Research Computing
The Apollo HPC Cluster

Tom Armour
Jeremy Maris
Research Computing
IT Services
University of Sussex

Apollo Cluster - Aims

- Computing provision beyond capability of desktop
- Priority access determined by HPC Advisory Group
- HPC User Group to be formed next term
- Shared infrastructure and support from IT Services
- Extension by departments
  - Storage (adding to Lustre, access to SAN)
  - CPU
  - Software Licenses
- Expansion by IT Services as budgets allow

High Performance Computing ?

- Generically, computing systems comprised of multiple processors linked together in a single system. Used to solve problems beyond the scope of the desktop.

- High performance computing
  - Maximising number of cycles per second, usually parallel
- High throughput computing
  - Maximising number of cycles per year, usually serial
- Facilitating the storage, access and processing of data
  - Coping with the massive growth in data

High Performance Computing ?

- Single problem split across many processors
  - tasks must run quickly
  - tightly coupled, task parallel,
  - communication between threads
    - Weather forecasting
    - Theoretical chemistry
    - Imaging processing
      - 3D image reconstruction
      - 4D visualisation
    - Sequence assembly
    - Whole genome analysis

High Throughput Computing

- A lot of work done over a long time frame
  - one program run many times, eg searching a large data set
  - loosely coupled (data parallel, embarrassingly parallel)
    - ATLAS analysis
    - Genomics (sequence alignment, BLAST etc)
    - Virtual Screening (eg in drug discovery)
    - Parameter exploration (simulations)
    - Statistical analysis (eg bootstrap analysis,)

Apollo Cluster - Hardware

- Total 488 cores
  - 22 x 12 core 2.67GHz Intel nodes - 264 cores
  - 2 x 48 core 2.2GHz AMD nodes – 96 cores
  - 17 blades , GigE (informatics) - 128 cores
  - 48 GB RAM for Intel nodes
  - 256 GB RAM for AMD nodes
- 20 TB Home NFS file system (backed up)
- 80 TB Lustre scratch file system (not backed up)
- QDR (40Gbs) Infiniband interconnect

Apollo Cluster - Filesystems

- Home - 20 TB, RAID 6 set of 12 x 2TB disks
  - Exported via NFS
  - Backed up, keep your valuable data here
  - Easily overloaded if many cores read/write at same time
- Lustre parallel file system 80TB
  - Redundant metadata server
  - Three object servers, each with 3 x 10 TB RAID6 OST
  - Data striping configured by file or directory
  - Can stripe across all 108 disks. Aggregate data rate ~ 3.8GB/s
  - NOT backed up, for temporary storage. /mnt/lustre/scratch
- Local 400GB scratch disk per node

Apollo Cluster - Lustre Performance

Servers	All-OSS		OSS1	OSS2	OSS3
Clients	18		6	6	6
metric	KB/s	IOPS	KB/s	KB/s	KB/s
write	3,854,980	60,235	1,273,054	1,272,119	1,273,239
rewrite	3,005,168	46,956	1,039,489	1,117,662	1,020,844
read	2,982,143	46,596	1,143,591	1,164,641	1,165,076
reread	18,411,516	287,680	9,127,979	8,918,498	7,724,083
random read	151,069	2,361			
random write	261,675	4,089			
reverse read	251,587	3,932			
stride read	303,852	4,748			
mixed	280,632	4,385			

IOZONE Benchmark Summary

Apollo Cluster - Software

- Module system used for defining paths and libraries
  - Need to load software required
  - Module avail, module add XX, module unload XX
  - Access optimised math libraries, MPI stacks, compilers etc
  - Latest version of packages, eg python, gcc easily installed
- Intel parallel studio suite
  - C C++ Fortran, MKL, performance analysis etc
- gcc and Open64 compilers
- Jeremy or Tom will compile/install software for users

Apollo Cluster - Software

Compilers/tools	Libraries	Programs
ant	acml	adf
gcc suite	atlas	aimpro
Intel (c, c++, fortran)	blas, gotoblas	alberta
git	cloog	FSL
jdk 1.6_024	fftw2, fftw3	gadget
mercurial	gmp	gap
open64	gsl	Gaussian 09
python 2.6.6	hpl	hdf5
sbcl	lapack, scalapack	idl
	MVAPICH, MVAPICH2	matlab
	mpfr	mricron
	nag	netcdf
	openMPI	paraview
	ppl	stata
		WRF

