

Information Technology Division



## HPCC New User Training

Getting Started on HPCC Resources

(Part 2/2)

Misha Ahmadian *High Performance Computing Center* 

*Summer 2022* 

### Outline



### **Part 2:**

- Resource Allocation and Job Submission with SLURM
- Software builds and installation
- ✤ HPCC Policies
- Getting Help



Information Technology Division

# Resource Allocation and Job Submission with



### Introduction to Slurm



- Simple Linux Utility for Resource Management (SLURM):
  - Primary HPCC scheduler used in RedRaider
  - Main entities:
    - **1.** Nodes: Physical computing resources
    - 2. Partition: A logical set of nodes
    - **3.** Jobs: Allocations of resources assigned to a user for a specified amount of time
    - 4. Job Steps: sets of (possibly parallel) tasks within a job
    - 5. Tasks: Implies the requested/allocated computing resources to process(es) per job or job step (By default, each task refers to <u>1 CPU core</u>)



### Slurm Commands



TEXAS TECH UNIVERSITY Information Technology Division

- Useful Slurm Commands:
  - sinfo:
    - View information about <u>nodes</u> and <u>partitions</u>.
  - squeue:
    - View information about <u>jobs</u> located in partitions.
    - Useful options to filter the output:
      - -u <user>, --user=<user>: Shows the list of jobs or job steps that belong to a specific user
      - --me: Shows the list of jobs or job steps that belongs to you (owner)
      - -p <partition>, --partition=<partition>: Filters the jobs within a partition.
  - srun:
    - submits a job for execution or initiates job steps in real time.
    - srun has the same options as sbatch with a few more. (Please see the man page)
    - srun works similar to the "mpirun" and it can be replaced with "mpirun" as well.



Information Technology Division<sup>-</sup>

#### • sbatch:

- submits a job script for later execution.
  - The submitted job stays in the queue until the requested resources become available.
  - The job submission script is a text file that contains "#SBATCH" hints with sbatch command line options.

```
#!/bin/bash
#SBATCH -J MPI_test
#SBATCH -N 2
#SBATCH -ntasks-per-node=128
#SBATCH -o %x.%j.out
#SBATCH -o %x.%j.out
#SBATCH -e %x.%j.err
#SBATCH -p nocona
module load gcc/10.1.0 openmpi/3.1.6
mpirun ./my mpi
```



Information Technology Division

• Job Submission Script Layout:

Description	SLURM
Set the name for job	-J,job-name= <jobname></jobname>
The name of the standard output file	-o,output= <filename pattern=""></filename>
The name of the standard error file	-e,error= <filename pattern=""></filename>
Define the queue (partition) name	-p,partition= <partition_names></partition_names>
Type of parallel env for job/task allocation	-N,nodes=<# of nodes> ntasks-per-node= <ntasks></ntasks>
Reserve memory per slot	mem-per-cpu= <size[k m g t]></size[k m g t]>
Set the maximum job run time	-t,time= <hh:mm:ss></hh:mm:ss>
Specify the cluster policy for this job	-A,account= <account>   -q,qos</account>



- Select a partition:
  - Partition in Slurm groups physical nodes into a logical set and allows jobs to request for nodes' resources from that partitions.
    - -p, --partition=<partition\_name>

Name	# of Nodes	Туре	Nodes	#Core/Node	#Mem/Node	#Mem/Core	#GPU/node
nocona	240	AMD ROME CPU	cpu-[23-26]-[1-60]	128	503 GB	3.9 GB	N/A
matador	20	Intel/Nvidia V100 GPU	gpu-[20-21]-[1-10]	40	376 GB	9.4 GB	2
gpu-build	1	Intel/Nvidia V100 GPU	gpu-20-11	32	187 GB	5.9 GB	1
toreador	11	AMD/Nvidia A100	gpu-20-[12-15],gpu-21-[11-17]	16	188 GB	11.8 GB	3
quanah	467	Intel Xeon Broadwell	cpu-[1-10]-[*]	36	188 GB	5.3 GB	N/A
xlquanah	16	Intel Xeon Broadwell	cpu-19-[1-16]	36	251GB	6.9 GB	N/A



TEXAS TECH UNIVERSITY Information Technology Division

- Requesting CPU:
  - In Slurm, users must define the following options in their job submissions in order to request for CPU resources:
    - 1. Number of nodes: How many total nodes for the job?
      - -N, --nodes=<number of nodes>
    - 2. Number of tasks per node: (*Recommended*) (By default, each task consumes 1x CPU core)
      - --ntasks-per-nodes=<number of task per node>

OR Number of total tasks: How many task across the nodes?

