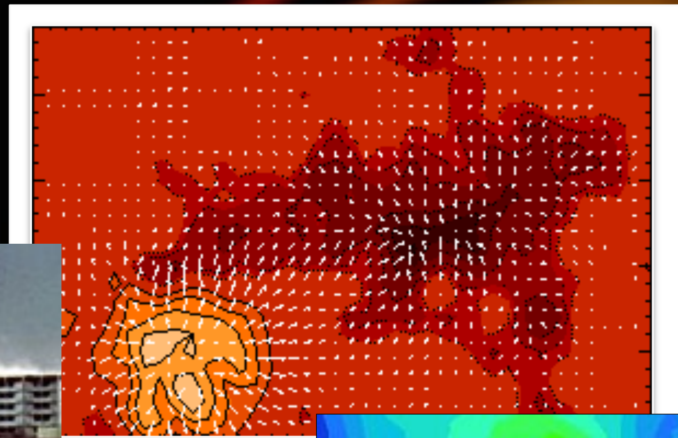
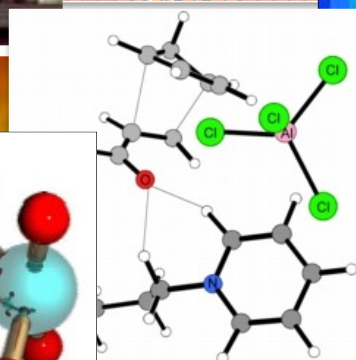
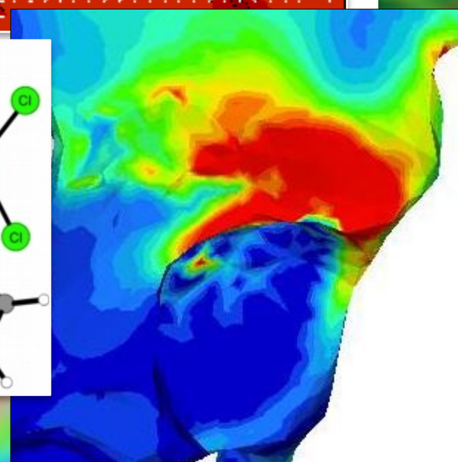
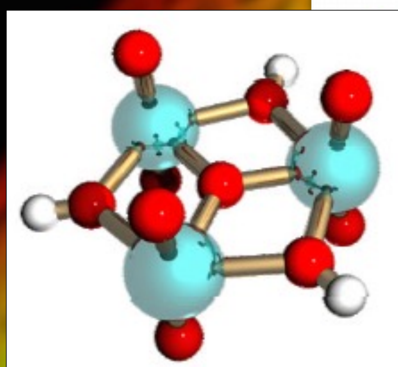
Alabama Supercomputer Center
Alabama Research and Education Network

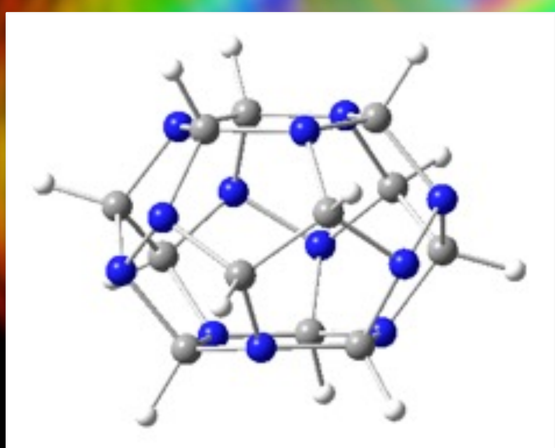
Who uses HPC?





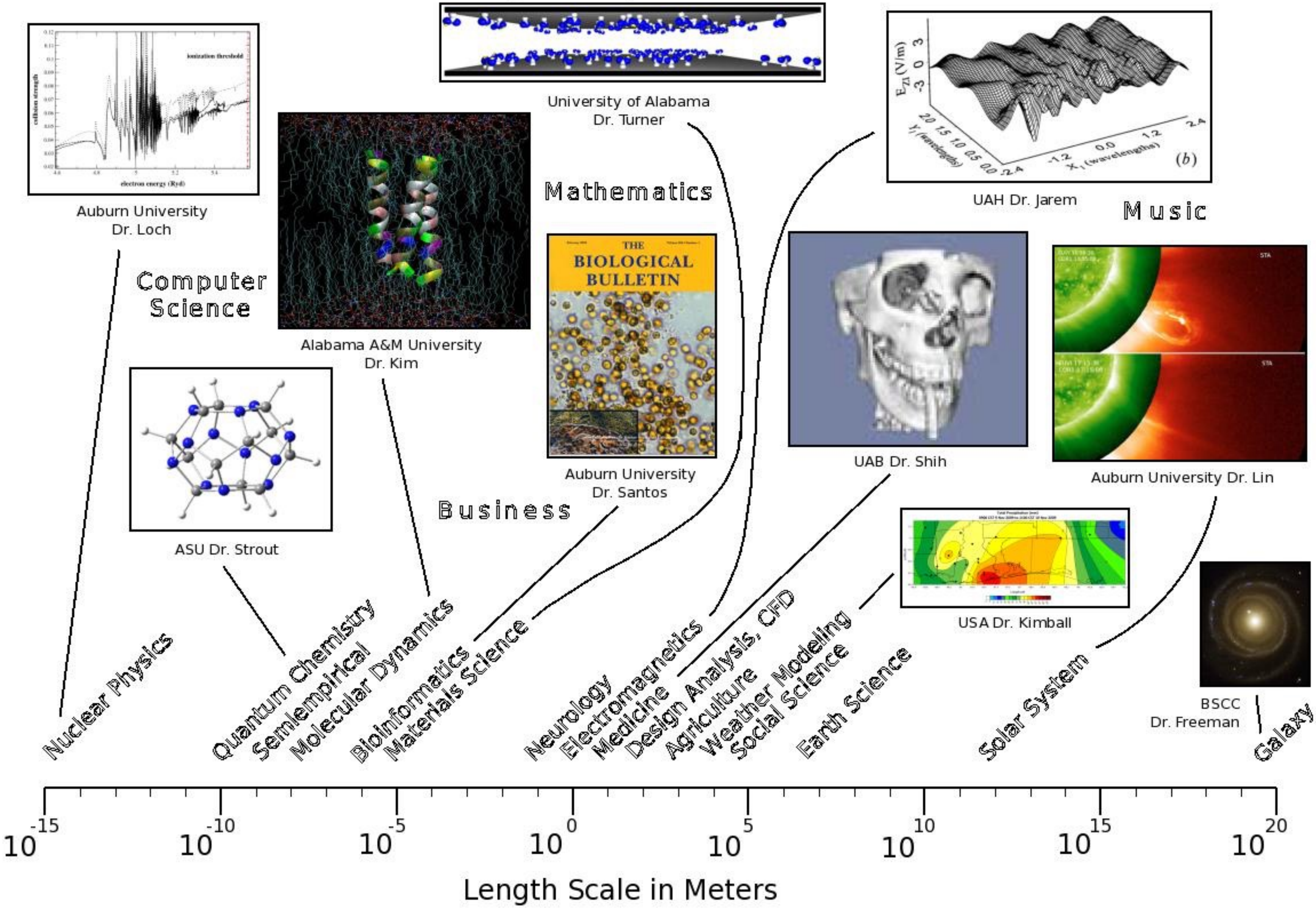






Alabama A&M University
Alabama State University
Athens State University
Auburn University
Auburn University in Montgomery
Bevill State College
Intel Corporation
Jacksonville State University
NASA
Operon Biotechnologies
Time Domain
Troy University
Tuskegee University
U.S. Air Force
U.S. Army
University of Alabama
University of Alabama at Birmingham
University of Alabama in Huntsville
University of Montevallo
University of South Alabama
University of West Alabama
ATA Engineering

