



NAS HECC Hardware Resources Overview

April 2, 2014

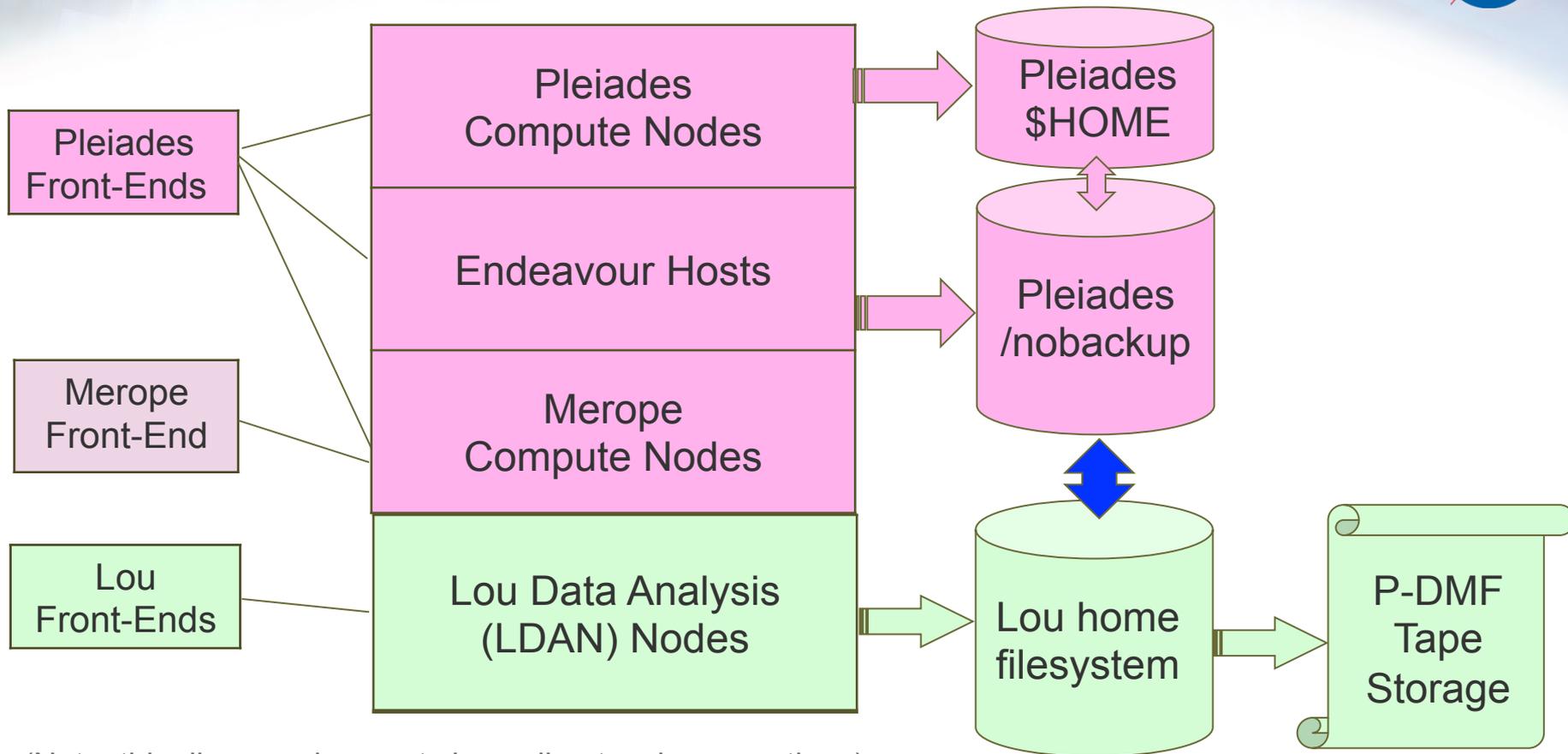
NASA Advanced Supercomputing Division

Prologue



- Why this webinar?
 - Many changes in the NAS HECC resources for the past year or so:
 - with simplified work flow, and
 - increased capacity and capability in computing power and data storage
 - Web pages published, emails sent, but you may have missed some
- Outline: High level overview on current status
 - Pleiades recent augmentation (and some less known bigmem nodes)
 - Merope cluster (with repurposed Pleiades processors)
 - Endeavour shared memory systems
 - New Lustre and NFS filesystems
 - New Lou cluster for long term storage and data processing (Idan nodes)

NAS HECC Hardware Resources



(Note: this diagram does not show all network connections)

Simplified workflow

- All compute systems use the same home and /nobackup filesystems
- Can use common front-ends (pfe's) for all PBS jobs submission/checking
- On Lou, can do local copy (instead of remote copy) between /nobackup and Lou



