MAKING THE MOST OF FARBER: STORAGE & STANDBY

Dr. Jeffrey Frey
University of Delaware, IT

Dare to be first.

GOALS

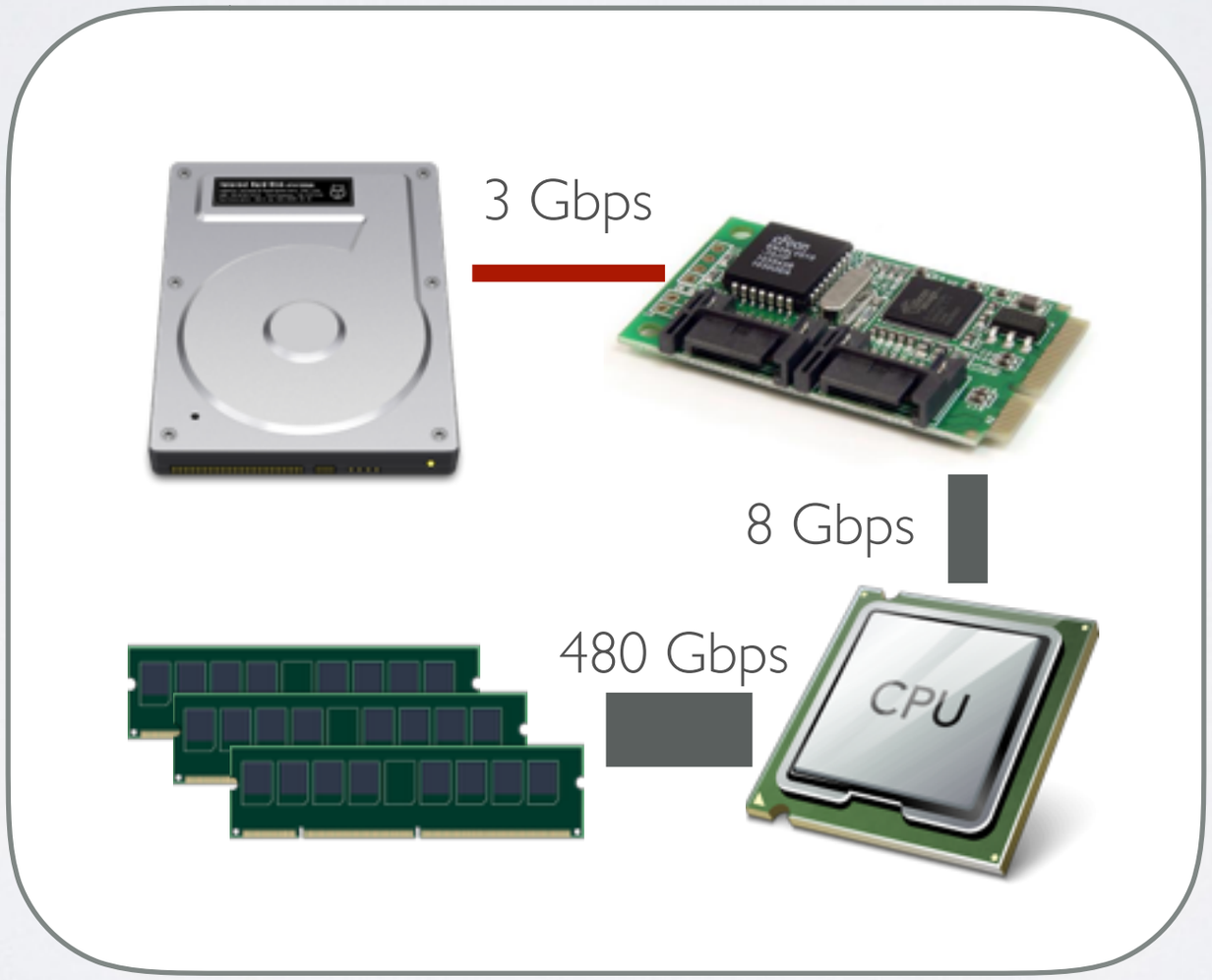
- Storage
  - What storage resources are available on Farber?
  - What are their differences? Limitations?
  - When should each be used?

GOALS

- Storage
- Standby queues
- Regular "owned" queues: limitations, properties
- Design tenets of a community cluster…
- …and how they lead to standby queues

