PACE SLURM Orientation

An Introduction to the New Phoenix-Slurm Environment

Jeffrey Valdez, M.S.
Adapted from slides by Deepa Phanish, Ph.D.

PACE – Research Computing Facilitation (RCF) Team

Georgia Tech

PACE Team

Partnership for an **A**dvanced **C**omputing **E**nvironment provides faculty participants sustainable leading-edge advanced research computing resources with technical support services, infrastructure, software, and more.

Please attend the PACE Clusters Orientation for more information: https://pace.gatech.edu/content/orientation

Ken Suda, Sr. Applications developer

Michael Brandon, Management Support

Pam Buffington, Director

Craig Moseley, Team lead Customer Exp

Allan Metts, Digital Learning Principal Architect

Matt Guidry, Sr. Systems Support Eng

Greg Beyer, Sr. Systems Support Eng

Philip R. Florio, Systems IT Architect

Paul Manno, RBW-IT Professional

Ruben Lara, Sr. Systems Support Mgr

Dee Womack, Systems support Eng

Eric Harmon, Systems Support Engineer Lead

Peter Wan, Sr. Systems Support Eng

Dan (Ann) Zhou, Storage Architect

Aaron Jezghani, PhD, Scheduler Architect

Semir Sarajlic, Team Lead, ASE

Trever Nightingale, Cloud Architect

Ronald Rahaman, Research Software Eng

Eric Coulter, PhD, CI Integration Architect

Deepa Phanish, PhD, Research Computing Facilitator

Fang (Cherry) Liu, PhD, Lead Research Software Eng

Marian Zvada, Research Computing Facilitator

Jeffrey Valdez, Research Computing Facilitator

Michael Weiner, PhD, Research Computing Facilitator

Grigori Yourganov, PhD, Research Computing Facilitator

IO500

Open Science Grid

NSF

XSEDE Extreme Science and Engineering Discovery Environment

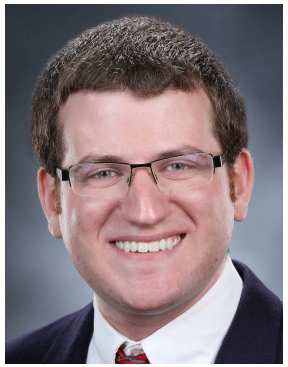
Georgia Tech

PACE-RCF

Meet the **R**esearch **C**omputing **F**acilitation team! We interact with the advanced computing research community at Georgia Tech, respond to a wide range of requests submitted by faculty & student researchers.

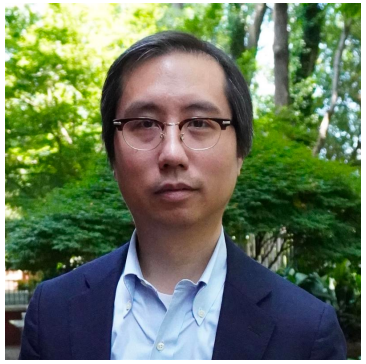
Fang (Cherry) Liu, PhD
RCF Team Lead

"HPC brings you to an amazing parallel universe!"



Michael D. Weiner, PhD
Research Computing Facilitator

"First used HPC in my research, still in awe of its power after over a decade!"



Jeffrey Valdez, MS, MBE
Research Computing Facilitator

"Always happy to help the research community at GT!"

Deepa Phanish, PhD
Research Computing Facilitator

"I love GT! Just jumped over the PACE fence!"

Marian Zvada, MSc, MBA
Research Computing Facilitator

"I like to refine business and enhance PACE customer experience. Always ready to play chess, too!"

Grigori Yourganov, PhD
Research Computing Facilitator

"I am very excited to join the PACE team and help GT researchers find optimal HPC strategies for their research work!"

Georgia Tech

Outline – Migration to SLURM!

Simple **L**inux **U**tility for **R**esource **M**anagement **(SLURM)*** is a resource manager with scheduling logic integrated into it. Phoenix will be the second cluster in PACE's transition from Torque/Moab to Slurm after Hive. We expect the new scheduler to provide improved stability and reliability, offering a better user experience.

