FENIX

RESEARCH INFRASTRUCTURE

Introduction to the ICEI resources at CEA

7th Fenix Research Infrastructure Webinar

Thomas Leibovici (CEA) - September 22, 2020

The ICEI project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 800858.

Agenda

- ICEI/FENIX resources types and implementation at CEA/TGCC
- Interactive computing and storage systems
  - Systems description
  - Access and use
- Virtual machine services
  - System description and use
- Getting help
  - Documentation and services to TGCC users

Overview of FENIX services available at CEA/TGCC

- **Interactive Computing** Service (IAC)
  - Compute nodes equipped GPUs, large to extra large memory
  - Provides quick access to compute servers to analyse and visualise data interactively. Also usable for HPC simulations.
- **Virtual Machine** (VM) Service
  - OpenStack Cluster to manage and run virtual machines accessible from the Internet
  - Service for deploying VMs in a stable and controlled environment, e.g. platform services like collaboratory, websites, databases…
- **Active Data Repositories** (ACD)
  - Lustre parallel filesystems *work* and *flash*
  - High performance site-local data repositories for working on large data sets
- **Archival Data Repositories** (ARD)
  - *Store*: Hierarchical storage system with Lustre as top-level (POSIX interface)
  - *OpenIO* object store, accessible through an OpenStack Swift interface
  - Federated data stores for long-term storage and sharing of large data sets

Interactive computing and storage systems

Systems description

Interactive computing cluster (IAC) : hardware

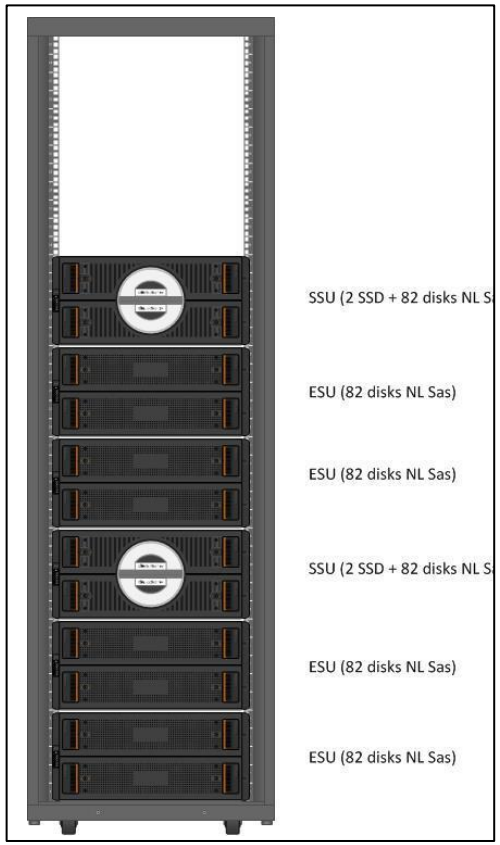
- 30 compute nodes:
  - 2 CPUs Intel Cascade Lake G-6240 (each 18 cores @ 2.6 GHz)
  - 384 GB of RAM
  - 1 GPU NVidia V100 with 32GB of memory
- 2 compute nodes with extra large memory:
  - 4 CPUs Intel Cascade Lake G-6240 (each 18 cores @ 2.6 GHz)
  - 3,072 GB of RAM
  - 1 GPU NVidia V100 with 32GB of memory
- Interconnection network: InfiniBand HDR (100-200Gb)
  - Low-latency & high bandwidth network for MPI applications
- Total I/O throughput: 150 GBytes/sec (1200 Gbits)

Work filesystem (ACD) configuration

- ClusterStor L300 appliance
- Lustre 2.12 Filesystem
- 492 Hard Drive Disks of 10TeraBytes (SAS)