STORAGE

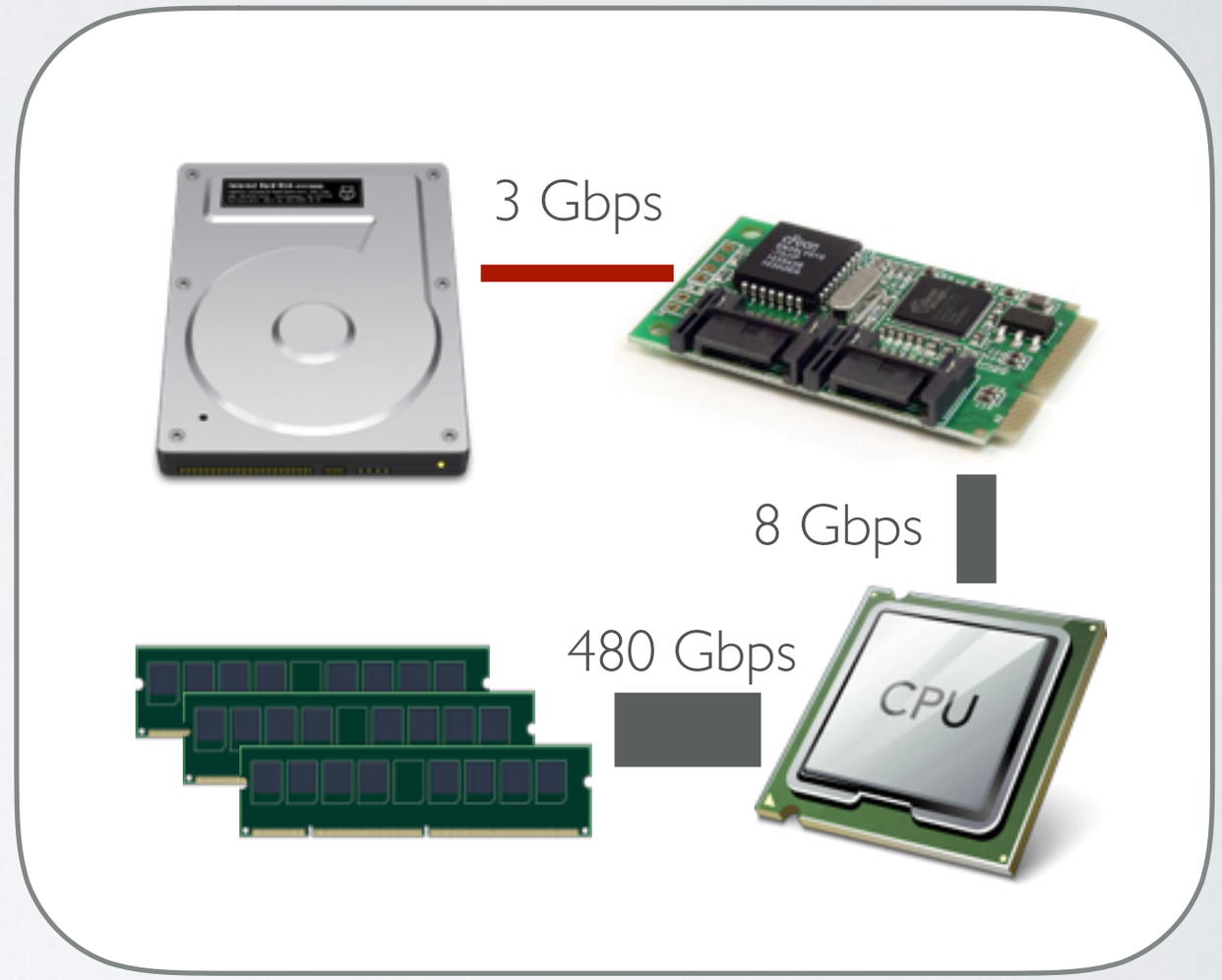
- Local storage is relatively simple



Dare to be first.
UNIVERSITY OF DELAWARE

STORAGE

- Maximum bandwidth governed by slowest component
  - SATA hard disk
  - 3 Gbps → 300 MBps



Base signaling rate:
3 Gbps = 3000 Mbps

Encoding: 8data+2parity
80%(3000 Mbps)(1 B/8 b)
= 300 MBps

Dare to be first.
UNIVERSITY OF DELAWARE

DEMO

- Use Linux "dd" command to write 671 MB to local disk:

```
[root@n001 ~]# dd if=/dev/zero of=/scratch/bigzero bs=4k count=163840
163840+0 records in
163840+0 records out
671088640 bytes (671 MB) copied, 0.505866 s, 1.3 GB/s
```

Question:

Didn't you just say the limit was ~ 300 MB/s? How can the computer report 4x that?

Answer:

Caching. The OS accepts the data and stores it in memory, then moves it to disk as quickly as it can. Node has 64 GB of RAM, so easy to accommodate 671 MB cache.

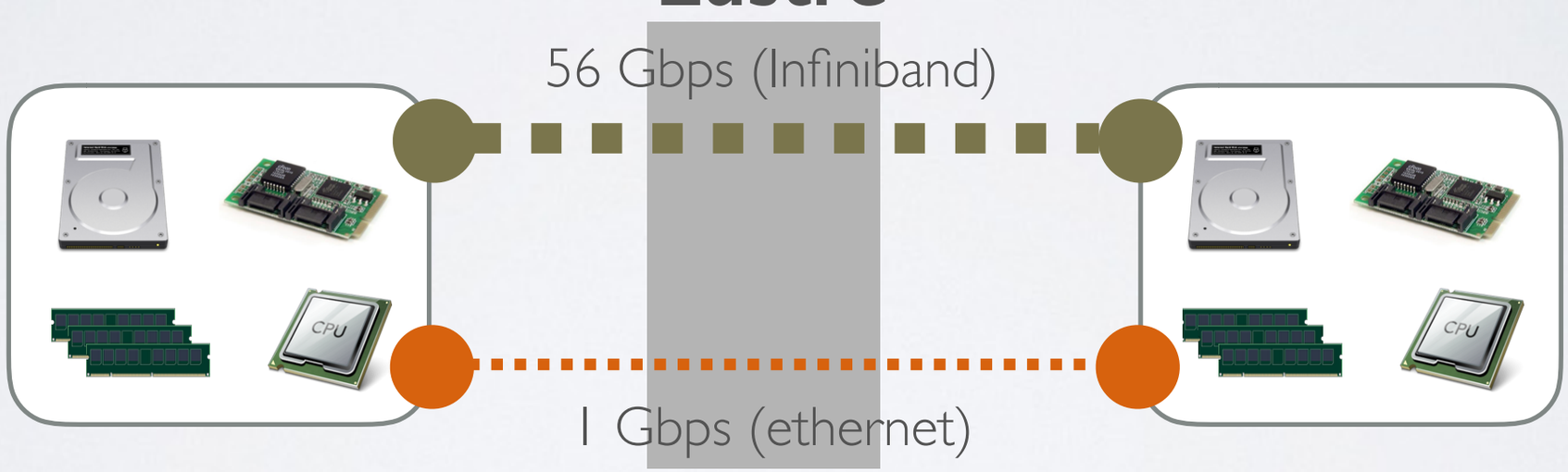
DEMO

- Use Linux "dd" command to write 671 MB to local disk:

```
[root@n001 ~]# dd if=/dev/zero of=/scratch/bigzero bs=4k count=163840
163840+0 records in
163840+0 records out
671088640 bytes (671 MB) copied, 0.505866 s, 1.3 GB/s
[root@n001 ~]# dd if=/dev/zero of=/scratch/bigzero bs=4k count=163840
163840+0 records in
163840+0 records out
671088640 bytes (671 MB) copied, 6.48713 s, 103 MB/s
[root@n001 ~]# dd if=/dev/zero of=/scratch/bigzero bs=4k count=163840
163840+0 records in
163840+0 records out
671088640 bytes (671 MB) copied, 6.31107 s, 106 MB/s
```

