Example parallel job scripts on the Linux-Cluster

- Introductory remarks
- Shared Memory jobs
  - On the CoolMUC-2 cluster
  - On the CoolMUC-3 cluster
- MPI jobs
  - On the CoolMUC-2 cluster
  - Notes
  - On the CoolMUC-3 cluster
  - Notes
  - General comments
- Special job configurations
  - Job Farming (starting multiple serial jobs on a shared memory system)

Introductory remarks

The job scripts for SLURM partitions are provided as templates which you can adapt for your own settings. In particular, you should account for the following points:

- Some entries are placeholders, which you must replace with correct, user-specific settings. In particular, path specifications and e-Mail addresses must be adapted. Always specify the appropriate directories instead of the names with the three periods in the following examples!
- For recommendations on how to do large-scale I/O please refer to the description of the file systems available on the cluster. It is recommended to keep executables within your HOME file system, in particular for parallel jobs. The example jobs reflect this, assuming that files are opened with relative path names from within the executed program.
- Because you usually have to work with the environment modules package in your batch script, sourcing the file /etc/profile.d/modules.sh is included in the example scripts.

Shared Memory jobs

This job type uses a single shared memory node of the designated SLURM partition. Parallelization can be achieved either via (POSIX) thread programming or directive-based OpenMP programming.

In the following, example scripts for starting an OpenMP program are provided. Please note that these scripts are usually not useful for MPI applications; scripts for such programs are given in subsequent sections.

On the CoolMUC-2 cluster

```
#!/bin/bash
#SBATCH -J job_name
#SBATCH -o ./%x.%j.%N.out
#SBATCH -D ./
#SBATCH --get-user-env
#SBATCH --clusters=cm2
#SBATCH --partition=cm2_tiny
#SBATCH --qos=cm2_tiny
#SBATCH --nodes=1-1
#SBATCH --cpus-per-task=28
#SBATCH --mail-type=end
#SBATCH --mail-user=insert_your_email_here
#SBATCH --export=NONE
#SBATCH --time=08:00:00
module load slurm_setup
export OMP_NUM_THREADS=${SLURM_CPUS_PER_TASK}
./my_openmp_program.exe
```

On the CoolMUC-3 cluster
MPI jobs

For MPI documentation please consult the [MPI page](#) on the LRZ web server. On current cluster systems, Intel MPI is used as the default environment.

MPI jobs may be jobs that use MPI only for parallelization ("MPP-style"), or jobs that combine usage of MPI and OpenMP ("hybrid")

On the CoolMUC-2 cluster

<table>
<thead>
<tr>
<th>CoolMUC-2 MPP-style job</th>
<th>CoolMUC-2 hybrid MPI+OpenMP job</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>#!/bin/bash</code></td>
<td><code>#!/bin/bash</code></td>
</tr>
<tr>
<td><code>#SBATCH --J job_name</code></td>
<td><code>#SBATCH --J job_name</code></td>
</tr>
<tr>
<td><code>#SBATCH -o ./%x.%j.%N.out</code></td>
<td><code>#SBATCH -o ./%x.%j.%N.out</code></td>
</tr>
<tr>
<td><code>#SBATCH --D .</code></td>
<td><code>#SBATCH --D .</code></td>
</tr>
<tr>
<td><code>#SBATCH --get-user-env</code></td>
<td><code>#SBATCH --get-user-env</code></td>
</tr>
<tr>
<td><code>#SBATCH --clusters=cm2</code></td>
<td><code>#SBATCH --clusters=cm2</code></td>
</tr>
<tr>
<td><code>#SBATCH --partition=cm2_std</code></td>
<td><code>#SBATCH --partition=cm2_std</code></td>
</tr>
<tr>
<td><code>#SBATCH --nodes=8</code></td>
<td><code>#SBATCH --nodes=8</code></td>
</tr>
<tr>
<td><code>#SBATCH --ntasks-per-node=28</code></td>
<td><code>#SBATCH --ntasks-per-node=4</code></td>
</tr>
<tr>
<td><code>#SBATCH --mail-type=end</code></td>
<td><code>#SBATCH --mail-type=end</code></td>
</tr>
<tr>
<td><code>#SBATCH --mail-user=insert_your_email_here</code></td>
<td><code>#SBATCH --mail-user=insert_your_email_here</code></td>
</tr>
<tr>
<td><code>#SBATCH --export=NONE</code></td>
<td><code>#SBATCH --export=NONE</code></td>
</tr>
<tr>
<td><code>#SBATCH --time=08:00:00</code></td>
<td><code>#SBATCH --time=08:00:00</code></td>
</tr>
<tr>
<td>module load slurm_setup</td>
<td>module load slurm_setup</td>
</tr>
<tr>
<td><code>export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK</code></td>
<td><code>export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK</code></td>
</tr>
<tr>
<td>mpiexec -n $SLURM_NTASKS ./my_openmp_program.exe</td>
<td>mpiexec -n $SLURM_NTASKS ./my_mpi_program.exe</td>
</tr>
</tbody>
</table>

The example will start 224 MPI tasks distributed over 8 nodes.