How Supercomputers Are Used

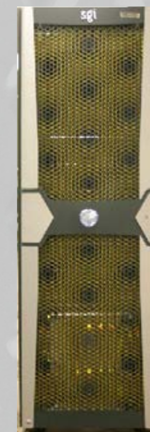


Alabama Supercomputer Authority
Historical Perspective

Cray X-MP
1987

Cray C90
1994

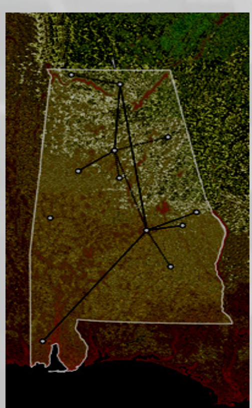
SGI Altix 350
2004



Altix 450
2006

SGI UV 2000
2012

An eternity in computer years

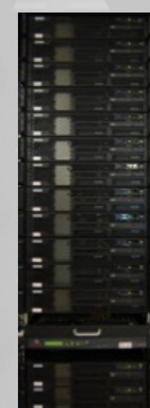


9 node
network

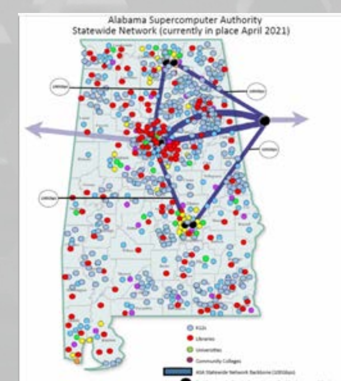
nCube
1991

Cray SV1
1999

Cray XD1
2004



DMC
2008



State-wide
network

Dense Memory Cluster

- Currently 3,740 x86-64 Processors (Intel)
- Shared/Distributed Memory Architecture
  - InfiniBand high speed/low latency network
- Memory (48GB-6000GB per node)
  ~26.1 TB memory available total
- Disk Storage
  ~162 TB internal, 842 TB shared via GPFS & BeeGFS

A Cluster of Nodes

Today almost all HPC systems are a group of servers that are used as one big computer. The servers are called “nodes” and the whole configuration is called a “cluster”. Some node types are;

- **Login nodes** let the users connect from their campus, and submit compute jobs to the cluster.
- **Compute nodes**
  - Conventional compute nodes typically have tens of processor cores and gigabytes of memory.
  - Big compute nodes may have hundreds of processor cores and terabytes of memory
  - Some have GPU math coprocessor boards
- **Infrastructure servers**
  - These servers handle functions like serving out passwords, managing queues, security, software licenses, time synchronization, monitoring, and backup.

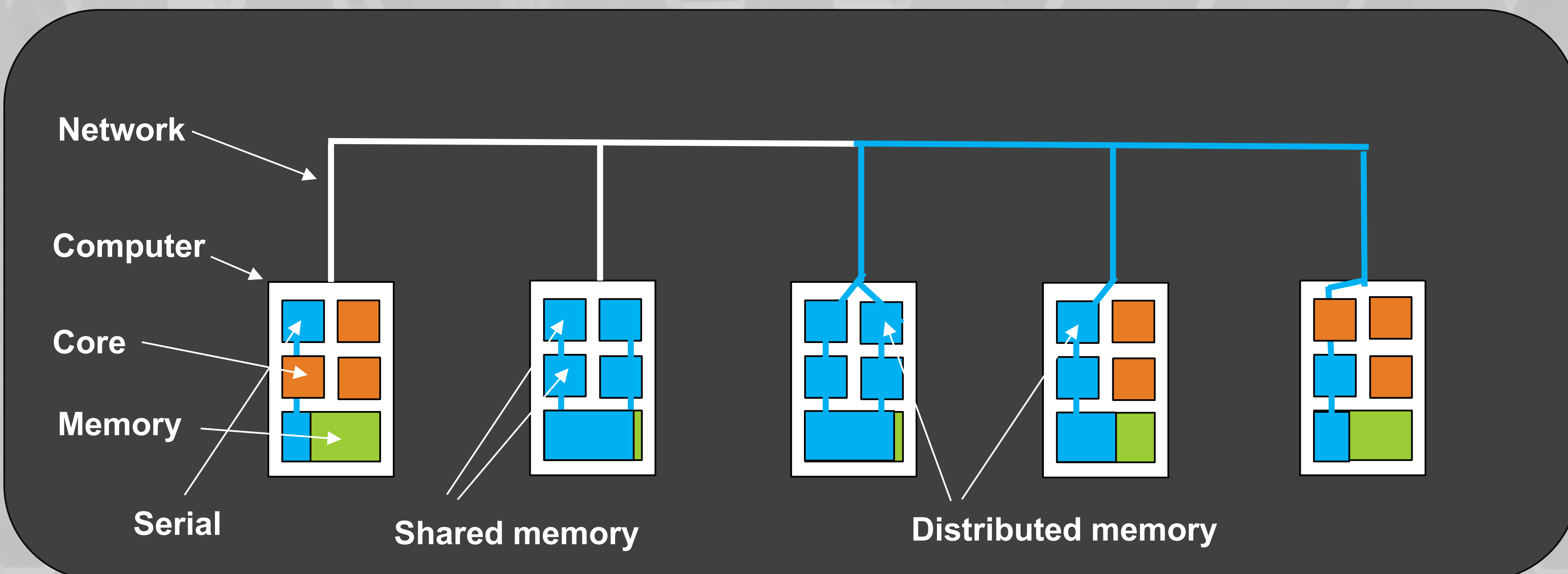
HPC Compared to Email/Web Servers

HPC System	Email/web Servers
CPU runs at 100% capacity	CPU runs at 10% capacity
Long running jobs	Many small, instantaneous tasks
Heavy communication between computers in the same building via InfiniBand	Frequent communication with computers outside the building via Ethernet
Access through ssh, scp, sftp	Access through http, imap, smtp
Software optimized for run time	Software optimized for latency
A small number of users	A large number of users
Some users need terabytes of disk storage	Many users need only megabytes of disk storage
Math coprocessors can improve performance	Redundant servers can improve performance
One job (calculation) can use many cores	Multiple virtual servers may share a core

Why the node size matters

Nodes

- **Serial Processing** – Traditionally, most software has used a single computer processor core.
- **Shared Memory Parallelism** – Software that runs on multiple processor cores that can access the same memory using programming tools like OpenMP.
- **Distributed Memory Parallelism** – Software that utilizes multiple computers on a network using programming tools like MPI parallel messaging library.
  - MPI software runs most efficiently if the network is fast with low latency, and if all of the nodes have the same model processor.



