Short Slurm Introduction

Cluster management and job scheduling system for CLAIX.

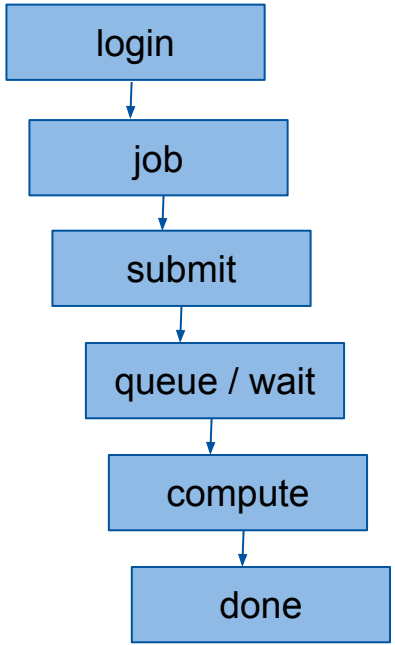
NHR4 CES
NHR for Computational Engineering Science

slurm workload manager

RWTH AACHEN UNIVERSITY

Batch system for CLAIX

- We use Slurm
- Batch system:
  - Share resources “fairly”
  - **Queue**’s user programs as **jobs**
  - Considers **priority** to decide order
  - **Allocates** time and resources to **“jobs”** from users.
  - Starts, executes, and monitors **jobs**.

