Example parallel job scripts on the Linux-Cluster

- Introductory remarks
- Shared Memory jobs
- MPI jobs
  - On the CooLMUC-2 cluster
  - On the CooLMUC-3 cluster
  - On the Myrinet cluster
- 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.

<table>
<thead>
<tr>
<th>CooLMUC-2 cluster</th>
<th>CooLMUC-3 cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>#!/bin/bash</td>
<td>#!/bin/bash</td>
</tr>
<tr>
<td>#SBATCH --o /home/hpc/.../.../myjob.%j.%N.out</td>
<td>#SBATCH --o /home/hpc/.../.../myjob.%j.%N.out</td>
</tr>
<tr>
<td>#SBATCH --D /home/hpc/.../.../mydir</td>
<td>#SBATCH --D /home/hpc/.../.../mydir</td>
</tr>
<tr>
<td>#SBATCH --J jobname</td>
<td>#SBATCH --J jobname</td>
</tr>
<tr>
<td>#SBATCH --get-user-env</td>
<td>#SBATCH --get-user-env</td>
</tr>
<tr>
<td>#SBATCH --clusters=mpp2</td>
<td>#SBATCH --clusters=mpp3</td>
</tr>
<tr>
<td>#SBATCH --nodes=1-1</td>
<td>#SBATCH --nodes=1-1</td>
</tr>
<tr>
<td>#SBATCH --cpus-per-task=28</td>
<td>#SBATCH --cpus-per-task=64</td>
</tr>
<tr>
<td>#SBATCH --mail-type=end</td>
<td>#SBATCH --mail-type=end</td>
</tr>
<tr>
<td>#SBATCH --mail-user=<a href="mailto:xyz@xyz.de">xyz@xyz.de</a></td>
<td>#SBATCH --mail-user=<a href="mailto:xyz@xyz.de">xyz@xyz.de</a></td>
</tr>
<tr>
<td>#SBATCH --export=NONE</td>
<td>#SBATCH --export=NONE</td>
</tr>
<tr>
<td>source /etc/profile.d/modules.sh</td>
<td>source /etc/profile.d/modules.sh</td>
</tr>
<tr>
<td>cd mydir</td>
<td>cd mydir</td>
</tr>
<tr>
<td>export OMP_NUM_THREADS=28</td>
<td>export OMP_NUM_THREADS=64</td>
</tr>
<tr>
<td># 28 is the maximum reasonable value for CooLMUC-2</td>
<td># 256 is the maximum reasonable value for CooLMUC-3</td>
</tr>
<tr>
<td>./myprog.exe</td>
<td>./myprog.exe</td>
</tr>
</tbody>
</table>

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. The following examples illustrate jobs that use four computational nodes.

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 job</th>
</tr>
</thead>
</table>
#!/bin/bash
#SBATCH --o /home/hpc/.../.../myjob.%j.%N.out
#SBATCH --D /home/hpc/.../.../mydir
#SBATCH --J Jobname
#SBATCH --get-user-env
#SBATCH --clusters=mpp2
#SBATCH --ntasks=112
#SBATCH --cpus-per-task=7
#SBATCH --mail-type=end
#SBATCH --mail-user=xyz@xyz.de
#SBATCH --export=NONE
#SBATCH --time=08:00:00
source /etc/profile.d/modules.sh
cd $SCRATCH/mydata
mpiexec $HOME/exedir/myprog.exe

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!

On the CooLMUC-3 cluster

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.

