A Hardware Based Cluster Control and Management System

R. Panse, V. Lindenstruth, T. Alt, H. Tilsner, L. Hess,
Kirchhoff Institute of Physics, Im Neuenheimer Feld 227, D-69120 Heidelberg, Germany
(panse@kip.uni-heidelberg.de)

Abstract

Super computers will be replaced more and more by PC cluster systems. Also, future LHC experiments will use large PC cluster. These clusters will consist of off-the-shelf PCs, which in general are not built to run in a PC farm. Configuring, monitoring, and controlling such clusters requires a serious amount of time consuming and administrative effort. We propose a cheap and easy hardware solution for this issue. The main item of our cluster control system is the Cluster Interface Agent card (CIA). The CIA card is a low-cost PCI expansion card equipped with a network interface. With the aid of the CIA card the computer can be fully controlled remotely, independent of the state of the node itself. The card combines a number of features needed for the remote control, including power management and reset. The card operates entirely independent of the PC and can remain powered while the PC may even be powered down. It offers a wide range of automatization features, including automatic installation of the operating system, changing BIOS settings or booting a rescue disk and as well as monitoring and debugging of the node. With the aid of PCI scans and hardware tests errors and pending failures can be easily detected in an early stage. Working prototypes of the card already exist. The paper will outline the status of the project and first implementation results of the preproduction devices, currently being built.

INTRODUCTION

Massive computer farms are used to solve complex tasks, which can not be dealt with quickly by a single computer. Future collider detectors will produce huge amounts of data, which has to be processed in a short period of time. For the collider experiments like ALICE[1] or CMS[2] at the LHC at CERN hundreds of PC will be used. These PC have to be installed and configured to run in a cluster network. However, this is a difficult administration task. The task will be even more difficult if one uses heterogeneous cluster nodes. The disk image of the nodes cannot be used for all computers in the cluster. Therefore, individual disk images have to be built. Another challenging task for the cluster administration is the monitoring of the cluster. Every computer is prone to errors. There are many sources of errors such as hardware errors and software failures. It is important that a faulty computer node does not stop the cluster while it is running. Pending errors have to be detected to prevent data loss or to generally prevent the cluster being disturbed whilst running. Especially the processing of data for a collider experiment is very expensive, because a collider experiment is very cost-intensive. Because of those high costs, the cluster should always run smoothly. However, sometimes a cluster node will fail and thus it has to be debugged and repaired. Finding sources of error can take a long time. Faulty hardware can fail to work for an unpredictably long time. Test programs can help to check selected hardware such as the memory or hard disk.

Cluster Control

There are several tools to handle and manage computer clusters. One of interest is the remote control of the computer nodes. In the last years there were a couple of developments with regard to remotely control computer systems. Using Wake On LAN (WOL) for example, one is able to remotely start up the computer. Disk images can help to install an operating system over a network to a fixed computer system. However, the disk image is built for a particular computer environment. That is the reason why heterogeneous computer workstations have to be individually configured. Sometimes one want to change the BIOS settings, which are not remotely accessible by default. Some motherboards support a BIOS serial console. Thereby you can access the BIOS via a serial interface (COM-Port) of the main board. This serial connection can be used to plug in a remote-controlled device, which is used to control the BIOS remotely. Nevertheless, most operating systems support remote access. VNC[4], Terminal server and SSH are one of the famous remote access applications. These applications facilitate working with the operating system via the network interface. Sometimes it will be necessary to initiate a hardware reset of a computer, because the operating system crashed. Remote access is not possible anymore by a failed system, the machine has to be rebooted manually. There are remote switchable power sockets to solve this problem. But most times, several computers are supplied by the same power socket. Remote administration of a file or a application server is more and more demanding. A server is normally much more expensive as an ordinary computer. Therefore, more expensive hardware based remote control systems justify the higher cost. Remote management boards for example, permit remote control of computer servers, but are very expensive.

CLUSTER INTERFACE AGENT

Software based computer control systems depend on the operating system and do not work if the system crashed.

