Site Update: Georgia Institute of Technology

Marian Zvada, MSc, MBA (zvada@gatech.edu)
Aaron Jezghani, PhD (ajezghani3@gatech.edu)

Partnership for **A**dvanced **C**omputing **E**nvironment
www.pace.gatech.edu

Georgia Tech

The Partnership for an Advanced Computing Environment

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

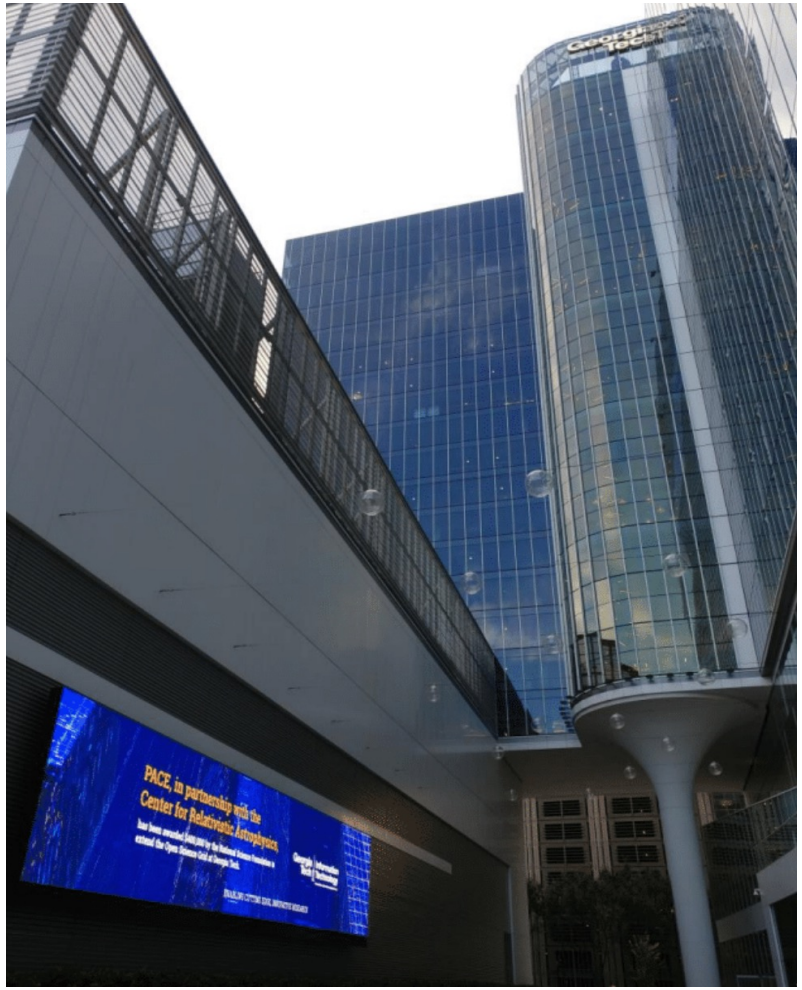
SOFTWARE LICENSES / TOOLS

Isolated Special Purpose Clusters (e.g., Hive)

Shared pooled nodes funded by the Institute and Faculty via cost model

BACKUP / STORAGE / NETWORKING / INTERCONNECT

INFRASTRUCTURE
(Space / Cooling / Power)