- Total capacity: **3.5 PetaBytes**
- Total throughput: **70 GigaBytes/sec**

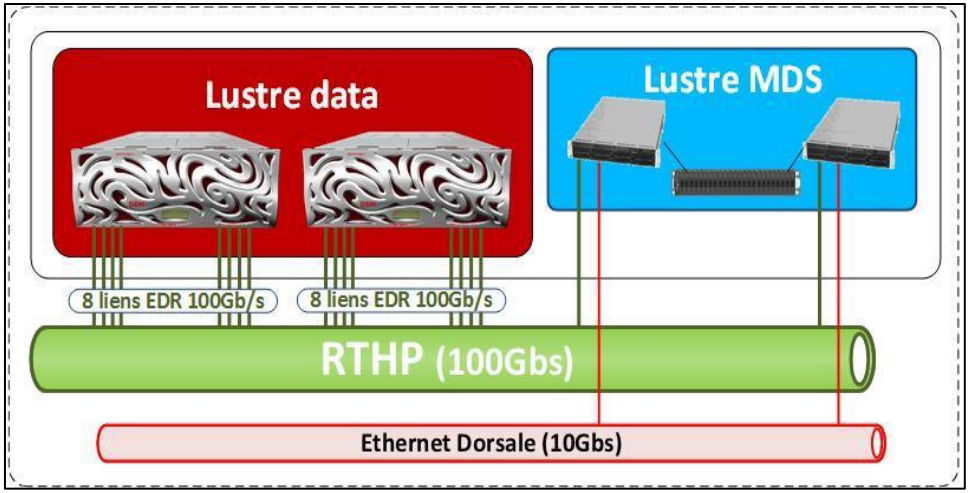
- HDD only, suitable for sequential IO workloads



FENIX RI

Flash filesystem (ACD) configuration

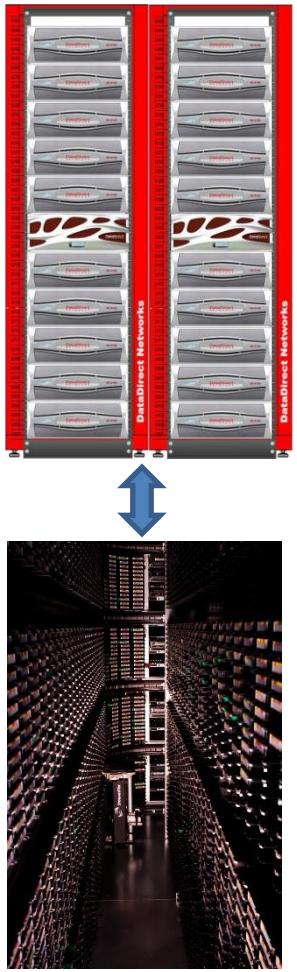
- Full-SSD filesystem for intensive IO patterns (data analysis, AI…)
  - 2 DDN SFA 18KXe controllers with embedded IO servers
  - 84 SSD drives of 15.38 TeraBytes (SAS 12Gbits/s)
  - Metadata: 2 MDS. Capacity: up to 1 billion files.
  - Lustre 2.12 filesystem
- Total capacity: 970 TeraBytes
- Total throughput: 120 GigaBytes/sec
- I/O operations per seconds (random 4k): 1,400,000 read, 232,000 write

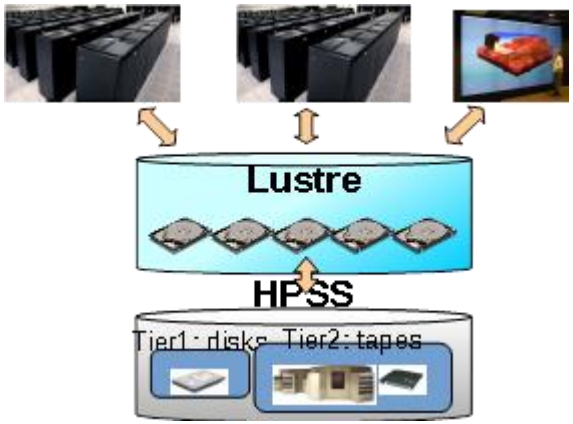


FENIX RI

Store filesystem (ARD) configuration

- Archival purpose: long-term storage of datasets
- Hierarchical storage system:
  - Top-tier: Lustre filesystem (4.8 PetaBytes @ 70GB/s)
  - Bottom-tier: Tape storage managed by HPSS (extendable)
  - Transparent migration between disks and tapes
- Accessible like a common filesystem
  - POSIX interface
  - Automatic reload from tape at first I/O
- Store should also be accessible through Swift as the outcome of the ICEI R&D