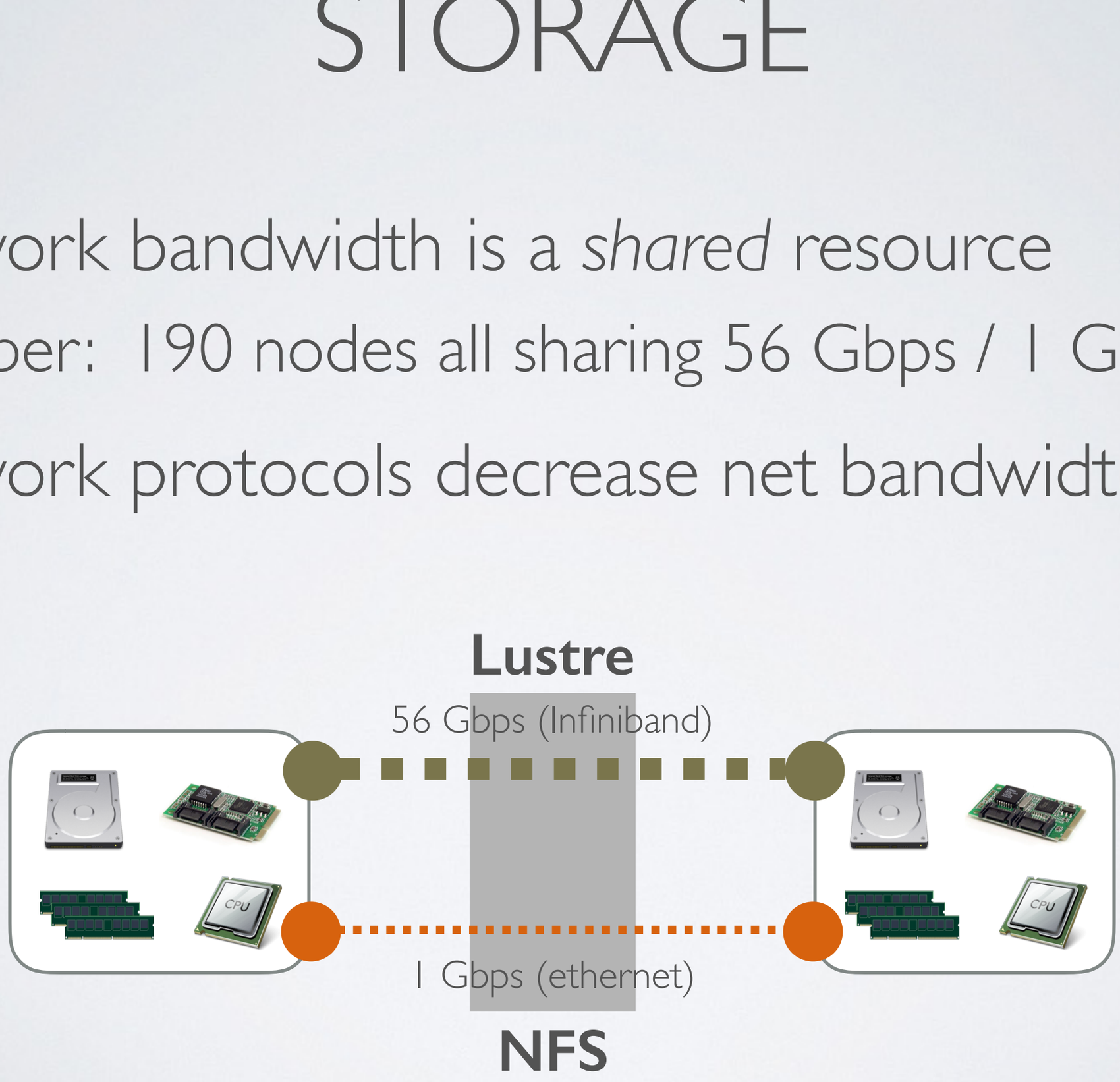
STORAGE

- Network-attached storage introduces additional bottlenecks



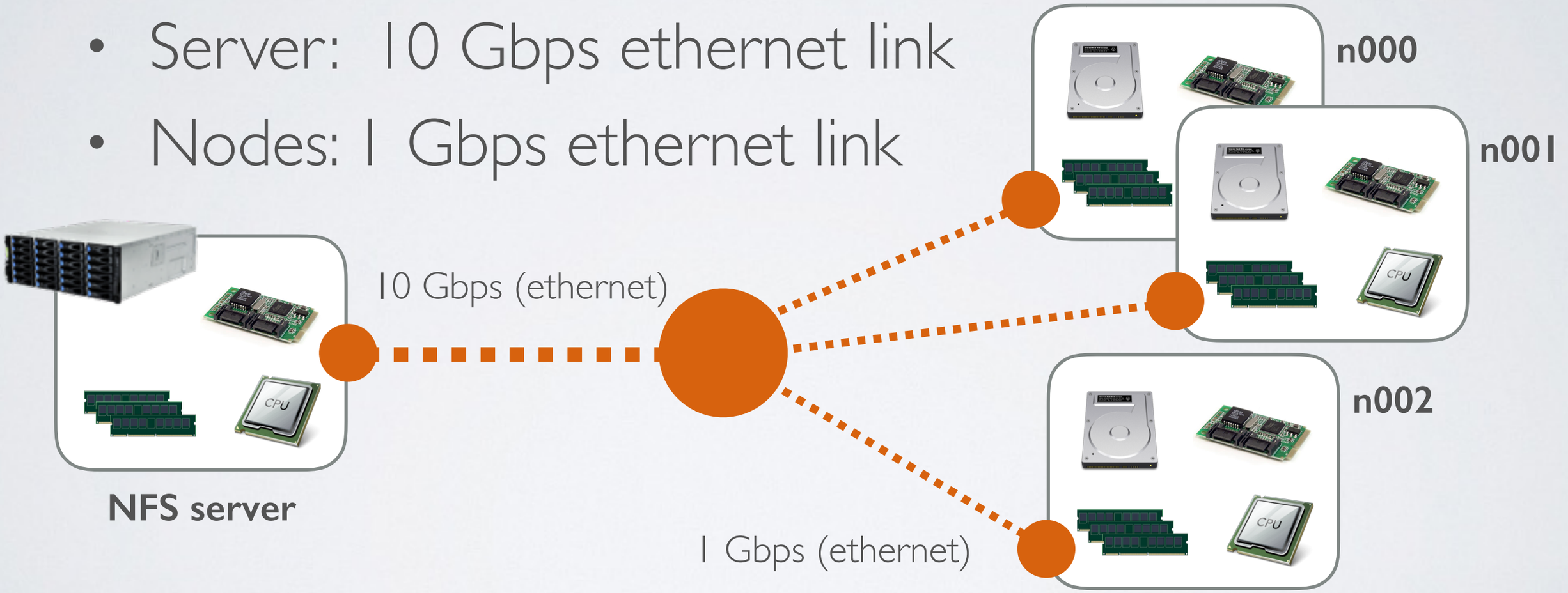
STORAGE

- Network bandwidth is a *shared* resource
  - Farber: 190 nodes all sharing 56 Gbps / 1 Gbps
- Network protocols decrease net bandwidth