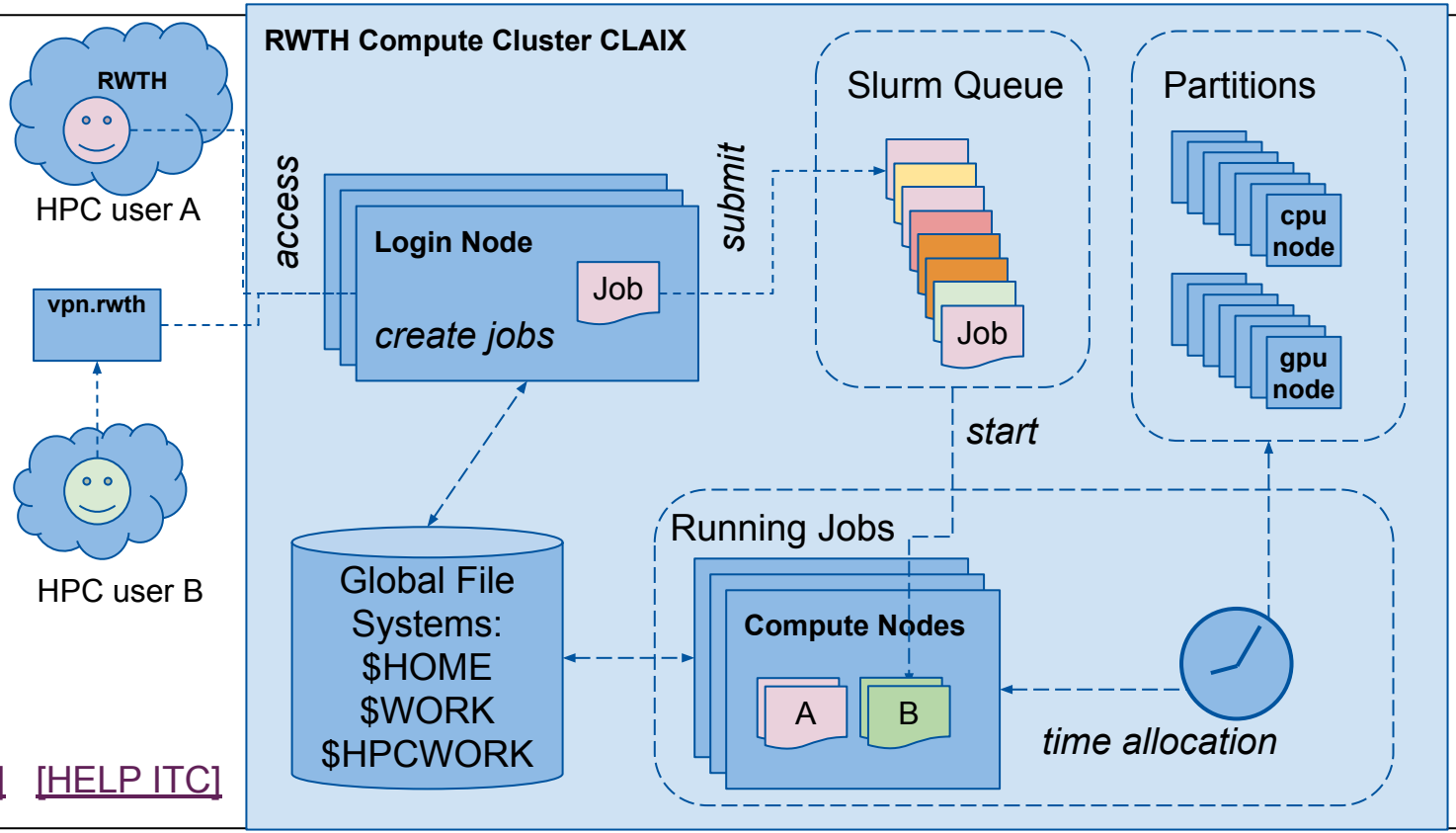


22.05.4

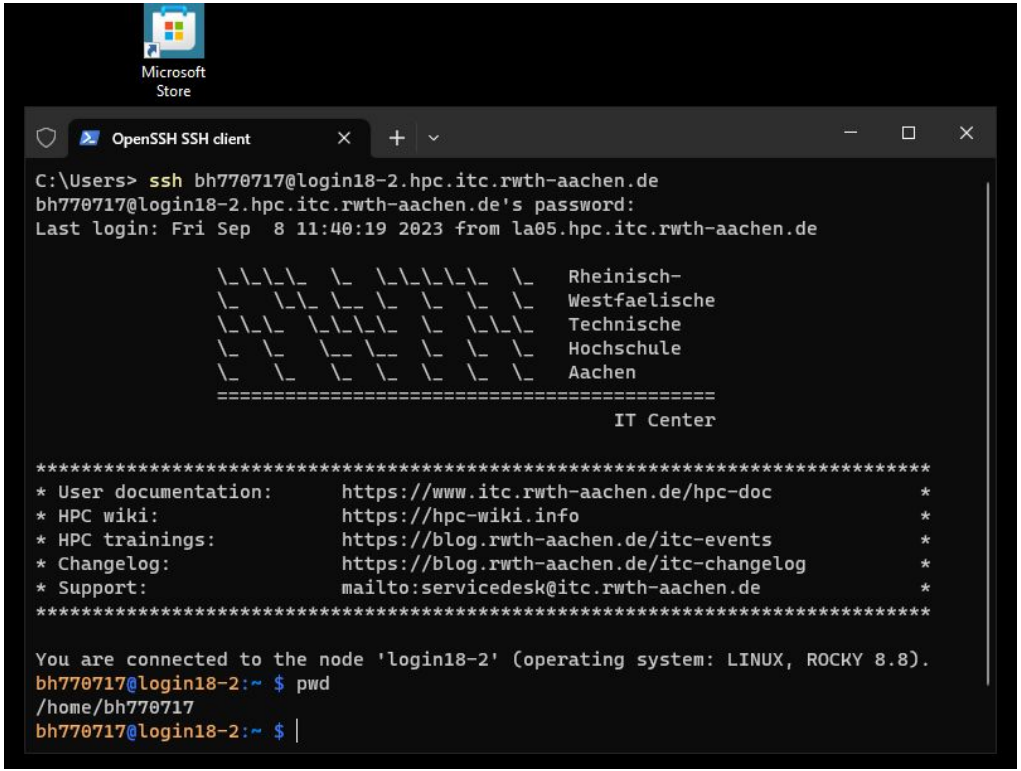
[LINK]

SLURM: Resource Manager + Scheduler

Batch system for CLAIX



Login - SSH



Login Nodes Do's:

- Code, Compile, Edit
- SSH to other nodes
- Slurm commands
- Basic GUI programs

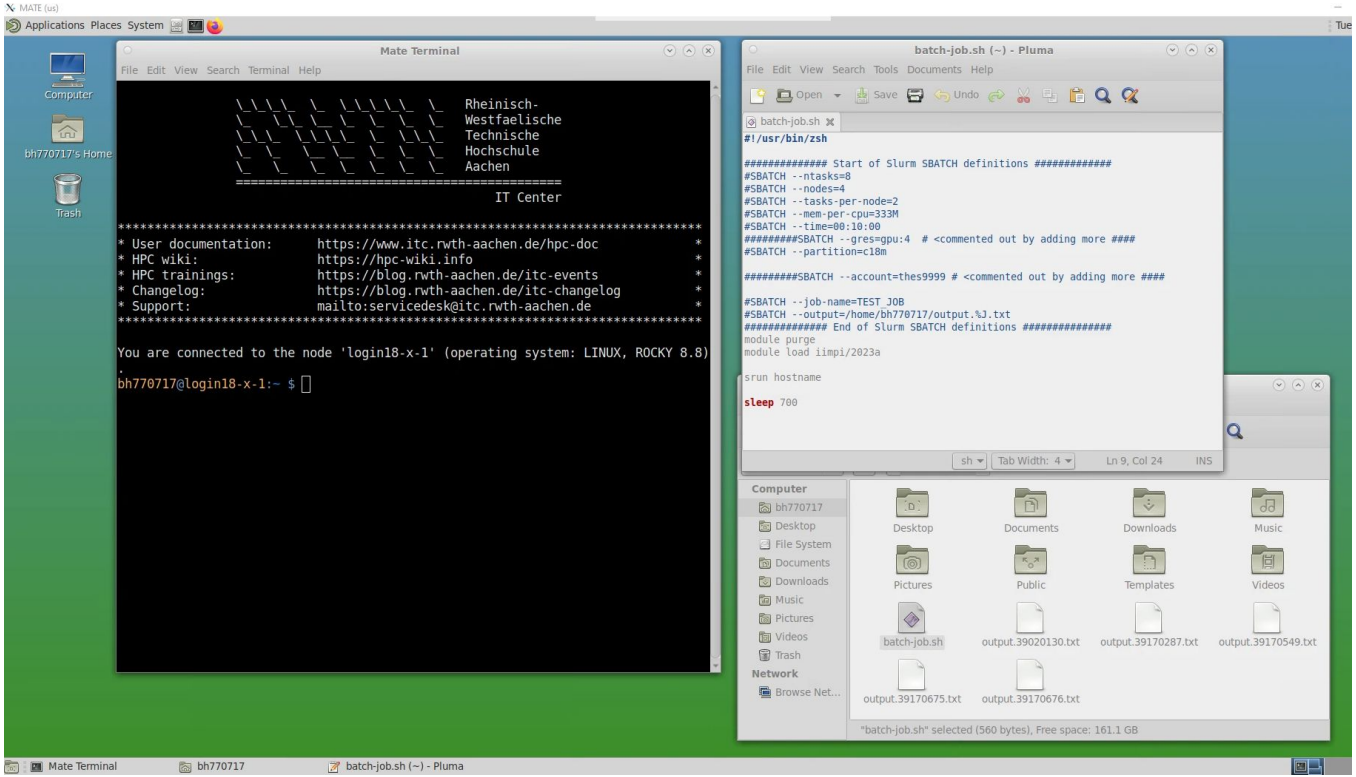
Login Nodes Dont's:

- Large data processing
- MPI, PARALLEL, intensive programs
- Use all resources
- Monitoring

- **We Limit Cores + Mem**
- **We will kill the programs and you will lose the work**
- **Login nodes are restarted mondays**

Login - FastX

[LINK]

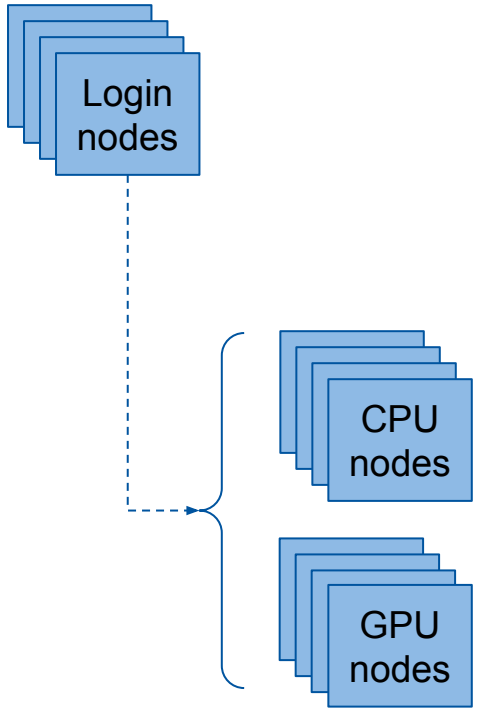


Compute nodes

- Submit jobs:
  - **Submit** programs to compute nodes
  - **Compute nodes** is where computations/tests/work are done
  - Queue/wait for requested resources

- Commands
  - **salloc** <args> run cmd's interactively
  - **sbatch** script.sh submit batch script job

- The **login-nodes** are NOT the "compute cluster"
  - Still you can:
    - Compile, run, test programs and work



Compute nodes - salloc

- An interactive session with **salloc** needs:
  - **Resources:** CPUs, GPUs, Nodes, Time
  - **Optional partition**: Group of nodes: use **devel** for quick tests!

