PACE Clusters Orientation

An Introduction to PACE's Phoenix, Hive, and Firebird High-Performance Computing Research Clusters

PACE – Research Computing Facilitation Team

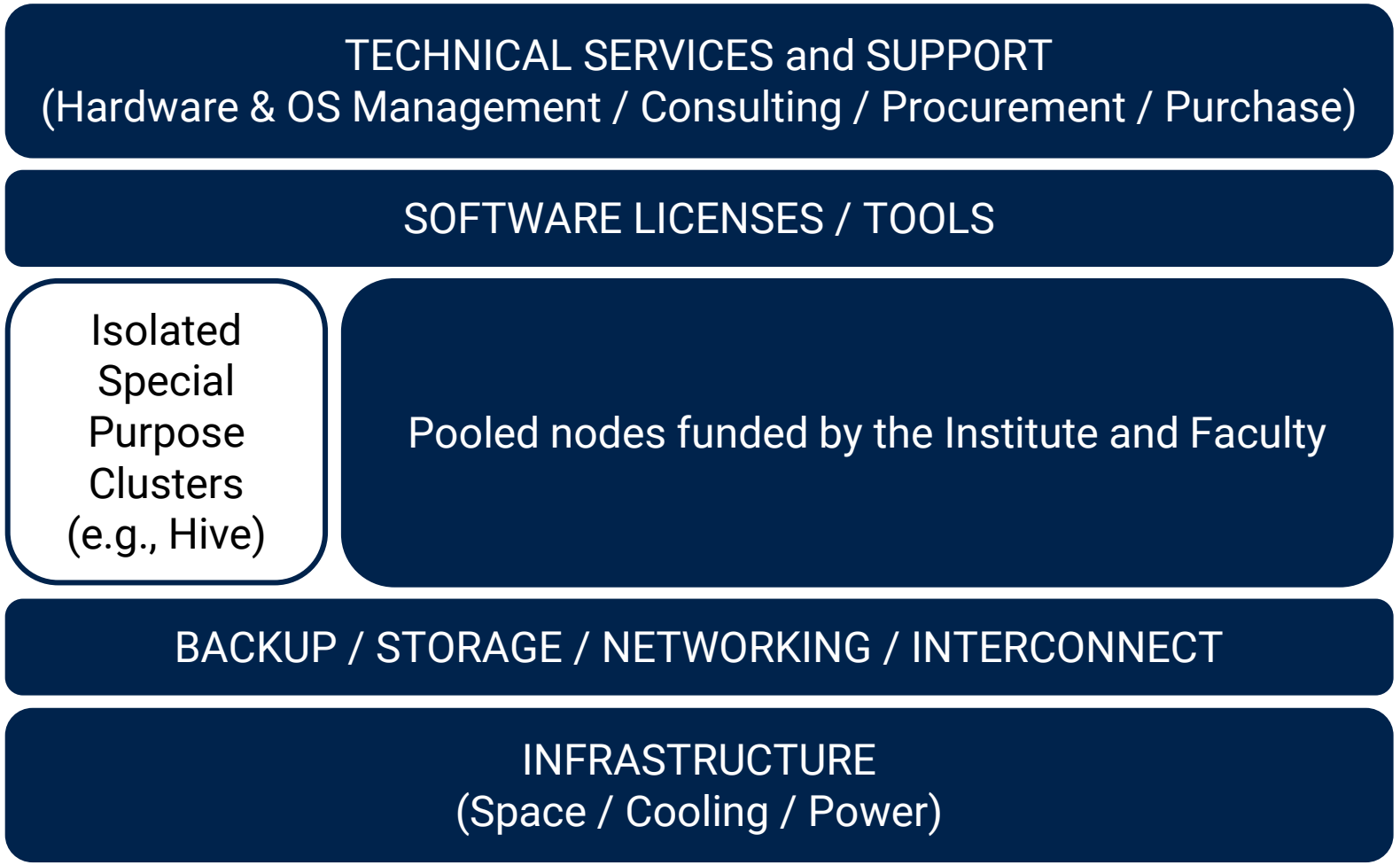
Georgia Tech

Partnership for an Advanced Computing Environment

- PACE provides faculty participants sustainable leading-edge advanced research computing resources with technical support services, infrastructure, software, and more.
- **Free tier provides any GT academic or research faculty the equivalent of 10,000 CPU-hours per month on a 192GB node (or 456 GPU-hours per month on a RTX6000 GPU node) and 1 TB of project storage at no cost on Phoenix**
- Participation and calculator: https://pace.gatech.edu/participation
  - Learn more about paid options for compute & storage
- Virtual tour of Coda datacenter hosting PACE resources: https://pace.gatech.edu/coda-datacenter-360-virtual-tour