Pleiades Front-End Nodes Status

	Pfe[1-12]	Bridge[1-2]	Bridge[3-4]	Pfe[20-27]
Processor	Harpertown	Harpertown	Nehalem X7560	Sandy Bridge
# of physical cores /host	8	8	32	16
GB Memory/ host	16	64	256	64
Network (GigE)	1	1 and 10	1 and 10	1 and 10

- Pfe[20-27] replaced pfe[1-12] in Oct 2012
- For remote file transfers, we used to suggest using bridge nodes (with the faster 10 GigE). Now you can use either pfe[20-27] or bridge nodes
- Use bridge[3-4] (or the Idan nodes) if you need more memory
- Use pfe's or bridge nodes to submit/manage PBS jobs for Pleiades, Endeavour (and Merope).

Pleiades Compute Nodes



Recent Augmentation (Dec/2013 – Mar/2014)

	Nehalem 8 cores, 24GB/node	Westmere 12 cores, 24GB/node	Sandy Bridge 16 cores, 32GB/node	Ivy Bridge 20 cores, 64GB/node
# of Racks	20	74* -> 62*	26	46@ -> 75#
# of Nodes	1,280	4,672 -> 3,904	1,872	3,312 -> 5,400
# of Cores	10,240	56,064 -> 46,848	29,952	66,240 -> 108,000
Peak TFlops	120	658 -> 549	623	1,484 -> 2,419

- Include 2 half-populated Westmere + GPU racks
- @ These 46 Ivy Bridge racks replaced 64 Harpertown racks in Aug, 2013
- # 3 racks currently reserved for IB testing
- Removed 20 Nehalem, 12 Westmere racks; added 29 Ivy Bridge racks
- Total: 163 racks, 11,176 nodes, 184,800 cores, ~502TB memory, 3.59 Pflops/s
- More detailed Pleiades Configuration: Knowledge Based article 77
- *#PBS -l select=xx:ncpus=xx:model={wes,san,ivy}* (default is wes for most queues)
- Ask for fewer Ivy Bridge (20 cores, 64 GB) nodes than Westmere or Sandy Bridge nodes

Pleiades Big Memory Compute Nodes Status



	Westmere	Sandy Bridge	Ivy Bridge
Regular	24 GB	32 GB	64 GB
Big Memory	48 GB (17 nodes) 96 GB (4 nodes)	96 GB (3 nodes) 256 GB (5 nodes)	128 GB (3 nodes) 256 GB (4 nodes)*
#PBS -l select=... .	<code>:model=wes:bigmem=true:mem=90GB</code>	<code>:model=ldan:mem=252GB</code> (Do not add :bigmem=true)	<code>:model=ivy:bigmem=true</code>

* The 256 GB Ivy Bridge bigmem nodes are reserved for a project which procured these nodes.

- To request 1 big memory node + 10 regular nodes

Westmere: `#PBS -l select=1:ncpus=12:bigmem=true:model=wes+10:ncpus=12:model=wes`

Ivy Bridge: `#PBS -l select=1:ncpus=20:bigmem=true:model=ivy+10:ncpus=20:model=ivy`

Sandy Bridge: `#PBS -l select=1:ncpus=16:model=ldan+10:ncpus=16:model=san`

- Normal, long, debug queues can use all (Westmere, Sandy Bridge, Ivy Bridge) big memory nodes
- Devel queue can use only use the 96 GB Westmere and 256 GB Sandy Bridge nodes. No Ivy Bridge bigmem nodes are available for the devel queue.
- Info on How to Get More Memory for Your Job: Knowledge Based article 222

Merope Cluster Status



- Contains repurposed (out of maintenance) processors retired from Pleiades
 - Housed in building 233, a few km away from the main building 258
- In production since Sept 2013
- No local filesystem. Can access Pleiades filesystems.
- Currently with 640 Harpertown nodes and running CentOS instead of SLES11
 - SLES11 executable may work. If not, recompiling may be needed
- Front end: merope-fe1 (currently CentOS)
- Compute nodes will change and use SLES11SP3
 - Harpertown (currently, 640 nodes) -> Westmere + Nehalem (Date: TBD)
- Merope PBS jobs use your Pleiades allocation
- Server pbspl233; no devel queue; debug 2 hrs, normal 8 hr, long 16 hours
- Managing Merope PBS jobs
 - merope-fe1% qsub -q queue_name job_script (\$PBS_DEFAULT is pbspl233)*
 - pfe's% qsub -q queue_name@pbspl233 job_script (\$PBS_DEFAULT is pbspl1)*
- Incentives to use Merope: Much faster turnaround
- Drawback: Slow IO, processors (in 233) are far away from filesystems (in 258)
- Merope info: Knowledge Based articles 448, 449

