HTCF Basics

CGS_SB High Throughput Computing Facility

Spring 2023

Brian Koebbe - Systems Manager

Eric Martin - Operations & Systems Specialist

Couch Biomedical Research Building - Room 5214

https://htcf-users.slack.com

What? Why?

A cluster is:

Set of computers (servers/nodes/hosts) that work together such that they can be viewed as a single system.

Why a cluster? Scale!

- Biological data is large and complicated.
- Many things done on a personal computer can be done in a fraction of the time on a cluster.

Brian Koebbe, Eric Martin

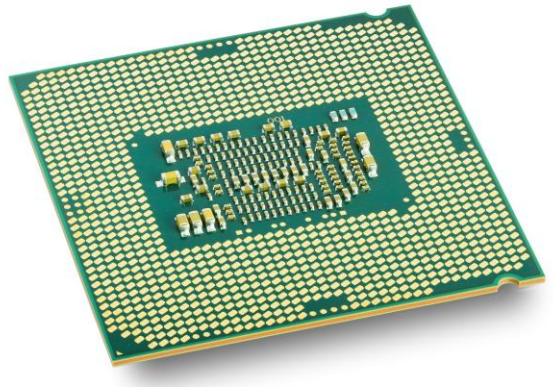
- Operating System Adminstrators
- HW & SW Architects / Engineers / Technicians:
  - Network
  - Storage
  - Computation
- R&D
- QA
- Support

What we'll cover

- **Cluster Anatomy**
  - **Server Hardware Anatomy**
  - **Cluster Systems Design/Layout**
  - **How the Systems Work Together**
- HTCF Specifics
  - Storage
  - Software
  - Scheduler
- Usage Example
- Q&A

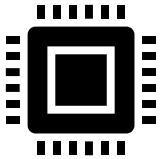
Server Anatomy

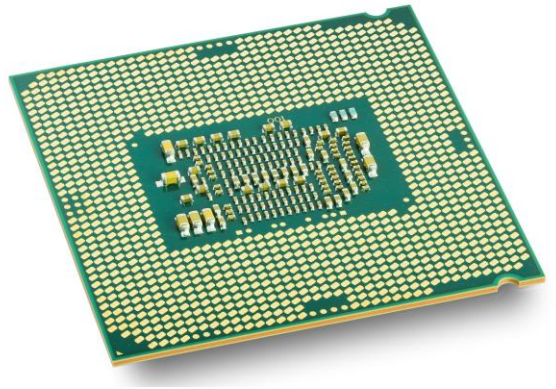
- **Processor**
  - **CPU** vs Cores
- Memory
  - Memory vs Storage
- Storage
  - Local Disk
- Network

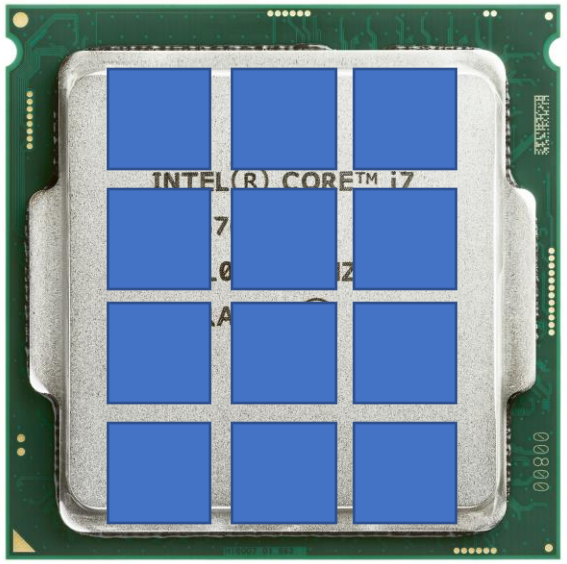


Server Anatomy

- **Processor**
  - CPU vs **Cores**
- Memory
  - Memory vs Storage
- Storage
  - Local Disk
- Network

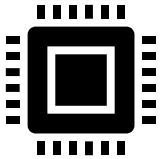


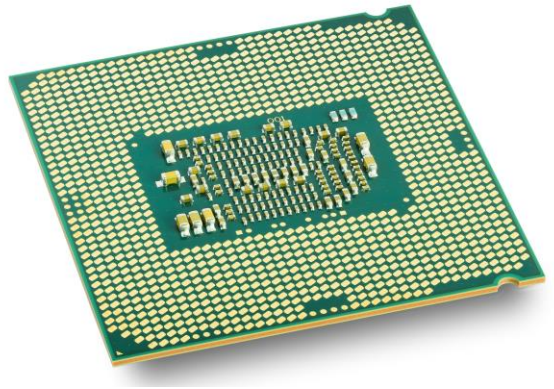


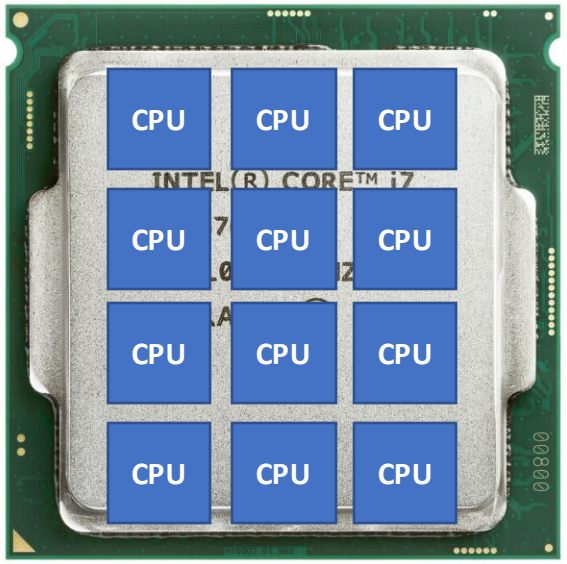


Server Anatomy

- **Processor**
  - **CPU = Cores**
- Memory
  - Memory vs Storage
- Storage
  - Local Disk
- Network

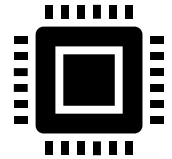






Server Anatomy

- Processor
  - CPU = Cores
- **Memory (RAM)**
  - **Memory** vs Storage
- Storage
  - Local Disk
- Network

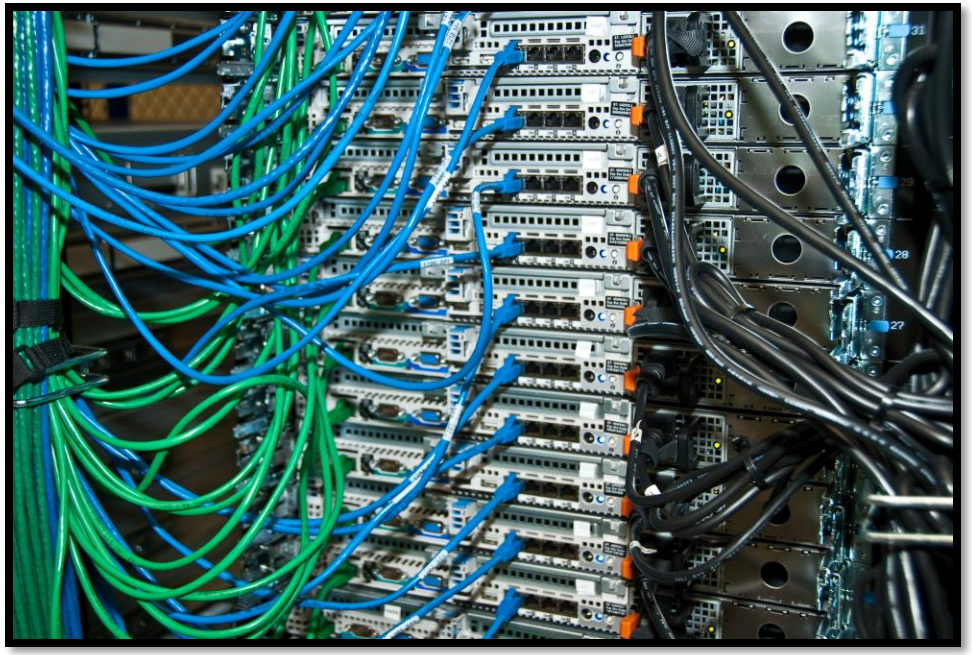


Server Anatomy

- Processor
  - CPU = Cores
- Memory (RAM)
  - Memory vs Storage
- **Storage**
  - **Local Disk**
- Network

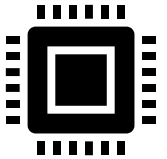
Server Anatomy

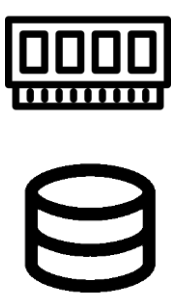
- Processor
  - CPU = Cores
- Memory (RAM)
  - Memory vs Storage
- Storage
  - Local Disk
- **Network**

