7BA005 Drive 17.
[I] Carrier released.

  - Remove carrier.
  - Replace disk in location G7BA005-17 with FRU 01EJ599.
  - Reinsert carrier.
  - Issue the following command:

      mmchcarrier rg_gssio2 --replace --pdisk 'e1s17'
```

Example: MMVDISK method with expected output:

```
# mmvdisk pdisk replace --prepare --recovery-group ESS01L --pdisk e6d1s02

mmvdisk: Suspending pdisk e6d1s02 of RG ESS01L in location SV50918970-1-2.
mmvdisk: Location SV50918970-1-2 is Enclosure SV50918970 Drawer 1 Slot 2.
mmvdisk: Carrier released.
mmvdisk:
mmvdisk: - Remove carrier.
mmvdisk: - Replace disk in location SV50918970-1-2 with type '38L6721’.
mmvdisk: - Reinsert carrier. mmvdisk: - Issue the following command:
mmvdisk:
mmvdisk: mmvdisk pdisk replace --recovery-group ESS01L --pdisk 'e6d1s02'
```

Step 2 – Physically Replace the pdisk

- You can request an IBM SSR be dispatched to your location and replace the drive
  - Note if you choose to have an SSR replace the drive the drive may be shipped to an alternate location
  - SSR will be scheduled based on availability and skills

- Physical pdisks are customer replaceable (CRU) if you are comfortable performing the physical replacement process

Step 3 – Complete the Replacement Process

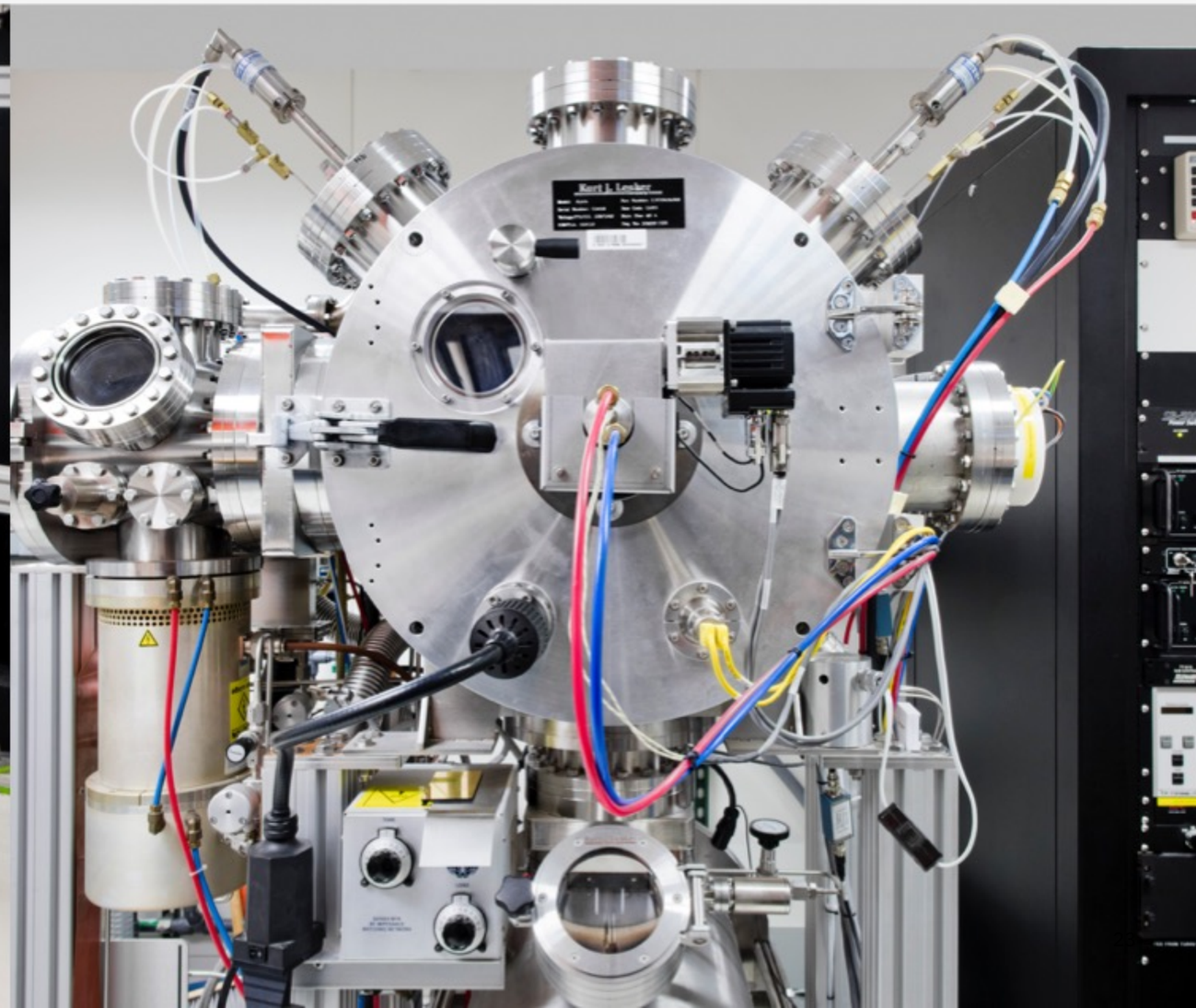
- Example: Legacy - Non-MMVDISK with expected output

