Resource limits for parallel jobs on Linux Cluster

This subdocument contains a description of constraints under which parallel jobs execute on the cluster systems: maximum run times, maximum memory and other SLURM-imposed parameters.

- Resource limits for interactive jobs
- Resource Limits for batch jobs
- Resource limits on housed systems
- Details on Policies

Resource limits for interactive jobs

<table>
<thead>
<tr>
<th>Partition</th>
<th>Core counts and remarks</th>
<th>Run time limit (hours)</th>
<th>Memory limit (GB)</th>
</tr>
</thead>
</table>
| interactive nodes on CooLMUC-2 | Maximum number of cores per job on the back end: **56**. **Notes:**  
- Please do not use resources in this partition to run regular production jobs! This partition is meant for testing!  
- A given user account should not run more than one job at a time. | 2  
(default is 15 minutes) | 2 per physical core |
| interactive nodes on CooLMUC-3 | Maximum number of physical cores per job on the back end: **192**. **Notes:**  
- Please do not use resources in this partition to run regular production jobs! This partition is meant for testing!  
- A given user account should not run more than one job at a time. | 2  
(default is 15 minutes) | 1.4 per physical core |

Resource Limits for batch jobs

The following is an overview of the resource limits imposed for various classes of jobs. These are comprised of run time limits, limits on core counts for parallel jobs, and memory limits. Please consult the SLURM specifications subdocument for a more detailed explanation of parallel environments, in particular how to correctly specify memory requirements. With respect to run time limits it is recommended to always specify a target run time via the --time switch; this in particular for smaller jobs may allow the scheduler to perform backfilling.

- The designation "shared memory" for parallel jobs assumes that a number of cores assigned by SLURM will be used by threads; typically a command like export OMP_NUM_THREADS=<number> should be issued to achieve this.
- The designation "distributed memory" for parallel jobs assumes that MPI is used to start one single-threaded MPI task per core assigned by SLURM. In principle it is also possible to run hybrid MPI + threaded programs, in which case the number of cores assigned by the system will be equal to the product (# of MPI tasks) * (# of threads), rounded up if necessary.

<table>
<thead>
<tr>
<th>Job Type</th>
<th>Architecture (System)</th>
<th>Core counts and remarks</th>
<th>Run time limit (hours)</th>
<th>Memory limit (GB)</th>
</tr>
</thead>
</table>
| Distributed memory parallel (MPI) jobs | 28-way Haswell-EP nodes with Infiniband FDR14 interconnect (CooLMUC-2) | Please specify the cluster  
--clusters=mpp2  
as well as a core count of at least 28 (see example jobs for how to do this). Jobs specifying more than **1680 cores** (or 60 nodes) will not start. | 72  
2 per task |
| Shared memory parallel job | One 28-way Haswell-EP node (CooLMUC-2) | Please specify the cluster  
--clusters=mpp2  
and specify using a single node (see example jobs for how to do this) | 72  
56 per node |
| Distributed memory parallel job | 64-way KNL 7210-F node with Intel Omnipath interconnect (CooLMUC-3; submission node must be lxlogin8 or lxlogin10) | Please specify the cluster  
--clusters=mpp3  
as well as a core count of at least 64 (see example jobs for how to do this). Jobs specifying more than **2048 physical cores** (or 32 nodes) will not start. | 48  
~90 DDR, plus 16 HBM per node |
Distributed memory parallel job

- 8-way Ivy Bridge node with Infiniband FDR14 interconnect (IvyMUC; submission node must be lxlogin8 or lxlogin10)
- Please specify the cluster --clusters=ivy_muc
- Jobs specifying more than 192 physical cores (or 12 nodes) will not start
- lxlogin8 or lxlogin10

72
~56 per node

Shared memory thread-parallel job

- A subset of the HP DL580 shared memory system (actually considered part of the serial cluster, but information is replicated here)
- Please specify the cluster --clusters=inter and the partition --partition=teramem_inter
- as well as the number of cores needed for the single task to be started (see example jobs for how to do this). At most 64 logical cores can be used for a single job.

48
(default 8)
~60 per physical core (each physical core has 2 hyperthreads)

If a job appears to not use resources properly, it may get deleted at LRZ staff's or surveillance system's discretion.

Resource limits on housed systems

The clusters and partitions listed in this section are only available for institutes that have a housing contract with LRZ.

