Lessons Learnt in the Management of CDF Hardware and Service Contract

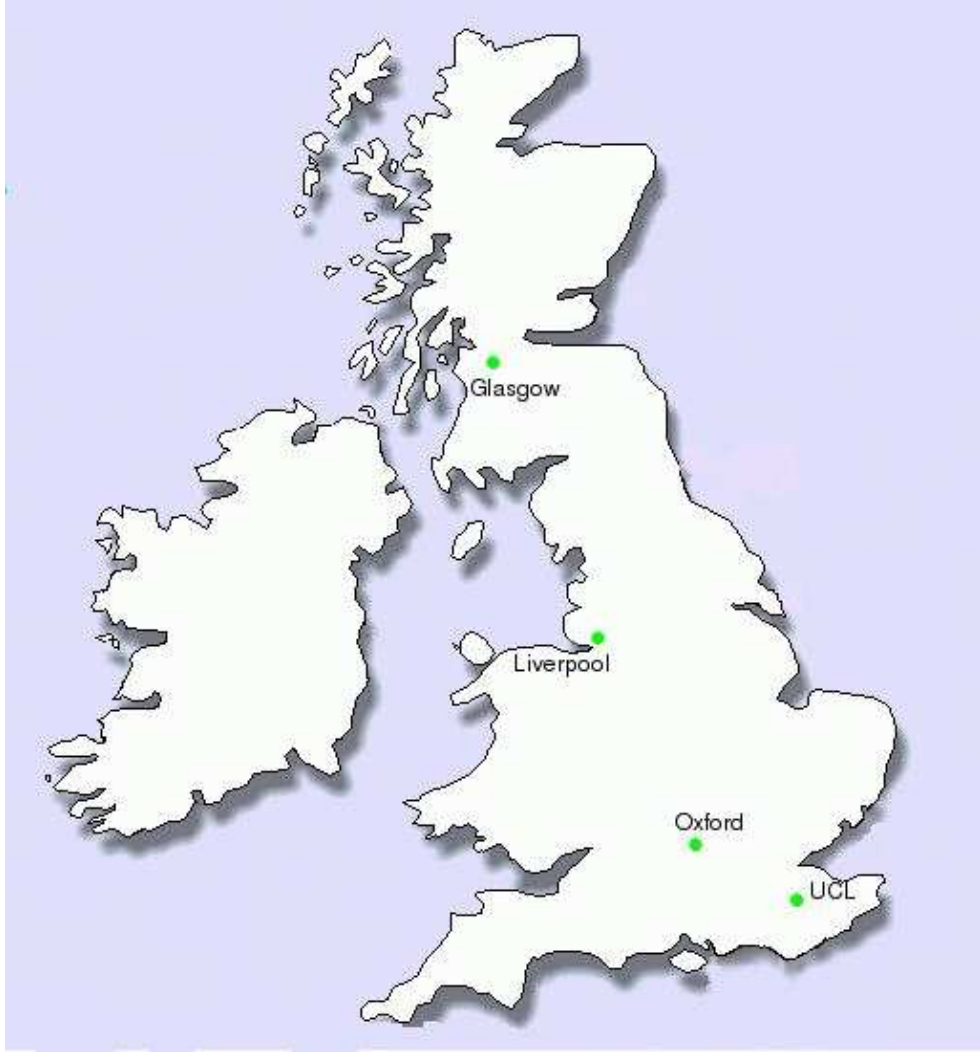
Mòrag Burgon-Lyon, Paul Millar, Richard St.Denis, A. Stan Thompson

University of Glasgow, CDF collaboration

Contents

Introduction

4 UK Institutions own almost identical clusters



Introduction

- Equipment for Collider Detector at Fermilab (CDF) Experiment.

Introduction

- Equipment for Collider Detector at Fermilab (CDF) Experiment.
- Clusters originally comprised: a server; 9 worker nodes; and 4 RAID systems, each of 1.7TB. An additional RAID and node were added in November 2003.

Introduction

- Equipment for Collider Detector at Fermilab (CDF) Experiment.
- Clusters originally comprised: a server; 9 worker nodes; and 4 RAID systems, each of 1.7TB. An additional RAID and node were added in November 2003.
- Server and worker nodes running Fermi Linux 7.3 (based on Red Hat 7.3).

Introduction

- Equipment for Collider Detector at Fermilab (CDF) Experiment.
- Clusters originally comprised: a server; 9 worker nodes; and 4 RAID systems, each of 1.7TB. An additional RAID and node were added in November 2003.
- Server and worker nodes running Fermi Linux 7.3 (based on Red Hat 7.3).
- Kerberos installed, a mandatory requirement for access to Fermilab computers.

Introduction

- Equipment for Collider Detector at Fermilab (CDF) Experiment.
- Clusters originally comprised: a server; 9 worker nodes; and 4 RAID systems, each of 1.7TB. An additional RAID and node were added in November 2003.
- Server and worker nodes running Fermi Linux 7.3 (based on Red Hat 7.3).
- Kerberos installed, a mandatory requirement for access to Fermilab computers.
- The problems were encountered on the original 4 RAID systems, formatted with Reiser file system.