Endeavour Status



- SGI UV shared memory systems replaced Columbia; in production since Feb 2013
- endeavour1: 512 cores, 2 TB endeavour2: 1024 cores, 4TB
- Has Pleiades \$HOME, /nobackuppX, /nobackupnfs2
- Front ends: same as Pleiades
- What should run on Endeavour:
 - Jobs using more than 252 GB for one process can't run on Pleiades
 - OpenMP Jobs with >40 threads
- Endeavour allocation is needed; can request transfer some Pleiades allocation over
- PBS server: pbspl3
- Managing PBS jobs

#PBS -lncpus=xx,mem=yyGB

or

#PBS -lselect=ncpus=xx:mem=yyGB

Endeavour queues are `e_normal`, `e_long`, `e_vlong`, `e_debug`

pfe's% qsub -q queue_name[@pbspl3] job_script

pbspl3% qsub -q queue_name job_script

- Endeavour info: Knowledge Based articles 410, 411, 425

Pleiades /nobackup Status



New and reconfigured Lustre /nobackuppX

	p7	p8	Old p1	Old p2	p9	p3-p6
# of OSS	16	26	8	8	16	8 each
# of OST	84	312	120	120	240	60 - 120
Size (TB)/OST	21.8	21.8	14.1	14.1	14.1	7.1 to 14.1
Total Size (PB)	1.8	6.7	1.7	1.7	3.4	varies

- Completed migrating almost all users from old p1/p2 to p7/p8 (Feb of 2014)
- p9 created from old p1/p2 (Feb 2014); moving p6 users to p9 (in progress)
- Decisions about p3, p4, p6 yet to be made; p5 will stay as is

- With new p7/p8, total space doubled to ~15 PB (Aug, 2013)
- Default quota: software limit 0.5TB, hard limit 1 TB.
- Checking Lustre quota (dynamic disk allocation, not actual usage)

pfe% lfs quota -u username /nobackup/username (display total disk allocated and limits)

pfe% lfs quota -v -u username /nobackup/username (include per OST info)

- Can provide larger quota if needed. Please clean up unnecessary files. When system is over 80% full, a few top users are notified.



New Lustre /nobackupp7-8 (using Netapp instead of DDN)

- Has data validation capability between Lustre server and storage device
- Better handle of random IO; Better performance than old /nobackupp1-6
Benchmark on p8 with 1200 threads, each doing IO on its own file, files spread among 312 OSTs, getting IO bandwidth 35 – 60 GB/s
- Default stripe count = 1, stripe size = 4MB. To use larger stripe count:
pfe% lfs setstripe -c 32 -s 4m dir1
- Using a larger stripe count for a file decreases the chances to hit quota limit on a single disk
- Lustre Best Practices: Knowledge Base article 226

