

Icahn School of Medicine at Mount Sinai

The Trials and Travails of multi-vendor and multi-version upgrades

Francesca Tartaglione

HPC Systems Admin, Mt Sinai

Spectrum Scale Day – Oct 24, 2018

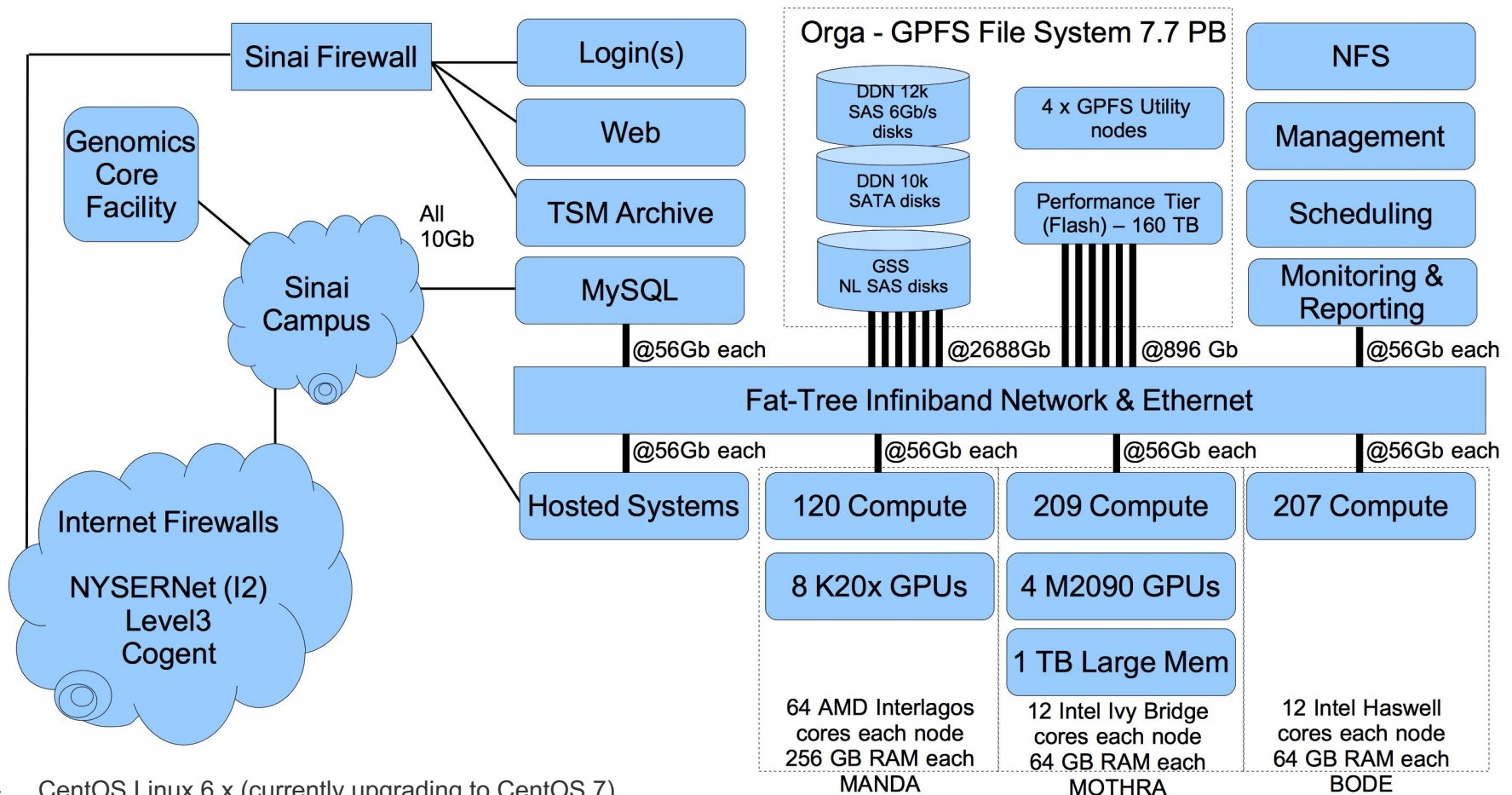


**Mount
Sinai**

Summary

1. Storage and cluster overview
2. 2017 upgrade
3. 2018 upgrade
4. Open Issues/Conclusions

Minerva as of first quarter 2017 – before the upgrades

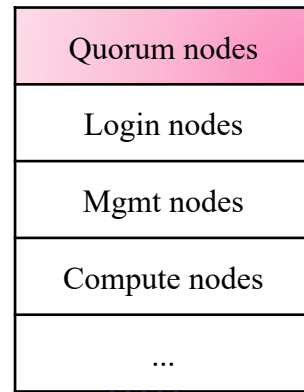


- CentOS Linux 6.x (currently upgrading to CentOS 7)
- Mellanox (FDR 56Gbps) Fat-Tree Infiniband
- 10 Gigabit Ethernet to Mt Sinai Campus Network
- LSF Scheduler
- Archival storage with Spectrum Protect (~ **16 PB** of archived data)