Nodes

DMC Nodes

Nodes	Cores	Memory	Processors
dmcvlogin2 - 4	8	16 GB	VM emulating Ivy Bridge, but running on a 2.3 GHz Haswell
dmc5-dmc40	20	128 GB	2.5 GHz Ivy Bridge (10 core)
dmc41-dmc52	36	128 GB	2.1 GHz Broadwell (18 core)
dmc53	192	6 TB	2.1 GHz Platinum Skylake-SP (24 core)
dmc54-dmc77	36	48 GB	2.7 GHz Gold Skylake-SP (18 core)
dmc78-dmc88	128	1 TB	2.0 GHz EPYC 7713 Milan (64 core)
dmc201	24	128 GB	2.3 GHz Haswell + two P100 GPUs
dmc202	24	90 GB	2.2 GHz Broadwell + four V100 GPUs
dmc203-204	128	1 TB	2.0 GHz Milan + four A100 GPUs

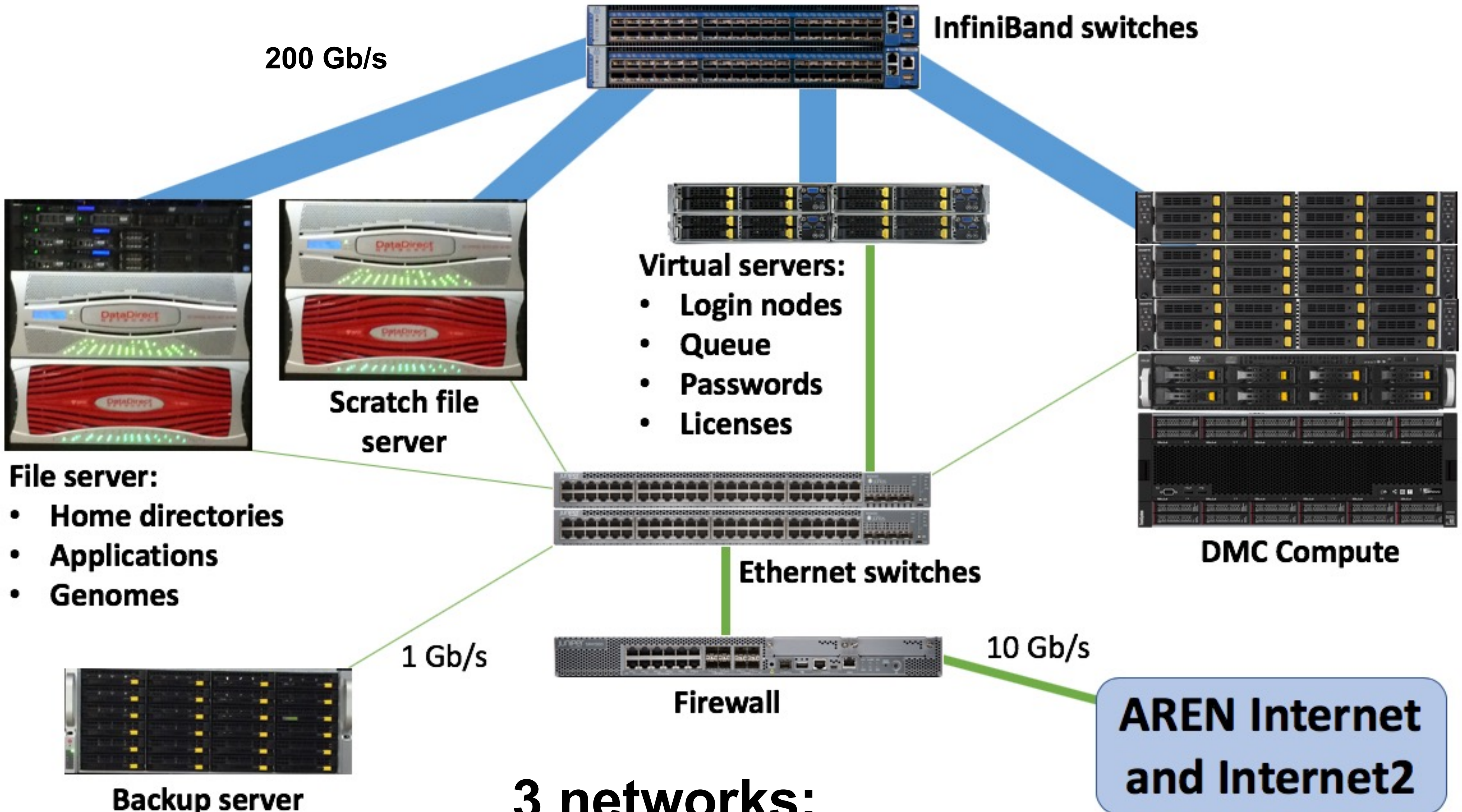
NVIDIA GPUs (DMC)

Pascal P100
- 2 Pascal GPUs
- 16 GB memory/GPU
- 3584 cores per GPU

Volta V100
- 4 Volta GPUs
- 32 GB memory/GPU
- 5120 cores per GPU

Ampere A100
- 8 Ampere GPUs
- 40 GB memory/GPU
- 6912 cores per GPU

The HPC Network



3 networks;
IB, Ethernet, IPMI

HPC Management Challenges

- **Policies and procedures**
- **Security**
- **Tight integration**
- **User account management**
- **Bleeding edge technology**
- **Facility issues**
- **Software installation, testing, and removal**
- **System monitoring, reporting, accounting**
- **Limited standardization**
  - Nearly every HPC system is customized for a given need.
  - Quantum chemistry, AI, molecular mechanics, bioinformatics all have different optimal hardware configurations.

