Let us learn SLURM

A Hands-on Workshop for SLURM on UCF ARCC Clusters

Armando Fandango
R. Paul wiegand
Nicholas Lucas

Advanced Research Computing Center
Institute for Simulation & Training
University of Central Florida

January 2015

Topics for Today

What is SLURM?

- **S**imple **L**inux **U**tility for **R**esource **M**anagement

What is SLURM?

- **S**imple **L**inux **U**tility for **R**esource **M**anagement
- 2-in-1: resource manager and scheduler, and so replaces Torque & Moab

What is SLURM?

- **S**imple **L**inux **U**tility for **R**esource **M**anagement
- 2-in-1: resource manager and scheduler, and so replaces Torque & Moab
- Specially designed for Linux clusters

What is SLURM?

- **S**imple **L**inux **U**tility for **R**esource **M**anagement
- 2-in-1: resource manager and scheduler, and so replaces Torque & Moab
- Specially designed for Linux clusters
- Designed to be portable, scalable, fault-tolerant, and *simple*

What is SLURM?

- **S**imple **L**inux **U**tility for **R**esource **M**anagement
- 2-in-1: resource manager and scheduler, and so replaces Torque & Moab
- Specially designed for Linux clusters
- Designed to be portable, scalable, fault-tolerant, and *simple*
- Open-Source : `git://github.com/SchedMD/slurm.git`

What is SLURM?

- **S**imple **L**inux **U**tility for **R**esource **M**anagement
- 2-in-1: resource manager and scheduler, and so replaces Torque & Moab
- Specially designed for Linux clusters
- Designed to be portable, scalable, fault-tolerant, and *simple*
- Open-Source : `git://github.com/SchedMD/slurm.git`
- Widely adopted by HPCs around the country (and the world)

What is the change for STOKES Users?

- All basic scheduler commands are different

What is the change for STOKES Users?

- All basic scheduler commands are different
- Submit script directives & syntax must be changed

Yes, you will have to rewrite your submit scripts

What is the change for STOKES Users?

- All basic scheduler commands are different
- Submit script directives & syntax must be changed

  Yes, you will have to rewrite your submit scripts

- That's what we are going to learn today

  SLURM commands and Scripts :-)

So.. Open your computers and login to stokes.

How to run Jobs?

- srun
- sbatch
- salloc

srun

srun

- Try this: `srun hostname`

srun

- Try this: `srun hostname`
- Try this: `srun -N 2 hostname`

srun

- Try this: `srun hostname`
- Try this: `srun -N 2 hostname`
- Try this: `srun -n 2 hostname`

srun

- Try this: `srun hostname`
- Try this: `srun -N 2 hostname`
- Try this: `srun -n 2 hostname`
- Try this: `srun -N 2 -n 2 hostname`

srun

- Try this: `srun hostname`
- Try this: `srun -N 2 hostname`
- Try this: `srun -n 2 hostname`
- Try this: `srun -N 2 -n 2 hostname`
- Try this: `srun -N 2 -n 4 hostname`

srun

- Try this: `srun hostname`
- Try this: `srun -N 2 hostname`
- Try this: `srun -n 2 hostname`
- Try this: `srun -N 2 -n 2 hostname`
- Try this: `srun -N 2 -n 4 hostname`
- `-N` or `--nodes` : number of nodes

srun

- Try this: `srun hostname`
- Try this: `srun -N 2 hostname`
- Try this: `srun -n 2 hostname`
- Try this: `srun -N 2 -n 2 hostname`
- Try this: `srun -N 2 -n 4 hostname`
- `-N` or `--nodes` : number of nodes
- `-n` or `--ntasks` : number of tasks

srun

- Try this: `srun hostname`
- Try this: `srun -N 2 hostname`
- Try this: `srun -n 2 hostname`
- Try this: `srun -N 2 -n 2 hostname`
- Try this: `srun -N 2 -n 4 hostname`
- `-N` or `--nodes` : number of nodes
- `-n` or `--ntasks` : number of tasks
- Best Practice: use either -n or -N with –ntasks-per-node

srun - more options

srun - more options

- `--label` : prepend task-id to output

srun - more options

- `--label` : prepend task-id to output
- Try this: `srun -N 2 -n 4 --label hostname`

srun - more options

- `--label` : prepend task-id to output
- Try this: `srun -N 2 -n 4 --label hostname`
- `--ntasks-per-node` : number of tasks per node