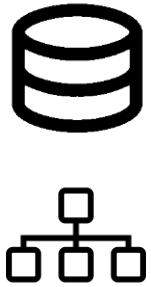


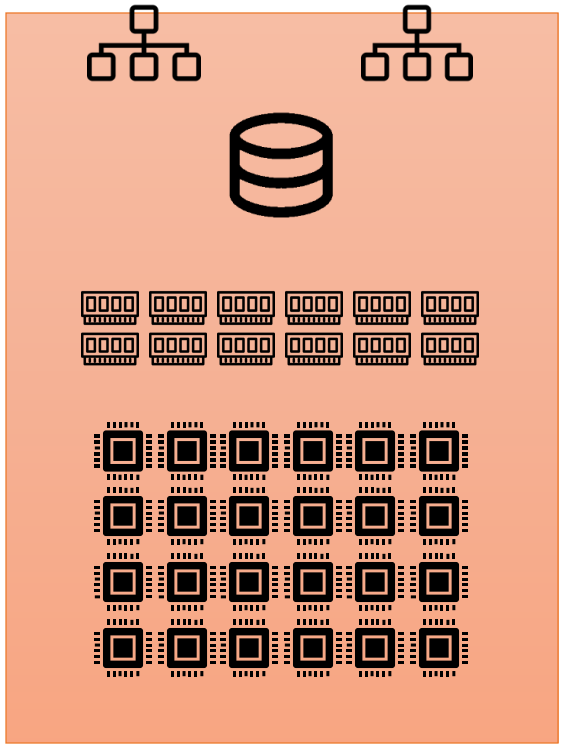
Node Anatomy

- Processor
  - 24 Cores
- Memory (RAM)
  - 768 GB
- Storage
  - 1 Disk
- Network
  - 2 connections
  - 3 IP Addresses



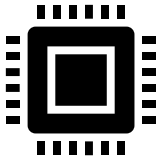


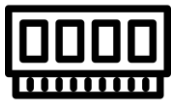


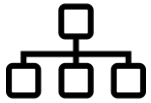


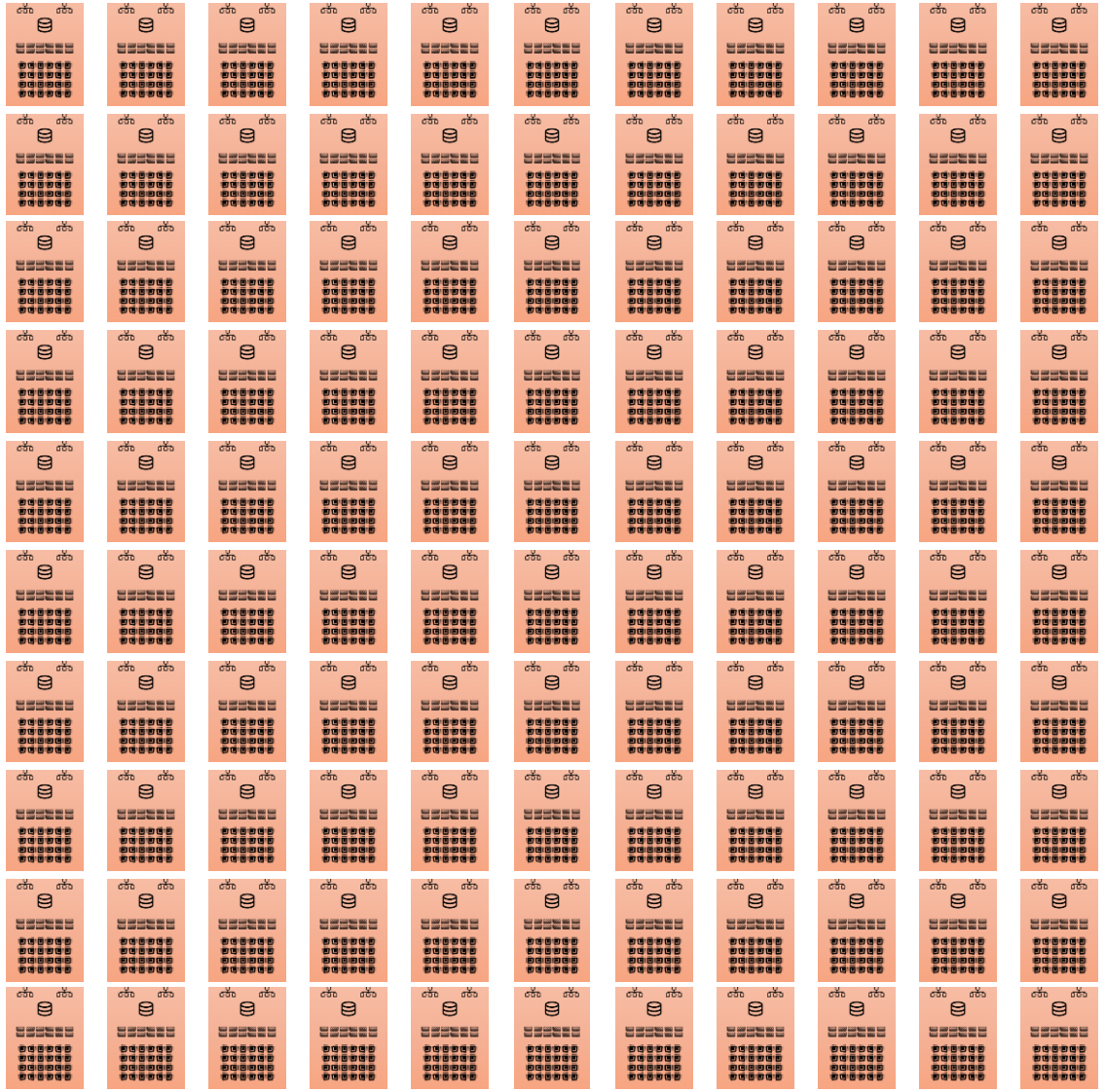
Node Anatomy: 100+ Servers

- Processor
  - 2000+ Cores
- Memory (RAM)
  - 32+TB
- Storage
  - 100 Disk
- Network
  - 200 connections
  - 300 IP Addresses



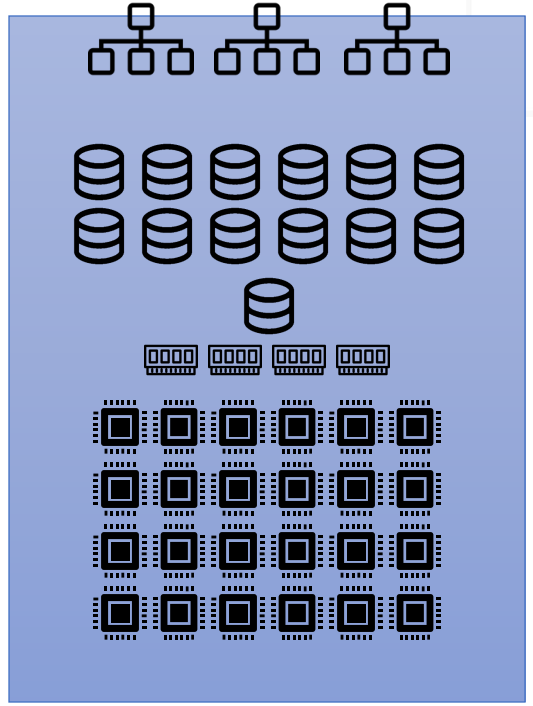






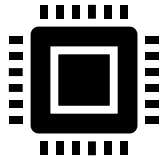
LTS Anatomy

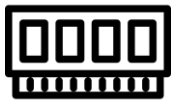
- Processor
  - 24 Cores
- Memory (RAM)
  - 196GB
- Storage
  - 12 Disk
- Network
  - 3 connections
  - 4 IP Addresses