Georgia Tech

The Big Picture



Georgia Tech

Getting Help is Easy:

- Email to open tickets:
  - pace-support@oit.gatech.edu
- Visit our documentation: docs.pace.gatech.edu
- Come to a PACE Consulting Session: https://docs.pace.gatech.edu/training/consulting/

Sign up for our hands-on workshops taught by PACE Research Scientists: **Linux 101**, **Linux 102**, **Python 101: Intro to Data Analysis with NumPy**, **Git 101**, **Optimization 101**, and **Applications of Machine Learning** plus **OSG Orientation**

For more detailed questions about your work, attend our weekly **PACE Consulting Session**

Schedules: https://pace.gatech.edu/training

Georgia Tech

Open Science Grid

- The Open Science Grid (OSG) is a network of computing clusters designed for distributed High Throughput Computing (dHTC).
- Available to all researchers at **all US institutions** at no charge via **OSG Connect**, an access point to the OSPool
- PACE can help GT researchers use OSG
- **High Throughput Computing (HTC):** Workflows involving **many** jobs, where many **independent** tasks must be completed
  - Independent tasks can run at different times, on different speed and type processors
  - Generally does not involve parallel programming (one CPU or sometimes a few)
  - The goal is to run lots of computations at once, to reduce time to science

- Some examples of types of research using HTC that could benefit from OSG
  - Text analysis
  - Genomics
  - Parameter sweeps
  - Analysis of many separate input files, like images
  - Model optimization, including Monte Carlo
  - Anything involving many single-core calculations
  - Any workflow that can be modified to fit
- Any field of research – life science, physical science, engineering, social science, humanities, etc.
- Attend **PACE OSG Orientation** to learn more

The Phoenix Cluster

- 3.4 PF peak 64-bit performance (9.1 PF peak 32-bit performance)
  - #463 on the November 2022 Top500 list
- Compute Nodes
  - Intel Cascade Lake CPU nodes with 192GB/384GB/768GB configurations
    - 1.6TB NVMe or 8TB SAS
  - Nvidia Tesla V100 GPU nodes with 2x GPUs/node and 192GB/384GB/768GB configurations
  - Nvidia Quadro Pro RTX6000 GPU nodes with 4x GPUs/node and 384GB/768GB configurations
  - AMD Epyc CPU nodes, some with 2x Nvidia A100 GPUs
- 100 Gbps InfiniBand interconnect fabric
- 5 PB Lustre project storage + 1.5 PB Lustre scatch storage
- #76 on Nov 2022 for the IO500 list and #122 on Nov 2022 for the Graph500 list

TOP 500 The List.

