FAQ

General FAQ

Connecting to the Clusters

Why can't I connect to the clusters from home?

You can but to do so requires passing via the EPFL VPN service. See http://network.epfl.ch/vpn for how to use this service.

Users preferring a command line tool might also wish to consider the tremplin SSH proxy tunnel service:. You can find the Linux and Windows procedure here.

Why am I asked for a password while sshing from the frontend to a node?

Once logged in to a frontend of a cluster, you can ssh directly to the node(s) running your job(s). You can prevent to be asked for the Gaspar password again by creating a passwordless ssh key.

Run the following command **only once** in any of the clusters:

```
ssh-keygen -b 4096 -t rsa
ssh-copy-id 127.0.0.1
```

top

Batch System Questions

What's the maximum run time of a job?

If you have a free account it's 6 hours. For paid accounts it's 3 days but you can ask to run for longer by contacting us and explaining why you need to run for more than 3 days.

How do I submit a job that requires a run time of more than three days?

Labs with signed contracts can use the "--qos=week" flag to ask for up to 7 days.

Why is my job using the `--qos=week` flag pending with reason `QOSGrpNodeLimit` ?

There is a global limit on how many nodes can be running very long jobs (week+). This is to ensure the scheduler can reassign resources to groups with higher priority in a reasonable time frame.

Can I submit array jobs and, if so, how?

Yes, with the `--array` directive to sbatch. See http://slurm.schedmd.com/job_array.html for the official documentation and our scitas-examples git for several examples.

Can I have a paid and free account at the same time?

The short answer is no. Once a user is part of a group that pays for access (*Premium* or *Commitment*) then they are removed from the *Free* group.

What is the difference between hpc-lab and lab ?

hpc-lab is the name of the group to manage user access to the cluster (in the groups.epfl.ch sense). *lab* is the name of the Slurm account, automatically populated with users from the *hpc-lab* group. You have to use the account name *lab* in your batch scripts.

Is it safe to share nodes with other users?

Yes! We use *cgroups* to limit the amount of CPU and memory assigned to users. There is no way for users to adversely affect each other.

Is there a debug queue?

Not as such. In SLURM the concept of queues doesn't exist so to have priority access for debugging there is a partition which gives access to dedicated nodes:

```
sbatch --partition debug myjobscript
```

The limits on the debug partition vary by cluster but in general the maximum run time is one hour (default is 15 minutes) and users are only allowed one job at a time. 2 nodes max. Interactive jobs are allowed.

I have a premium account and I have run on the debug partition. Do I have to pay for debug time?

No. Debug time is free of charge.

What is a <job id>?

It's the unique numerical identifier of a job and is given when you submit the job:

```
[user@cluster jobs]$ sbatch s1.job
Submitted batch job 1234567
```

It can also be seen using squeue:

```
[user@cluster jobs]$ squeue
JOBID   PARTITION NAME   USER ST TIME    NODES NODELIST(REASON)
1234567 serial    s1.job user R  INVALID 1     c03
```

How do I display the used CPU time for my account since a certain point in time?

You can use the *sreport* tool. Here is an example of query where the used time is reported in core hours. Just replace *2018-01-01T00:00:00* with the start time you wish and *scitas-ge* with your account name.

```
$ sreport cluster AccountUtilizationByUser -t Hour --parsable2 start=2018-01-01T00:00:00  Accounts=scitas-
ge  Format=Cluster,Account,Login,Used
--------------------------------------------------------------------------------
Cluster/Account/User Utilization 2018-01-01T00:00:00 - 2018-04-29T23:59:59 (10278000 secs)
Use reported in TRES Hours
--------------------------------------------------------------------------------
Cluster|Account|Login|Used
fidis|scitas-ge||156349
fidis|scitas-ge|user1|7
fidis|scitas-ge|user2|22834
fidis|scitas-ge|user3|0
```

How do I specify that my multi-node MPI job should use nodes connected to the same switch?

You can specify the maximum number of switches to be used as follows (in this case one switch)

```
#SBATCH --switches=1
```

Please note that jobs with such requirements may take much longer to schedule than those than can be spread across the cluster. This option should only be used in very specific cases!

Is any form of simultaneous multithreading (SMT) (such as Intel's 'Hyper-Threading' or 'HT') enabled on the clusters?

In general SMT can decrease performance if there are any shared resources in the CPU. Considering parallel codes typically all perform similar operations any such shared resources would quickly become a bottleneck. As such SMT/HT is as a general rule disabled in all SCITAS clusters.

Why does my job fail immediately without leaving any trace (output)?

This usually happens when one specifies a non-existing working directory (for example by using: `--chdir /path/that/does/not/exist`).