- -n, --ntasks=<number of tasks>
- Requesting the right number of cores is key to optimizing throughput





TEXAS TECH UNIVERSITY Information Technology Division<sup>-</sup>

- **Requesting Memory:** 
  - One can specify the size of the consumable Memory in two ways in Slurm: •
    - Memory per core (*Recommended*): 1.
      - --mem-per-cpu=<size[M|G]>
    - 2. Memory per node:
      - --mem=<size[M|G]>



If no memory size was specified, Slurm will assign the default memory per core to your job. ٠

Nocona: 4027 MB (3.9 GB) per core Matador: 9639 MB (9.4 GB) per core Quanah: 5370 MB (5.3 GB) per core

- Make sure you won't exceed the total memory per node: ٠
  - -p nocona -N 1 -n 128 --mem-per-cpu=100G 🗙





Information Technology Division

- Requesting Runtime limits:
  - Recommended that you set the max runtime you expect a job will take.
    - -t, --time=<time>
    - <time> can be:
      - *minute*
      - *minute:seconds*
      - hours:minutes:seconds
      - days-hours
      - *days-hours:minutes*
      - *days-hours:minutes:seconds*
    - E.g., --time=24:00:00
  - Please note that there is a 48-hour default time limit per job and exceeding this amount will end up with rejecting your job submission.





TEXAS TECH UNIVERSITY Information Technology Division

- Requesting GPU:
  - GPUs are available by requesting any node in the <u>matador partition</u>.
    - Number of GPUs per node (*Recommended*):
      - --gpus-per-node=[<type>:]<number>
    - Total number of GPUs:
      - -G, --gpus=<# of gpus> (Currently Unavailable on RedRaider cluster)
  - There is only one type of GPU in RedRaider Cluster (v100) and is optional to be specified.
  - It is <u>required</u> to requesting at least **one GPU per node** when submitting a job to Matador.
  - Make sure you do not exceed more than 2 GPUs per node during the job submission.
    - --partition=matador --nodes=2 -gpus-per-node=4 🗙
    - --partition=matador --nodes=2 --gpu-per-node=2





- Accounts on RedRaider Cluster:
  - Accounts, in Slurm, imposes a set of pre-defined resource limits and assigns the usage/fair-share policies to each job.

• -A, --account=<account>

Account -A, account	Default Runtime	Maximum Runtime	CPU/Mem Limit per job	Total # Jobs Per User	# Running Jobs / User	Allowed Partitions	Priority	Special Account
default *	48 hours	48 hours	No limit	2000	No limit	All Partitions	normal	No
xlquanah **	72 hours	14 days	36 cores / 251GB	100	3	xlquanah	normal	Yes
Dedicated resource users	72 hours	No limit	Up to the total available resources	No limit	No limit	Nocona/Quanah	high	Yes

(\*) The system will assign the default Account/QoS if the user does not define them in their job submissions.

(\*\*) Request for special access is required. There should be a valid use case to approve the access to the "xlquanah" partition.





