

LINUX cluster with CompactPCI as an example of an Internet server

The basic principles of (Linux) clustering are explained using an Internet server as an example and configuration aids for adapted CompactPCI systems are given. This example can also be transferred to other fields in communications and automation.

INTRODUCTION

The demand is growing for system solutions with maximum availability. This is also partly due to the rapid growth of Internet and communications technologies.

With the support of new hardware and software technologies a core system which can be easily expanded for various applications can be formed from reliable CompactPCI hardware and easily accessible operating system technology with expanded cluster management.

According to the latest findings Internet service providers must double their performance capability about every 100 days. In order that the space requirements do not also increase in proportion, the reduction in system size will in the future become an important factor in a provider's performance capability and competitiveness.

The basic principles of (Linux) clustering are explained using an Internet server as an example and configuration aids for adapted CompactPCI systems are given.

This example can also be transferred to other fields in communications and automation.

HIGH AVAILABILITY AND CLUSTERING

High-availability system solutions are employed both for critical company applications and in the industrial field. In particular the increasing complexity and the substantial growth in the requirement for telecommunications services demand high availability systems as an economic necessity.

High availability depends on many factors - including hardware and software as well as the operation of the system and also the environment in which a system is to be used.

To assess hardware and software systems with the objective of obtaining continuous availability, generally different availability levels are used which include the requirements on a system or the effects of the failure of any component.

Looking at the field of telecommunications, often systems are required with an availability of 99.999% (Assured Availability, Level 4), whereas for other fields an availability of 99.9% (Basic Fail-Over, Level 2) may be adequate (Source: IDC, Availability Spectrum).

Here should be noted that the availability of a com-

plete system is involved and not that of the individual components.

System availability	Failure in minutes/year
99.9%	525.6
99.999%	5.256
99.9999%	0.5256

Table 1.

The main solution methods for obtaining continuous availability are described in the following.

In the normal case all the methods of solution assume redundancy of important hardware components at the system level, buffering against power variations and failure using uninterruptible power supplies, server and application security against detrimental user intervention and, last but not least, properly functioning data back-up.

The method of combining a number of disk drives to form RAID arrays (Redundant Array of Inexpensive Disks) is very popular. Here depending on the operating mode, the data of one disk is mirrored to another or the cross-check sum of the data of one disk is stored on another one. In both cases with the failure of a disk, the original information can be restored by its replacement at run time. It is important that operation can continue despite the replacement of a disk.

Data mirroring with fail-over

In its basic form this concept provides a back-up server for an active, primary server (hot stand-by). The back-up mirrors the data resources of the primary server via a separate server network and it can restore resources for user access after failure of the primary.

The mirroring of data often takes place at the file level, which can lead to long delays and, in some cases, also to the loss of data.

The users address the resources - irrespective of the actual active server platform - as "virtual servers" via the network.

Fault-tolerant server systems and high-availability clusters

Fault-tolerant and fail-safe systems are, as previously, the domain of specialists (e.g. the Himalaya servers

INFRASTRUCTURE

from Compaq/Tandem). However, "ordinary" manufacturers also offer high-availability solutions as expansions to standard operating systems.

Basically, it can be said that redundancy attempts to prevent single points of failure, (SPOF), developing in the complete system - a method that has been commonly used in aerospace for decades.

A common feature of all systems which work with the doubling up of components and hot stand-by is that they leave half of the available resources unused. The second system which is waiting in hot stand-by for the failure of the first one remains unused when the primary is operating.

Other methods which are not so wasteful with the available resources - redundancy is expensive! - link fail-safety with load distribution. The principal idea behind architectures which can be combined under the term Cluster is based on the even distribution of the load between the participating systems.

The even distribution of the tasks does not just avoid unused redundancy, but also increases the overall throughput. The failure of a subcomponent is only the extreme case of the continual load balancing which occurs amongst the active systems.

Clusters: Principles and applications

What is a cluster?

A cluster is a combination of nodes which are all used for the solution of a task. The nodes might be individual PCs, but is also possible that computers with many hundreds or thousands of processors may form a cluster.

Communication occurs via specialised connections. These may be direct links between the individual processors or a high speed network such as the SCSI interface - or simply even a Fast Ethernet link.

