



# Introduction to the New Shared Memory System Endeavour

Feb. 27, 2013

NASA Advanced Supercomputing Division

# Endeavour: New Shared Memory System



- Endeavour1-2 will replace Columbia21-24 to support applications with
  - large shared memory need
  - high OpenMP thread counts

Use Pleiades for all other applications

	Endeavour		Columbia			
Host Name	Endeavour1	Endeavour2	C21	C22	C23	C24
Architecture	SGI Ultra Violet (UV) 2000		SGI Altix 4700			
Processor	Xeon Sandy Bridge		Itanium Montecito/Montvale			
# of sockets	64	128	256	1024	512	512
# of cores	512	1024	512	2048	1024	1024
Memory (TB)	2	4	1	4	2	2
Memory/Core	4 GB		2 GB			
Clock Speed	2.6 GHz		1.6 – 1.67 GHz			
Interconnect	NUMALink 6		NUMALink 4			
Peak Performance	32 Tflops/s		30 Tflops/s			

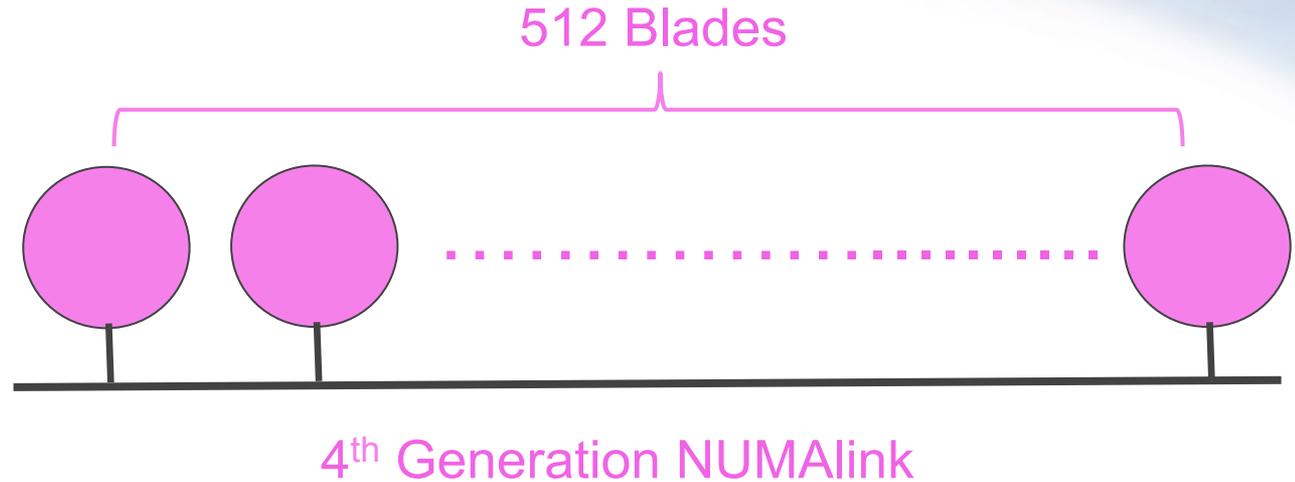
# Cache-Coherent Global Shared Memory



Columbia22  
2048 cores  
4 TB



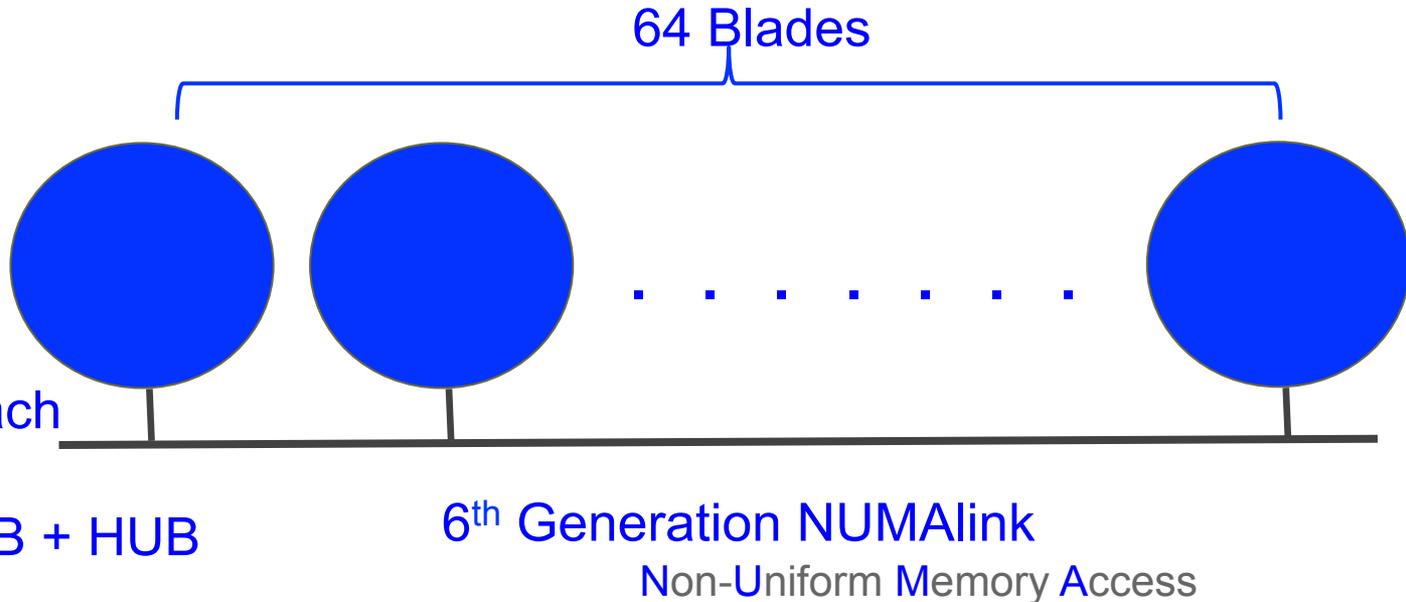
1 blade with  
4 cores/8 GB  
+ SHUB



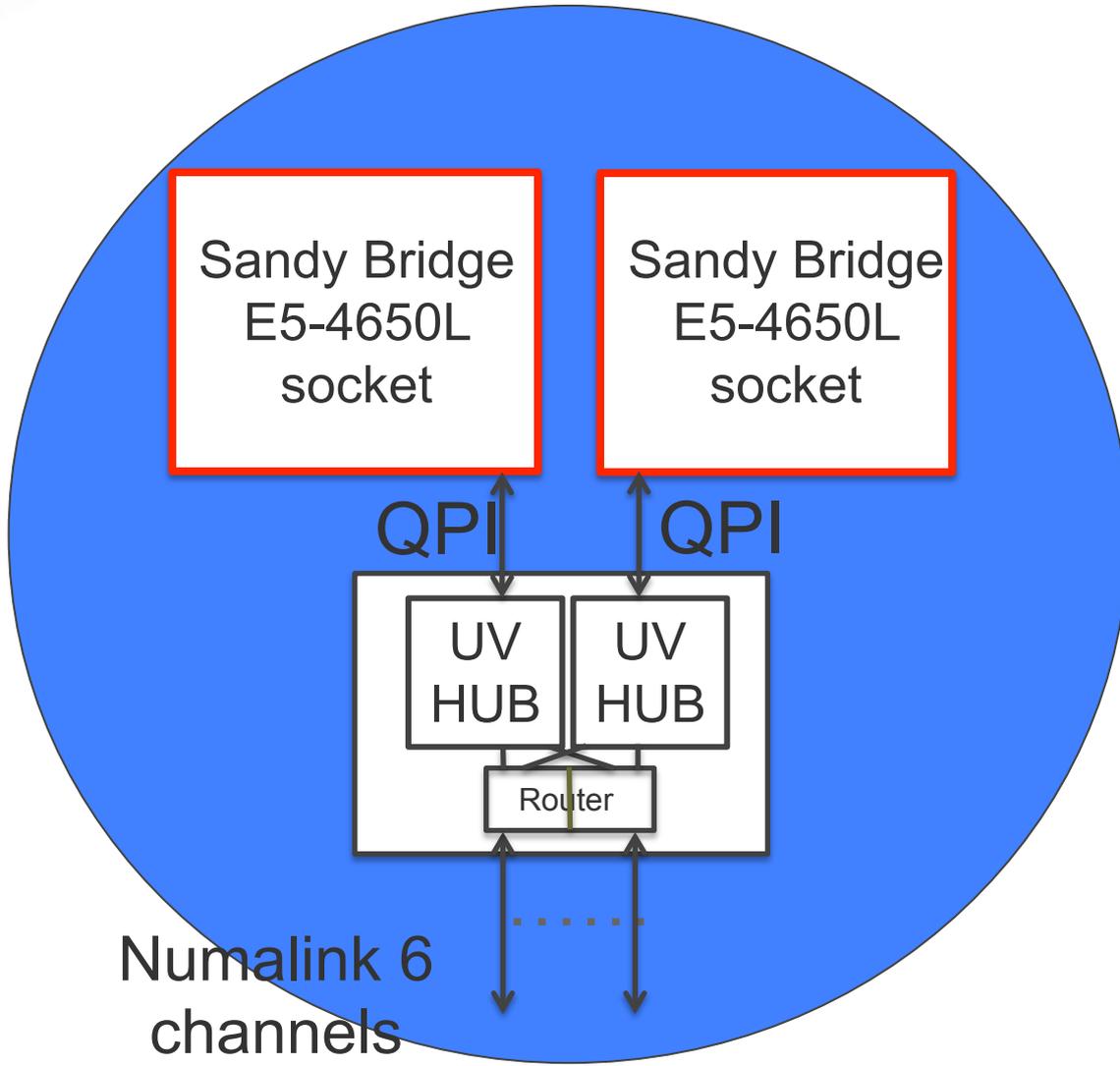
Endeavour2  
1024 cores  
4 TB



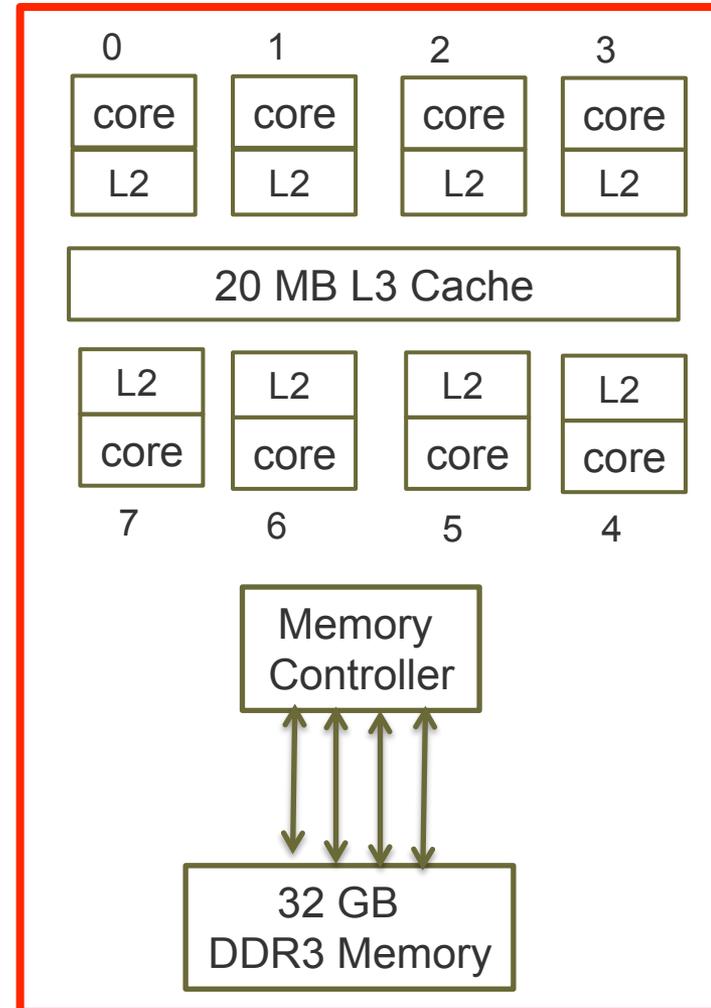
1 blade with  
2 sockets. Each  
socket has  
8 cores/32 GB + HUB



