PACE Clusters Orientation

PACE – Outreach and Faculty Interaction Team
A Partnership for an Advanced Computing Environment

- PACE is a collaboration between Georgia Tech Faculty and the OIT that provides faculty participants a sustainable leading-edge advanced research computing resources with technical support services.

- Infrastructure
- Commodity Services
- Technical Services & Support
- Software
- Access to up to 10,000 CPU hours/month and 1TB of storage on the Phoenix Cluster at no cost

Sign up for our hands-on workshops: Linux 101, Linux 102, Python 101: Intro to Data Analysis with NumPy, Optimization 101, Applications to Machine Learning, Introduction to ML, and Introduction to Deep Learning

For more detailed questions about your work attend our biweekly PACE Consulting Session
The New Cost Model and Participation in PACE

• PACE is transitioning from a service that purchases nodes with equipment funds, to a service which operates as a Cost Center
  • Also, a free tier that provides any PI (academic faculty or research faculty) the equivalent of 10,000 CPU-hours (per month) on a 192GB compute node and 1 TB of project storage at no cost on the Phoenix cluster
    • The Phoenix cluster is a heterogeneous cluster with 31,104 cores that delivers over 1.8 LINPAC petaflops of computing power, and it is ranked at #277 in the November 2020 Top500 list
    • [https://pace.gatech.edu/update-gts-research-cyberinfrastructure-cost-model](https://pace.gatech.edu/update-gts-research-cyberinfrastructure-cost-model)

• Participation in PACE and the participation calculator
  • [http://docs.pace.gatech.edu/moreInformation/participation](http://docs.pace.gatech.edu/moreInformation/participation)

• Virtual tour of Coda datacenter hosting PACE resources can be accessed at: [https://pace.gatech.edu/coda-datacenter-360-virtual-tour](https://pace.gatech.edu/coda-datacenter-360-virtual-tour)
The Big Picture

- TECHNICAL SERVICES and SUPPORT
  (Hardware & OS Management / Consulting / Procurement / Purchase)

- SOFTWARE LICENSES / TOOLS

- BACKUP / STORAGE / NETWORKING / INTERCONNECT

- INFRASTRUCTURE
  (Space / Cooling / Power)

Isolated Special Purpose Clusters (e.g., Hive)
Getting Help is Easy:

• Email to open tickets:
  pace-support@oit.gatech.edu

• Preferred - run this script on the cluster
  pace-support.sh
Accessing Clusters

• You will need an SSH Client (a.k.a. terminal):
  - Windows: MobaXterm, PuTTy, **Powershell**, WSL, Xming (free), X-win32 (software.oit.gatech.edu)
  - MacOS: iTERM2, Terminal, XQuartz
  - Linux: System-default terminal (gnome/KDE)

• SSH access to PACE clusters:
  ```
  ssh -Y <GT_user_ID>@<headnode>.pace.gatech.edu
  ```
  Phoenix - `login-phoenix.pace.gatech.edu`
  Hive - `login-hive.pace.gatech.edu`

• You need to be connected to VPN to access the resources
  For information VPN access, see:
  http://faq.oit.gatech.edu/search/node/vpn
The Phoenix Cluster

• A heterogeneous cluster featuring Intel’s Cascade Lake processors and Nvidia’s Volta and Quadro Pro 6000 GPUs with the following node configuration available:
  • 24-core, 192GB RAM
  • 24-core, 384GB RAM
  • 24-core, 768GB RAM
  • 24-core, 384GB RAM, SAS drives (8TB)
  • 24-core, 768GB RAM, SAS drives (8TB)
  • 24-core, 192GB RAM, V100 GPUs
  • 24-core, 384GB RAM, V100 GPUs
  • 24-core, 768GB RAM, V100 GPUs
  • 24-core, 384GB RAM, RTX6000 GPUs
  • 24-core, 768GB RAM, RTX6000 GPUs
PACE Queues and the Scheduler

- A queue can be a small part of a cluster (hive-gpu), or span the entire cluster (e.g., hive-all), as on Hive cluster. On Phoenix inferno and embers queues span the entire Phoenix cluster.
- MAM does the accounting on the Phoenix cluster
- The scheduler (Moab) allows a fair use of shared resources within a queue (dynamic priorities, limits on walltime, CPU and RAM)
Queues on the Phoenix Cluster