- What is changing on Phoenix?
- How to login into Phoenix-Slurm?
- How to use Slurm?
- Where to find help?
- What's next? How can you help us?

*SchedMD is the primary source for the Slurm distribution.

Georgia Tech

Outline – Migration to SLURM!

Simple **L**inux **U**tility for **R**esource **M**anagement **(SLURM)*** is a resource manager with scheduling logic integrated into it. Phoenix will be the second cluster in PACE's transition from Torque/Moab to Slurm after Hive. We expect the new scheduler to provide improved stability and reliability, offering a better user experience.

- What is changing on Phoenix?
- How to login into Phoenix-Slurm?
- How to use Slurm?
- Where to find help?
- What's next? How can you help us?

*SchedMD is the primary source for the Slurm distribution.

Introducing Slurm to Phoenix – What Changes?

- **All Phoenix compute nodes (~1319) have been migrated to our new "Phoenix-Slurm" cluster**
  - Researchers no longer have access to the "Phoenix" (Moab/Torque) cluster after last Maintenance Period (started January 31st, 2023)

- The nodes represent each existing resource pool as before

- **Phoenix-Slurm cluster features a revised application software stack***
  - Researchers installing their own software will need to recompile applications to reflect new MPI and other libraries.
  - Use OnDemand to access Jupyter notebooks and VNC sessions via online portal. Command line interfaces `pace-jupyter-notebook` and `pace-vnc-job` have been retired.

* Review this list of software and let us know if anything is missing!

Georgia Tech

QOS and Resource Pool/Partition – What Changes?

- Quality of Service (QOS) instead of queues
  - inferno (paid) or embers (free backfill)
- Resource Pool/partition targeted and charged based on QOS (if inferno), resources requested, and account (internal vs. external)

cpu-small	cpu-med	cpu-large	cpu-sas	gpu-v100	gpu-rtx6000	cpu-amd	gpu-a100

- No longer charge cpu-sas and gpu-* partitions extra based on memory
- New resources include cpu-amd and gpu-a100 partitions
- More details on Resource Pool/Partition Allocation can be found on Phoenix Cluster Resources
- Service costs for Phoenix-Slurm listed on Rate Study

Introducing Slurm to Phoenix – What Does Not Change?

- Phoenix-Slurm jobs will use charging accounts (paid with inferno QOS and free backfill with embers QOS)
- Storage and quotas remain same as Phoenix
  - home (10GB), project storage (1TB), and scratch (15TB)
- Modularized access of software packages remain the same
  - You will continue using `module spider, avail, list, load, rm and purge`
- Underlying hardware, QOS names (replacing queues - inferno and embers), and their wallclock limits and prioritization are same as before

Outline – Migration to SLURM!

Simple **L**inux **U**tility for **R**esource **M**anagement **(SLURM)*** is a resource manager with scheduling logic integrated into it. Phoenix will be the second cluster in PACE's transition from Torque/Moab to Slurm after Hive. We expect the new scheduler to provide improved stability and reliability, offering a better user experience.

- What is changing on Phoenix?
- How to login into Phoenix-Slurm?
- How to use Slurm?
- Where to find help?
- What's next? How can you help us?

*SchedMD is the primary source for the Slurm distribution.