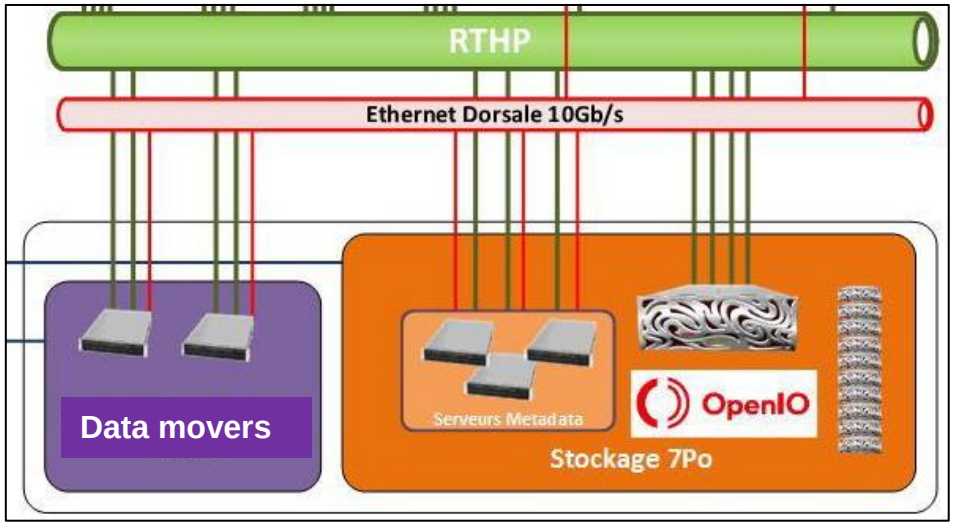




FENIX RI

OpenIO object store (ARD) configuration

- Data storage:
  - 1 DDN SFA18KXe controller + 10 drawers SS9012 (up to 90 disks per drawer)
  - 650 hard drive disks of 14TB
  - Capacity: 7 PetaBytes
- Total bandwidth: 15 GBytes/s
- Metadata management :
  - 3 servers, each with 4 SSD of 1.9TB
  - Capacity: up to 2.5 billion objects
- Software: *OpenIO*
- Swift interface
  - and possibly S3