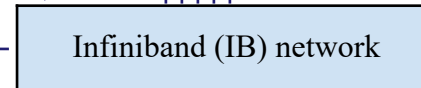
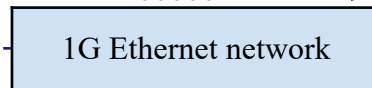
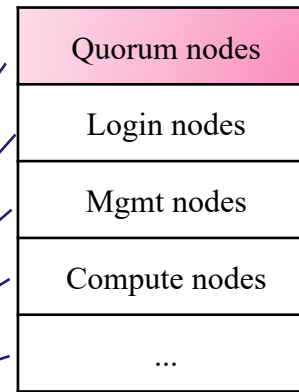
Storage and File System as of first quarter 2017

Client side:

Cluster Manda
(GPFS 3.5.0-25)

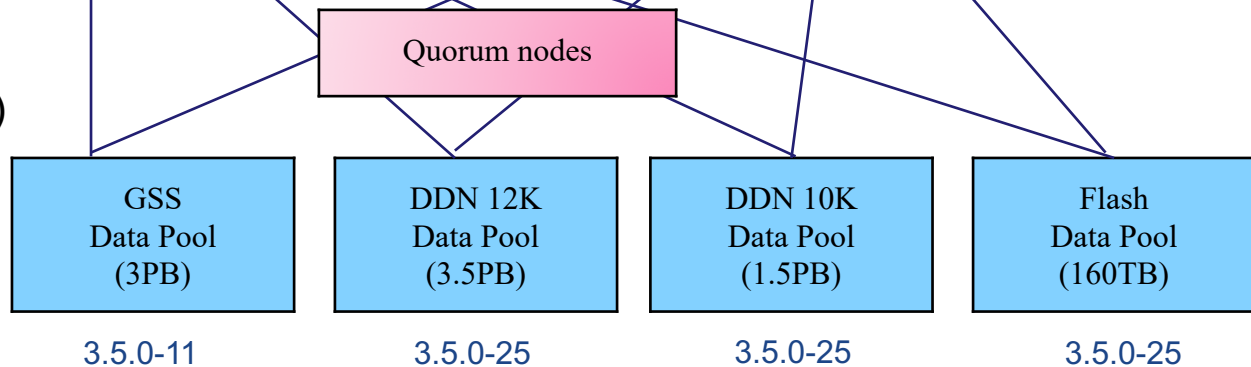


Cluster Mothra
(GPFS 3.5.0-25)



File-system side:

/sc/orga:
(GPFS 3.5.0-X)



- About 600 nodes mounting a single file system across 3 multi-clusters.
- More than 1.3 billions of files.
- Average file size is still in the KB's.

The Urgency of the Upgrade

Shortage of storage space:

5.7 PB was used (~75% full) and was fully allocated.

Reaching end of disk life:

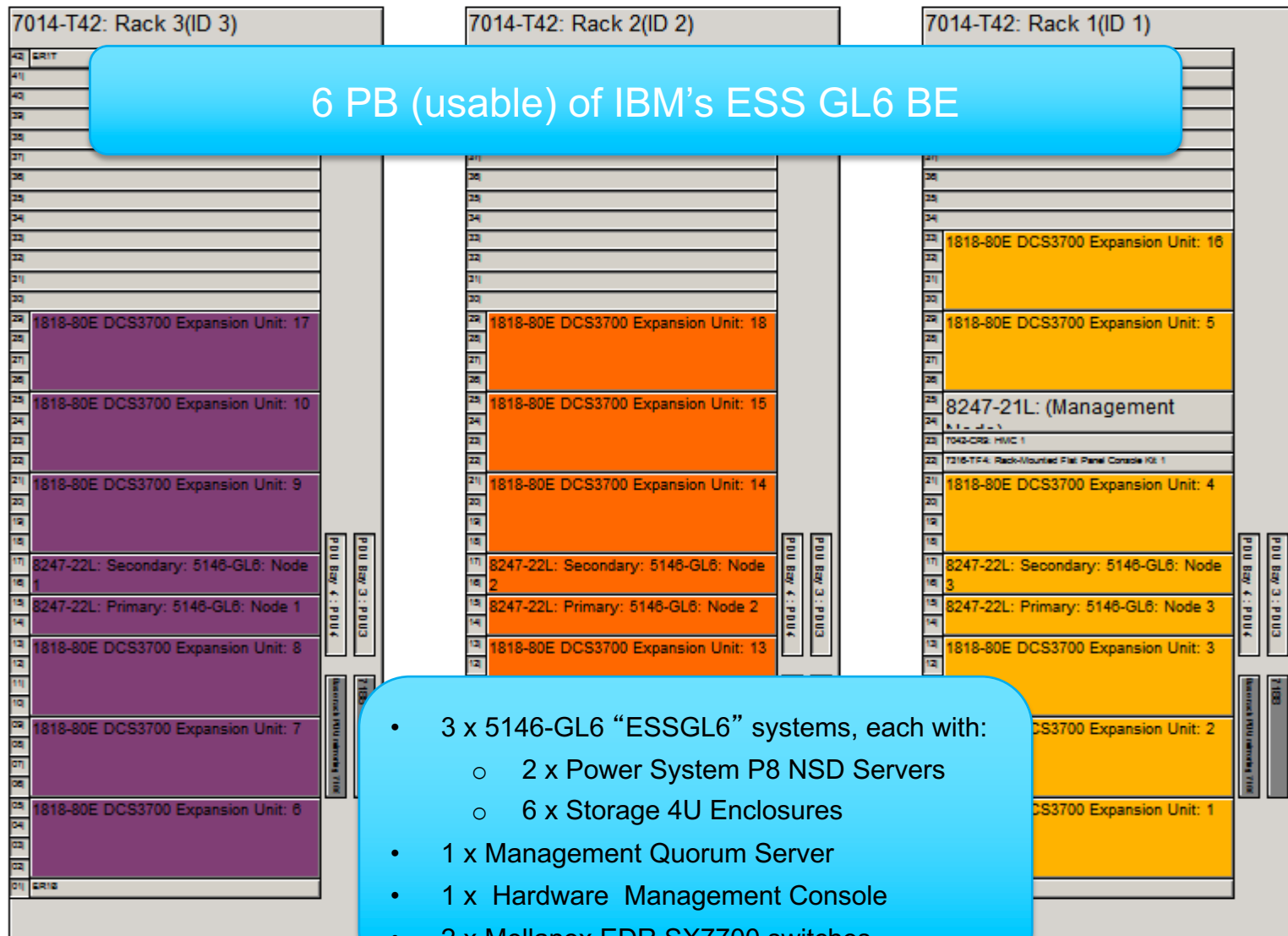
- ▶ IBM GSS was purchased at 2013
and neither IBM nor Lenovo will provide any more support for it.
- ▶ DDN 10K was purchased in 2012
and DDN will not provide any more support for it next year.

Out of support:

Spectrum Scale (GPFS) 3.5 was out of support.



Minerva 2017 Spectrum Scale Upgrade – OUR SOLUTION



ESS Upgrade PLAN Overview

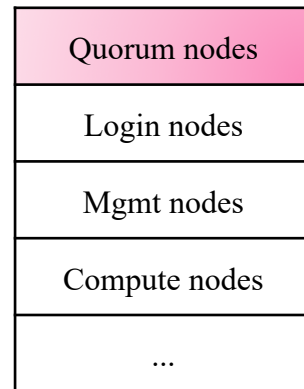
1. Install ESS with GPFS downgraded to 4.1 and test ESS as standalone cluster.
2. Upgrade all cluster components of the production Spectrum Scale cluster to Spectrum Scale 4.1.1-8 (except the GSS – no support from IBM for GPFS 4.1.1-8)
3. Delete temporary standalone cluster and add ESS to production Spectrum Scale cluster (using three recovery groups, one from each ESS building block) 
4. Migrate GSS file data to ESS
5. Migrate DDN 10K file data to ESS
6. Retire DDN 10K and GSS
7. === Switch whole cluster to **GPFS 4.1.1-8** =====
8. Upgrade client cluster and file system cluster to GPFS 4.2
9. === Switch whole cluster to **GPFS 4.2** =====

Goal: Minimum number of outages

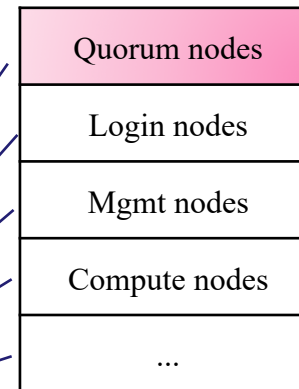
Storage and File System Upgrade Plan

Client side:

Cluster Manda
(GPFS 4.1=>4.2)



Cluster Mothra
(GPFS 4.1=>4.2)



1G Ethernet network

Infiniband (IB) network

File-system side:

/sc/orga:
(GPFS 4.1=>4.2)

Quorum nodes

ESS
Data Pool
(6PB)

~~GSS
Data Pool
(3PB)~~

DDN 12K
Data Pool
(3.5PB)

~~DDN 10K
Data Pool
(1.5PB)~~

Flash
Data Pool
(160TB)