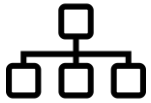


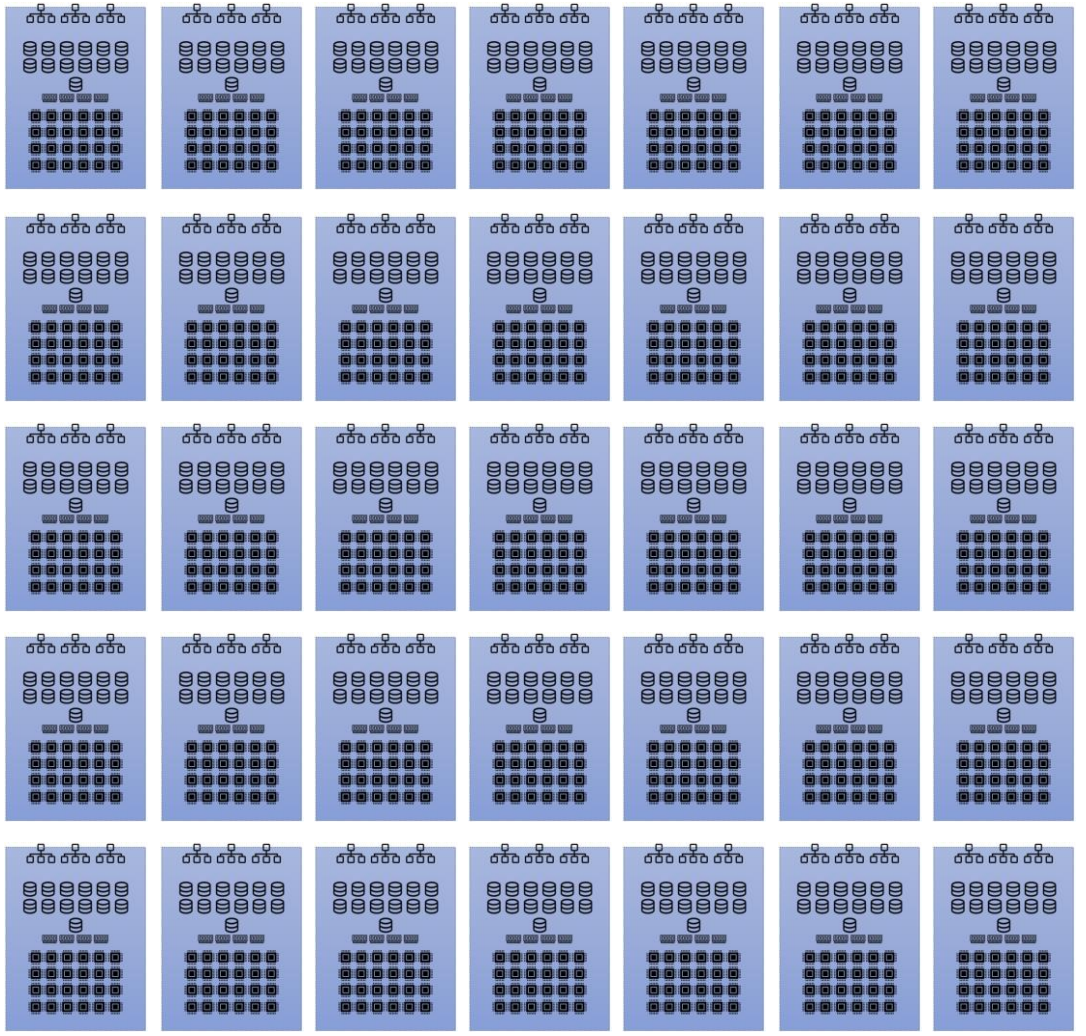
LTS Anatomy: ~35 Servers

- Processor
  - 840 Cores
- Memory (RAM)
  - 6.8 TB
- Storage
  - ~420 disks
- Network
  - 105 connections
  - 140 IP Addresses



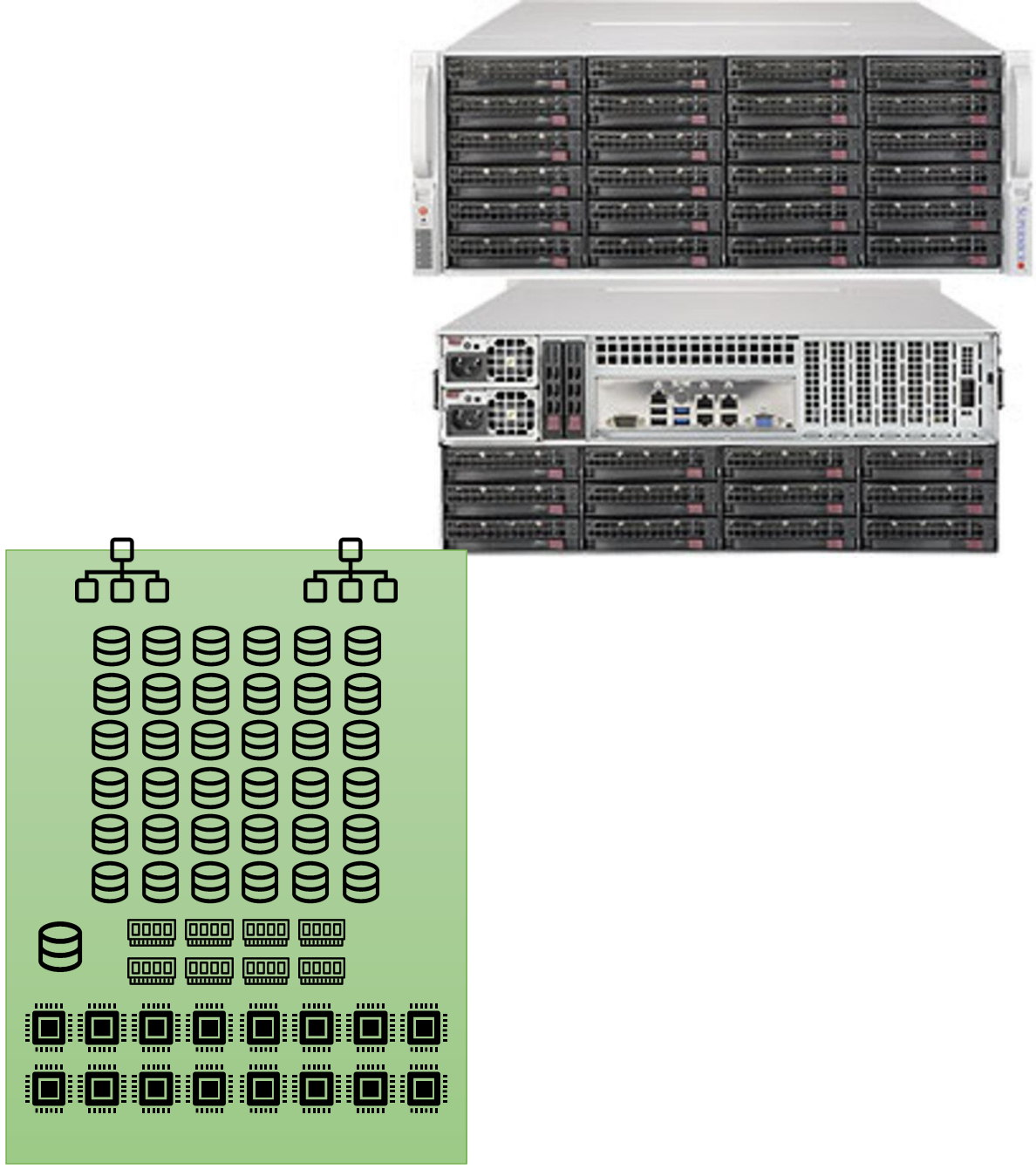






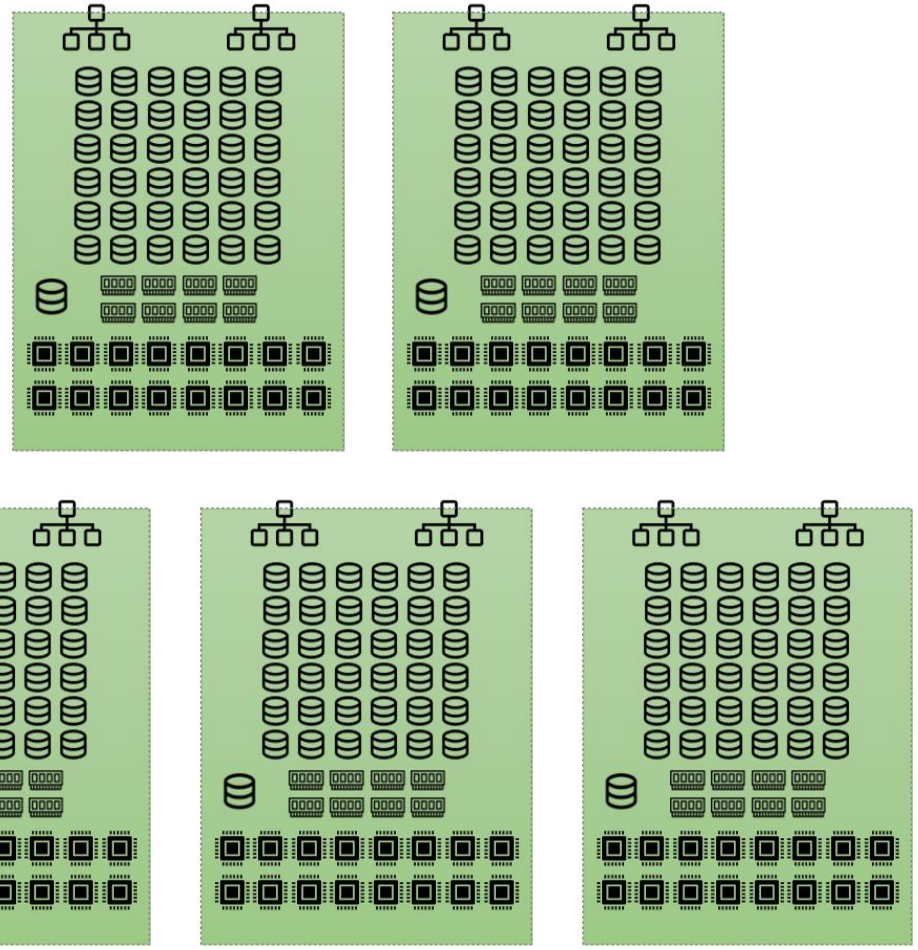
HTS Anatomy

- Processor
  - 16 Cores
- Memory (RAM)
  - 256 GB
- Storage
  - 36 disks
- Network
  - 2 connections
  - 3 IP Addresses