- The Phoenix cluster has two queues unlike the prior PACE managed resources, and QOS is used to navigate jobs to appropriate resources:
  - Inferno: the primary queue
    - 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 account
  - Embers: the backfill queue
    - 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
  - For details refer to [http://docs.pace.gatech.edu/phoenix_cluster/submit_jobs_phnx/](http://docs.pace.gatech.edu/phoenix_cluster/submit_jobs_phnx/)
- The Hive cluster's queues are detailed at the provided link [http://docs.pace.gatech.edu/hive/resources/](http://docs.pace.gatech.edu/hive/resources/)
NOTE: Wait hours per job by queue presented are for October - November 2020 and are meant to serve as a guide. This metric is based on the queue utilization, which varies over time.
Steps to Use the PACE Clusters

1. Connect to GT VPN
2. Login to headnode using SSH
3. Submit jobs to Moab using “qsub” command
4. Moab decides which compute node to use
5. Your job starts on the node that Moab picks

Summary

- login-phoenix
- login-hive
- login-testflight-coda
Head Nodes vs. Compute Nodes

• **Head Nodes**: The machines you use to log in
  - Good for compiling, editing, debugging, 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
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 and 5GB on Hive), backed up daily
  • **project storage**
    • The Phoenix cluster – quota depends on the amount of storage purchased by the PI. Project storage is no longer referred to as **data**, the new naming schema will be *p-<pi-username>-<number>* (e.g., *p-jdoe4-0*). Backed up daily.
    • **Note:** Project storage quotas on the Phoenix cluster will be enforced in early 2021 after all user data migrations and storage audit is complete
  • The Hive cluster - project storage is referred to as **data**, limited to 2M files or directories, backed up daily
  • **scratch**
    • The Phoenix cluster – quota is set to 15TB, and files > 60 days are deleted. Limited to 1M files or directories. Scratch is **not backed up**!
    • The Hive cluster – quota is set to 7TB, and files > 60 days are deleted. Limited to 1M files or directories. Scratch is **not backed up**!
Data Transfers in/out

• For small file copies, you may use `scp` to move files from computer to the Phoenix cluster or other clusters
  • `scp -r ~/mylocalstuff <username>@<login-node>.pace.gatech.edu:~/
• For fast and reliable data migration, please use Globus ([https://www.globus.org](https://www.globus.org)) service along with these endpoints for the following clusters:
  • Hive – PACE Hive
  • Phoenix – PACE Phoenix
  • [http://docs.pace.gatech.edu/storage/globus/#using-globus-phoenix-pace-phoenix](http://docs.pace.gatech.edu/storage/globus/#using-globus-phoenix-pace-phoenix)
• 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
Running Jobs: Overview

• Users make requests to **Moab scheduler** specifying the requirements of the code:
  • The number of Nodes and/or Cores per node.
  • The total Memory or Memory-per-core.
  • An estimated Runtime (walltime, not CPU time)
  • Specific hardware resources, e.g. GPU

http://docs.pace.gatech.edu/faqs/pbs/

• Allocated resources can only be used by the user for the duration of requested walltime. This is the only time users can directly login to compute nodes.
Operation Modes

Two modes of operation:

- **Batch**: Submit & forget. Job waits in the queue until resources become available, runs, emails user on exit.

- **Interactive**: Allows interactive use, no different than remotely using any workstation.
  - Required for using GUI, such as MATLAB, R, COMSOL, ANSYS, visualization, etc.
Submitting Batch Jobs

• Everything needs to be scripted. Not for codes that require user interaction (e.g. press ‘y’ to continue).
• A ‘PBS script’ that includes resource requirements, environmental settings, and tasks.
• Use ‘qsub’ to submit the job.

```
qsub example_PBS_Script.pbs
```
• The output and error logs are printed on files, as they would appear on the screen.
PBS Script Example