```
# mmchcarrier rg_gssio2 --replace --pdisk e1s17
[I] The following pdisks will be formatted on node server1:
    /dev/sdmi
[I] Pdisk e1s17 of RG rg_gssio2 successfully replaced.
[I] Resuming pdisk e1s17#018 of RG rg_gssio2.
[I] Carrier resumed.
```

- Example: MMVDISK method with expected output

```
# mmvdisk pdisk replace --recovery-group ESS01L --pdisk e6d1s02
mmvdisk: mmvdisk: Preparing a new pdisk for use may take many minutes.
mmvdisk:
mmvdisk: The following pdisks will be formatted on node ess01io1:
mmvdisk: /dev/sdrk mmvdisk: mmvdisk: Location SV50918970-1-2 is Enclosure SV50918970 Drawer 1 Slot 2.
mmvdisk: Pdisk e6d1s02 of RG ESS01L successfully replaced.
mmvdisk: Resuming pdisk e6d1s02#047 of RG ESS01L.
mmvdisk: Carrier resumed.
```

What Happens Next?



Rebuild and Rebalance

- When the replace command returns successfully
  - GNR will begin rebuilding and rebalancing the data strips onto the new disk or disks
  - The failed disk may remain in a temporary form (by name only) until all of the data from it rebuilds
  - Old disk information for the failed disk will be permanently deleted

Example of the Recovery Group During Rebuild

```
# mmlsrecoverygroup BB1RGL -L --pdisk
declustered
recovery group arrays vdisks pdisks
---------------- ----------- ------ ------
BB1RGL 3 5 121
declustered needs replace scrub background activity
array service vdisks pdisks spares threshold free space duration task progress priority
----------- ------ ------ ------ ------ --------- ---------- -------- -------------------------

LOG no 1 3 0 1 534 GiB 14 days scrub 1% low
DA1 no 2 60 2 2 3647 GiB 14 days rebuild-1r 4% low
DA2 no 2 58 2 2 1024 MiB 14 days scrub 27% low
n. active, declustered user state,
pdisk total paths array free space condition remarks
----------- ----------- ----------- ---------- ----------- -------
[...]
e1d4s06 2, 4 DA1 62 GiB normal ok
e1d5s01 2, 4 DA1 1843 GiB normal ok
e1d5s01#026 0, 0 DA1 70 GiB draining slow/noPath/systemDrain/adminDrain/noRGD/noVCD
e1d5s02 2, 4 DA1 64 GiB normal ok
e1d5s03 2, 4 DA1 63 GiB normal ok
e1d5s04 2, 4 DA1 1853 GiB normal ok
e1d5s04#029 0, 0 DA1 64 GiB draining failing/noPath/systemDrain/adminDrain/noRGD/noVCD
e1d5s05 2, 4 DA1 62 GiB normal ok
[...]
```

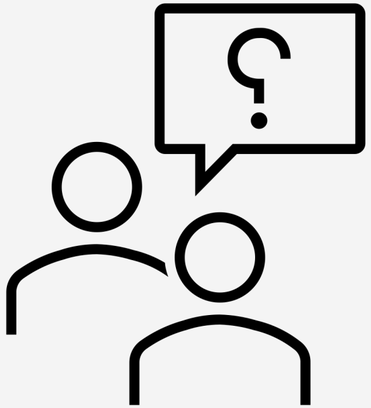
Temporary device Information that is stored until the rebuild and rebalance has completed

Reference links:

- https://www.ibm.com/docs/en/ess-p8/5.3.7?topic=command-mmvdisk-pdisk

- https://www.ibm.com/docs/en/ess-p8/6.1.2?topic=ess-problem-determination-guide

Live Q&A



Use Raise Hand feature in Webex for us to unmute your line.

Thank you for joining our webinar!