Information Technology Division<sup>-</sup>

- Example of a simple job to submit an MPI program to Slurm:
  - Create a job submission script file (e.g., submit.sh):

```
#!/bin/bash
#SBATCH -J MPI_test
#SBATCH -N 2
#SBATCH -ntasks-per-node=128
#SBATCH -o %x.%j.out
#SBATCH -e %x.%j.err
#SBATCH -p nocona
```

```
module load gcc/10.1.0 openmpi/3.1.6
mpirun ./my_mpi
```

- Submit the job with sbatch:
  - sbatch submit.sh
- Monitor the job with squeue:
  - squeue --me
  - squeue -u <username>
- Cancel the job with scancel:
  - scancel job\_id

**Example:** /lustre/work/examples/nocona

	Jot	b Submission with Slurm			~ະ#2
login-20-25:/slurm_test/ Submitted batch job 1246 login-20-25:/slurm_test/	′mpi∕test\$ sbatch sub 9 ′mpi∕test\$ squeue -u	mit.sh mah			
JOBID PARTITION P	RIORI ST USER	NAME TIME	NODES CPUS	5 NODELIST(REASON)	
12469 test	22153 R mah	Misha_MPI 0:04	2 20	) cpu-23-[26-27]	
login-20-25:/slurm_test/	′mpi/test\$				

### Exercise #2



- 1. Make sure you're already Logged in to the "login.hpcc.ttu.edu" using your eraider account.
- 2. Go to your home directory and copy the following directory into your home directory:
  - \$ cp -r /lustre/work/examples/nocona/training/mpi ~/
- 3. Go into the 'mpi' directory on your home directory:
  - a) List the contents of the directory
  - b) Print the contents of the 'makefile' file
  - c) Load the proper modules for "GCC 10.1.0" and "OpenMPI 4.0.4"
  - d) Run the "make" command to compile the "mpi\_hello\_world.c" code
  - e) Modify the 'mpi\_slurm.sh' file as follows:
    - *i.* Request 1 node from 'nocona' partition with 2 tasks (CPU cores) per node
    - ii. Load the right modules that will work properly with the "mpi\_hello\_world"
  - f) Submit the 'mpi\_slurm.sh' job script
  - g) Check the current status of your jobs
  - h) Check the job's output/error files after it finished.

### Interactive Session



TEXAS TECH UNIVERSITY Information Technology Division

#### • interactive:

- Starts an interactive session/job:
  - interactive -c 2 -p nocona
  - See the interactive -h for all the available options.
- Make sure the prompt changes to cpu-#-#.
- Make sure you run "exit" when you're finished.
- Keep in mind resource/runtime limits apply to interactive based on the selected account.
- The interactive command will forward the X11 if the SSH session was established with -X or -Y.
- Please note that direct SSH to any worker nodes not part of your job is blocked on the RedRaider cluster.

Available modules in Nocona partition	7
login-20-26:\$ interactive -h	
Usage: interactive [-A] [-c] [-p] [-J] [-w] [-g] [-r] [-t] [-h]	
Optional arguments:	
-A: the account name	
-c: number of CPU cores to request (default: 1)	
<ul><li>* -p: the partition name (MANDATORY)</li></ul>	
-N: number of nodes (default: 1)	
-m: Memory per CPU core	
-J: JOD name (default: INTERACTIVE)	
-w: node name	
-y. number of aros per node to request	
-t: The max runtime for the interactive session (limits will be applied)	
-h: show this usage info	
(*) Mandatory options.	

login-20-26:\$

### The 'gpu-build' Partition



- Building and Testing GPU applications:
  - The gpu-build partition contains one Intel/GPU node with 1x Nvidia V100 GPU device, 32x Intel CPU cores and 192 GB RAM, which allows users to:
    - Build their own GPU applications.
    - Test GPU applications and the environment setup before submitting a job to Matador partition.
    - Accessing the Lmod Module environment for GPU compilers/applications.
  - In order to access the 'gpu-build' node, you need to establish an interactive session:
    - \$ interactive -p gpu-build -c 2
  - Limitations:

Partition	Max Runtime	Max CPU per user	Max Mem per user	Max interactive session
	(per job)	(in total)	(in total)	per user
gpu-build	5 hours	6	36006 MB (35 GB)	2

### Debugging a Finished Job



#### • sacct:

- reports accounting information about active or completed jobs or job steps.
  - sacct -j <jobid>
- More filter options are available by checking the -e, --helpformat options of sacct command.
  - sacct -j <jobid> --format=partition,jobid,ntasks,nodelist,maxrss,maxvmsize,exitcode
- When debugging:
  - Check the output and error files
  - Check the output of sacct for:
    - ✓ Memory usage
    - $\checkmark$  Exit code
    - $\checkmark$  Start and end time.

