

# Curnagl

## Kesako?

Curnagl (Romanche), or Chocard à bec jaune in French, is a sociable bird known for its acrobatic exploits and is found throughout the alpine region. More information is available at

<https://www.vogelwarte.ch/fr/oiseaux/les-oiseaux-de-suisse/chocard-a-bec-jaune>

It's also the name of the HPC cluster managed by the DCSR for the UNIL research community.

A concise description if you need to describe the cluster is:

*Curnagl is a 96 node HPC cluster based on AMD Zen2/3 CPUs providing a total of 4608 compute cores and 54TB of memory. 8 machines are equipped with 2 A100 GPUs and all nodes have 100Gb/s HDR Infiniband and 100Gb/s Ethernet network connections in a fat-tree topology. The principal storage is a 2PB disk backed filesystem and a 150TB SSD based scratch system. Additionally all nodes have 1.6 TB local NVMe drives.*

If you experience unexpected behaviour or need assistance please contact us via [helpdesk@unil.ch](mailto:helpdesk@unil.ch) starting the mail subject with DCSR Curnagl

## How to connect

The login node is **curnagl.dcsr.unil.ch**

For full details on how to connect using SSH please read [the documentation](#)

Before connecting we recommend that you add the host's key to your list of known hosts:

```
echo "curnagl.dcsr.unil.ch ecdsa-sha2-nistp256
AAAAE2VjZHNhLXNoYTItbmlzdHAyNTYAAAAIbmlzdHAyNTYAAABBBcUnvgFAN/X/8b1FEIxy8p3u9jgff
0NgCI7CX4ZmqIhaYis2p7AQ34foIXemaw2wT+Pq1V9dCUh18mWXnDsJGrg=" >>
~/.ssh/known_hosts
```

You can also type "yes" during the first connection to accept the host key but this is less secure.

Please be aware that you must be connected to the VPN if you are not on the campus network.

Then simply `ssh username@curnagl.dcsr.unil.ch` where username is your UNIL account

The login node must not be used for any form of compute or memory intensive task apart from software compilation and data transfer. Any such tasks will be killed without warning.

# Hardware

## Compute

The cluster is composed of 96 compute nodes of which eight have GPUs.

Number of nodes	Memory	CPU	GPU
52	512 GB	2 x AMD Epyc2 7402	-
12	1024 GB	2 x AMD Epyc2 7402	-
8	512 GB	2 x AMD Epyc2 7402	2 x NVIDIA A100
24	512 GB	2 x AMD Epyc3 7443	

## Network

The nodes are connected with both HDR Infiniband and 100 Gb Ethernet. The Infiniband is the primary interconnect for storage and inter-node communication.

# Partitions

There are 3 main partitions on the cluster:

## interactive

The interactive partition allows rapid access to resources but comes with a number of restrictions, the main ones being:

- Only one job per user at a time
- Maximum run time of 8 hours but this decreases if you ask for lots of resources.

For example:

CPU cores requested	Memory requested	GPUs requested	Run Time Allowed
4	32	1	8 hours
8	64	1	4 hours
16	128	1	2 hours
32	256	1	1 hour

We recommend that users access this using the `sinteractive` command. This partition should also be used for compiling codes.

This partition can also be accessed using the following sbatch directive:

```
#SBATCH -p interactive
```

### Note on GPUs in the interactive partition

There is one node with GPUs in the interactive partition and in order to allow multiple users to work at the same time these A100 cards have been partitioned into 2 instances each with 20GB of memory for a total of 4 GPUs.

The maximum time limit for requesting a GPU is 8 hours with the CPU and memory limits applying.

For longer jobs and to have whole A100 GPUs please submit batch jobs to the gpu partition.

Please do not block resources if you are not using them as this prevents other people from working.

If you request too many resources then you will see the following error:

```
salloc: error: QOSMaxCpuMinutesPerJobLimit
```

```
salloc: error: Job submit/allocate failed: Job violates accounting/QOS policy (job submit limit, user's size and/or time limits)
```

Please reduce either the time or the cpu / memory / gpu requested.

## cpu

This is the main partition and includes the majority of the compute nodes. Interactive jobs are not permitted. The partition is configured to prevent long running jobs from using all available resources and to allow multi-node jobs to start within a reasonable delay.

The limits are:

Normal jobs - 3 days

Short jobs - 12 hours

Normal jobs are restricted to  $\sim 2/3$  of the resources which prevents the cluster being blocked by long running jobs.

In exceptional cases wall time extensions may be granted but for this you need to contact us with a justification before submitting your jobs!

The cpu partition is the default partition so there is no need to specify it but if you wish to do so then use the following sbatch directive

```
#SBATCH -p cpu
```

## gpu

This contains the GPU equipped nodes.

To request resources in the gpu partition please use the following sbatch directive:

```
#SBATCH -p gpu
```

The limits are:

Normal jobs - 3 days

Short jobs - 12 hours

Normal jobs are restricted to  $\sim 2/3$  of the resources which prevents the cluster being blocked by long running jobs.

```
--gres=gpu:N
```

where N is 1 or 2.

## Software

For information on the DCSR software stack see the following link:

<https://wiki.unil.ch/ci/books/high-performance-computing-hpc/page/dcsr-software-stack>

# Storage

The storage is provided by a Lenovo DSS system and the Spectrum Scale (GPFS) parallel filesystem.

## /users

Your home space is at /users/username and there is a per user quota of 50 GB and 100,000 files.

We would like to remind you that all scripts and code should be stored in a Git repository.

## /scratch

The scratch filesystem is the primary working space for running calculations.

The scratch space runs on SSD storage and **has an automatic cleaning policy** so in case of a shortage of free space files older than 2 weeks (starting with the oldest first) will be deleted.

Initially this cleanup will be triggered if the space is more than 90% used and this limit will be reviewed as we gain experience with the usage patterns.

The space is per user and there are no quotas (\*). Your scratch space can be found at /scratch/username

e.g. `/scratch/ulambda`

Use of this space is not charged for as it is now classed as temporary storage.

*\* There is a quota of 50% of the total space per user to prevent runaway jobs wreaking havoc*

## /work

The work space is for storing data that is being actively worked on as part of a research project. Projects have quotas assigned and while we will not delete data in this space there is no backup so all critical data must also be kept on the DCSR NAS.

The structure is:

`/ work / FAC / FACULTY / INSTITUTE / PI / PROJECT`

This space can, and should, be used for the installation of any research group specific software tools including python virtual environments.

---

Révision #16

Créé 20 mai 2021 09:15:38 par Ewan Roche

Mis à jour 6 avril 2023 08:39:11 par Ewan Roche