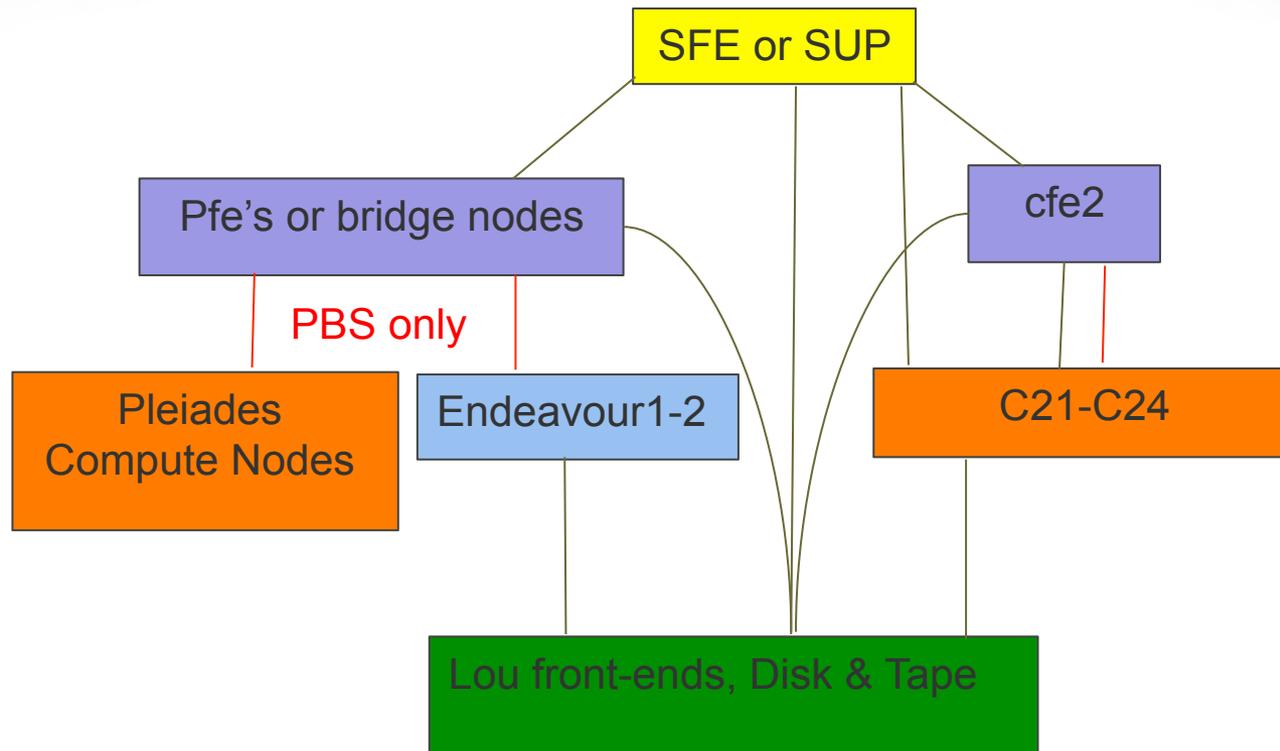
# An Endeavour Blade Has 2 San Sockets like the Pleiades Sandy Bridge Blade



