Using the general purpose computing clusters at EPFL

scitas.epfl.ch

October 31, 2018
Welcome

What you will learn

- What is a cluster
- What is a scheduler
- How the environment is organised
- How to run simple jobs on our clusters

What you will not learn

- Writing and (compiling) codes
- Parallelising code
What is a cluster?
What is a cluster?
## Our clusters

<table>
<thead>
<tr>
<th>Login Hosts</th>
<th>Cores</th>
<th>RAM</th>
<th>Network</th>
<th>Storage</th>
<th>Arch</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>deneb{1,2}.epfl.ch</strong></td>
<td>16x2.6</td>
<td>64 GB</td>
<td>40 Gbit/s</td>
<td>350 TB</td>
<td>E5v2</td>
</tr>
<tr>
<td></td>
<td>24x2.5</td>
<td></td>
<td></td>
<td></td>
<td>E5v3</td>
</tr>
<tr>
<td></td>
<td>16x2.6</td>
<td>+ 4x NVidia K40</td>
<td></td>
<td></td>
<td>E5v2</td>
</tr>
<tr>
<td></td>
<td>16x2.6</td>
<td></td>
<td></td>
<td></td>
<td>E5v2</td>
</tr>
<tr>
<td></td>
<td>32x2.6</td>
<td></td>
<td></td>
<td></td>
<td>E5v2</td>
</tr>
<tr>
<td><strong>fidis.epfl.ch</strong></td>
<td>28x2.6</td>
<td>128 GB</td>
<td>56 Gbit/s</td>
<td>375 TB</td>
<td>E5v4</td>
</tr>
<tr>
<td></td>
<td>28x2.6</td>
<td></td>
<td></td>
<td></td>
<td>E5v4</td>
</tr>
<tr>
<td></td>
<td>28x2.6</td>
<td></td>
<td></td>
<td></td>
<td>s6g1</td>
</tr>
</tbody>
</table>
What is each cluster optimized for?

### Distinctive characteristics

- **Network**: Fast vs. Slow, Low latency vs. High latency
- **Storage**: Standard vs. High performance parallel
- **Accelerators**: GPUs, Xeon Phi, etc...

### Parallel multi-node workloads: deneb and fidis

- **Network**: fast low latency network interconnect (Infiniband)
- **Allocation**: whole nodes

### Single node and serial workload: deneb *serial*

- **Allocation**: individual cores
Shared Storage (cluster)

/scratch

- high performance temporary space
- not backed up
- low redundancy, built for performance
- local to each cluster
- automatically cleanup procedure deletes files (older than 2 weeks) without warning (when occupancy reaches a threshold)
- for disposable files: intermediary results, temporary files
## Shared Storage (global)

### /home
- per user quotas of 100GB
- backed up to a remote site
- available on all clusters
- for important files: source code, final results, theses

### /work
- per group quotas
- 50GB for free
- for more: 300CHF/TB for 3 years (+ 100CHF/TB for backup)
- available on all clusters
- for common files: software, datasets
Connecting to a cluster [Hands-on!]

**ssh username@deneb2.epfl.ch**

- Linux: simply connect using `ssh`
- Windows: install git-bash, connect using `ssh` from git-bash
- OSX: connect using `ssh`

**Basic shell commands, moving around**

- `id`
- `pwd`
- `ls /scratch/\<username\>`
- `cd /scratch/\<username\>`
Batch Systems

Batch
Goal: to take a list of jobs and execute them according to a priority when appropriate resources become available
Interactive use is possible but it not the principal way of running jobs!

SLURM
We use SLURM on all our clusters. It’s widely used in the HPC world and open source.
http://slurm.schedmd.com
The fundamental command is `sbatch` which submits jobs to the batch system.

**Workflow**

A typical workflow to get your computation done is:

- create a short job-script
- submit it to the batch system
- *it will get executed*
- look at the output

The job **will wait in the queue** until resources are available to run it.
Workflow of a job
Workflow of a job
Workflow of a job
Workflow of a job
Exercise 1: sbatch [Hands-on!]

Prerequisites
Copy the examples to your working directory:

```bash
cp -r /scratch/examples/using-the-clusters .
```

Open and edit the first exercise
Open the file ex1.sh with your editor of choice:
- nano
- emacs
- vim
- gedit
- ...