Hardware based control systems are mostly particularly for one purpose or very expensive. Therefore we have developed the Cluster Interface Agent (CIA) card. With the aid of the CIA card one is able to remotely configure and control a single computer or a whole cluster. The CIA card is planned to be used for the HLT[3] cluster of the ALICE experiment. The low profile PCI form factor of the card permits using it in a two inch server rack. The card is equipped with a 10/100 Mbit/s network interface and can act completely independent from the cluster node. No additional software is needed at the cluster node. There are two main task of the CIA card. First, the card returns information about the host computer without disturbing the work of the host. For example, it can give information about the CPU state, hardware scans or can read the content of the memory. With the aid of this information, the computer in which the CIA card is plugged in, can be monitored to detect failures or can be configured by suitable programs. On the other hand, the CIA card can actively control the computer. The card emulates a couple of IO-devices such as mouse, keyboard or floppy drive. The interfaces of the CIA card with the host computer will be the PCI and the USB bus. Being powered by the standby power of the mainboard or by an external power supply permits the CIA card to be work even if the computer is switched off.

CIA Network

Every node on the cluster will be equipped by a CIA card. As shown in figure 1 the card can be part of a second network. This network is completely independent from the cluster network used to connect the nodes. User application running on the nodes are neither disturbed nor is their network bandwidth being reduced . The administrator has full hardware control of the nodes via the network of the CIA cards. He can install new operating systems, change BIOS settings or monitor the nodes. Another possibility are remote activated power cycles of the computers.

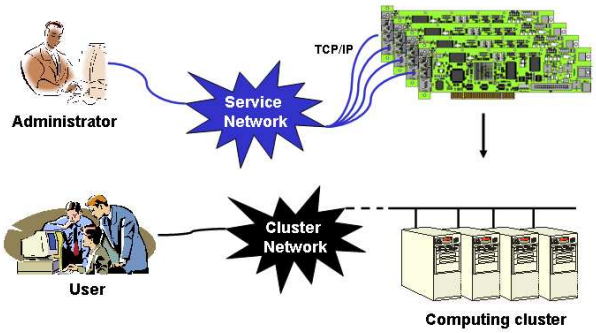

Figure 1: CIA Service Network.

Embedded System

The CIA card is controlled by an integrated embedded system. Its heart is an Altera Excalibur[5] device, which is a FPGA with an embedded hardcore processor. The processor is an industry-standard ARM922T processor. It has a system performance of up to 200 MHz (210 Dhrystone MIPS) . The main memory of the system is a 32 MB SDRAM. An 8 MB flash memory stores the root file system and the kernel of the Linux running on the embedded system, as well as the configuration data of the FPGA. The Linux is remote accessible with the network interface of the CIA card. The embedded system controls all components of the card. PCI bus cycles, USB requests or power cycles of the node is managed by special software running on the card. Furthermore, the FPGA can be used to easily integrate additional hardware in order to adapt the CIA card to different hardware.

Data Control

The floppy interface of the CIA card is directly connected to the main board of the node. Thereby the card emulates a floppy drive. The data which is accessible via the emulated drive is provided by the embedded system. New disk images can easily be distributed to the node.In addition the CIA card is equipped with an USB interface. The USB interface of the card will be connected to the USB interface of the computer. The card register itself as an USB mass storage device like a USB stick. Thus boot devices can be emulated by the card via USB or via the floppy interface.

Video Control

The CIA card functions as a video card and is registered as a default VGA card by the PCI bus. No video cards other than the CIA card will be required for the computer system. The CIA card processes the video data and the data will be made available via a VNC Server running in the embedded system. Using a VNC viewer, the administrator can work on the node as sitting in front of it. The keyboard and mouse interaction sent from the VNC viewer will be translated to PCI accesses to the keyboard and mouse controller of the main board . The CIA card is a PCI master and can initiate all kind of PCI cycles.

Power and Reset Switch

To trigger a hardware based reset or power cycle, the CIA card can be connected to the power and reset switch of the computer. The CIA card therefore makes remote power cycles possible.