Infrastructure servers

- Documentation / web *
- Bastion server
- Network traffic monitoring
- Log aggregation / SIEM
- Data transfer node *
- Queue system *
- License management *
- Directory (pswd, uid, gid) *
- Time synchronization *
- Configuration management
- MFA (i.e. Duo) *
- Linux repo / provisioning

- Backup
- System monitoring *
- System metrics *
- Security scanning *
- File system mgt *
- File system metadata *
- File system storage *
- Database *
- Cluster management
- Virtual machine host *
- Container server *

* Some HPC systems have multiple of these

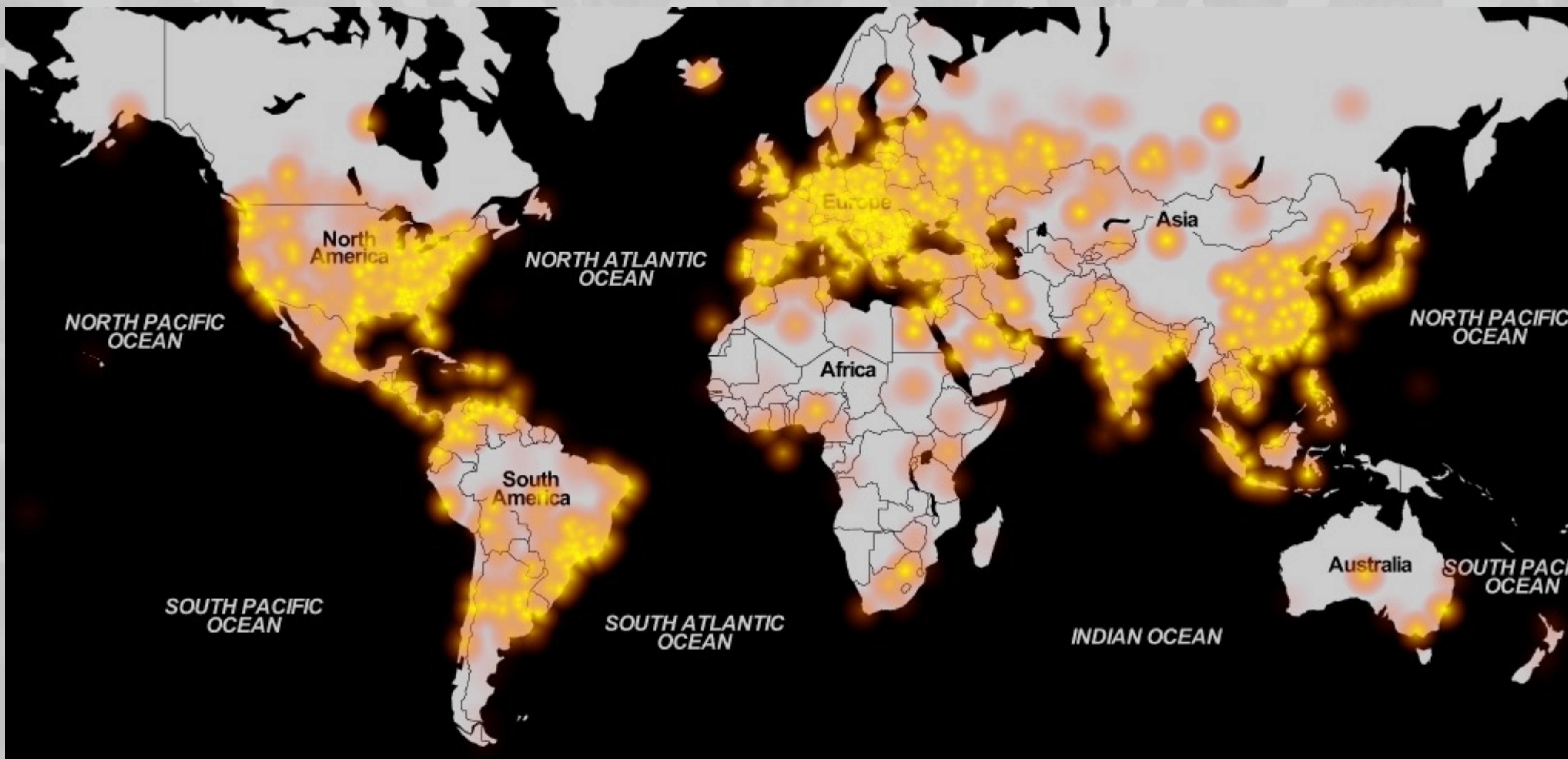
HPC Security

Secure

- Users have login access.
- HPC systems run thousands of pieces of software that can't easily be validated.
- There are many security standards; ITAR, CMMC, CIS, Top Secret, NIST SP 800-171
- Security standards are intentionally worded vaguely.
- NIST SP 800-171 has 110 requirements.
- Operating systems aren't compliant.
- Compliance configurations (STIGs) aren't optimized for HPC.
  - Most turn on SELinux. SELinux gives 50% performance degradation for metadata heavy work.

Unauthorized connection attempts blocked by the HPC firewall in a typical 24 hour period

Secure



The firewall is one of multiple layers of security.

Linux Integration

- **HPC systems have specialized hardware and software which must all work together.**
  - InfiniBand network
  - Parallel file system
  - Queue system
  - GPUs
  - Environment module system
  - Stable & secure
- **Often you choose an operating system that these support (i.e. Rocky Linux)**
- **Then you fight through other issues (i.e. getting software written on Ubuntu to work on Rocky)**
- **Configure kernel for usage**

Installed Applications Software

Apps

Anaconda packages	7828
Spack packages	4185
Perl modules	1132
LMOD modules	1020
Compiled from source	782
R modules	358
Singularity containers	50
Ruby gems	33

Apptainer

SINGULARITYCE

Singularity has split into two projects named Apptainer and SingularityCE (community edition).

- These numbers include libraries and utilities, as well as the core packages.
- This includes duplicates if same item installed in two versions of anaconda, etc.
- This does not include software installed in home directories.
- Does not include software from operating system distribution.

HPC Software installation without an installer

Apps

- **Much of the software is open source, research software.**
  - No professional development staff
  - Poorly documented
  - Incomplete list of prerequisites
  - Not tested on more than one Linux distribution
  - No examples, unit tests, or functional tests included
  - No support available
  - No bug fix or security updates
  - Source or precompiled for a different Linux version
  - May use make, cmake, script, none provided
- **You may want to change the compile flags to optimize for your hardware configuration.**

Compilers and Programming

Apps

- **Compilers**
  - GNU C/C++ Fortran 77/90/95
  - Intel C/C++ Fortran 77/90/95
  - NVIDIA HPC SDK C/C++ Fortran 77/90/HP
- **Parallel Programming**
  - Shared memory: OpenMP, Pthreads, Java threads
  - Distributed memory: MPI
  - Math libraries: ACML, GMP, MKL, Atlas, GSL
  - GPU: CUDA, OpenCL, OpenACC

Software install tools

Apps

- **Spack –** compiles software and prerequisites from source in an isolated environment
- **Python**
  – conda is better at managing prerequisites
  – pip compiles from source
- **LMOD environment modules**
  – Have multiple version of the same thing
  – Manage prerequisites
- **Containers (singularity, docker)**
  – Runs the user layer of a different OS
  – Can't fix kernel compatibility
  – Another prerequisite system
  – Can create new security issues

File storage

- Home directories are **100 GB** by default.
- Home directories can be increased to **7 TB** (7000 GB) by request, at no charge.
- Additional home directory space can be purchased.
- There are areas outside your home directory for installed software and publicly available genomes.
- **/scratch** areas store terabytes of data, which is automatically erased one week after the job completes. Visible to all nodes on the cluster.
- **/scratch-local** areas store 1-8 TB of data, but are erased when the job completes. /tmp in a separate partition
- Home directories are backed up, but scratch areas are not.