### Current Status of the Job Scheduler



Information Technology Division

• You can check the current status of the Slurm Job Scheduler at this <u>Link</u>.

Texas Tech University						Directory Raiderlink A-Z Index 📴 Trans	late 🔾
High Performance Com	puting Center	Resources 🗸	Operations 🗸	About HPCC 🗸	RedRaider Cluster $\smallsetminus$	HPCC Support	
TTU / High Performance Computing Center		Refreshing a	automatically every 2 minutes 🤁				
		SUMMARY	JOBS NODE	S			
		F	Partition Status				
	NOCONA	MATADOR	ADOR GPU-BUILD QUAN	CC-RESERVED			
			Destition				
			lotal				
	Cluster Allocation		Job Status		Node Status		
	CPU Allocation CPU Idle Other	Pending Running Completing Other	0.00%	Fully Allocated Partially Allocat Idle Drained Draining Down Reserved Other	ted 18.78%		





- 1. Make sure you're already Logged in to the "login.hpcc.ttu.edu" using your eraider account.
  - For this example, please assure your SSH session has the X11 forwarding enabled!
- 2. Make an interactive session to one of the Nocona nodes:
  - Use the 'interactive' command.
  - Request for 1 CPU core from 'nocona' partition.
  - You can use the same temporary reservation as you used in the last exercise.
- 3. Once the worker node was allocated, locate the "MATLAB" module
- 4. Try to run the MATLAB graphical user interface (GUI) on the cluster:
  - cpu-#-#\$ matlab
- 5. Close the MATLAB window to exit the program.
- 6. Exit the interactive session.



Information Technology Division

# Software builds and installation



### HPCC RedRaider Cluster – Overall Look



Information Technology Division<sup>-</sup>



### HPCC RedRaider Cluster - CPU Architectures



Information Technology Division<sup>-</sup>

• Multiple partitions – Multiple architectures:



### Software builds on HPCC Clusters



Information Technology Division

- Modules & compiled code are different on each of the RedRaider partitions!
  - Each CPU architecture may bring a different set of features and instructions.
  - Compiled programs (C/C++/Fortran) need to be re-compiled to match each CPU architecture.
  - E.g., programs that are compiled on **Intel** nodes may not work properly/efficiently on **AMD** nodes.
  - Different Compilers/Math libraries optimize the programs differently on each architecture.

Compiler	AMD ROME	Intel Broadwell	Intel Ivy Bridge	Intel Cascade Lake	Nvidia GPU
GNU/GCC	GCC 10+	GCC 4+	GCC 4+	GCC 10+	GCC 8+
Intel	May work	Optimized	Optimized	Optimized	Intel 19+
AOCC	Optimized	Not Applicable	Not Applicable	Not Applicable	N/A
MKL	May work	Optimized	Optimized	Optimized	MKL 19+
AOCL	Optimized	Not Applicable	Not Applicable	Not Applicable	N/A
CUDA	N/A	N/A	N/A	N/A	<b>CUDA 10+</b>

### Software builds on HPCC Clusters



Information Technology Division

### • Tips and Recommendations:

- 1. Create a separate directory for each CPU architecture, and make a copy from your code/program and place it under each directory:
  - mkdir nocona matador quanah
- 2. Login to the RedRaider login node, and for each CPU architecture make an interactive session to the corresponding worker node:
  - interactive -p nocona -c 10
- 3. Go to the directory of you code that has the same name as the current session's partition:
  - cd nocona
- 4. Load a proper compiler module and recompile your code:
  - module load gcc/10.1.0
- 5. If applicable, add the -O3 optimization flag to all the CFLAGS, CPPFLAGS, CXXFLAGS, FFLAGS.
  - CFLAGS=-03 FFLAGS=-03 make -j 10 all

### Software builds on HPCC Clusters



Information Technology Division

- Tips and Recommendations:
  - 5. We recommend mapping the MPI jobs to the L3-cache memory on **Nocona (AMD)** nodes:
    - mpirun -map-by l3cache ./mpi\_app
  - 6. HPCC will not support Python v2 on Nocona and Matador nodes with CentOS 8. (This rule will be applied to Quanah and Ivy in the near future.)
    - Users can still get Python v2 from Conda (Anaconda/Miniconda)
    - Python 2 is NOT RECEIVING SECURITY UPDATES and should be retired from your workflows ASAP.
  - 7. Python applications (including the applications from Condo repo) will continue working with different architectures without recompiling them.