Apollo Cluster - Queues 1

- Sun Grid Engine used for batch system (soon UNIVA)
- parallel.q for MPI and OpenMP jobs
  - Intel nodes
  - Slot limit of 36 cores per user at present
- serial.q for serial, OpenMP and MPI jobs
  - AMD nodes
  - Slot limit of 36 cores per user
- Informatics only queues inf.q + others
- No other job limits – need user input for configuration

Apollo Cluster - serial job script

- Queue is by default the serial.q

```
#!/bin/sh
#$ -N sleep
#$ -S /bin/sh
#$ -cwd
#$ -q serial.q
#$ -M user@sussex.ac.uk
#$ -m bea
echo Starting at: `date`
sleep 60
echo Now it is: `date`
```

Apollo Cluster - parallel job script

- For parallel jobs you must specify the pe - parallel environment.
- parallel environments: openmpi, openmp, mvapich2 often less efficient.

```
#!/bin/sh
#$ -N JobName
#$ -M user@sussex.ac.uk
#$ -m bea
#$ -cwd
#$ -pe openmpi NUMBER_OF_CPUS # eg 12-36
#$ -q parallel.q
#$ -S /bin/bash
# source modules environment:
. /etc/profile.d/modules.sh
module add gcc/4.3.4 qlogic/openmpi/gcc
mpirun -np $NSLOTS -machinefile $TMPDIR/machines /path/to/exec
```

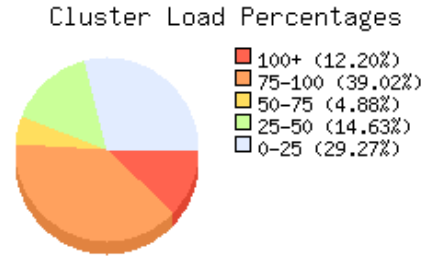
Apollo Cluster - Monitoring

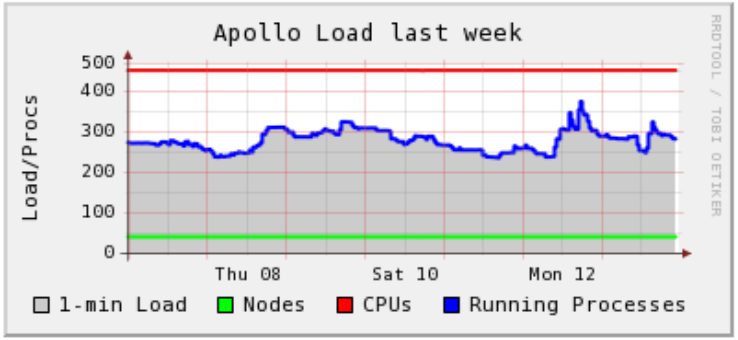
Ganglia statistics for apollo and feynman (the EPP cluster) are at http://feynman.hpc.susx.ac.uk

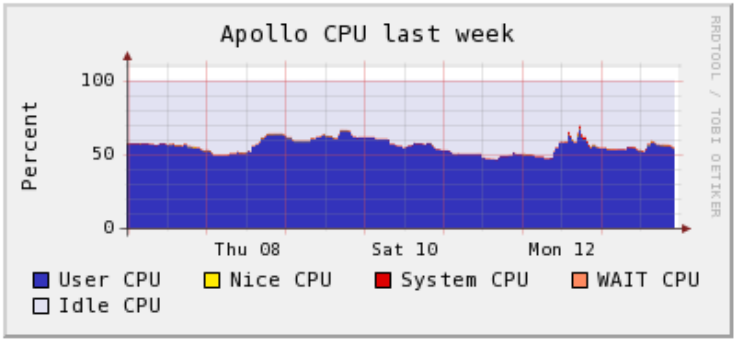
CPUs Total:	**452**
Hosts up:	**41**
Hosts down:	**0**

