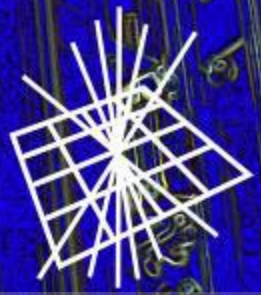


GridPP

UK Computing for Particle Physics

Site Report

Liverpool

Steve Jones, John Bland, Rob Fay

Introduction

- Liverpool HEP group host a general “Physics Computing Facility” with many local users, and a T2 grid site. Many of the resources are shared.
- We've restructured our grid site somewhat, so I'll give a summary of what we bought and how it looks now.
- Eighteen months ago, it was all Cream/Torque/Maui, with support for 25 VOs or so, traffic from (say) 12.
- Now it's Cream/Torque/Maui/Arc/Condor/VAC/VAC (tbd), still supporting 25 VOs....
- And we've put some more disk and cpu online ,many services are virtualised and we have plans to beef up the monitoring.
- Obviously, that's a lot of change, so I'll show a trick that helps us manage it.

What we bought

- We expanded our CPU by buying
  - 5 x E5-2630V3 (2 cpu, 32 cores, 11.07 HS06 per core, 4.12 GB RAM per core, total 1771 HS06, i.e. 9 % boost)
  -
- We grew our storage by buying
  - 4 new storage nodes: 132TB, 96GB RAM, 2x10Gb dual-homed
  - Using 6TB WD NAS PRO SATA drives.
  - Small gamble as not supported in 24bay NAS but allowed us to afford an extra server.
  - Our suspicion is NAS PRO is Enterprise hardware with slightly modified firmware and a new label.

Overall

- Overall
  - Site hs06: 21329
  - Site slots: 1986
- Logical grid site Comprises two physical sites
  - Physics Cluster Room (large), which we call HAMMER
  - Central Cluster Room (1 x rack), which we call Chadwick
  -
- Special routing required to join them together.

CPU Types

Node types:

Name,	CPUs,	Slots,	HS06 (Slot),	RAM (Slot),	Scale factor
BASELINE,	0,	0,	10.00,	0.00,	0.0000
L5530,	2,	7,	12.34,	3.50,	1.2340
L5420,	2,	8,	8.90,	2.00,	0.8900
E5620,	2,	12,	10.63,	2.00,	1.0633
X5650,	2,	24,	8.66,	2.08,	0.8660
E5-2630,	2,	23,	11.28,	2.17,	1.1280
E5-2630V3,	2,	32,	11.07,	4.12,	1.1070
E5620-VAC,	2,	10,	12.05,	2.40,	1.2050

ARC Condor Cluster

Cluster: CONDOR_BATCH_HAMMER

Set label	Nodetype	Number	Slots	Slot HS06	HS06
21	E5620	4	12	10.63	510.36
21X	X5650	16	24	8.66	3325.44
22	E5620	20	12	10.63	2551.80
23p1	E5620	10	12	10.63	1275.90
26	E5-2630	4	23	11.28	1037.76
26L	L5420	7	8	8.90	498.40
26V	E5-2630V3	5	32	11.07	1771.20

ARC Condor Cluster

Cluster properties:

- HS06 : 10970
- Physical CPUs : 132
- Logical CPUs (slots): 1100
- Cores: 8.333
- Benchmark: 9.974
- CE_SI00: 2493
- CPUScalingReferenceSI00: 2500.000

Cream/Torque Cluster

Cluster: TORQUE_BATCH_HAMMER

Set label, Nodetype, Number, Slots, Slot HS06, HS06
25p2, E5-2630, 10, 23, 11.28, 2594.40

Cluster properties:

- HS06 : 2594
- Physical CPUs : 20
- Logical CPUs (slots): 230
- Cores: 11.500
- Benchmark: 11.280
- CE_SI00: 2820
- CPUScalingReferenceSI00: 2500.000

VAC Hammer Cluster

Cluster: VAC_CLOUD_HAMMER

Set label	Nodetype	Number	Slots	Slot HS06	HS06
23p2	E5620-VAC	10	10	12.05	1205.00
24	E5620-VAC	20	10	12.05	2410.00
25p1	E5-2630	10	23	11.28	2594.40

VAC Hammer Cluster

Cluster properties:

- HS06 : 6209
- Physical CPUs : 80
- Logical CPUs (slots): 530
- Cores: 6.625
- Benchmark: 11.716
- CE_SI00: 2929
- CPUScalingReferenceSI00: 2500.000

VAC Chadwick Cluster

Cluster: VAC_CLOUD_CHADWICK

Set label, Nodetype, Number, Slots, Slot HS06, HS06

comp5xx, L5530, 18, 7, 12.34, 1554.84

Cluster properties:

- HS06 : 1554
- Physical CPUs : 36
- Logical CPUs (slots): 126
- Cores: 3.500
- Benchmark: 12.340
- CE_SI00: 3085
- CPUScalingReferenceSI00: 2500.000

Why VAC?