### Local Python Package Installation



Information Technology Division

- Install a Python package into your home folder:
  - \$ module load intel python
  - \$ pip install --user <package name>
    - Example: pip install --user matplotlib
- Install a local copy of Python using Conda:
  - \$ /lustre/work/examples/InstallPython.sh
  - \$ . \$HOME/conda/etc/profile.d/conda.sh
  - \$ conda activate
  - \$ conda install <package name>
    - Example: conda install biopython



### Local R Package Installation



Information Technology Division

- Install an R package into your home folder:
  - Example (On Quanah Node:)
    - \$ module load intel R
    - \$ R
    - \$ install.packages('<package name>')

Example: install.packages('readr')

- Select a mirror
- The R application will ask if you want to install it locally the first time you do this.





Information Technology Division<sup>-</sup>

# **HPCC** Policies







- Login nodes (login.hpcc.ttu.edu, quanah.hpcc.ttu.edu):
  - No jobs are allowed to run on the login node.
- SSH Access:
  - No direct SSH access allowed to a node(s) if you have no job running on the node(s)
- Software Installation:
  - Software requests are handled on a case-by-case basis
  - Requesting software does not guarantee it will be installed "cluster-wide".
  - May take two or more weeks to complete your request.
- Scratch Purge Policy:
  - Scratch will be purged monthly by removing of all files not accessed within the past year, or sufficient to bring total scratch space usage across all accounts down to 80% of capacity.



Information Technology Division<sup>-</sup>

# Getting Help



### Further Assistance



#### TEXAS TECH UNIVERSITY Information Technology Division

- Visit Our Website:
  - Most up-to-date user guide documents
    - <u>https://www.depts.ttu.edu/hpcc/userguides/index.php</u>
  - Job scheduler and resource allocation status page
    - <u>https://www.depts.ttu.edu/hpcc/status/slurm\_web.php</u>
  - Current status of all HPCC services
    - <u>https://www.depts.ttu.edu/hpcc/status/cachet.php</u>
- Read the documentation!
  - <u>https://slurm.schedmd.com/documentation.html</u>
- Submit a support ticket:
  - Send an email to <u>hpccsupport@ttu.edu</u>

cas Tech University					Directory Raide	link A-Z Index 隆 Transl
High Performance	e Computing Center	Resources $\lor$	Operations 🗸	About HPCC $\smallsetminus$	RedRaider Cluster $\lor$	HPCC Support
10 / High Performance Computing Center	Service Monitoring Pa User Guides Check out our user guides. Getting Help If you need any further assistance ple All systems are operational	i <b>ge for</b> Texas Tec ase take a look at our FAQ a	the bottom of the page or cor	nlact us.		
	© Login Nodes © SLURM Scheduler				•	
	<ul><li>⊕ Nodes Status</li><li>⊕ Storage Servers</li></ul>				•	
	<ul><li>⊙ Networking</li><li>⊙ Other Services</li></ul>				•	
	FAQ					
	I can not login to the HPCC sy	rstem			+	
	I can't log into TTU VPN and S	SSH gateway but I am ab	ole to log into Raiderlink		*	
	I am able to log into both Raid	erlink and TTU VPN but	cannot log into the HPCC s	system	+	

### Quick Reminder



### HPCC Training Courses

- Please check the website for upcoming User Training workshops
  - <u>http://www.depts.ttu.edu/hpcc/about/training.php</u>
- ShortCourse Survey
  - Looking forward to have your feedback on this Training Workshop
    - You will receive a survey in your inbox from TTU ShortCourse
- The PowerPoint slides are available online
  - <u>http://www.depts.ttu.edu/hpcc/about/training.php</u>



# Information Technology Division<sup>\*\*</sup>