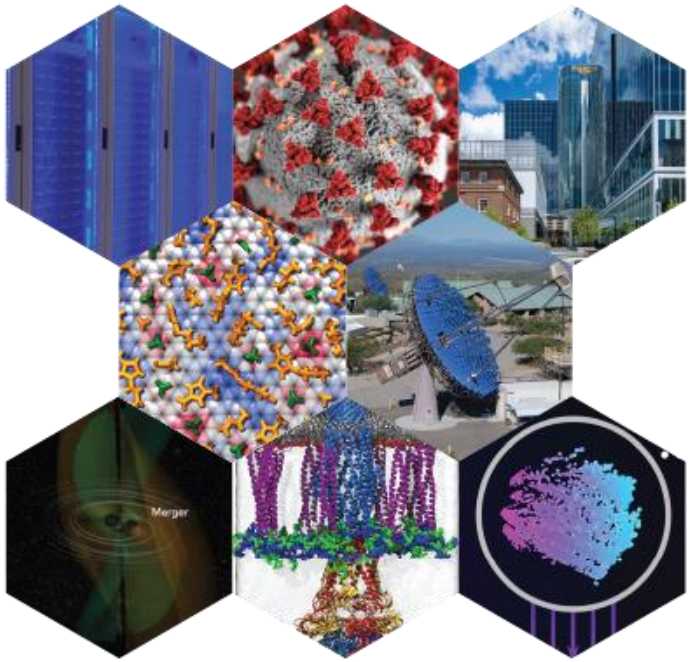
GRAPH 500

IO 500

Georgia Tech

The Hive Cluster

- Funded under an NSF MRI award (1828187) for certain areas of research
- 484 Compute nodes
  - 16 with 4x Tesla V100 GPUs
- 100 Gbps InfiniBand interconnect fabric
- 2.5 PB GPFS project and scratch storage



The Firebird Cluster

- Controlled Unclassified Information (CUI) and export controlled (ITAR) research projects supported
- Nodes similar to those on Phoenix
- Independent storage for each research project

Accessing Clusters

- You need to be connected to Georgia Tech's VPN to access PACE resources
  - For information about VPN access, see http://docs.pace.gatech.edu/gettingStarted/vpn/
- You will need an SSH Client (a.k.a. terminal). Recommended options:
  - Windows: **Powershell** (built-in on Windows 10), **Putty**, or Windows Subsystem for Linux (WSL)
  - MacOS: **Terminal** (built-in; located in Applications -> Utilities)
  - Linux: System-default **terminal** (gnome/KDE)
- Never seen this before? Take PACE's **Linux 101** workshop
- Alternatively, use Open OnDemand in your browser
- SSH access to PACE clusters:

```
ssh <GT_user_ID>@<headnode>.pace.gatech.edu
```

Phoenix - login-phoenix.pace.gatech.edu

Hive - login-hive.pace.gatech.edu

Georgia Tech

Head Nodes vs. Compute Nodes

- **Head Nodes: The machines you use to log in**
  - Shared resource
  - Good for editing, data management, etc.
  - Not good for actual computations or visualizations!
  - Named like “login-phoenix.pace.gatech.edu”

- **Compute Nodes: The machines that run all computations**
  - No direct access by users
  - Allocated per-job by the scheduler
  - Should be used to compile code

Storage and Quotas

- Your data are accessible from all nodes (head and compute nodes)
- Three storage directories:
  - **home**
    - 10GB quota for all users on Phoenix or Firebird and 5GB on Hive, backed up daily
  - **project storage**
    - Phoenix – quota depends on the amount of storage purchased by the PI and is a single quota for your entire research group. Project storage is linked from your home directory as *p-<pi-username>-<number>* (e.g., *p-jdoe4-0*). Backed up daily.
    - Hive – quota is a single quota for your entire research group. Project storage is linked from your home directory as *data*, limited to 2M files or directories per user, backed up daily
  - **scratch**
    - quota is set to 15TB (Phoenix) or 7TB (Hive), and files > 60 days are deleted. Limited to 1M files or directories. Scratch is not backed up!
    - not available on Firebird

Georgia Tech

Data Transfers in/out

- For fast and reliable data migration, please use **Globus** (https://www.globus.org) via these endpoints:
  - Hive – PACE Hive
  - Phoenix – PACE Phoenix
  - http://docs.pace.gatech.edu/storage/globus/
- For small file copies, you may use scp
  - scp -r ~/mylocalstuff <username>@<login-node>.pace.gatech.edu:~/
- Any **SFTP** client will work with PACE. FileZilla is a free FTP tool for Windows, macOS, and Linux
  - Use "<login-node>.pace.gatech.edu" for configuring any of these clients
- Use the "Files" tab in OpenOnDemand web browser application

Georgia Tech

Requesting Local Disk

- Some applications can benefit from a local disk for faster I/O.
- All PACE machines have local disks (NVMEs and SAS)
- Each job automatically creates a directory under /scratch, e.g.:

  `/scratch/21034470`

  which is automatically deleted after job completes! (no cleanup needed)
- Use `${TMPDIR}` to access it, e.g. `cd ${TMPDIR}`
- To guarantee availability of local disk space when requesting a partial node:

  `#SBATCH --tmp=<size>[units, default MB]`

Georgia Tech

PACE Software Stack

- Licensed software packages:
  - Common license: Matlab, Fluent, Abaqus, …
  - Individual license: Vasp, Gaussian, …
- Open source packages and HPC libraries:
  - BLAS, PETSc, NAMD, NetCDF, FFTW, LAMMPS, …
- Compilers:
  - C/C++ & Fortran: GCC and Intel, both with OpenMP support
  - MPI Libraries: MVAPICH and OpenMPI
  - GPU: CUDA
- Scripting Languages: Python, Perl, R, …
- **Anaconda** recommended for Python and package management
  - Build custom environments and install your own Python packages
  - Be sure to set up following instructions before first use
  - https://docs.pace.gatech.edu/software/anacondaEnv/
- Request new software for all via the Software Request Form
- Install your own software!

Georgia Tech

Modules: How to Access Software on PACE

- Painless configuration for software environment and switching between different versions:
- Main commands:
  - `module spider`: Lists all software and its available versions on cluster
  - `module avail`: Lists all available modules that can be loaded with current environment
  - `module list`: Displays all the modules that are currently loaded
  - `module load`: Loads a module to the environment
  - `module rm`: Removes a module from the environment
  - `module purge`: Removes all loaded modules

```
$ module load matlab/r2021a
```

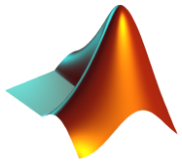
- **Must-Read**: PACE-specific use cases and examples

  http://docs.pace.gatech.edu

Georgia Tech

Open OnDemand

- Access PACE clusters through a web browser
  - Requires VPN
- Especially useful for graphical interactive jobs
- Documentation: https://docs.pace.gatech.edu/ood/guide/
- Access:
  - Phoenix: https://ondemand-phoenix.pace.gatech.edu/
  - Hive: https://ondemand-hive.pace.gatech.edu/



- Supported Apps:
  - **Jupyter**: For Python, Julia, and other languages
  - **Matlab**
  - **VSCode**
  - **Rstudio**
  - **Interactive Shell**
  - **Interactive Desktop**: graphical desktop on a PACE compute node for use with any graphical software