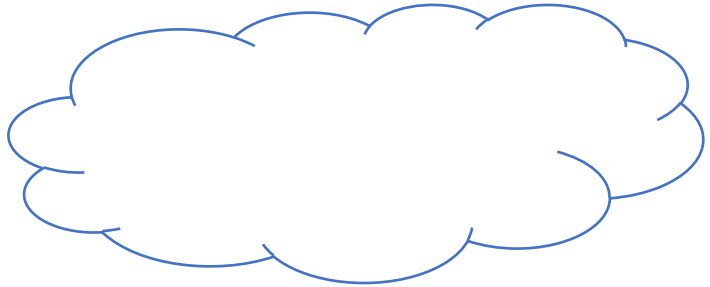


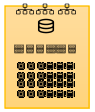
HTS Anatomy: 5 Servers

- Processor
  - 80 Cores
- Memory (RAM)
  - 1280 GB
- Storage
  - 180 disks
- Network
  - 10 connections
  - 15 IP addresses

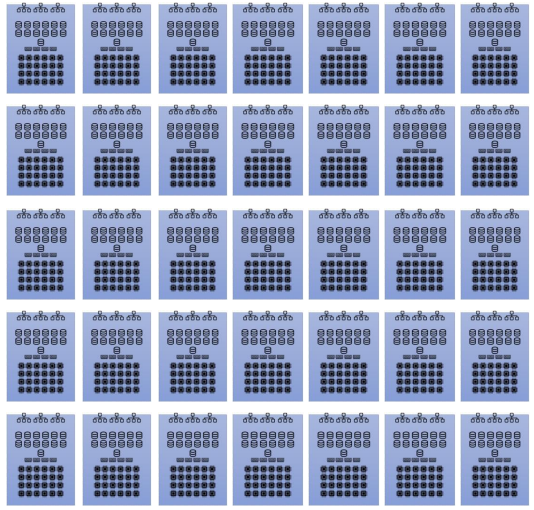


HTCF

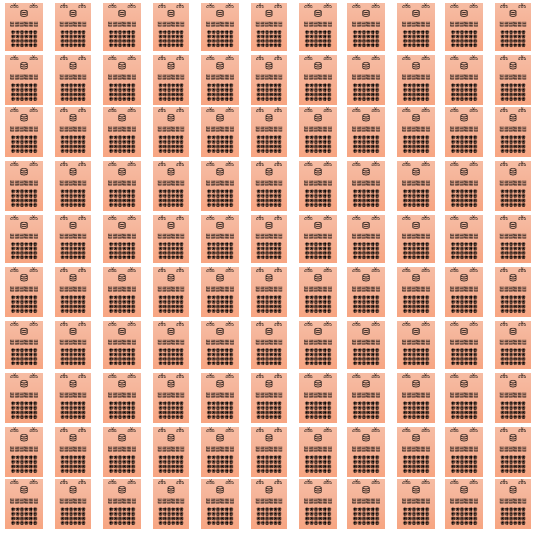




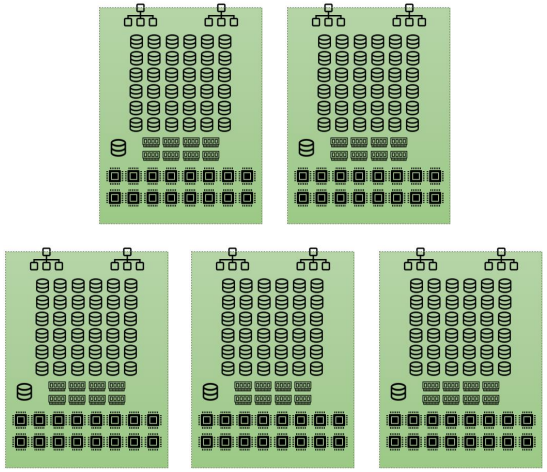
login.htcf.wustl.edu



LTS

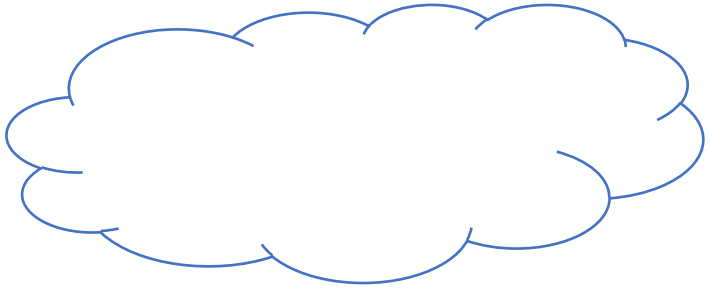


Nodes



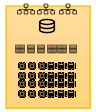
HTS

HTCF



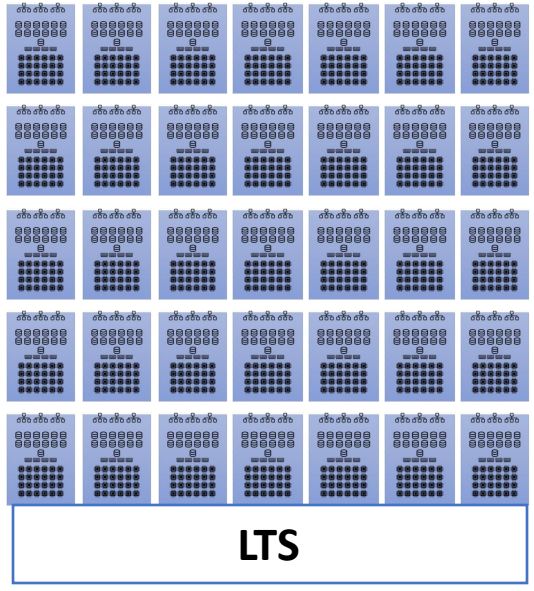
WUSTL

ADMIN



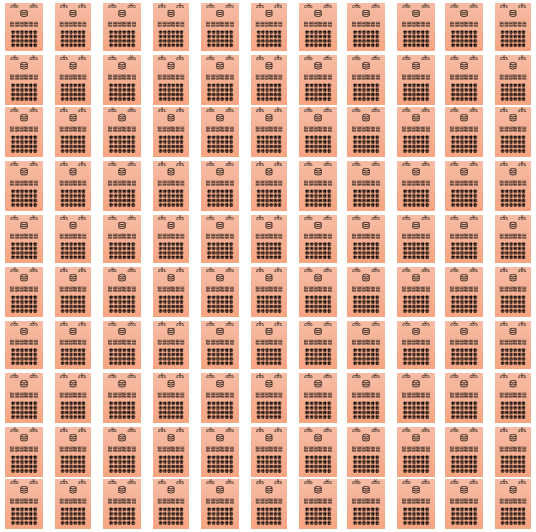
login.htcf.wustl.edu

HTS

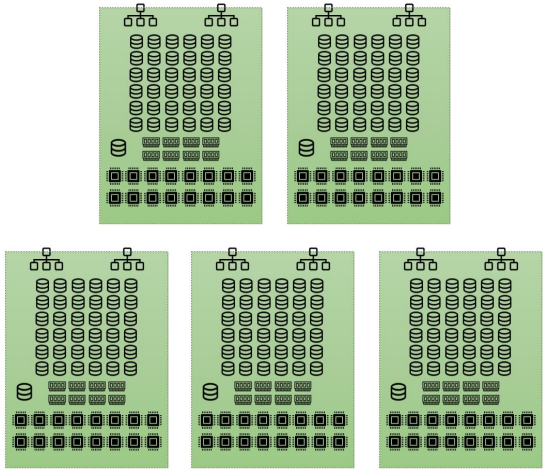