FENIX RI

Interactive computing and storage systems

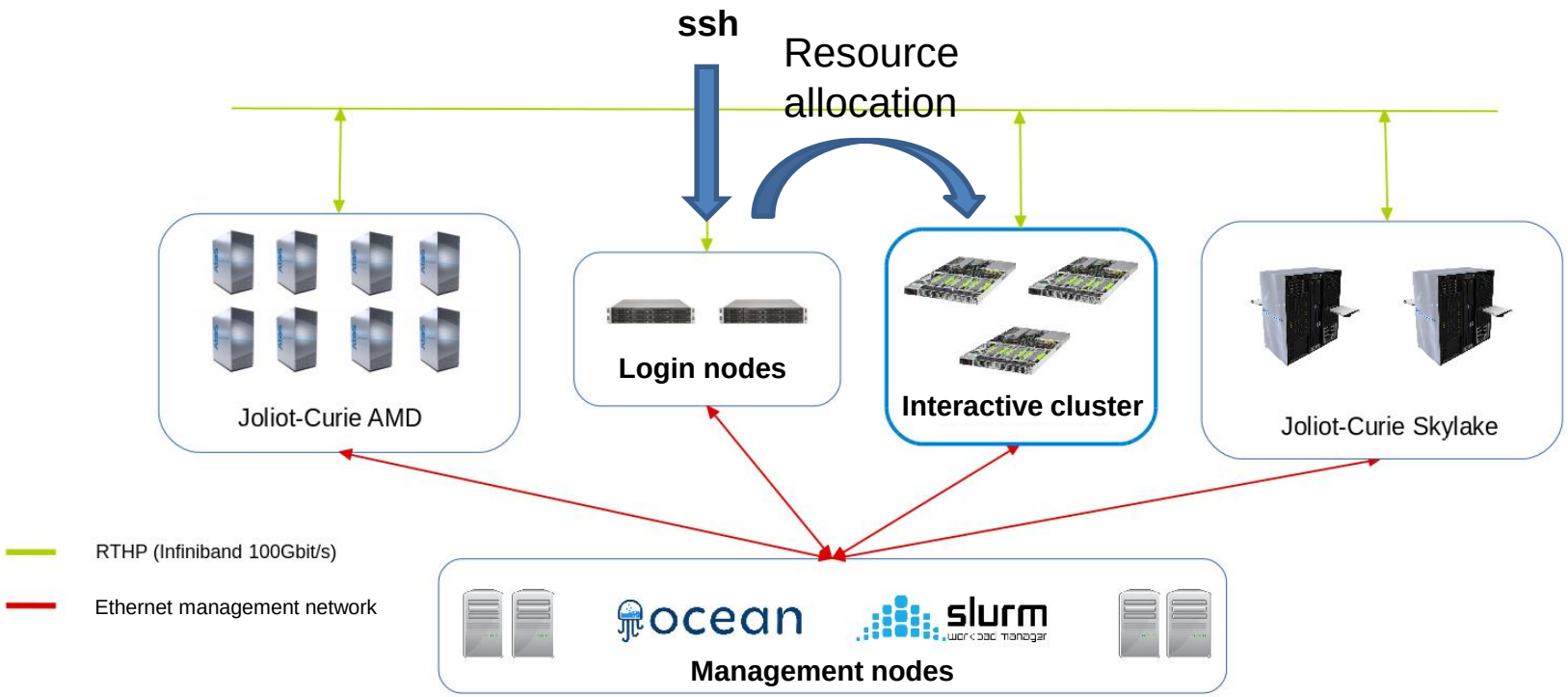
Access and use

Getting access to the systems

1. Information to request an access: https://fenix-ri.eu/access
   - Further details at the end of this webinar
2. Peer review and technical assessment
3. Notification that your resource request is accepted
4. The operations team of TGCC contacts you and provides you with instructions to access the systems

Interactive computing cluster: site integration

- Interactive nodes are installed within the Joliot-Curie supercomputer



Accessing the interactive computing cluster

- SSH access points :
  - **fenix-iac.ccc.cea.fr**
  - main DNS entry, load-balanced across login nodes
  - Connection stickiness: a same user always arrive on the same login node when multiple connections are opened
  - Alternate DNS entries to target particular login nodes:
    - **fenix-iac-log1.ccc.cea.fr**
    - **fenix-iac-log2.ccc.cea.fr**
  - Can be used for file transfers with sftp/scp/rsync
- SSH credentials
  - login/password only, no public key authentication
- Login nodes :
  - 2x nodes with 2x AMD EPYC 7502, 256GB DDR4
  - to be used for job submissions and data transfers only

Security of incoming/outgoing connections

- Registering a host to connect to the computing center
  - The source host must comply with the following security rules:
    - located in a **European country**
    - **public DNS** record associated to a **public FQDN**
    - host security guaranteed by the **security officer** of your organization
  - Information validated by the project leader and transmitted to the **TGCC hotline**.
- Accessing an external service from the computing center
  - Same conditions as above.
  - Only **encrypted** flows to **authenticated** services are eligible.

Software environment: modules

- Extensive collection of HPC software provided by TGCC
  - Compilers, libraries, tools, applications ...
  - On top of the base OS: Atos SCS5 (Red Hat Enterprise Linux 7 derivative)
- Accessible through environment modules
  - List available software products:

```
module avail
```

  - Make a product available in your environment:

```
module load <product>[/version]
```

  - Display currently loaded modules:

```
module list
```

  - And much more (see documentation).
- Modules handle software dependencies and conflicts
  - Dependencies are automatically loaded

Software environment: toolchains

- A toolchain is a common set of tools and libraries used to build products
- The current default toolchain is composed of:
  - Intel 19 compiler (soon updated to Intel 20)
  - OpenMPI 4 library
- Many products are provided for multiple toolchains

```
$ module help openfoam
[…]
    Software configuration(s):
    0 : module load flavor/buildcompiler/intel/19 flavor/buildmpi/openmpi/4.0
    1 : module load flavor/buildcompiler/intel/20 flavor/buildmpi/openmpi/4.0
```

