SuperMUC-NG Job Processing with SLURM

- General
- List of relevant commands
- Queues (SLURM partitions) and their limits
- `srun` and `mpiexec`
- `salloc`, `srun` for interactive processing
- `sbatch` command / `#SBATCH` option
  - Batch Job Examples
    - General options applicable for all jobs
    - Options for resources and execution (click to expand)
  - Input Environment Variables
  - Output Environment Variables
  - Guidelines for resource selection
  - SLURM Documentation

General

The batch system on SuperMUC-NG is the open-source workload manager SLURM (Simple Linux Utility for Resource management). For details about the SLURM batch system, see Slurm Workload Manager.

Submit hosts are usually login nodes that permit to submit and manage batch jobs.

Intel processors on SuperMUC-NG support the hyperthreading mode which might increase the performance of your application. With hyperthreading, you have to increase the number of MPI tasks per node from 48 to 96 in your job script. Please be aware that with 96 MPI tasks per node each process gets only half of the memory by default. If you need more memory, you have to specify it in your job script and use the fat nodes (see example batch scripts).

List of relevant commands

<table>
<thead>
<tr>
<th>Command’s name</th>
<th>Functionality</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>sbatch</code></td>
<td>submit a job script</td>
</tr>
<tr>
<td><code>scancel</code></td>
<td>delete or terminate a queued or running job</td>
</tr>
<tr>
<td><code>squeue</code></td>
<td>print table of submitted jobs and their state. <strong>Note:</strong> non-privileged users can only see their own jobs.</td>
</tr>
<tr>
<td><code>salloc</code></td>
<td>create an interactive SLURM shell</td>
</tr>
<tr>
<td><code>srun</code></td>
<td>execute argument command on the resources assigned to a job. <strong>Note:</strong> must be executed inside an active job (script or interactive environment). <code>mpiexec</code> is an alternative and preferred on LRZ system</td>
</tr>
<tr>
<td><code>sinfo</code></td>
<td>provide overview of cluster status</td>
</tr>
<tr>
<td><code>scontrol</code></td>
<td>query and modify SLURM state</td>
</tr>
</tbody>
</table>

Queues (SLURM partitions) and their limits

- Batch queues are called partitions in SLURM.
- The allocation granularity is multiples of one node (only complete nodes are allocated and accounted for).
- Scheduling and prioritization is based on a multifactor scheme including wait time, job size, partition, and required quality of service.

The following partitions are available. Check with `sinfo` for more details and special partitions:

<table>
<thead>
<tr>
<th>partition</th>
<th>min-max islands</th>
<th>min-max nodes per job</th>
<th>max usable memory</th>
<th>cores per node</th>
<th>max run time (hours)</th>
<th>max running jobs per user</th>
</tr>
</thead>
<tbody>
<tr>
<td>test</td>
<td>1</td>
<td>1-16</td>
<td>90 GB</td>
<td>48</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(also used for interactive access with <code>salloc</code>)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>micro</td>
<td>1</td>
<td>1-16</td>
<td>90 GB</td>
<td>48</td>
<td>48</td>
<td>20</td>
</tr>
<tr>
<td>general</td>
<td>1</td>
<td>17-792</td>
<td>90 GB</td>
<td>48</td>
<td>48</td>
<td>20</td>
</tr>
<tr>
<td>large</td>
<td>2-6</td>
<td>793-3168 (half of system)</td>
<td>760 GB</td>
<td>48</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>fat</td>
<td>1</td>
<td>1-128</td>
<td>760 GB</td>
<td>48</td>
<td>48</td>
<td>2</td>
</tr>
</tbody>
</table>
srn and mpiexec

With SLURM `srun` command users can spawn any kind of application, process or task inside a job allocation or directly start executing a parallel job (and indirectly ask SLURM to create the appropriate allocation). It can be a shell command, any single-/multi-threaded executable in binary or script format, MPI application or hybrid application with MPI and OpenMP. When no allocation options are defined with srun command the options from sbatch or salloc are inherited.

mpiexec at LRZ internally calls srun to start the application. Therefore the usual behaviour of mpiexec can be expected.

salloc / srun for interactive processing

`salloc` is used the allocated nodes for interactive processing. The options for resource specification in `salloc/srun/sbatch` are the same.
`srun` can be used instead of `mpiexec`. `srun` or `mpiexec` execute on the nodes previously allocated by the `salloc`.
There is no advantage by using `salloc` over `sbatch --partition=test` in terms of wait time.