Phoenix-Slurm Access

- Login using `ssh <GT-username>@login-phoenix-slurm.pace.gatech.edu`

- Specify charging account for jobs using `-A gts-<PI-username>`
  - Run `pace-quota` to find your charging account

```
jvaldez8@login-phoenix-slurm-1:~                                    ⌥⌘3
[jvaldez8@login-phoenix-slurm-1 ~]$ pace-quota

Gathering storage and job accounting information for user: jvaldez8

  ** Please note that the information and display format of this tool **
  ** is subject to change and should *not* be used for scripting.    **


================================================================================
                          Welcome to the Phoenix Cluster!
================================================================================
 * Your Name (as PACE knows it)          :   Jeffrey Nolasco Valdez
 * UserID                                :   657470
 * Username                              :   jvaldez8
 * Your Email (for PACE contact)         :   valdez@gatech.edu

================================================================================
                    Phoenix Storage with Individual User Quota
================================================================================
Filesystem                                    Usage (GB)    Limit      %   File Count      Limit      %
Home:/storage/home/hcodaman1/jvaldez8              0.4      10.0   4.2%         1276    1000000   0.1%
Scratch:/storage/scratch1/0/jvaldez8              19.1   15360.0   0.1%         7307    1000000   0.7%

================================================================================
                    Phoenix Storage with Research Group Quota
================================================================================
Filesystem                                    Usage (GB)    Limit      %   File Count      Limit      %
/storage/coda1/pace-admins                       63665.9      0.0   0.0%     67159487          0   0.0%
/storage/coda1/p-jvaldez8/0                         14.8   1024.0   1.4%       200980          0   0.0%

================================================================================
                          Job Charge Account Balances
================================================================================
Name                          Balance         Reserved        Available
gts-jvaldez8                  68.00           0.00            68.00
phx-pace-staff                Infinity        0.00            Infinity
[jvaldez8@login-phoenix-slurm-1 ~]$
```


Georgia Tech

Outline – Migration to SLURM!

Simple **L**inux **U**tility for **R**esource **M**anagement **(SLURM)*** is a resource manager with scheduling logic integrated into it. Phoenix will be the second cluster in PACE's transition from Torque/Moab to Slurm after Hive. We expect the new scheduler to provide improved stability and reliability, offering a better user experience.

- What is changing on Phoenix?
- How to login into Phoenix-Slurm?
- How to use Slurm?
- Where to find help?
- What's next? How can you help us?

*SchedMD is the primary source for the Slurm distribution.

Phoenix-SLURM – Info commands

If you want…	Here's an example…
To check job status	`squeue -u <GT-username>`
To cancel a job	`scancel <job id>`
Info on completed jobs	`sacct -j <job id>`
Node utilization overview	`pace-check-queue <partition>`
Account Info	`pace-quota`
Review completed jobs	`pace-job-summary <job-id>`

* Refer to this guide for an exhaustive list of Slurm commands and options.

