



National Infrastructure Storage and Data

Scientific Computing @UiB





Agenda

<https://documentation.sigma2.no>

- NIRD
- Fram
- Policies
- Performance

Lorand Janos Szentannai





NIRD

National Infrastructure for Research Data

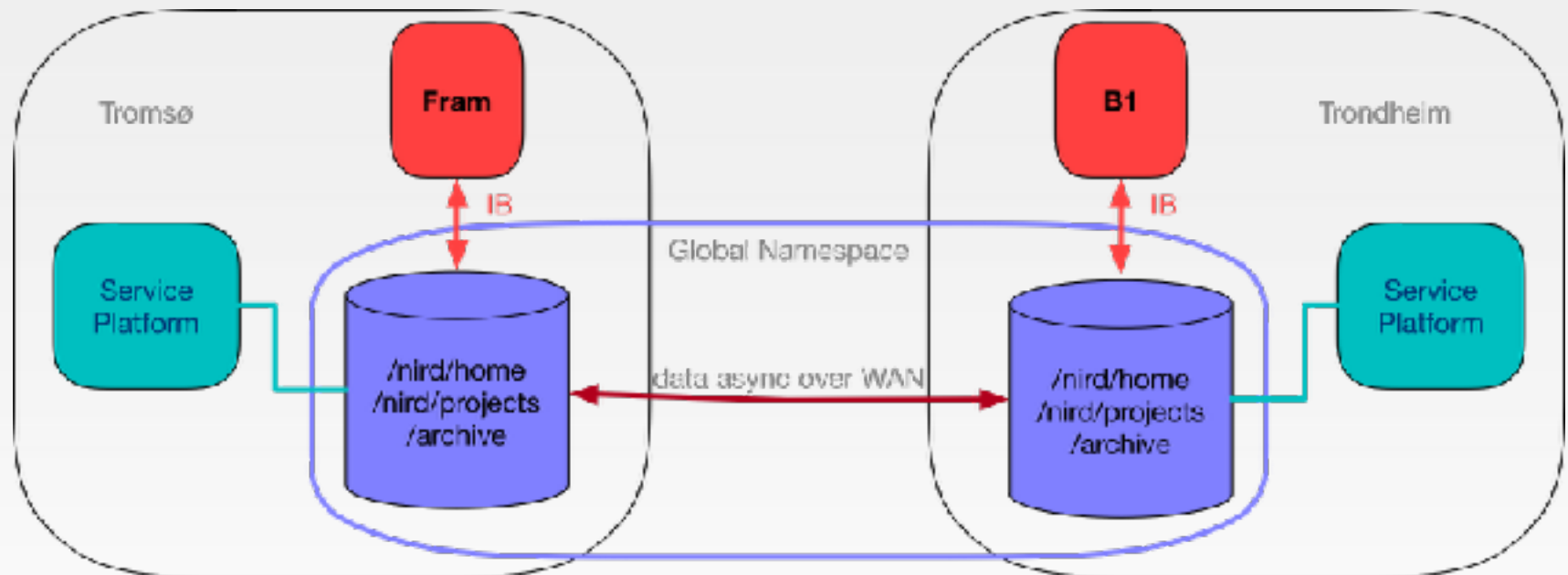
- data-centric approach
- 12PB total storage
- geo-distributed model
 - Tromsø
 - Trondheim
- GridScaler (GPFS)
- yearly capacity upgrades





NIRD

National Infrastructure for Research Data





Fram

HW Info

Details	Fram
System	Lenovo NeXtScale nx360
Number of Cores	32256
Number of nodes	1006
CPU type	Intel E5-2683v4 2.1 GHz Intel E7-4850v4 2.1 GHz (hugemem)
Max Floating point performance, double	1.1 Petaflop/s
Total memory	78 TiB
Total disc capacity	2.5 PB





Fram

HW Info

- 1004 dual socket and 2 quad socket nodes
- high-bandwidth low-latency Infiniband network
- island topology, with 9216 cores in each
- 3 types of computes
 - A: 2 x 16 core Broadwell CPU/ 64GB RAM
 - B: 8 pcs. large memory node 512GB RAM
 - C: 2 pcs. huge memory node 6TB RAM





Fram

Available storage areas

/cluster	2.5PB
/nird/home	0.5PB
/nird/projects	





Fram

Available storage areas

`/cluster/work/jobs/$SLURM_JOB_ID`

`/cluster/work/users/$USER`

`/cluster/projects/<project_name>`

`/cluster/shared/<meta-project_name>`

`/nird/home/$USER`

`/nird/projects/nird/NSxxxxxK`





/nird/home file system

- 0.5PB storage space
- 20GB default quota per user
- available only on login nodes, computes, NIRD

