NDGF Site Report

2013-04-15, Hepix spring 2013, Bologna

Overview

New organisation

Storage & Computing

Subsite reports

CSC

NSC

HPC2N

Unige

UiO

New organisation

New slides template!

SW maintenace for Tier1 software contracted

Generic area and BMS area coordinators hired

Storage and computing

Distributed dCache with storage in 8 sites at 5PB

One storage incident last 6 months

After scheduled power cycle 4 raid6 sets came up with 2 disks missing

All of the disks were Seagate Constellation ES ST32000444SS. These disks have a bug in their firmware (even at the latests, KS6B), causing them not to spin up after power cycles.

The remaining disks had uncorrectable read errors at some block making recovery difficult

This in part due to raid parity check cron jobs not run due to root missting form /etc/shadow

In the end, lots of manual labour rescued 3 pools, the 4th had too many filesystem errors to recover within reasonable time

Storage and computing

Computing

ARC

Atlas ND cloud:

ATLAS Grid Monitor

2013-04-15 CEST 09:30:15

Processes: Grid Local

Country	Site	CPUs	Load (processes: Grid+local)	Queueing
Denmark	**Steno Tier 1 (DCSC/KU)**	5332	799+3202	**1190**+0
Germany	**LRZ-LMU**	1844	1429+59	**12**+0
Norway	**Abel C (UiO/USIT)**	10736	0+6943	**0**+0
	EPF (UiO/FI)	120	0+1	**0**+0
Slovenia	**Arctur-1**	24	0+0	**0**+0
	Arnes	2020	1779+0	**449**+0
	SiGNET	2106	1816+8	**633**+0
Sweden	**Alarik (SweGrid, Luna>**	3776	400+2474	**133**+0
	Ritsem (SweGrid, HPC2>	544	382+0	**145**+0
	Siri (SweGrid, Lunarc)	456	70+379	**1474**+45
	Smokerings (NSC)	520	325+149	**717**+20
	Tintin (SweGrid, Uppm>	2624	0+2367	**0**+855
Switzerland	**Bern ATLAS T2**	532	103+0	**144**+0
	Geneva ATLAS T3	405	0+372	**0**+242
	Manno PHOENIX T2	2240	623+1265	**90**+173
	Manno PHOENIX T2	2240	589+1299	**89**+174
UK	**UKI-LT2-IC-HEP Grid C>**	4	360+2734	**97**+1412
	UKI-SCOTGRID-GLASGOW	1980	0+0 (no queue info)	**0**+0
TOTAL	***18 sites***	***37503***	***8675 + 21252***	***5173 + 2921***

CSC site “report”

- Issues having 2TB disks in Sun x4540 Thumpers
- Upgraded firmware, OS, etc
- Still, not all disks come up when rebooting, but what has helped is:
  - Use lsiutil and:
    - 1.) set all ports in the controller to 1.5/1.5 or 3.0/3.0
    - 2.) on each disks that did not come online or came online late, we changed the port speed to _something else_ (1.5/1.5, 1.5/3.0 or 3.0/3.0)
  - Counters indicate "physical reset problem"
  - Anyone with experience in this, contact me later

NSC site report

The new computer room building is nearing completion. Builders claim it will be ready to move into in June (knock on wood).

Ordered another 2 petabyte of disk storage (Lustre) for the Swedish met-office (12 HP SL4540 × 60 disks à 3 Tbyte). Hardware expected to arrive in a week or two.

Problems with Mellanox optical FDR InfiniBand cables in new clusters. Have replaced ~5% of the cables over 8 months.

HPC2N site report

HP DL180G6 dCache pools, still having HDD issues such as elevated failure rate and repeated occurrances of drives that trigger spontaneous rebuild or even loop rebuilds. Issue escalated within HP, they have now collected 5 HDDs with the loop/spontaneous rebuild issue to perform special analysis on them.

The HDDs are HP 2T SATA 7kRPM LFF MB2000EAMZF, aka Seagate Constellation ES ST32000644NS. These were delivered in September 2010 in a shipment of 34 HP DL180G6:s each with 12 HDDs. A later shipment of 15 DL180G6:s with 3T MB3000EBKAB HDDs have not shown any elevated failure rates.

Replacement drives are predominantly HP MB2000EBZQC, aka Seagate ST2000NM0011. We haven't seen any abnormal issues with these drives, but haven't been able to get HP to replace all our drives pending analysis. The issue is mainly that they don't fail frequently enough to warrant that action yet. We'll se what they say after analysis.

For those of you with any kind of Seagate HDD, be advised that there are a LOT of firmware bugs in them and you really really want to make sure that you follow firmware relases and get fresh firmware installed.