STORAGE

- NFS
  - Server: 10 Gbps ethernet link
  - Nodes: 1 Gbps ethernet link



Dare to be first.
UNIVERSITY OF DELAWARE

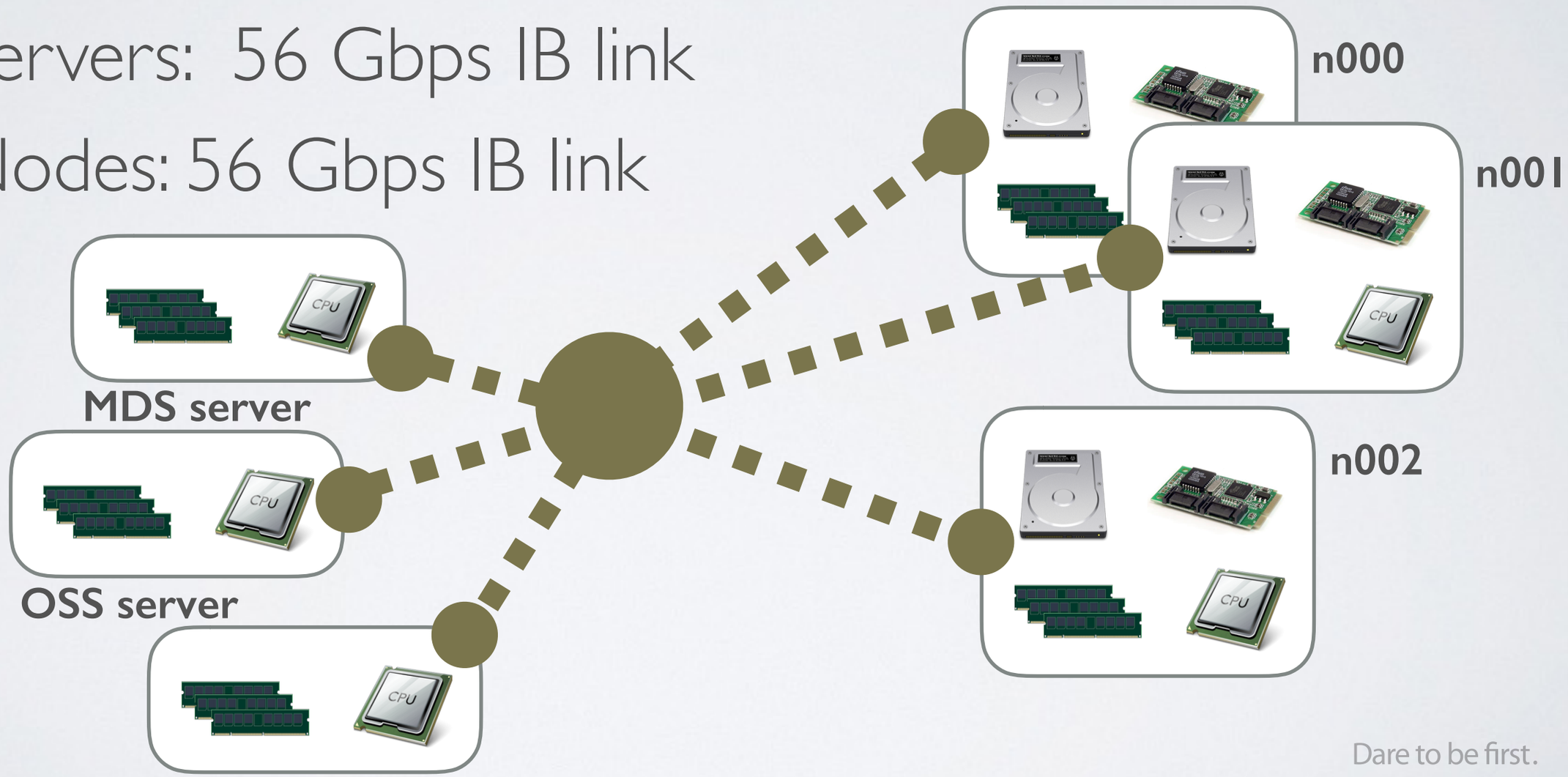
DEMO

- Use Linux "dd" command to write 671 MB to NFS home directory:

```
[frey@n000 ~]$ dd if=/dev/zero of=bigzero  bs=4k count=163840
163840+0 records in
163840+0 records out
671088640 bytes (671 MB) copied, 5.92785 s, 113 MB/s
[frey@n000 ~]$ dd if=/dev/zero of=bigzero  bs=4k count=163840
163840+0 records in
163840+0 records out
671088640 bytes (671 MB) copied, 5.9354 s, 113 MB/s
[frey@n000 ~]$ dd if=/dev/zero of=bigzero  bs=4k count=163840
163840+0 records in
163840+0 records out
671088640 bytes (671 MB) copied, 5.94937 s, 113 MB/s
```

