Patagonia - A Dual Use Cluster of PCs for Computation and Education

Felix Rauch   Christian Kurmann   Blanca Maria Müller-Lagunez   Thomas M. Stricker

Laboratory for Computer Systems

ETH - Swiss Institute of Technology

CH-8092 Zürich, Switzerland

{rauch,kurmann,bmc,tomstr}@inf.ethz.ch

Abstract

In this paper we present the installation and the configuration of Patagonia, a novel "Cluster of PCs" that can be used alternatively as a compute farm for computation or as a classroom for education. The Patagonia cluster is built from off-the-shelf PC hardware but is equipped with a high-speed network between the computers to enable parallel and distributed computation. Based on the availability of fast disks and the high-speed network we propose "partition cloning" as a novel approach to software distribution and operating system maintenance in the cluster. We also discuss the multi-boot setup that allows us to configure and operate different operating systems for research and education with the least possible interference. Our work includes a security setup that protects the operating systems against each other and a network server concept that can provide one and the same individual filespace for users (i.e. home directories) transparently to a UNIX and Windows NT workstation clients. We evaluate the performance of our software distribution by "partition cloning" over the Gigabit Ethernet network and found an important application of high-speed networking technology that goes beyond parallel computing. As an example of a dedicated software utility for clusters we introduce our cluster administration tool, that was written at our site to maintain and reboot the cluster with ease. The Patagonia setup shows that it is possible to find a common denominator among the different requirements of a PC installation for computation and for education or more general business use. A dual use installation of this kind can greatly improve the utilization of fast PC workstations over their entire lifecycle.

1   Introduction

Computer clusters in major research universities are used for research and education. While both kinds of installations can be called "clusters" in English, the installations are far from looking identical. The PCs of an education cluster are workstations that are sparsely distributed across tables in a classroom and do have screens, keyboards and mice, while clusters of PCs for high performance computing are densely packed rack mounted systems kept in a cooled machine room and wired to one central operator console. Despite the completely different look of the two kinds of clusters, the current trends in technology mandate that they are built with nearly the same components, namely commodity PCs leveraging off an ever increasing compute performance found in single chip microprocessors, high volume DRAM memories and low cost disk drives of several GB capacity. Besides the technical specifications, both types of PC clusters also have in common that they have a low utilization compared to traditional mainframes and supercomputers. Typically the education clusters are only used during the day and maybe in the evenings whereas the compute clusters are often idle while the programmers work hard to improve the code of their parallel applications. We admit that in both usage models the management could fill the capacity of the idling processors with embarrassingly parallel computations in cryptography, number theory or combinatorics for research clusters or alternatively bring in PC users from the street that are looking for free text processors or internet access — but such complementary use would hardly improve the cost effectiveness of the installations.

As a consequence we initiated the Patagonia cluster project at ETH Zürich to build a cluster that fits both the needs of education and the needs of research. While education has priority during the day on our cluster, research has priority during the night and with some limitations during vacations. The Patagonia cluster is a collection of high-end PCs in a classroom interconnected by a fully switched Gigabit Ethernet. Dual processors, large memories and fast disks make the machines fit for experiments in parallel and distributed computing while fast graphics cards, 17 inch monitors, mice and keyboards make them suitable for instruction in our advanced CS curriculum.

While our ideas and the corresponding experimen-

tal study clearly originates from a university environment we would like to mention the striking similarities to many modern corporate computing environments using PCs. In those environments the computing needs are quite similar. Most companies rely on a rapidly growing number of compute intensive tasks in data mining, combinatorial optimization and process simulations in addition to the typical personal computing needs of a large number of employees at their desks. A usage model that can cover personal computing and occasional compute-intensive tasks with a single infrastructure would be a great advantage to those businesses. However such a usage model explicitly excludes the traditional transaction processing because it has extremely high requirements regarding reliability and data security that clusters of PCs can not meet at this time.

In our subsequent contribution to the workshop we will describe the requirements and how we managed to find a common system software basis for the *dual use* cluster based on a multi-boot setup of several carefully isolated operating system images. We were successful in installing a system software setup with several fully sandboxed operating systems that are bootable in a networked environment. The setup includes multiple Windows NT installations which contradicts the recommendations of the manufacturer, who is claiming technical unfeasibility. In our Patagonia project we particularly focus on the benefits of the high end facilities provided by the research part of the cluster (mainly the high speed disks and the high speed network) that also help tremendously with the installation and maintenance of the education partitions of the cluster. Although the true cost of such an installation based on university pricing remains very hard to estimate, we are convinced that the combined installation could be cheaper and easier to maintain than the installation of two specialized clusters for scientific computing and for education.

2 The Different Requirements for Education and Research

In the Patagonia cluster, we plan for research users in computer and computational sciences as well as for general educational users with entirely different requirements in terms of system software, application software and account management.

The different usage modes require different administrative setups for the educational and the computational installations.

2.1 Hardware Requirements

Research users of the cluster need a high performance computing platform; therefore fast processors and a large main memory are required. Fast processors are a great plus in education as well, as they offer the flexibility of using some compute intensive programs as demonstration cases in advanced lectures.

Local hard disk drives in the 10 GB range offer enough swap space for virtual memory or enough space for the storage of large data sets in compute intensive applications, as well as for the local installation of most of the software packages used in education.

