SLURM

Stefan Kemnitz[1]

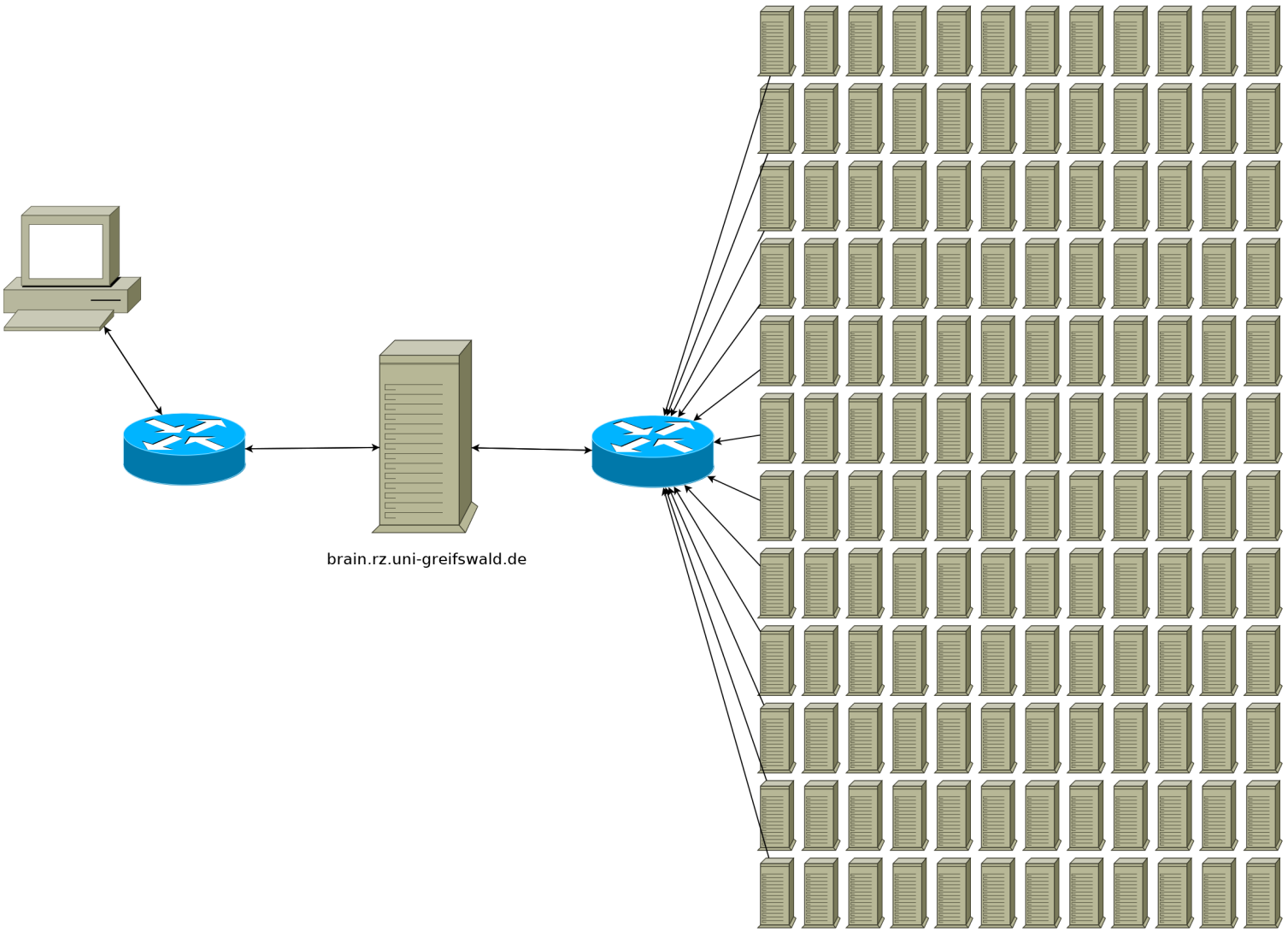
[1]Department of distributed high performance computing
University of Rostock

SLURM, 2019

Simple Linux Utility for Resource Management

"Slurm is the workload manager on six of the top ten systems including the number 1 system, Sunway TaihuLight with 10,649,600 computing cores" wikipedia

Cluster Diagram



Partitions

- group machines by hardware specification
  - compute (public)
  - batch (public)
  - gpu (public)
  - atlas (closed)
  - iapetos (closed)
- jobs can not be spread across multiple partitions

Slurm Commands

command	result
sinfo	show state of the system
squeue	show running jobs
sbatch	submit job script
scancel	stop job by number
sshare	show used ressources in cpu minutes

Example output sinfo

```
PARTITION AVAIL  TIMELIMIT  NODES  STATE NODELIST
all          up 3-00:00:00      2 maint* node[079-080]
all          up 3-00:00:00      2  drain node[086,088]
all          up 3-00:00:00     91    mix gpu[02-09],node[001-030,032-071,081-083,085,089-092,094,098,101-103]
all          up 3-00:00:00      9  alloc node[031,084,087,093,095-097,099-100]
all          up 3-00:00:00      9   idle gpu[01,10],node[072-078]
atlas        up 3-00:00:00      4 drain* node[112-115]
atlas        up 3-00:00:00     16   idle node[104-111,116-123]
batch        up 3-00:00:00      2 maint* node[079-080]
batch        up 3-00:00:00      2  drain node[086,088]
batch        up 3-00:00:00     13    mix node[081-083,085,089-092,094,098,101-103]
batch        up 3-00:00:00      8  alloc node[084,087,093,095-097,099-100]
compute*     up 3-00:00:00     70    mix node[001-030,032-071]
compute*     up 3-00:00:00      1  alloc node031
compute*     up 3-00:00:00      7   idle node[072-078]
gpu          up 3-00:00:00      8    mix gpu[02-09]
gpu          up 3-00:00:00      2   idle gpu[01,10]
iapetos      up 3-00:00:00      9  down* node[127,129-132,141-144]
iapetos      up 3-00:00:00      5    mix node[136,145-148]
iapetos      up 3-00:00:00      6  alloc node[125-126,128,133-135]
iapetos      up 3-00:00:00     16   idle node[137-140,149-160]
```