srun - more options

- `--label` : prepend task-id to output
- Try this: `srun -N 2 -n 4 --label hostname`
- `--ntasks-per-node` : number of tasks per node
- `-c` or `--cpus-per-task` : number of cores per task

srun - more options

- `--label` : prepend task-id to output
- Try this: `srun -N 2 -n 4 --label hostname`
- `--ntasks-per-node` : number of tasks per node
- `-c` or `--cpus-per-task` : number of cores per task
- `-w` or `--nodelist` : specific nodes to be allocated

srun - more options

- `--label` : prepend task-id to output
- Try this: `srun -N 2 -n 4 --label hostname`
- `--ntasks-per-node` : number of tasks per node
- `-c` or `--cpus-per-task` : number of cores per task
- `-w` or `--nodelist` : specific nodes to be allocated
- `-x` or `--exclude` : specific nodes not to be allocated

srun - even more options

srun - even more options

- `--time`

srun - even more options

- `--time`
- `--account`

srun - even more options

- `--time`
- `--account`
- `--output`

srun - even more options

- `--time`
- `--account`
- `--output`
- `--job-name`

srun - even more options

- `--time`
- `--account`
- `--output`
- `--job-name`
- `--mem`

srun - even more options

- `--time`
- `--account`
- `--output`
- `--job-name`
- `--mem`
- `--mem-per-cpu`

srun - even more options

- `--time`
- `--account`
- `--output`
- `--job-name`
- `--mem`
- `--mem-per-cpu`
- `man srun` or `srun --help` : to read manual for more options

salloc

salloc

- Try this: `salloc hostname`

salloc

- Try this: `salloc hostname`
- Try this: `salloc`

salloc

- Try this: `salloc hostname`
- Try this: `salloc`
- Try this: `srun hostname`

salloc

- Try this: `salloc hostname`
- Try this: `salloc`
- Try this: `srun hostname`
- salloc is almost same as srun !!!!

salloc

- Try this: `salloc hostname`
- Try this: `salloc`
- Try this: `srun hostname`
- salloc is almost same as srun !!!!
- salloc only reserves nodes and starts a shell

salloc

- Try this: `salloc hostname`
- Try this: `salloc`
- Try this: `srun hostname`
- salloc is almost same as srun !!!!
- salloc only reserves nodes and starts a shell
- srun runs the –number-of-tasks– copies of your app

interactive sessions with srun

interactive sessions with srun

- Try this: `srun --pty bash`

interactive sessions with salloc

- Get allocation and note job-id: `salloc`

interactive sessions with salloc

- Get allocation and note job-id: `salloc`
- Get node: `squeue --job <job-id>`

interactive sessions with salloc

- Get allocation and note job-id: `salloc`
- Get node: `squeue --job <job-id>`
- ssh into node. `ssh ec<node-id>`

interactive sessions with salloc

- Get allocation and note job-id: `salloc`
- Get node: `squeue --job <job-id>`
- ssh into node. `ssh ec<node-id>`
- load module: `module load starccm`

interactive sessions with salloc

- Get allocation and note job-id: `salloc`
- Get node: `squeue --job <job-id>`
- ssh into node. `ssh ec<node-id>`
- load module: `module load starccm`
- run: `starccm+`

interactive sessions with salloc

- Get allocation and note job-id: `salloc`
- Get node: `squeue --job <job-id>`
- ssh into node. `ssh ec<node-id>`
- load module: `module load starccm`
- run: `starccm+`
- ssh into node. `ssh -X ec<node-id>`

interactive sessions with salloc

- Get allocation and note job-id: `salloc`
- Get node: `squeue --job <job-id>`
- ssh into node. `ssh ec<node-id>`
- load module: `module load starccm`
- run: `starccm+`
- ssh into node. `ssh -X ec<node-id>`
- load module: `module load starccm`

interactive sessions with salloc

- Get allocation and note job-id: `salloc`
- Get node: `squeue --job <job-id>`
- ssh into node. `ssh ec<node-id>`
- load module: `module load starccm`
- run: `starccm+`
- ssh into node. `ssh -X ec<node-id>`
- load module: `module load starccm`
- run: `starccm+`

squeue

squeue

- Open two terminal windows side by side and try following commands

squeue

- Open two terminal windows side by side and try following commands
- Terminal Window 1
  - `watch "squeue -t all"`

squeue

- Open two terminal windows side by side and try following commands
- Terminal Window 1
  - `watch "squeue -t all"`
- Terminal Window 2
  - `srun -N 2 sleep 60 &`
  - `srun -N 4 uptime`
  - `srun -N 2 hostname`

squeue

- Open two terminal windows side by side and try following commands
- Terminal Window 1
  - `watch "squeue -t all"`
- Terminal Window 2
  - `srun -N 2 sleep 60 &`
  - `srun -N 4 uptime`
  - `srun -N 2 hostname`
- Do you see the jobs come and go in Window 1 ?