Example

```bash
$ srun --nodes=2 --ntasks=2 --partition=test
hostname
i01r04c06s08
i01r04c06s08
i01r04c06s08
i01r04c06s10
i01r04c06s10
$ salloc --nodes=2 --ntasks=5 --partition=test
salloc: Granted job allocation 45932
# will be executed on loginode
$ hostname
login01e
#will be executed on the allocated nodes
$ mpiexec -n 5 hostname
i01r04c06s08
i01r04c06s08
i01r04c06s08
i01r04c06s10
i01r04c06s10
$ srun -n 5 hostname
i01r04c06s08
i01r04c06s08
i01r04c06s08
i01r04c06s08
i01r04c06s08
i01r04c06s10
i01r04c06s10
```

sbatch Command / #SBATCH option

Batch job options and resources can be given as command line switches to `sbatch` (in which case they override script-provided values), or they can be embedded into a SLURM job script as a comment line of the form.

Batch Job Examples
General options applicable for all jobs

```bash
#!/bin/bash
# Job Name and Files (also --job-name)
#SBATCH -J jobname
# Output and error (also --output, --error):
#SBATCH -o ./%x.%j.out
#SBATCH -e ./%x.%j.err
# Initial working directory (also --chdir):
#SBATCH -D ./
# Notification and type
#SBATCH --mail-type=END
#SBATCH --mail-user=<youremail>
# Wall clock limit:
#SBATCH --time=24:00:00
#SBATCH --no-requeue
# Setup of execution environment
#SBATCH --export=NONE
#SBATCH --get-user-env
#SBATCH --account=<projectID>
```

Hints and Explanations:

Replacement patterns in filenames:

- `%J`: jobid.stepid of the running job. (e.g. “128.0”)
- `%j`: jobid of the running job.
- `%s`: stepid of the running job.
- `%t`: task identifier (rank) relative to current job.
  this will create a separate IO file per task.
- `%u`: User name.
- `%x`: Job name.
- `%a`: Job array ID

Notification types:

NONE, BEGIN, END, FAIL, REQUEUE

`no-requeue`:
Wether the job should eligible to being requeue or not. When a job is requeued, the batch script is initiated from its beginning. `no-requeue` specifies that the batch job should never be requeued under any circumstances.

Environment:

Do not export the variables of the submitting shell into the job. With would make debugging of errors nearly impossible for LRZ.

`account`:
Resources used by this job are substracted from budget of this project. The billing unit is core-hours. Make sure that you use the right project.

Options for resources and execution (click to expand)

- **MPI without hyperthreading using tasks per node**

  ```bash
  #... (general part)
  #SBATCH --
  partition=general
  #Number of nodes and MPI tasks per node:
  #SBATCH --nodes=128
  #SBATCH --ntasks-per-node=48
  #SBATCH --cpus-per-task=1
  # Run the program:
  mpiexec -n 6144 ./myprog
  ```
MPI with hyperthreading using tasks per node

```bash
#SBATCH --partition=general
#Number of nodes and MPI tasks per node:
#SBATCH --nodes=128
#SBATCH --ntasks-per-node=96
#SBATCH --ntasks-per-core=2
#SBATCH --cpus-per-task=1
#Run the program:
mpiexec -n 6144 ./myprog
```

Hybrid MPI/OpenMP without hyperthreading

```bash
#SBATCH --partition=general
#Number of nodes and MPI tasks per node:
#SBATCH --nodes=128
#SBATCH --ntasks-per-node=8
#SBATCH --cpus-per-task=6
#Run the program:
export OMP_NUM_THREADS=6
export OMP_PLACES=cores
mpiexec -n 1024 ./myprog
```

Hybrid MPI/OpenMP with hyperthreading

```bash
#SBATCH --partition=general
#Number of nodes and MPI tasks per node:
#SBATCH --nodes=128
#SBATCH --ntasks-per-node=8
#SBATCH --cpus-per-task=12
#Run the program:
export OMP_NUM_THREADS=12
export OMP_PLACES=threads
mpiexec -n 1024 ./myprog
```

Resource Specifications:

- `nodes=<minnodes[-maxnodes]>`
  - Request that a minimum of `minnodes` nodes be allocated to this job. A maximum node count may also be specified with `maxnodes`. If only one number is specified, this is used as both the minimum and maximum node count. The default behavior is to allocate enough nodes to satisfy the requirements of the `ntasks` and `cpus-per-task` options.