HPC2N site report

Firmware changelog for MB2000EAMZF:

Version: HPG5 (26 Oct 2012) After long term use of the HDD, a rare condition might occur following a power cycle where the drive heads may land on areas of the disk containing data, which could potentially cause data loss or mechanical damage. Firmware version HPG5 prevents this condition from occurring.

Version: HPG4 (16 Jan 2012) Firmware version HPG4 prevents a low likelihood condition where the drive will stop responding after a frame error and loss of sync. If this condition occurs, a drive power cycle is required to recover.

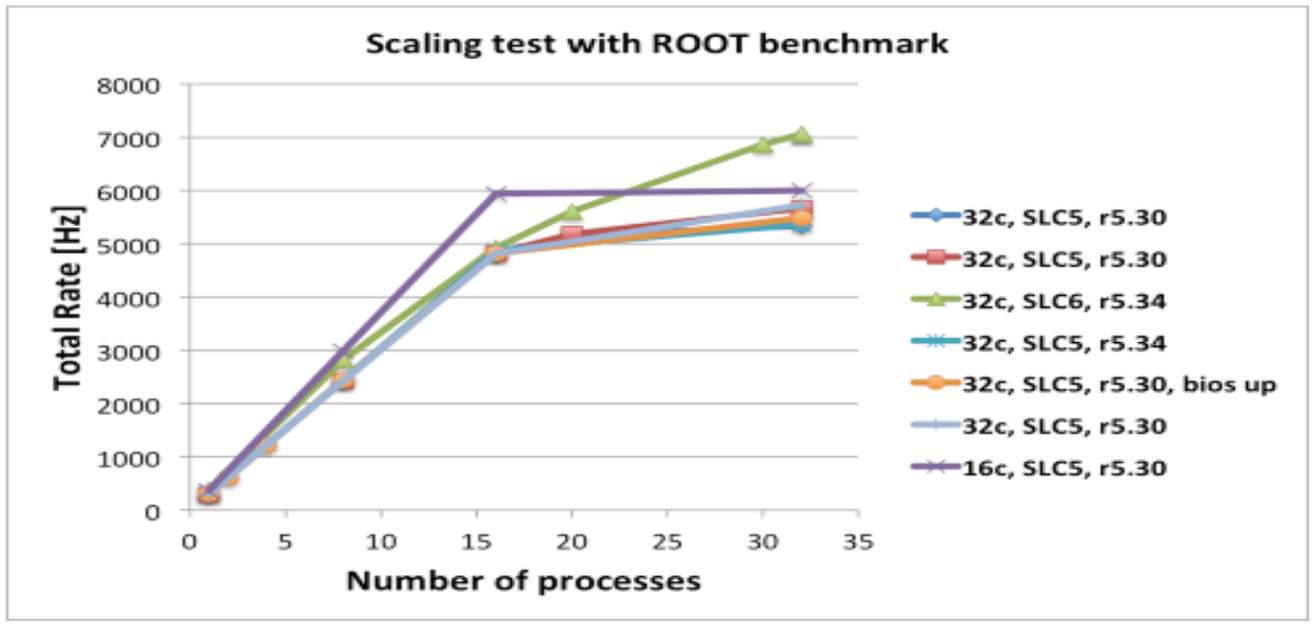
Version: HPG3 (8 Feb 2011) Corrects a drive hang that causes long command response times and drive resets

Version: HPG2 (22 Oct 2010) Corrects background task mishandling that can lead to drive resets and device fault condition.

Unige site report

Oddity in benchmarks on latest 2x16-core AMD nodes

Performance of 32-core machines



- an older 16-core machine behaves as one would expect
- latest 32-core do not
- ROOT version makes no difference
- upgrade from SLC5 to SLC6 helps

New Grid Resources at UiO

Abel

"Not just a group theory!"

Key facts Abel / new UiO set-up

- General purpose HPC cluster, MEGWARE MiriQuid
- 10k physical cores, Intel SNB E5-2670, 4 GB RAM per core(!)
- #96 June '12 Top500
- 10 Gbps connectivity to LHCOPN
- Full IB (FDR 56 Gbps) connectivity
- Parallel FS† for grid over IB
- HDS Enterprise Storage
  - HNAS 3090
    - 1PiB for dCache disk pools
  - HUS VM
    - 50 TiB for Operational FS
  - Spectra T-Finity
    - 579 TiB (Jaguar) dCache tape pools