What happened

- Glasgow cluster was installed in February 2003.

What happened

- Glasgow cluster was installed in February 2003.
- Initial testing was carried out on Oxford hardware. On completion the others clusters were cloned from the Oxford station. No burn-in tests were carried out on the Glasgow hardware.

What happened

- Glasgow cluster was installed in February 2003.
- Initial testing was carried out on Oxford hardware. On completion the others clusters were cloned from the Oxford station. No burn-in tests were carried out on the Glasgow hardware.
- Problems began on 24th January 2004 when a data transfer of the order of 1TB from Fermilab to a cache on one of the RAID disks was attempted.

What happened

- Glasgow cluster was installed in February 2003.
- Initial testing was carried out on Oxford hardware. On completion the others clusters were cloned from the Oxford station. No burn-in tests were carried out on the Glasgow hardware.
- Problems began on 24th January 2004 when a data transfer of the order of 1TB from Fermilab to a cache on one of the RAID disks was attempted.
- Over a period of 2 months the RAID systems encountered problems of increasing severity.

What happened

- System administration of the Glasgow cluster is a secondary role as budget covered equipment and service agreement only.
- A large amount of time was spent in communication with the vendor speculating about possible causes, suggesting investigation tools and potential solutions.
- Tests took a long time to run, e.g. a single pass of read-write test with bad blocks tool takes 4 days for 1.7Tb.
- Evidence points to poor hardware installation as root cause. Vendor eventually replaced SCSI cables that were found to be damaged, the SCSI adaptor card, 4 disks and the RAID controller firmware.

Support Requirement

- FTE=A + B*N + C*log(N) (N=number of nodes)
- A = Time required to learn and remain familiar with setup, software and help users with problems, etc.
- B = Hardware problems
- C = Software problems (e.g. grid tools, security updates, etc). This scales with nodes, but far more gently as many nodes are clones.

Lessons Learnt

Our experience with the RAID system has taught the following lessons:

- Hardware problems can be difficult to diagnose.

Lessons Learnt

Our experience with the RAID system has taught the following lessons:

- Hardware problems can be difficult to diagnose.
- Service contracts alone are not enough to solve problems efficiently.

Lessons Learnt

Our experience with the RAID system has taught the following lessons:

- Hardware problems can be difficult to diagnose.
- Service contracts alone are not enough to solve problems efficiently.
- Need to be able to quickly clarify when the problem is the vendor's responsibility.

Lessons Learnt

Our experience with the RAID system has taught the following lessons:

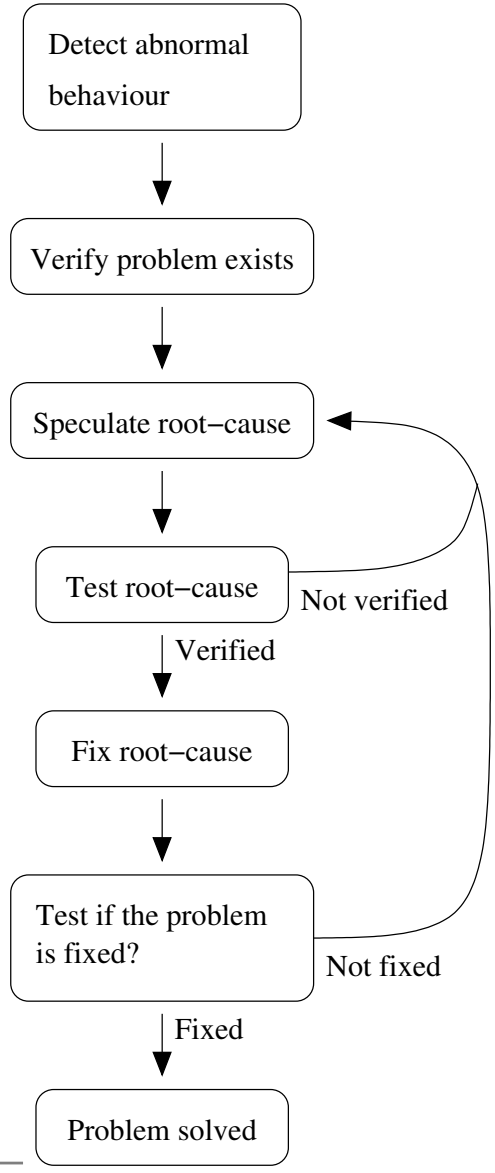
- Hardware problems can be difficult to diagnose.
- Service contracts alone are not enough to solve problems efficiently.
- Need to be able to quickly clarify when the problem is the vendor's responsibility.
- Cluster needs a thorough testing workout before users come on the system.

Lessons Learnt

Our experience with the RAID system has taught the following lessons:

- Hardware problems can be difficult to diagnose.
- Service contracts alone are not enough to solve problems efficiently.
- Need to be able to quickly clarify when the problem is the vendor's responsibility.
- Cluster needs a thorough testing workout before users come on the system.
- The same tests can then be used to aid problem diagnosis.