Data
Transfer



Our journey through the upgrade – issues

Issue: Two ESS nodes (ess1 and ess2) began arbitrating and several servers went into an unknown state.

```
[root@ess1 ~]# tail /var/adm/ras/mmfs.log.latest
Fri Aug 4 11:39:53.348 2017: [D] Pdisk n2s01 of RG rg_ess1-ib0 path //ess2-ib0/dev/sda10: SCSI op=0x00 Test Unit Ready: Iocctl or RPC Failed: err=233.
Fri Aug 4 11:40:48.507 2017: [D] Pdisk n1s01 of RG rg_ess1-ib0 path //ess1-ib0/dev/sda10: SCSI op=0x00 Test Unit Ready: Iocctl or RPC Failed: err=668.
Fri Aug 4 11:42:53.354 2017: [D] Pdisk n2s01 of RG rg_ess1-ib0 path //ess2-ib0/dev/sda10: SCSI op=0x00 Test Unit Ready: Iocctl or RPC Failed: err=233.
Fri Aug 4 11:43:48.531 2017: [D] Pdisk n1s01 of RG rg_ess1-ib0 path //ess1-ib0/dev/sda10: SCSI op=0x00 Test Unit Ready: Iocctl or RPC Failed: err=668.
Fri Aug 4 11:45:53.477 2017: [D] Pdisk n2s01 of RG rg_ess1-ib0 path //ess2-ib0/dev/sda10: SCSI op=0x00 Test Unit Ready: Iocctl or RPC Failed: err=233.
Fri Aug 4 11:46:48.734 2017: [D] Pdisk n1s01 of RG rg_ess1-ib0 path //ess1-ib0/dev/sda10: SCSI op=0x00 Test Unit Ready: Iocctl or RPC Failed: err=668.
Fri Aug 4 11:48:53.590 2017: [D] Pdisk n2s01 of RG rg_ess1-ib0 path //ess2-ib0/dev/sda10: SCSI op=0x00 Test Unit Ready: Iocctl or RPC Failed: err=233.
Fri Aug 4 11:49:48.747 2017: [D] Pdisk n1s01 of RG rg_ess1-ib0 path //ess1-ib0/dev/sda10: SCSI op=0x00 Test Unit Ready: Iocctl or RPC Failed: err=668.
Fri Aug 4 11:51:53.676 2017: [D] Pdisk n2s01 of RG rg_ess1-ib0 path //ess2-ib0/dev/sda10: SCSI op=0x00 Test Unit Ready: Iocctl or RPC Failed: err=233.
Fri Aug 4 11:52:48.834 2017: [D] Pdisk n1s01 of RG rg_ess1-ib0 path //ess1-ib0/dev/sda10: SCSI op=0x00 Test Unit Ready: Iocctl or RPC Failed: err=668.

[root@ess2 ~]# tail /var/adm/ras/mmfs.log.latest
Fri Aug 4 11:44:01.933 2017: [E] Pdisk e3d2s06 of RG rg_ess1-ib0 failed; state missing/noPath/systemDrain location 'SV63045348-2-6' FRU '00VJ361' WWID 5000cca2608d
baec
Fri Aug 4 11:44:01.934 2017: [D] Pdisk e3d2s06 of RG rg_ess1-ib0 state changed from diagnosing/noPath to missing/noPath/systemDrain.
Fri Aug 4 11:44:01.935 2017: [D] Pdisk e3d2s06 of RG rg_ess1-ib0 turning failed light on.
Fri Aug 4 11:46:13.144 2017: [N] Scheduled takeover of recovery group rg_ess2-ib0 in approximately 180.0 seconds.
Fri Aug 4 11:49:13.467 2017: [N] Scheduled takeover of recovery group rg_ess2-ib0 in approximately 180.0 seconds.
Fri Aug 4 11:51:06.087 2017: [D] Pdisk e1d3s04 of RG rg_ess1-ib0 path //ess2-ib0/dev/sdps status changed from active to notEnabled.
Fri Aug 4 11:51:06.088 2017: [D] Pdisk e1d3s04 of RG rg_ess1-ib0 path //ess2-ib0/dev/sdys status changed from active to notEnabled.
Fri Aug 4 11:51:06.089 2017: [D] Pdisk e1d3s04 of RG rg_ess1-ib0 state changed from ok to diagnosing/noPath.
Fri Aug 4 11:51:06.090 2017: [D] Pdisk e1d3s04 of RG rg_ess1-ib0 state changed from ok to diagnosing/noPath.
Fri Aug 4 11:52:13.985 2017: [N] Scheduled takeover of recovery group rg_ess2-ib0 in approximately 180.0 seconds.
```

- ▶ Started with jamming on IB communication of ESS servers, no panic or reboot, but no failovers. Happened with system loaded with jobs and we eventually lost some disks. Needed several hours of recovery time.