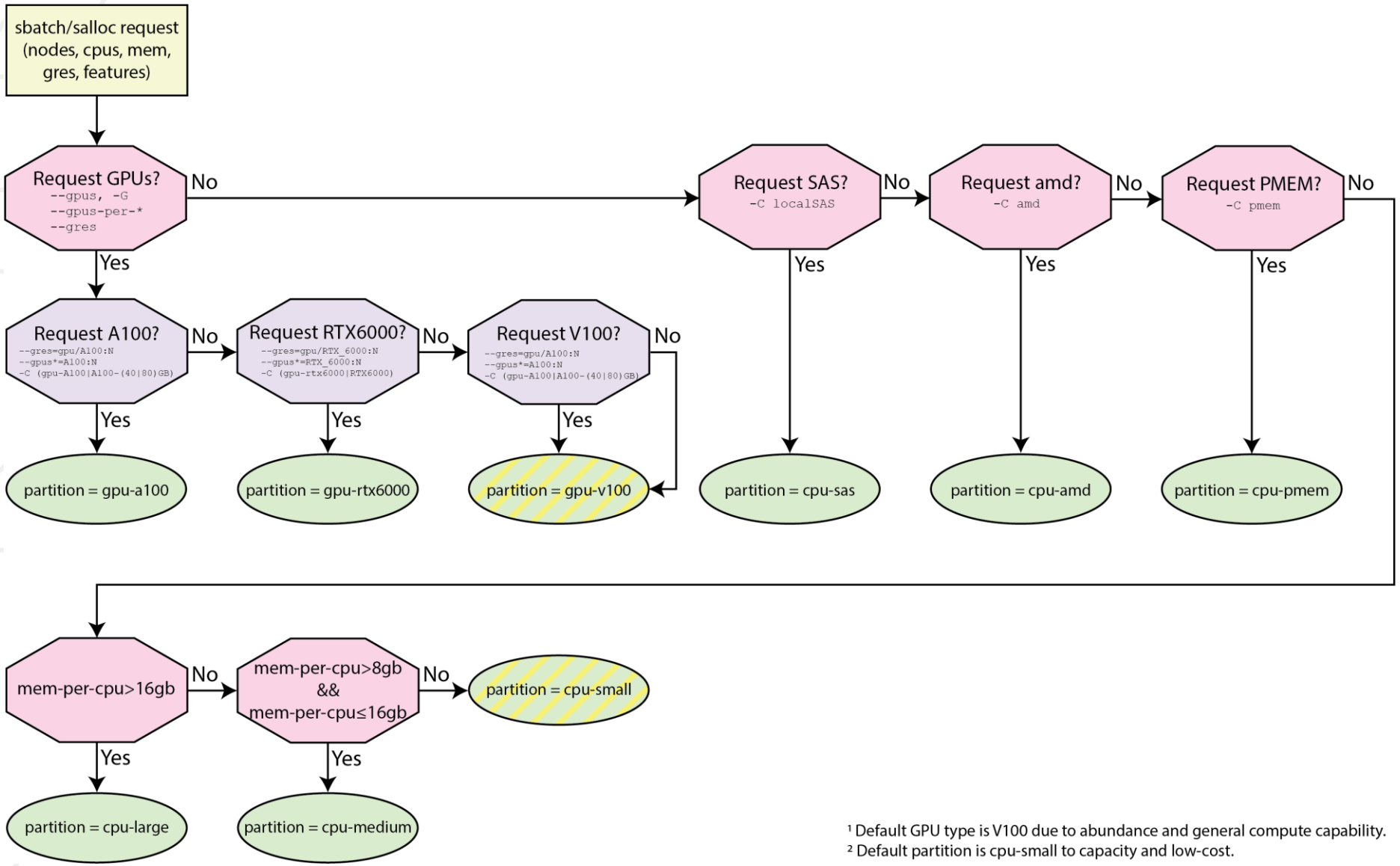
Georgia Tech

Using the Scheduler

- *Note: **Firebird** has not yet migrated to the Slurm scheduler. Join a PACE Consulting Session or visit https://docs.pace.gatech.edu/firebird_cluster/firebird/#job-submissions-on-firebird to learn more about using Firebird's Moab/Torque scheduler.*
- The scheduler allows a fair use of shared resources.
  - Request the needed number of CPUs, memory, walltime, and any specific hardware resources (e.g., GPUs).
  - Allocated resources can be accessed for the requested duration.
- On **Hive**, jobs are submitted to a specific partition based on the resource needs (e.g., *hive*, *hive-gpu*, *hive-nvme*).
  - Visit https://docs.pace.gatech.edu/hive/slurm_guide/ for more details.
  - Tracking accounts are used for recording usage.
- On **Phoenix** and **Firebird**, jobs are submitted to a QOS, each of which spans the entire cluster. Partitions are assigned automatically.
  - Visit https://docs.pace.gatech.edu/phoenix_cluster/slurm_guide_phnx/ for more details.
  - Charge accounts are used for accounting (charging faculty for use).

Georgia Tech

Resource Requests and Phoenix Partitions



spillover between node classes allows waiting jobs to run on underutilized, more capable nodes rather than those requested, requiring no user action, at no additional charge

GPU requests default to **V100** if not specified

More details at https://docs.pace.gatech.edu/phoenix_cluster/slurm_resources_phnx/

[1] Default GPU type is V100 due to abundance and general compute capability.
[2] Default partition is cpu-small to capacity and low-cost.

Georgia Tech

QOS on Phoenix

- The Phoenix cluster has two levels of Quality of Service (QOS):
  - Inferno: the primary QOS [Firebird equivalent: blaze]
    - Base Priority = 250,000
    - Max jobs per user = 500
      - Max eligible jobs per user = 500 (but for GPU jobs, it is 10)
    - Wallclock limit = 21 days
      - Wallclock limit on GPUs = 3 days
    - Jobs are charged to the provided charge account
  - Embers: the backfill QOS [Firebird equivalent: cinders]
    - Base Priority = 0
    - Max jobs per user = 50
      - Max eligible jobs per user = 1
    - Wallclock limit = 8 hours
    - Eligible for preemption after 1 hour
    - Jobs are not charged to an account