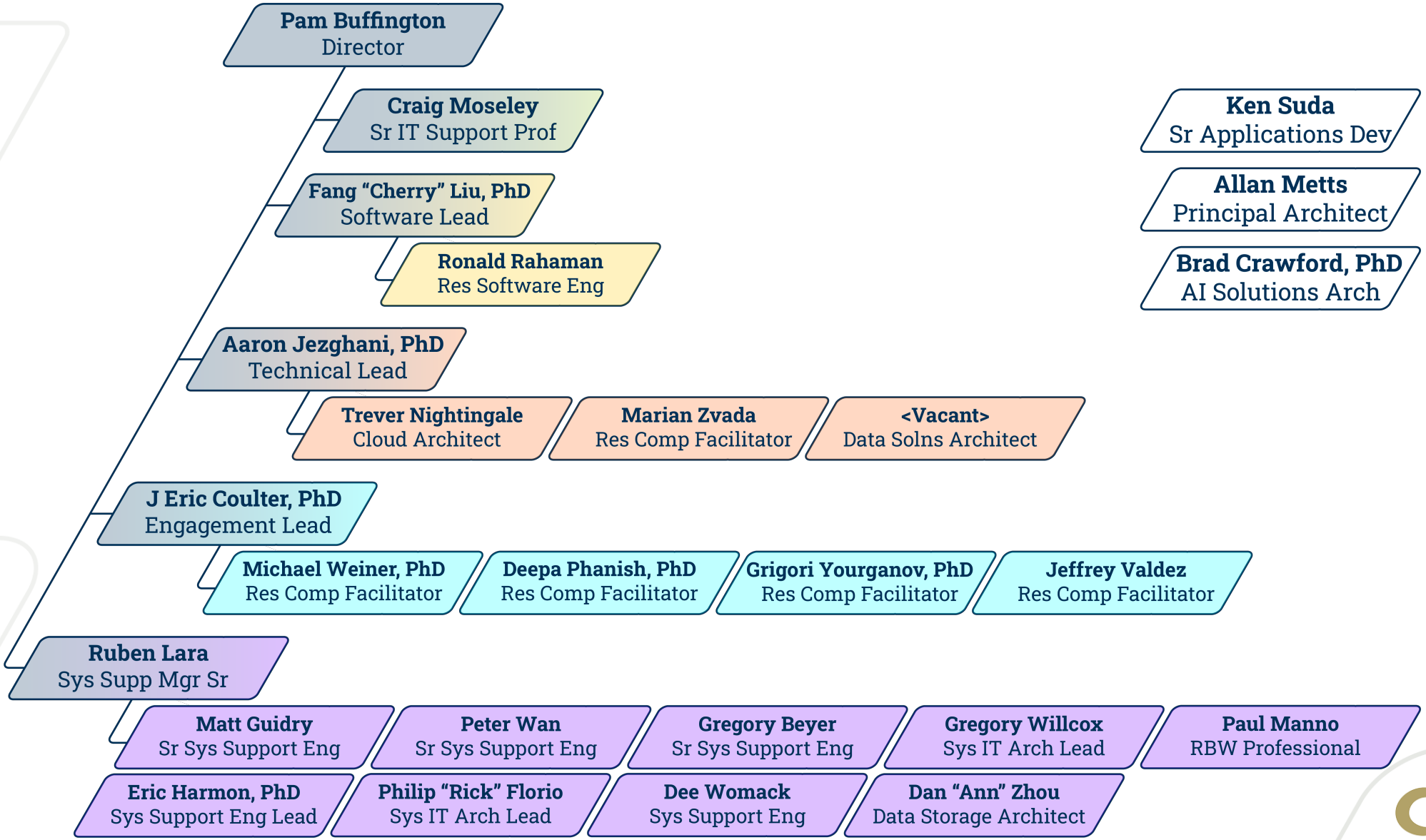


Virtual tour of Coda datacenter hosting PACE resources:
https://pace.gatech.edu/coda-datacenter-360-virtual-tour