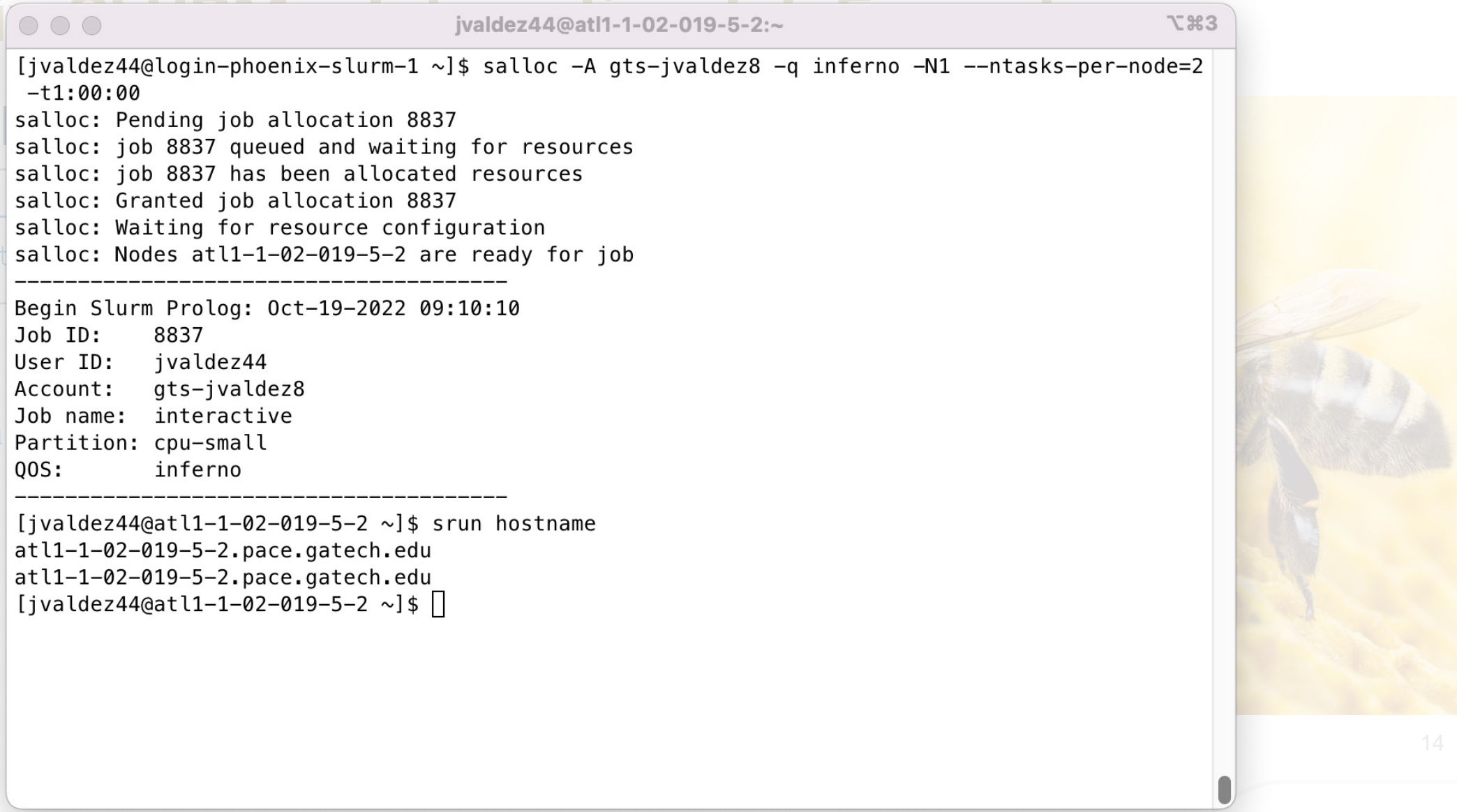
Phoenix-SLURM – Interactive Job Example

- Use salloc to allocate resources

```
salloc -A gts-gburdell3 -q inferno -N1
    --ntasks-per-node=2 -t1:00:00
```

- Use `srun <myprogram>` to execute

`!! salloc` is the recommended workflow by SchedMD because of recent changes with the way the interactive shell is set up in the slurm.conf since 20.11. The `srun --pty` will not work anymore



Georgia Tech

Phoenix-SLURM – Batch Jobs

- Create example.sbatch script

```
#!/bin/bash
#SBATCH -Jexample
#SBATCH --account=gts-gburdell3
#SBATCH -N2 --ntasks-per-node=2
#SBATCH --mem-per-cpu=1G
#SBATCH -t15
#SBATCH -q inferno
#SBATCH -oReport-%j.out
#SBATCH --mail-type=BEGIN,END,FAIL
#SBATCH --mail-user=gburdell3@gatech.edu
cd $SLURM_SUBMIT_DIR
module load anaconda3/2022.05
srun python test.py
```

- Use `sbatch example.sbatch` to submit

jvaldez44@login-phoenix-slurm-1:~/slurm_batch_example ⌥⌘3

```
[jvaldez44@login-phoenix-slurm-1 slurm_batch_example]$ cat Report-8834.out
--------------------------------------
Begin Slurm Prolog: Oct-19-2022 09:10:17
Job ID:    8834
User ID:   jvaldez44
Account:   gts-jvaldez8
Job name:  SlurmPythonExample
Partition: cpu-small
QOS:       inferno
--------------------------------------
Result of 2 ^ 2: 4
Result of 2 ^ 2: 4
Result of 2 ^ 2: 4
Result of 2 ^ 2: 4
--------------------------------------
Begin Slurm Epilog: Oct-19-2022 09:10:21
Job ID:        8834
Array Job ID:  _4294967294
User ID:       jvaldez44
Account:       gts-jvaldez8
Job name:      allocation
Resources:     cpu=4,mem=4G,node=2
Rsrc Used:     cput=00:00:16,vmem=1028K,walltime=00:00:04,mem=0,energy_used=0
Partition:     cpu-small
QOS:           Unknown
Nodes:         atl1-1-02-019-5-2,atl1-1-02-019-12-1
--------------------------------------
[jvaldez44@login-phoenix-slurm-1 slurm_batch_example]$
```


