Submitting a SLURM Job Script

Modified on: Fri, Dec 16, 2016 at 12:09 PM

Please review this article about batch computing before trying to submit SLURM jobs to the clusters!
(https://ubccr.freshdesk.com/solution/articles/5000686927)

NOTE: the term "script" is used throughout this documentation to mean an executable file that you create and submit to the job scheduler to run on a node or collection of nodes. The script will include a list of SLURM directives (or commands) to tell the job scheduler what to do. Details and options for these scripts are below.

The job flags are used with SBATCH command. The syntax for the SLURM directive in a script is "#SBATCH <flag>". Some of the flags are used with the srun and salloc commands, as well as the fisbatch wrapper script for interactive jobs.

Resource	Flag Syntax	Description	Notes
partition	--partition=general-compute	Partition is a queue for jobs.	default on ub-hpc is general-compute
qos	--qos=general-compute	**QOS is quality of service value** (https://ubccr.freshdesk.com/solution/articles/13000015771) (limits or priority boost)	default on ub-hpc is general-compute
time	--time=01:00:00	Time limit for the job.	1 hour; default is 72 hours
nodes	--nodes=2	Number of compute nodes for the job.	default is 1; **compute nodes** (http://www.buffalo.edu/ccr/support/research_facilities/general_compute.html)
cpus/cores	--ntasks-per-node=8	Corresponds to number of cores on the compute node.	default is 1
node type	--constraint=IB or --constraint=IB&CPU-E564	Node type feature. IB requests nodes with InfiniBand	default is no node type specified; **compute nodes** (http://www.buffalo.edu/ccr/support/research_facilities/general_compute.html)
resource feature	--gres=gpu:2	Request use of GPUs on compute nodes	default is no feature specified;
memory	--mem=24000	Memory limit per compute node for the job. Do not use with mem-per-cpu flag.	memory in MB; default limit is 3000MB per core
memory	--mem-per-cpu=4000	Per core memory limit. Do not use the mem flag,	memory in MB; default limit is 3000MB per core
account	--account=group-slurm-account	Users may belong to groups or accounts.	default is the user's primary group.
job name	--job-name="hello_test"	Name of job.	default is the JobID
output file	--output=test.out	Name of file for stdout.	default is the JobID
email address	--mail-user=username@buffalo.edu	User's email address	required
email notification	--mail-type=ALL --mail-type=END	When email is sent to user.	omit for no email
access	--exclusive	Exclusive acccess to compute nodes.	default is sharing nodes

Helpful Hints:

Requesting nodes with InfiniBand

MPI jobs should request nodes that have InfiniBand.

```
#SBATCH --constraint=IB
```

Using the Debug Nodes

There are 7 compute nodes in the debug partition.

- 2 8-core nodes
- 4 12-core nodes
- 1 16-core node with 2 GPUs and no IB

The maximum time for a job is 1 hour. Submit a job to this partition using "--partition=debug".

Using Large Memory Nodes

Jobs that require 32-cores per node or over 128GB of memory should be submitted to the largemem partition.
This partition contains only the 32-cores nodes and has higher priority.

```
#SBATCH --partition=largemem --qos=largemem
#SBATCH --ntasks-per-node=32
#SBATCH --mem=250000
```

Using GPUs

```
#SBATCH --partition=gpu --qos=gpu
#SBATCH --gres=gpu:2
```

Requesting Specific Nodes

```
#SBATCH -w, --nodelist=cpn-f16-35,cpn-f16-37
```

Excluding Specific Nodes

```
#SBATCH --exclude=cpn-f16-35,cpn-f16-37
```

Creating a Node List

```
export NODELIST=nodelist.$$
srun -l bash -c 'hostname' | sort | awk '{print $2}' > $NODELIST
cat $NODELIST
```

Create a SLURM script using an editor such as vi or emacs using steps 1 through 3. The script (or file) can be called anything you want but should end in .sh (i.e. myscript.sh). If you are unfamiliar with the UNIX commands to edit files, please **read this article** (https://ubccr.freshdesk.com/support/solutions/articles/5000686120)

Step 1: Resource Specification

```
#!/bin/sh
#SBATCH --partition=general-compute --qos=general-compute
#SBATCH --time=00:15:00
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=8
#SBATCH --constraint=IB
#SBATCH --mem=23000
# Memory per node specification is in MB. It is optional.
# The default limit is 3000MB per core.
#SBATCH --job-name="hello_test"
#SBATCH --output=test-srun.out
#SBATCH --mail-user=username@buffalo.edu
#SBATCH --mail-type=ALL
##SBATCH --requeue
#Specifies that the job will be requeued after a node failure.
#The default is that the job will not be requeued.
```

Step 2: Variables, Paths and Modules

```
echo "SLURM_JOBID="$SLURM_JOBID
echo "SLURM_JOB_NODELIST"=$SLURM_JOB_NODELIST
echo "SLURM_NNODES"=$SLURM_NNODES
echo "SLURMTMPDIR="$SLURMTMPDIR

echo "working directory = "$SLURM_SUBMIT_DIR

module load intel/13.1
module load intel-mpi/4.1.3
module list
ulimit -s unlimited
#
```

Step 3: Launch Application

```
# The initial srun will trigger the SLURM prologue on the compute nodes.
NPROCS=`srun --nodes=${SLURM_NNODES} bash -c 'hostname' |wc -l`
echo NPROCS=$NPROCS
echo "Launch helloworld with srun"
#The PMI library is necessary for srun
export I_MPI_PMI_LIBRARY=/usr/lib64/libpmi.so
srun ./helloworld
#
echo "All Done!"
```

Step 4: Submit job

[cdc@rush:/ifs/user/cdc]$ sbatch slurmHelloWorld-srun
Submitted batch job 91487

Step 5: Check Status of Job

```
[cdc@rush:/ifs/user/cdc]$ squeue -u cdc
  JOBID PARTITION     NAME     USER  ST       TIME  NODES NODELIST(REASON)
  88915 general-c GPU_test      cdc  PD       0:00      1 (Priority)
  91487 general-c hello_te      cdc  PD       0:00      2 (Priority)
[cdc@rush:/ifs/user/cdc]$ squeue -j 91487
  JOBID PARTITION     NAME     USER  ST       TIME  NODES NODELIST(REASON)
  91487 general-c hello_te      cdc  PD       0:00      2 (Priority)
```

R indicates that the job is running.
PD indicates that the job is pending. The job is waiting in the queue.

Cancel a Job

A queued or running job can be cancelled.

```
 [cdc@rush:/ifs/user/cdc]$ squeue -u cdc
  JOBID PARTITION     NAME     USER  ST       TIME  NODES NODELIST(REASON)
  92321     debug     test      cdc   R       0:00      2 d09n29s02,d16n02
  88915 general-c GPU_test      cdc  PD       0:00      1 (Priority)
  91716 general-c hello_te      cdc  PD       0:00      2 (Priority)
  91791 general-c hello_te      cdc  PD       0:00      2 (Priority)
  91792 general-c hello_te      cdc  PD       0:00      2 (Priority)

[cdc@rush:/ifs/user/cdc]$ scancel 92321

[cdc@rush:/ifs/user/cdc]$ squeue -u cdc
  JOBID PARTITION     NAME     USER  ST       TIME  NODES NODELIST(REASON)
  88915 general-c GPU_test      cdc  PD       0:00      1 (Priority)
  91716 general-c hello_te      cdc  PD       0:00      2 (Priority)
  91791 general-c hello_te      cdc  PD       0:00      2 (Priority)
  91792 general-c hello_te      cdc  PD       0:00      2 (Priority)
```