† Session, RTE directories

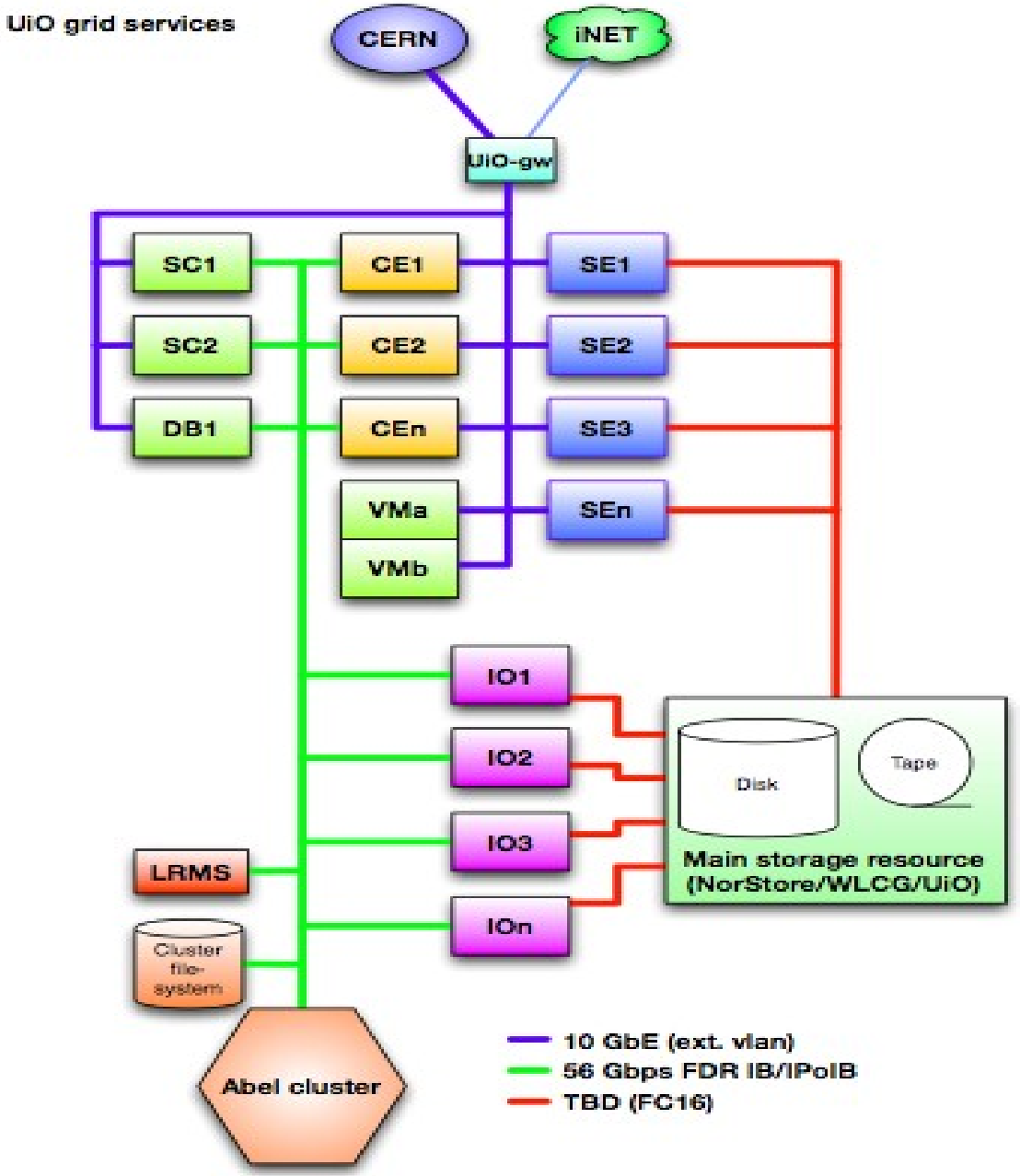
Main questions

Why IB on grid infrastructure?

Why Enterprise Storage?

Abel is part of the Norwegian HPC infrastructure and is a general purpose HPC cluster with an inclination towards life sciences. The design is optimized for IO and large memory jobs, a good match for sharing of resources with HEP.

Being part of a large investment means that the added cost of IB and enterprise storage is removed. In addition to performance the advantage is in the TCO, with greatly reduced operational costs and improved availability.