Usage:

- personal files (not private)
- scripts
- keys
- sources
- files for which backup is needed





/nird/home/\$USER

Policy

- not a shared area
- all data stored in \$HOME has to be treated as being private
- directory permissions are set to mode 0700
- periodically controlled and adjusted





/nird/home/\$USER

Backup

- daily snapshots for the last 7 days
- weekly snapshots for the last 6 weeks

/nird/home/.snapshots





/cluster/work/jobs/\$SLURM_JOB_ID

- job data
- set to \$SCRATCH
- no quota
- no backup
- cleaned up as soon as job finishes





/cluster/work/users/\$USER

- staging and cross-job data
- used for job output
- set to \$USERWORK
- automatic clean-up
- no quota
- no backup





/cluster/work/users/\$USER

Policy

- automatic cleanup always on
- the oldest files will be deleted first
- weekly scan removes files older than 42 days
- at 70%, files older than 21 days are subject to automatic deletion





/cluster/projects/nnXXXXk

- Notur project data
- local to the HPC cluster
- backup (snapshots)
- 1TB quota
- group enforcement*

* to be activated in the coming weeks





/cluster/projects/nnXXXXk

Backup

- daily sync over to NIRD
- on NIRD:
 - snapshots for the last 7 days
 - weekly snapshots for the last 6 weeks
 - ask support for file restore





/nird/projects/nird/NSxxxxxK

- Norstore project data
- not available on compute nodes
- backup (snapshots)
- min. 10TB quota





/nird/projects/nird/NSxxxxxK

- daily snapshots for the last 7 days
- weekly snapshots for the last 6 weeks

/nird/projects/nird/NSxxxxxK/.snapshots





/cluster/shared/<meta-project_name>

- project data shared between more projects
- input files
- group enforcement*
- quota enforced**

* to be activated in the coming weeks

** policy to be clarified in following weeks





Data handling and storage policy

Overview

	/nird/home	/cluster	/nird/projects
Available on login nodes	Y	Y	Y
Available on compute nodes	Y	Y	N
Available on NIRD	Y	N	Y
Automatic deletion	N	Y ¹	N
Backup	Y	N ²	N
Group enforcement	N	Y ³	N
Mode enforcement	Y	N	N
Performance (Fram)	II	I	II
Quota	Y	Y	Y

¹ /cluster/work

² except /cluster/projects & /cluster/shared

³ /cluster/projects & /cluster/shared





Lustre

Basics

- open-source
- parallel file system
- scalable
 - I/O throughput
 - capacity
 - 100k clients
- supports high-performance networks
- serving the most powerful super-computers in the world





Lustre

Features

- performance
- object base storage
- scalable
- stable
- POSIX compatible
- user, group and project* quotas
- RDMA

*not available on Hexagon





Lustre

Basics

Good for:

- large files
- sequential throughput
- parallel applications writing to different parts of file

Not good for:

- lots of small files
- high number of meta-data requests (i.e. `stat()`)