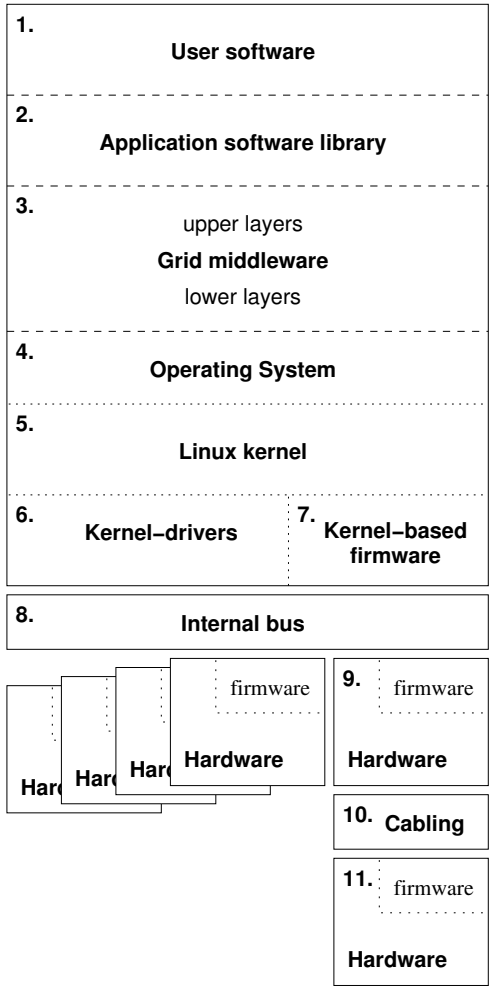
Solving hardware problems



- When abnormal behaviour is observed, the first step is to speculate as to the root-cause, followed up with investigation.
- If the cause is verified, a fix can be applied and tested.
- If not verified or fixed, the problem is examined again with further speculation as to the cause.
- All these steps incur a time penalty for the system administrator.