Steps taken:

- ✓ Recovered servers, disks
- ✓ Reseated IB cables from ESS side and started IB network review.

Our journey through the upgrade – issues

Issue: Another two sets of ESS nodes began arbitrating and several servers went into an unknown state.

- ▶ Tried to take a gpfs.snap → cancelled it as computing nodes were starting to unmount gpfs (Stale Handle)
- ▶ Some ESS pairs began arbitrating and we needed to call IBM again to recover the disks.
- ▶ Waiters from old (QDR cards with OFED 1.5.3) based clients in Manda AMD GPFS cluster
- ▶ dmesg with a lot of IB errors
- ▶ users reported hangs

Steps taken:

- ✓ Recovered servers and disks, and several stale clients.
- ✓ Upgraded Utility nodes to GPFS 4.1.1-15
- ✓ Upgraded (rolling fashion)
 - firmware and Mellanox OFED on ESS EDR HCAs
 - SX7700 switches firmware
- ✓ Revisited RDMA parameters [pagepool, verbsSendBufferMemoryMB] on server and client clusters
- ✓ Set modes uniformly to datagram on ESS servers

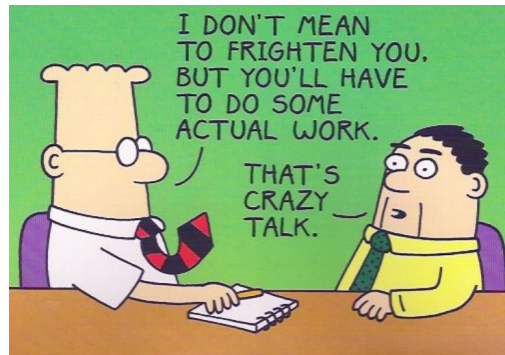
We suspended the ESS disks and resumed all GSS disks

Our journey through the upgrade



One-day planned maintenance

- ▶ Upgraded production Spectrum Scale cluster to Spectrum Scale 4.1.1-15
 - 8 x 3650 NSD Flash servers
 - 4 x DDN12K NSDs (upgrade only GPFS, not GridScaler)
 - 4 x DDN 10K NSDs (upgrade only GPFS, not GridScaler)
 - all compute nodes (~ 600 nodes)
- ▶ Migrated ESS file data back to GSS
 - re-stripe to move data from ESS NSDs to GSS NSDs (mmrestripefs command)
 - remove ESS NSDs from file system



Our journey through the upgrade



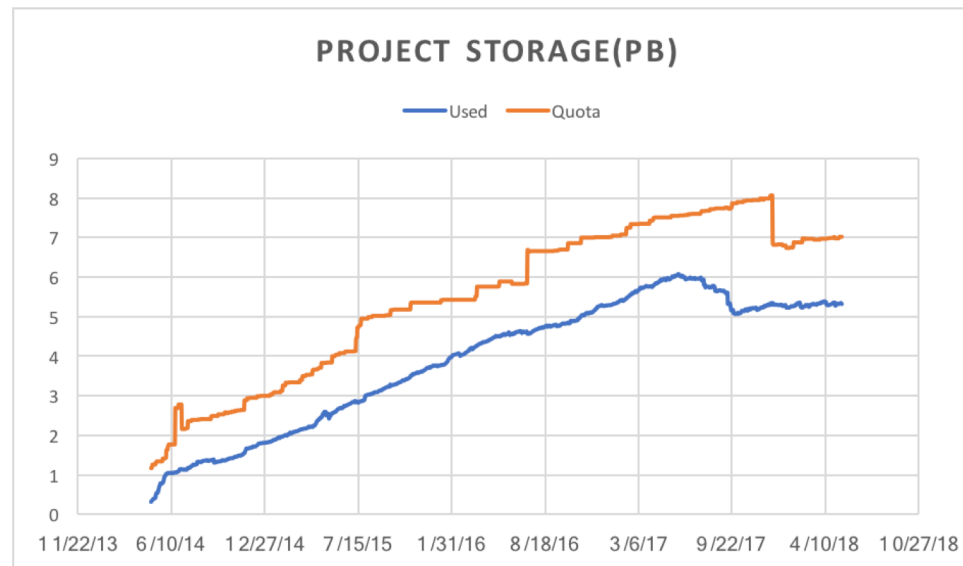
1. Created new cluster with EMS and ESS IO nodes → heavy testing
 - Delete ESS NSDs, vdisks and recovery groups
 - Upgrade ESS to current code (5.1.1) → then downgrade Spectrum Scale to 4.1.1-15
2. Upgraded GPFS on GSS from 3.5.0-11 to 3.5.0-25 (latest GPFS version tested on GSS)
3. Integrated the ESS into the file system
4. Final migration of the data
 - *DDN10k 700TB @ 50TB/day in 2 weeks*
 - *GSS 2PB @ 35TB/day in 7 weeks*
5. Removed GSS and DDN10K