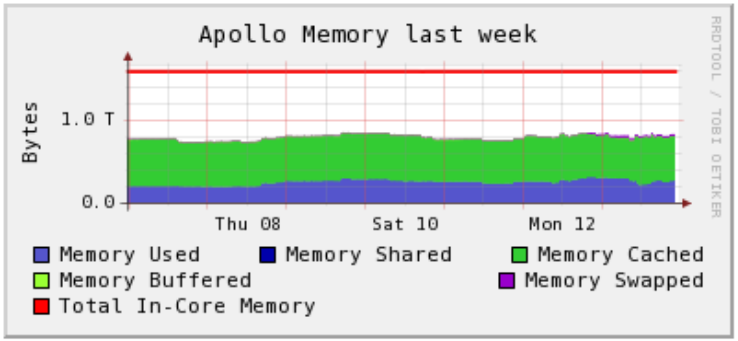
Avg Load (15, 5, 1m):
62%, 61%, 61%

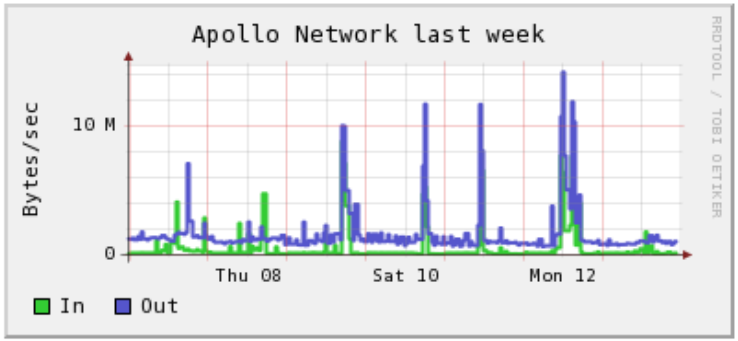
Localtime:
2011-12-13 23:26



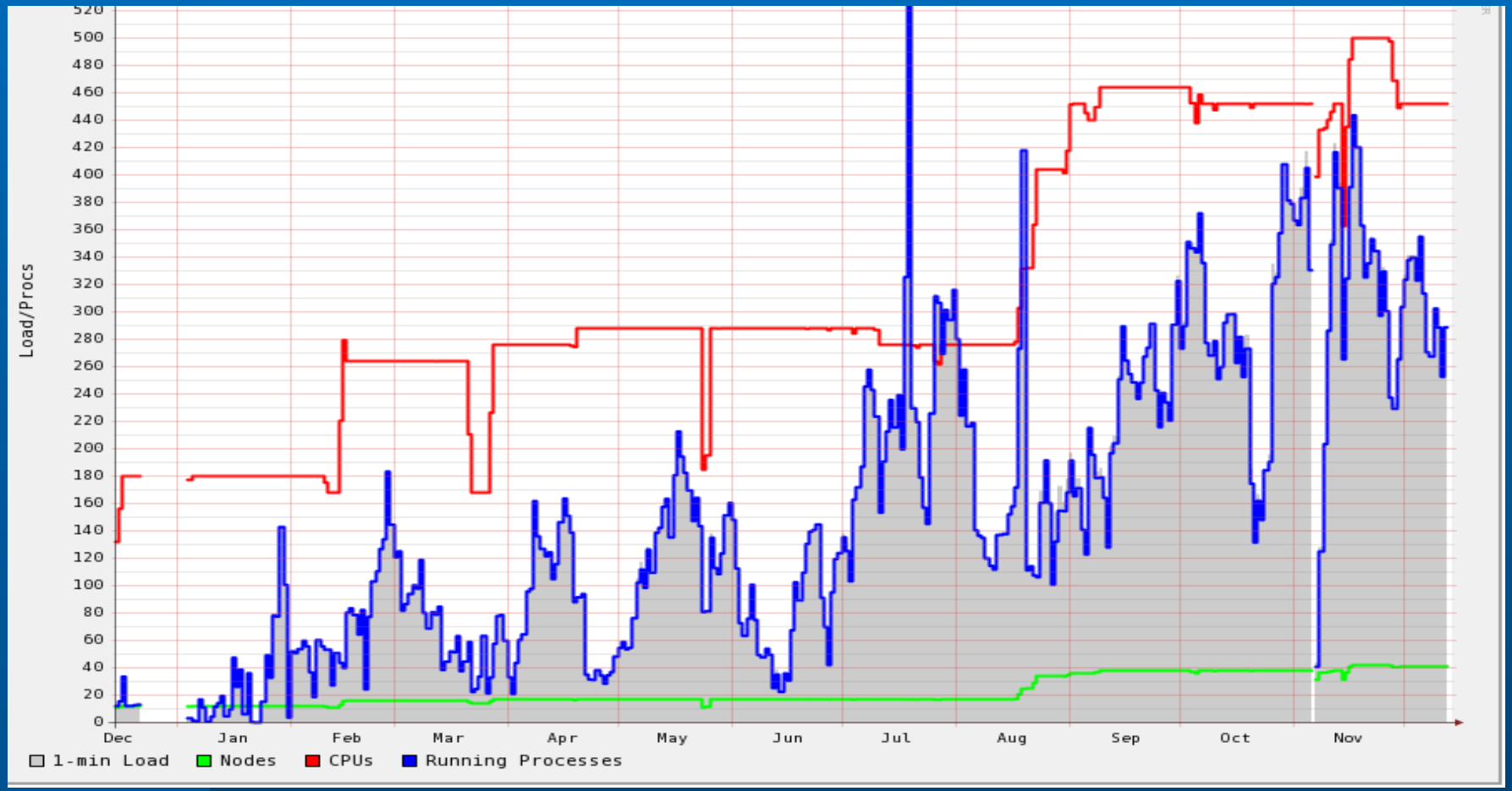








Apollo Cluster - Historic load



Apollo Cluster - Accounting

- Simply, qacct –o
- Accounting monitored to show fair usage, etc
- UNIVA's UniSight reporting tool coming soon.
- Largest users of CPU are Theoretical Chemistry and Informatics
- Other usage growing, eg Engineering (CFD), weather modelling (Geography) and fMRI Image analyis (BSMS and Psychology)
- Use the AMD nodes (serial.q and inf.q) !

Apollo Cluster - Documentation

- For queries and accounts on the cluster, email researchsupport@its.sussex.ac.uk
  - Jeremy and Tom monitor this; your email won't get lost!
- Userguides and example scripts are available on the cluster in /cm/shared/docs/USERGUIDE
- Other documentation in /cm/shared/docs path
- Workshop slides available from wiki and http://www.sussex.ac.uk/its/services/research
- NEW- Wiki at https://www.hpc.sussex.ac.uk

Apollo Cluster - Adding nodes

- C6100 chassis + four servers each 2.67 GHz, 12 core, 48GB RAM, IB card, licences ~ £14,600
- R815 48 core 2.3GHz AMD 128GB ~£9,300
- C6145 New AMD Interlagos – 2 x64 core nodes each with 128GB ~£15,500
- Departments guaranteed 90% exclusive use of their nodes, 10% sharing with others, plus back fill of idle time.