- `ntasks`:
  - The default is one task per node, but note that the `cpus-per-task` option will change this default.

- `ntasks-per-node`:
  - Request that `ntasks` be invoked on each node. If used with the `ntasks` option, the `ntasks` option will take precedence and the `ntasks-per-node` will be treated as a maximum count of tasks per node. Meant to be used with the `nodes` option.

- `ntasks-per-core`:
  - Request that the maximum `ntasks` be invoked on each core.

- `cpus-per-task`:
  - Without this option, the controller will just try to allocate one core per task.

- `switches=<number>[@waittime hh:mm:ss]`
  - Maximum count of switches desired for the job allocation and optionally the maximum time to wait for that number of switches.

- `array`:
  - Submit a job array, multiple jobs to be executed with identical parameters.

- `mpiexec`:
  - In most cases `mpiexec` can be used without specifying the number of tasks, because this is inherited from the `sbatch` command. Slurm output variables can also be used e.g.,
    ```bash
    mpiexec -n $SLURM_NTASKS ./myprog
    ```
  - If SLURM can detect the number of tasks form its settings it is sufficient to use `mpiexec` without further parameters e.g.,
    ```bash
    mpiexec ./myprog
    ```
Large Job Memory Jobs (>90 GB/node)

```bash
#... (general part)
#SBATCH --partition=fat
#SBATCH --nodes=64
#SBATCH --ntasks-per-node=48
#SBATCH --cpus-per-task=1
Run the program:
mpiexec -n 3072 ./myprog
```

OpenMP Job

```bash
#... (general part)
#SBATCH --partition=general
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=8
export OMP_NUM_THREADS=8
#For pinning threads
export OMP_PLACES=cores
#Spreading for better Memory Balance:
#Thread 0 goes to core 0
#Thread 1 goes to core 6
export OMP_PROC_BIND=spread
mpiexec -n 1024 ./myprog
```

Large MPI Jobs (>792 nodes)

```bash
#... (general part)
#SBATCH --partition=fat
#SBATCH --switches=2@24:00:00
#SBATCH --nodes=1024
#SBATCH --ntasks-per-node=48
#SBATCH --cpus-per-task=1
mpiexec -n 49152 ./myprog
```
Array Jobs

```bash
#!/bin/bash

# Chain of batch jobs with dependencies
NR_OF_JOBS=6
JOB_SCRIPT=./my_batch_script

echo "Submitting chain of $NR_OF_JOBS jobs for batch script $JOB_SCRIPT"

#submit and get JOBID
JOBID=$( sbatch $(JOB_SCRIPT) 2>&1 | awk '{print $NF}' )

I=1
while [ $I -lt $NR_OF_JOBS ]
    do
        JOBID=$( sbatch --dependency=afterok:$JOBID
                  $JOB_SCRIPT 2>&1
                  | awk '{print $NF}' )
        I=$(( $I+1 ))
    done
```

dependency=<dependency_list>
Defer the start of this job until the specified dependencies have been satisfied completed. <dependency_list> is of the form:<type:job_id[:job_id][type:job_id[:job_id]]>

- after:job_id[:jobid...] job can begin execution after the specified jobs have begun execution.
- afterany:job_id[:jobid...] job can begin execution after the specified jobs have terminated.
- afternotok:job_id[:jobid...] job can begin execution after the specified jobs have terminated in some failed state (non-zero exit code, node failure, timed out, etc).
- afterok:job_id[:jobid...] job can begin execution after the specified jobs have successfully executed (ran to completion with an exit code of zero).