Methods of using a cluster:

There are various fields of application which are suitable for clusters and which necessitate an appropriate architecture of the cluster (or corresponding combinations):

- Clusters for obtaining the highest possible computing power (e.g. weather forecasting). With this type of cluster the main feature is the highest possible computing power of the individual processors and very fast communications between the individual nodes.
- Clusters for achieving the highest possible fail-safety (controllers in atomic power stations, telecommunications applications - key words: Fail-safe clusters, high availability and fault tolerance, high-availability clusters). With the failure of one or more computers in the conglomeration the functions of the cluster must continue to be maintained. The full performance capability can no longer be obtained, but the applications continue to run even on the partial system. To obtain the fail-safety level individual nodes can be switched in as substitute systems or the faulty parts are just switched off. The server nodes combined to form a cluster generally have direct access to common hard disk systems, e.g. RAID subsystems which are connected to the participating servers, for example, via a common SCSI bus. Only one server in each system has exclusive rights of access to the appropriate resources ("shared disk"). Server resources may move from cluster node to cluster node during the failure of a server or also for static, manual load distribution. Here, HA clusters achieve a substantially finer fail-over granularity as solutions with network-based data mirroring which exclusively supports the fail-over of complete systems.
- Clusters for load distribution (load balancing) for load critical applications (data bases, web servers). Various applications require performance data which cannot be effectively attained by single computers. In this case an application is shared by a number of systems which appear externally as one computer. It is then possible, for example, to operate a web server which processes many millions of accesses per hour. Here, the individual requests are distributed over the various computers in the order in which they arrive. This application particularly involves high performance input and output units

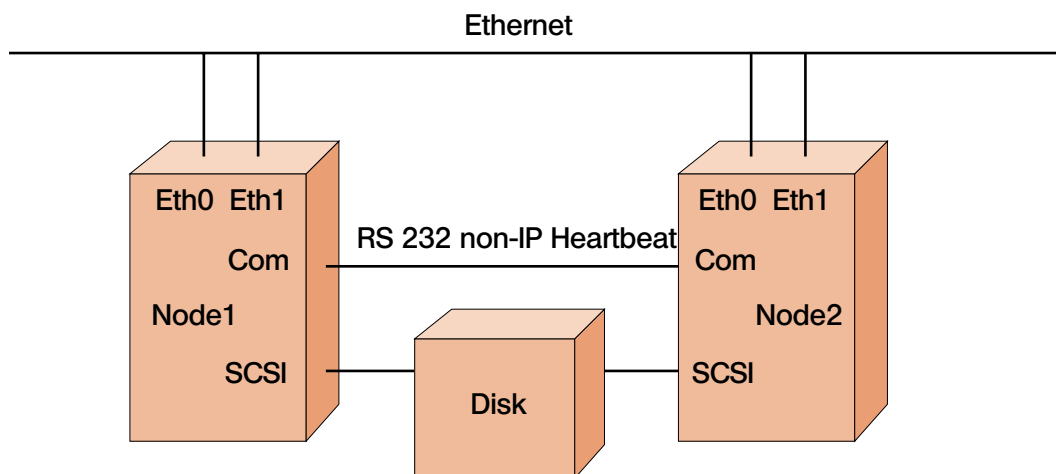


Figure 1. Configuration example of a cluster.

(hard disks, networks).

CLUSTERING WITH LINUX

Brief overview of Linux

As with any other Unix, Linux is a highly stable, but (still) not very user-friendly operating system which is preferentially employed on servers (but also increasingly on standard workstations) and which already enjoys a very high level of acceptance.

Linux is structured similarly to other Unix derivatives. The main difference and the real advantage is the freely available Linux source code.

In 1998 Linux had a share of 17.2 % of the total market (world-wide, referred to deliveries, totalling 4.3 million servers) of server operating systems which cannot be ignored when it is taken into account that in 1992 there were approx. 1,000 users and in 1991 just six users.

Currently, Linux is the server operating system with the greatest growth rate. By 2003 market researchers expect an average annual total growth rate of 25%.

Currently, there are more than 10 million users of Linux throughout the world.

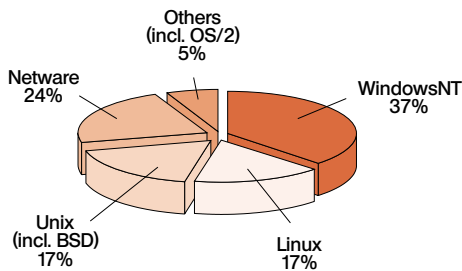


Figure 2. Server Market 1998.

Source: International Data Corp. (IDC, Framingham, Mass).