LTS

HTS Backend

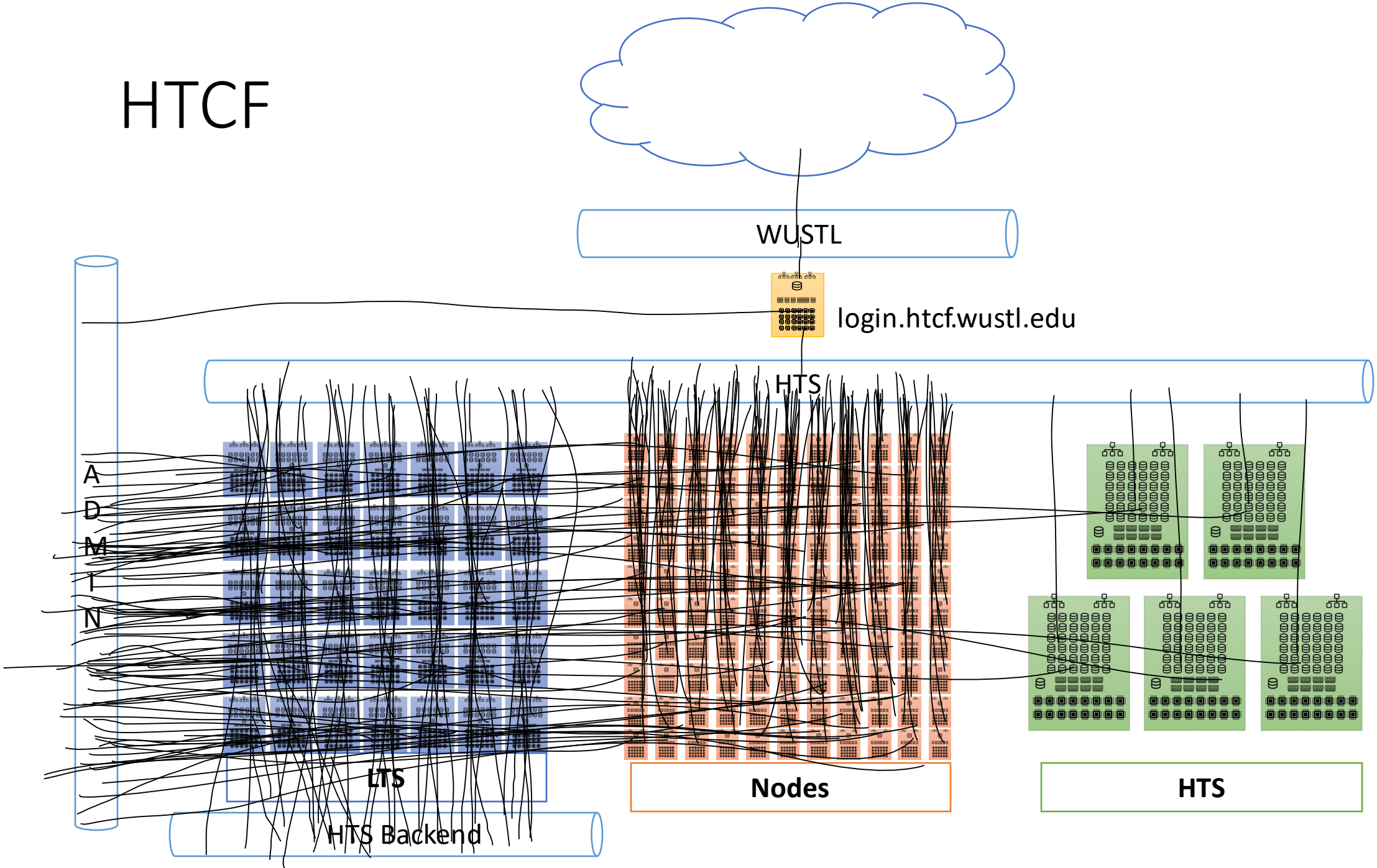


Nodes



HTS

HTCF

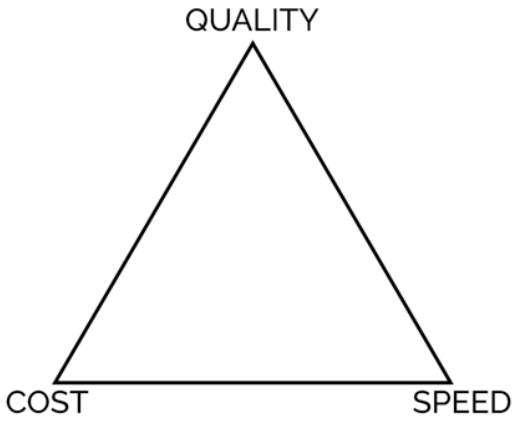

WUSTL
login.htcf.wustl.edu
HTS
ADMIN
LTS
HTS Backend
Nodes
HTS

What we'll cover

- Cluster Anatomy
  - Server Hardware Anatomy
  - Cluster Systems Design/Layout
  - How the Systems Work Together
- **HTCF Specifics**
  - **Storage**
  - **Software**
  - **Scheduler**
- Usage Example
- Q&A

Storage

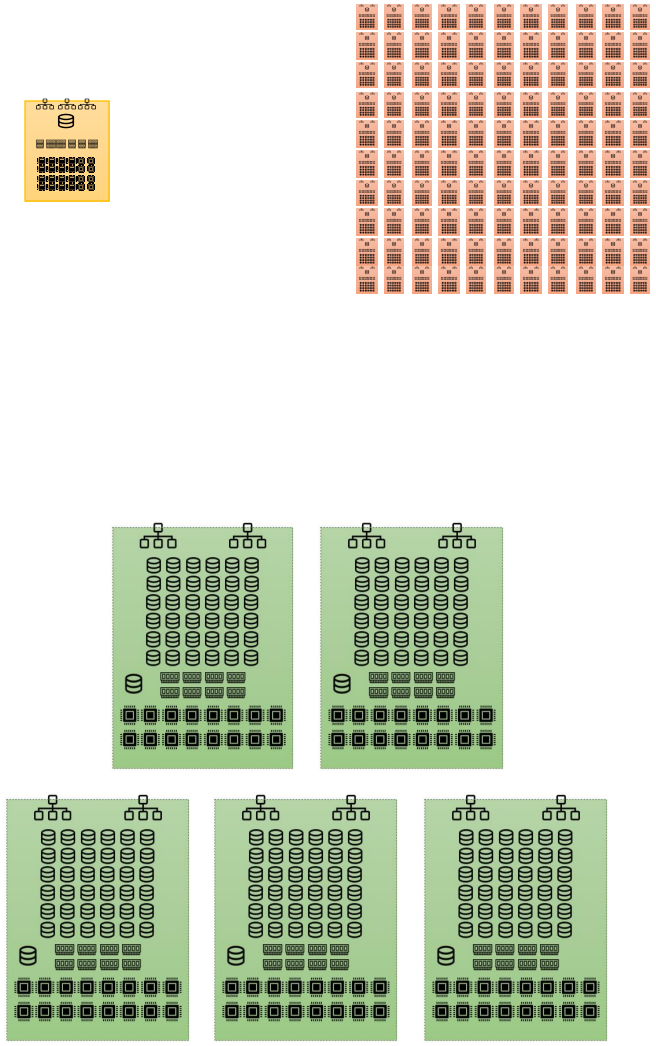
- Faster / Slower
- Temporary / Permanent (backups)
- Larger / Smaller
- Good for small files / large files
- High / Limited accessibility
- **All are shared**



HTS: /scratch

Temporary working space for jobs.