# This is an example PBS script
#PBS -N hello
#PBS -l nodes=2:ppn=4
#PBS -l pmem=2gb
#PBS -l walltime=15:00:00
#PBS -A <specify your account>
#PBS -q embers
#PBS -j oe
#PBS -o myjob.out
#PBS -m abe
#PBS -M youremail@gatech.edu

# A name for this run, can be anything
# 2 nodes, 4 cores in each
# 2GB memory per core (16GB total)
# 15 hrs “max”, after which job is killed!!
# Account to which the job is charged (e.g., GT-gburdell3)
# submitting to queue named “embers”
# Put output and error files in specified format
# Event notifications set to start, finish or error, via email

cd $PBS_O_WORKDIR
echo "Started on `/bin/hostname`"
module purge
module load gcc/4.9.0 mvapich2/2.1
mpirun -np 8 ./hello

Actual Computation
Examples for Specific Resource Requests on the Phoenix Cluster

- **CPU-192GB**: less than 8 GB per processor, no features
  - #PBS -l nodes=1:ppn=4:pmem=6gb

- **CPU-384GB**: less than 16 GB per processor, no features
  - #PBS -l nodes=1:ppn=4:pmem=9gb

- **CPU-768GB**: more than 16 GB per processor, no features
  - #PBS -l nodes=1:ppn=4:pmem=22gb

- **CPU-384GB**: less than 16 GB per processor, "local-sas" feature request
  - #PBS -l nodes=1:ppn=4:local-sas,pmem=5gb

- **CPU-768GB**: more than 16 GB per processor, "local-sas" feature request
  - #PBS -l nodes=1:ppn=4:local-sas,pmem=26gb

- **GPU-192GB-V100**: less than 8 GB per processor, 1-2 GPUs and optionally Tesla V100-16GB or TeslaV100-32GB feature
  - #PBS -l nodes=1:ppn=4:gpus=1:TeslaV100-16GB

- **GPU-384GB-V100**: less than 16 GB per processor, 1-2x GPUs and optionally Tesla V100-16GB or TeslaV100-32GB feature
  - #PBS -l nodes=1:ppn=4:gpus=1,mem=40gb

- **GPU-768GB-V100**: more than 16 GB per processor, 1-2x GPUs and optionally Tesla V100-16GB or TeslaV100-32GB feature
  - #PBS -l nodes=1:ppn=2:gpus=2,pmem=20gb

- **GPU-384GB-RTX6000**: less than 16 GB per processor, 1-4x GPUs and RTX6000 feature
  - #PBS -l nodes=1:ppn=2:gpus=1:RTX6000

- **GPU-768GB-RTX6000**: more than 16 GB per processor, 1-4x GPUs and RTX6000 feature
  - #PBS -l nodes=1:ppn=6:gpus=1:RTX6000,mem=128gb
Checking Account Balance on the Phoenix Cluster

[puser32@login-phoenix-3 ~]$ mam-balance

<table>
<thead>
<tr>
<th>Name</th>
<th>Balance</th>
<th>Reserved</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT-gburdell3</td>
<td>68.0000</td>
<td>0.0000</td>
<td>68.0000</td>
</tr>
<tr>
<td>GT-gburdell3-CODA20</td>
<td>43011.3219</td>
<td>0.0000</td>
<td>43011.3219</td>
</tr>
<tr>
<td>GT-gburdell3-FY20Phase2</td>
<td>17860.7500</td>
<td>0.0000</td>
<td>17860.7500</td>
</tr>
<tr>
<td>GT-gburdell3-phys</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>GT-phys-CODA20</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

• Further details about the job accounting may be accessed from the following documentation:
  http://docs.pace.gatech.edu/phoenix_cluster/submit_jobs_phnx/#job-accounting

• The mam-balance does not apply to Hive cluster
Some applications can benefit from a local disk for faster I/O.