- Select a toolchain by loading the corresponding modules first:

```
$ module load flavor/buildcompiler/intel/20 flavor/buildmpi/openmpi/4.0
$ module load openfoam
```

Software environment: compilers

- The available compilers on the cluster are:
  - Intel Compiler suite (icc, icpc, ifort)
    - Default version: 19.0.5.281
    - Latest version: 20.0.2
  - GNU compiler suite (gcc, g++, gfortran)
    - Default version: 7.3.0
    - Latest version: 10.1.0
  - PGI compiler suite (pgcc, pgCC, pgf90)
    - Default version: 18.7
    - Latest version: 20.4
- Compilers can be selected using modules, like other products
  - The corresponding toolchain is automatically selected
  - The Intel compiler suite is recommended
  - Default versions will soon change: Intel 20, PGI 20
  - We recommend to already use these future defaults for new projects!

Computing on the Interactive cluster: resources

- Do not run computations on login nodes
  - Your tasks will be automatically throttled and/or terminated
- Allocate resources on the interactive cluster using the batch manager
  - **SLURM**: common scheduler for all nodes of the Joliot-Curie supercomputer
  - Allows both submitting batch jobs and running interactive tasks
  - Compute nodes are split in two partitions:
    - **v100l**: dual-socket interactive nodes (1 GPU, 384GB RAM)
    - **v100xl**: quad-socket interactive nodes (1 GPU, 3TB RAM)
- SLURM serves resource allocation requests by order of priority
  - Priority is computed by comparing:
    - The current rate of resource consumption (over a few weeks)
    - The target rate for consuming a project's resources steadily over its lifetime
  - Under-consumers get a higher priority than over-consumers
  - Use your project's resources regularly!
- Backfilling is enabled to fill scheduling gaps with lower priority jobs

Computing on the Interactive cluster: billing

- Default consumption: (*core count*) x (*elapsed time*)
  - Each core is dedicated to a single job
  - Default allocation time: 2 hours
  - Use of GPUs: require to allocate full nodes
    - Work in progress to enable sharing GPUs
- Oversubscribed mode
  - Up to 4 jobs and/or users per core
  - Default memory: 1/4 of available memory per core
    - Users can request more if needed
    - No impact on CPU oversubscription ratio
  - Billed time based on the memory allocation ratio
  - Under evaluation, may evolve depending on actual use

Interactive cluster: jobs management

- Using ccc_* commands (AKA Bridge) is recommended for managing resources and jobs
  - Abstraction layer for batch systems and resource managers
  - Facilitates running tasks in TGCC's environment
  - Easy to transpose if you already know SLURM
- Example commands :
  - Listing available partitions: `ccc_mpinfo`
  - Listing available QoS: `ccc_mqinfo`
  - Listing jobs: `ccc_mpp`
  - Getting information on a job: `ccc_macct`
  - Submitting a batch job: `ccc_msub`
  - Running tasks within allocated resources: `ccc_mprun`

Computing on the Interactive cluster: examples

- Example commands for interactive use cases :

- Start a shell on a dedicated quad-socket node:

```
ccc_mprun -p v100xl -N 1 -x -s
```

- Start a shell on 8 cores of a dual-socket node allowing oversubscription:

```
ccc_mprun -p v100l-os -c 8 -s
```

- Start a shell on 4 dedicated cores with X11 forwarding for ten hours

```
ccc_mprun -p v100l -T 36000 -c 4 -X first -s
```

- A GPU-accelerated remote visualization system will soon be provided
  - Likely Nice DCV
  - No specific client app or plugin required

File systems

- Use of file systems:
  - *SCRATCH:* work space for temporary data (purged 60 days after last access)
  - *WORK*: permanent work data (no purge, quotas, no backup)
    - Legacy work (HDD) or flash (SSD), depending on your project request
  - *STORE:* for archival data (high capacity, high latency: tape backend)
    - Recommended file size: 10GB to 1TB

- To specify the project space to work in:
  - `module switch dfldatadir/project_name`
  - Note: this also applies to Swift