(Cluster still at 4.1 as Spectrum Scale version)

Minerva 2018 Spectrum Scale Upgrade

Reasons:

- ▶ Storage usage continuing to grow
- ▶ After deleting GSS (3PB) and DDN 10K(1.5PB) we still needed storage
- ▶ Old flash → End Of Service after 2018



Minerva 2018 Spectrum Scale Upgrade - Solution



42	ER1T
41	8831-S52: Switch 1: Switch - 1
40	
39	

Additional
Storage

4 PB (usable) of ESS GL6S - LE

- 2 x 5148-22L ES NSD systems
- 1 x 5148-21L ESS Management Server
- 5 x 5U84 drive enclosures with 82x10TB NL SAS drives
- 1 x 5U84 drive enclosure with 82x10TB NL SAS drives + 2x800GB SSD (for internal usage)

22	
21	
20	5148-22L: ESS Data Server 1 (Qty = 2(1))
19	
18	5148-22L: ESS Data Server 1 (Qty = 2(2))
17	
16	5147-084: ESS 5U84 Secondary Storage for Elastic Storage Server 1 (Qty = 5(4))
15	
14	
13	
12	
11	5147-084: ESS 5U84 Secondary Storage for Elastic Storage Server 1 (Qty = 5(5))
10	
09	
08	
07	
06	5147-084: ESS 5U84 Primary Storage for Elastic Storage Server 1
05	
04	
03	
02	
01	ER1B

5147-024: ESS 2U24 Storage for Elastic Storage Server 1 (Qty = 2(2))
5148-22L: ESS GS2S Data Server 2 (Qty = 2(2))
5148-22L: ESS GS2S Data Server 1 (Qty = 2(1))
5147-024: ESS 2U24 Storage for Elastic Storage Server 1 (Qty = 2(1))

264 TB(usable) of ESS GS2S - LE

To replace
older flash

- 2 x 5148-22L ES NSD systems
- 2 x 2U24 drive enclosures with 48x3.84TB SSDs

Minerva 2018 Spectrum Scale Upgrade - Plan

- ▶ Basic ESS installation from IBM Lab Services
- ▶ Add ESS GL6S and GS2S to production Spectrum Scale cluster
 - Migrate metadata pool from old flash to ESS GS2S 
- ▶ Upgrade production cluster to Spectrum Scale 4.2



Our journey through the upgrade – main issue

- Issues with IB → ib0 (IPoIB interface) going down randomly from ibdev2netdev (but ibstat State: Active) on the BE systems

```
[root@ess4 ~]# ibdev2netdev
```

```
mlx5_0 port 1 ==> ib0 (Down)
```

```
mlx5_1 port 1 ==> ib1 (Up)
```

```
mlx5_2 port 1 ==> ib2 (Up)
```

```
mlx5_3 port 1 ==> ib3 (Up)
```

```
mlx5_4 port 1 ==> ib4 (Up)
```

```
mlx5_5 port 1 ==> ib5 (Up)
```

- ESS node/s in arbitrating mode
 - service gpfs stop
 - ifdown/ifup or service openibd restart
 - service gpfs start
- Log ASSERT after GPFS restart → other storage nodes to go arbitrating



Cluster DOWN!!!!



Our journey through the upgrade

Actions taken:

- ▶ Removed GS2S NSDs from the cluster
 - migrated back the metadata on them
- ▶ Upgraded:
 - Client clusters to Spectrum Scale 4.2.3-8 (~600 nodes)
 - All ESS (BE and LE) to gss-toolkit 5.2.2 (and GPFS 4.2.3-8)
- ▶ Upgraded OFED
 - Utility node that was running the Master SM had a very old driver version (MLNX_OFED_LINUX-2.0-3.0.0.3, opensm-4.0.5) → upgraded to MOFED 4.1.1
 - Everywhere else to MOFED 4.1 where possible
- ▶ Finally got rid of the assert but...
 - ... ib0 still going down on the ESS nodes from time to time (randomly)

At least we could restart gpfs without causing system wide outages!