Parallel Shared File Systems

Files

- **HPC main stays**
  - Lustre
  - Spectrum Scale (formerly GPFS) - scratch
  - BeeGFS - home
  - Panasas
- **Honorable mention**
  - WekaFS – performant, limited quotas, cloud
  - Tiered systems – Spectralogic, DMF, TSM/Tivoli
  - DAOS, PNFS – still in early development stage
  - All SSD – Pure, Vast
- **Dogs (low performance)**
  - GlusterFS, NFS, EMC, everything else

Job Queue system: SLURM!

- Has nothing to do with the drink from Futurama.
- Operates much like a game of Tetris.
  - As jobs are submitted to the queue system, the scheduler picks available nodes that can complete your job. If a node is not available, it will wait until one is and then run your job.
- Can inform you when your job has begun, when it ends and if it errored anytime during the process.

Before the invention of a queue system, users would have to show up at their allotted time to use the supercomputer… which could be 2AM on a Sunday morning.

ASC queue list

Queue

Queue	Wall Time	Mem	# Cores
express	4:00:00	16gb	1-4
small	60:00:00	4gb	1-8
medium	150:00:00	16gb	1-16
large	360:00:00	120gb	1-64
bigmem	360:00:00	130-500gb	1-32
benchmark	24:00:00	120gb	1-64
gpu	360:00:00	20gb	1-2
class	12:00:00	64gb	1-60
sysadm	168:00:00	4tb	1-1000
special	1008:00:00	2tb	1-128

Running scripts

Queue

- Create a script to run the software, like this.

```
#!/bin/sh
source /opt/asn/etc/asn-bash-profiles-special/modules.sh
module load wrf/3.5.1_parallel
export OMP_NUM_THREADS=2
./compile em_b_wave
cd test/em_b_wave
./run_me_first.csh
./ideal.exe
./wrf.exe
```

- Submit with chmod & run_script
- Or teach every user sbatch

```
sbatch --qos=small -J lstestSCRIPT --begin=2022-09-09T21:03:53 --requeue --
mail-user=dyoung@asc.edu -o lstestSCRIPT.o%A --mail-
type=FAIL,END,TIME_LIMIT,FAIL,REQUEUE -t 60:00:00 -N 1-1 -n 2 --mem-per-
cpu=1000mb --constraint=milan
```

Error log file & jobinfo

Queue

- The queue creates a log file with a name ending with the job number, such as “water2comG16.o21239”
- Job performance information can be seen with the command “`jobinfo -j JOB_NUMBER`”

```
[screen 0: bash]                                                ⌥⌘4
asndcy@dmcvlogin4:~> jobinfo -j 756407
####################################################################
#         Alabama Supercomputer Center - SLURM Epilog
# Your username for this job is:                asndcy
# Your account for this job is:                 users
# Your group for this job is:                   analyst
# Your job ID is:                               756407
# Your job name is:                             parallel8comG16
# Your partition for this job is:               dmc
# Your architecture for this job is:            ivy
# Your job submit QOS is:                       large
# Your job ran on nodes:                        dmc28
# Your number of processors used:               8
# Your job was submitted at:                    2022-09-12T14:13:32
# Your job started at:                          2022-09-12T14:13:32
# Your job ended at:                            2022-09-12T14:22:39
# Your job elapsed time is:                     00:09:07
# Your job dedicated time is:                   01:12:56
# Your requested wall time is:                  15-00:00:00
# Your job cpu parallel efficiency is:          80.94%
# Your requested memory is:                     32000M
# Your max memory used:                         3121516K
# Your job memory efficiency is:                9.53%
# Your job state is:                            COMPLETED
# Your job exit code is:                        0:0
# Your requested resource was:                  billing=1,cpu=8,mem=32000M,node=1
# Your job commercial value is:                 $ 1.45867
####################################################################
asndcy@dmcvlogin4:~>
```

Queue scheduling strategies

Use

- Schedule by node, core, or lane (hyperthreads)
- FIFO – first in first out
- Partition based schedule
- Resource pool scheduling - DMC
- Reservation scheduling
- Cycle scavenging
- Gang scheduling

Queue scheduler terms

Use

- **Dynamic Priority –** quantifies job order
- **Static Priority –** attached to queue or hardware
- **Fairshare –** by user, research group, or project
- **Limits –** memory, time, jobs per user/running
  - Wall time vs CPU time
- **Constraints -** This job must run on A100 GPU
- **Backfill –** small jobs utilize resources without delaying high priority
- **Starving jobs –** size/priority prevents job from running… ever
- **Preemption –** kill a job to restart later
- **ACL** – Access Control List
- **Dozens more**

Popular Queue Systems

Use

- **SLURM –** open source with paid support, good scheduling, scalable, current growing pains, limited documentation – currently on DMC
- **PBS Pro –** mature, well supported, recent list of new features
- **OpenPBS, Torque/Moab, Torque/Maui, Torque –** others from the PBS code tree
- **Grid Engine –** was popular in academia when it was free
- **Condor –** Cycle scavenging
- **IBM Spectrum LSF –** very full featured and expensive