Input Environment Variables
Upon startup, sbatch will read and handle the options set in the following environment variables. Note that environment variables will override any options set in a batch script, and command line options will override any environment variables. Some which may be used by you in $HOME/.profile:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBATCH_ACCOUNT</td>
<td>--account</td>
</tr>
<tr>
<td>SBATCH_JOB_NAME</td>
<td>--jobid</td>
</tr>
<tr>
<td>SBATCH_REQUEUE</td>
<td>--requeue</td>
</tr>
<tr>
<td>SBATCH_NOREQUEUE</td>
<td>--no-requeue</td>
</tr>
</tbody>
</table>

Output Environment Variables

The Slurm controller will set the variables in the environment of the batch script:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLURM_JOB_ACCOUNT</td>
<td>Account name associated of the job allocation</td>
</tr>
<tr>
<td>SLURM_JOB_NUM_NODES</td>
<td>Number of nodes.</td>
</tr>
<tr>
<td>SLURM_JOB_NODELIST</td>
<td>To convert the Slurm compressed format into a full list: scontrol show hostname $SLURM_JOB_NODELIST</td>
</tr>
<tr>
<td>SLURM_NTASKS</td>
<td>Number of tasks. Example of usage: mpiexec -n $SLURM_TASKS</td>
</tr>
<tr>
<td>SLURM_NTASKS</td>
<td>These variables are only set if the corresponding sbatch option was given. Example of usage: export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK</td>
</tr>
<tr>
<td>SLURM_PROCID</td>
<td>The MPI rank (or relative process ID) of the current process. Can be used in wrapper scripts if [ $SLURM_PROCID ] then ./master else ./slave fi</td>
</tr>
</tbody>
</table>

Guidelines for resource selection

Processing Mode

- Jobs that only use one or at most a few hardware cores perform serial processing and are not supported on SuperMUC-NG. Use the SuperMUC-Cloud for such purposes.
- Bunches of multiple independent tasks can be bundled into one job, using one or more nodes.

Run time limits

- Please note that all job classes impose a maximum run time limit. It can be adjusted downward for any individual job. Since the scheduler uses a backfill algorithm, the better you specify a realistic runtime limit, the better throughput of your job may be achieved.

Islands/Switches

- When a tree topology is used, this defines the maximum count of switches desired for the job allocation and optionally the maximum time to wait for that number of switches. If Slurm finds an allocation containing more switches than the count specified, the job remains pending until it either finds an allocation with desired (lower) switch count or the time limit expires. If there is no switch count limit, there is no delay in starting the job. This trades off better performance vs. shorter wait time in the queue.

Memory Requirements

- The total memory available in user space for the set of nodes requested by the job must not be exceeded.
- The memory used on each individual node must not be exceeded by all tasks run on that node.
- Applications exist for which the memory usage is unsymmetric. In this case it may become necessary to work with a variable number of tasks per node. One relevant scenario is a master-worker scheme where the master may need an order of magnitude more memory and therefore requires a node of its own, while worker nodes can share a node. LRZ provides the "mixed" partition for using thin and fat nodes concurrently.

Disk and I/O Requirements
The disk and I/O requirements are not controlled by the batch scheduling system, but rely on parallel shared file systems, which provide system-global services with respect to bandwidth - this means that the total I/O bandwidth is shared between all users. The consequence is that all I/O may be significantly slowed down if heavily used by multiple users at the same time, or even - for large scale parallel jobs - by a single user. At present, LRZ cannot make any Quality of Service assurance for I/O bandwidth.

The appropriate usage of the parallel file systems is essential.

Please consult SuperMUC-NG File Systems for more detailed technical information.

Licences

- Some jobs may make use of licensed software, either from the LRZ software application stack, or of software installed in the user's HOME directory. In many cases, the software needs to access a license server because there exist limits on how many instances of the software may run and who may access it at all.
- There is no connection from SuperMUC-NG to the outside. Check with LRZ if you are in need of such licenses.
- LRZ is currently not able to manage license contingents. The reason is that a significant additional effort is required, not only with suitable configuration of SLURM, but also with how the license servers are managed. The situation implies that a job will fail if the usage limit of a licensed software is exceeded when the job starts.

Conversion of scripts from LoadLeveler and other Workload Managers table

- see: List of the most common command, environment variables, and job specification options used by the major workload management systems

SLURM Documentation

- SLURM Workload Manager at LRZ
- Command/option Summary (two pages)
- Documentation for SLURM at SchedMD
- The manual pages slurm(1), sinfo(1), squeue(1), scontrol(1), scancel(1), sview(1)