zvada@gatech.edu · ajezghani3@gatech.edu · pace.gatech.edu

Georgia Tech

The PACE Team



zvada@gatech.edu · ajezghani3@gatech.edu · pace.gatech.edu

Georgia Tech

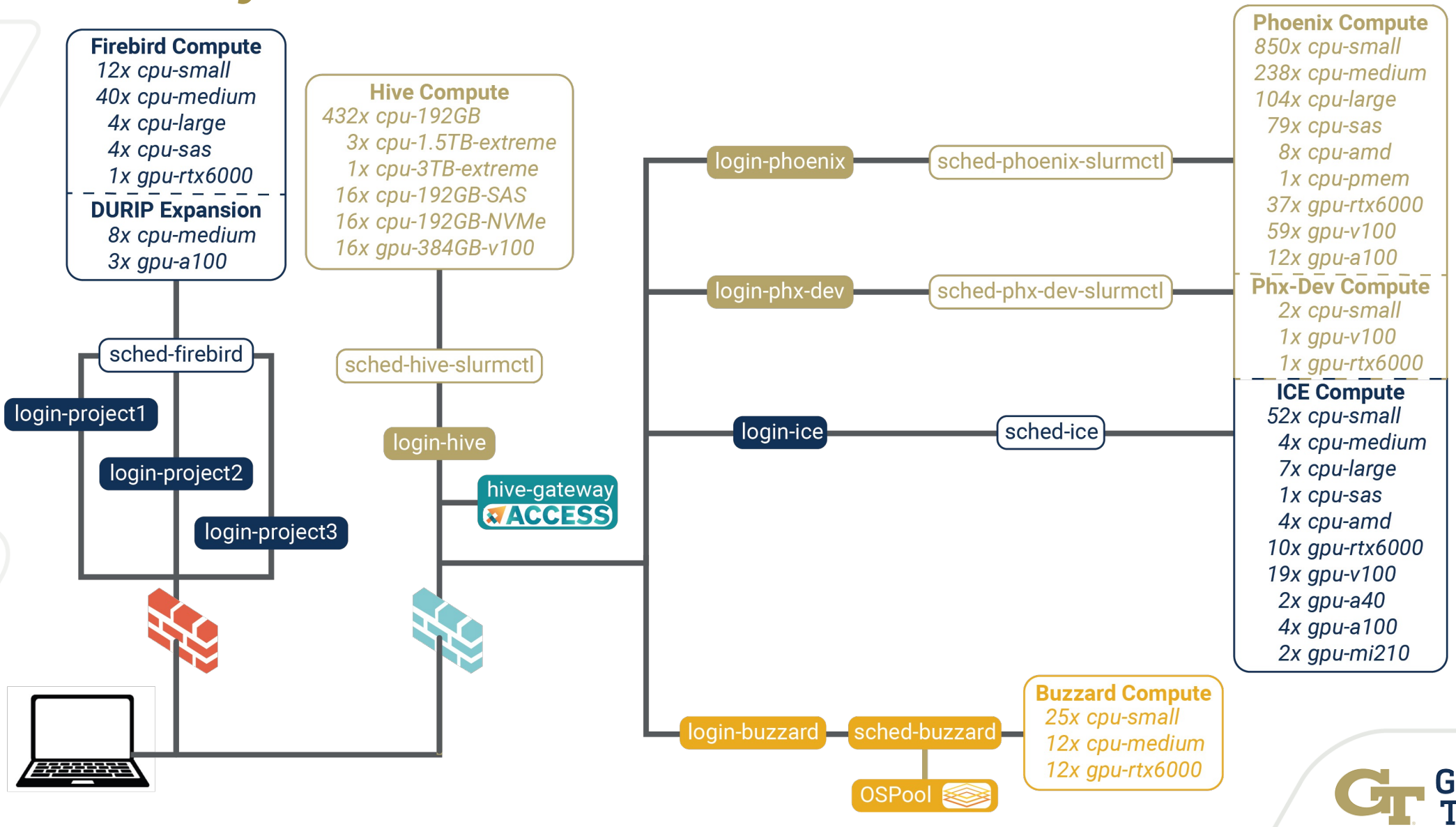
PACE Clusters

- Phoenix: *Campus-Wide Access*
  1,389 nodes · 34,968 CLX/Milan CPUs · 292 GPUs (RTX6000/V100/A100) · 6.5 PB Lustre
- Hive: *Special-Purpose Research (NSF MRI)*
  484 nodes · 11,636 CLX CPUs · 64 GPUs (V100) · 2.5 PB GPFS
- ICE: *Instructional Access*
  101 nodes · 3,032 CLX/Milan CPUs · 98 GPUs (RTX6000/V100/A40/A100/MI210) · 4.1 PB Lustre
- Firebird: *CUI/ITAR-Compliant Cluster*
  72 nodes · 1,888 CLX/ICX CPUs · 16 GPUs (RTX6000/A100) · 1 PB JBOD
- Buzzard: *Open Science Grid (OSG) HTC Cluster*
  49 nodes · 1,176 CLX CPUs · 48 GPUs (RTX6000) · 415 TB StashCache