### Sandy Bridge E5-4650L

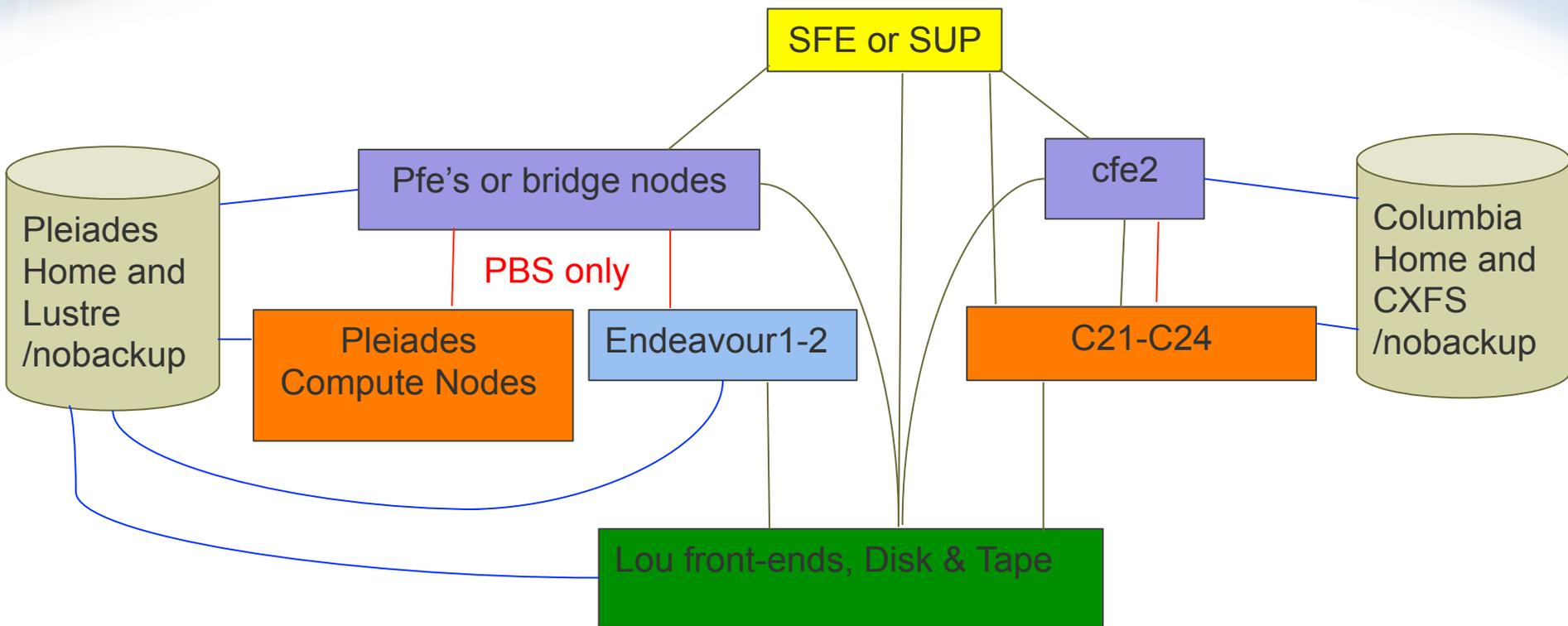


# Endeavour Network Environment



- Unlike Columbia, Endeavour does not have its own front-end
- Endeavour uses Pleiades front-ends (pfe20-27 and bridge1-4)
- Like the Pleiades compute nodes, interactive login (from anywhere) will be killed unless you have a PBS job running on Endeavour
- You need a Pleiades account and an Endeavour allocation (GID in PBS ACL)
- Cannot run a PBS job across Pleiades compute nodes and Endeavour

# Endeavour Network Environment (cont'd)



- Pleiades home and Lustre filesystems are mounted on Endeavour
- Columbia home and CXFS filesystems are NOT mounted
- Any Columbia data you need for Endeavour have to be copied over to the Pleiades home or Lustre filesystems

# Moving Columbia Data to Pleiades

- Default Pleiades quota: \$HOME 8GB/10GB, /nobackup 500GB/1TB
- **Default stripe count on Pleiades /nobackup is 1.** You may want to create subdirectory with a larger stripe count if your files are large

```
pfe% cd /nobackup/username
```

```
pfe% mkdir bigstripe_dir
```

```
pfe% lfs setstripe -c 32 bigstripe_dir
```

- Copy Columbia data to Pleiades using **shiftc**, bbftp or scp

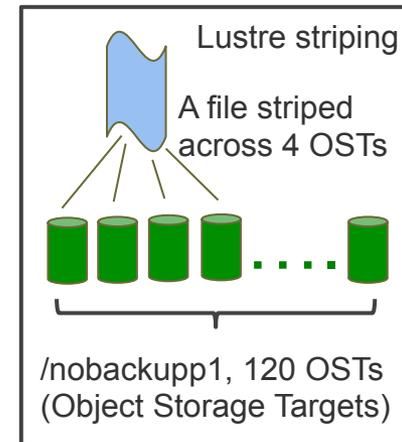
```
pfe% shiftc -rp cfe2:/nobackup2a/username/dir1 /nobackup/username/bigstripe_dir
```