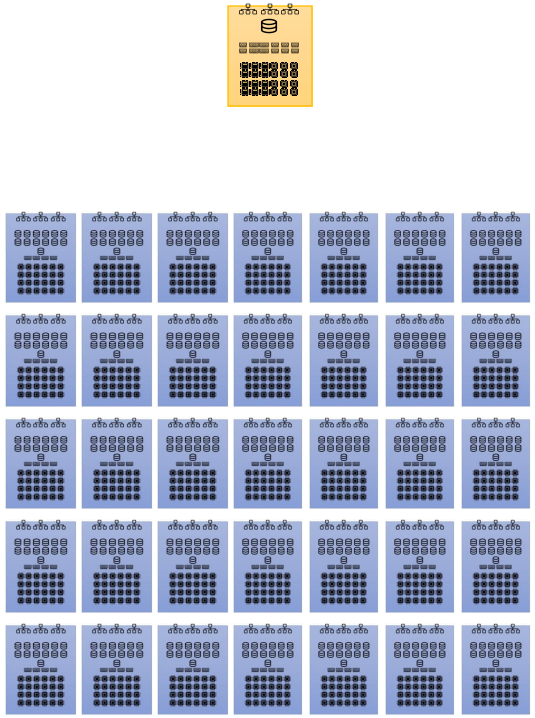
- Fastest with large reads/writes
- Slow with small reads/writes
- Size: 750TB
  - 84 million files/directories
- Speed: 10+ GB/s
- NOT backed up
- 2TB per account



LTS: filesystem storage

/lts

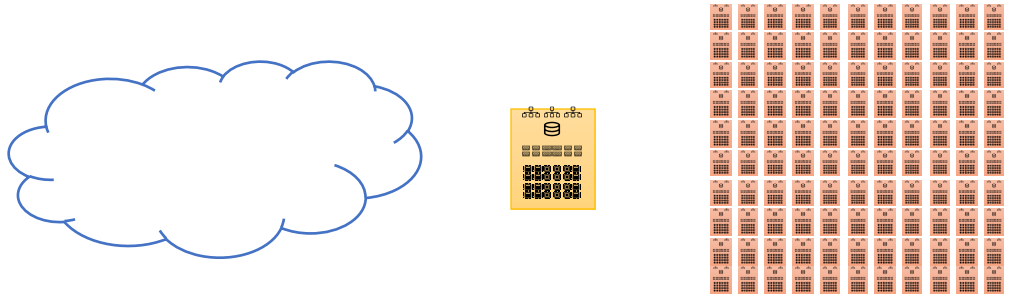
For raw data/finished product data



Slower, permanent, backed up

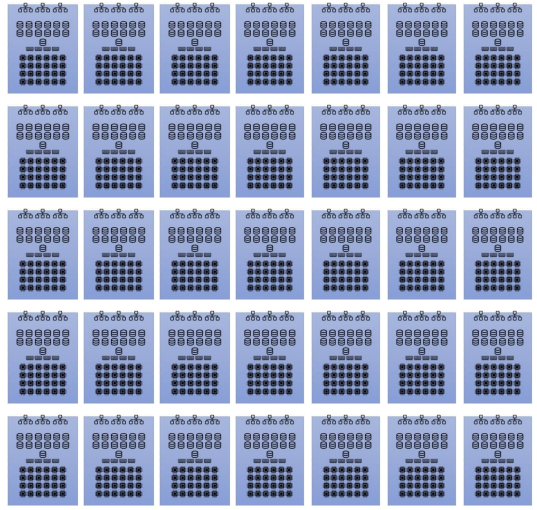
- Accessed via login server
- Size: 4.2PB
- Speed: ~150 MB/s
- Purchased in "buckets"
  - 1TB increments
  - Recommended max 15TB
  - $7.77/TB/month

LTS: Object Storage

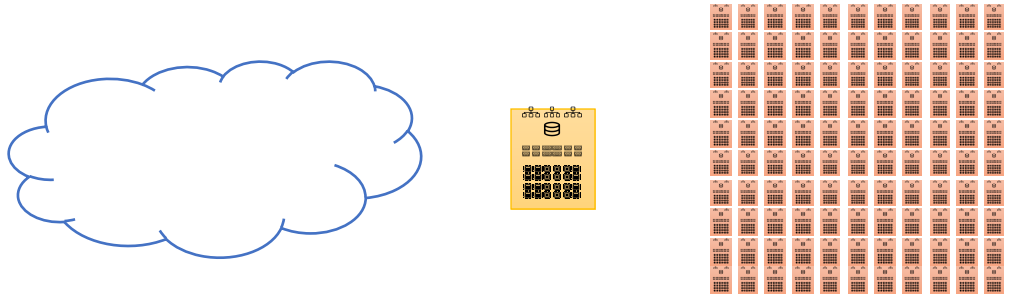


Like LTS filesystem storage but accessible via the S3 (cloud storage) protocol, not filesystem.

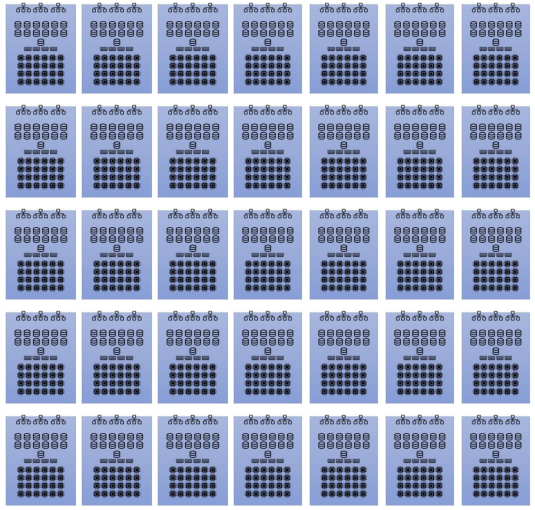
- Accessible
  - Internet
  - Login
  - Nodes
- Size: 4.2PB
- Speed: 300+MB R/W
- No bucket size limit
  - 1TB increments
  - $7.77/TB/month



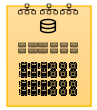
LTS: Object Storage

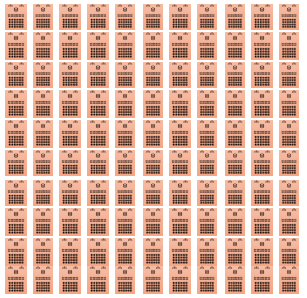


```
[janesmith@login ~]$ s3cmd mb s3://jane1
[janesmith@login ~]$ s3cmd ls

2023-03-01 12:34  s3://jane1

[janesmith@login ~]$ s3cmd put somefile.fastq s3://jane1/somefile.fastq

somefile.fastq -> s3://jane1/somefile.fastq  [1 of 1]
123456 of 123456   100% in    2s    51.75 kB/s  done

[janesmith@login ~]$ s3cmd put --recursive dir1 dir2 s3://jane1/somewhere/

File 'dir1/file1-1.txt' stored as 's3://jane1/somewhere/dir1/file1-1.txt' [1 of 5]
File 'dir1/file1-2.txt' stored as 's3://jane1/somewhere/dir1/file1-2.txt' [2 of 5]
File 'dir1/file1-3.log' stored as 's3://jane1/somewhere/dir1/file1-3.log' [3 of 5]
File 'dir2/file2-1.bin' stored as 's3://jane1/somewhere/dir2/file2-1.bin' [4 of 5]
File 'dir2/file2-2.txt' stored as 's3://jane1/somewhere/dir2/file2-2.txt' [5 of 5]
```



LTS: Reference

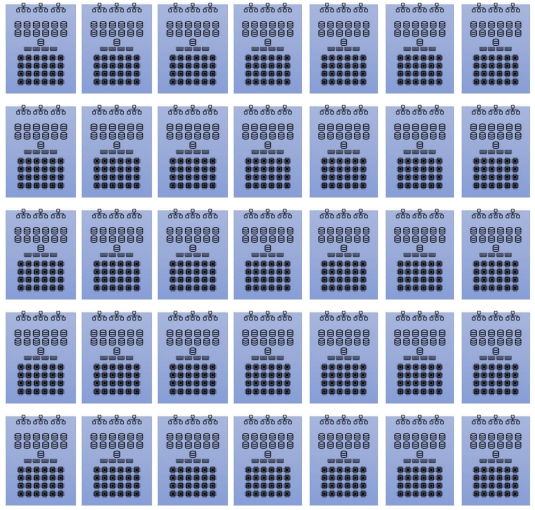




/ref

Reference space for software and read-only data sets

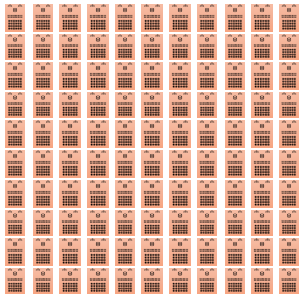
- Faster small reads than /scratch
- Slower writes
- Size: 4.2PB
- Speed: 3GB/300MB R/W
- Each groups begins with 1TB
  - 1TB increments
  - $7.77/TB/month



Storage: /home

Personal space for shell configuration files, small scripts

- Slow
- Accessible via login, nodes
- 20GB
- Not meant to handle job workloads



Software

- Installed and managed by labs
- Can be installed in various ways
  - Software specific instructions
  - Spack/Easybuild
  - R Libraries/Python Modules
  - Conda
  - Compiling source code
- Install Location
  - /ref/<lab>/software/
  - $HOME (limited space, slow)

Software

Some software installation is easy

Much software installation is **hard**

Management software (spack, conda) may help,

but **no** system is foolproof.

Slack, Google, software mailing lists, forums are your friend.

We are also here to help.