Georgia Tech

zvada@gatech.edu · ajezghani3@gatech.edu · pace.gatech.edu

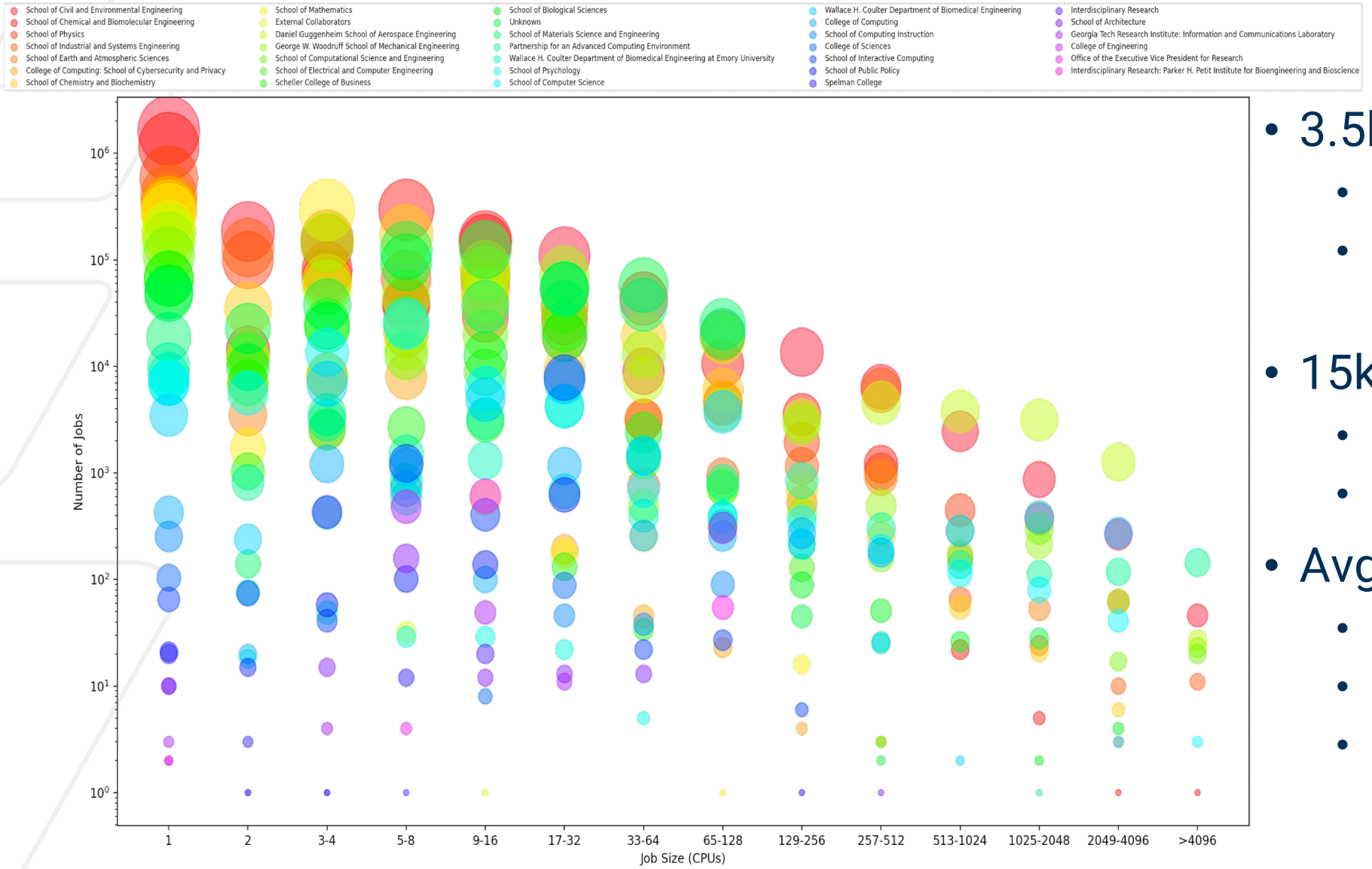
Cluster Layout



zvada@gatech.edu · ajezghani3@gatech.edu · pace.gatech.edu

Georgia Tech

Cluster User and Workflow Composition



- 3.5k registered users
  - 300-400 active/month
  - Full spectrum of GT community
- 15k students on ICE
  - 50-600 active/month
  - Highly variable activity
- Avg. 10k jobs/day
  - Frequent bursts to 20k+