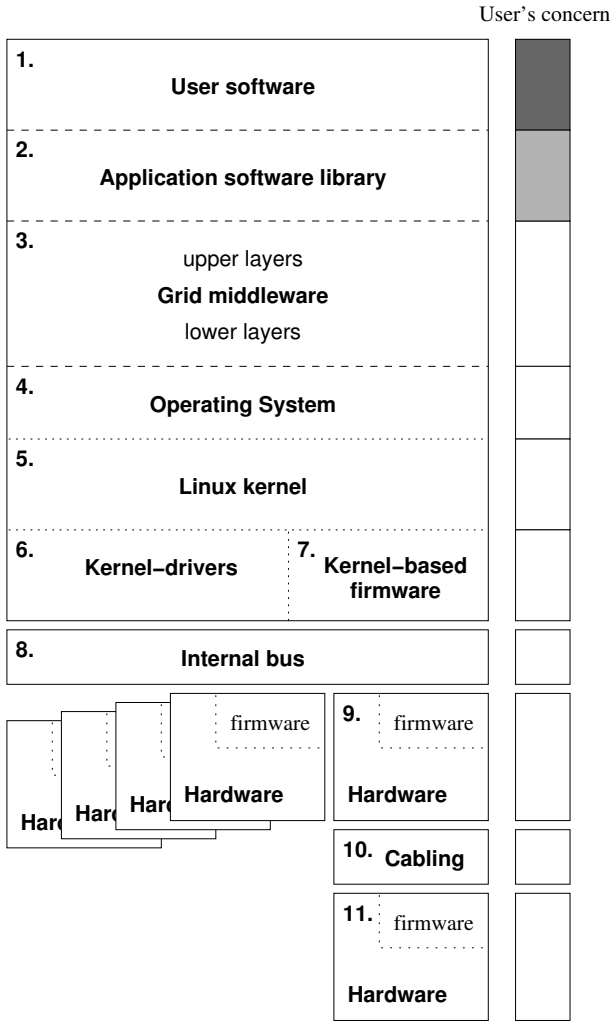
Layers in a cluster

Divisions of responsibility between user, system administrator and vendor



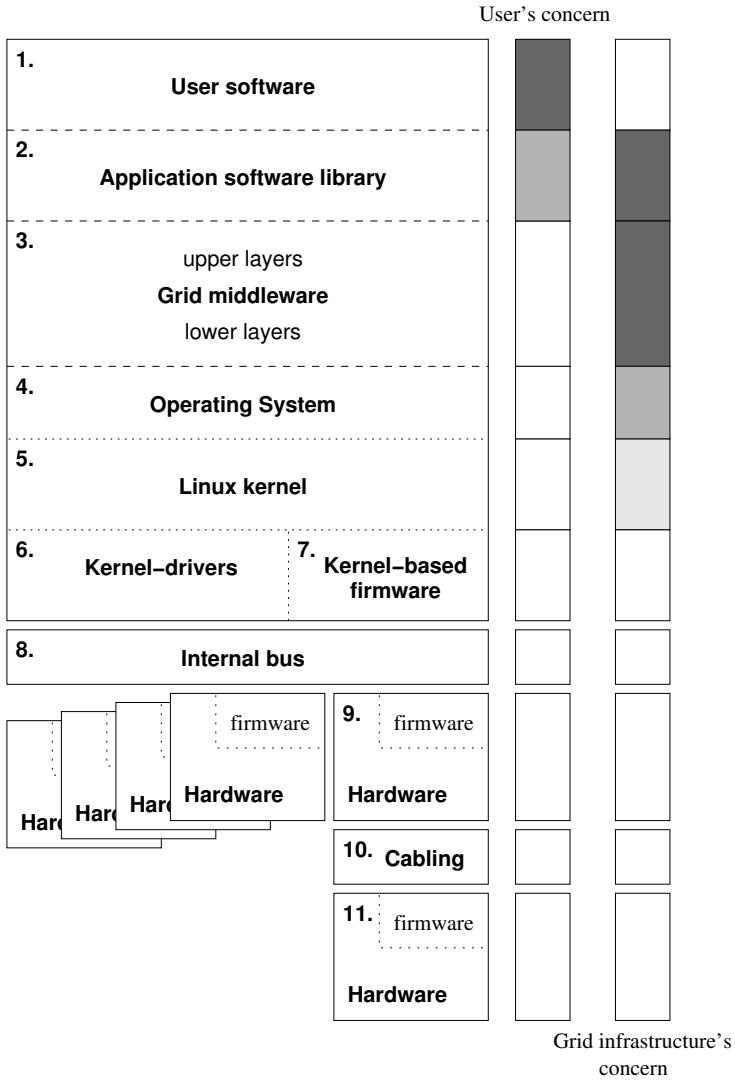
Layers in a cluster

Divisions of responsibility between user, system administrator and vendor



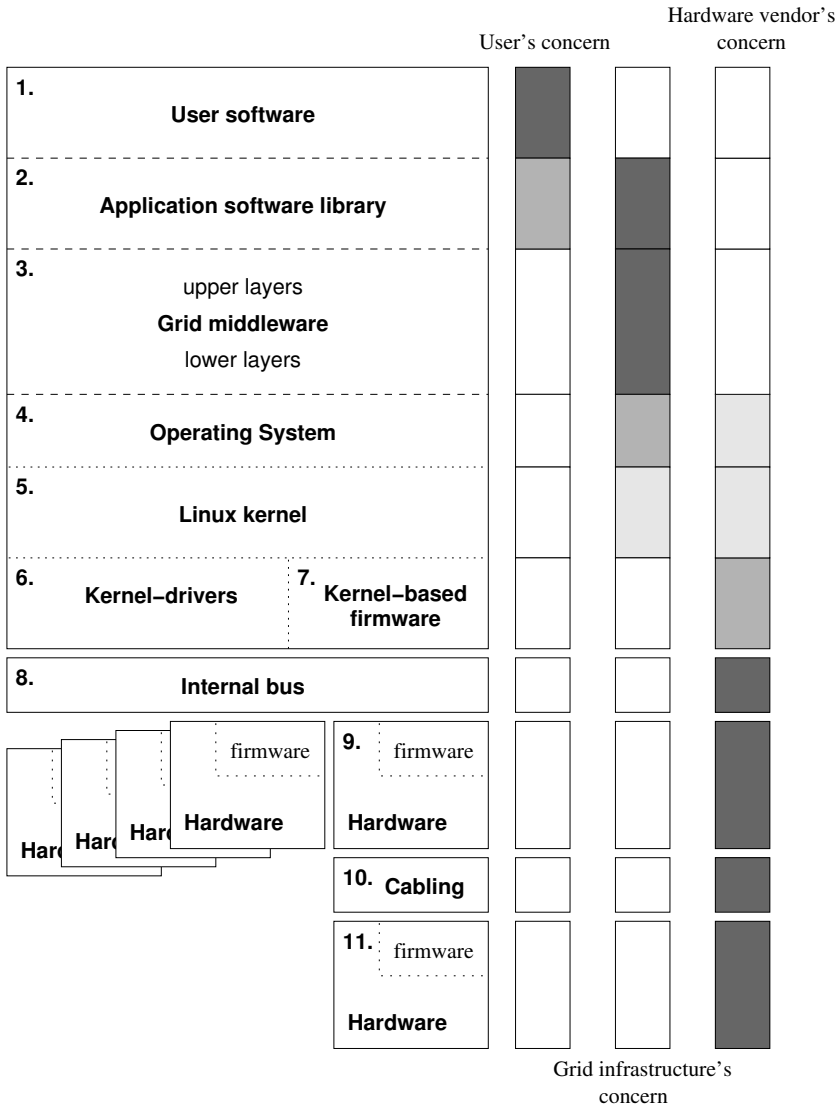
Layers in a cluster

Divisions of responsibility between user, system administrator and vendor



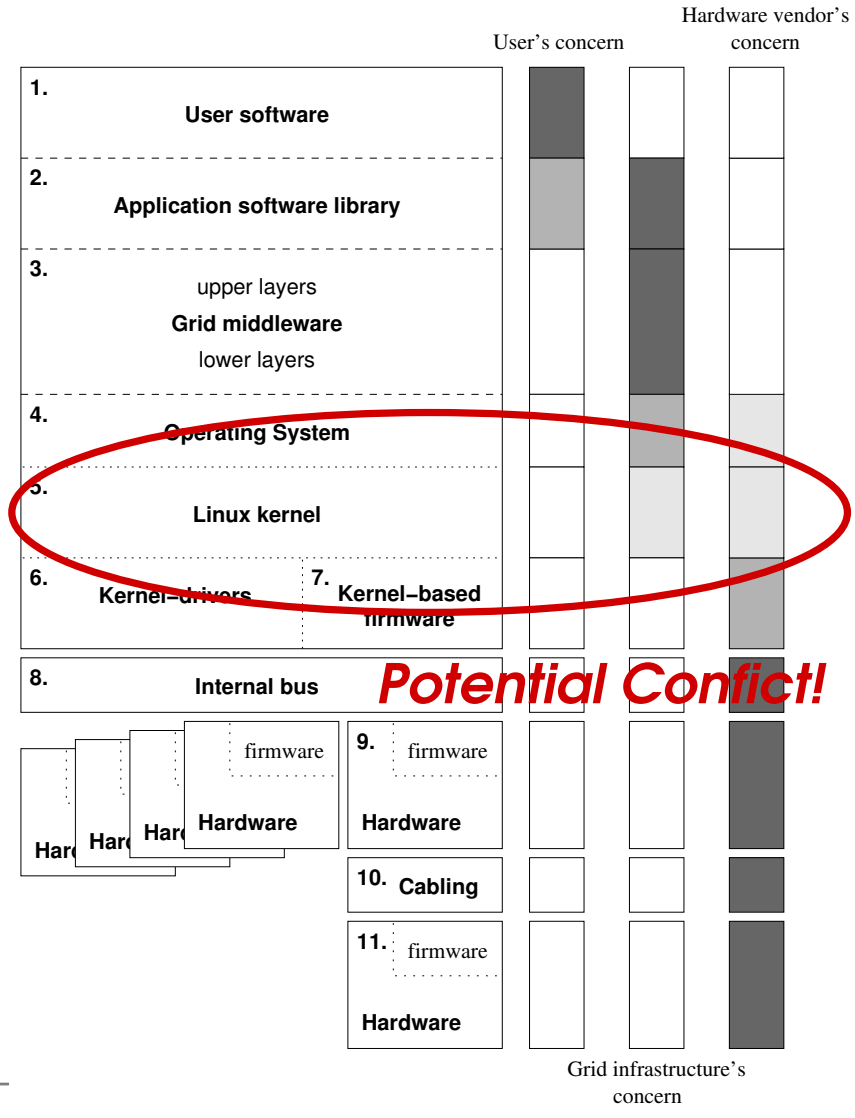
Layers in a cluster

Divisions of responsibility between user, system administrator and vendor



Layers in a cluster

Divisions of responsibility between user, system administrator and vendor



Division of Responsibility

The hardware is covered by a service agreement with the vendor. Based on our recent experience we suggest the following division of responsibility that the cluster is functioning correctly:

Division of Responsibility

The hardware is covered by a service agreement with the vendor. Based on our recent experience we suggest the following division of responsibility that the cluster is functioning correctly:

- Hardware (CPU, memory, cabling, . . .) is the responsibility of the vendor.

Division of Responsibility

The hardware is covered by a service agreement with the vendor. Based on our recent experience we suggest the following division of responsibility that the cluster is functioning correctly:

- Hardware (CPU, memory, cabling, . . .) is the responsibility of the vendor.
- The local system administrator is responsible for installation of Kernel based firmware and firmware drivers, but it is the vendor's responsibility to provide working versions.

Division of Responsibility

The hardware is covered by a service agreement with the vendor. Based on our recent experience we suggest the following division of responsibility that the cluster is functioning correctly:

- Hardware (CPU, memory, cabling, . . .) is the responsibility of the vendor.
- The local system administrator is responsible for installation of Kernel based firmware and firmware drivers, but it is the vendor's responsibility to provide working versions.
- The Linux Kernel and operating system are of concern to both people providing the Grid infrastructure, and the hardware vendor. The system administrator must mediate between constraints.

Division of Responsibility

The hardware is covered by a service agreement with the vendor. Based on our recent experience we suggest the following division of responsibility that the cluster is functioning correctly:

- Hardware (CPU, memory, cabling, . . .) is the responsibility of the vendor.
- The local system administrator is responsible for installation of Kernel based firmware and firmware drivers, but it is the vendor's responsibility to provide working versions.
- The Linux Kernel and operating system are of concern to both people providing the Grid infrastructure, and the hardware vendor. The system administrator must mediate between constraints.
- It is the Grid infrastructure people's responsibility to provide working code, whereas the system administrator's remit covers correct installation of this software.