Exercise 1: sbatch [Hands-on!]

```bash
#!/bin/bash
#SBATCH --workdir /scratch/<put-your-username-here>
#SBATCH --nodes 1
#SBATCH --ntasks 1
#SBATCH --cpus-per-task 1
#SBATCH --mem 1G
#SBATCH --partition serial
#SBATCH --account scitas-courses
#SBATCH --reservation intro2clusters

sleep 10
echo "hello from $(hostname)"
sleep 10
```
Exercise 1: SLURM directives

#SBATCH: directive to the batch system

--nodes 1
# the number of nodes to use – on Castor this is limited to 1

--ntasks 1
# the number of tasks (in an MPI sense) to run per job

--cpu-per-task 1
# the number of cores per aforementioned task

--mem 4096
# the memory required per node in MB

--time 12:00:00
--time 2-6
# the time required (12 hours and 2 days 6 hours respectively)

--partition build
# select a particular partition serial, parallel, debug, build
Exercise 1: submit ex1.sh to the batch system [Hands-on!]

No time was specified so it defaults to 15 minutes

$ sbatch ex1.sh
Submitted batch job 1509281

$ cat /scratch/<username>/slurm-1509281.out
hello from r02-node12

Remember the Job ID

The number returned by `sbatch` is known as the **Job ID** and is the unique identifier for a task. If you ever need to ask for help you’ll need to know this number.
### Sjob job-id

$ sjob 1451763

<table>
<thead>
<tr>
<th>JobID</th>
<th>JobName</th>
<th>Cluster</th>
<th>Account</th>
<th>Partition</th>
<th>TimeLimit</th>
<th>User</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1451763</td>
<td>ex1.sh</td>
<td>deneb</td>
<td>scitas-ge</td>
<td>debug</td>
<td>00:15:00</td>
<td>rmsilva</td>
<td>scitas-ge</td>
</tr>
<tr>
<td>1451763.bat+</td>
<td>batch</td>
<td>deneb</td>
<td>scitas-ge</td>
<td>debug</td>
<td>00:15:00</td>
<td>rmsilva</td>
<td>scitas-ge</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Submit</th>
<th>Eligible</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Elapsed</th>
<th>ExitCode</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00:20</td>
<td>0:0</td>
<td>COMPLETED</td>
</tr>
<tr>
<td>00:00:20</td>
<td>0:0</td>
<td>COMPLETED</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NCPUS</th>
<th>NTasks</th>
<th>NodeList</th>
<th>UserCPU</th>
<th>SystemCPU</th>
<th>AveCPU</th>
<th>MaxVMSize</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>1</td>
<td>r02-node02</td>
<td>00:00.002</td>
<td>00:00.025</td>
<td>00:00:00</td>
<td>145216K</td>
</tr>
</tbody>
</table>
Socket, Core, Threads vs CPU

**Socket**
Receptacle on the motherboard for one physically packaged processor (each of which can contain one or more cores).

**Core**
A complete private set of registers, execution units, and retirement queues needed to execute programs.

**Threads**
One or more hardware contexts within a single core. Each thread has attributes of one core; managed & scheduled as a single logical processor by the OS.
Cancelling jobs

**scancel**

To cancel a specific job:

```bash
scancel <JOB_ID>
```

To cancel all your jobs:

```bash
scancel -u <username>
```

To cancel all your jobs that are not yet running:

```bash
scancel -u <username> -t PENDING
```
Exercise 2: squeue [Hands-on!]

59GB of memory required

```bash
#!/bin/bash
#SBATCH --workdir /scratch/<username>
#SBATCH --nodes 1
#SBATCH --ntasks 1
#SBATCH --cpus-per-task 8
#SBATCH --mem 59G
#SBATCH --time 00:30:00
#SBATCH --partition serial
#SBATCH --account scitas-courses
#SBATCH --reservation intro2clusters

cd /scratch/examples/linpack/
./runme_xeon64
```
Exercise 2: what’s going on? [Hands-on!]

squeue

With no arguments squeue will list all jobs currently in the queue! The output and information shown can be refined somewhat by giving options.

- `squeue -t R -u username`
- `squeue -t PD -u username`
- `squeue -t PD -u username --start`

Queue

Squeue is a custom squeue that shows only your jobs with more useful information.

scontrol: more details

`scontrol -dd show job <jobid>`
Why is a module system needed

- The OS version is restricted to an older one due to compatibility requirements of storage systems and specialized interconnects.
- The above is often in direct conflict with the needs of the HPC community, for which newer versions bring performance improvements and support for newer hardware (new CPU features).
- Many scientific codes are not even packaged under most Linux distributions.