  - 10% >24 hour walltime
  - ~1 hour avg. wait

Georgia Tech

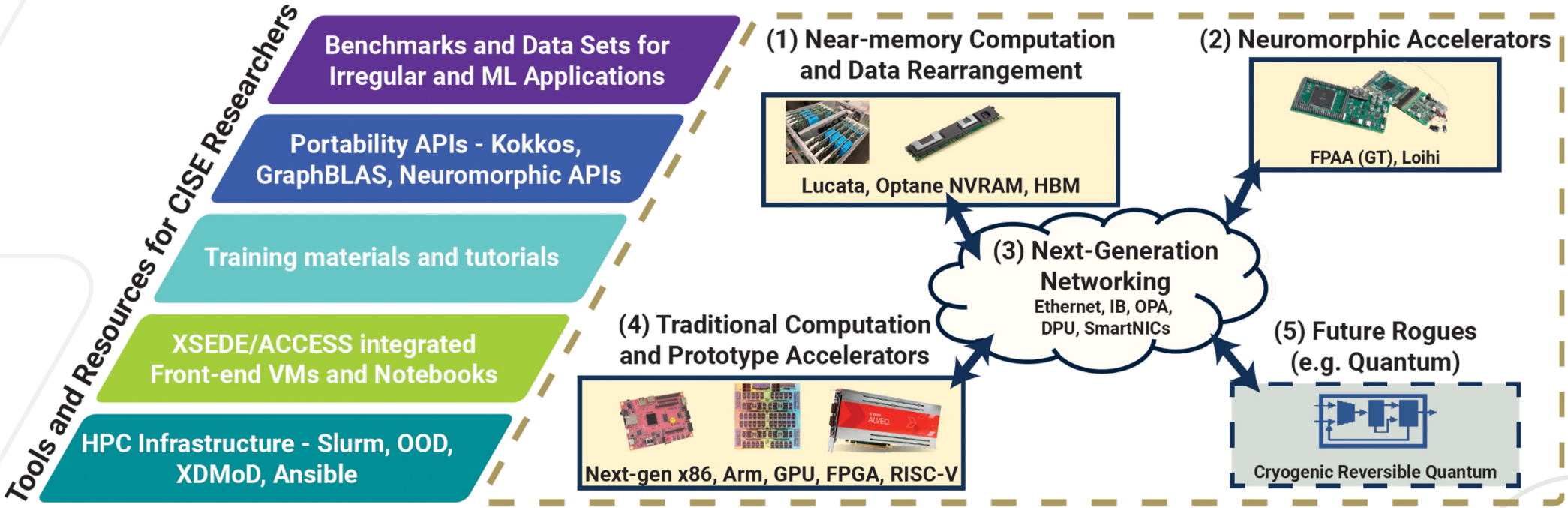
zvada@gatech.edu · ajezghani3@gatech.edu · pace.gatech.edu

Additional Hardware In-bound

- Next-generation x86 architectures
  - Intel Sapphire Rapids
  - AMD Genoa
    - Dual-InfiniBand NICs
- AI-focused GPU infrastructure
  - H100-HGX and H100-DGX platforms
- Planned refresh of all compute infrastructure
  - Cascade Lake -> ???
  - RTX6000/V100 -> ???