STORAGE

- Lustre
  - Servers: 56 Gbps IB link
  - Nodes: 56 Gbps IB link

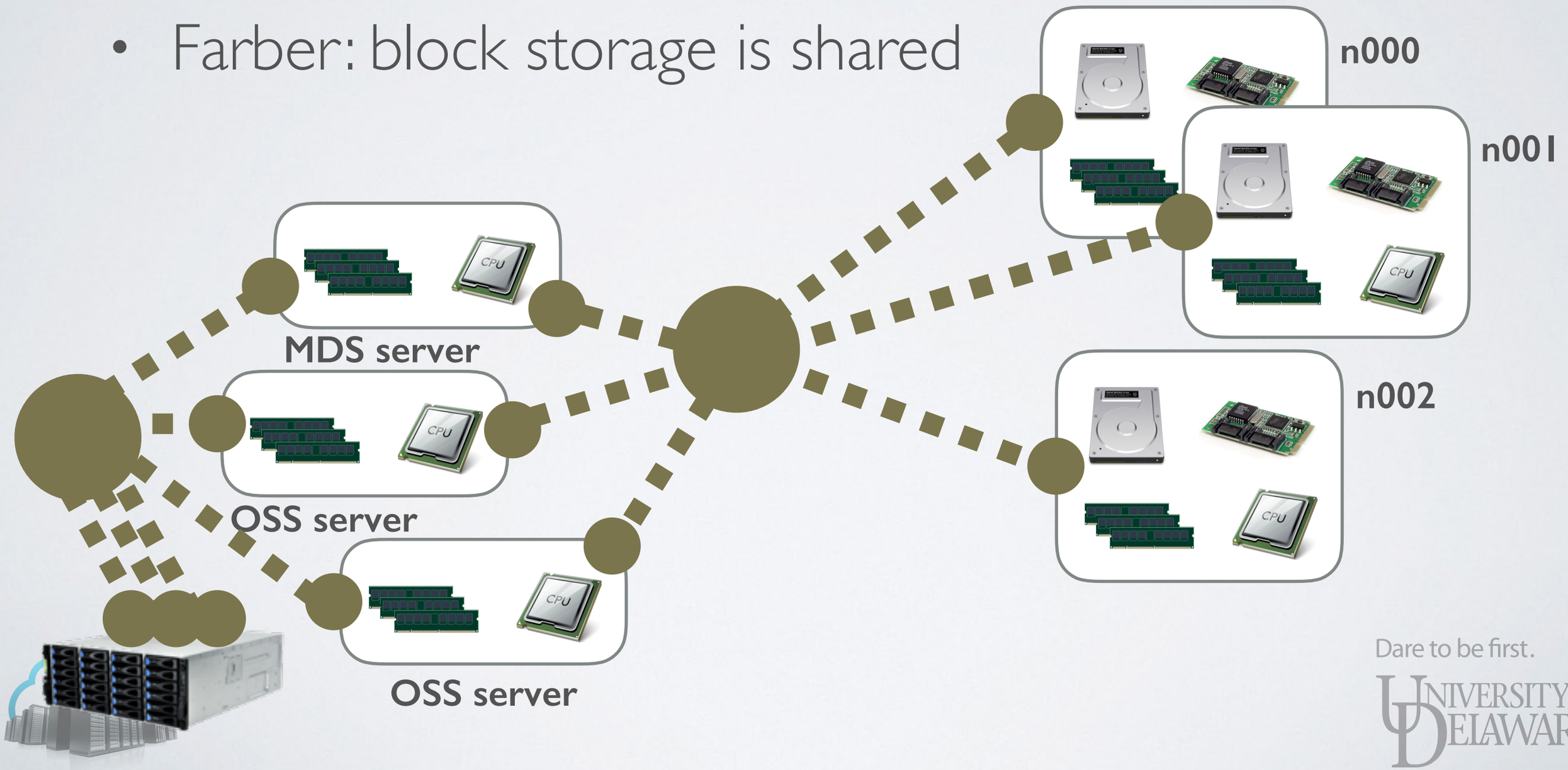


Dare to be first.

UNIVERSITY OF DELAWARE

STORAGE

- Each OSS is meant to augment bandwidth, but…
  - Farber: block storage is shared



Dare to be first.

UNIVERSITY OF DELAWARE

DEMO

- Use Linux "dd" command to write 671 MB to Lustre scratch directory:

```
[frey@n000 ~]$ dd if=/dev/zero of=/lustre/scratch/frey/bigzero bs=4k count=163840
163840+0 records in
163840+0 records out
671088640 bytes (671 MB) copied, 3.83286 s, 175 MB/s
[frey@n000 ~]$ dd if=/dev/zero of=/lustre/scratch/frey/bigzero bs=4k count=163840
163840+0 records in
163840+0 records out
671088640 bytes (671 MB) copied, 4.42253 s, 152 MB/s
[frey@n000 ~]$ dd if=/dev/zero of=/lustre/scratch/frey/bigzero bs=4k count=1638400
1638400+0 records in
1638400+0 records out
6710886400 bytes (6.7 GB) copied, 47.1157 s, 142 MB/s
```

DEMO

- Use Linux "dd" command to write four 671 MB files to L

total data / total time = 4 (6.7 GB) / 89 s
= 286 MB/s

```
[frey@n000 frey]$ for idx in 0 1 2 3; do dd if=/dev/zero of=bigzero-${idx} bs=4k
count=1638400 & ; done
[frey@n000 frey]$
1638400+0 records in
1638400+0 records out
6710886400 bytes (6.7 GB) copied, 59.111 s, 114 MB/s
1638400+0 records in
1638400+0 records out
6710886400 bytes (6.7 GB) copied, 63.9794 s, 105 MB/s
1638400+0 records in
1638400+0 records out
6710886400 bytes (6.7 GB) copied, 87.698 s, 76.5 MB/s
1638400+0 records in
1638400+0 records out
6710886400 bytes (6.7 GB) copied, 89.4624 s, 75.0 MB/s
```

DEMO

- By comparison, same four-file test on local disk

total data / total time = 4 (6.7 GB) / 227.06 s
= 112 MB/s

```
[frey@n000 frey]$ for idx in 0 1 2 3; do dd if=/dev/zero of=bigzero-${idx} bs=4k
count=1638400 & ; done
[frey@n000 frey]$
1638400+0 records in
1638400+0 records out
6710886400 bytes (6.7 GB) copied, 110.561 s, 60.7 MB/s
1638400+0 records in
1638400+0 records out
6710886400 bytes (6.7 GB) copied, 205.875 s, 32.6 MB/s
1638400+0 records in
1638400+0 records out
6710886400 bytes (6.7 GB) copied, 220.848 s, 30.4 MB/s
1638400+0 records in
1638400+0 records out
6710886400 bytes (6.7 GB) copied, 227.06 s, 29.6 MB/s
```

STORAGE SUMMARY

- Local disk (/tmp)
  - Bandwidth not shared with other nodes
    - Predictable performance
  - Grid Engine jobs: $TMPDIR references a directory on the local disk

STORAGE SUMMARY

- Local disk (/tmp)
- NFS (/home, /home/work)
  - Higher-performance disk on server
  - 10 Gbps bandwidth cap on network interface
  - User home directory, $HOME
    - 20 GB quota
  - Workgroup storage, $WORKDIR
    - Quota depends on node quantity purchased

DEMO

```
[user@farber ~]$ cd $HOME
[user@farber ~]$ df -k .
Filesystem           1K-blocks    Used Available Use% Mounted on
storage-nfs1:/export/home/1032
                      20971520 1243136  19728384   6% /home/1032

[(group:user)@farber ~]$ cd $WORKDIR
[(group:user)@farber group]$ df -k .
Filesystem           1K-blocks       Used Available Use% Mounted on
storage-nfs1:/export/work/group
                     1879048192 1442929664 436118528  77% /home/work/group

[(group:user)@farber group]$ df -k $WORKDIR
Filesystem           1K-blocks       Used Available Use% Mounted on
storage-nfs1:/export/work/group
                     1879048192 1442929664 436118528  77% /home/work/group
```

Dare to be first.