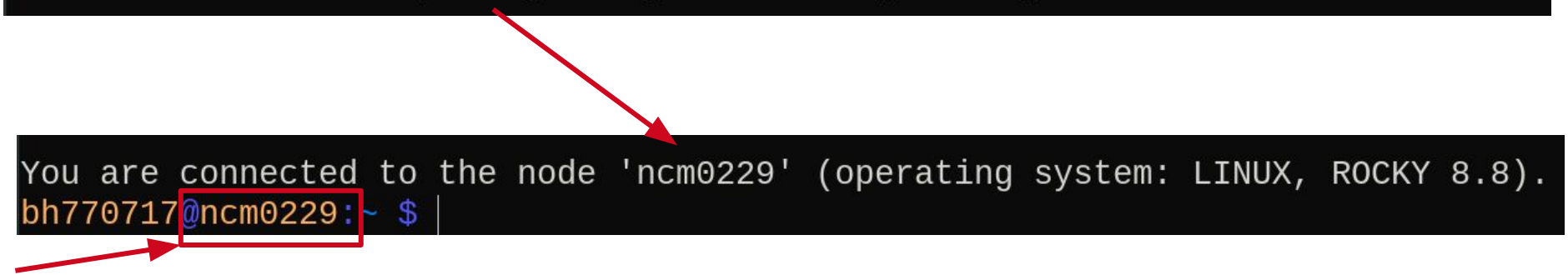
```
bh770717@login18-1:~ $ salloc -n 8 -N 2 -p c18m --time=01:00:00 --mem=7G --account=supp0001
salloc: [I] No output file given, set to: output_%j.txt
salloc: Pending job allocation 39234937
salloc: job 39234937 queued and waiting for resources
salloc: job 39234937 has been allocated resources
salloc: Granted job allocation 39234937
salloc: Waiting for resource configuration
salloc: Nodes ncm[0707,0713] are ready for job
```

Partition devel (no account):

```
bh770717@login18-1:~ $ salloc -n 8 -N 2 -p devel --time=00:25:00 --mem=7G
```

Compute nodes - salloc

```
bh770717@login18-1:~ $ salloc -n 8 -N 2 --time=01:00:00 --mem=7G
salloc: [I] No output file given, set to: output_%j.txt
salloc: Pending job allocation 39214003
salloc: job 39214003 queued and waiting for resources
salloc: job 39214003 has been allocated resources
salloc: Granted job allocation 39214003
salloc: Waiting for resource configuration
salloc: Nodes ncm[0229,0238] are ready for job
```



Compute nodes - salloc

```
You are connected to the node 'ncm0229' (operating system: LINUX, ROCKY 8.8).
bh770717@ncm0229:~ $ exit
--------------------------------------------------------------------------------
Bis zum naechsten Mal, bh770717!
--------------------------------------------------------------------------------
salloc: Relinquishing job allocation 39214003
bh770717@login18-1:~ $
```

Compute nodes - sbatch

- Define a Slurm Job (a file):
  - **Resources:** CPUs, GPUs, Nodes, Time
  - **Partition**: Group of nodes
  - **Account**: Project account, resources get billed
  - **Modules:** Our provided libraries and programs
  - **Your Program:** commands in a script file.
- Code it using either:
  - Console editors: vim, nano, pico, emacs, …
  - GUI editors, PLUMA, EMACS, …

DEMO

Compute nodes - sbatch

```
#!/usr/bin/zsh

## > start SBATCH

## Request Resources
## More info slurm.schedmd.com/sbatch.html

#SBATCH -n 1           # Use 1 CPU ...
#SBATCH -N 1           # within 1 Node ...
#SBATCH --mem 2500MB   # and 2500MB of main memory ...
#SBATCH -t 00:15:00    # for up to 15 Minutes.
#SBATCH -o %J-out.txt # Output: path to file must exist!

## < end SBATCH


## Your program

srun hostname
```

