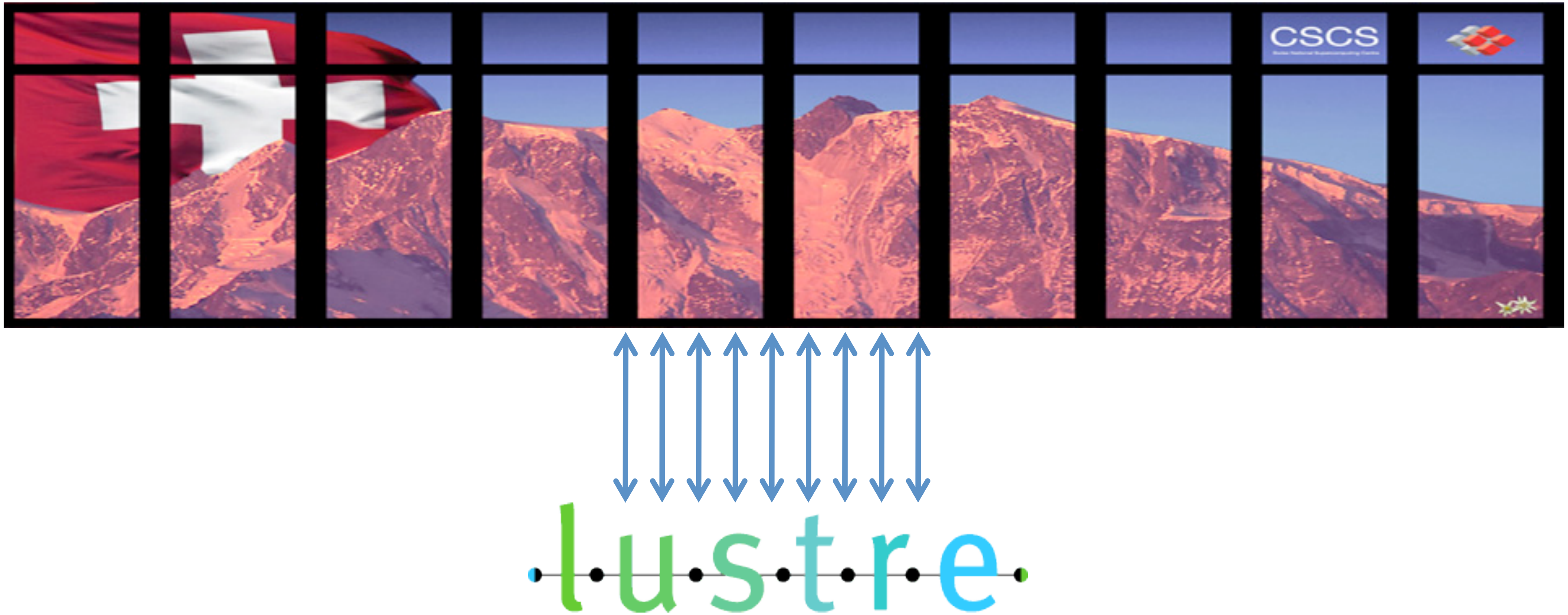
I/O on Rosa

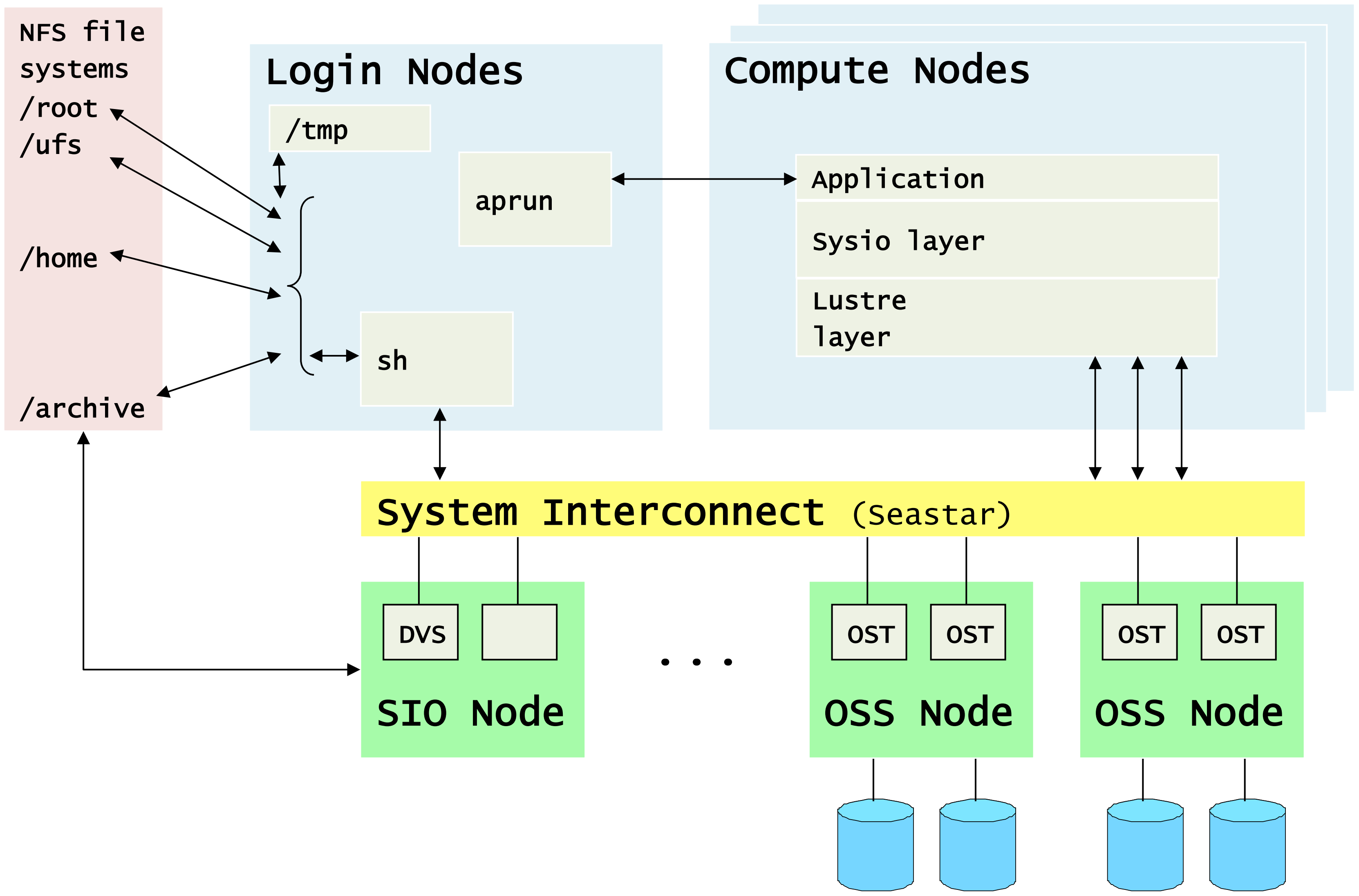


Jason Temple

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre

Cray XT I/O architecture

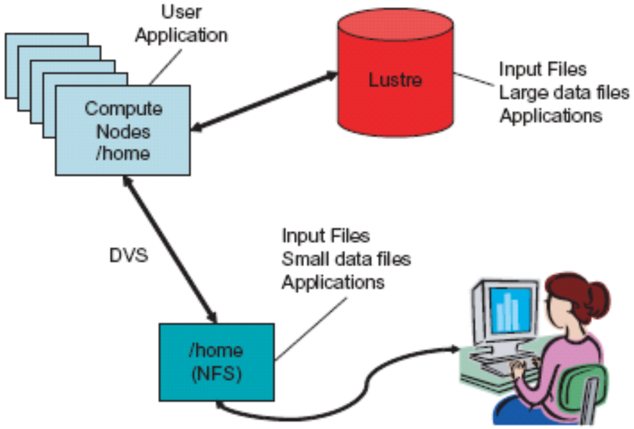


Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

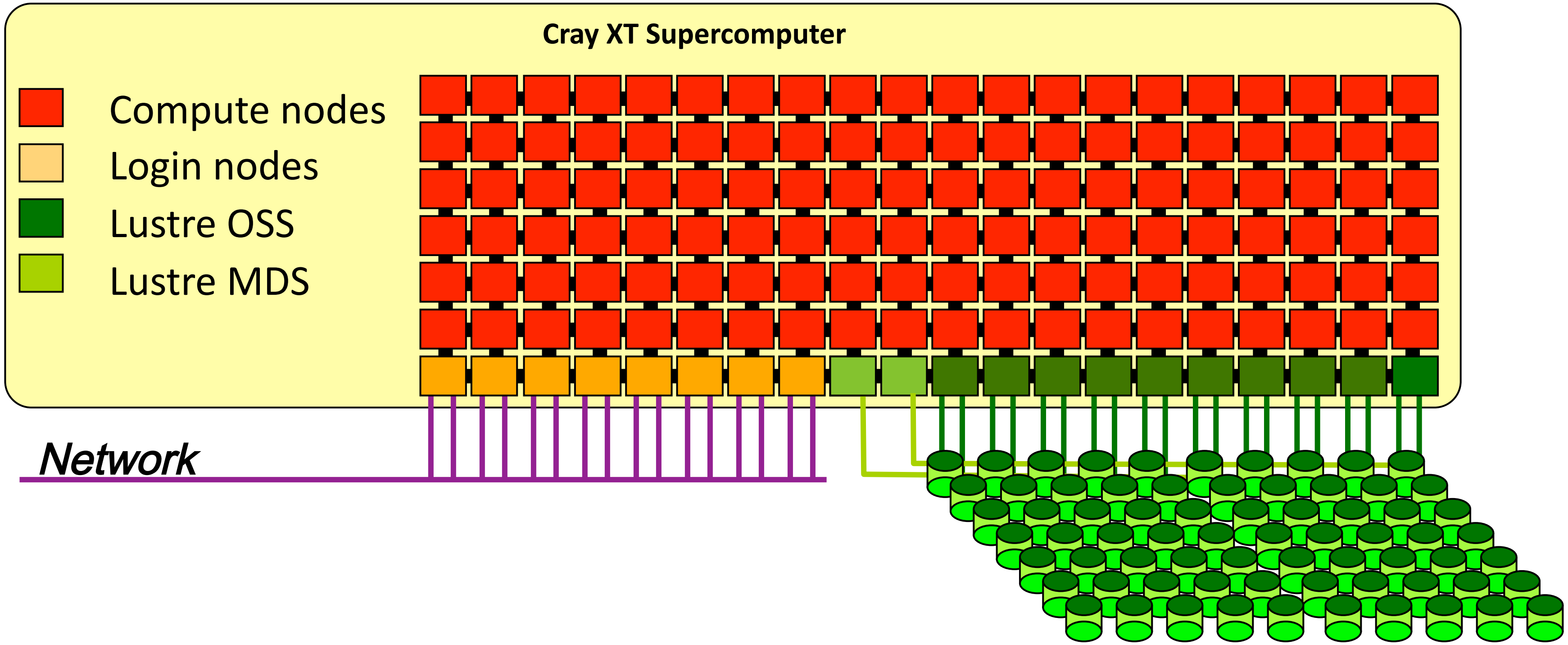
CSCS
Swiss National Supercomputing Centre

Cray XT I/O architecture

- All I/O is offloaded to service nodes
- Lustre