<table>
<thead>
<tr>
<th>Job Type</th>
<th>Architecture</th>
<th>Core counts and remarks</th>
<th>Run time limit (hours)</th>
<th>Memory limit (GByte)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributed memory parallel (MPI) jobs</td>
<td>28-way Haswell-EP nodes with Infiniband FDR14 interconnect</td>
<td>Please specify the cluster --clusters=tum_chem and one of the partitions --partition=[tum_chem_batch, tum_chem_test]</td>
<td>384 (test queue: 12)</td>
<td>2 per task (in MPP mode, using 1 physical core/task)</td>
</tr>
<tr>
<td>Distributed memory parallel (MPI) jobs</td>
<td>28-way Haswell-EP nodes with Infiniband FDR14 interconnect</td>
<td>Please specify the cluster --clusters=tum_ch2</td>
<td>48</td>
<td>2 per task (in MPP mode, using 1 physical core/task)</td>
</tr>
<tr>
<td>Distributed memory parallel (MPI) jobs</td>
<td>28-way Haswell-EP nodes with Infiniband FDR14 interconnect</td>
<td>Please specify the cluster --clusters=hm_mech</td>
<td>336</td>
<td>18 per task (in MPP mode, using 1 physical core/task)</td>
</tr>
<tr>
<td>Serial or shared memory jobs</td>
<td>sgi UV2K / UV20 / UV30 systems</td>
<td>Please specify the cluster --clusters=lmu_exc and select the partition according to the desired machine type.</td>
<td>depends on selected partition</td>
<td>depends on selected partition</td>
</tr>
<tr>
<td>Serial or shared memory jobs</td>
<td>28-way Haswell-EP nodes with Ethernet interconnect</td>
<td>Please specify the cluster --clusters=tum_geodesy</td>
<td>240</td>
<td>2 per task / 60 per node</td>
</tr>
<tr>
<td>Shared memory parallel job</td>
<td>Intel- or AMD-based shared memory systems</td>
<td>Please specify the cluster --clusters=myri and as well as one of the partitions --partition=myri_[p,u]</td>
<td>144</td>
<td>3.9 per core</td>
</tr>
</tbody>
</table>
Details on Policies

Policies for interactive jobs

Limitations

- On login shells, parallel programs should not be started directly. Please always use the salloc command to initialize a time-limited interactive parallel environment. Note that the shell initialized by the salloc command will still run on the login node, but executables started with srun will be started up on the interactive partitioned which was assigned.
- Usage of the cron system (e.g. via /usr/bin/crontab) as well as the /usr/bin/at or /usr/bin/batch commands is not allowed. You are required to use the documented SLURM commands.

Policies for queued batch jobs

General restrictions

- The job name should not exceed 10 characters. If no job name is specified, please do not use excessively long script names.
- Do not use the xargs command to generate command line arguments at submission time. Instead, generate any necessary arguments inside your script.

Scheduling

For parallel jobs, it is recommended to explicitly specify the run time limit. This may shorten the waiting time, since the job might be run in backfill mode (in other words: use resources that are free while the scheduler tries to fit another large job into the system). Your specification gives the scheduler the information required to organize this.

Jobs in Hold

Jobs in user hold will be removed at the LRZ administrators' discretion if older than 8 weeks.

Job Submissions

- There exists a maximum number of jobs that can be submitted by a user. This limit may change over time, dependent on the cluster load.
- Submission of large numbers of jobs (>100, including array jobs) with very short run time (< 1min) is considered a misuse of resources. It causes both waste of computational resources and - if mail notifications are used - disruption of the notification system. Users that submit such jobs will be banned from further use of the batch system. Bundle the individual jobs into a much bigger one!

Memory use

Jobs exceeding the physical memory available on the selected node(s) will be removed, either by SLURM itself, or the OOM ("out of memory") killer in the operating system kernel, or at LRZ's discretion since such a usage typically has a negative impact on system stability.

Limits on queued jobs

In order to prevent monopolization of the clusters by a single user, a limit of 50 queued jobs is imposed on both CooLMUC-2 and CooLMUC-3

Software licenses

Many commercial software packages have been licensed for usage on the cluster; most of these require the use of so-called floating licenses, only a limited amount of which are typically available. Since it is not possible to check whether a license is available before a batch job starts, LRZ cannot provide any guarantees that a batch job requesting use of such a license will run successfully.