scontrol show

scontrol show

- Try this: `scontrol show partition`

scontrol show

- Try this: `scontrol show partition`
- Try this: `scontrol show job`

scontrol show

- Try this: `scontrol show partition`
- Try this: `scontrol show job`
- Run this: `srun -N 2 sleep 60 &`

scontrol show

- Try this: `scontrol show partition`
- Try this: `scontrol show job`
- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`

scontrol show

- Try this: `scontrol show partition`
- Try this: `scontrol show job`
- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`
- Now run this: `scontrol --detail show job <job-id>`

sinfo

- a "queue" is called *partition* in SLURM

sinfo

- a "queue" is called *partition* in SLURM
- Try this : `sinfo`

sinfo

- a "queue" is called *partition* in SLURM
- Try this : `sinfo`
- Try this: `sinfo -all`

sinfo

- a "queue" is called *partition* in SLURM
- Try this : `sinfo`
- Try this: `sinfo -all`
- Try this: `sinfo -Node`

sacct

sacct

- Try this: `sacct`

sacct

- Try this: `sacct`
- Try this: `sacct -j <job-id>`

sacct

- Try this: `sacct`
- Try this: `sacct -j <job-id>`
- Try this: `sacct --long -j <job-id>`

sshare & sreport

sshare & sreport

- Try this: `sshare`

sshare & sreport

- Try this: `sshare`
- Try this: `sshare -l`

sshare & sreport

- Try this: `sshare`
- Try this: `sshare -l`
- Try this: `sreport -a cluster AccountUtilizationByUser Tree`

scancel

scancel

- Run this: `srun -N 2 sleep 60 &`

scancel

- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`

scancel

- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`
- `scancel <job-id>`

scancel

- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`
- `scancel <job-id>`
- Now run this: `squeue --user=‘whoami‘`

scontrol

scontrol

- Run this: `srun -N 2 sleep 60 &`

scontrol

- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`

scontrol

- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`
- `scontrol hold <job-id>`

scontrol

- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`
- `scontrol hold <job-id>`
- Now run this: `squeue --user=‘whoami‘`

scontrol

- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`
- `scontrol hold <job-id>`
- Now run this: `squeue --user=‘whoami‘`
- `scontrol release <job-id>`

scontrol

- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`
- `scontrol hold <job-id>`
- Now run this: `squeue --user=‘whoami‘`
- `scontrol release <job-id>`
- Now run this: `squeue --user=‘whoami‘`

scontrol

- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`
- `scontrol hold <job-id>`
- Now run this: `squeue --user=‘whoami‘`
- `scontrol release <job-id>`
- Now run this: `squeue --user=‘whoami‘`
- `scontrol requeue <job-id>`

scontrol

- Run this: `srun -N 2 sleep 60 &`
- Get the job id: `squeue --user=‘whoami‘`
- `scontrol hold <job-id>`
- Now run this: `squeue --user=‘whoami‘`
- `scontrol release <job-id>`
- Now run this: `squeue --user=‘whoami‘`
- `scontrol requeue <job-id>`
- Now run this: `squeue --user=‘whoami‘`

Batch Job Script

Its nothing new !!!! Put together what you learnt so far.

```
#!/bin/bash
#SBATCH --account=arcc
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=2
#SBATCH --time=00:10:00
#SBATCH --error=rpwslurm-%J.err
#SBATCH --output=rpwslurm-%J.out
#SBATCH --job-name=PaulSlurmMPIJob
#SBATCH --mail-type=FAIL
#SBATCH --mail-type=BEGIN
#SBATCH --mail-type=END
#SBATCH --mail-user rpwiegand@gmail.com

# Load modules
module load openmpi/openmpi

# Output Information
echo "Slurm nodes: $SLURM_JOB_NODELIST"

# Run your app
mpirun ./hello
```