- The systems we have deployed have successfully absorbed the huge amount of work generated by the experiments.
- But due to coupling between components, the systems themselves have absorbed huge amounts of effort by admins, engineers, scientists, project managers, planners. If we reduced the coupling, the systems might be more flexible to change, and hence more reliable and easier to install and manage.
- To be fair, reliability of 'traditional' cluster technology has become much better lately (fewer releases.)
- But Virtual Machines dramatically reduce coupling. Unavoidable coupling that remains (which I regard as non-functional from a SW POV) includes:

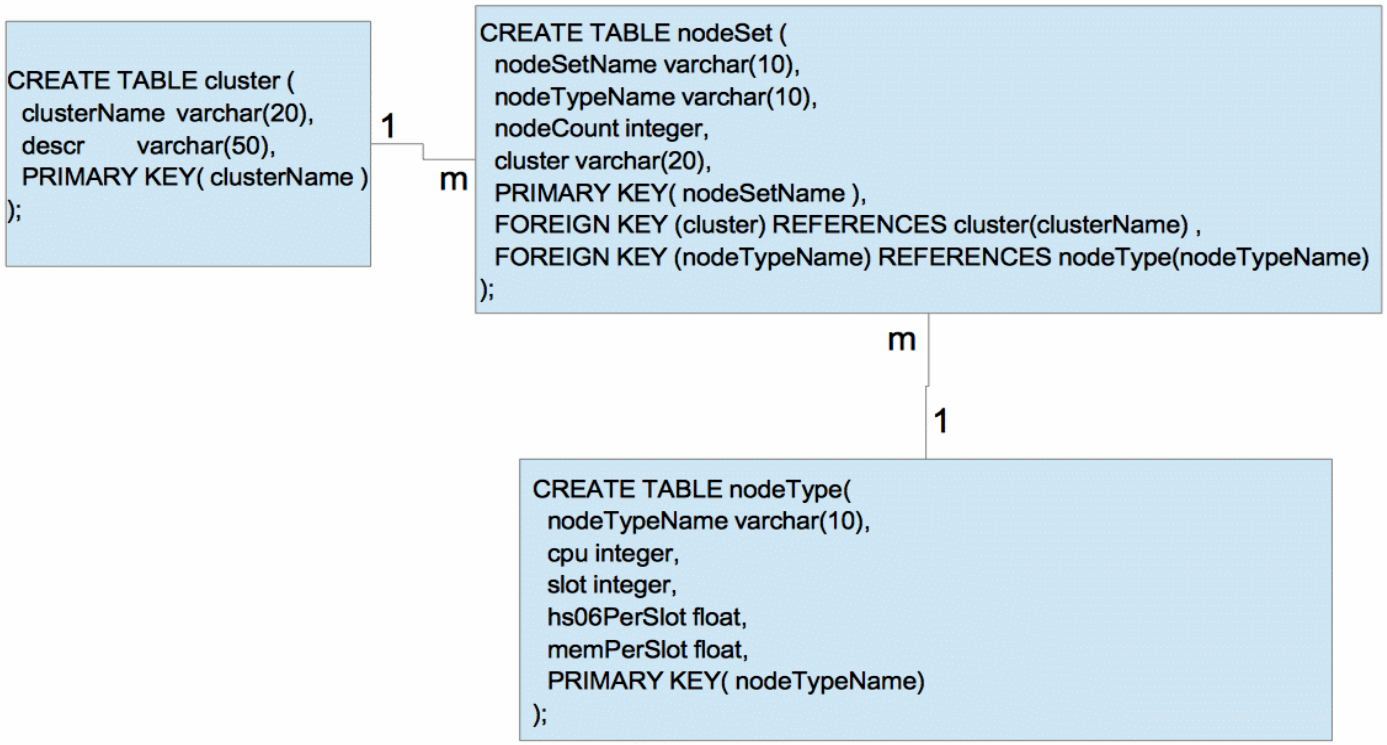
  Walltime, CPUs/Inst. set, RAM, (network, storage, security …?)
- All (almost) of the SW functional complexity is transferred into the VM.