Contact researchsupport@its.sussex.ac.uk for pricing. We can get quotes for your research proposals.

Feynman Cluster

- Initially a “Tier-3” local facility for Particle Physics
- Same system architecture as Apollo, logically separate
- Used to analyse data from the ATLAS experiment at the Large Hadron Collider at CERN
- Serial workload with high IO and storage requirements
- 96 cores and 2/3 of Lustre storage (54 TB)
- Joining GridPP as a Tier-2 national facility, with expansion of storage and CPU
- System managed by Emyr James, in collaboration with IT Services

Future Developments

- Wiki for collaborative documentation
- UNIVA Grid Engine and Hadoop/Map Reduce
- Moving cluster to new racks to for allow expansion
- Adding new Infiniband switches
- Updating Lustre to 1.8.7
- Adding 60 TB more Lustre storage for EPP
- Consolidating clusters (Feynman, Zeus)
- Two days downtime sometime in February or early March

Questions?

Questions?

- Problems and Issues
- Access to resources
  - Who should have priority access ?
  - Who should have a lesser priority ?
  - Fair share access for all?
- Queue design
  - Time limits?
  - Minimum and/or maximum slots on parallel.q ?
- Requests ?

Questions?

- Is this a useful day ?
- Do you want another one ?
  - What should we cover ?
  - When?
- Shorter “hands on” training sessions?
- Session on compilers and code optimisation ?