    - High performance parallel I/O file system
    - Direct data transfer between compute nodes and files
- DVS
    - Virtualization service
    - Allows compute nodes to access NFS mounted on service node
- No local disks
- /tmp is a MEMORY file system, on each login node



The Storage Environment



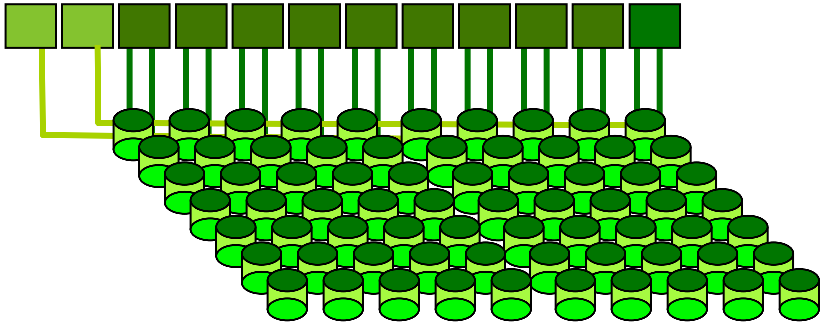
Lustre

high performance

parallel filesystem

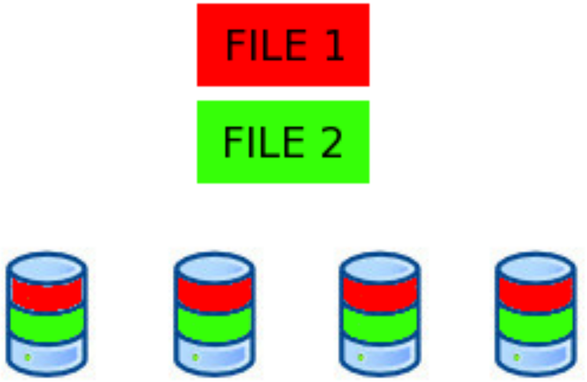
lustre

- A scalable cluster file system for Linux
  - Developed by Cluster File Systems, Inc.
  - Name derives from “Linux Cluster”
  - The Lustre file system consists of software subsystems, storage, and an associated network
- **MDS** – metadata server
  - Handles information about files and directories (**MDT**)
- **OSS** – Object Storage Server
  - The hardware entity
  - The server node
  - Support multiple OSTs
- **OST** – Object Storage Target
  - The software entity
  - This is the software interface to the backend volume