## References

[http://www.nas.nasa.gov/hecc/support/kb/Lustre-Basics\\_224.html](http://www.nas.nasa.gov/hecc/support/kb/Lustre-Basics_224.html)

[http://www.nas.nasa.gov/hecc/support/kb/Lustre-Best-Practices\\_226.html](http://www.nas.nasa.gov/hecc/support/kb/Lustre-Best-Practices_226.html)

[http://www.nas.nasa.gov/hecc/support/kb/Use-Shift-for-Reliable-Local-and-Remote-File-Transfers\\_300.html](http://www.nas.nasa.gov/hecc/support/kb/Use-Shift-for-Reliable-Local-and-Remote-File-Transfers_300.html)



# Endeavour Software Environment



- Linux Based Operating System

Pleiades: SLES11SP1 (Linux 2.6)    Endeavour: SLES11SP2 (Linux 3.0)

- Public software modules under Pleiades /nasa

Most modules built under SLES11SP1 are expected to work under SLES11SP2

Totalview/8.9.2-1 (limit to 256 processes)

- Unsupported software under /u/scicon/tools/bin

Some should work for both Pleiades and Endeavour – *mbind.x*, *gm.x*

Some are meant for Pleiades only – *node\_stat.sh*, *qs*, *qtop.pl*

# Compiling and Running Your Code



- Columbia (IA-64) executable will not run on Endeavour (x86-64)
- Unlike Columbia, there are no compiler, math or MPI libraries modules loaded by default. Recommend

```
pfe% module load comp-intel/2012.0.032
```

```
pfe% module load mpi-sgi/mpt.2.06rp16 (this may change)
```

- SGI's SCSL is not available on Endeavour. Use Intel MKL instead. No need to load explicit MKL module if you load v11.1 or later compiler.

```
pfe% ifort -O2 program.f -mkl
```

[http://www.nas.nasa.gov/hecc/support/kb/MKL\\_90.html](http://www.nas.nasa.gov/hecc/support/kb/MKL_90.html)

- Pin OpenMP threads to avoid multiple threads running on the same core and thread migration

[http://www.nas.nasa.gov/hecc/support/kb/ProcessThread-Pinning-Overview\\_259.html](http://www.nas.nasa.gov/hecc/support/kb/ProcessThread-Pinning-Overview_259.html)

- Use *mpiexec* instead of *mpirun* for MPI applications (for NAS systems)

# Definition of an MAU for PBS Jobs



MAU = Minimum Allocatable Unit

- Columbia : **1 blade** with 4 cores and ~7.6 GB
- Endeavour1-2: **1 socket** with 8 cores and ~30 GB
- Pleiades Sandy Bridge : **1 blade** (node) with 16 cores and ~30 GB

Small amount of memory (<2GB) in each Endeavour socket is reserved for system usage. Also, 1 socket in Endeavour1 and 2 sockets in Endeavour2 are used by the boot cpusets.

# PBS Server and Queues



- Total PBS resources

Endeavour1: 63 MAUs = 504 cores / 1,890GB

Endeavour2: 126 MAUs = 1,008 cores / 3,780GB

- PBS server is pbspl3 (pbspl1 is for Pleiades)
- Queues for general use (still needs gid listed in PBS ACL)

*pfe20% qstat -Q @pbspl3*

Queue	Max Ncpus/def	MaxTime/def	pr
e_normal	--/ 8	08:00/01:00	0
e_long	--/ 8	72:00/36:00	0
e_vlong	--/ 8	600:00/24:00	0
e_debug	128/ 8	02:00/00:30	15

**-q e\_debug** is required if you want to use the e\_debug queue.

For other queues, -q is optional. PBS will route your job to the proper queue based on your walltime request.

# PBS Resource Request Examples



*#PBS -lncpus=48*

This gives you 6 MAUs (48cores and 180GB) on Endeavour1 or 2

*#PBS -lmem=128GB*

This gives you 5 MAUs (40 cores and 150GB) on Endeavour 1 or 2

*#PBS -lncpus=1,mem=128GB*

or

*#PBS -lselect=ncpus=1:mem=128GB*

This gives you 5 MAUs (40 cores and 150 GB) on Endeavour 1 or 2

*#PBS -lselect=host=endeavour1:ncpus=48:mem=128GB*

This gives you 6 MAUs (48cores and 180GB) on Endeavour1

*#PBS -W group\_list=your\_endeavour\_gid*

**Note:** Like on Columbia, if your job starts to use more memory than allocated, it will be killed by a policykill daemon and you will receive an email.