<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>

<table>
<thead>
<tr>
<th>#!/bin/bash</th>
<th>#!/bin/bash</th>
<th>#!/bin/bash</th>
</tr>
</thead>
<tbody>
<tr>
<td>#SBATCH --o /home/hpc/.../.../myjob.%j.%N.out</td>
<td>#SBATCH --o /home/hpc/.../.../myjob.%j.%N.out</td>
<td>#SBATCH --o /home/hpc/.../.../myjob.%j.%N.out</td>
</tr>
<tr>
<td>#SBATCH --D /home/hpc/.../.../mydir</td>
<td>#SBATCH --D /home/hpc/.../.../mydir</td>
<td>#SBATCH --D /home/hpc/.../.../mydir</td>
</tr>
<tr>
<td>#SBATCH --J Jobname</td>
<td>#SBATCH --J Jobname</td>
<td>#SBATCH --J Jobname</td>
</tr>
<tr>
<td>#SBATCH --get-user-env</td>
<td>#SBATCH --get-user-env</td>
<td>#SBATCH --get-user-env</td>
</tr>
<tr>
<td>#SBATCH --clusters=mpp3</td>
<td>#SBATCH --clusters=mpp3</td>
<td>#SBATCH --clusters=mpp3</td>
</tr>
<tr>
<td>#SBATCH --nodes=8</td>
<td>#SBATCH --nodes=8</td>
<td>#SBATCH --nodes=8</td>
</tr>
<tr>
<td>#SBATCH --ntasks-per-node=64</td>
<td>#SBATCH --ntasks-per-node=16</td>
<td>#SBATCH --ntasks-per-node=16</td>
</tr>
<tr>
<td>#SBATCH --mail-type=end</td>
<td>#SBATCH --mail-type=end</td>
<td>#SBATCH --mail-type=end</td>
</tr>
<tr>
<td>#SBATCH --export=NONE</td>
<td>#SBATCH --export=NONE</td>
<td>#SBATCH --export=NONE</td>
</tr>
<tr>
<td>#SBATCH --time=08:00:00</td>
<td>#SBATCH --time=08:00:00</td>
<td>#SBATCH --time=08:00:00</td>
</tr>
<tr>
<td>source /etc/profile.d/modules.sh</td>
<td>source /etc/profile.d/modules.sh</td>
<td>source /etc/profile.d/modules.sh</td>
</tr>
<tr>
<td>cd $SCRATCH/mydata</td>
<td>cd $SCRATCH/mydata</td>
<td>cd $SCRATCH/mydata</td>
</tr>
<tr>
<td>mpiexec $HOME/exedir/myprog.exe</td>
<td>mpiexec $HOME/exedir/myprog.exe</td>
<td>mpiexec $HOME/exedir/myprog.exe</td>
</tr>
<tr>
<td># will start 512 MPI tasks</td>
<td># will start 128 MPI tasks with 4 threads each</td>
<td># will start 128 MPI tasks with 16 threads each</td>
</tr>
</tbody>
</table>

On the Myrinet cluster

Myrinet 10 GE 32-way systems
Please note:

For some software packages, it is also possible to use SLURM's own \texttt{srun} command; this will however not work well for programs compiled against Intel MPI.

It is also possible to use the --nodes keyword in combination with --tasks-per-node (instead of --ntasks) to configure parallel jobs.

\section*{Special job configurations}

\subsection*{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 MATLAB jobs within a shared memory parallel SLURM script. Note that the various subdirectories subdir\_1, ..., subdir\_8 must exist and contain the needed input data.

\begin{verbatim}
#!/bin/bash
#SBATCH -o /home/hpc/.../.../myjob.%j.%N.out
#SBATCH -D /home/hpc/.../.../mydir
#SBATCH -J Jobname
#SBATCH --get-user-env
#SBATCH --clusters=myri
#SBATCH --partition=myri_large
# MA-TUM users can also specify matum_p or matum_u above
#SBATCH --ntasks=32
#SBATCH --mail-type=end
#SBATCH --mail-user=xyz@xyz.de
#SBATCH --export=NONE
#SBATCH --time=36:00:00
source /etc/profile.d/modules.sh
cd $OPT_TMP/mydata
mpiexec $HOME/exedir/myprog.exe
# At most 32 cores can be used in myri_large.
# More may be possible in the matum\_? queues

Please note:

For some software packages, it is also possible to use SLURM's own \texttt{srun} command; this will however not work well for programs compiled against Intel MPI.

It is also possible to use the --nodes keyword in combination with --tasks-per-node (instead of --ntasks) to configure parallel jobs.

\subsection*{Multi-Serial Example}

\begin{verbatim}
#!/bin/bash
#SBATCH -o /home/hpc/.../.../myjob.%j.%N.out
#SBATCH -D /home/hpc/.../.../mydir
#SBATCH -J Jobname
#SBATCH --get-user-env
#SBATCH --clusters=myri
#SBATCH --partition=myri_std
#SBATCH --nodes=1-1
#SBATCH --mail-type=end
#SBATCH --mail-user=xyz@xyz.de
#SBATCH --export=NONE
#SBATCH --time=08:00:00
source /etc/profile.d/modules.sh

module load matlab

# Start as many background serial jobs as there are cores available on the node:
for i in $(seq 1 8) ; do
cd subdir_${i}
matlab -nodesktop input.m > output.res &
cd ..
done
wait
# for completion of background tasks
\end{verbatim}
\end{verbatim}

Please note:

For some software packages, it is also possible to use SLURM's own \texttt{srun} command; this will however not work well for programs compiled against Intel MPI.

It is also possible to use the --nodes keyword in combination with --tasks-per-node (instead of --ntasks) to configure parallel jobs.