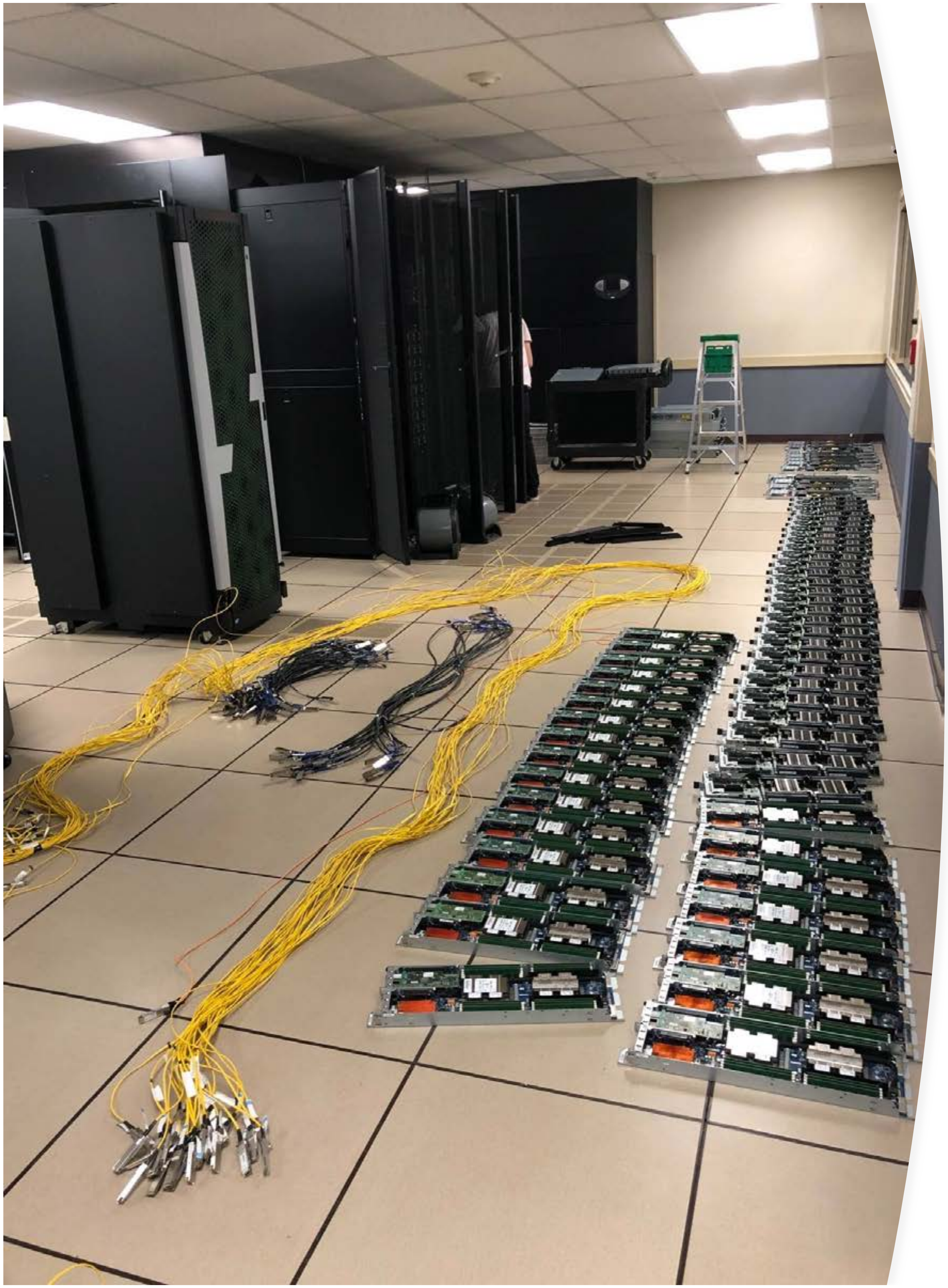
slurm workload manager

PBS Professional®

System reporting

Use

- **How many hours did user X use last month?**
- **How much did UAH use the system?**
- **How much did UAH Computer Science use?**
- **What is the dollar value of that use?**
- **How is storage usage growing over time?**
- **How many hours was each application used?**
- **How many hours were GPUs used?**
- **What are average wait times in queue?**
- **How do we allocate resources to users, and cut them off when exhausted?**
- **How do we bill users?**
- **How accurate? -** Jiffy counters, microstate