Georgia Tech

Charge accounts on Phoenix

- On CPU nodes, charge rates are based on CPU-hours (total number of procs * walltime) allocated
- On GPU nodes, charge rates are based on GPU-hours (total number of GPUs * walltime) allocated for the job, with a fixed ratio of CPUs included per GPU
  - V100 includes 12 CPUs
  - RTX6000 includes 6 CPUs
  - A100 includes 32 AMD CPUs
- **Firebird** follows a similar pattern
- Rates for Phoenix (GEN) and Firebird (CUI): https://docs.pace.gatech.edu/moreInformation/participation/#rate-study

Account Name Syntax	Description	Example
gts-<PI UID>	An institute-sponsored account that provides 10k CPU hours on a base CPU-192GB node, although the credits can be used on any node class. These credits reset on the 1st of the month.	gts-gburdell3
gts-<PI UID>-CODA20	Account for 2020 hardware refresh with the move to Coda	gts-gburdell3-CODA20
gts-<PI UID>-FY20PhaseN	Account for compute resources purchased in FY20	gts-gburdell3-FY20Phase2
gts-<PI UID>-<group>	PI-specific child account for a shared (multi-PI or school-owned) account. This is the account to which jobs should be charged. Depending on the arrangement made by the shared account's managers, there may be a fixed value assigned to each PI, or you may have access to the full shared balance. The visible balance may be a total lifetime value or a value reset each month, depending on the managers' preference.	gts-gburdell3-phys
gts-<PI UID>-<custom>	Account opened in Phoenix on the postpaid billing model. PIs are billed based on actual usage each month and may set limits if preferred.	gts-gburdell3-paid
gts-<PI UID>-<custom>	Account opened in Phoenix on the prepaid billing model for state funds. PIs deposit funds in advance.	gts-gburdell3-startup

Georgia Tech

Checking Accounts, Balances, and Storage with pace-quota

```
[gburdell3@login-phoenix-2 ~]$ pace-quota
================================================================================================================
                                        Welcome to the Phoenix Cluster!
================================================================================================================
 * Your Name (as PACE knows it)          :   George P. Burdell
 * UserID                                :   123456
 * Username                              :   gburdell3
 * Your Email (for PACE contact)         :   burdell@physics.gatech.edu

================================================================================================================
                                Phoenix Storage with Individual User Quota
================================================================================================================
Filesystem                                          Usage (GB)    Limit      %   File Count      Limit      %
Home:/storage/home/hcoda1/2/gburdell3                      0.9     10.0   9.2%          208    1000000   0.0%
Scratch:/storage/scratch1/2/gburdell3                      0.0  15360.0   0.0%            1    1000000   0.0%

================================================================================================================
                                Phoenix Storage with Research Group Quota
================================================================================================================
Filesystem                                          Usage (GB)    Limit      %   File Count      Limit      %
/storage/coda1/p-jdoe43/0                               309.1   1024.0  30.2%       119992          0   0.0%

================================================================================================================
                                        Job Charge Account Balances
================================================================================================================
Name                                    Balance       Reserved      Available
gts-jdoe43-CODA20                     291798.90        3329.35      288469.67
gts-jdoe43-phys                       241264.01          69.44      241194.66
gts-jdoe43                                41.72           0.00          41.72
```

On Hive, tracking accounts are shown.

Georgia Tech

Operation Modes

- **Batch**: Submit & forget. Job waits in the queue until resources become available, then runs on its own on the background.
  - Everything needs to be scripted. Not for codes that require user interaction (e.g., “press ‘y’ to continue”).
  - A script includes resource requirements, environment settings, and tasks.