VAC Tips

- Do put squids on the VAC Factories for use by CVMFS. Before we did this, the site squids coped, but cause a torrent of logging which filled our disks.
- Do put /var on it's own huge partition (or use logical volumes/volume groups). VAC downloads large image files which quickly consume small partitions.
- Do monitor your space somehow (see above). Make sure you get emails when things are filling up.

Admin Automation

- We have our own boot build system based on Kickstart with some ideas by Peter Gronbech. We changed to Puppet 3, for Hiera parameterisation. This helps at lot, more modularity, less 'if then else' logic in the puppet modules, less coupling.
- To removes puzzlement over the meaning of Physical CPUs, Logical CPUs (slots), Cores, Benchmark, CE_SI00 and CPUScalingReferenceSI00 etc., we made a rough and ready configuration database that has proved very valuable for allocating nodes and measuring HS06 for the information system.
- I introduce that next, but I haven't got a name for the system yet, nor even a UI. So let's call it "TINDAT", because it “cuts my effort There Is No Doubt About That ….


CREATE TABLE cluster (
clusterName varchar(20),
descr varchar(50),
PRIMARY KEY(clusterName)
);
1
m
CREATE TABLE nodeSet (
nodeSetName varchar(10),
nodeTypeName varchar(10),
nodeCount integer,
cluster varchar(20),
PRIMARY KEY(nodeSetName),
FOREIGN KEY (cluster) REFERENCES cluster(clusterName) ,
FOREIGN KEY (nodeTypeName) REFERENCES nodeType(nodeTypeName)
);
m
1
CREATE TABLE nodeType(
nodeTypeName varchar(10),
cpu integer,
slot integer,
hs06PerSlot float,
memPerSlot float,
PRIMARY KEY(nodeTypeName)
);

TINDAT

```
INSERT INTO nodeType VALUES ("L5530"   ,  2,    7,    13.370 ,  3.4285);
INSERT INTO nodeType VALUES ("L5420"   ,  2,    8,     8.960 ,  2.000);
INSERT INTO nodeType VALUES ("E5620"   ,  2,   10,    12.050 ,  2.500);
etc...

INSERT INTO cluster VALUES ("VAC_CLOUD_CHADWICK","Vac  nodes in  Chadwick room");
INSERT INTO cluster VALUES ("TORQUE_BATCH_HAMMER","Torque nodes ");
etc.

INSERT INTO nodeSet VALUES ("comp5xx", "L5530", 18,  "VAC_CLOUD_CHADWICK");
INSERT INTO nodeSet VALUES ("21", "E5620"     ,  4,  "CONDOR_BATCH_HAMMER");
INSERT INTO nodeSet VALUES ("21X", "X5650"    , 16,  "CONDOR_BATCH_HAMMER");
etc.
```

No UI (yet?); maintain with SQL INSERT, DELETE and UPDATE.

TINDAT

- But what does it do? It uses a python script, clusterReport.py, to select and format the data from the TINDAT database to report:
  - Scaling values, rack layouts, HS06 breakdowns, rack and cluster breakdowns and totals for each of our four different clusters (see slides 3 to 10)
- The data is in the appropriate format(s) to feed into both the YAIM config management system, as well as the ARC system and VAC, thus automating the cluster layout, capacity and power metrics for the BDII system.
- My specifications for interfaces data are in:
  - https://www.gridpp.ac.uk/wiki/Publishing_tutorial
  - https://www.gridpp.ac.uk/wiki/Example_Build_of_an_ARC/Condor_Cluster

Liverpool is Fragmented

- Condor is there because ATLAS needed facilities for multicore jobs, and we wanted to change from Torque anyway because of software engineering issues.
- So that makes about half the HS06, at present.
- In late 2015, we were loaned ~ 18 worker-nodes from a central services department. Since the worker-nodes reside on another network in a remote server room with limited access and no head node, we set them up with VAC.
- In 2016, we were asked to act as a VAC test site. We converted a sizeable portion of that cluster to VAC, which is currently ~ 660 slots, perhaps going to 800 “eventually”.
- Hence three different cluster technologies over two sites. This is a nightmare to admin manually.