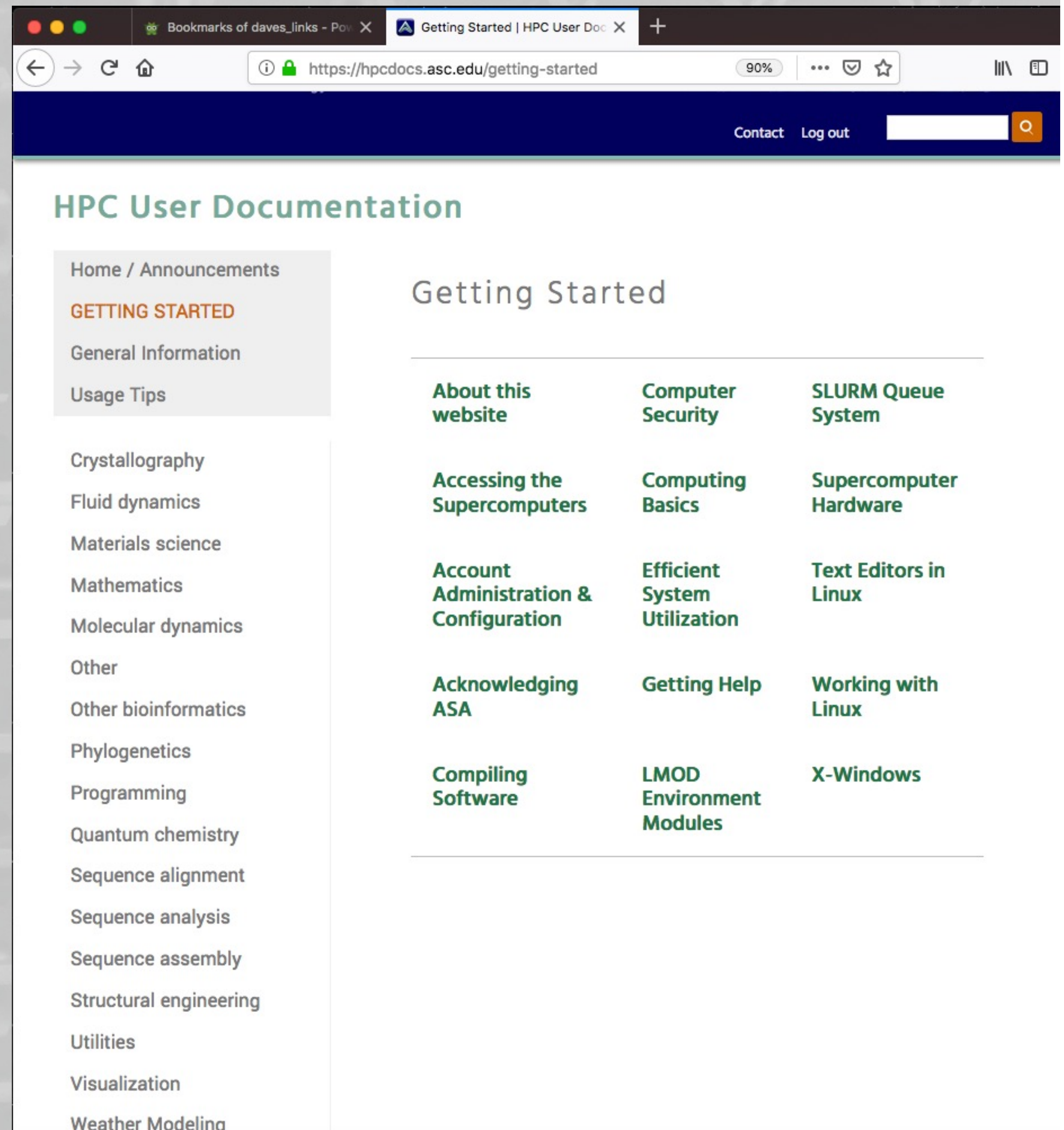


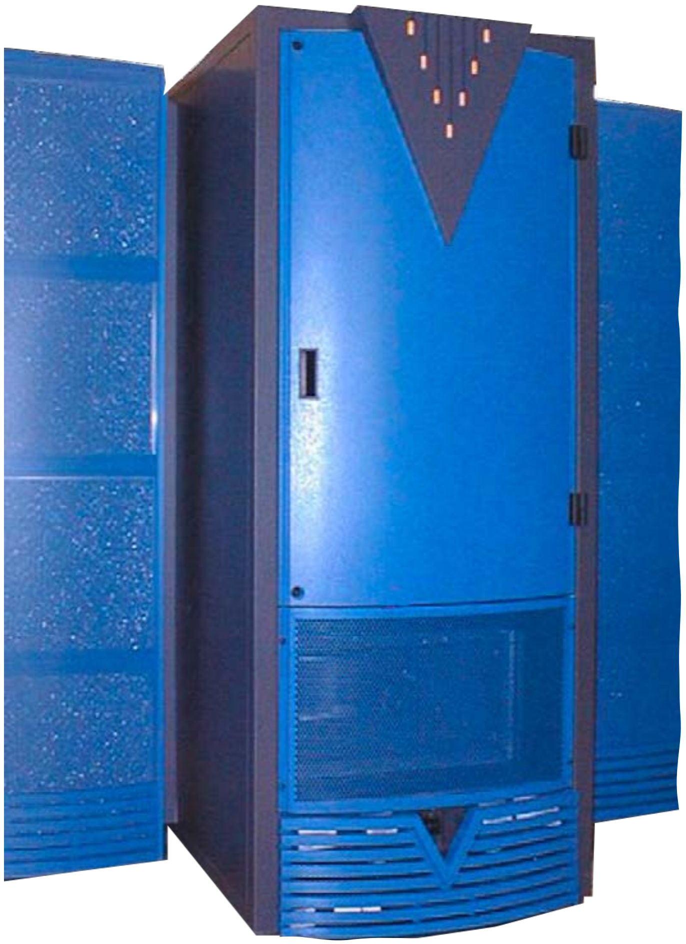
Ways to manage an HPC system

- **Unsupported**
  - Part time system administrator
- **Supported**
  - System administrators
  - Software analysts
- **Supported and Development**
  - System administrators
  - Software analysts
  - Software developers

User Support

- **Documentation**
  - https://hpcdocs.asc.edu
  - Man pages
  - Programming examples
  - Best practices white papers
- **Queue scripts**
  - A uniform front end for submitting all jobs to the queue that hides the details of queue command syntax.
- **Technical support staff**
  - Our HPC staff have degrees in chemistry, mathematics, business, and computer science.
  - **hpc@asc.edu**
- **Software installation**





Bleeding Edge

- **The Alabama Supercomputer Center has had**
  - A hypercube architecture
  - Early GPUs
  - FPGA chips
  - A scalar–vector machine
  - Cray SV1 was serial number 1
  - Cray XD1 tied Oak Ridge for first in the country on the same day.
  - Knights Landing processor test bed
- **We get NDA briefs, trial access to new hardware/software, sometimes even prototype hardware on loan**
- **Quantum computers? – not yet**

HPC Trends 1

Trend

- **The preferred Linux distribution changes over time** (support for IB, file system, etc.)
- **Higher power density**
  - Power distribution and cooling problems
- **More cores per server**
- **More GPU use**
- **More vector performance per core**
- **Web interface for queue system**
- **Intel and AMD GPUs**
- **Arms race mentality**
- **Auto parallelizing compilers**
- **More data / storage**

HPC Trends 2

Trend

- **Cloud computing**
  - Off site, on site, hybrid
  - Cloud burst
  - Cloud infrastructure servers
  - Cloud storage / backup
  - Cloud use of pay-by-hour software licenses
  - Disaster recovery
- **Checkpointing**
- **Specialized processors – AI chips**
- **Correctness – automated numerical analysis**
- **ARM chips, Power chips, RISC-V ???**
- **Composable computing**
- **Liquid cooling**

Possible futures for ASC Trend

- **Replace CentOS**
- **Queue web interface**
- **New / different file system / hardware**
- **More security**
- **More GPUs**
- **Data transfer node –** i.e. Globus
- **Faster networks**
- **Whole new cluster**
- **More technical staff**
- **Water cooling**
- **More containers**
- **Different monitoring software**

Supercomputer accounts

- Regular accounts can be obtained free for academic use at
  - https://www.asc.edu/hpc/ASA-HPC-Annual-Grant-Request-Form
- Class accounts can be obtained by having the instructor email **hpc@asc.edu**
- For commercial accounts or paid services, contact Nichole Gipson, ASA Client Services Assistant and E-rate Coordinator, at **ngipson@asc.edu** or **(334)659-4777**

What it costs

- Academic usage is free for faculty and students at the public universities in the state of Alabama. Academic use includes;
  - Course work
  - Thesis work
  - Research to be published

- CPU time can be purchased to do work that will become the property of the funding organization. This is defined as commercial work.

- Additional fees must be paid to use some software packages for commercial work.