- **Interactive**: Allows interactive use, no different than remotely using any workstation. Job waits in the queue until resources become available, then logs you into a compute node.
  - Great for debugging and trying out workflows

Batch Jobs

- Create example.sbatch script

```
#!/bin/bash
#SBATCH -Jexample
#SBATCH --account=hive-gburdell3
#SBATCH -N2 --ntasks-per-node=2
#SBATCH --mem-per-cpu=6G
#SBATCH -t1:00:00
#SBATCH -qembers
#SBATCH -oReport-%j.out
#SBATCH --mail-type=BEGIN,END,FAIL
#SBATCH --mail-user=gburdell3@gatech.edu

module load anaconda3/2022.05
python test.py
```

Shell
A name for this job, can be anything
Charge account (Phoenix) or tracking account (Hive)
2 nodes, 2 cores in each
6GB memory per core (24GB total)
15 hours max, after which job is stopped!
QOS "embers" (on Hive, specify a partition with -p instead of a QOS)
name output file (includes STDOUT and STDERR)
when to receive email notifications
email address for notifications

For AMD CPUs:

Add `-C amd`

- Use `sbatch example.sbatch` to submit

Slurm info commands

To check job status	`squeue -u <GT-username>`
To cancel a job	`scancel <job id>`
Info on completed jobs	`sacct -j <job id>`
Review completed jobs	`pace-job-summary <job-id>`

MPI, Arrays, and GPUs

- MPI Jobs using srun!

```
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

```
#SBATCH -N1 --gres=gpu:1
```

Other GPU types (Phoenix):

Add `-C RTX6000` **or** `-C A100-40GB`

Default Values

--ntasks (-n)	• 1
--ntasks-per-node	• 12 for V100 (fixed) • 6 on Hive • 6 for RTX6000 (fixed) • 32 for A100 (fixed) • 1 for non-GPU
--mem-per-cpu	• 1GB
--cpus-per-task (-c)	• 1

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)

Georgia Tech

Interactive Command-Line Jobs

- Also available in OnDemand as “Interactive Shell”
- Same Slurm commands, but this time on the command line with **salloc**:

Hive: `salloc -A hive-gburdell3 -phive -N1 --ntasks-per-node=2 -t1:00:00`
Phoenix: `salloc -A gts-gburdell3 -qinferno -N1 --ntasks-per-node=2 -t1:00:00`

- User waits in real-time until space is assigned
- The scheduler logs the user onto a compute node when the resources become available
- Use srun inside interactive job to execute
- Session is terminated:
  - The user exits
  - The terminal is closed (or internet connection is lost)
  - The walltime is exceeded

Georgia Tech

Checking Partition Status

Use the command with a specific partition to see nodes and utilization of that partition