Our journey through the upgrade

- ▶ Worked together with IBM and Mellanox
 - High LinkErrorRecoveryCounter and PortRcvErrors counters
 - /var/adm/ras/mmfs.log.latest

```
Mon Jul 16 08:01:13.172 2018: [I] VERBS RDMA async event IBV_EVENT_PORT_ACTIVE on mlx5_0 port 1.
Mon Jul 16 08:01:13.222 2018: [I] VERBS RDMA async event IBV_EVENT_PORT_ACTIVE on mlx5_0 port 1.
Mon Jul 16 08:01:29.447 2018: [W] VERBS RDMA async event IBV_EVENT_PORT_ERR on mlx5_5 port 1.
Mon Jul 16 08:01:29.449 2018: [W] VERBS RDMA port state changed from IBV_PORT_ACTIVE to IBV_PORT_DOWN for device mlx5_5 port 1.
Mon Jul 16 08:01:32.522 2018: [I] VERBS RDMA async event IBV_EVENT_PORT_ACTIVE on mlx5_0 port 1.
Mon Jul 16 08:01:32.571 2018: [I] VERBS RDMA async event IBV_EVENT_PORT_ACTIVE on mlx5_0 port 1.
Mon Jul 16 08:01:34.660 2018: [W] VERBS RDMA async event IBV_EVENT_CLIENT_REREGISTER on mlx5_5 port 1.
Mon Jul 16 08:01:34.661 2018: [W] VERBS RDMA port state changed from IBV_PORT_DOWN to IBV_PORT_INIT for device mlx5_5 port 1.
Mon Jul 16 08:01:34.932 2018: [I] VERBS RDMA async event IBV_EVENT_PORT_ACTIVE on mlx5_5 port 1.
Mon Jul 16 08:01:34.933 2018: [I] VERBS RDMA port state changed from IBV_PORT_INIT to IBV_PORT_ACTIVE for device mlx5_5 port 1.
Mon Jul 16 08:01:39.922 2018: [I] VERBS RDMA async event IBV_EVENT_PORT_ACTIVE on mlx5_0 port 1.
Mon Jul 16 08:01:39.971 2018: [I] VERBS RDMA async event IBV_EVENT_PORT_ACTIVE on mlx5_0 port 1.
```

- /var/log/messages

```
[Tue Jul 24 17:45:29 2018][ 38.941779] systemd-networkd[8773]: ib0: gained carrier
[Tue Jul 24 17:45:29 2018][ 38.942647] systemd-networkd[8773]: ib0: lost carrier

[Sun Jul 29 10:30:29 2018][405936.261382] systemd-networkd[71361]: Enumeration completed
[Sun Jul 29 10:30:29 2018][405936.261530] systemd[1]: Started Network Service.
[Sun Jul 29 10:30:29 2018][405936.262461] systemd-networkd[71361]: ib0 : gained carrier
[Sun Jul 29 10:30:37 2018][405944.745035] systemd-networkd[71361]: ib0 : lost carrier
[Sun Jul 29 10:30:37 2018][405944.751840] systemd-networkd[71361]: ib0 : gained carrier
[Sun Jul 29 10:31:16 2018][405983.745194] systemd[1]: Starting Network Service...
[Sun Jul 29 10:31:16 2018][405983.760190] systemd-networkd[71839]: Enumeration completed
[Sun Jul 29 10:31:16 2018][405983.760333] systemd-networkd[71839]: ib0 : lost carrier
[Sun Jul 29 10:31:16 2018][405983.760457] systemd[1]: Started Network Service.
[Sun Jul 29 10:31:16 2018][405983.760905] systemd-networkd[71839]: ib0 : gained carrier
```

Our journey through the upgrade

Actions suggested and taken:

- Reseat cables (both ends) on the ib0 interfaces
- Change the ifcfg files for all IPoIB interfaces for all ESS to NM_CONTROLLED=no

```
[root@ess1 ~]# cat /etc/sysconfig/network-scripts/ifcfg-ib0
DEVICE=ib0
BOOTPROTO=none
IPADDR=172.29.43.67
PREFIX=22
ONBOOT=yes
TYPE=Infiniband
CONNECTED_MODE=no
NM_CONTROLLED=no
MTU=2044
IPV6INIT=no
```

But after the changes, we could still see high counters (LinkErrorRecoveryCounter, PortRcvErrors)