Compute nodes - scancel

```
bh770717@login18-1:~/hpc-intro/single-cpu $ sbatch single-cpu.sh
Submitted batch job 43221626
bh770717@login18-1:~/hpc-intro/single-cpu $ cat 43221626-out.txt
ncm0014.hpc.itc.rwth-aachen.de
bh770717@login18-1:~/hpc-intro/single-cpu $ s
```

Compute nodes - scancel

```
bh770717@login18-1:~ $ sbatch batch-job.sh
Submitted batch job 39170549
bh770717@login18-1:~ $ squeue --me
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          39170549      c18m TEST_JOB bh770717  R       0:24      4 ncm[0710,0713,0718,0720]
bh770717@login18-1:~ $ scancel 39170549
bh770717@login18-1:~ $ squeue --me
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
bh770717@login18-1:~ $
```

Compute nodes - squeue

```
bh770717@login18-1:~ $ squeue --me
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
          39170287      c18m TEST_JOB bh770717  R       8:46      4 ncm[0077,0080,0149,0151]
```

```
bh770717@login18-1:~ $ scontrol show job 39170287
JobId=39170287 JobName=TEST_JOB
   UserId=bh770717(45727) GroupId=bh770717(45727) MCS_label=N/A
   Priority=425469 Nice=0 Account=default QOS=normal
   JobState=RUNNING Reason=None Dependency=(null)
   Requeue=0 Restarts=0 BatchFlag=1 Reboot=0 ExitCode=0:0
   RunTime=00:10:13 TimeLimit=00:10:00 TimeMin=N/A
   SubmitTime=2023-09-15T12:04:29 EligibleTime=2023-09-15T12:04:29
   AccrueTime=2023-09-15T12:04:29
   StartTime=2023-09-15T12:04:30 EndTime=2023-09-15T12:14:30 Deadline=N/A
   SuspendTime=None SecsPreSuspend=0 LastSchedEval=2023-09-15T12:04:30 Scheduler=Main
   Partition=c18m AllocNode:Sid=login18-1:34317
   ReqNodeList=(null) ExcNodeList=(null)
   NodeList=ncm[0077,0080,0149,0151]
   BatchHost=ncm0077
   NumNodes=4 NumCPUs=8 NumTasks=8 CPUs/Task=1 ReqB:S:C:T=0:0:*:*
   TRES=cpu=8,mem=2664M,node=4,billing=8
   Socks/Node=* NtasksPerN:B:S:C=2:0:*:* CoreSpec=*
   MinCPUsNode=2 MinMemoryCPU=333M MinTmpDiskNode=0
   Features=Rocky8&hostok DelayBoot=00:00:00
   OverSubscribe=OK Contiguous=0 Licenses=(null) Network=(null)
   Command=/rwthfs/rz/cluster/home/bh770717/batch-job.sh
   WorkDir=/rwthfs/rz/cluster/home/bh770717
   AdminComment=##itc exclusive=false itc##
   StdErr=/home/bh770717/output.%J.txt
   StdIn=/dev/null
   StdOut=/home/bh770717/output.%J.txt
```

GPU jobs

optional

```
bh770717@login18-x-1:~ $ salloc -n 24 -N 1 --gres=gpu:1 --time=00:10:00 -p c18g
salloc: [I] No output file given, set to: output_%j.txt
salloc: Pending job allocation 39243549
salloc: job 39243549 queued and waiting for resources
salloc: job 39243549 has been allocated resources
salloc: Granted job allocation 39243549
salloc: Waiting for resource configuration
salloc: Nodes nrg04 are ready for job
```

Slurm as the batch system for CLAIX

Pending Reasons

- None
  - The job has not been in a schedule run of Slurm up to now
- Priority
  - At least one other job has a higher priority and will run first on the same resources
- Resources
  - The job is waiting for resources to become free
- AssocMaxWallDurationPerJobLimit
  - The job requested a longer runtime than it is allowed
- AssocMaxCpuPerJobLimit
  - The job requested more cpus than it is allowed
- JobArrayTaskLimit
  - The job array has more running tasks than allowed
- Dependency
  - The job is waiting for another specific job to end

Projects (--account) and Quota

Disk Quota - r_quota

Going over disk Quota is bad!