2.2 Networking Requirements

Parallel and distributed programs usually require frequent and fast data transfers among the processing nodes. A high-speed network is therefore important. In the educational case, a fast network eases the administration of the machines and ensures the scalability of server based applications in the cluster during peak hours (e.g. noon time).

The high-speed networks specifically designed for compute clusters need to overcome only small distances within a systems cabinet, since their PCs are rack mounted machines. The driver software can be greatly simplified if the networking technology offers link level flow control, reliable transmission and switches that can switch with the full link speed. Most short distance interconnects can provide these services. Unlike compute clusters the classroom clusters are spread over larger distances, maybe within an entire building. A high-speed networking technology for a dual use cluster must therefore be able to deal with LAN distances. Gigabit Ethernet could make it possible to have the high speed of a Gigabit/s network as well as the flexible cabling of a commodity LAN. A modern *switched* Gigabit Ethernet can provide a total switching capacity for 16–24 Gigabit/s and a fully switched non-blocking backplane for a dual use cluster.

2.3 Operating System

Researchers in the computational and computer sciences need the freedom to try different parameters and improvements to their application code including the OS; therefore an open and flexible operating system with a readily available full source code is required.

Educational users on the other hand might be simple users, so a complex and unhandy operating system or difficult booting procedure is unacceptable. Even for a pure educational cluster the option to boot different operating systems is a big advantage as different lectures might have different needs (e.g. our non-standard operating system Oberon is the programming environment of choice for several courses here at ETHZ).

In an educational environment with a large user base (over 1000 students in our case) it is also important that users can individually configure their accounts and that all customization state is associated with the account and not with the machines.

2.4 Programming Tools and Application Software

In both usage modes the customers require a set of basic, common programming tools and application software, but they will also execute their own programs. It would be highly desirable to offer the choice of either installing application programs on a central server and loading them over a (fast) network at startup-time or to replicate the software on the local harddisk drives of each machine at installation time, which results in better performance.

2.5 Security and Maintenance

Computational users typically work in the same group or in the same project and usually trust each other. They are also fairly competent users who know what they are doing. In this mode of operation, security and foolproofness of the installation is not very important. But once a cluster is also used by students, some of them will inevitably try to stretch its limits of allowed use, some will try to break into installed applications or crack the entire system security. Also some simple minded users might not understand the system enough to realize when they are doing something harmful or cause permanent damage to the installation. An educational or dual-use system must therefore protect itself from unwanted changes of state in the operating system images. The different operating systems need to be properly sandboxed (protected from each other) by maintaining a reasonable level of overall operating system security.

Research clusters in closed rooms do not need any additional physical security besides a lock on the door. Apart from a few selected operators no one has access to the machines to open them or to boot other OSes from external devices. Machines in educational clusters are located in public rooms and need to be secured against theft and the booting of any other OS. The machines must be closed and attached to the tables with locks. A controlled booting environment can be achieved with an advanced BIOS setup.

Maintenance of the computational installation is mostly done by the computational scientists themselves, but the educational environment will have designated professional system administrators. Both groups should be able to maintain their own parts without much interference.

3 Classes of Operating Systems

3.1 Operating Systems with Security and Well Organized System and User State

Most traditional time-sharing operating systems for mainframes and UNIX-like operating systems for PCs offer great flexibility to have a home directory for each user on a central server. All user files and individual configuration settings can be properly contained. Application software can be installed in the normal directory tree of a file system. The server exports these files through a shared file system (e.g. NFS) which is mounted by the clients. It is therefore easy to install new software for use on all machines on the centralized server. There are some conventions on where to place and how to name configuration files and the system specific customizations are cleanly separated from the user customizations. A simple but effective protection mechanism with read, write and execute permissions for user, group and world access enforces the separation of state between user and system settings. Capability based access control can be used to separate login authentication and data access and to secure a server even against untrusted clients (e.g. the virtue and vice principle in AFS [1]).

3.2 Operating Systems with Security and Disorganized State

Other operating systems such as Windows NT are more difficult to handle in clusters since the applications store some user state on the local machine, on the server, in the application directory as well as in the local registry databases on the client. Storing programs on a central server is therefore extremely difficult. Storing the user's files on a central server is possible, but due to the fact that the corresponding registry entries must be loaded from the server at login and written back after each logout, it is not very efficient. The registry, which is a simple version of a database, might offer more consistency and fine grained access control than a file system and could potentially result in better and more consistent management of system and user settings. However the incoherent combination of a registry holding configurations and a file system holding the data causes an unsurpassed mess of user and system state in networked Windows NT installations.

3.3 Operating Systems without Security

Operating systems which were kept simple for educational purposes or which were meant as a run-time system on a personal machine do not offer any security at all. The only protection might involve a harddisk partition, containing a clean copy of the operating system including all system files, that is mounted read-only. Any changeable files including the user data must be stored on a different volume (e.g. a ram disk). Most Windows 95 and MacOS classic cluster installations work with read-only copies and copy to some scratch disks if files need to be changed. Due to lacking protection mechanisms in the CPU (i.e. user and privileged modes)

a programmer with enough skills still has the possibility to directly access the harddisk drive or the network card, but he needs the desire to do so, good programming skills and has to invest a lot of effort and time to get anywhere. Therefore all PC operating systems that must change some state within the OS installation or within application packages for user customizations are generally a bad idea for public access. They were designed to run on personal computers and not on public computers. Unfortunately there are many interesting application programs that are only available for those common operating systems without security.