Lmod

- **Lmod** is a utility that allows multiple, often incompatible, tools and libraries to co-exist on a system.
How software modules are organised

- Packages are organized hierarchically: Compiler / MPI / blas
- **Lmod** is designed to maintain the environment consistent
- **Lmod** does everything possible to automatically reload any software when one of the hierarchy layers is changed

Basic commands

- `module av(ailable)`
- `module load / unload <module-name>`
- `module spider <name>`
- `module save / restore <mnemonic-name>`
- `module purge`
Exercise 3: modules [Hands-on!]

ex3.sh: using module files

```bash
#!/bin/bash
#SBATCH --ntasks 1
#SBATCH --cpus-per-task 1
#SBATCH --nodes 1
#SBATCH --mem 4G
#SBATCH --time 00:05:00
#SBATCH --partition serial
#SBATCH --account scitas-courses
#SBATCH --reservation intro2clusters

echo STARTING AT $(date)
module purge
module load matlab
matlab -nodesktop -nojvm -r mymfile

echo FINISHED AT $(date)
```
Exercise 4: MPI parallel jobs [Hands-on!]

ex4.sh: srun to run parallel jobs

```bash
#!/bin/bash

#SBATCH --nodes 2
#SBATCH --ntasks 32
#SBATCH --cpus-per-task 1

module purge
module load intel
module load intel-mpi

cd osu
make clean
make

srun ./osu_allgather
```
Interactive access

Why interactive?

For debugging or running applications such as Matlab interactively we don’t want to submit a batch job.

Sinteract or salloc

There are two main ways of getting access depending on what you want to achieve:

- **salloc** - standard tool for an interactive allocation
- **Sinteract** - custom tool to access a node

Behind the scenes both use the same mechanism as sbatch to get access to resources.
Sinteract --help


options:
   -c cores    cores per task (default: 1)
   -n tasks    number of tasks (default: 1)
   -t time     as hh:mm:ss (default: 00:30:00)
   -m memory   as #[K|M|G] (default: 4G)
   -p partition (default: parallel)
   -a account  (default: scitas-ge)
   -q qos      as [normal|gpu|gpu_free|mic|...] (default: )
   -g resource as [gpu|mic][:count] (default is empty)
   -r reservation reservation name (default is empty)
   -s constraints list of required features (default is empty)

examples:
/usr/bin/Sinteract -c 16 -p serial
/usr/bin/Sinteract -p gpu -q gpu_free -g gpu:1
Sinteract [Hands-on!]

Sinteract

[<user>@deneb2 ~]$ Sinteract -p debug
Cores: 1
Tasks: 1
Time: 00:30:00
Memory: 4G
Partition: debug
Account: scitas-ge
Jobname: interact
Resource:
QOS: scitas
Reservation:

salloc: Granted job allocation 1451785
salloc: Waiting for resource configuration
salloc: Nodes r02-node02 are ready for job
[<user>@r02-node02 ~]$
salloc [Hands-on!]

salloc then srun for MPI tasks

```
[<user>@deneb2 ]$ salloc -N 1 -n 2 --mem 2048 -p debug
salloc: Granted job allocation 1451788
salloc: Waiting for resource configuration
salloc: Nodes r02-node02 are ready for job

[<user>@deneb2 ]$ hostname
deneb2

[<user>@deneb2 ]$ srun hostname
r02-node02
r02-node02

[<user>@deneb2 ]$ exit
salloc: Relinquishing job allocation 1451788
```
Sinteract: storage [Hands-on!]

Storage locations

The different storage locations are accessible via environment variables.

- `$SCRATCH` - your scratch folder on the current cluster `/scratch` filesystem
- `$TMPDIR` - a temporary folder in a local filesystem (generally `/tmp`)
- `$WORK` - your groups’ folder in the global `/work` filesystem

Try it: `Sinteract -p debug`

[rmsilva@r02-node02 ~]$ echo $SCRATCH
/scratch/rmsilva

[rmsilva@r02-node02 ~]$ echo $TMPDIR
/tmp/1451790

[rmsilva@r02-node02 ~]$ echo $WORK
/work/scitas-ge
The debug partition

`--partition=debug`

All the clusters have a few nodes that only allow short jobs and are intended to give you quick access to debug jobs:

- `salloc -p debug`
- `#SBATCH -p debug`
- `Sinteract -p debug`

The number of nodes in the partition and the limits vary by cluster.
Not everyone is equal

A group’s priority on the clusters is related to the number of shares (percentage of the cluster) that they have committed to use (paid) and their recent usage relative to their shares.