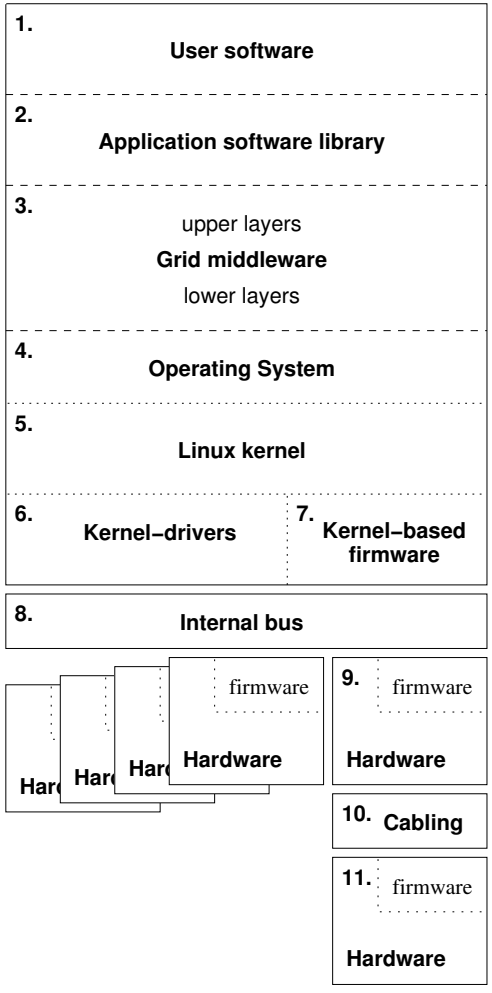
Division of Responsibility

The hardware is covered by a service agreement with the vendor. Based on our recent experience we suggest the following division of responsibility that the cluster is functioning correctly:

- Hardware (CPU, memory, cabling, . . .) is the responsibility of the vendor.
- The local system administrator is responsible for installation of Kernel based firmware and firmware drivers, but it is the vendor's responsibility to provide working versions.
- The Linux Kernel and operating system are of concern to both people providing the Grid infrastructure, and the hardware vendor. The system administrator must mediate between constraints.
- It is the Grid infrastructure people's responsibility to provide working code, whereas the system administrator's remit covers correct installation of this software.
- Users' code is their own responsibility.

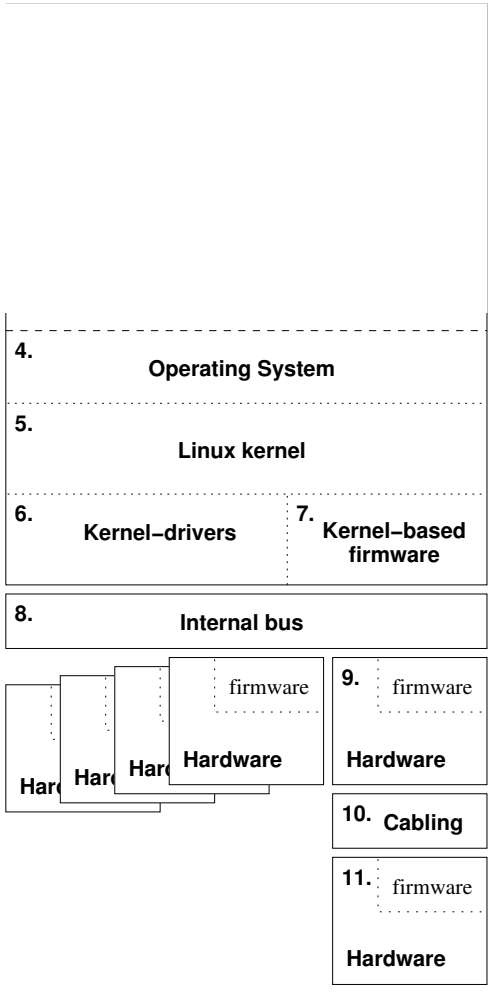
Automated Approach

To facilitate rapid diagnosis of hardware (without large amounts of system administrator's time) the following is proposed:



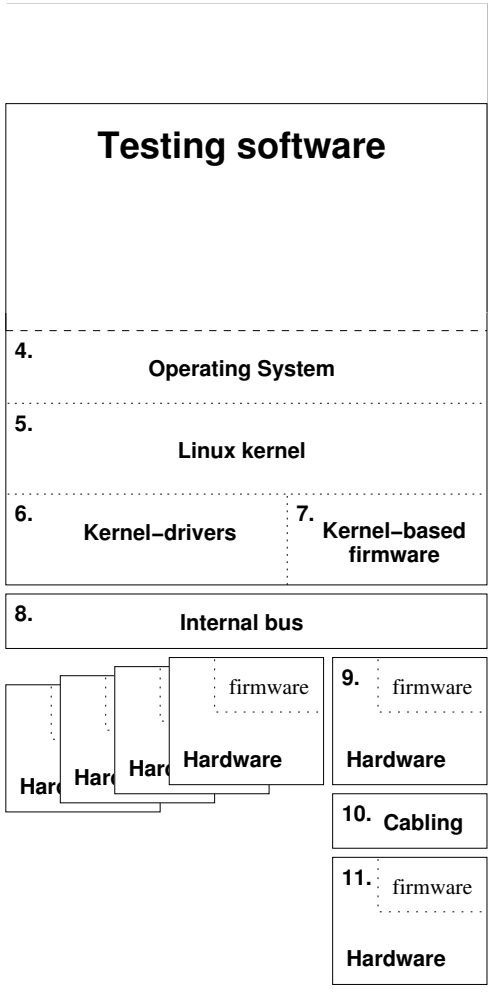
Automated Approach

To facilitate rapid diagnosis of hardware (without large amounts of system administrator's time) the following is proposed:



Automated Approach

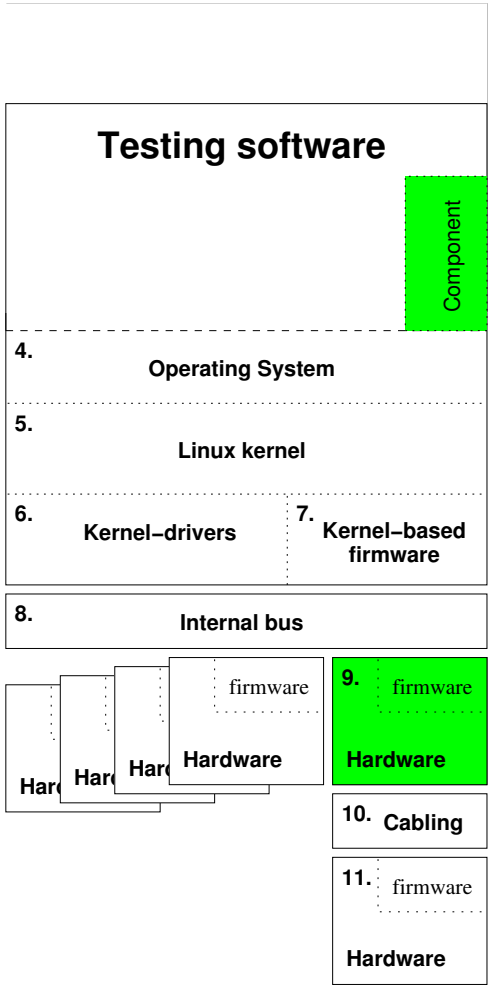
To facilitate rapid diagnosis of hardware (without large amounts of system administrator's time) the following is proposed:



- A test environment is developed to test the hardware.

Automated Approach

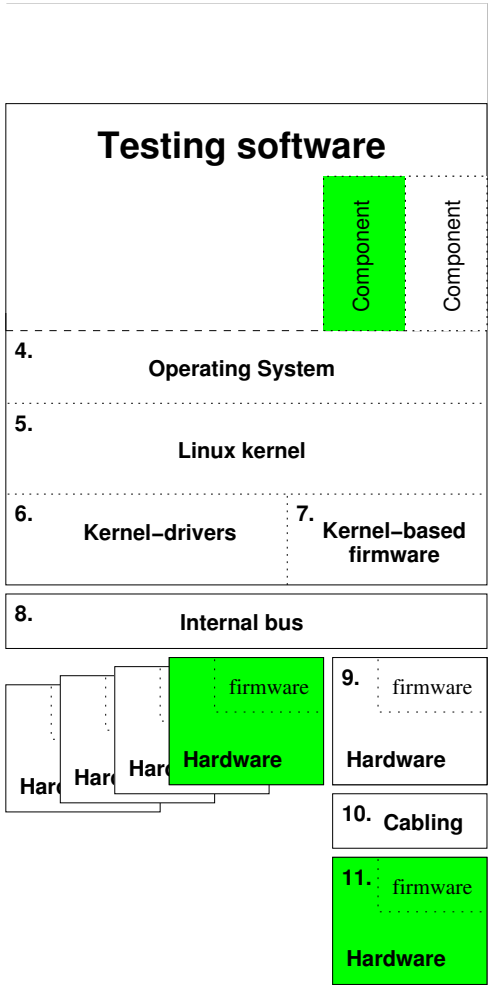
To facilitate rapid diagnosis of hardware (without large amounts of system administrator's time) the following is proposed:



- A test environment is developed to test the hardware.
- Environment has components that test some feature of the hardware.