Getting NodeList

My application requires NODEFILE or MACHINEFILE

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

```
echo $SLURM_JOB_NODELIST
```

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

```
echo $SLURM_JOB_NODELIST
ec[3,6]
```

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

```
echo $SLURM_JOB_NODELIST
ec[3,6]

export NODELIST=‘srun hostname | sort‘ ; echo $NODELIST
```

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

```
echo $SLURM_JOB_NODELIST
ec[3,6]

export NODELIST=‘srun hostname | sort‘ ; echo $NODELIST
ec3 ec3 ec3 ec3 ec6 ec6
```

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

```
echo $SLURM_JOB_NODELIST
ec[3,6]

export NODELIST=‘srun hostname | sort‘ ; echo $NODELIST
ec3 ec3 ec3 ec3 ec6 ec6

export NODELIST=‘srun hostname | sort -u‘ ; echo $NODELIST
```

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

```
echo $SLURM_JOB_NODELIST
ec[3,6]

export NODELIST=‘srun hostname | sort‘ ; echo $NODELIST
ec3 ec3 ec3 ec3 ec6 ec6

export NODELIST=‘srun hostname | sort -u‘ ; echo $NODELIST
ec3 ec6
```

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

```
echo $SLURM_JOB_NODELIST
ec[3,6]

export NODELIST=‘srun hostname | sort‘ ; echo $NODELIST
ec3 ec3 ec3 ec3 ec6 ec6

export NODELIST=‘srun hostname | sort -u‘ ; echo $NODELIST
ec3 ec6

export NODELIST=‘srun hostname | sort -u | sed s/"ec"/"ic"/ ‘ ; echo $NODELIST
```

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

```
echo $SLURM_JOB_NODELIST
ec[3,6]

export NODELIST=‘srun hostname | sort‘ ; echo $NODELIST
ec3 ec3 ec3 ec3 ec6 ec6

export NODELIST=‘srun hostname | sort -u‘ ; echo $NODELIST
ec3 ec6

export NODELIST=‘srun hostname | sort -u | sed s/"ec"/"ic"/ ‘ ; echo $NODELIST
ic3 ic6
```

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

```
echo $SLURM_JOB_NODELIST
ec[3,6]

export NODELIST=‘srun hostname | sort‘ ; echo $NODELIST
ec3 ec3 ec3 ec3 ec6 ec6

export NODELIST=‘srun hostname | sort -u‘ ; echo $NODELIST
ec3 ec6

export NODELIST=‘srun hostname | sort -u | sed s/"ec"/"ic"/ ‘ ; echo $NODELIST
ic3 ic6

export NODELIST=‘srun hostname | sort -u | sed s/"ec"/"ic"/ |
                            tr ’\n’ ’,’ | sed s/",$"//‘ ; echo $NODELIST
```

Getting NodeList

My application requires NODEFILE or MACHINEFILE

- Get nodes: `salloc -N 2 -n 6`

```
echo $SLURM_JOB_NODELIST
ec[3,6]

export NODELIST=‘srun hostname | sort‘ ; echo $NODELIST
ec3 ec3 ec3 ec3 ec6 ec6

export NODELIST=‘srun hostname | sort -u‘ ; echo $NODELIST
ec3 ec6

export NODELIST=‘srun hostname | sort -u | sed s/"ec"/"ic"/ ‘ ; echo $NODELIST
ic3 ic6

export NODELIST=‘srun hostname | sort -u | sed s/"ec"/"ic"/ |
                            tr ’\n’ ’,’ | sed s/",$"//‘ ; echo $NODELIST

ic3,ic6
```

Final Thoughts

We hope this will enable you to start writing your own slurm scripts.

Final Thoughts

We hope this will enable you to start writing your own slurm scripts.

Use the SLURM documentation at `http://slurm.schedmd.com/`.

Final Thoughts

We hope this will enable you to start writing your own slurm scripts.

Use the SLURM documentation at `http://slurm.schedmd.com/`.

Stay Tuned:
Advanced SLURM workshop : Job Arrays, Newton (GPGPU and Phi), Job Sequencing, Task-Core-Node Allocation Strategies, and more.

Questions

Thanks

Thanks for coming today.

email us: `request-stokes@ist.ucf.edu`