# PBS Commands Examples



Commands can be issued on pfe's or bridge nodes

For Endeavour jobs, `qsub` requires at least either `queue_name` or `@pbspl3`; otherwise the job is sent to run on Pleiades compute nodes instead

```
pfe20% qsub -q queue_name @pbspl3 job_script
```

```
pfe20% qstat -nu username @pbspl3
```

```
pfe20% qstat 1234.pbspl3
```

```
pfe20% qdel 1234.pbspl3
```

If your workload is primary on Endeavour, do

```
pfe20% setenv PBS_DEFAULT pbspl3
```

```
pfe20% qsub job_script [-q queue_name]
```

```
pfe20% qstat -nu username
```

```
pfe20% qstat 1234
```

```
pfe20% qdel 1234
```

# SBU Rate and Job Accounting



System	Pleiades Harpertown	Pleiades Nehalem	Pleiades Westmere	Pleiades Sandy Bridge	Endeavour	Columbia
# of cores per MAU	8	8	12	16	8	4
SBU rate	0.45	0.80	1.00	1.82	0.74*	0.18

\*Provisional rate based on best performance obtained for the SBU benchmark codes: ENZO, WRF, GEOS-5, Overflow, USM3D, FUN3D

- In general, expect better performance than Columbia but worse performance than the Pleiades Sandy Bridge nodes
- To check allocation status:

```
pfe20% acct_ytd -c endeavour your_gid
```

Project	Host/Group	Fiscal Year	Used	%	Limit	Remain	Linear YTD Usage	Project Exp Date
Your_gid	endeavour	2012	0.000	0.00	10000.000	10000.000	N/A	04/30/13

- To check SBU usage by job, completed on endeavour2, for a specific date:

```
pfe20% acct_query -c endeavour2 -d 02/22/13 -u your_username -olow
```

- Training videos provided by SGI (John Baron)

## *Optimizing Parallel Performance on SGI UV*

- ✓ memory architecture of UV 2000
- ✓ data locality by parallel data initialization
- ✓ pinning for pure OpenMP or MPI/OpenMP hybrid codes

6 videos, each less than 10 min

[http://www.sgi.com/partners/developers/training\\_videos.html#opp](http://www.sgi.com/partners/developers/training_videos.html#opp)

- Endeavour is still being tuned

- ✓ May experience runtime variability (topology, IO, buffer cache, MPT version/settings )
- ✓ Topology-aware PBS scheduling to be tested
- ✓ IO fabric to be improved
- ✓ Optimum MPT version and environment variables to be settled

Some of these experiments may require dedicated time

# Can I Run on Endeavour?



- Endeavour should be used to run jobs that need large shared memory or large thread counts which cannot be accommodated by the Pleiades resources
  - ✓ OpenMP code with  $> 16$  threads
  - ✓ serial or OpenMP code needs  $> 90$  GB
  - ✓ MPI code whose rank 0 needs  $> 90$  GB and/or more ranks need  $> 30$  GB per rank
- Columbia users: decide between Pleiades or Endeavour
  - ✓ send email to [support@nas.nasa.gov](mailto:support@nas.nasa.gov)
    - system to transition to, Columbia GID, username, SBUs to move, MPI code ?
- Pleiades users: Individual request to run on Endeavour will be reviewed



# User Documentation

<http://www.nas.nasa.gov/hecc/support/kb/entry/410>  
Endeavour Configuration Details

<http://www.nas.nasa.gov/hecc/support/kb/entry/411>  
Endeavour Quickstart Guide



# Questions?

A PDF and recording of this webinar will be available within 48 hours at:

<http://www.nas.nasa.gov/hecc/support/training.html>

Next Webinar

**“Advanced Features of the SHIFT Transfer Tool”**

tentatively scheduled for **Wednesday** March 27, 2013 at 11am

Suggestions for future webinar topics are welcome