Monitoring

Every computer is error-prone. The cluster has to be protected from pending failures of computer nodes. In order to that the nodes have to be monitored online. Even though software failures are very unpredictable, hardware failures can be detected before data is lost or before the whole running cluster has to be stopped. Very high temperature of an entity such as a hard disk or a power supply is an indication of a pending failure. High voltage can damage

the main board and can be an indication of a faulty component of the board. With the help of the CIA card, one is able to inspect and monitor these parameter of the computer. An integrated analogue digital converter (ADC) is used to measure voltage, temperature or acoustics such as hard disk noise to detect damaged components. In addition, the CIA card can inspect faulty entities via the PCI bus. Additionally, PCI scans support finding hardware to separate the location of the failure. If a device does not respond to the PCI request, the CIA card can detect it and inform the administrator.

Debugging

Sometimes a computer node fails so that this node has to be made available to the cluster system as soon as possible. However, debbuging and finding the source of error can be quite time consuming. Because of the PCI master functionality of the CIA card, it can help to find hardware based failures. The CIA card has full memory access via the PCI bus. Memory mapped devices can be accessed and tested. Starting memory, hard disk or CPU tests are possible by emulating disk images of test programs or by directly access via PCI cycles. A further source of debug information are the POST codes which are issued by the BIOS of the monitored computer. These codes are available on an dedicated IO port of the processor and are captured by the CIA card. If a test failed, the CIA card will record and inform the administrator.

CIA CARD PROTOTYPE

Figure 2 shows one of the first prototypes of the CIA board. It is realized as a PCI expansion card with low profile form factor which allows its usage even in 2HE server racks.

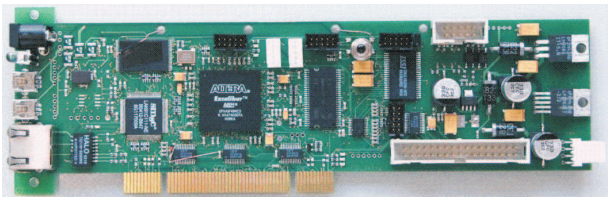

Figure 2: CIA card prototype

Following features are implemented:

- PCI scan (detecting hardware)
- Read out BIOS DMI information
- Video card functionality
- Web based GUI
- Remote control via a VNC viewer (text mode[1])
- Hardware based remote power cycle and resets
- Full PCI access (memory and memory mapped devices)
- Mock-up USB device

Following features are going to be implemented:

- Temperature monitoring
- Fans monitoring
- Voltage monitoring
- Infrared access to CIA card
- Bluetooth access to CIA card
- LCD Display
- Microphone for detecting unusual sounds from fans or disks

Figure 3 shows the CIA card control applications. The card can be commanded by a web browser. PCI scans can be initiated, which provide a list of the devices plugged in the PCI bus. Also, receiving POST codes and accessing the main memory is facilitated by the web interface. The picture in the middle of figure 3 shows a VNC viewer, which displays the BIOS settings. The picture to the right in figure 3 is a JAVA application, which possesses the same features such as the web interface, but a couple of CIA cards can also be controlled at the same time. Every control application can be done by any computer that is connected to the CIA network.

Figure 3: CIA Control Application

CONCLUSION

The CIA card is a PCI expansion card developed to have full remote access to a cluster node. It allows the remote installation of operating systems and the manipulation of BIOS settings. Controlling the power and reset switches of the computer is one of his feature. The CIA card is completely independent from the computer and can work even if the computer is switched off, as the card is supplied by the standby power of the main board. With the aid of the PCI cycles, the card can monitor and scan the computer. Temperature, voltage and acoustics can be measured and monitored by the card too.

REFERENCES

[1] ALICE Experiment, see for example:, http://alice.web.cern.ch/Alice/AliceNew/

[1] Video cards have two modes, the text mode for console applications and video mode used for GUI based applications

[2] CMS Experiment, see for example:, http://cmsdoc.cern.ch/cms/outreach/html/index.shtml

[3] The ALICE Collaboration, “ALICE - Technical Design Report of the Trigger, Data Acquisition, High-Level Trigger, and Control System”,
CERN/LHCC/2003-062, January 2004.

[4] Real VNC website, http://www.realvnc.com

[5] Altera website, http://www.altera.com