Table of Contents

Information about heterogeneous and modular jobs

Heterogeneous jobs

As of version 17.11 of Slurm, heterogeneous jobs are supported. For example, the user can run:

```
srun --account=deep --partition=dp-cn -N 1 -n 1 hostname : --partition=dp-dam -N 1 -n 1 hostname
dp-cn01
dp-dam01
```

Please notice the `:` separating the definitions for each sub-job of the heterogeneous job. Also, please be aware that it is possible to have more than two sub-jobs in a heterogeneous job.

The user can also request several sets of nodes in a heterogeneous allocation using `salloc`. For example:

```
salloc --partition=dp-cn -N 2 : --partition=dp-dam -N 4
```

In order to submit a heterogeneous job via `sbatch`, the user needs to set the batch script similar to the following one:

```
#!/bin/bash

#SBATCH --job-name=imb_execute_1
#SBATCH --account=deep
#SBATCH --mail-user=
#SBATCH --mail-type=ALL
#SBATCH --output=job.out
#SBATCH --error=job.err
#SBATCH --time=00:02:00

#SBATCH --partition=dp-cn
#SBATCH --nodes=1
#SBATCH --ntasks=12
#SBATCH --ntasks-per-node=12
#SBATCH --cpus-per-task=1

#SBATCH packjob

#SBATCH --partition=dp-dam
#SBATCH --constraint=
#SBATCH --nodes=1
#SBATCH --ntasks=12
#SBATCH --ntasks-per-node=12
#SBATCH --cpus-per-task=1

srun ./app_cn : ./app_dam
```

Here the `packjob` keyword allows to define Slurm parameters for each sub-job of the heterogeneous job. Some Slurm options can be defined once at the beginning of the script and are automatically propagated to all sub-jobs of the heterogeneous job, while some others (i.e. `--nodes` or `--ntasks`) must be defined for each sub-job. You can find a list of the propagated options on the ?Slurm documentation.

When submitting a heterogeneous job with this colon notation using ParaStationMPI, a unique `MPI_COMM_WORLD` is created, spanning across the two partitions. If this is not desired, one can use the `--pack-group` key to submit independent job steps to the different node-groups of a heterogeneous allocation:

```
srun --pack-group=0 ./app_cn ; srun --pack-group=1 ./app_dam
```

Using this configuration implies that inter-communication must be established manually by the applications during run time, if needed.

For more information about heterogeneous jobs please refer to the ?relevant page of the Slurm documentation.

Heterogeneous jobs with MPI communication across modules

In order to establish MPI communication across modules using different interconnect technologies, some special Gateway nodes must be used. On the DEEP-EST system, MPI communication across gateways is needed only between Infiniband and Extoll interconnects.

Attention: Only ParaStation MPI supports MPI communication across gateway nodes.

This is an example job script for setting up an Intel MPI benchmark between a Cluster and a DAM node using a IB ↔ Extoll gateway for MPI communication:

```
#!/bin/bash

# Script to launch IMB PingPong between DAM-CN using 1 Gateway
# Use the gateway allocation provided by SLURM
# Use the packjob feature to launch separately CM and DAM executable


# General configuration of the job
#SBATCH --job-name=modular-imb
#SBATCH --account=deep
#SBATCH --time=00:10:00
#SBATCH --output=modular-imb-%j.out
#SBATCH --error=modular-imb-%j.err

# Configure the gateway daemon
#SBATCH --gw_num=1
#SBATCH --gw_psgwd_per_node=1

# Configure node and process count on the CM
#SBATCH --partition=dp-cn
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1

#SBATCH packjob

# Configure node and process count on the DAM
#SBATCH --partition=dp-dam-ext
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1

# Echo job configuration
echo "DEBUG: SLURM_JOB_NODELIST=$SLURM_JOB_NODELIST"
echo "DEBUG: SLURM_NNODES=$SLURM_NNODES"
echo "DEBUG: SLURM_TASKS_PER_NODE=$SLURM_TASKS_PER_NODE"


# Set the environment to use PS-MPI
module --force purge
module use $OTHERSTAGES
module load Stages/Devel-2019a
module load Intel
module load ParaStationMPI

# Show the hosts we are running on
srun hostname : hostname

# Execute
APP="./IMB-MPI1 Uniband"
srun ${APP}  : ${APP}
```

Attention: During the first part of 2020, only the DAM nodes will have Extoll interconnect (and only the nodes belonging to the `deep-dam-ext` partition will have Extoll active), while the CM and the ESB nodes will be connected via Infiniband. This will change later during the course of the project (expected

end of Summer 2020), when the ESB will be equipped with Extoll connectivity (Infiniband will be removed from the ESB and left only for the CM).

A general description of how the user can request and use gateway nodes is provided at ?this section of the JURECA documentation.

Attention: some information provided on the JURECA documentation do not apply for the DEEP system. In particular:

- as of 31/03/2020, the DEEP system has 2 gateway nodes.
- As of 09/01/2020 the gateway nodes are exclusive to the job requesting them. Given the limited number of gateway nodes available on the system, this may change in the future.
- As of 09/04/2020 the `xenv` utility (necessary on JURECA to load modules for different architectures - Haswell and KNL) is not needed any more on DEEP when using the latest version of ParaStationMPI (currently available in the `Devel-2019a` stage and soon available on the default production stage).