UiO grid services
CERN
iNET
UiO-gw
SC1
SC2
DB1
CE1
CE2
CEn
VMa
VMb
SE1
SE2
SE3
SEn
IO1
IO2
IO3
IOn
Disk
Tape
Main storage resource
(NorStore/WLCG/UiO)
LRMS
Cluster
file-
system
Abel cluster
10 GbE (ext. vlan)
56 Gbps FDR IB/IPoIB
TBD (FC16)

Abel at UiO

Enterprise Storage

HNAS 3090

Achieved performance: 105000 IOPS

GRID allowance: 1PiB

Block storage

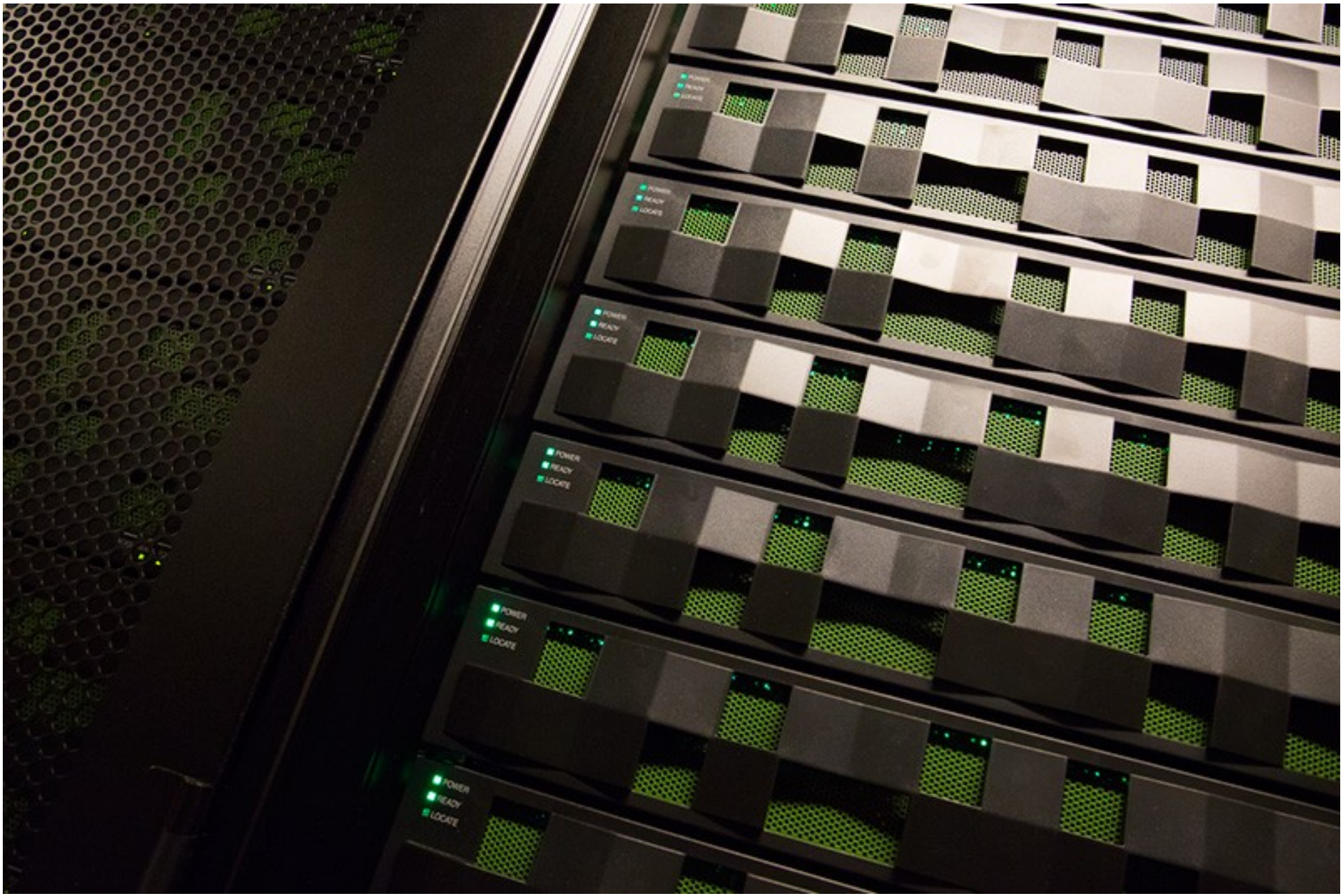
Achieved performance: 75000 IOPS

GRID Operational FS:

Spectra T-Finity TS1140 technology

4.0 TB native
12.0 TB compressed
Performance native up to 250 MBps
Performance compressed up to 650 Mbps
Reliable 237000 hours MTBF
Larger buffers – main buffer 1 GB
Secure – supports encryption and key management
Expandable to 3659 PB

GRID allowance: 579 TiB



HDS HNAS storage box at UiO

Spectra T-Finity tape library at UiO