Georgia Tech

MPIs Arrays GPUs

- MPI Jobs using `srun!`

```
mpicc mpi_program.c -o mpi_program
srun {-n4 -c1} mpi_program program_arguments
```

- Array Jobs indexing

```
#SBATCH --array=1-10
#SBATCH -o %A_%a.out
srun <myprogram> data${SLURM_ARRAY_TASK_ID}
```

- GPU Jobs on Nvidia Tesla V100 32GB

```
#SBATCH -N1 --gres=gpu:V100:1
#SBATCH -C V100-32GB
#SBATCH --mem-per-gpu=12G
```

Default Values

--ntasks (-n)	• 1
--ntasks-per-node	• 6 or 12 for GPU (rigid) • 1 for non-GPU
--mem-per-cpu	• 1GB
--cpus-per-task (-c)	• 1
---gres=gpu:1 or –gpus=1	• 1 Nvidia Tesla V100 GPU

Recommendations

Single-threaded: -N1 --ntasks-per-node=1 -c1
Multi-threaded: -N1 --ntasks-per-node=1 -cn
Single-threaded MPI: -Nx --ntasks-per-node=m -c1
Multi-threaded MPI: -Nx --ntasks-per-node=m -cn
(n*m<=24)

Check out our Slurm guide for detailed description and examples.
Check out our Conversion guide for examples on converting PBS scripts to Slurm.

Georgia Tech

Outline – Migration to SLURM!

Simple **L**inux **U**tility for **R**esource **M**anagement **(SLURM)*** is a resource manager with scheduling logic integrated into it. Phoenix will be the second cluster in PACE's transition from Torque/Moab to Slurm after Hive. We expect the new scheduler to provide improved stability and reliability, offering a better user experience.

- What is changing on Phoenix?
- How to login into Phoenix-Slurm?
- How to use Slurm?
- **Where to find help?**
- What's next? How can you help us?

*SchedMD is the primary source for the Slurm distribution.

Getting Help is Easy!

- Visit our documentation to start converting scripts

https://docs.pace.gatech.edu/phoenix_cluster/slurm_guide_phnx/

- Email to open tickets on specific queries

pace-support@oit.gatech.edu

- Come to our Consulting Sessions if you have issues

https://docs.pace.gatech.edu/training/consulting/

Outline – Migration to SLURM!

Simple **L**inux **U**tility for **R**esource **M**anagement **(SLURM)*** is a resource manager with scheduling logic integrated into it. Phoenix will be the first cluster in PACE's transition from Torque/Moab to Slurm. We expect the new scheduler to provide improved stability and reliability, offering a better user experience.

- What is changing on Phoenix?
- How to login into Phoenix-Slurm?
- How to use Slurm?
- Where to find help?
- What's next? How can you help us?

*SchedMD is the primary source for the Slurm distribution.

THANK YOU!

What's next?

- Monitor user experience with the test deployment of Phoenix-Slurm
- Continue migration efforts with Firebird and ICE clusters

Two ways you can help!

- Try Phoenix-Slurm, shift your workflows, and let us know!

  pace-support@oit.gatech.edu

- We also welcome your feedback on this orientation!
  https://b.gatech.edu/3K6NPwQ

Q&A

Link to slides:

https://docs.pace.gatech.edu/training/slurm-orientation/

Survey:

https://b.gatech.edu/3K6NPwQ

Georgia Tech