4   The Hardware of the Patagonia Cluster

The classroom for the cluster offers space for 24 PCs configured as workstations with keyboard and CRT monitors. The Intel based PCs installed are of the type *Dell Precision 410* with 17 inch *EIZO* monitors. Each machine contains a Pentium II processor running at 400 MHz, 128 MB SDRAM memory connected to a 100 MHz front side bus and a built-in *3Com Cyclone* Fast Ethernet network interface card (NIC). Sixteen machines are equipped with an additional second Pentium II processor, a total of 256 MB of main memory and a *Packet Engines GNIC-II* Gigabit Ethernet NIC connected to a central *Cabletron* Gigabit switch by fiberoptic cables. All machines are equipped with a *Seagate Cheetah* Ultra2 SCSI harddisk drive with a capacity of 9 GB.

The high end hardware configuration of the systems results in a few performance characteristics that are crucial to the success of our proposed installation, setup and maintenance procedures for the operating systems. First even the Fast Ethernet backup LAN can transfer an aggregate of about 10 MByte/s between machines; at this time the Gigabit Ethernet LAN can transfer up to 38 MByte/s under Linux and 15–16 MByte/s under Windows NT, but this indicates that the drivers are still unstable and of questionable quality. Second the 400 MHz CPUs can uncompress the data of compressed raw disk partitions at about 9 MByte/s. And third, the Ultra2 SCSI disks can write data streams with almost 20 MByte/s and read them back with 16 MByte/s. The local memory system is capable of copying data with almost 100 MByte/s (full memcpy() with a contiguous read and write stream).

The design goals of a dual use cluster put some unusual requirements on the networking technologies used in the Patagonia cluster. For rack mounted computation-only clusters, *Myrinet* or *SCI (Scalable Coherent Interface)* proved to be the best choice of a Gigabit/s networking technology. Both technologies are fast, reliable and cheap as long as only SAN (System Area Network) cabling is used. Although LAN (Local Area Network) cabling exists for Myrinet and SCI, it is still expensive and its installation remains specific to a particular technology. Permanent cabling of a classroom requires longer distances and therefore we decided to use fully switched Gigabit Ethernet as high-speed network based on universal fiberoptic cabling. Gigabit Ethernet with its cabling flexibility and its great price due to the high volume comes at the cost of reliable transfers and limited flow control options on the links and in the switches. In our CoPs project [12] we plan to develop better driver software to compensate for this deficiency.

The Dell-machines are desktops, which can easily be placed on the tables. Minitowers placed under the table are in the path of the cleaning crews and therefore their cabling is exposed to damage. The PCs have hooks to secure the equipment and support Advanced Power Management. Controlling the power consumption adaptively is important as the room is already heated up by the human users during peak hours. Thus the occasional powering down of monitors and harddisk drives of unused machines during the day helps to save energy and keep the room at a more reasonable working temperature. For operation in compute mode our air-conditioning units are capable of cooling the room with all machines running at full speed, but with the CRT monitors turned off and only few people in the room.

The harddisk drives have a size of 9 GB, allowing several operating system images including application software to be installed at the same time. The disks shipped by the original manufacturers had to be replaced under warranty against the quiet disks agreed upon in the order and the fans of the network switches had to be replaced — otherwise the classroom would have been at risk of loosing its attraction as a decent working place because of too much noise. The cost per workstation in the dual use Patagonia cluster is estimated to stay at about 1.5 times the cost of our pure education clusters (i.e. uni-processor 400 MHz PCs with just 64 MB memory and cheap IDE disks). Given the high cost of the office space in Zürich and the cost of the human resources to plan, install and maintain such a cluster the costs of the PCs are less important than they might appear.

For now copies of all the software packages are stored on the local disks. For the central storage of user data an old Windows NT server based on an Intel 80486 processor was taken over from a previous installation. All functions of this server will soon be replaced by a high performance UNIX server running Samba. The new server can handle user file systems for UNIX and Windows NT accounts in a uniform way. The research partitions are served by a separate Linux server with a raid storage system directly connected to the Gigabit Ethernet. Account management is performed through NIS over the network exactly the same way it is done for the personal workstations of the researchers in our group.

5 The System Software of the Patagonia Cluster

For educational mode of operation, two types of operating systems are required: Windows NT and Oberon. Windows NT is a well-known commodity OS that is most suitable for introduction courses involving the use of word processors, spreadsheets, databases and global information systems such as WWW and e-mail. A single language for the user interfaces is not sufficient to cover the education of students in computer science and other departments. Computer scientists work on several platforms simultaneously and prefer a consistent English user interface across all platforms, while most other students learn faster if the course materials written in German are consistent with the user interface of the operating system and the application software. Until all common operating systems and common applications support switching between multiple languages on the fly, separate installations must be maintained. The different nationalized versions of most application programs can not be installed concurrently and therefore the Patagonia cluster features two fully isolated Windows NT installations for education in German and in English. These images are provided in addition to the Windows NT image with experimental drivers and development tools for research. Again the OS sandboxing techniques created for the dual purpose clusters offer new possibilities for the educational mode of operation.

For the time being, a working set consisting of the most common software is installed on the local disk drives for faster access, ease of installation and to take as much load as possible from the (old and soon to be replaced) central Windows NT server. We plan to migrate more applications back to the server once the new servers become operational. All user home directories are stored on a central server running Windows NT. The boot-partition as well as all other local partitions are hidden and protected from the users access. The partition of the active operating system is remapped and appears to the users of any Windows NT Education image as if there were only one single partition on the C: drive.