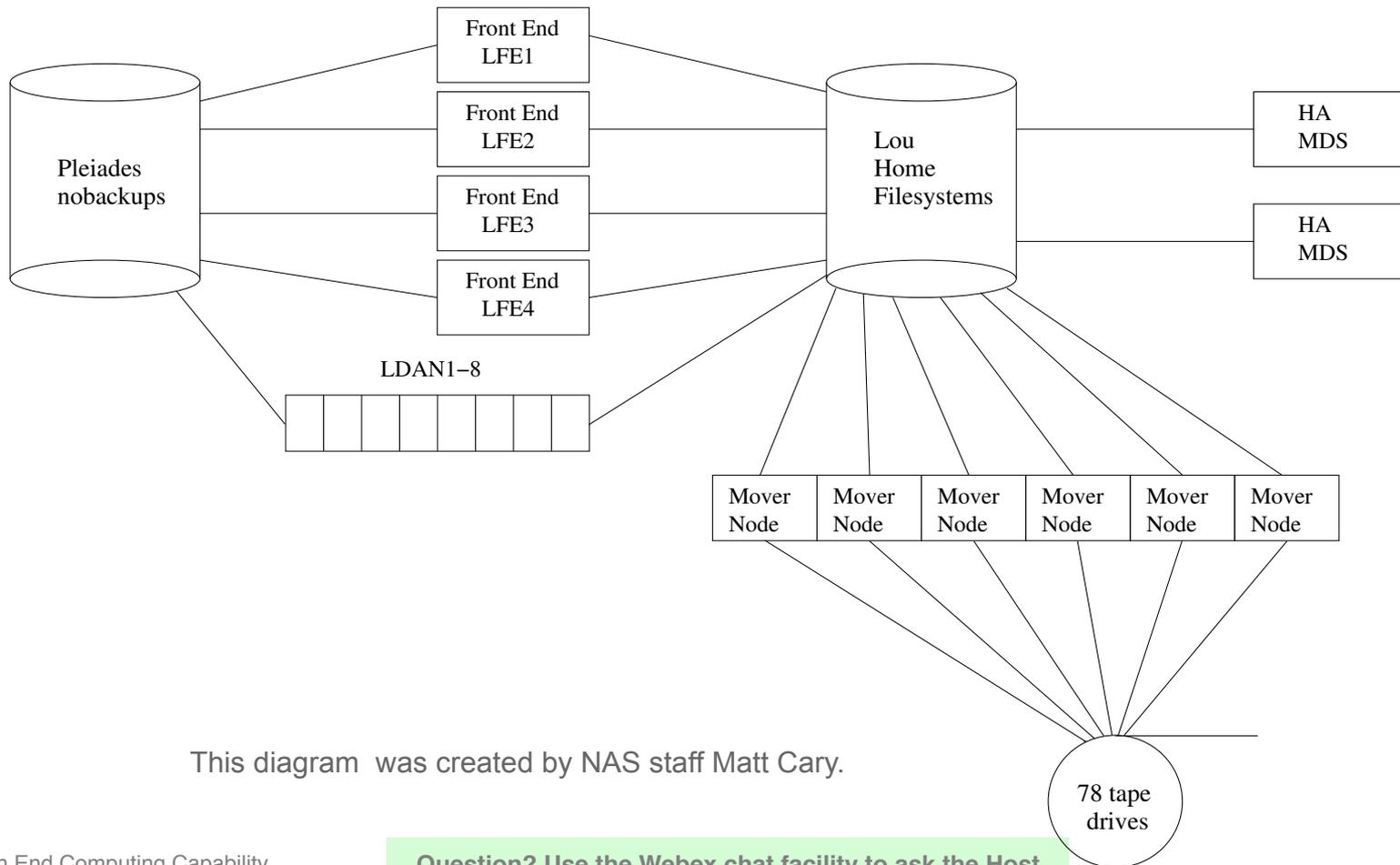
NFS /nobackupnfs2 (to complement /nobackuppX)

- For codes that do lots of small IO and do not perform well on Lustre
- Has data validation capability between NFS server and storage device
- ~ 700 TB (60x more than old /nobackupnfs1) (Aug, 2013)

Lou Archive System Status



- Completed parallel DMF upgrade and migration of all Lou users (Nov 2013)
- Provide high performance and scalability for managing archive data
 - Hardware: Itanium based (old Lou1/2) -> Sandy Bridge x86 based cluster
 - disk: ~3 PB, tape: ~115 PB (for 2 copies, unique data ~60PB)



This diagram was created by NAS staff Matt Cary.

Lou Archive System Status (cont'd)



- Lou front-ends (San, 16 cores/64 GB each): 2 (Dec 2012) -> 4 (now)

%ssh lfe3 (with specific lfe, for example)

%ssh lfe (with load balancer)

- may need to modify your `.ssh/config` to allow load balancer to work.

- Download `config_nas.txt` from <http://www.nas.nasa.gov/hecc/support/kb/files/>

- All Pleiades Lustre `/nobackup` are mounted on lfe's; Easy disk-to-disk file transfer between pfe's and lfe's (no need to do scp, or going through bridge nodes any more – as for transferring data between Columbia and Lou)

lfe% ls -ld \$HOME

..... /u/your_name -> /sxx/your_name (a lou local disk space)

lfe% cd /nobackup/your_name; pwd

/pleiades/nobackup/your_name

lfe% shiftc /u/your_name/file1 /nobackup/your_name

- Getting more than 400 MB/s should be common; retrieving data from tape will take extra time *(lfe3,4 may be faster until Lustre software upgrade on lfe1,2)*

- Recent webinar: Effective Use of the Lou Storage Cluster (Dec 11, 2013)

http://www.nas.nasa.gov/hecc/support/past_webinars.html



Status of LDAN Nodes

- Lou back-end nodes (Lou Data Analysis Nodes, or Idan nodes):
8 Idan nodes (Sandy Bridge processors, March 2013)
 - Idan[1-3]: 96 GB
 - Idan[4-8]: 256 GB
- Now serve two purposes: Idan mode and bigmem mode (Jan 2014)
 - Idan mode to process data on Lou (using lou's \$HOME)
pfe or lfe% qsub -q Idan -l...:mem=252GB (limit 2 jobs per user)
Make sure to have needed modules loaded (in Lou's dot files or PBS scripts)
 - bigmem mode for Pleiades compute jobs (using Pleiades \$HOME)
pfe% qsub -q {normal,long,debug,devel} -l...:model=Idan:mem=252GB
- Knowledge Based articles 222 and 413