Why does my job fail after submission with error "Invalid generic resource (gres) specification"?

Because on Izar it's necessary to specify the --gres=gpu: X flag. Where X is the number of GPUs per node required for sbatch.

How do I set up job notification emails?

Add **both** following commands to your submission script to set the email address:

```
#SBATCH --mail-user=$FIRST_NAME.$LAST_NAME@epfl.ch
#SBATCH --mail-type=$NOTIFICATION_TYPE
```

1. A valid email address, preferably one provided by EPFL.
2. A type of notification. Valid type values are NONE, BEGIN, END, FAIL, REQUEUE, ALL (equivalent to BEGIN, END, FAIL, REQUEUE, and STAGE_OUT), STAGE_OUT (burst buffer stage out and teardown completed), TIME_LIMIT. Multiple *type* values may be specified in a comma-separated list.

Why does my job fail after requeuing with the error "Requested operation is presently disabled for job JOBID"?

The requeueing possibility must be explicitly requested by the user by adding the option --requeue to the batch script:

```
#SBATCH --requeue
```

Later, a job executed with this option can use the "scontrol requeue JobID" command to be dispatched again.

ⓘ You can also check the official documentation for more at https://slurm.schedmd.com/sbatch.html

top

File System Questions

Where is my scratch space?

`/scratch/<user name>` - e.g. `/scratch/jmenu`

Can you recover an important file that was on my scratch area?

NO. /scratch is not backed up so the file is gone forever. Please note that we automatically delete files on scratch to prevent it from filling up!

I've deleted a file on /home or /work - How can I recover it?

If it was deleted in the last seven days then you can use the daily snapshots to get it back. These can be found at:

- `/home/.snapshots/<date>/<username>/`
- `/work/.snapshots/<date>/<laboratory or group>/`

e.g. `/home/.snapshots/2015-11-11/bob/`

The home filesystem is backed up onto tape. If the file was deleted more than a week before we may be able to help. The /work filesystem is not backed up by default.

How to display scratch quota and usage information?

A. There are no quotas on scratch, as files older than 2 weeks may be deleted without notice as the filesystems fills up. However, you can see scratch usage for using the fsu command:

```
fsu /scratch
```

How to display quota and usage information for the /home and /work file systems?

1. **/home**:
   to get user quota and file system usage for your group members, use the following command:

   ```
   fsu -q /home
   ```

   You can also see an overview of /home usage and quota here.

2. **/work**:
   to get group quota and file system usage for your group members, use the following command:

   ```
   fsu -q /work
   ```

You can also see an overview of /work file system usage and quota here.

How can I edit a file in the clusters using an application on my computer?

If you wish to manipulate file on the remote filesystem by using a software that is installed on your workstation
you can mount the remote filesystem by using sshfs. After install it, you can type from a terminal:

```
$sudo mkdir /media/dest
$sudo sshfs -o allow_other USER@MACHINE:/scratch/USER /media/dest
```

where USER is your gaspar account and MACHINE the cluster fs you wish to mount.

top

Software Questions

I want to use an Intel software on my own machine/server. How can I do it?

Please check this page.

Why do I get the error "module: command not found" or "slmodules: command not found"?

This is probably because you have `tcsh` as your login shell and the environment isn't propagated to the compute nodes.

In order to fix the issue please change the first line of your job script as follows:

```
#!/bin/bash -l
```

or

```
#!/bin/tcsh -l
```

The -l option tells `bash`/`tcsh` to launch an interactive shell which correctly sources the files in `/etc/profile.d/`

How can I change my default shell?

Most systems use bash by default and most of our documentation assumes your default shell is bash. You can change your default shell on this page.

Which options should I use to link with the Intel MKL?

Ask the Intel Math Kernel link line advisor

If you use the Intel compilers then you can pass the -mkl flag which will do the hard work for you.

What compilers/MPI combination do you support?

SCITAS supports *Intel compilers and Intel MPI* (full proprietary) or *GCC compilers and MVAPICH2* (full free). Other combinations are not supported (if provided they will be supported on a *best-effort* basis).

Why do some system tools stop working after the `intel` module is loaded?

> ⓘ While this was true for the *Cornalin* (July 2017) release, it no longer applies as of the *Paien* (July 2018) release.

The Intel compiler includes its own versions of many libraries (and those take precedence over the system ones). Sometimes these libraries do not include symbols which are needed by some system tools and these will no longer work. Examples of these could be: `git`, `rsync`, etc.

If a module exists providing the same tool one can just load that module.

If no module exists you will have to `module unload intel` before using the command and `module load intel` afterwards (any modules which depend on intel will simply become *inactive* and will be restored automatically).