Our journey through the upgrade

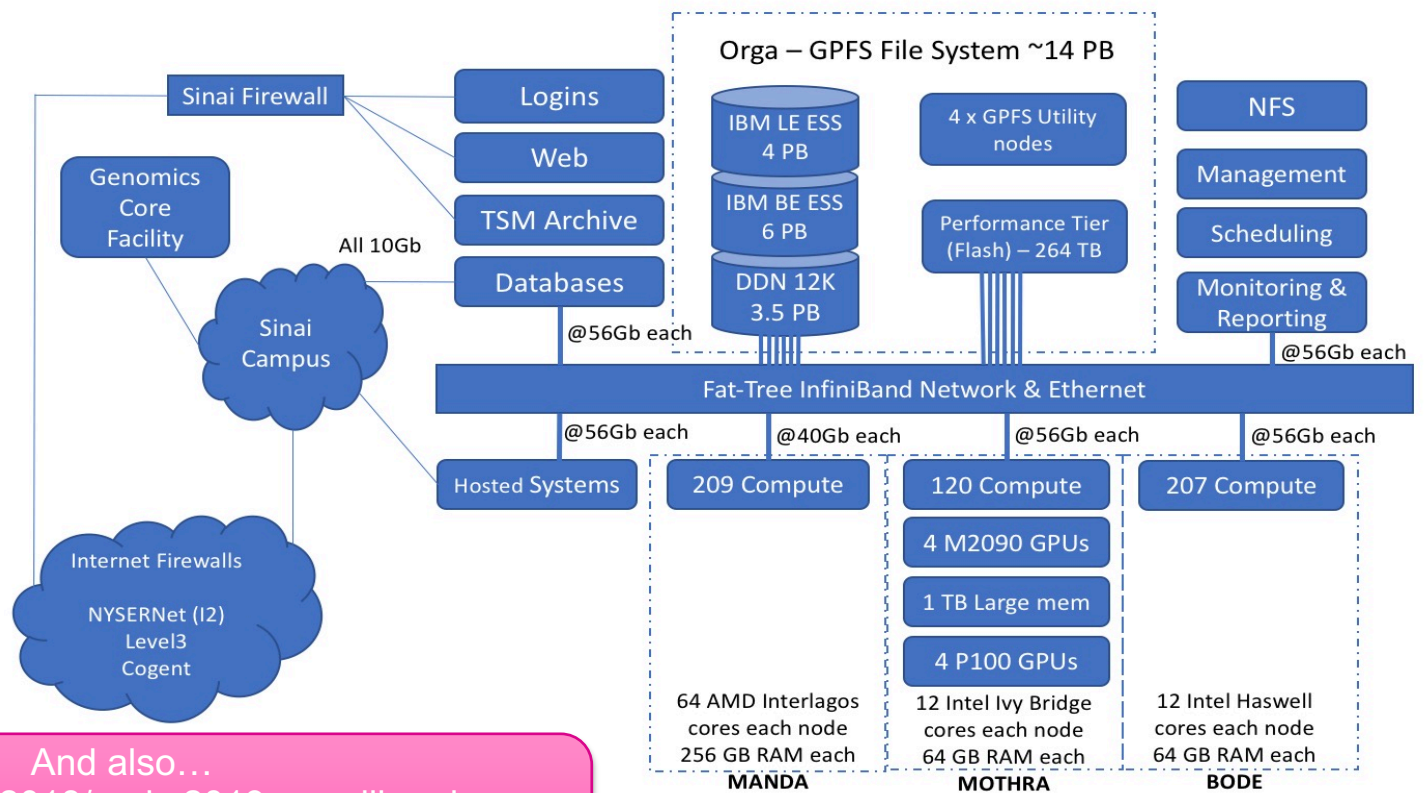
- ▶ Onsite visit from Mellanox → identified cooling issue
 - Fiber cables too long (20 meters) → really sensitive to temperature
 - SFP at both ends of the cables were burning hot
 - HCA temperatures were above 70°C, sometimes even 90°C! (max recommended 55°C)
 - EDR switches installed with the airflow in the wrong direction
 - Not sufficient airflow in the back of the ESS racks (due to tight cabling and the sound proof panel)
- ▶ Actions taken:
 - Replaced (where possible) fiber with copper cables
 - Reverse EDR switches
 - Removed the sound proof panel from the back of the racks

Since then we haven't seen ib0 going down anymore and
the cluster has been stable
(still some high counters happening on a couple of ESS)



Minerva as of today – after the upgrades

- ▶ Integrated back the new ESS GS2S into the production cluster
- ▶ Migrated all the metadata
- ▶ Removed the old flash
- ▶ **Next steps:** Integrate the last 2 ESS GL6S → **Spectrum Scale 5**



And also...
... by the end of 2018/early 2019 we will replace
most of the compute and infrastructure nodes
(~15,000 cores)

Open issues

- ▶ Still some high *LinkErrorRecoveryCounter* and *PortRcvErrors* counters on a couple of ESS NSDs
- ▶ HCA temperatures still above 55°C:

HCA port	ess1	ess2	ess3	ess4	ess5	ess6	ess7	ess8	nsd01	nsd02	nsd03	nsd04
ib0	67	66	64	62	62	63	62	65	64	60	67	65
ib1	67	66	64	62	62	63	62	65	64	60	67	65
ib2	68	69	62	67	60	61	69	64	63	62	64	64
ib3	68	69	62	67	60	61	69	64	63	62	64	64
ib4	66	66	62	66	62	60	65	67	68	57	62	60
ib5	66	66	62	66	62	60	65	66	68	57	62	60

- ▶ Still investigating with IBM and Mellanox
 - Evaluating to reroute the ESS NSDs network cables to let the air flow properly
- ▶ While taking gpfs.snap the ESS reboot themselves randomly

Minerva Spectrum Scale Upgrades



To the IBM Lab Services and Support teams
for the continuous support and assistance during the upgrades



Minerva Spectrum Scale Upgrades



- ▶ Please feel free to share your experiences and recommendations with us !!!