UNIVERSITY OF DELAWARE

STORAGE SUMMARY

- Local disk (/tmp)
- NFS (/home, /home/work)
- Server uses ZFS filesystems
- Snapshots = point-in-time copy of the directory
- Second quota (40 GB) controls how many snapshots users have available

DEMO

```
[frey@farber ~]$ cd ~/.zfs
[frey@farber .zfs]$ ls -l
total 4
dr-xr-xr-x 2 root root 3 Jul 18  2014 shares
dr-xr-xr-x 2 root root 2 Jan 20 12:15 snapshot
[frey@farber .zfs]$ cd snapshot
[frey@farber snapshot]$ ls -l
total 3183
drwx------ 36 frey everyone 57 Nov 22  2014 20141128-2215
drwx------ 37 frey everyone 61 Dec 23  2014 20141226-2215
drwx------ 38 frey everyone 64 Jan 22  2015 20150123-2215
drwx------ 38 frey everyone 66 Feb 19  2015 20150220-2215
   :
drwx------ 47 frey everyone 78 Jan 19 14:05 20160119-2215
drwx------ 47 frey everyone 78 Jan 19 14:05 20160120-0615
drwx------ 48 frey everyone 79 Jan 20 11:19 20160120-1215

[frey@farber snapshot]$
```

DEMO

```
[frey@farber snapshot]$ cd 20141226-2215
[frey@farber 20141226-2215]$ ls -l
total 50
drwxr-xr-x  2 frey everyone    10 Dec  9  2014 bin
-rwxr-xr-x  1 frey everyone  6788 Dec 18  2014 chew_up_memory
-rw-r--r--  1 frey everyone   416 Dec 18  2014 chew_up_memory.c
-r--------  1 frey everyone   148 Jul 23  2014 default.pvctxt
drwxr-xr-x  3 frey everyone     7 Sep 15  2014 Developer
drwxr-xr-x  3 frey everyone     4 Dec 10  2014 gaussian_tests
drwxr-xr-x  3 frey everyone     3 Dec  5  2014 intel
drwxr-xr-x 12 frey everyone    14 Nov  5  2014 samples
-rwxr-xr-x  1 frey everyone 21659 Dec 18  2014 UGE_OOMWatcher
-rw-r--r--  1 frey everyone 13115 Dec 18  2014 UGE_OOMWatcher.c
[frey@farber 20141226-2215]$ tail -6 UGE_OOMWatcher.c
  } else {
    perror("Failed to initialize memory threshold event");
  }
  return rc;
}

[frey@farber 20141226-2215]$
```

Dare to be first.

UNIVERSITY OF DELAWARE

STORAGE SUMMARY

- Local disk (/tmp)
- NFS (/home, /home/work)
- Lustre (/lustre/scratch)
  - High-performance disk behind servers
  - Multiple 56 Gbps network connections
  - Large capacity (275 TB)
  - After configuring your workgroup, $SCRATCH

STORAGE SUMMARY

- Local disk (/tmp)
- NFS (/home, /home/work)
- Lustre (/lustre/scratch)
  - Data is NOT backed up
  - Weekly removal of "old" data (Sunday night)
    - Directories older than 30 days

Striping maps fixed-size chunks of a file to different Lustre storage targets. Applications that can access that file asynchronously or concurrently can see increased bandwidth. In this example, the first 4 MB of the file reside on OST 9 (OSS 4); the next 4 MB on OST 3 (OSS 2); the fifth 4 MB chunk is again on OST 9 (OSS 4).

Setting stripe parameters on a directory affect all files subsequently created in that directory.

```
[(
[(

[(
[(
4x
lm
lm
lmm_layout_gen:      0
lmm_stripe_offset:   9
     obdidx         objid         objid         group
          9          60397489        0x39997b1                  0
          3          64070609        0x3d1a3d1                  0
          5          63724385        0x3cc5b61                  0
          8          60884727        0x3a106f7                  0
```

DEMO

```
[(it_nss:frey)@farber frey]$ lfs setstripe --count=2 --size=8M 4x4M.db
error on ioctl 0x4008669a for '4x4M.db' (3): stripe already set
error: setstripe: create stripe file '4x4M.db' failed
```

Once a file has been created, it's striping cannot be modified. You must create a second file with the alternate striping properties and copy the first file's content to it.

DEMO

```
[(it_nss:frey)@farber frey]$ lfs setstripe --count=2 --size=8M 4x4M.db
error on ioctl 0x4008669a for '4x4M.db' (3): stripe already set
error: setstripe: create stripe file '4x4M.db' failed

[(it_nss:frey)@farber frey]$ lfs setstripe --count=2 --size=8M 2x8M.db
[(it_nss:frey)@farber frey]$ dd if=4x4M.db of=2x8M.db ibs=4M obs=8M
100+0 records in
50+0 records out
419430400 bytes (419 MB) copied, 0.952191 s, 440 MB/s

[(it_nss:frey)@farber frey]$
```

OWNED QUEUES

- Each workgroup on Farber is assigned a set of "owned" nodes
- Each workgroup is given an "owned" queue that feeds jobs to "owned" nodes
  - Only accepts jobs from workgroup members
  - No limit on wall time

DEMO

```
[(it_nss:frey)@farber ~]$ qconf -shgrp @it_nss
group_name @it_nss
hostlist n000 n001 n002 n003 n004

[(it_nss:frey)@farber ~]$ qhostgrp
HOSTNAME                ARCH         NCPU NSOC NCOR NTHR NLOAD  MEMTOT  MEMUSE  SWAPTO  SWAPUS
----------------------------------------------------------------------------------------------
global                  -               -    -    -    -     -       -       -       -       -
n000                    lx-amd64       20    2   20   20  0.01   63.0G    8.6G    2.0G     0.0
n001                    lx-amd64       20    2   20   20  0.00   63.0G    1.3G    2.0G     0.0
n002                    lx-amd64       20    2   20   20  0.00   63.0G    4.8G    2.0G     0.0
n003                    lx-amd64       20    2   20   20  0.00   63.0G   14.4G    2.0G     0.0
n004                    lx-amd64       20    2   20   20  0.00   63.0G    1.2G    2.0G     0.0

[(it_nss:frey)@farber ~]$ qhostgrp -g kirby
HOSTNAME                ARCH         NCPU NSOC NCOR NTHR NLOAD  MEMTOT  MEMUSE  SWAPTO  SWAPUS
----------------------------------------------------------------------------------------------
global                  -               -    -    -    -     -       -       -       -       -
n078                    lx-amd64       20    2   20   20  0.00   63.0G   12.9G    2.0G     0.0
n079                    lx-amd64       20    2   20   20  0.00   63.0G    9.6G    2.0G     0.0
n106                    lx-amd64       20    2   20   20  0.00  126.1G    8.5G    2.0G     0.0
n108                    lx-amd64       20    2   20   20  0.00  126.1G   26.7G    2.0G     0.0
n109                    lx-amd64       20    2   20   20  0.00  126.1G   11.7G    2.0G     0.0
n110                    lx-amd64       20    2   20   20  0.00  126.1G   25.3G    2.0G     0.0
```

DEMO

```
[(it_nss:frey)@farber ~]$ qconf -shgrpl
@128G
@256G
@afwallace
@allhosts
@arce
@archer
@base
@cadsr_cluster
@cbcbcore
@ccei_biomass
@ccm_gillespi
@clouds_wind_climate
@dditoro
@disasters
@ececis_research
@geography
@gpu
@hmichael
@ifsa
@it_css
@it_nss
@jayaraman_lab
@jneun
@kirby
@kukulka_lab
   :
```