Where should I compile code?

The best place to compile code is on the dedicated build nodes - see the build nodes section of the compiling codes documentation for more details.

Why does my COMSOL job fail to get a license?

Occasionally your COMSOL jobs might fail with a message such as:

```
Could not obtain license for COMSOL ...

License error: -5.
 No such product exists.
 No such feature exists.
 Feature:       COMSOL
 License path:  ...
 FlexNet Licensing error:-5,414
```

There are particularly few license tokens for some COMSOL features (5 or 10 tokens are common).

One common issue is caused by having COMSOL running on your laptop/workstation using the tokens for the same features you need for your jobs. Please close any COMSOL sessions you are not actively using and try again.

(Alternatively, if possible for the task you are doing, you can try to user other equivalent software packages like ANSYS.)

top

GPUs

Which cluster can I use to run jobs using GPUs?

At the moment, we have one GPU accelerated cluster:

- Izar

What kind of access can I have?

There are two different kind of access, depending on your account type:

1. Premium users have access to GPUs via --qos=gpu
2. Free **LIMITED RESSOURCES** users have access to GPUs via --qos=gpu_free

> ⚠ Non paying access (–qos=gpu_free) is limited to a maximum run time of 12 hours and only one node.

How do I submit jobs to the GPU nodes?

If you have a paid access then you need to pass the following options:

```
--partition=gpu --qos=gpu --gres=gpu:X
```

Where X is the number of GPUs required **per node**.

If you have been granted a free access then you need to pass the these options:

```
--partition=gpu --qos=gpu_free --gres=gpu:X
```

Where X is the number of GPUs required **per node**..

top

Cluster / Partition specific FAQ

Deneb

How do I use one of the nodes with more than 64GB of memory?

Specify the amount of memory required with "--mem <quantity in MB>" either on the command line or in your job script.

How do I ask to use a specific processor type (Ivy Bridge, 16 cores or Haswell, 24 cores)?

For Ivy Bridge please give the option "`--constraint=E5v2`" and for Haswell "`--constraint=E5v3`". If you do not specify a constraint it may run on either but a multi-node job will never span both architectures.

How do I access the Deneb serial partition?

In order to replace Castor there is now a specific partition on Deneb for serial tasks that allows one to request individual cores.

To use this please pass the "--partition=serial” directive to SLURM:

```
#SBATCH --partition=serial
```

Why did my job get sent to the serial partition?

Jobs that do not require a whole node are automatically routed to the serial partition at submission time so as to prevent resources being wasted. If you really do want your one CPU job to run on the parallel partition then pass the "`--exclusive`" option to sbatch.

top

Fidis

How many nodes are there in Fidis and what are their characteristics ?

There are 336 nodes with 128 GB of memory and 72 nodes with 256 GB of memory.

Each node have two Intel Broadwell processors running at 2.6 GHz. Each processor has 14 cores which makes 28 cores per node. A 800 GB local SSD disk (with 200GB allocated to /tmp) makes local checkpoints very fast.

All 408 nodes are interconnected with a non-blocking FDR infiniband network.

The Gacrux extension consists of 216 nodes, each with 192 GB of memory.

Each node has two Intel Skylake processors (Xeon 6132) each with 14 cores @ 2.6 GHz. The Gacrux nodes are interconnected with EDR Infiniband interconnect.

The nodes are arranged in non-blocking groups of 24 and the *Skylake* architecture offers significantly increased memory bandwidth.

Please note that to make use of the new AVX-512 instructions your codes will need to be recompiled. The centrally provided codes and libraries available through modules have been optimised for the new architecture.

How can I specify a particular type of node (Fidis or Gacrux)?

If you wish to specifically ask for the Gacrux/Skylake nodes then please use the following SLURM directive:

```
#SBATCH --constraint=s6g1
```

If you wish to use only the Fidis/Broadwell nodes then please specify:

```
#SBATCH --constraint=E5v4
```

If you do not specify a constraint then jobs may run on either partition but they will never span different architectures.

Are there debug nodes?

Two of the Gacrux nodes are available through the debug partition along with four Fidis nodes.

Why can't I connect to <random port>@<external server> from Fidis?

Fidis' network topology does not allow for direct connections to servers outside the EPFL network. Connections must always be done through an proxy server. This is already configured on all nodes for the most common use cases (ssh, git, http, https, ftp).

For some cases you need to use specific versions of some tools. For example, to download a file from an FTP server (and if you can't use the Data Transfer Nodes) you need to use a more recent version of `curl` (by doing `module load curl`) and use the `--proxytunnel` option.

If you find an use case which you depend on and does not work out of the box please contact us.

top