Besides the two Windows NT partitions for education in German and English, the disks of the Patagonia cluster also host the Oberon System [9]. Oberon is a programming language, a run-time- and operating system with an integrated development environment for object oriented or structured programming. The Oberon system is kept lean, simple and easily fits into a very small disk partition or main memory. All data is stored on the local disk, which is mounted read only. The modules and objects of the system are only copied to a RAM-disk when required. As there are no user accounts in Oberon and no home directories, no server is needed. Network access is used only for common Internet services like e-mail, WWW and printing. The programming work of students can be kept on a simple floppy disk or on a ZIP drive.

For the computational operations we chose to install Linux and Windows NT as operating systems. Linux is an *Open Source* OS which offers great flexibility to researchers [8]. Some software and drivers for our advanced networking hardware are unfortunately only available for Windows NT, therefore this OS is also installed for selected applications and performance testing in the research mode of our cluster. For maximum flexibility, the Windows NT used in research is installed in a separate partition that is completely isolated from educational use and is password protected as an entire image. Additionally this partition is also maintained by the scientists and file protection is not enabled since it is not needed.

6 Installation, Security and Setup for Maintenance

6.1 Initial Installation

As a first step to initial installation the 9 GB harddisks are partioned into a small 20 MB partition for the bootmanager (see below), a partition for Windows NT education in English, in German and a Windows NT partition for scientific work, each 2 GB in size are. Further, we add two partitions for the ETH specific Oberon System, one for Oberon education and one for Oberon research respectively, each of which is 100 MB in size. As a Linux setup for the computational users we install a 1 GB partition for the root file system and a smaller 128 MB swap partition. A 1 GB partition is left as spare partition for future operating systems to be installed later (such as e.g. Solaris x86, NetBSD or Rhapsody) or for more swap space.

The booting procedure is accomplished by a commercial boot utility named *System Commander*. A startup screen permits the users to select a partition with the desired OS image from a list of options and allows password protection for certain partitions as well as for removable devices such as the floppy or the ZIP drive. The tool boots the mentioned OS without any problems or special configuration.

6.2 Replication of OS Images by "Cloning"

The basic setup for multi-boot and each OS are installed on a master machine. To generate the master we proceed to the point where the OS images had to be configured for each machine with unique names and addresses for network operation. Once the master is set up, we duplicate the internal disk image of the Master to an identical external disk drive. To do so, we boot a service OS (i.e. a minimal Linux system) from a zip drive or a

boot floppy disk and then copy the master image on the internal disk to an external one block-wise with the dd command. Copies of entire partitions are referred to as *cloning*. Subsequently we attach the external disk to all other machines using the dd command in the other direction to copy the external disk onto the internal one. This is only required for the initial setup. For further installation we can use the network to distribute copies.

To keep the individual configuration of the operating systems at a minium, we introduced a DHCP server which serves IP addresses and names derived from the unique Ethernet MAC address in the built in network interface. In addition to DHCP a few simple scripts permit to setup the individual setup of the different machines (e.g. we have a few dual-processor and a few single-processor machines).

Oberon images configure themselves at startup statically based on a table with unique identifiers contained in the primary Ethernet interface. For daily operation the software distribution scheme by cloning also works remotely over the standard and the high-speed networks and is much more comfortable than the initial setup (see Section 6.4).

6.3 Details of the Security Setup

The most important goals of a successful security setup are not to inconvenience the users with unknown and therefore distracting operational procedures of unknown OSes. The purpose of a security setup is protect the integrity of the system installation from corruption and the different users from each other. In a multi-boot setup multiple levels of security are needed:

The first level of this security setup handles the booting procedure. The *System Commander*[1] utility offers menu controlled selection of the operating system of choice and protects research systems and administration setup with a sophisticated password protection scheme, which allows user groups to be defined. This protects educational users from booting unsupported OSes with which they would experience unexpected authentication failures. Further, it protects from booting from external attached devices and floppy disks.

The second level of system security deals with the visibility (or better non-visibility) of other non-active operating systems, their partitions including boot partition with its administrative tools. It should remain impossible to access or modify data from alternate partitions at least not for a normal, mortal user. In order to dissuade educational users from cracking system security, a shareware tool, called *DeviceLock*[2] hides unused partitions and denies access to them from Windows NT.

The same task is accomplished in UNIX systems by configuring the corresponding mounting tables.

For the purpose of allowing the same application links (and shortcuts stored in the user profiles) across all Windows NT installations on campus, we enlist the device name remapping mechanism of Windows NT and remap the partitions in a way that the partition of the active OS image is always the C: drive. With this setup it is even possible to switch between two different Windows NT systems transparently for an English or German version, while maintaining the same user profiles and application links. All these tricks and treats would not be necessary under a well designed OS like Linux that offers mount points in the file system and keeps mount a privileged operation.

As the cluster is also used by many students, the system must make a reasonable effort to protect itself from being modified or damaged. Within an executing OS image some protection on the file level can be accomplished by invoking the appropriate security setup mechanisms offered by UNIX and Windows NT. In the education environment a Windows NT Domain Server authenticates users and controls access to the local and remote files in the cluster — for the research environment a Linux server handles all user authentication employing the NIS protocol. The Oberon system copies itself onto a ram disk upon startup and the access to the disks is restricted to read-only access at the driver level.