Within a group the relative consumption of the members decides who has more priority.

http://slurm.schedmd.com/fair_tree.html
http://slurm.schedmd.com/priority_multifactor.html
Current usage and priority (truncated)

<table>
<thead>
<tr>
<th>Account</th>
<th>User</th>
<th>RawShares</th>
<th>NormShares</th>
<th>RawUsage</th>
<th>NormUsage</th>
<th>EffectvUsage</th>
<th>FairShare</th>
</tr>
</thead>
<tbody>
<tr>
<td>scitas-ge</td>
<td>aubort</td>
<td>1</td>
<td>0.062500</td>
<td>0</td>
<td>0</td>
<td>0.0000000</td>
<td>0.195841</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>clemenco</td>
<td>1</td>
<td>0.062500</td>
<td>12486</td>
<td>0.000027</td>
<td>0.138814</td>
<td>0.171577</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>culpo</td>
<td>1</td>
<td>0.062500</td>
<td>186</td>
<td>0.000000</td>
<td>0.002074</td>
<td>0.173310</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>ddossant</td>
<td>1</td>
<td>0.062500</td>
<td>0</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.195841</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>degiorgi</td>
<td>1</td>
<td>0.062500</td>
<td>0</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.195841</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>desbioll</td>
<td>1</td>
<td>0.062500</td>
<td>0</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.183709</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>eroche</td>
<td>1</td>
<td>0.062500</td>
<td>0</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.195841</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>foureste</td>
<td>1</td>
<td>0.062500</td>
<td>0</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.181976</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>guglielm</td>
<td>1</td>
<td>0.062500</td>
<td>0</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.195841</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>nvarini</td>
<td>1</td>
<td>0.062500</td>
<td>0</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.195841</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>query</td>
<td>1</td>
<td>0.062500</td>
<td>1</td>
<td>0.000000</td>
<td>0.000012</td>
<td>0.178510</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>rezzonic</td>
<td>1</td>
<td>0.062500</td>
<td>80</td>
<td>0.000000</td>
<td>0.000897</td>
<td>0.175043</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>richart</td>
<td>1</td>
<td>0.062500</td>
<td>1</td>
<td>0.000000</td>
<td>0.000019</td>
<td>0.176776</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>rmsilva</td>
<td>1</td>
<td>0.062500</td>
<td>0</td>
<td>0.000000</td>
<td>0.000004</td>
<td>0.180243</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>scitasbui+</td>
<td>1</td>
<td>0.062500</td>
<td>77195</td>
<td>0.000170</td>
<td>0.858180</td>
<td>0.169844</td>
</tr>
<tr>
<td>scitas-ge</td>
<td>vkeller</td>
<td>1</td>
<td>0.062500</td>
<td>0</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.195841</td>
</tr>
</tbody>
</table>

Level FS shows you if you are under (>1) or over (<1) served.
Helping yourself

**man pages are your friends!**

- man sbatch
- man sacct
- man gcc
- module load intel; man ifort
Getting help

1234@epfl.ch

- send a mail to 1234@epfl.ch
- start the subject with HPC

We need to know as many of the following as possible

- the Job ID
- the directory location and name of the submission script
- where the “slurm-*.out” file is to be found
- how the “sbatch” command was used to submit it
- the output from “env” and “module list” commands
Going further

SCITAS offers courses in

- Compiling code and using MPI
- MPI, an introduction to parallel programming
- MPI, advanced parallel programming
- Introduction to profiling and software optimisation
- Computing on GPUs
- Data Management: code and large files
- Introduction to Linux
Useful links

Change your shell at:

https://cadiwww.epfl.ch/cgi-bin/accountprefs/

SCITAS web site:

http://scitas.epfl.ch

(in particular) SCITAS documentation space:

http://scitas-data.epfl.ch/kb

SLURM man pages:

http://slurm.schedmd.com/man_index.html