Lustre File Striping

- A Stripe defines the number of OSTs to write the file across
  - Can be set on a per file or directory basis



- CRAY recommends that the default be set to
  - not striping across all OSTs, but
  - set default stripe count of one to four

- But not always the best for application performance.
  As a general rule of thumb :
  - If you have one large file: stripe over all OSTs
  - If you have a large number of files (~2 times #OSTs): turn off striping

Rosa’s Lustre Configuration

- 1 MetaData Server
- 20 Object Storage Servers
- 80 Object Storage Targets
- This filesystem is called **/scratch/rosa**

- No failover capability in this version (1.6)
- With the coming upgrade, failover will be available

Each OST is capable of writing up to 200MB/s

This gives us the ability to write an aggregate speed of

16 GB/s !!!

Lustre lfs command

- lfs is a lustre utility that can be used to create a file with a specific striping pattern, displays file striping patterns, and find file locations
- The most used options are :
    - lfs setstripe
    - lfs getstripe
    - lfs df
- For help execute lfs without any arguments
- `$ lfs`
- `lfs> help`
- `Available commands are:`

```
setstripe
        find
getstripe
        check
```

- ....

```
jtemple@rosa1:Fri Jun 11-11:06:~ > lfs df -h
UUID                  bytes    Used Available  Use% Mounted on
scratch-MDT0000_UUID     1.7T     5.6G     1.6T    0% /scratch/rosa[MDT:0]
scratch-OST0000_UUID     3.6T     1.8T     1.6T   49% /scratch/rosa[OST:0]
scratch-OST0001_UUID     3.6T     1.9T     1.5T   51% /scratch/rosa[OST:1]
scratch-OST0002_UUID     3.6T     1.8T     1.6T   49% /scratch/rosa[OST:2]
…
scratch-OST004d_UUID     3.6T     1.8T     1.6T   51% /scratch/rosa[OST:77]
scratch-OST004e_UUID     3.6T     1.9T     1.5T   51% /scratch/rosa[OST:78]
scratch-OST004f_UUID     3.6T     1.8T     1.6T   50% /scratch/rosa[OST:79]

filesystem summary:    286.2T   146.2T   125.5T   51% /scratch/rosa
```

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

CSCS
Swiss National Supercomputing Centre

lfs setstripe

- Sets the stripe for a file or a directory
- lfssetstripe<file|dir><-s size><-i start><-c count>
  - stripe size: Number of bytes on each OST (0 filesystem default)
  - stripe start: OST index of first stripe (-1 filesystem default)
  - stripe count: Number of OSTs to stripe over (0 default, -1 all)
- Comments
  - The stripe of a file is given when the file is created. It is not possible to change it afterwards.
  - If needed, use lfs to create an empty file with the stripes you want (like the touch command)
- Rosa /scratch configuration:
  - 80 OSTs
  - Default count: 4
  - Default size: 1MByte

```
jtemple@rosa1:Fri Jun 11-11:12:/scratch/rosa/jtemple > lfs getstripe .
…
./stripe_all
stripe_count: -1 stripe_size: 0 stripe_offset: -1
./stripe_one
stripe_count: 1 stripe_size: 0 stripe_offset: -1
./stripe_default
(Default) stripe_count: 4 stripe_size: 1048576 stripe_offset: 0
```

Lustre striping hints

- For maximum aggregate performance: Keep all OSTs occupied

- Many clients, many files: Don’t stripe
  - If number of clients and/or number of files >> number of OSTs:
  - Better to put each object (file) on only a single OST.
- Many clients, one file: Do stripe
  - When multiple processes are all accessing one large file:
  - Better to stripe that single file over all of the available OSTs.
- Some clients, few large files: Do stripe
  - When a few processes access large files in large chunks:
  - Stripe over enough OSTs to keep the OSTs busy on both write and read paths.

THANKS!