6.4 Planned Mode of Daily Operation

The Patagonia cluster is in full operation since the beginning of the year 1999 but some fine tuning (especially for the educational Windows NT setup) is still ongoing, so we do not yet have much experience in daily maintenance.

For improved maintenance in the parallel computing mode a *Cluster Administration Tool* was developed at our site (see Figure 4 in appendix A, [11]). The tool visualizes the state of each machine in the cluster and indicates how it is used (i.e. shows the current load, operating system and user). The tool is designed for a smaller experimental research cluster and even allows to reboot a specified OS remotely provided a reboot daemon is running in all OSes available on the cluster. While the query functions pose no problem, the tool needs to be adapted to work around the device lock of the secure Windows NT installations in the Patagonia cluster.

With the help of the cluster administration tool and the fast network it would be easy to remotely boot a suspect machine with a potentially destroyed education partition into a service OS (e.g. a minimal Linux) and restore the suspect partition from a default image stored on a server.

Should an entire disk image be damaged, a new disk image can be installed in the same way as in the initial installation from an externally attached disk drive

---

[1] System Commander, © V-Communications,
http://www.v-com.com/
[2] DeviceLock for Windows NT, © SmartLine,
http://www.protect-me.com/dl/index.htm

6

or, with a well configured boot floppy disk containing the minimal service OS, even over the network.

As the reinstalled OS supports DHCP, it is not required to configure the installation to the IP address of the reinstalled machine as the system will get its address from the DHCP server. Note that unlike automated installs or configuration scripts our technique is completely OS independent and will therefore work for future releases of Linux and Windows NT. With these mechanisms we could achieve a fully automated maintenance that protects our cluster from the worst OS failures and many damaging manipulations.

6.5 A Network Account Setup for UNIX and Windows NT

The Patagonia cluster integrates the two most widely used operating systems available today, Windows NT and UNIX. All students in our department are given a UNIX and a Windows NT account. In the past the two accounts have been served by two distinct servers. This implies, among other things, that users are now using space on both servers that could very well be integrated into a single one, with a corresponding gain in flexibility and saving in administrative effort. This can be done through the use of the UNIX Samba package, which implements the SMB (Short Message Block) protocol within the Microsoft Networks/OpenNET File Sharing Protocol family, which supports the sharing of file systems, printers and communication abstractions, such as named pipes [5, 10]. Samba enables UNIX file systems to be shared on the PC in the same way as shared file systems are mounted from an genuine NT server. It requires operation of an SMB daemon on the file server and SMB clients on each PC. Two different services of a Windows NT server are important for our cluster: The authentication service as a primary domain controller (PDC) and the simple network file system service. The Samba design for UNIX allows a variety of configurations:

For full integration of account management, the configuration of Samba as a true PDC with authentication and file service for a Windows NT protection domain would be most attractive. However there are severe doubts as to whether this will be practical and robust with the current software releases, since they do not support the full functionality of a PDC. The UNIX solution lacks e.g. support for redundancy in a backup domain controller (BDC) or user groups for example. A conservative deployment requires that we still maintain a genuine Windows NT server as PDC but handle as many services as possible with the existing powerful Sun SPARC infrastructure run under Solaris 2.6. Our current solution is now described in more detail.

In an NT network the user can log in either locally to the PC or to any of the domains in which the PC participates. The user accounts database entry associated with the login defines among other things, the following main attributes: (1) user login and password, (2) groups to which user belongs, (3) location of user home directory, (4) location of user profile.

To allow the user to login from any PC in the domain, the user profile must be of the roaming type, i.e. located on the file server. In the case of the file server providing SMB services from a UNIX machine, the user's home directory registered with the security accounts manager is on the file server and the user profile is located within the user's home directory. In this setup the user has only one home directory for use on both platforms. To cope with this, the following subdirectories are located within the UNIX home directory for use by Windows NT:

NTdata used to store files created directly by the users.

NTprofile holds the user profile tree including ntuser.dat.

Ntdocs holds all the preferences of the different applications run on the Windows NT workstation like MSOffice or Netscape.

Since the Windows NT server stores only rights and privileges the bulk of the user's data is being served from UNIX and the disk quota for UNIX is also applicable for working under Windows NT. It is currently set to 50 MB per user. User profiles will be originally created on the Windows NT server, tested and then placed on the UNIX server. The Windows NT Policy Editor can be used for this task.

This approach entails the installation of an account on UNIX and on Windows NT. The UNIX account is set-up according to well-established procedures and independent of the Windows NT account by the UNIX administrators. For accounting purposes, a single user database is used to record the user's account. However installation remains a 2-step process first on UNIX and then on Windows NT.

The Windows NT subdirectories on the Samba share are set up by a Windows NT login script, i.e., if they do not exist, they will be dynamically created and the appropriate configuration scripts with initial preferences like the Netscape settings, will be copied from a template which is dependent on which group the user belongs to. This means that Samba will take the Home and Profile user's variables from the PDC.

One important issue to be tackled in this approach is automatic password synchronization between the Windows NT PDC and the Solaris NIS setup. With the current configuration featuring a Windows NT Server as a PDC, a Windows NT user password change will be updated on the PDC and BDC (primary and backup NT server) but not for authentication of the file system services provided by the UNIX server under Samba. The Windows NT authentication module has builtin

hooks for integrating library modules implementing third-party authentication protocols, such as Samba. We are currently evaluating a third party software solution that provides proper password synchronization between Windows NT and UNIX.