FENIX RI

File systems: access paths

- After datadir's module switch or load, filesystems paths are available through **environment variables**
- For **personal** folders :
  - `$CCCSCRATCHDIR` for SCRATCH filesystem
  - `$CCCWORKDIR` for WORK or FLASH filesystem
    (depending on your resource request)
  - `$CCCSTOREDIR` for STORE filesystem
- For team **shared** folders :
  - `$ALL_CCCSCRATCHDIR` for SCRATCH filesystem
  - `$ALL_CCCWORKDIR` for WORK or FLASH filesystem
    (depending on your resource request)
  - `$ALL_CCCSTOREDIR` for STORE filesystem

Use of Swift object storage

- Access from the computing centre:

```
# specify project storage space
module switch dfldatadir/project_name

# load swift environment
module load swift

# swift commands: no specific argument required
swift list
swift upload container object
swift download container object
```

- Swift client libraries available for most languages
  - e.g. python-swiftclient
- Access from the Internet available in November 2020

FENIX RI

Useful computing center commands (1/2)

- ccc_quota: user and projects disk usage

```
$ ccc_quota
Disk quotas for user loginname (uid 24256):

             ============ SPACE =============  ============ INODE =============
 Filesystem     usage    soft    hard   grace   entries    soft    hard   grace
 ----------  -------- ------- ------- -------  -------- ------- ------- -------
       HOME      1.9G      5G      5G       -       304       -       -       -
      STORE        4k       -       -       -         1       -       -       -
       WORK   140.71G       -       -       -       222       -       -       -
    SCRATCH        8k       -       -       -         2       -       -       -

Disk quotas for data space groupname (gid 5893):

             ============ SPACE =============  ============ INODE =============
 Filesystem     usage    soft    hard   grace   entries    soft    hard   grace
 ----------  -------- ------- ------- -------  -------- ------- ------- -------
       HOME         -     20G     20G       -         -       -       -       -
      STORE     4.75T       -       -       -     3.09k    100k  100.1k       -
       WORK    11.45T     20T     20T       -     2.97M      5M      5M       -
    SCRATCH     56.8G    100T  100.1T       -    71.38k      2M   2.05M       -
```

Useful computing center commands (2/2)

