Job Submission in Ibex Cluster

Passant Hafez
HPC Applications Specialist
Supercomputing Core Lab

1) Job Submission
2) Creating a Job Submission Script
3) Job Monitoring

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

1) Job Submission

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

- Resource Manager: SLURM v 19.05.2
- SLURM Resource Allocation Methods:
    - srun
    - salloc
    - sbatch

    (most options are common between them)

- Don’t SSH directly to the compute nodes.

General Notes:

- Requested resources vs waiting time (also --exclusive)
- 2 queues:
  - batch queue (default): up to 14 days (more than that job will be pending FOREVER)
  - debug queue: up to 2 hrs

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

SLURM Options/Directives

-t or --time=(default it's no. of **mins**, format D-HH:MM)
-n or --ntasks=
-N or --nodes=

--partition= or -p (default **batch**)

-J or --job-name= (default is **job script's name**)
-C or --constraint=

--mem (default **2 GB/core**, > **~362 GB** job is marked as large-memory)
--mem-per-cpu=
-e or --error=
-o or --output=
(default both standard output and standard error are directed to a file of the name "**slurm-%j.out**"
and Filename patterns (%J, %j, %A, %a, %x, %N)

--gres=gpu:p100:4
or
--gres=gpu:4 --constraint=[p100]

--mail-type=ALL (equivalent to BEGIN, END, FAIL, REQUEUE, and STAGE_OUT)
TIME_LIMIT
TIME_LIMIT_90 (reached 90% of time limit)
TIME_LIMIT_80 (reached 80% of time limit)
TIME_LIMIT_50 (reached 50% of time limit)
(useful for extensions)

--mail-user=

-x or --exclude
-w or --nodelist=cn603-11-r,cn603-12-l or db202-02-[1-8]

-c or --cpus-per-task= (default is 1)
--ntasks-per-core=
--ntasks-per-socket=
--ntasks-per-node=
--threads-per-core=

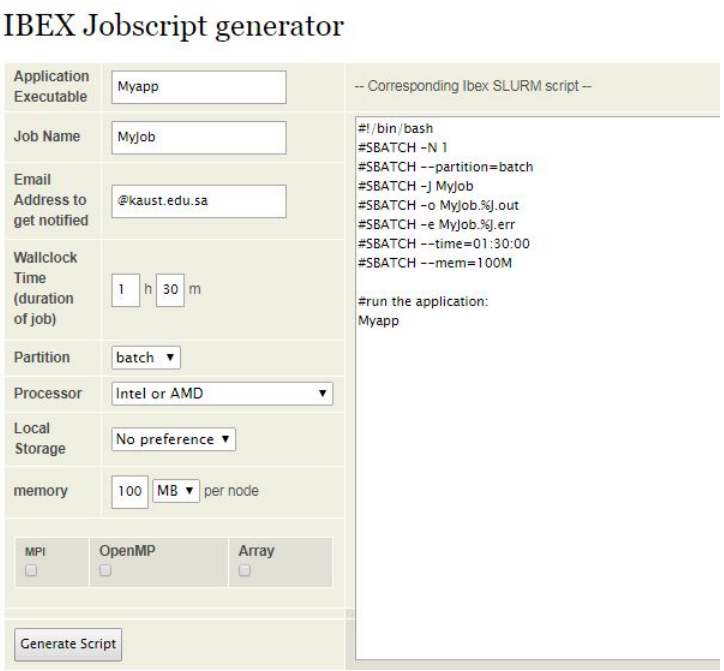
-a or --array=

-d or --dependency=
--no-requeue

2) Creating a Job Submission Script

جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

- Use Ibex Job Script Generator: https://www.hpc.kaust.edu.sa/ibex/job



Notes:

- Make sure you're not loading any conflicting modules before submitting the job, or just add **module purge** before loading other modules in the script.
- Use **#SBATCH** with any of the previously mentioned options

  For example: #SBATCH --time=00:30:00

SLURM Output Environment Variables:

SLURM controller sets some variables in the environment of the batch script, for example:

- **SLURM_JOB_ID**
- **SLURM_JOB_NODELIST**
- **SLURM_ARRAY_JOB_ID** → Array's master job ID number
- **SLURM_ARRAY_TASK_ID** → Job array ID (index) number

Filename Pattern:

sbatch allows for a filename pattern to contain one or more replacement symbols, which are a percent sign "%" followed by a letter, for example:

%j represents jobid of the running job.

%x Job name.

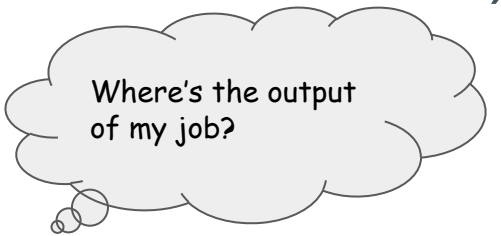
%A Job array's master job allocation number.

%a Job array ID (index) number.

For more patterns check: https://slurm.schedmd.com/sbatch.html#lbAH

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

3) Job Monitoring



- squeue (for running and pending jobs)
- scontrol show job <job_id>
- sacct -j <job_id> (for any job state)
- seff <job_id> (for finished jobs)

3) Job Monitoring

sacct --format

```
Fields available:

Account             AdminComment        AllocCPUS           AllocGRES
AllocNodes          AllocTRES           AssocID             AveCPU
AveCPUFreq          AveDiskRead         AveDiskWrite        AvePages
AveRSS              AveVMSize           BlockID             Cluster
Comment             Constraints         ConsumedEnergy      ConsumedEnergyRaw
CPUTime             CPUTimeRAW          DerivedExitCode     Elapsed
ElapsedRaw          Eligible            End                 ExitCode
Flags               GID                 Group               JobID
JobIDRaw            JobName             Layout              MaxDiskRead
MaxDiskReadNode     MaxDiskReadTask     MaxDiskWrite        MaxDiskWriteNode
MaxDiskWriteTask    MaxPages            MaxPagesNode        MaxPagesTask
MaxRSS              MaxRSSNode          MaxRSSTask          MaxVMSize
MaxVMSizeNode       MaxVMSizeTask       McsLabel            MinCPU
MinCPUNode          MinCPUTask          NCPUS               NNodes
NodeList            NTasks              Priority            Partition
QOS                 QOSRAW              Reason              ReqCPUFreq
ReqCPUFreqMin       ReqCPUFreqMax       ReqCPUFreqGov       ReqCPUS
ReqGRES             ReqMem              ReqNodes            ReqTRES
Reservation         ReservationId       Reserved            ResvCPU
ResvCPURAW          Start               State               Submit
Suspended           SystemCPU           SystemComment       Timelimit
TimelimitRaw        TotalCPU            TRESUsageInAve      TRESUsageInMax
TRESUsageInMaxNode  TRESUsageInMaxTask  TRESUsageInMin      TRESUsageInMinNode
TRESUsageInMinTask  TRESUsageInTot      TRESUsageOutAve     TRESUsageOutMax
TRESUsageOutMaxNode TRESUsageOutMaxTask TRESUsageOutMin     TRESUsageOutMinNode
TRESUsageOutMinTask TRESUsageOutTot     UID                 User
UserCPU             WCKey               WCKeyID             WorkDir
```