For the time being, accounts are being created on UNIX and a list is sent to the Windows NT server where the accounts are created with the appropriate rights by running batch scripts. In the very near future we expect to be able to generate accounts on UNIX with installation scripts that will be copied to the PDC in a specific subdirectory and which will run at regular intervals to install accounts automatically.

7 Evaluation of OS Installation by "Partition Cloning"

In this section we will discuss the measured performance of our approach to OS installation by partition cloning and experimentally show that this is an OS independent, fast and scalable approach to better systems maintenance.

Over standard Fast Ethernet (100BaseT) the speed of software distribution is limited by network bandwidth and the particular hub/switch topology of the network in a classroom. The 24 machines in the Patagonia cluster are connected in a way that the there are four rows of tables with space for six machines each. Underneath every row is a twelve port Ethernet hub which is connected to a Ethernet switch in the communication room. With the hub/switch combination all machines on a table share the bandwidth over a single string to the backbone. This network is the base network as it is supported by all used OSes. The classroom setup is quite different from a rack mounted cluster, where the machines are usually close enough to each other for an easier and more homogenous cabling.

For experiments with the high-speed network configuration sixteen machines of our cluster are directly connected to a Gigabit Ethernet switch by fiberoptic cables. The fiberoptic network allows a star network to a switch backplane that can route up to 24 GBit/s of total traffic. The measurements reported here are done on only twelve Gigabit Ethernet NICs, since four cards are presently in use for software development.

One of the PCs in the cluster is equipped with an additional 18 GB Seagate SCSI disk drive which contains the master images used for reinstalling/restoring OS images by partition cloning. This machine serves as a network file system (NFS) server and offers a choice of OS images as image files to all the other machines in the cluster. For the measurements, the clients are brought up under Linux and the shared file system with the images is mounted over the Fast Ethernet or Gigabit Ethernet respectively. An NFS block size of 4096 bytes was used,

as it provides higher performance than 1024 bytes. The data transfers measured for cloning are accomplished by reading and uncompressing the compressed image of a 2 GB Windows NT partition with *gzip* on the clients and pipeing the resulting data stream to the raw device or a *dd* command, which eventually writes the data to the clients partition. In the case of a distribution with plain OS images, the *gzip* command for decompression is not required and *dd* can read and write the data from the NFS mounted partition directly to the local partition which is to be cloned.
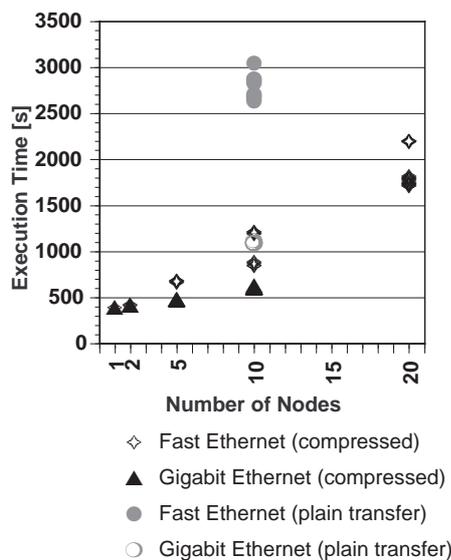

Figure 1: Total execution times for distributing a 2 GB Windows NT Operating System partition simulaneously to 1, 2, 5, 10, 20 machines by partition cloning in the Patagonia cluster.

The measured execution times of a few partition cloning runs with different numbers of clients over the two networks are shown in Figure 1. The numbers indicate that distributing an OS patition over the switched Gigabit Ethernet takes less time than over Fast Ethernet and scales far better with an increasing number of clients, as was expected from the topology. We expect perfect scalability from 10 to 15 clients but an extensive run with all 15 clients connected to Gigabit Ethernet is still to be done. The distribution technique with compressed images seems to reduce the cloning times, although this is not obvious from the basic performance date of the machines. The decompression rate of a 400 MHz CPU, the network bandwidth and the disk write bandwidth of an Ultra2 SCSI drive are sufficiently close to each other that a prediction of performance is interesting. For runs with 10 client machines distributing an uncompressed image over Fast Ethernet synchronization was lost and the server showed an unsual high disk activity which could be responsible for the disproportionally long execution time.
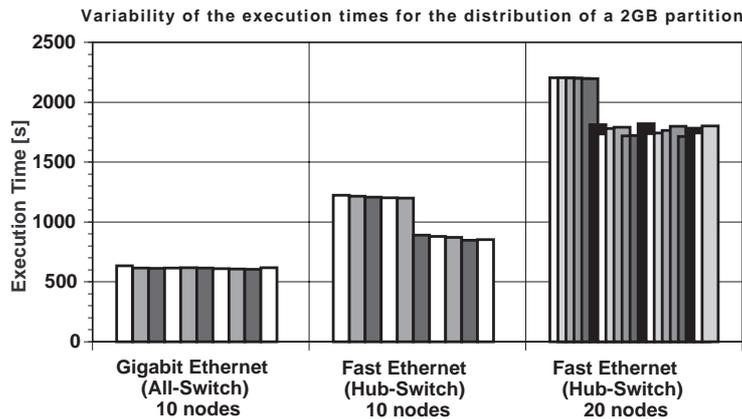
Figure 2: Variability of the execution times for distributing a 2 GB Windows NT Operating System partition simultaneously to 10 and 20 machines by partition cloning over Gigabit and Fast Ethernet.