sbatch

```
sbatch my_job_script.sh arg0 arg1 arg2 arg3
```

1. makes a copy of my_job_script .sh
2. stores all environment settings
   - which path you are in
   - loaded modules
   - exported variables
3. sends this to the controllers queue
4. returns an ID which represents your job

Example output squeue

```
 JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
411687       gpu R-TIGER2 ps114092 PD       0:00      8 (Dependency)
411688       gpu R-TIGER2 ps114092 PD       0:00      8 (Dependency)
411689       gpu R-TIGER2 ps114092 PD       0:00      8 (Dependency)
411690       gpu R-TIGER2 ps114092 PD       0:00      8 (Dependency)
411691       gpu R-TIGER2 ps114092 PD       0:00      8 (Dependency)
411860   compute submitc.    walbr PD       0:00      2 (Dependency)
411861   compute submitc.    walbr PD       0:00      2 (Dependency)
411870   compute submitc.    walbr PD       0:00      2 (Dependency)
411871   compute submitc.    walbr PD       0:00      2 (Dependency)
411686       gpu R-TIGER2 ps114092  R 1-19:49:51      8 gpu[02-09]
411851   compute submitc.    walbr  R    9:16:28      2 node[032-033]
411850   compute submitc.    walbr  R 1-10:29:43      2 node[005-006]
411867   compute submitc.    walbr  R 2-07:05:48      2 node[001-002]
411864   compute submitc.    walbr  R 2-23:44:31      2 node[007-008]
424284   iapetos K-1J0Jb0   langef  R   13:04:46      4 node[145-148]
424361   iapetos t0.125a1   langef  R    1:13:36      1 node128
424362   iapetos t0.125a1   langef  R    1:13:36      1 node128
424363   iapetos t0.175a1   langef  R    1:13:36      1 node128
424364   iapetos t0.175a1   langef  R    1:13:36      1 node128
424365   iapetos t0.225a1   langef  R    1:13:36      1 node128
424366   iapetos t0.225a1   langef  R    1:13:36      1 node128
424367   iapetos t0.275a1   langef  R    1:13:36      1 node128
424368   iapetos t0.275a1   langef  R    1:13:36      1 node128
424237     batch FreeSurf frenzels  R      25:20      1 node095
420220   compute  DP1_S30 pm101481  R    2:10:31      4 node[036-039]
417025   compute DP1_S60_ pm101481  R 2-00:00:50      2 node[009-010]
417027   compute DP1_S60_ pm101481  R 2-00:00:50      2 node[011-012]
417023   compute DP1_S100 pm101481  R 2-00:01:55      2 node[028-029]
417021   compute DP1_S100 pm101481  R 2-00:02:53      2 node[003-004]
```

Controller

- squeue shows all jobs
- filter your jobs by squeue -u username
- as soon as a machine is ready sends your script to that machine

Fair sharing

- every minute you allocate on a core is recorded
- will be added to your account and your group account
- your fraction of the total minutes used will decide who comes next in the queue

Example output sshare

Account	User	RawShares	NormShares	RawUsage	EffectvUsage	FairShare
root			1.000000	1427785127	0.000000	1.000000
biochemie		1	0.045455	403990430	0.283021	0.013355
bioinformatik		1	0.045455	9458	0.000007	0.999899
bwl		1	0.045455	15522275	0.010873	0.847216
chem		1	0.045455	0	0.000000	1.000000
compus		1	0.045455	213088425	0.149262	0.102681
compus	kemnitzs	1	0.002525	2723	0.008294	0.102630
compus_ext		1	0.045455	16190025	0.011342	0.841172
genomforschung		1	0.045455	72411077	0.050729	0.461363
ipp		1	0.045455	0	0.000000	1.000000
mathematik_informa+		1	0.045455	12107744	0.008472	0.878809
none		1	0.045455	23353275	0.016360	0.779203
pharmazie		1	0.045455	11593872	0.008119	0.883551
physik_fehske		1	0.045455	19547286	0.013681	0.811701
physik_henning		1	0.045455	361	0.000000	0.999996
physik_ihle		1	0.045455	76966833	0.053917	0.439469
psychiatrie		1	0.045455	477842251	0.334558	0.006086
radiologie		1	0.045455	13506474	0.009462	0.865636
trash		1	0.045455	0	0.000000	1.000000
urz		1	0.045455	0	0.000000	1.000000
zoologie		1	0.045455	71648586	0.050194	0.465136

sbatch options

- specify restrictions you want to apply
  - –time "days-hours:minutes:seconds"
  - –mem "amount in MB"
  - –partition "partition name"
- specify how many elements you want to have
  - –nodes "number"
  - –cpus_per_task "number"
  - –ntasks "number"

Submit file

```
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus_per_task=1
#SBATCH --mem 1G
#SBATCH --time 20
#SBATCH --partition batch

./my_program.sh
```