Lustre

Naming

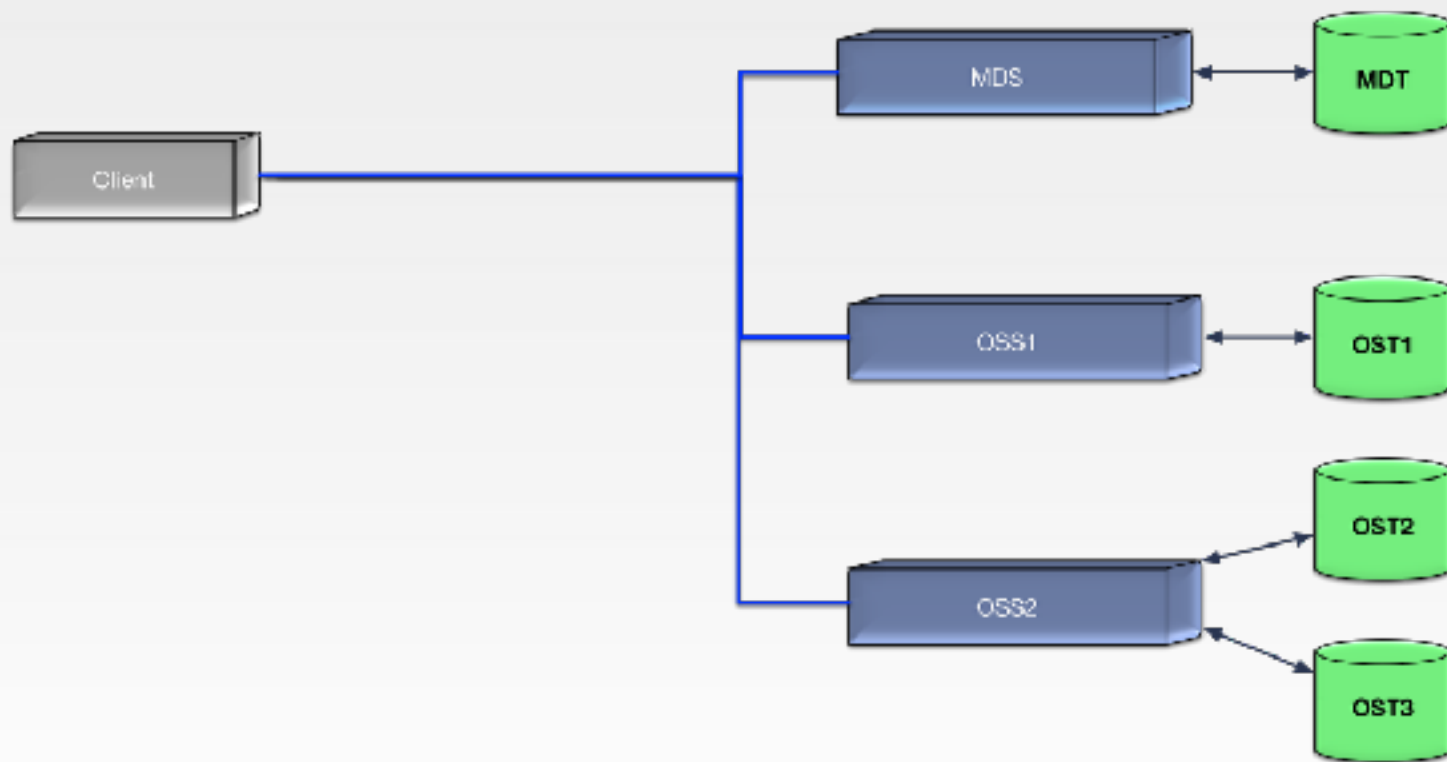
- MDS - Meta-data Server
 - MDT - Meta-data Target
- OSS - Object Storage Server
 - OST - Object Storage Target





Lustre

Diagram





Lustre

Striping

- set striping for new file or directory

```
lfs setstripe -c 2 new_file
```

- assign file to specific OST

```
lfs setstripe -c 1 --index 1 new_file
```

- get striping for a file

```
lfs getstripe my_file
```

```
lmm_stripe_count: 2
```

```
lmm_stripe_size: 1048576
```

```
lmm_pattern: 1
```

```
lmm_layout_gen: 0
```

```
lmm_stripe_offset: 9
```

obdidx	objid	objid	group
9	122006	0x1dc96	0
7	122006	0x1dc96	0





Lustre

Striping

Default stripe counts and sizes:

- /cluster 1 1M

Lustre setup for /cluster FS

- 2 MDSes in HA
- 8 OSSes
- 80 OSTs





Lustre

Striping

hexagon-4> lfs df /work

UUID	1K-blocks	Used	Available	Use%	Mounted on
bergen-MDT0000_UUID	438584324	1379684	407960224	0%	/work[MDT:0]
bergen-OST0000_UUID	15614268268	3612643624	11220464516	24%	/work[OST:0]
bergen-OST0001_UUID	15614268268	2479085728	12354028284	17%	/work[OST:1]
bergen-OST0002_UUID	15614268268	3143063288	11690053244	21%	/work[OST:2]
bergen-OST0003_UUID	15614268268	1849733268	12983382308	12%	/work[OST:3]
bergen-OST0004_UUID	15614268268	2928858204	11904255136	20%	/work[OST:4]
bergen-OST0005_UUID	15614268268	2686063596	12147044764	18%	/work[OST:5]
bergen-OST0006_UUID	15614268268	3069261116	11763849648	21%	/work[OST:6]
bergen-OST0007_UUID	15614268268	3004054760	11829055652	20%	/work[OST:7]
bergen-OST0008_UUID	15614268268	2556402700	12276707356	17%	/work[OST:8]
bergen-OST0009_UUID	15614268268	2590198964	12242906408	17%	/work[OST:9]
bergen-OST000a_UUID	15614268268	2145642356	12687475804	14%	/work[OST:10]
bergen-OST000b_UUID	15614268268	2449161200	12383957936	17%	/work[OST:11]





Resources

Contact & Documentation

Documentation: <https://documentation.sigma2.no>

Announcements: <https://opslog.sigma2.no>

 @MetacenterOps

Support mail: support@metacenter.no





Questions?



UNIVERSITY OF BERGEN