Interactive Use Limits

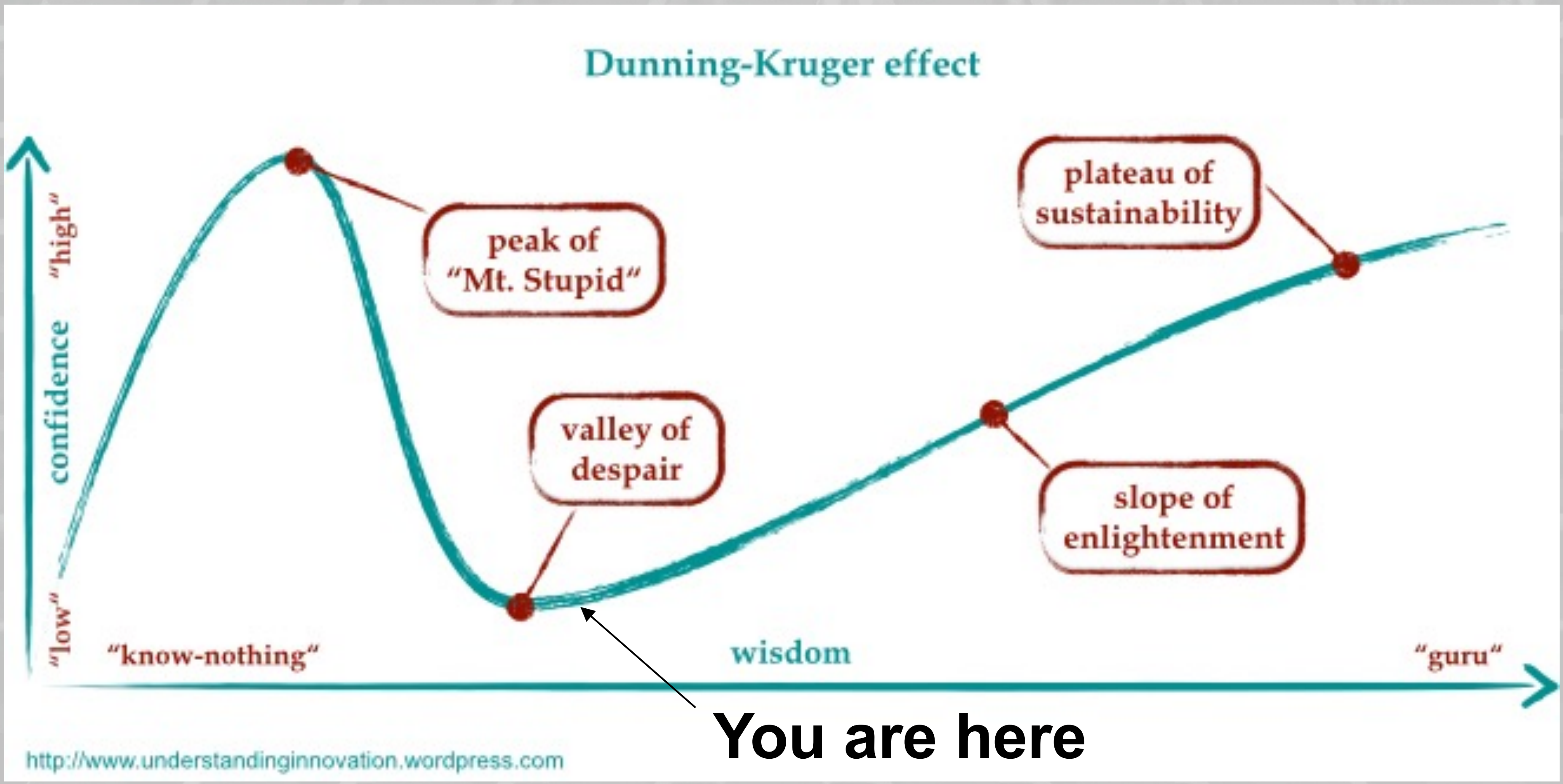
- The login nodes are meant for interactive work, such as setting up calculations and examining results. In order to enforce this, the following limits apply.
  – 10 minutes of CPU time
  – 10 GB of disk I/O
  – 4 GB of memory
  – Exceptions for scp, sftp, tar, gzip, compilers

- Jobs larger than this must be submitted to the job queue system.

Using graphic interfaces

- X-Windows is the remote access graphic interface for Linux.
- Free X-Windows clients include X11 (Linux or Mac), and MobaXterm (Windows).
- Some campuses have site licenses for commercial X-Windows clients.
- The responsiveness of X-Windows applications can be limited by network bandwidth and latency.

Learning Curve

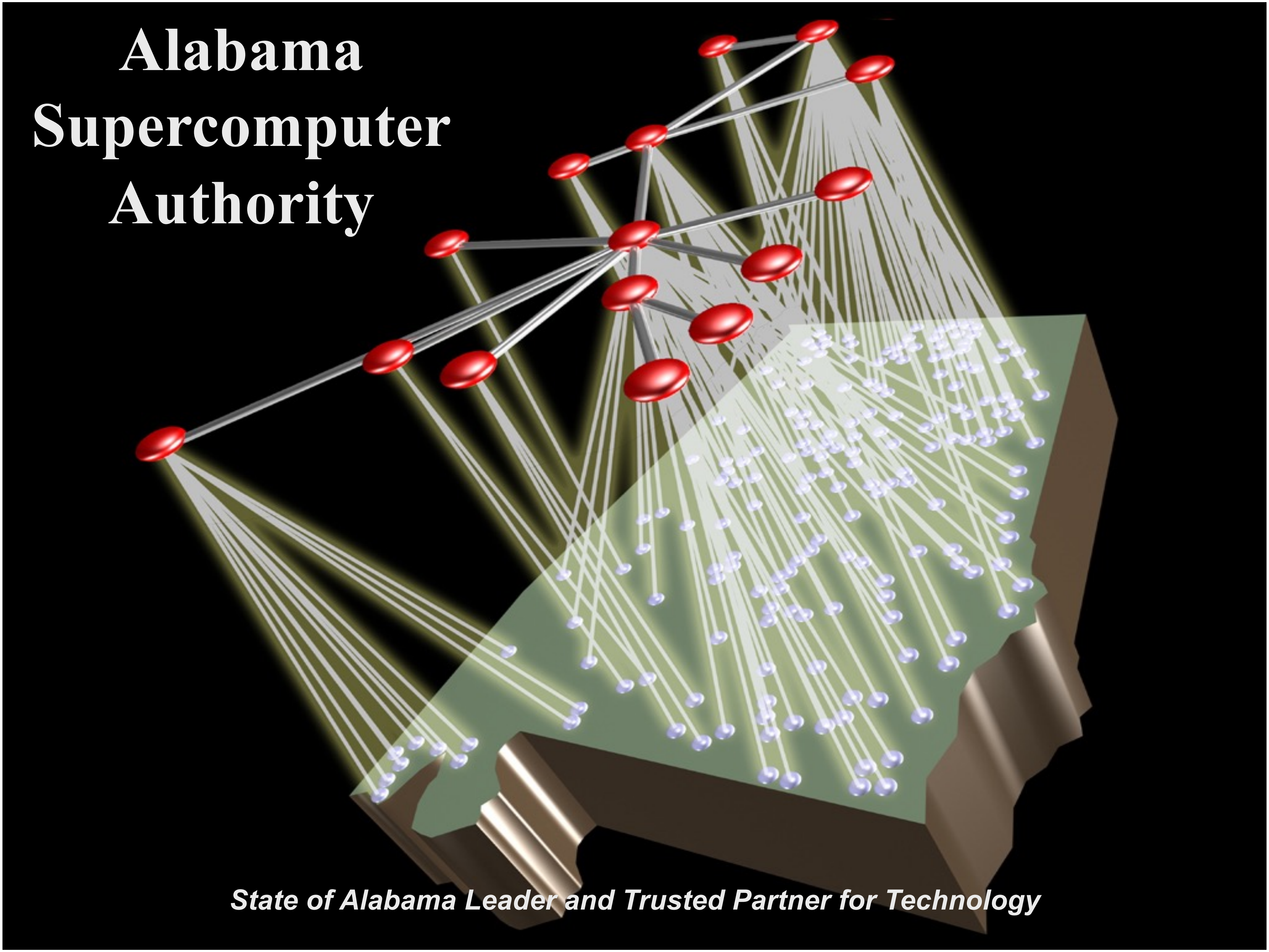


- It takes about eight months to turn a really experienced email/web Linux administrator into an HPC Linux administrator.

Summary

- The Alabama Supercomputer Authority provides a high performance computing system.
- This is free of charge for academic use by state funded educational institutions in Alabama.
- Setting up and maintaining and HPC system is a big, complicated job. It can also be an interesting career.
- Supercomputers are cool !

- Send your resume to
dyoung@asc.edu


Alabama
Supercomputer
Authority
State of Alabama Leader and Trusted Partner for Technology