Clustering and high availability under Linux

The subjects of clustering and high availability are being increasingly addressed by the Linux community and appropriate activities are being undertaken. Details about this can be found, for example, as a White Paper in the Linux High Availability HOWTO (<http://metalab.unc.edu/pub/Linux/ALPHA/linux-ha/High-Availability-HOWTO.html>).

Apart from concepts for a high availability Linux solution, associated subject fields such as mass storage technologies (e.g. with SCSI), file systems and redundant networks are dealt with. In particular the subject of file systems is a critical point when a fault occurs and another cluster node takes over (fail-over). In addition, references are given for commercially available solutions.

Various high availability software solutions based on Linux are now being offered by a number of different companies.

- The system from Technaut, for example, monitors

operating systems and applications and is quoted as guaranteeing an availability of 99.99%. When a fault occurs, a second system takes over within a few seconds. (<http://www.technauts.com>)

- Wizard Software offers a "light" version of its cluster software Watchdog for Linux, which is particularly intended for services such as web or mail servers. (<http://www.wizard.de>)
- Amongst their products, TurboLinux with their TurboCluster Server offers the possibility of setting up a high availability cluster with automatic load distribution using a number of Intel (or also Alpha) computers under Linux. The TurboCluster Server combines the Linux operating system with high availability software and configuration tools. However, TurboLinux had to make changes to the Linux core for its cluster solution and is therefore adding to the fragmentation of the "Linux Standard". In this field TurboLinux co-operates with various industrial partners, e.g. Compaq, Linuxcare or the Cubix Corp. - the latter will bundle its server from the Density Series with the TurboCluster Server. (<http://www.turbolinux.com>)

This listing does not claim to be complete, but shows a random selection of possible software solutions.

LINUX CLUSTER SOLUTION ON COMPACTPCI

CompactPCI

CompactPCI brings together the advantages of PCI bus technology with those of a rugged 19" system which can also be employed in an industrial environment (extended temperature range, EMC). Boards (CPUs etc.) in a single or double height Eurocard format are used together with a passive backplane. The PICMG (<http://www.picmg.com>) specifications form the basis of the CompactPCI standard. In these specifications expansions such as hot swap for simple and fast board replacement and H.110 for the telecommunications field are taken into account.

In particular, hot swap technology, which is the method of replacing or supplementing boards in running operation, in combination with rear I/O, i.e. with cabling on the system backplane, is an important feature for forming high availability systems. Hot swap technology demands, of course, appropriate support from the software and the operating system.

The CompactPCI system platform enables the construction of rugged systems. The MTBF (Mean Time Between Failures) and MTTR (Mean Time To Repair) values are significantly longer than with normal PCs due to the system and board design and due to the application of selected components.

Very efficient high availability systems can be formed on the smallest space due to the compactness of the CompactPCI system construction.

Linux cluster on CompactPCI

The high availability of computer systems plays a dominating role in the rapidly growing Internet community.

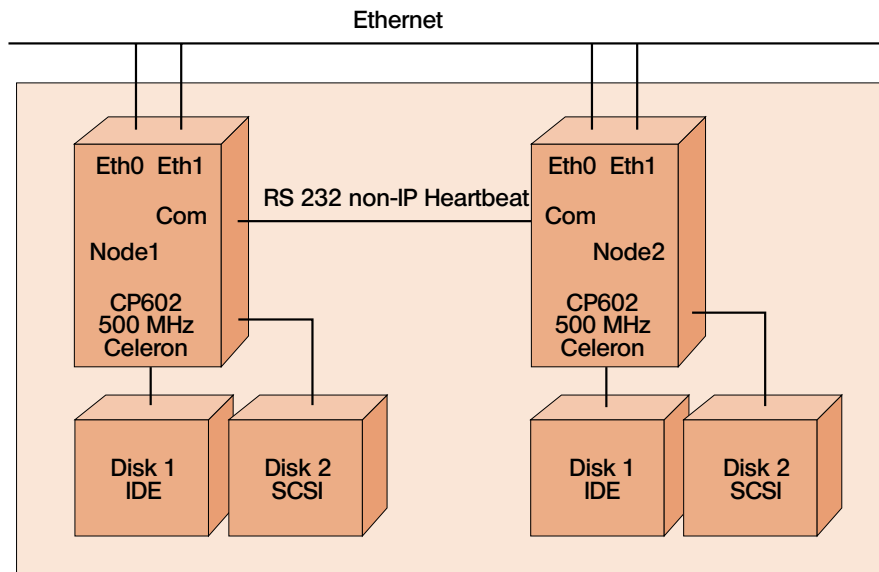


Figure 3. Configuration example of a cluster server using CompactPCI.