DEMO

```
[(it_nss:frey)@farber ~]$ qhostgrp -g ccei_biomass -j
HOSTNAME                ARCH         NCPU NSOC NCOR NTHR NLOAD  MEMTOT  MEMUSE  SWAPTO  SWAPUS
----------------------------------------------------------------------------------------------
global                  -               -    -    -    -     -       -       -       -       -
n005                    lx-amd64       20    2   20   20  0.99   63.0G    3.6G    2.0G     0.0
   job-ID     prior   name       user         state submit/start at     queue      master ja-task-ID
   ------------------------------------------------------------------------------------------------
       401819 0.50340 aa-aa8.qs  emiliord     r     01/20/2016 11:53:40 ccei_bioma MASTER
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
                                                                       ccei_bioma SLAVE
```

Just like the *qhost* command, adding –j to *qhostgrp* displays jobs running on each node listed.

STANDBY QUEUES

- One design tenet for a community cluster:
  - Users should have opportunistic access to a greater portion of cluster resources than they own
- With the major percentage of nodes sold, this means using nodes owned by another workgroup
  - Wall time **must** be limited so owners can regain access to their nodes

STANDBY QUEUES

- Standby queue limits (per-user):
  - 8-hour runtime limit
    - 200 cores / job
  - 4-hour runtime limit
    - 800 cores / job
  - 800 cores total (all standby jobs)
- Standby jobs always receive lower scheduling priority versus "owned" jobs

STANDBY QUEUES

- The standby queues on Farber extend to all compute nodes…
  - …except nodes with coprocessors (GPU, Intel Phi)

- Submit a job to run in the 8-hour standby queue:

```
[(it_nss:frey)@farber ~]$ qsub -pe mpi 200 -l standby=1 my_job.qs
```

- Submit a job to run in the 4-hour standby queue:

```
[(it_nss:frey)@farber ~]$ qsub -pe mpi 600 -l standby=1 -l h_rt=4:00:00 my_job.qs
```

STANDBY QUEUES

- Standby queues are or
  larger-than-owned por
  - Pros:
    - Cooperate well with
    - Limited wall time prev
    - Great way for users to perform "scale-up" tests

Run a job with varying core counts for a limited time. Compare how far the job progresses in 8 (or 4) hours versus the core count used. Resulting information informs applications for CPU time on national/government supercomputers.

STANDBY QUEUES

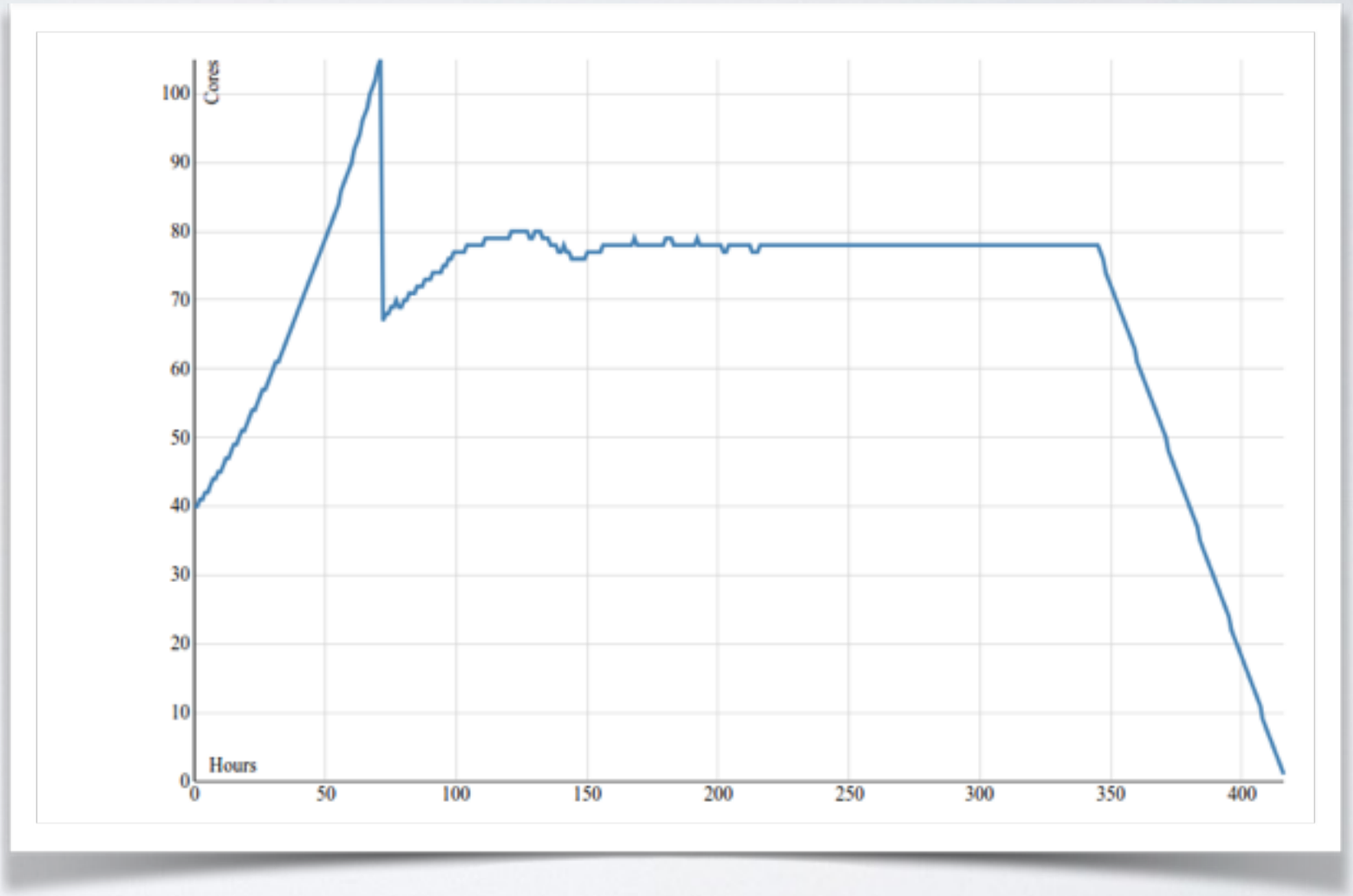
- Standby queues are one way to grant access to a larger-than-owned portion of a cluster
  - Cons:
    - Per-user core-count limit can leave jobs queued while adequate resources are idle
    - Walltime limits may be too restrictive for some users

STANDBY QUEUES

- Standby queues are one way to grant access to a larger-than-owned portion of a cluster
- *Fair share* scheduling (A.K.A. *sharetree* in Grid Engine) is another way
  - Groups are guaranteed a percentage of resources (their *share*) ***over time***
    - Immediate use may exceed *share*, but future jobs will have lower priority to meet others' *shares*

STANDBY QUEUES

- Group depicted has 79 cores as its share
  - At 70 hours, job priority is lowered due to other group's jobs queued and waiting
- Priority recovers by 120 hours



Dare to be first.