All PACE machines have local disks (NVMEs and SAS)

**Standard location:** /scratch

**Different than:** ~/scratch !

Scheduler monitors free space

Each job **automatically creates** a directory under /scratch, e.g.:

```
scratch/21034470.sched-torque.pace.gatech.edu
```

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, request it in this format:

```
#PBS -l nodes=1:ppn=4:local-sas,pmem=5gb
#PBS -l file=10gb (per processor!)
```
Interactive Jobs

• Same PBS commands, but this time on the command line:

  qsub -I -X -q embers -A GT-gburdell3 -l nodes=2:ppn=4,walltime=4:00:00,pmem=2gb
  • -X allows for GUI
  • -A is for the account the job will be charged
  • "," are for binding multiple values for a parameter (-l)

• The scheduler logs the user onto a compute node when the resources become available

• Session is terminated when:
  • The user exits; The terminal is closed; The walltime is exceeded

• Alternatively, use the **pace-vnc-job** command to launch a VNC job

  pace-vnc-job -q embers -A GT-gburdell3 #submits 1hour default job

• Note: You will need a VNC client installed on your computer to access the session by following the 3 steps provided on the terminal

• Documentation: [http://docs.pace.gatech.edu/interactiveJobs/setupVNC_Session/](http://docs.pace.gatech.edu/interactiveJobs/setupVNC_Session/)

• Also, **pace-jupyter-notebook** command available!
## Monitoring Jobs

`qstat` lists your queued jobs and their state:

```
qstat -u <UserName> -n
```

sched-torque.pace.gatech.edu:

<table>
<thead>
<tr>
<th>Job ID</th>
<th>Username</th>
<th>Queue</th>
<th>Jobname</th>
<th>SessID</th>
<th>NDS</th>
<th>TSK</th>
<th>Req'd Memory</th>
<th>Req'd Time</th>
<th>S</th>
<th>Elap Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>95521.sched-torque.pacatl1-1-02-010-14-l</td>
<td>ssarajlic3</td>
<td>inferno</td>
<td>pace-vnc-job.pbs</td>
<td>337870</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>01:00:00</td>
<td>R</td>
<td>00:01:42</td>
</tr>
<tr>
<td>95522.sched-torque.pacatl1-1-02-010-14-l</td>
<td>ssarajlic3</td>
<td>inferno</td>
<td>pace-vnc-job.pbs</td>
<td>339184</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>01:00:00</td>
<td>R</td>
<td>00:01:08</td>
</tr>
<tr>
<td>95526.sched-torque.pacatl1-1-02-010-14-l</td>
<td>ssarajlic3</td>
<td>inferno</td>
<td>pace-jupyter-not</td>
<td>340707</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>01:00:00</td>
<td>R</td>
<td>00:00:34</td>
</tr>
</tbody>
</table>
More on PBS Jobs

• Cancelling a submitted job

qdel <jobID>

• Querying for specific users/queues

showq -w class=<QueueName>,user=<UserID>  ➔ All jobs for the queue & user
showq -r -w user=<UserID>  ➔ All “running” jobs for the given user
showq -b -w class=<QueueName>  ➔ All “blocked” jobs for the given queue
showq -w qos.cpu=192GB  ➔ All jobs for the given qos
### Checking the Queue or Node Class Status

- **High level overview of the Phoenix cluster that summarizes the utilization per node class**

```bash
done. check-queue inferno -b
```

---

#### inferno Queue Summary

- **Last Update:** 12/01/2020 17:00:05
- **Number of Nodes (Accepting Jobs/Total):** 689/1261 (54.64%)
- **Number of Cores (Used/Total):** 13802/30240 (45.64%)
- **Amount of Memory (Used/Total) (GB):** 19833/359951 (5.51%)