The Rogues Gallery: A Post-Moore Testbed

- Hosted by the Center for Research into Novel Computing Hierarchies
- NSF-funded, open-access cluster for novel compute
  - x86, GPU, Arm, SOC, FPGA, DPU, SmartNIC, Lucata Pathfinder, etc.



Georgia Tech

zvada@gatech.edu · ajezghani3@gatech.edu · pace.gatech.edu

Motivating a Change in Scheduler

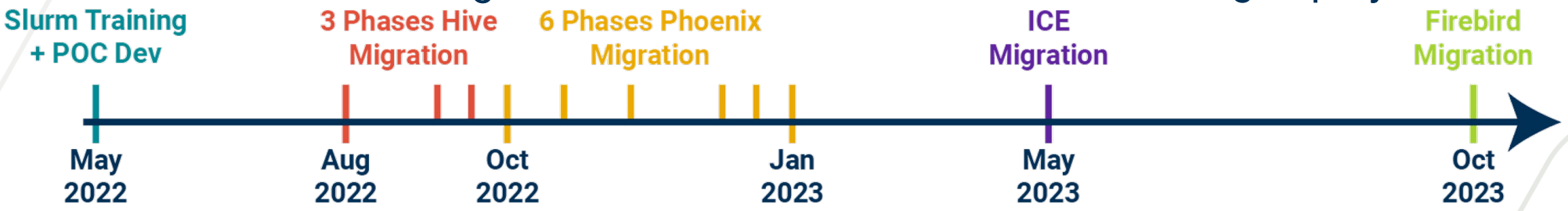
- Previously used Moab/MAM/Torque to manage cluster resources and user workload
  - Queue and QOS to provide multifaceted scheduling
  - 2-dimensional resource limits (e.g. CPUs and CPU-time) to consolidate queues
  - Architecture-based cost recovery using MAM – short time to production
- “Interesting” solutions to manage clusters
  - “Topology-aware” scheduling via NODESETLIST (preferred node features)
  - Submit filter script to abstract scheduler interface (lack of API -> lots of conditional statements)
  - Manage Firebird node reboots via epilog scripts (problematic as scheduler is unaware of reboot)

Motivating a Change in Scheduler

- Moab/MAM/Torque communications required 10-minute timeout to prevent corrupt job accounting
  - Throttled job throughput considerably (<5k jobs/day)
  - Still inconsistent reporting between utilities, scheduler outages
- Limited policy definitions in Torque left much to be desired
  - Moab policy more robust, but rejections resulted in cryptic errors
- Lack of hardware support required custom scripting
  - Manage cgroups via prolog/epilog for Nvidia GPUDirect-RDMA and AMD GPUs
- Incompatible with RHEL8+
  - Required for AMD GPU drivers and CPU optimizations
- External sessions (e.g. VS Code) not confined to job allocation
  - Unexpected resource contention and job failures

Preparing to Migrate

- SchedMD training and engineering support
  - Team-wide training
    - Broader knowledge base
    - No single point of failure
  - Developer expertise to accelerate deployment
- Iterative transitions by cluster to build on past efforts
  1. Hive (3 phases)
     - “Traditional” HPC
  2. Phoenix (6 phases)
     - Cost-model accounting
     - Tiered scheduling
  3. ICE
     - Additional privacy (FERPA)
     - Further abstract scheduling
  4. Firebird
     - Slurm-managed project isolation



zvada@gatech.edu · ajezghani3@gatech.edu · pace.gatech.edu

Georgia Tech

Common Configurations Across the Cluster

- Updated hostnames to end in numbers for node expressions
- Prepare local RPMs with regular frequency to keep current with SchedMD releases
  - pmix v3/v4 support for OpenMPI and MVAPICH2 stacks
  - pmi2 for IntelMPI and older OpenMPI lacking pmix support
- VM + baremetal dev environments (per cluster) to validate updates/features before push to production
- Primary configuration files and RPM versions managed via Salt
  - Use configless Slurm to propagate configurations to clients to maintain consistency across cluster