Automated Approach

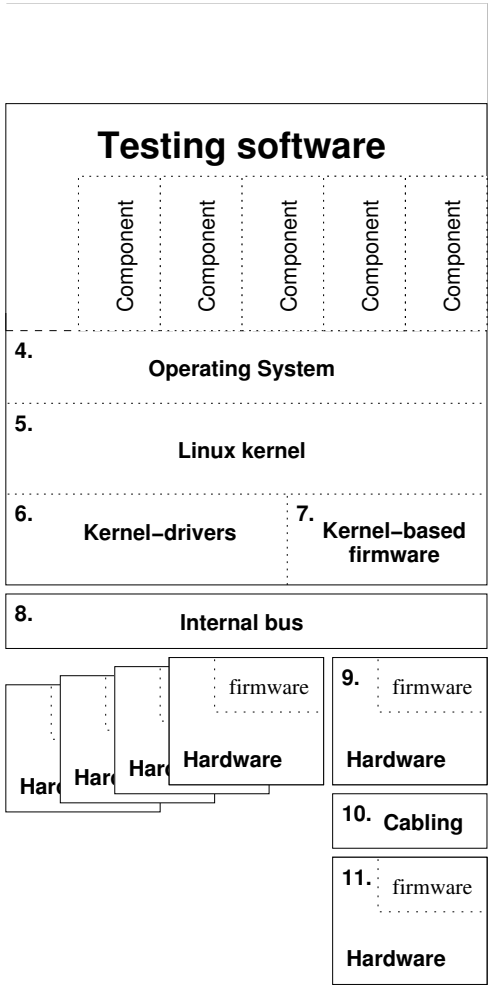
To facilitate rapid diagnosis of hardware (without large amounts of system administrator's time) the following is proposed:



- A test environment is developed to test the hardware.
- Environment has components that test some feature of the hardware.
- The order of testing is such that hardware a test assumes to be working correct is first tested.

Automated Approach

To facilitate rapid diagnosis of hardware (without large amounts of system administrator's time) the following is proposed:



- A test environment is developed to test the hardware.
- Environment has components that test some feature of the hardware.
- The order of testing is such that hardware a test assumes to be working correct is first tested.
- The test run autonomously, so no system administrator intervention is required.
- Passing all the tests is a necessary condition for the hardware to be "working".

Automated Approach

- Assume little about operating environment (DHCP, PXE, NFS?)

Automated Approach

- Assume little about operating environment (DHCP, PXE, NFS?)
- Minimum overhead and interaction with existing systems, install anywhere easily.

Automated Approach

- Assume little about operating environment (DHCP, PXE, NFS?)
- Minimum overhead and interaction with existing systems, install anywhere easily.
- Assume nothing about test-machine, need a remote service to collect results.

Automated Approach

- Assume little about operating environment (DHCP, PXE, NFS?)
- Minimum overhead and interaction with existing systems, install anywhere easily.
- Assume nothing about test-machine, need a remote service to collect results.
- A near-sufficient set of tests already exists:
  - memtest86+
  - badblocks
  - smartmontools
  - lm_sensors
  - cpuburn
  - ttcp/nttcp
  - . . .

Automated Approach

- Assume little about operating environment (DHCP, PXE, NFS?)
- Minimum overhead and interaction with existing systems, install anywhere easily.
- Assume nothing about test-machine, need a remote service to collect results.
- A near-sufficient set of tests already exists:
  - memtest86+
  - badblocks
  - smartmontools
  - lm_sensors
  - cpuburn
  - ttcp/nttcp
  - . . .
- Additional and more sophisticated tests can be developed later.

Tier 2 Example - ScotGRID

ScotGRID is a Tier 2 site providing GRID resources mainly to Particle Physics and BioInformatics:

- 59 dual 1GHz Pentium III worker nodes with 2GB memory.
- 28 dual 2.4GHz Xeon with 1.5G memory.
- over 6TB of disk storage.
- Dedicated Systems Administrator (David Martin).
- Compute nodes and disk from single vendor.
- Unscheduled downtimes are both rare and brief.

Summary

UNIVERSITY
of
GLASGOW

Summary

- Dedicated System Administrator, minimum 0.5 FTE, would be required to run the Glasgow cluster effectively.

Summary

- Dedicated System Administrator, minimum 0.5 FTE, would be required to run the Glasgow cluster effectively.
- Suite of diagnostic tools is useful. Automated tools would be better.

Summary

- Dedicated System Administrator, minimum 0.5 FTE, would be required to run the Glasgow cluster effectively.
- Suite of diagnostic tools is useful. Automated tools would be better.
- Run tests fully when new hardware arrives.

Summary

- Dedicated System Administrator, minimum 0.5 FTE, would be required to run the Glasgow cluster effectively.
- Suite of diagnostic tools is useful. Automated tools would be better.
- Run tests fully when new hardware arrives.
- Rerun tests should any problem arise.

Summary

- Dedicated System Administrator, minimum 0.5 FTE, would be required to run the Glasgow cluster effectively.
- Suite of diagnostic tools is useful. Automated tools would be better.
- Run tests fully when new hardware arrives.
- Rerun tests should any problem arise.
- Testing of hardware and RAIDs in particular is important. Different institutions have developed their own in-house solutions. However there isn't one well-known, strongly maintained software package.

Conclusion

Our experience shows that it is very difficult to maintain a small cluster without a dedicated systems administrator. Service contracts only go so far, especially when a vendor less familiar with a Linux operating system. Grid clusters combining hardware purchased from several budgets, but bought, tested and maintained by a dedicated systems administrator gives rise to a far more efficient and reliable system.