---

<table>
<thead>
<tr>
<th>Node Class</th>
<th>Total</th>
<th>CPUs</th>
<th>GPUs</th>
<th>Mem (GB)</th>
<th>Loc Drv (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nodes</td>
<td>Ded/Tot</td>
<td>Tot</td>
<td>Ded/Tot</td>
<td>Use/Ded/Tot</td>
</tr>
<tr>
<td>CPU-192GB</td>
<td>787</td>
<td>12310/18888</td>
<td>0/0</td>
<td>13155/51998/153465</td>
<td>16763/0/1109671</td>
</tr>
<tr>
<td>CPU-384GB</td>
<td>238</td>
<td>584/5712</td>
<td>0/0</td>
<td>4377/8768/91392</td>
<td>1123/0/292700</td>
</tr>
<tr>
<td>CPU-384GB-SAS</td>
<td>79</td>
<td>0/1896</td>
<td>0/0</td>
<td>585/0/30336</td>
<td>115/0/543664</td>
</tr>
<tr>
<td>CPU-768GB</td>
<td>64</td>
<td>80/1536</td>
<td>0/0</td>
<td>588/2400/48580</td>
<td>249/0/90240</td>
</tr>
<tr>
<td>CPU-768GB-SAS</td>
<td>4</td>
<td>0/96</td>
<td>0/0</td>
<td>40/0/3048</td>
<td>4/0/29796</td>
</tr>
<tr>
<td>GPU-192GB-V100</td>
<td>21</td>
<td>218/504</td>
<td>24/34</td>
<td>497/19/4095</td>
<td>41/0/39226</td>
</tr>
<tr>
<td>GPU-384GB-RTX6000</td>
<td>30</td>
<td>608/696</td>
<td>108/116</td>
<td>308/108/11136</td>
<td>211/0/40890</td>
</tr>
<tr>
<td>GPU-384GB-V100</td>
<td>22</td>
<td>0/528</td>
<td>0/44</td>
<td>154/0/8448</td>
<td>134/0/37576</td>
</tr>
<tr>
<td>GPU-768GB-RTX6000</td>
<td>5</td>
<td>2/120</td>
<td>2/20</td>
<td>44/256/3337</td>
<td>18/0/25167</td>
</tr>
<tr>
<td>GPU-768GB-V100</td>
<td>5</td>
<td>0/120</td>
<td>0/10</td>
<td>45/0/3810</td>
<td>8/0/8242</td>
</tr>
</tbody>
</table>

---

- **Utilization per Node Class**

```bash
done. check-queue -n=GPU-768GB-RTX6000 -c -s inferno
```
PACE Software Stack

Everything is in `/usr/local/pace-apps`

• Licensed software packages:
  • Common license: Matlab, Fluent, Mathematica, Abaqus, Comsol, ...
  • Individual license: Vasp, Gaussian, ...

• Open source packages and HPC libraries:
  • BLAS, PETSc, NAMD, NetCDF, FFTW, BLAST, LAMMPS, ...

• Compilers:
  • C/C++ & Fortran: GNU, Intel, PGI, NAG
  • Parallel Compilers: OpenMP, MPICH, MPICH2, MVAPICH
  • GPU Compilers: CUDA, PGI

• Scripting Languages: Python, Perl, R, ...
**Modules**

- Painless configuration for software environment and switching between different versions:
  No more editing `PATH`, `LD_LIBRARY_PATH`, etc!

- Main commands:
  - `module avail`: Lists all available modules that can be loaded
  - `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 (buggy)
  - `module spider`: Lists all software and its available versions on cluster

  ```
  $ module load matlab/r2020a
  ```

- Example PBS batch scripts command line utility (`pace-getexample`) for downloading example batch scripts for frequently used PACE software modules
  - `pace-getexample --list`: Displays a list of currently available examples
  - `pace-getexample --help`: Instructions on how to download and submit an example job to the queue

- **Must-Read**: PACE-specific use cases, examples, and gotchas
  - [http://docs.pace.gatech.edu](http://docs.pace.gatech.edu)
PACE Email Lists & Blog

- **User Lists:**
  - **All users:**
    - pace-availability
    - pace-discuss
  - **Cluster-specific lists:**
    - pace-hive

- **PACE blog:**
  http://blog.pace.gatech.edu
Meet our Research Scientist and Customer Experience Teams

Mehmet Belgin, PhD
Chris Blanton, PhD
Aaron Jezghani, PhD
Fang Liu, PhD
Kevin Manalo, PhD
Craig Moseley
Semir Sarajlic, MS, GISP
Chris Stone, PhD
Michael Weiner, PhD
Nuyun Zhang, PhD
THANK YOU!
We welcome your feedback!
https://b.gatech.edu/2LerSxZ

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

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

PACE Consulting Sessions:
Visit pace.gatech.edu for schedule of the weekly sessions and BlueJeans link to the session