- The example will start 32 MPI tasks with 7 threads each. Each node has 28 cores, so 4 tasks are started per host.
- Exploitation of hyperthreads via OpenMP could be done by setting `OMP_NUM_THREADS` to 14

<table>
<thead>
<tr>
<th>CoolMUC-2 MPP-style TINY job</th>
<th>CoolMUC-2 MPP-style LARGE job</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>#!/bin/bash</code></td>
<td></td>
</tr>
<tr>
<td><code>#SBATCH --J job_name</code></td>
<td></td>
</tr>
<tr>
<td><code>#SBATCH -o ./%x.%j.%N.out</code></td>
<td></td>
</tr>
<tr>
<td><code>#SBATCH --D .</code></td>
<td></td>
</tr>
<tr>
<td><code>#SBATCH --get-user-env</code></td>
<td></td>
</tr>
<tr>
<td><code>#SBATCH --clusters=cm2</code></td>
<td></td>
</tr>
<tr>
<td><code>#SBATCH --partition=cm2_std</code></td>
<td><code>#SBATCH --partition=cm2_std</code></td>
</tr>
<tr>
<td><code>#SBATCH --nodes=8</code></td>
<td></td>
</tr>
<tr>
<td><code>#SBATCH --ntasks-per-node=28</code></td>
<td><code>#SBATCH --ntasks-per-node=4</code></td>
</tr>
<tr>
<td><code>#SBATCH --mail-type=end</code></td>
<td></td>
</tr>
<tr>
<td><code>#SBATCH --mail-user=insert_your_email_here</code></td>
<td><code>#SBATCH --mail-user=insert_your_email_here</code></td>
</tr>
<tr>
<td><code>#SBATCH --export=NONE</code></td>
<td></td>
</tr>
<tr>
<td><code>#SBATCH --time=08:00:00</code></td>
<td></td>
</tr>
<tr>
<td>module load slurm_setup</td>
<td>module load slurm_setup</td>
</tr>
<tr>
<td><code>export OMP_NUM_THREADS=7</code></td>
<td><code>export OMP_NUM_THREADS=7</code></td>
</tr>
<tr>
<td>mpiexec -n $SLURM_NTASKS ./my_hybrid_program.exe</td>
<td>mpiexec -n $SLURM_NTASKS ./my_hybrid_program.exe</td>
</tr>
</tbody>
</table>
**Notes**

- A setup as for the hybrid job can also serve to provide more memory per MPI task without using OpenMP (e.g., by setting `OMP_NUM_THREADS=1`). Note that this will leave cores unused!
- Very small jobs (1-2 nodes) must use `cm2_tiny` instead, very large jobs (25-64 nodes) must use `cm2_large`.

**On the CoolMUC-3 cluster**

<table>
<thead>
<tr>
<th>CoolMUC-3 MPP-style job</th>
<th>CoolMUC-3 hybrid job</th>
<th>CoolMUC-3 hybrid job</th>
</tr>
</thead>
<tbody>
<tr>
<td>use physical cores only</td>
<td>use physical cores only</td>
<td>use hyperthreads</td>
</tr>
</tbody>
</table>

```
#!/bin/bash
#SBATCH --nodes=8
#SBATCH --ntasks-per-node=16
#SBATCH --constraint=quad,cache
#SBATCH --time=08:00:00
module load slurm_setup
export OMP_NUM_THREADS=16
mpiexec -n $SLURM_NTASKS ./my_hybrid_program.exe
```

The example will start 128 MPI tasks with 16 threads each; 4 hyperthreads per core are used.

**Notes**
starting more than 64 MPI tasks per KNL node is likely to cause startup failures.
- the --constraint option supplied in some of the scripts below is a suggestion. See the [KNL features documentation](#) for more details.

General comments

- For some software packages, it is also possible to use SLURM's own `srun` command; this will however not work well in all situations for programs compiled against Intel MPI.
- It is also possible to use the --ntasks keyword in combination with --cpus-per-task to configure parallel jobs; this specification replaces the --nodes/--tasks-per-node combination given in the scripts above.

Special job configurations

Job Farming (starting multiple serial jobs on a shared memory system)

Please use this with care! If the serial jobs are imbalanced with respect to run time, this usage pattern can waste CPU resources. At LRZ's discretion, unbalanced jobs may be removed forcibly. The example job script illustrates how to start up multiple serial jobs within a shared memory parallel SLURM script. Note that the various subdirectories subdir_1, ..., subdir_28 must exist and contain the needed input data.

```
#!/bin/bash
#SBATCH -J job_name
#SBATCH -o ./%x.%j.%N.out
#SBATCH -D ./
#SBATCH --get-user-env
#SBATCH --clusters=cm2
#SBATCH --partition=cm2_tiny
#SBATCH --qos=cm2_tiny
#SBATCH --nodes=1
#SBATCH --mail-type=end
#SBATCH --mail-user=insert_your_email_here
#SBATCH --export=NONE
#SBATCH --time=08:00:00
module load slurm_setup
MYPROG=path_to_my_exe/my_serial_program.exe
# Start as many background serial jobs as there are cores available on the node for ((i=1; i<=$SLURM_NTASKS; i++)); do
cd subdir_$i
$MYPROG &
cd ..
done
wait # for completion of background tasks
```

For more complex setups, please read the detailed [job farming document](#) (it is in the SuperMUC-NG section, but for the most part it applies for the Cluster environment as well).