- Cannot start jobs
- Cannot save work
- Will crash applications

```
bh770717@login18-1:~ $ r_quota
                        ------------ Blocks ------------- ------------- Files --
Object                    used   soft   hard        grace   used   soft   hard
/home/bh770717            512K      -   150G            -      -      -      -
/work/bh770717              0K      -   250G            -      -      -      -
/hpcwork/bh770717           4K   1000G  1100G            -      1     49K     50K
```

[LINK]

Quota system

Quota

- Projects have a quota of corehours **granted** to be used use **per month FOR the month**

- But: we provide a „**3-month-window**“ to use the quota:
  - This month you could use unused quota from the previous month and „borrow“ quota from the next month.

- WARNING! Using more than 3x your Monthly Quota:
  - Your jobs will **only** start, if no one else is using the CLAIX.

Quota system

Slurm Accounts (not = your user)

- Accounts have a **default** partition and **„allowed**“ partitions
  - The account **„default**“ has **„c18m**“ as default partition and is allowed **„c18g**“
- Submission without a project results in submission to the **„default**“ account
- In which projects am I involved?
  - Use „**r_wlm_usage -p <projectname/account>**“
  - Shows
    - Allowed partitions
    - Max usable cores per job
    - Max runtime limit per job
    - Consumed corehours of the last 4 weeks
    - Consumed corehours up to now and total granted corehours
- **r_wlm_usage -q** shows user quota information

r_wlm_usage

```
kz743613@login18-1:~/tmp/example $ r_wlm_usage -q
User:                                        kz743613
Status of user:                     RWTH Mitarbeiter
Quota monthly (core-h):                         2000
Remaining  core-h of prev. month:               1331
Consumed   core-h current month:                   5
Consumable core-h (%):                           166
Consumable core-h:                              5326
-------------------------------------------------
Consumed   core-h last 4 weeks:                  674
Consumed   core-h last year:                     826
```

```
You are involved in the following SLURM projects:

Account:                                     supp0001
Type:                                            supp
Start of Accounting Period:                08.05.2021
End of Accounting Period:                  07.05.2022
State of project:                              active
-------------------------------------------------
Quota monthly (core-h):                        100341
Remaining  core-h of prev. month:               58738
Consumed   core-h current month:                 6850
Consumed   core-h last 4 weeks:                 14028
Consumable core-h (%):                            152
Consumable core-h:                             252571
-------------------------------------------------
Total quota (core-h):                      1.200 Mio
Total consumed core-h so far:              0.415 Mio
-------------------------------------------------
Default partition:                               c18m
Allowed partitions:          c16g,c18m,c18g,c16m,c16s
Max. allowed wallclocktime:              120.0 hours
Max. allowed cores per job:                    57600
```

Accounts

```
bh770717@login18-1:~ $ salloc -n 8 -N 2 --time=01:00:00 --mem=7G --account=supp0001
salloc: [I] Partition set to 'c18m' due to chosen account 'supp0001'.
salloc: [I] No output file given, set to: output_%j.txt
salloc: Pending job allocation 39218039
salloc: job 39218039 queued and waiting for resources
salloc: job 39218039 has been allocated resources
salloc: Granted job allocation 39218039
salloc: Waiting for resource configuration
salloc: Nodes ncm[0177,0180] are ready for job
```

Resources to core-hours Quota per Node per hour (max 48 per node)

Type	Amount used	**Core-hours** billed for 1 Hour use
Cores	1	1
Cores	12	12
Cores	48	48
Memory	3.75 GB	1
Memory	180 GB	48
GPU	1	24
GPU	2	48

Thank you for your attention.
Any questions?

Best Practices

- Avoid loading modules in your `.zshrc` or `.bashrc`.
  - Do not automatically load conda environments.
- Use the **module** version of Programs provided by us when possible.
  - Avoid your own conda environments when possible.
  - Use module spider <name> to search for already installed software
- Purge modules (module purge) when uncertain of library dependencies
- Test your application on **devel** nodes with smaller problem sizes using **salloc**.
- Put all commands the program at the **end** of the batch job script file.
- Don't forget the module commands and the shebang (#!):
  - `#!/usr/bin/zsh`
- If you use relative paths: know the directory in which the job starts.
- Make the scripts executable (`chmod +x myscript.sh`)
- Confirm your program's memory / stack requirements fit in your job memory request.
- Avoid moving millions of tiny files between **GFS** and compute nodes.
  - Read once, distribute within the program.
- Do not run out of disk quotas!
- Use your project quota!