In our experiments we log the execution time of each client machine involved so we can study the balance of the load put on the network. The execution times are clustered into two groups of clients when Fast Ethernet is used. The effect is best visible in the bar graph of Figure 2, where a separate bar for each client is drawn. The five machines with longer execution times are the machines which are connected to the same hub as the server machine and therfore experience the overload of the network connection to the server.

The search for the limiting factor is best done with a simple pipeing analysis. For the analysis the effective transfer rates to the disks during a distribution are computed and compared to the maximal throughput of the different pieces of the system (e.g. the decompression speed or the disk write bandwidth). The chart in Figure 3 shows the total, aggregate bandwith of data transfers to the disk drives. Note that in some cases this is not the bandwidth as sent over the network because compression is used in some cases. Where compressed images are used, the data is uncompressed by the client after receiving it from the network and before writing it to the disk. Again, this figure shows that compressed images allow for a higher write data rate and that the fully switched Gigabit Ethernet scales better than partially switched Fast Ethernet.

Figure 3: Total (aggregate) transfer bandwidth achieved in distributing a 2 GB Windows NT Operation System partition simulaneously to 1,2,5,10,20 machines by partition cloning in the Patagonia cluster.

8   Conclusions

In our contribution to this years CC99 workshop we are able to report about the succesful installation of the Patagonia cluster at ETH Zürich. Patagonia is a personal computer cluster for both educational and computational use. The cluster is in educational use since the beginning of the year without major problems and the necessary software tools for research are about to be installed. The use of high performance hardware made the cluster immediately popular in our departement. Us-
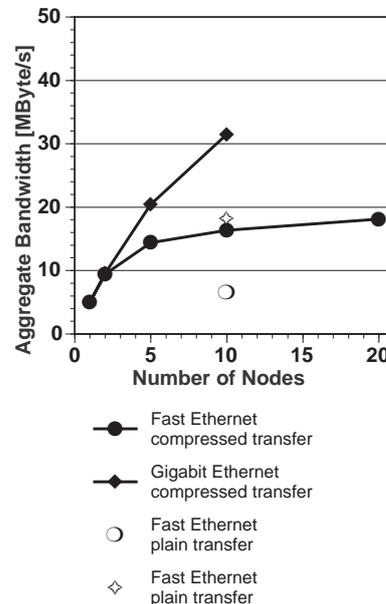
ing a third party boot manager with a good user interface helps computer illiterate users as well as more demanding users to boot their favored OS on the machines without much effort. The wish-list of improvements includes a standardized and more visible shutdown button and a better usage of the power saving features in all installed operating systems to keep the classroom cooler and more pleasant to work in.

The high end configuration of our research machines helped their purpose in several ways. The fast networks and fast disks simplifiy the installation, setup and main-

tenance procedures although using multiple operating systems can lead to highly increased complexity. During the initial configuration of our "master" machine, frequent backups of the entire disk or individual partition images to the big and fast disk drives for achival were an indispensable tool, as small changes in one partitions OS configurations could break the system so badly that a restart and further installations remained impossible. Especially the proper setup of Windows NT security involved much trial and error since the OS lacks transparency and the manufacturer offers little support for the configuration of a networked system with such advanced and ambitious requirements.

Despite contrary advice by several researchers in the OS community and the manufacturers the distribution of pre-configured OS images by cloning technique proved to be succesful and reliable to replicate operating system images to make frequent backups. Automatic maintenance and automatic restore of broken file systems can be done over the high speed network provided for cluster computing. We measured a copy performance of 3–4 Megabytes per machine while replicating OS images to up to 10 machines simulaneously.

Currently installed operating systems on the Patagonia cluster are Linux, Windows NT and the Native Oberon System. All three systems could be integrated into an existing Sun server infrastructure with all user accounts and user data directories served by those departemental servers. While the home directories could be provided by UNIX servers, the authentication and accounting information must still be replicated on a genuine Windows NT controller.

The resulting cluster of PCs is competitive and cost effective for computational science and highly interesting as a classrom for education. With its fast dual processors, the Gigabit/s network and its large memories the installation will certainly have a good window of opportunity as a research platform for parallel and distributed computing leading to many interesting research results in fast network file systems, fast middleware for distributed objects and applications in computational science. The cluster is expected to stay competitive for about two to three years in research mode (i.e. until the processing power of a single node has doubled) and last for about four to five years for educational use.

9 Acknowledgements

References

[1] W. Y. Arms. Reflections on Andrew (educational computing at Carnegie Mellon). *EDUCOM Review*, 25(3), 1990.

[2] Henri Bal. The Distributed ASCI Supercomputer (DAS). *http://www.cs.vu.nl/~bal/das.html*.

[3] D. J. Becker, D. Sterling, T.and Savarese, J. E. Dorband, U. A. Ranawake, and C. V. Packer. Beowulf: a parallel workstation for scientific computation. In *Proceedings of 1995 ICPP Workshop on Challenges for Parallel Processing*, Oconomowc, Wisconsin, U.S.A., August 1995. CRC Press.

[4] Nanette J. Boden, Robert E. Felderman, Alan E. Kulawik, Charles L. Seitz, Jakov N. Seizovic, and Wen-King Su. Myrinet - A Gigabit per Second Local Area Network. *IEEE-Micro*, 15(1):29–36, February 1995.

[5] Microsoft Corporation. Developer Tools and Information, Developer Relations Group. *ftp://ftp.microsoft.com/developr/drg/CIFS/*.