GridPP
UK Computing for Particle Physics

Storage: Hardware

- 4 new storage nodes: 132TB, 96GB RAM, 2x10Gb dual-homed
  - Using 6TB WD NAS PRO SATA drives
  - Small gamble as not supported in 24bay NAS but allowed us to afford an extra server
    - Our suspicion is NAS PRO is Enterprise hardware with slightly modified firmware and a new label (or vice versa)
- Total WLCG storage now 1173TB (+51TB Local)
  - ATLAS 769TB (138TB free)
  - LHCB 308TB (295TB free, just starting)
  - T2K/SNOPLUS 103TB (50TB free)

UNIVERSITY OF LIVERPOOL

Storage: Software

- New DPM pool nodes running DPM 1.8.10
  - Configured with Puppet and DPM modules
    - Just Works
- All systems on SL6
  - Tried Centos7 a while ago but too many problems
- Existing pool nodes and head node on 1.8.8 with YAIM
  - Aim to update or reinstall all pool nodes to 1.8.11, puppet
  - Once complete migrate to new DPM head node hardware
- WLCG and Local share DPM infrastructure for bulk data
  - No plans to set up separate clusterFS unless we get big lump of new hardware
    - As long as nobody needs more than XRootD we're golden.

Virtualisation: Upgrade

- Currently use KVM on SL6 for most service nodes
  - Resources shared between Grid and Local HEP
- Image files stored on NFS server
  - Performance is fine but reaching 100% capacity
- New KVM host and storage server
  - 20TB RAID10
  - Shared via NFS as standard
    - All guests migratable between 5 hosts
  - Also provide block access via libvirtd/LVM and libvirtd/iscsi if required
    - Don't see any real performance differences between NFS images and block devices, images being much easier to manage
  - Snapshot with LVM for backups

Virtualisation: continued

- Reliance on SPOF storage is not ideal but budget constrained
  - Cluster FS e.g. GlustreFS or similar would be better but outside the remaining available budget
    - Spread out budgets lead to continual small, inefficient upgrades
- Read starvation on heavy writes a concern for image storage
  - See GridPP Storage thread
  - Possibly always been an issue, is it Linux, RAID controllers, both? Does ZFS-on-Linux cope better?
- A number of retired DPM pool nodes available
  - Use for a Cloud computing or storage test-bed
  - Openstack, IPv6 etc

Network: Core

- 3xDell S4810 stacked giving virtual “chassis”
  - 2x20Gb failover WAN uplink to Campus
    - University JANET connection 2x10Gb (that we know of)
  - Dual-homed for most systems
    - External VLANs and Internal research VLANS
    - Internal VLANs have always used Jumbo Frames
  - Running short on ports
    - Chained a new 10G Switch to expand a local cluster
    - Don’t buy D-Link 10G switches, the interface is hopeless!
- New IP address range specifically for Grid/Research
  - In sort-of-DMZ with no University firewall filtering
  - Keep having to remind University network team that “no filters” means “no filters” not “some filters”

GridPP
UK Computing for Particle Physics

Network: IPv6

- Still waiting for University network core to upgrade firmware to support IPv6
  - Promised this will happen in August
  - All other hardware is supposed to be IPv6-ready
- We will then hopefully get our IPv6 allocation
  - Anything we can usefully do before then to prepare?

UNIVERSITY OF LIVERPOOL

Other things

- We're overhauling our configuration and monitoring infrastructure with a new one.
- We're still using puppet, but adding foreman for a dashboard and ENC.
- For monitoring side we're using sensu, ELK (Elasticsearch, Logstash, Kibana) with filebeat, collectd, graphite, grafana, and some custom software for managing alerts (in development.)

Conclusions

- Liverpool has changed a lot, and we've had to change our management practises to suit.
- In recent years, the buying round seems to have alternated between CPU and Disk. Perhaps that's an expected outcome. This year was Disk's turn.
- We continue to push for Ipv6 but are still waiting for third parties.
- Looking hard at Centos 7 now. Lots of porting and migration coming up.

Plans

- Cpu:
  - Ramp up to 800 slots of VAC. Tone down torque.
  - Continue to support ARC/Condor for
    - ATLAS Multicore, LHCb
    - VOs unable or unwilling to use pilot factories or develop VM dev. capabilities
- Storage:
  - Continue with DPM for the time being at least. The evolution of storage is an important topic that needs more definition.
- Long term:
  - Review evolution and respond accordingly. Too many imponderables to know, but (we think) VM technologies will open new paths.

End

- Questions?