Software: Notebook tools

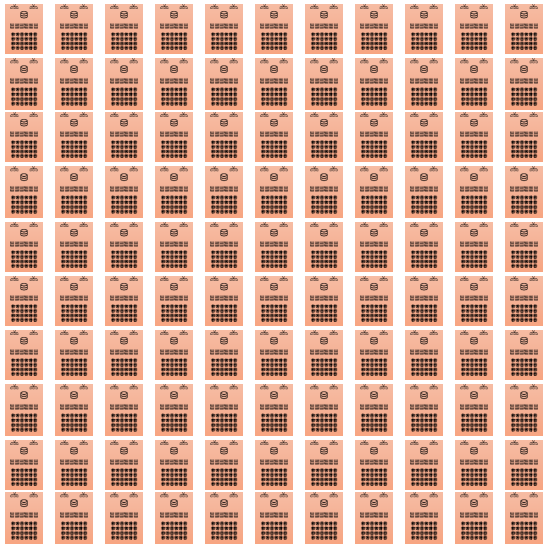
RStudio, Jupyter Lab

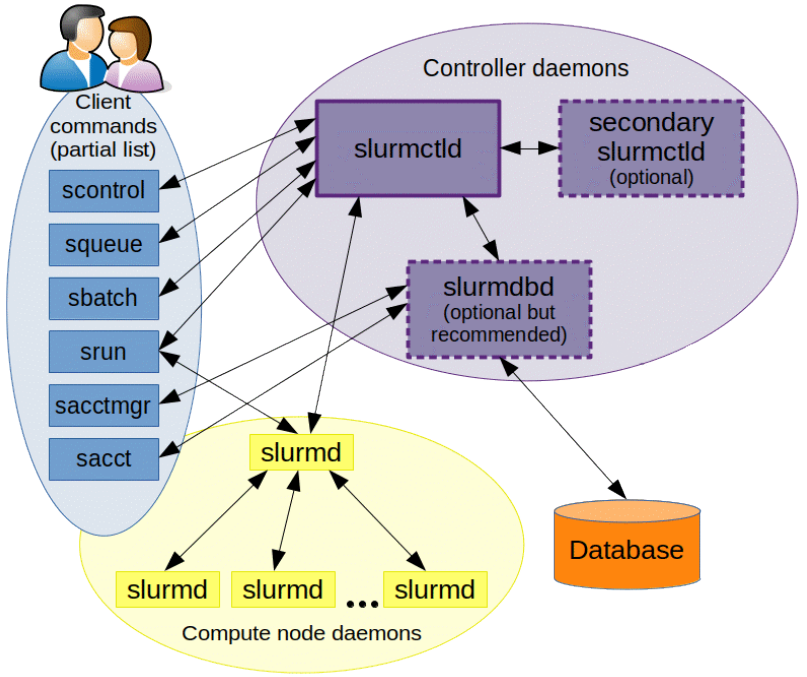
- Used on the cluster for interactive manipulation and analysis typically done by batch processing.
- These tools are really web servers.
- Web browser interface requires web access (via SSL tunnels) to nodes.
- Therefore, their use can be complicated and error prone.
- Installation (of RStudio) can be even more complicated and error prone.

Research is being done to see if we can make this simpler.

Slurm

- Resource management and job scheduling system
- Allocates requested resources
- Provides a framework for starting, executing, and monitoring work
- Arbitrates contention for resources by managing a queue of pending work





Example Usage

Will use bowie2 to align some sequencing reads to a set of long reference sequences.

- Login
- Grab some cluster resources (CPUs, Memory) with an interactive job
- Use Spack to install bowtie2 into /ref/jslab/software (LTS)
- Download HG19 data into /ref/jslab/data (LTS)
- Make Slurm sbatch script
- Launch job
- Monitor progress
- Look at completed job stats

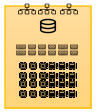
Logging in

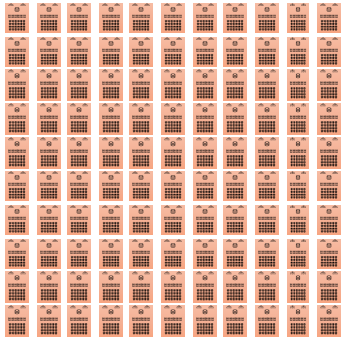
WUSTL Key Support:

https://wustlkey.wustl.edu

```
wustlkey: janesmith
lab: jslab

ssh janesmith@login.htcf.wustl.edu
[janesmith@login ~]$
```

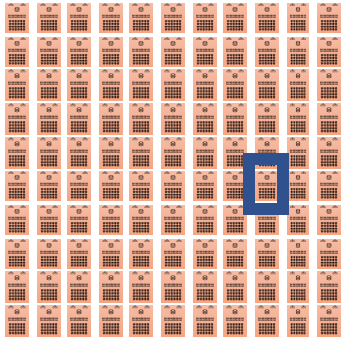
Slurm: Start Interactive Job



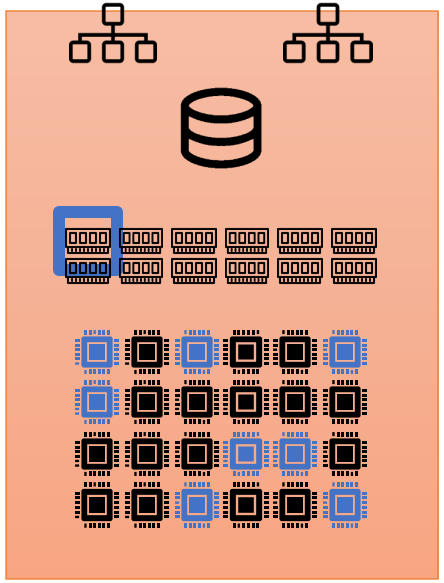


```
srun -c 8 --mem=8G -p interactive --pty /bin/bash -l
```

Slurm: Start Interactive Job

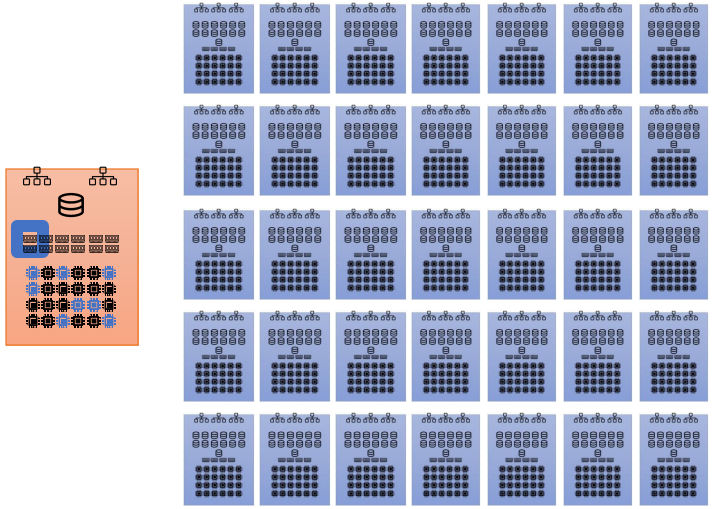


```
srun -c 8 --mem=8G -p interactive --pty /bin/bash -l
```



Install software if needed

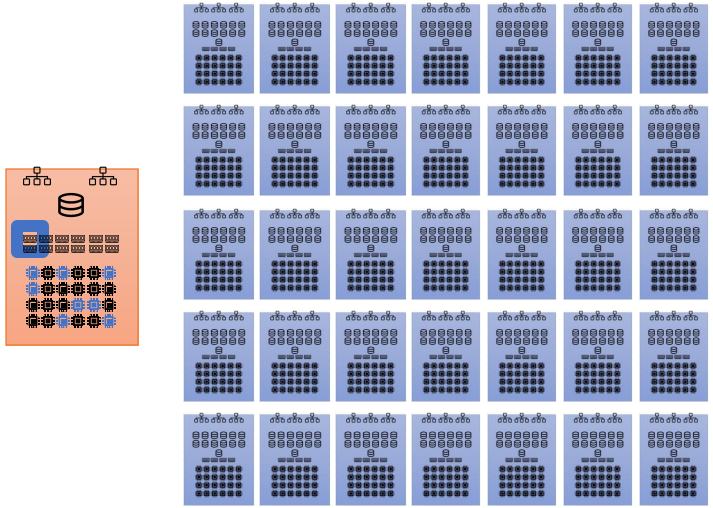
```
[janesmith@n067 ~]$ spack find bowtie2
==> No package matches the query: bowtie2
[janesmith@n067 ~]$
```