Common Configurations Across the Cluster

- High-availability setup to maximize scheduler uptime
  - Primary Slurmctld server + Secondary Slurmctld on Slurmdbd/Mariadb server
- Used node features to build partitions from nodesets
  - Easier management of partitionsAdded plugins to handle capabilities via high-level (API) interface
  - Lua job_submit to enforce job policy and abstract scheduling
  - Route/topology to map jobs efficiently across fabric
- Prolog/epilog to continue preamble/postamble in job output
  - Additionally control things like kernel paranoia for profiling

Cluster-Specific Considerations

- Hive: Isolate jobs from external Xsede/ACCESS users
  - Separate partitions atop same hardware as internal, but with EXCLUSIVE_USER set
- Phoenix: Per-partition TRESBillingWeights for cost recovery
  - Requires additional custom infrastructure/tooling for full accounting capabilities, but working well so far
  - Low-priority, preemptible QOS with 0 Usage for free backfill
- ICE: Soften boundaries between clusters using QOS partitions to respect hardware quantities rather than physical servers
- Firebird: Use node helper scripts to isolate projects and manage reboots
- Rogues Gallery: Federated clusters for incompatible settings across architectures

Updates to PACE utilities

- pace-check-queue: historical cluster status utility
  - Prior implementation involved parsing and caching the output from `checknode ALL`
  - Under Slurm, simply a wrapper around `sinfo --json`
    - Change in json schema in 23.02.x caused recent hiccups, but we've recovered

```
=== phoenix-all Partition Summary ===
        Last Update                          : 09/06/2023 13:01:35
        Next Maintenance Start               : 10/24/2023 06:00:00
        Number of Nodes (Accepting Jobs/Total) : 2/4 (50.00%)
        Number of Cores (Used/Total)         : 78/96 (133.33%)
        Amount of Memory (Used/Total) (GB)   : 215/1083 (19.85%)
```

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?
node1	18/24	53.84	0/0	8/ 144/ 178	4.00	-/ -/1393659	0.00	Yes
node2	24/24	98.92	0/0	9/ 176/ 177	5.00	-/ -/1393659	0.00	No (Busy)
node3	12/24	34.13	1/2	187/ 200/ 364	51.00	-/ -/844507	0.00	Yes
node4	24/24	12.78	2/2	11/ 40/ 364	3.00	-/ -/844507	0.00	No (Busy)

Updates to PACE utilities

- pace-quota: report on storage quotas, queue/account access/balances
  - Queue/account access was determined from Moab and MAM queries
  - Under Slurm, this is achieved with `sshare` and `sacctmgr`

```
Gathering storage and job accounting information for user: gburdell3

  ** 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)          :   George Burdell
 * UserID                                :   1234567
 * Username                              :   gburdell3
 * Your Email (for PACE contact)         :   gburdell3@gatech.edu

================================================================================================================
                                    Phoenix Storage with Individual User Quota
================================================================================================================
Filesystem                                      Usage (GB)    Limit      %   File Count      Limit      %
Home:                                                  0.0     10.0   0.0%          277    1000000   0.0%
Scratch:                                               0.0  15360.0   0.0%            1    1000000   0.0%

================================================================================================================
                                    Phoenix Storage with Research Group Quota
================================================================================================================
Filesystem                                      Usage (GB)    Limit      %   File Count      Limit      %
Project:                                            470.6   1024.0  46.0%       453577          0   0.0%

================================================================================================================
                                        Job Charge Account Balances
================================================================================================================
Name                                 Balance          Reserved          Available
free-tier                              67.99             20.79              47.20
```