```
[gburdell3@login-phoenix-slurm-1 ~]$ pace-check-queue cpu-large
=== cpu-large Partition Summary ===
          Last Update                          : 10/17/2022 18:15:01
          Next Maintenance Start               : 11/02/2022 06:00:00
          Number of Nodes (Accepting Jobs/Total) : 25/25 (100.00%)
          Number of Cores (Used/Total)         : 12/600 (2.00%)
          Amount of Memory (Used/Total) (GB)   : 80/3400 (2.35%)
```

Hostname	CPUs Ded/Tot	PhyCPU Load %	GPUs Ded/Tot	Mem (GB) Use/Ded/Tot	Mem % Util.	Loc Drv (GB) Use/Ded/Tot	Loc Drv % Util.	Accepting Jobs?
atl1-1-02-003-30-2	12/24	18.00	0/	20/ 80/ 753	10.00	-/ -/1360	0.00	Yes
atl1-1-03-004-6-1	0/24	2.00	0/	11/ 0/ 753	1.00	-/ -/1360	0.00	Yes
atl1-1-03-004-6-2	0/24	1.00	0/	11/ 0/ 753	1.00	-/ -/1360	0.00	Yes
atl1-1-03-004-7-1	0/24	2.00	0/	11/ 0/ 753	1.00	-/ -/1360	0.00	Yes
atl1-1-03-004-7-2	0/24	2.00	0/	11/ 0/ 753	1.00	-/ -/1360	0.00	Yes
atl1-1-03-004-8-2	0/24	3.00	0/	11/ 0/ 753	1.00	-/ -/1360	0.00	Yes

…

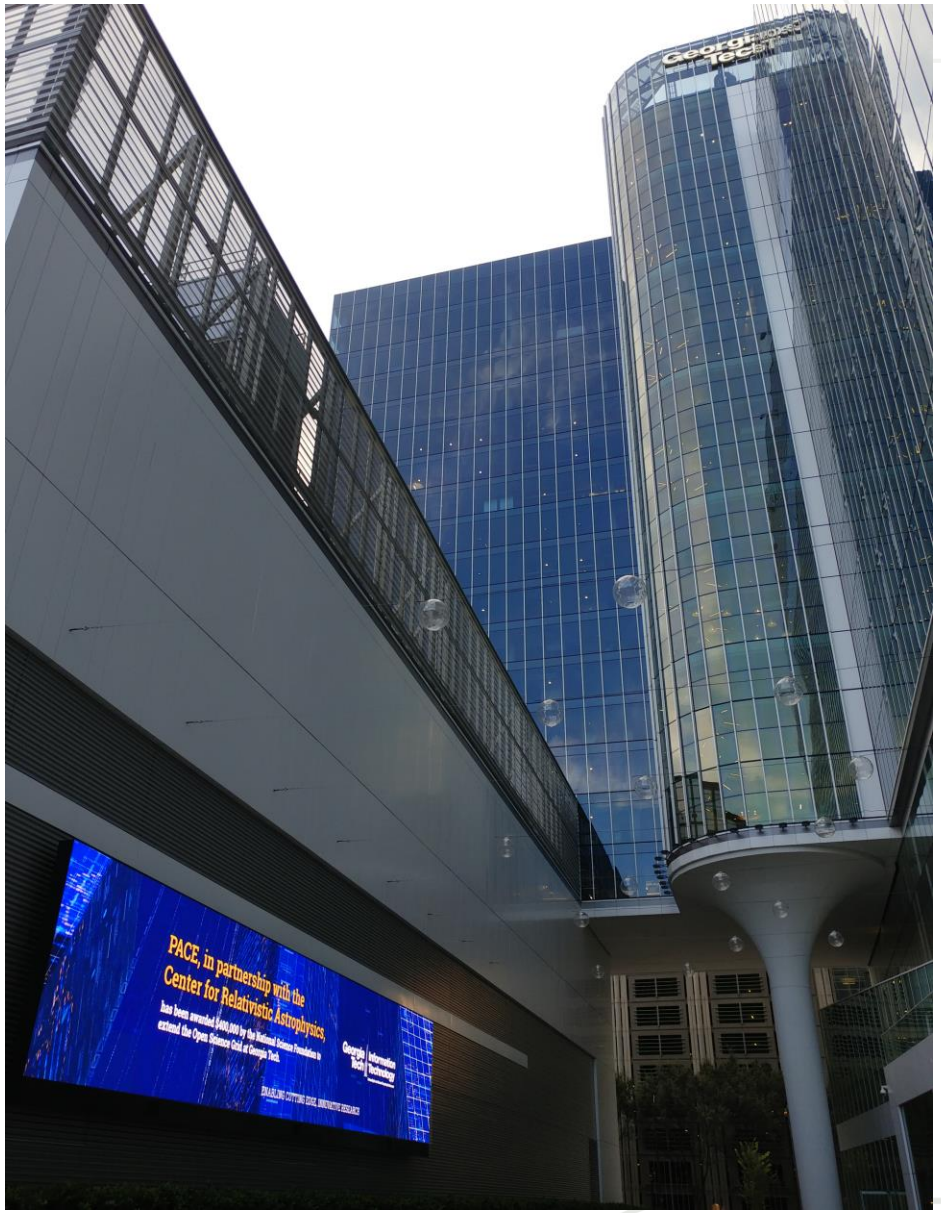
Georgia Tech

PACE Email Lists & Blog

- User Lists:
  - All users:
    - pace-availability (non-optional subscription)
    - pace-discuss (optional unsubscription)
  - Cluster-specific lists: (non-optional subscription)
    - pace-hive
    - pace-phoenix
    - pace-firebird

- PACE blog: http://blog.pace.gatech.edu

- PACE website: http://pace.gatech.edu
  - See next maintenance period



Georgia Tech

THANK YOU!

We welcome your feedback!

https://b.gatech.edu/2LerSxZ

Link to slides: http://www.pace.gatech.edu/content/orientation

PACE Documentation: http://docs.pace.gatech.edu

PACE Consulting Sessions: Visit https://docs.pace.gatech.edu/training/consulting/ for schedule & Zoom links for weekly sessions

Welcome to PACE!

Georgia Tech