Install software

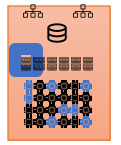
```
[janesmith@n067 ~]$ spack spec -I bowtie2
Input spec
--------------------------------
 -   bowtie2

Concretized
--------------------------------
 -   bowtie2@2.4.2%gcc@8.5.0 build_system=makefile arch=linux-rocky8-x86_64
[+]      ^intel-tbb@2021.7.0%gcc@8.5.0~ipo+shared+tm build_system=cmake build_type=RelWithDebInfo cxxstd=default arc
[+]          ^cmake@3.24.3%gcc@8.5.0~doc+ncurses+ownlibs~qt build_system=generic build_type=Release arch=linux-rocky
[+]          ^hwloc@2.8.0%gcc@8.5.0~cairo~cuda~gl~libudev+libxml2~netloc~nvml~oneapi-level-zero~opencl+pci~rocm buil
6_64
[+]              ^libpciaccess@0.16%gcc@8.5.0 build_system=autotools arch=linux-rocky8-x86_64
[+]                  ^libtool@2.4.7%gcc@8.5.0 build_system=autotools arch=linux-rocky8-x86_64
[+]                      ^m4@1.4.19%gcc@8.5.0+sigsegv build_system=autotools patches=9dc5fbd,bfdffa7 arch=linux-rock
[+]                          ^libsigsegv@2.13%gcc@8.5.0 build_system=autotools arch=linux-rocky8-x86_64
[+]                  ^util-macros@1.19.3%gcc@8.5.0 build_system=autotools arch=linux-rocky8-x86_64
[+]              ^libxml2@2.10.1%gcc@8.5.0~python build_system=autotools arch=linux-rocky8-x86_64
[+]      ^perl@5.36.0%gcc@8.5.0+cpanm+shared+threads build_system=generic arch=linux-rocky8-x86_64
[+]          ^berkeley-db@18.1.40%gcc@8.5.0+cxx~docs+stl build_system=autotools patches=26090f4,b231fcc arch=linux-r
[+]          ^bzip2@1.0.8%gcc@8.5.0~debug~pic+shared build_system=generic arch=linux-rocky8-x86_64
[+]              ^diffutils@3.8%gcc@8.5.0 build_system=autotools arch=linux-rocky8-x86_64
[+]          ^gdbm@1.23%gcc@8.5.0 build_system=autotools arch=linux-rocky8-x86_64
[+]      ^python@3.10.8%gcc@8.5.0+bz2+ctypes+dbm~debug+libxml2+lzma~nis~optimizations+pic+pyexpat+pythoncmd+readline
m=generic patches=0d98e93,7d40923,f2fd060 arch=linux-rocky8-x86_64
[+]          ^expat@2.4.8%gcc@8.5.0+libbsd build_system=autotools arch=linux-rocky8-x86_64
```

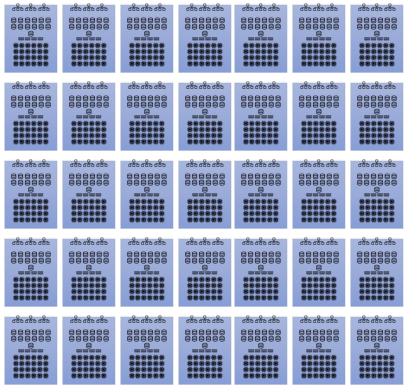
Install software



```
[janesmith@n067 ~]$ spack install -j ${SLURM_CPUS_ON_NODE} bowtie2

==> Installing bowtie2-2.4.2-4b5rw7qdsxvipu5rhvcrxdapmsbw6cx4
==> No binary for bowtie2-2.4.2-4b5rw7qdsxvipu5rhvcrxdapmsbw6cx4 found: installing from source
==> Using cached archive: /ref/htcfadmin/software/spack/spack-0.19.1/var/spack/cache/_source-cache/archive/4c/4cc555eeeeb8ae2d47aaa1551f3f01
0e90f462865027e.zip
==> No patches needed for bowtie2
==> bowtie2: Executing phase: 'edit'
==> bowtie2: Executing phase: 'build'
==> bowtie2: Executing phase: 'install'
==> bowtie2: Successfully installed bowtie2-2.4.2-4b5rw7qdsxvipu5rhvcrxdapmsbw6cx4
  Fetch: 0.23s.  Build: 2m 50.25s.  Total: 2m 50.48s.
[+] /ref/htcfadmin/software/spack/spack-0.19.1/opt/spack/linux-rocky8-x86_64/gcc-8.5.0/bowtie2-2.4.2-4b5rw7qdsxvipu5rhvcrxdapmsbw6cx4
[janesmith@n067 ~]$
```

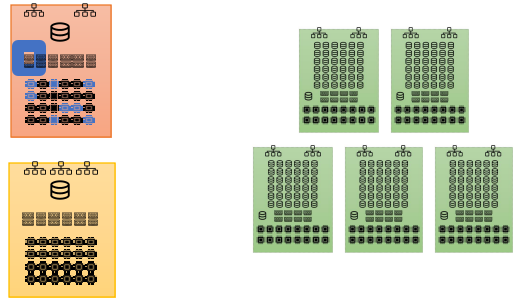
Download reference sequences





```
[janesmith@n197 data]$ wget https://genome-idx.s3.amazonaws.com/bt/hg19.zip
--2023-03-09 13:32:11--  https://genome-idx.s3.amazonaws.com/bt/hg19.zip
Resolving genome-idx.s3.amazonaws.com (genome-idx.s3.amazonaws.com)... 54.231.132.57, 3.5.29.224, 52.216
.50.1, ...
Connecting to genome-idx.s3.amazonaws.com (genome-idx.s3.amazonaws.com)|54.231.132.57|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3694403333 (3.4G) [application/zip]
Saving to: ‘hg19.zip’

hg19.zip                  100%[=====================================>]   3.44G  20.0MB/s    in 2m 43s

2023-03-09 13:34:54 (21.7 MB/s) - ‘hg19.zip’ saved [3694403333/3694403333]

[janesmith@n197 data]$ unzip hg19.zip
Archive:  hg19.zip
  inflating: hg19.1.bt2
  inflating: hg19.2.bt2
  inflating: hg19.3.bt2
  inflating: hg19.4.bt2
  inflating: hg19.rev.1.bt2
  inflating: hg19.rev.2.bt2
  inflating: make_hg19.sh
[janesmith@n197 data]$
```

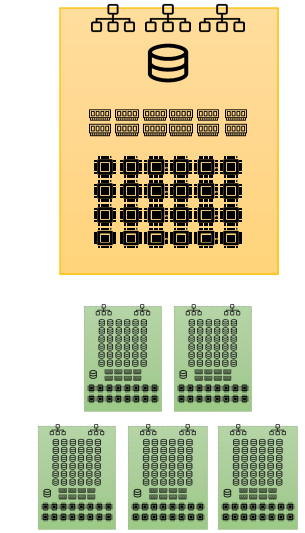
Create myjob.sbatch:



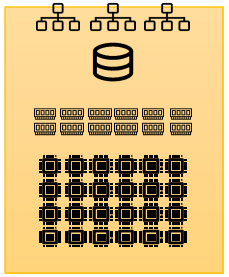
```
#!/bin/bash

#SBATCH -c 8
#SBATCH --mem=8G

eval $(spack load --sh bowtie2)

bowtie2 -x /ref/jslab/data/hg19 -p ${SLURM_CPUS_ON_NODE} /scratch/jslab/SRR23720981.fq > results.out
```

Submit job

```
[janesmith@login work]$ sbatch myjob.sbatch
```

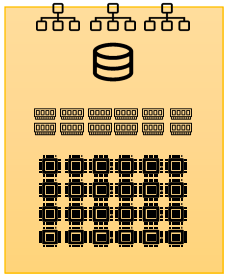


Monitor job: squeue



```
[janesmith@login ~]$ squeue -u janesmith
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           3124734   general job2.sba janesmith  R       0:41      1 n003
[janesmith@login ~]$
```

Check efficiency: seff



```
[janesmith@login ~]$ seff 3124734
Job ID: 3124734
Cluster: htcf
User/Group: janesmith/janesmith
State: COMPLETED (exit code 0)
Nodes: 1
Cores per node: 8
CPU Utilized: 00:40:32
CPU Efficiency: 95.30% of 00:42:32 core-walltime
Job Wall-clock time: 00:05:19
Memory Utilized: 3.21 GB
Memory Efficiency: 40.18% of 8.00 GB
```

Expectations

- Understand the job
  - What is being done?
  - What's the bottleneck?
  - What files are being read/written?

- Ask the question "Is this polite?"
- When there's a problem, help us help you...
  - Thoroughly describe the issue
    - Job ID
    - Full path to sbatch and files in question
  - Research
    - explain what's been done thus far to find a solution

What next

- **https://htcf-users.slack.com**
- We're available for lab meetings
- Future talks?

Q&A