New PACE utilities

- pace-job-summary: gather information about historical jobs
  - Provide a wrapper around `sacct` that only requires a Job ID
  - Yields resource utilization information, batch script, etc.
  - Users: reproduce/analyze past jobs
  - PACE: debug job/workflow issues

```
----------------------------------------
Begin Slurm Job Summary for 3023593
Query Executed on 2023-09-06 at 15:53:32
----------------------------------------
Job ID:     10000001
User ID:    gburdell3
Account:    free-tier
Job name:   myjob
Resources:  cpu=1,mem=80G,node=1
Rsrc Used:  cput=00:02:47,vmem=0.0M,walltime=00:02:47,mem=0.0M,energy_used=0
Exit Code:  0:0
Partition:  cpu-large
Nodes:      node1
QOS:        inferno
----------------------------------------
Batch Script for 10000001
----------------------------------------
#!/bin/bash
#SBATCH -J myjob                          # Job name
#SBATCH --account=free-tier               # charge account
#SBATCH -N1 --ntasks-per-node=1           # Number of nodes and cores per node required
#SBATCH --mem-per-cpu=80G                 # Memory per core
#SBATCH --time=48:00:00                   # Duration of the job (Ex: 10 mins)

cd $SLURM_SUBMIT_DIR                      # Change to working directory

#Load julia
module load julia/1.7.2

#run
julia ./somescript.jl ${run}


----------------------------------------
```


Georgia Tech

Managing PAM stack

- Strong desire for pam_slurm_adopt to avoid issues with external sessions
- But…administrative utilities necessitate pam_systemd
- Solution: account + pam_listfile and session + pam_succeed_if in /etc/pam.d/password-auth
  - If groupname is listed in /path/to/admingroupfile, skip pam_slurm_adopt
  - If in GID 0 (root) or 111111 (admins), use pam_systemd for session

```
.
.
.
account     sufficient    pam_listfile.so item=group sense=allow onerr=fail file=/path/to/admingroupfile
-account    required      pam_slurm_adopt.so
.
.
.
session    [default=1 success=ignore]    pam_succeed_if.so quiet gid in 0:111111
-session     optional      pam_systemd.so
.
.
.
```

Experiences so Far

- Phoenix: 35% increase in job throughput (10-15k vs. 7-10k daily jobs)
  - Stable through significantly higher bursts (30-40k/day,
  - Average job weight time 1.11 -> 0.71 hours
- Hive maintaining utilization as before
- ICE merger accommodated by policy robustness
- Much better stability in scheduler – 2 outages, with good reason
  - First resolved via tuned scheduling parameters to match shift in user workflow
  - Second resolved following SlurmDB association realignment + optimizations
- The additional features and hardware support in Slurm have been beneficial
  - Less home-brew scripting and more API calls
  - Able to support testbed architectures that would otherwise not be possible

Georgia Tech

The Slurm Roadmap for PACE and GT

- Last cluster transitioning to Slurm at end of October
- Upgrading to RHEL9 on all clusters
  - MariaDB 5.5 -> 10.x and concurrent SlurmDB migration
  - Cgroups v1 -> v2
- Adoption of additional Slurm capabilities
  - NSS for on-prem, cloud bursting
  - RESTFUL API and the wide array of opportunity it presents
  - scrun and Slurm container support
  - LUA burst buffer to stage job data?
- Using Slurm to manage access to all the new hardware
  - Secure enclaves, diversified accelerators, novel compute architectures, etc.

Georgia Tech

Conclusions

- In FY23, PACE underwent a transition to Slurm for resource management
  - Migration of roughly 2,000 servers across 4 clusters
  - Prioritized customer experience over expedience
    - Phased migrations with focus on user and development support
    - Avoided translation scripts to empower research community under Slurm
- Maintained prior capabilities but added many new features
  - Lower the barrier to entry
  - Enable and empower users in their research

Feel free to reach out with any questions!

zvada@gatech.edu · ajezghani3@gatech.edu · pace.gatech.edu

Georgia Tech