With the consistent 19" design of CompactPCI these PC systems match the control cabinet technology of modern Internet service providers and enable simple upgrading to satisfy the fast growing demands in this field.

PEP is therefore now offering Linux cluster support for its CompactPCI systems on the basis of TurboLinux. Using this cluster software many CPU units can be combined to form a virtual system which, when a CPU unit fails, distributes its function over the remaining CPU units. This is especially important when designing redundancy and load balancing for web servers.

The application of cluster software under Linux provides a simple method of setting up redundant systems using a flexible design to accommodate later improvements in performance. Therefore, clustering enables the setting up of high availability systems with less complex hardware technology than was previously the case. In addition, new systems and components can be expanded in a very user-friendly manner using

CompactPCI technology.

Apache Web Server

Web servers for intranet or Internet and smaller network servers with a permanently defined area of operation currently represent the fields application giving Linux the quickest path to enterprise environments.

The popular open-source web server, Apache, which has been running on Unix and Linux for a long time now, has the largest share of this success. The Apache web server is currently the most popular web server above all others. (<http://www.apache.org>)

Taking a look at the operating systems used on web servers, Linux is leading with 28.5%, followed by Windows 95/98 and NT with 24.4%. After that follows Solaris/SunOS with 17.7% and then the BSD Unix derivative with 15%. Following these, only Irex with 5.3% achieves a significant proportion - Mac, AIX, HP/UX, Sinix, Netware and the Unix derivatives from Digital and SCO are placed in falling ranking between

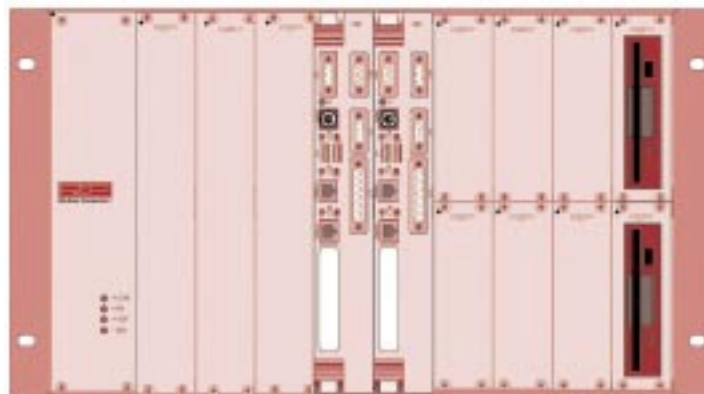


Figure 4. Linux TurboCluster server on CompactPCI in 19" double-height rack.

1.6 and 0.4%.

(Source: Evaluation of a total of 1,465,124 web servers up to April 1999 by the Internet Operation System Counter, <http://www.leb.net/hzo/ioscount/index.html>).

Configuration example

This system shows the capabilities of a redundant web server based on the PEP CompactPCI product range.

The server consists of two independently operating computer systems connected together using Fast Ethernet. If required, the computers may depend on common resources, such as for example, the voltage supply. However, many completely separate computers with voltage supply, drives, etc. can also be physically set up on one 19" rack.

The TurboCluster Server 1.2 Beta1 (June 12, 1999) from TurboLinux with an Apache web server runs on both CPU boards. (Further information on the CPU boards used, CP602, together with the other components can be found under <http://www.pep.com>).

The web server consists of two separate network computers which together form a common domain. Accesses to the Internet pages managed by them are shared between the two computers depending on the

loading. If a computer fails due to a hardware or software fault, the second system takes over the tasks of the previous computer within a few seconds without the user becoming aware of it.

PROSPECTS

Since the topics of clustering and high availability are currently being actively propagated by the Linux community and through various distribution channels, it is necessary to follow the on-going developments in this field.

The activities of the PICMG (Redundant System Slot Specification) are also of particular interest in the setting up of high availability CompactPCI systems.

PEP will in future offer support for Linux clusters on its CompactPCI system platform for the Internet and communications market. Servers and clusters based on CompactPCI can also be transferred to other fields such as automation engineering ■

Mr. Harald Müller is Project Manager Communication.

AD POLYHEDRA