UNIVERSITY OF DELAWARE

NODE SELECTION

- One reason for job scheduler is to NOT require users to match jobs to specific nodes
- Sometimes the user *does* want to choose specific nodes
  - Coprocessor (GPU, Phi)
  - Extended RAM (128 GB, 256 GB)
  - Processor version

COPROCESSORS

- Grid Engine is configured to know what nodes have a coprocessor; how many are installed
- Grid Engine keeps track of which coprocessor(s) have been assigned to jobs, which are free
- GPU: use the complex *nvidia_gpu*
- Phi: use the complex *intel_phi*

COPROCESSORS

- Request two full nodes, one GPU in each:

```
[(it_nss:frey)@farber ~]$
```

- Request ha … able:

```
[(it_nss:frey)@farber ~]$
```

- Request a …

```
[(it_nss:frey)@farber ~]$ qsub -q '*@g*' …
```

@g
[any queue] on [host name starts with g]

*@@gpu
[any queue] on [host in hostgroup @gpu]

@
[any queue] on [any host]

Dare to be first.
UNIVERSITY OF DELAWARE

COPROCESSORS

- Check GPU availability on hosts

```
[(it_nss:frey)@farber ~]$ qhostgrp -g gpu -F  nvidia_gpu
HOSTNAME                ARCH         NCPU NSOC NCOR NTHR NLOAD  MEMTOT  MEMUSE  SWAPTO  SWAPUS
----------------------------------------------------------------------------------------------
global                  -               -    -    -    -     -       -       -       -       -
g100                    lx-amd64       20    2   20   20  1.02   63.0G    1.9G    2.0G     0.0
    Host Resource(s):      hc:nvidia_gpu=1.000000
g101                    lx-amd64       20    2   20   20  0.82   63.0G    1.5G    2.0G     0.0
    Host Resource(s):      hc:nvidia_gpu=1.000000
g102                    lx-amd64       20    2   20   20  0.08   63.0G    1.3G    2.0G     0.0
    Host Resource(s):      hc:nvidia_gpu=1.000000
g103                    lx-amd64       20    2   20   20  0.01   63.0G    1.3G    2.0G     0.0
    Host Resource(s):      hc:nvidia_gpu=1.000000
g104                    lx-amd64       20    2   20   20  0.05   63.0G    1.3G    2.0G     0.0
    Host Resource(s):      hc:nvidia_gpu=1.000000
g105                    lx-amd64       20    2   20   20  0.86   63.0G    1.5G    2.0G     0.0
    Host Resource(s):      hc:nvidia_gpu=1.000000
g114                    lx-amd64       20    2   20   20  0.25  252.2G   53.4G    2.0G    4.3M
    Host Resource(s):      hc:nvidia_gpu=1.000000
g115                    lx-amd64       20    2   20   20  0.22  252.2G   58.8G    2.0G     0.0
    Host Resource(s):      hc:nvidia_gpu=1.000000
g189                    lx-amd64       20    2   20   20  0.00  125.9G    1.7G    2.0G     0.0
    Host Resource(s):      hc:nvidia_gpu=2.000000
test-gpu                lx-amd64       20    2   20   20  0.50   63.0G    5.5G    2.0G    2.1M
    Host Resource(s):      hc:nvidia_gpu=0.000000
test-gpu2               lx-amd64       24    2   24   24  0.00  252.2G    2.4G    2.0G     0.0
    Host Resource(s):      hc:nvidia_gpu=1.000000
```

EXTENDED RAM

- Requesting e.g. 4 GB per slot isn't enough to force a job to run on a node with 128 GB of RAM

```
[(it_nss:frey)@farber ~]$ qsub -l m_mem_free=4G -pe mpi 20
```

- This job *could* be run using 15 slots on one node and 5 on another

EXTENDED RAM

- Requesting e.g. 4 GB per slot isn't enough to force a job to run on a node with 128 GB of RAM
- Specifically request that the node have AT LEAST a certain amount of RAM present

```
[(it_nss:frey)@farber ~]$ qsub -l m_mem_free=4G,m_mem_total=80G -pe mpi 20
```

EXTENDED RAM

- Requesting e.g. 4 GB per slot isn't enough to force a job to run on a node with 128 GB of RAM
- Specifically request that the node have AT LEAST a certain amount of RAM present

```
[(it_nss:frey)@farber ~]$ qsub -l m_mem_free=4G,m_mem_total=80G -pe mpi 20
```

- Request the node come from one of the large-memory hostgroups

```
[(it_nss:frey)@farber ~]$ qsub -q '*@@128G' –l m_mem_free=4G -pe mpi 20
```

EXTENDED RAM

- Check memory complexes for a node

```
[(it_nss:frey)@farber ~]$ qhostgrp -g 128G -F
HOSTNAME                ARCH         NCPU NSOC NCOR NTHR NLOAD  MEMTOT  MEMUSE  SWAPTO  SWAPUS
----------------------------------------------------------------------------------------------
   :
n068                    lx-amd64       20    2   20   20  0.00  126.1G   19.7G    2.0G     0.0
   :
   hl:m_cache_l1=32.000K
   hl:m_cache_l2=256.000K
   hl:m_cache_l3=25.000M
   hl:m_mem_total=127.955G
   hl:m_mem_used=19.742G
   hl:m_mem_free=108.213G
   hl:m_numa_nodes=2.000000
   hl:m_mem_total_n0=63.956G
   hl:m_mem_free_n0=28.715G
   hl:m_mem_used_n0=35.240G
   hl:m_mem_total_n1=64.000G
   hl:m_mem_free_n1=41.730G
   hl:m_mem_used_n1=22.270G
    :
```

PROCESSOR VERSION

- Farber contains nodes with both v2 and v3 Ivy Bridge processors
  - The v3 processor includes additional vector extensions to the ISA
  - Some software may require v3 ISA extensions, so users will want to confine jobs to nodes with v3 processors

PROCESSOR VERSION

- The *cpu_ivy_bridge_revision* complex can be used to select nodes with v3 processors

```
[(it_nss:frey)@farber ~]$ qsub -l cpu_ivy_bridge_revision=3 …
```