- ccc_myproject: resource consumption reporting

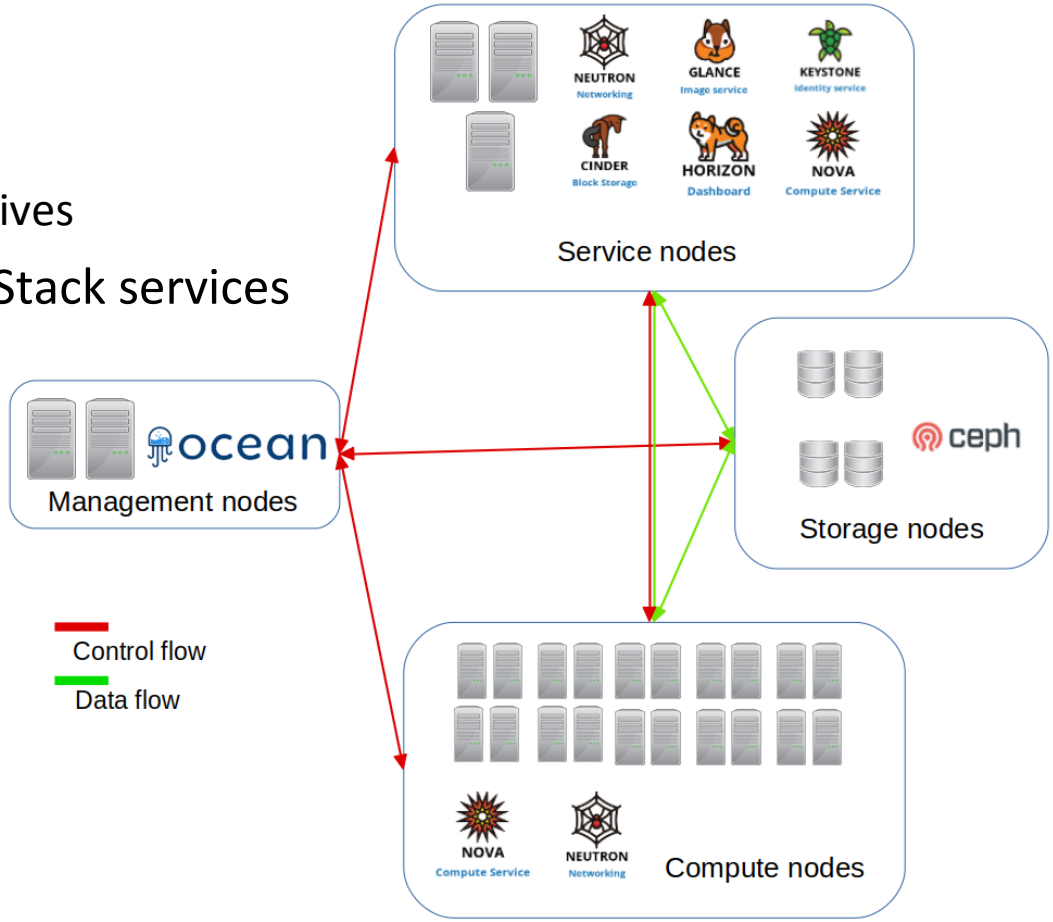
```
$ ccc_myproject

Accounting for project fnxh1234 on Irene v100l at 2020-09-14
Login                                Time in hours
loginnm1                                      0.00
loginnm2                                   7460.16
loginnm3                                5579201.07
loginnm4                                 426395.62
loginnm5                                1550152.76
loginnm6                               25857581.50
loginnm7                                 874023.20
loginnm8                                 419489.07
loginnm9                                1169639.57

Total                                  35883942.95
Allocated                              45000000.00
Suggested use at this time                  95.63%
Real use at this time                       79.74%

Project deadline 2020-10-01
```

Virtual machine services

System description and use

OpenStack cluster (VM service): hardware

- 20 hypervisors to run user VMs
  - 2 CPUs Intel Cascade Lake G-6240 (each 18 cores @ 2.6 GHz)
  - Memory: 192 GB of RAM
- 4 storage servers
  - 60.8 TB (raw) of SSDs
  - Ceph is used to store VM drives
- 3 service nodes to host OpenStack services
- 3 management nodes
- 10Gbits/s Network



OpenStack cluster (VM service): software

- Latest OpenStack release as of Sept. 2020: *Ussuri*
- Available VM profiles (tentative list):

Flavor	VCPUs	RAM	Disk
gpp.s	2	4GB	40GB
gpp.m	4	8GB	40GB
gpp.l	16	32GB	80GB
gpp.xl	32	64GB	80GB

  - All VM drives are backed by SSD storage and automatically encrypted
  - Each vCPU corresponds to a hyper-thread
  - No CPU oversubscription, fully dedicated cores
  - Optimized use of NUMA topologies and huge pages
- The VM service will be available starting from Nov. 2020

FENIX RI

Accessing other resources from VMs

- Access to other TGCC resources from VMs:
  - Access to **compute resources** through **ssh**
    - `ssh login@fenix-iac.ccc.cea.fr ccc_msub …`
  - Access to TGCC **filesystems** (work, flash, store) using **sshfs**
    - `sshfs login@fenix-iac.ccc.cea.fr:/ccc/work/... local_work_dir`
  - Access to **archival storage** using **Swift** protocol
    - `swift download/upload container object`
- Request a **service account** for these automated accesses to TGCC resources
  - Restricted access to a limited set of resources (e.g. compute only, specific storage spaces, read-only access…)
  - Limit impacts in case of security breach on a VM

Services to TGCC users

Services to TGCC users

- User documentation:
  - On the TGCC web site: https://www-eu.ccc.cea.fr
    (login/password required)
  - On command line: `machine.info`

- Hotline: multi-level support for users
  - From questions, user account management, to advanced applicative user support
  - Operating from 8:30 to 17:30 (CEST)
  - hotline.tgcc@cea.fr (+33 1 77 57 42 42)

Questions?

© CEA/DAM