[6] Christian Kurmann and Thomas Stricker. A Comparison of two Gigabit SAN/LAN technologies: Scalable Coherent Interface versus Mryinet. In *Proceedings of SCI Europe '98*, 1998. Also to appear in *SCI-Based Cluster Computing*, editors Hermann Hellwagner and Alexander Reinefeld.

[7] Burkhard Monien and Alexander Reinefeld. PC^2. Cluster Systems at Paderborn Center for Parallel Computing *http://www.uni-paderborn.de/pc2/systems/index.htm*.

[8] Eric S. Raimond. The open source initiative. *http://www.opensource.org/*.

[9] Martin Reiser. *The Oberon System*. ACM Press, 1991. ISBN 0-201-54422-9.

[10] David A. Solomon. *Inside Windows NT*. Microsoft Press, second edition, 1998.

[11] Rolf Spuler. Cluster Administration Tool. Internal report, ETH Zürich, July 1998.

[12] Thomas M. Stricker, Christian Kurmann, Michela Taufer, and Felix Rauch. CoPs — Clusters of PCs Project overview. *http://www.cs.inf.ethz.ch/CoPs/index.html*.

[13] Nicklaus Wirth and Jürg Gutknecht. *Project Oberon*. ACM Press, 1992. ISBN 0-201-54428-8.
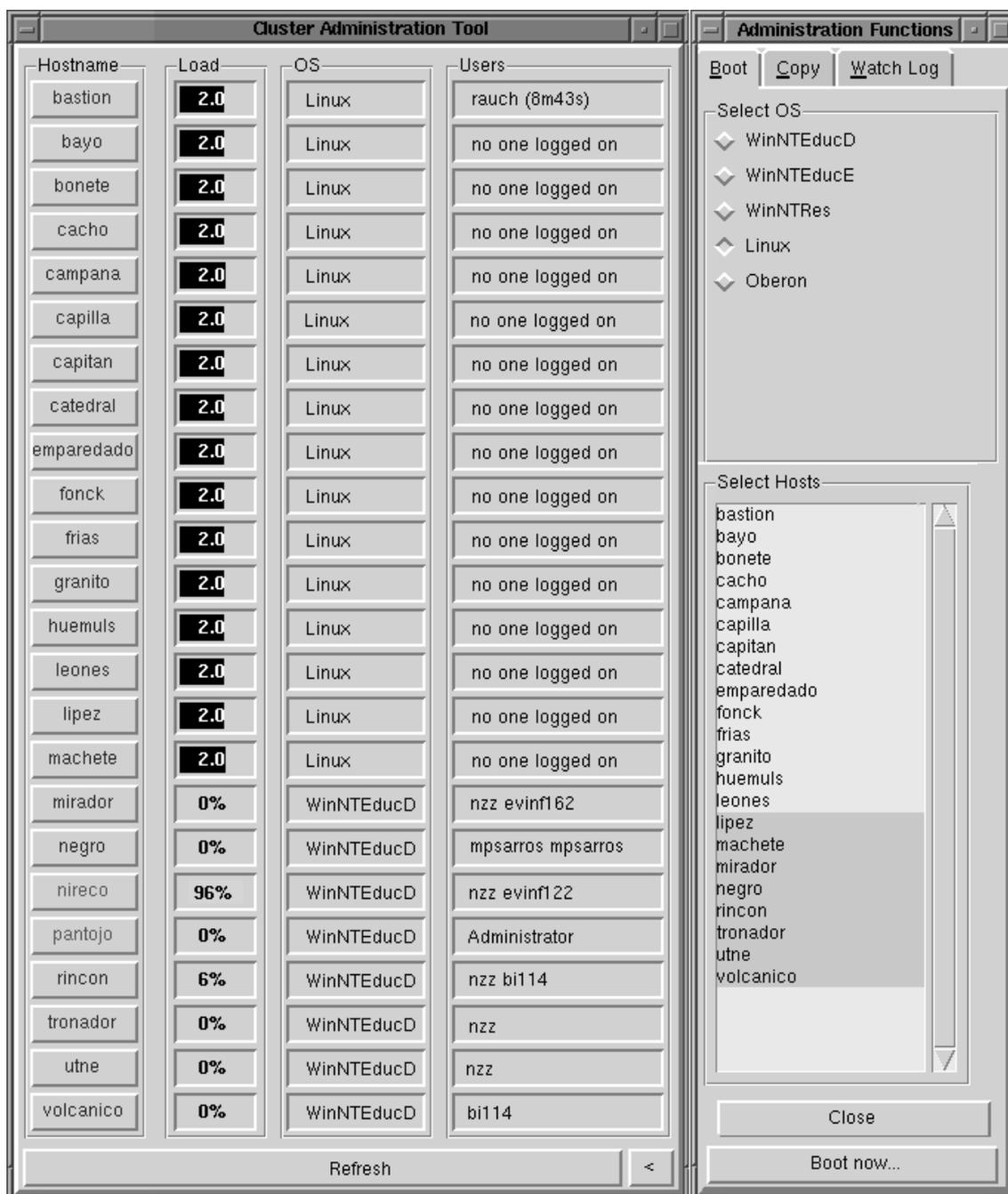
Appendix A: Illustrations

Figure 4: The Patagonia cluster administration tool allows remote supervision and operating system selection. The monitor and console windows are easily portable Tcl/Tk scripts. The maintenance functions are supported by catd (cluster administration tool deamon) which was ported to most operating systems in the cluster.