

Patagonia Cluster Project

Clusters of PCs Multi-Boot and Multi-Purpose?



Christian Kurmann, Felix Rauch,
Michela Taufer, Prof. Thomas M. Stricker
Laboratory for Computer Systems
ETHZ - Swiss Institute of Technology
CH-8092 Zurich

Research Cluster



Distributed Supercomputer, Henri E. Bal, U.Amsterdam

Characteristics of a Research Cluster

- fast processors (1 - 4 per node)
- large memory (RAM and disk)
- scalable high performance network (Switches and Gigabit/s)
 - e.g. Myrinet SAN, Giganet SAN, Swiss Tx-net, SCI
- usage pattern:
 - development during the day
 - **experiments throughout the night**

Education Cluster



Patagonia Cluster, CS Department, ETH Zürich

Characteristics of an Education Cluster

- large disks for multiple extensive software installations
- operating systems with standard and specialized configurations
- lots of room: LAN needed
- system security:
 - software installation
 - student data
 - hardware (theft)
- usage pattern:
 - **during the day only**

“Multimedia Collaborative Work”



Multimedia PCs at the Lab for Computer systems

⋮

Characteristics Multimedia Workstations

- large disks for large amount of software installations
 - operating systems in maintained default configuration and user specific configurations
 - lots of space: LAN needed
 - Gigabit Ethernet or Switcherland (Multimedia)
 - system security
 - installation and user data
 - usage pattern:
 - predominantly during the day
- ⇒ Widely the same requirements as in education!

7

⋮

Multi-user/Multi-purpose-Cluster

- **Observation:**
All three types of clusters have mostly the same requirements but totally different usage patterns.
- **Thesis:**
 - A single multi-boot cluster matches all needs!
 - OS and file system independent installation tool needed

8

⋮

Overview

- Multiboot / Operating Systems
 - Patagonia Project at DINFK/ETHZ
 - Patagonia Security / Maintenance
 - Installations with Cloning / Data Streaming
 - OS-Image Repositories
 - Evaluation Cloning / Repositories
 - Alternatives to Multiboot
 - Conclusions

9

⋮

The Right Operating System



10

⋮

System Software

Education:

- Windows NT German with file system security
- Windows NT English with file system security
- Oberon
- Solaris 2.7, same setup as Sun SPARC Cluster

Research:

- Linux
- Windows NT English without file system security
- Oberon, development system

11

⋮

Common Problem



Maintenance of software installations is hard:

- Different operating systems or applications in Cluster
 - Temporary installations: tests, experiments, courses
 - Software rejuvenation to combat software rotting proc.
- Manual Install:** days, **Network Installs:** hrs, **Cloning:** min

12

The Patagonia Project

Multi-boot, multi-purpose Cluster for the
Department of Computer Science at ETH Zurich

- 60 PCs (600-1000 MHz Dell, SCSI)
- 3 Computer Labs
- Windows NT, Linux, Oberon
- Used for Education and Research

13

Patagonia Installation

Replication through cloning

- Initial installation
 - configuration of a master disk
 - block wise copy of the master disk
- Cloning of single disk images / partitions
 - boot a service OS
 - block wise copy of the images over the network

14

Patagonia Installation II

Partitioning:

Boot-Partition	0.020 GB
Partitions for Windows NT 4.0 Education	2 x 2.0 GB
Partitions for Windows NT 4.0 Research	2.0 GB
Partitions for Linux	1.5 GB
Partitions for Oberon	2 x 0.1 GB
Spare partition (Solaris, Oracle...)	1.0 GB
Small Service Linux Partition	0.25 GB

approx. 9 GB

15

Patagonia Auto Configuration

Configuration of machine specific parameters:
IP number, Hostname, HostID, SystemID, SecurityID

- manual
- automatically through DHCP (with server)
- automatically through Ethernet MAC address (with static table)

DHCP = Dynamic Host Configuration Protocol

MAC = Media Access Control ID - Ethernet board built-in

16

Security in Patagonia

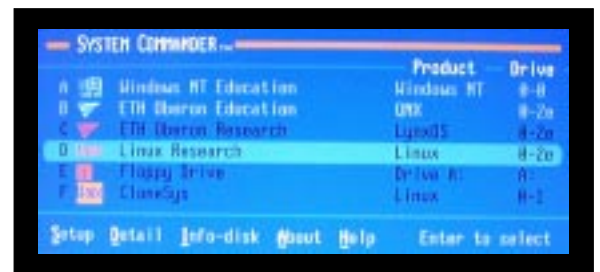
Goal: Security without interference for the users

⇒ reached through three levels

- booting with System Commander
- separate and hide partitions
 - lock partitions with Device Lock and resetting device IDs to C: (Windows NT)
 - mount tables (UNIX)
 - write-protected partitions plus RAM disk (Oberon)
- usage of access rights and clean authentication by a central server

17

Boot-Manager: System Commander



18

Accounts for UNIX and Windows

- Efficient Sun-Server for:
 - home directories over SMB (with Samba) resp. NFS (UNIX) from Sun-Server
 - authentication through a specific Windows NT-Server (NT-Clients) resp. NIS Server (UNIX-Clients)
- Account generation on UNIX, automatic generation for NT via scripts
- Password synchronisation with commercial tool *Passync*

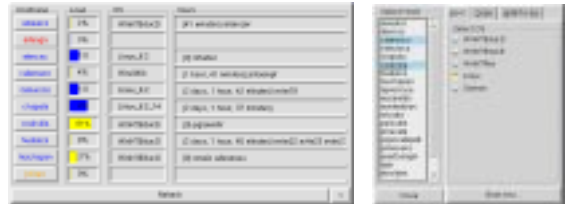
SMB = Server Message Block Protocol
 NFS = Network File System
 NIS = Network Information Service (former yellow pages)

19

Maintenance with CAT

Cluster Administration Tool (CAT):

- Cluster information system (running OS, users, load..)
- Remote boot to selected OS
- Installation daemon (*Dolly* for reliable data streaming)



20

Analytic Model for Data Distribution

Cloning through "partition-cast" can be done in many different ways:

Factors:

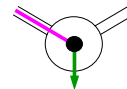
- data compression
- disk bandwidth / network bandwidth
- utilization of I/O bus, CPUs, memory system
- network topology

An analytic model [Rauch, EuroPar2000] helps to optimise for the best alternative.

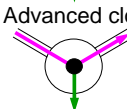
21

Active Nodes / Topologies

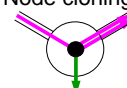
- Simple receiver for star topologies
- Advanced cloning node for multi drop chains
- Node cloning streams for general spanning tree



Simple Receiver



Multi-drop Chain



Universal Active Node

22

Tools for Partition-Cast

- *dd*/NFS, built-in OS function and network file system based on UDP/IP - simple - permits star topology only
- *Dolly*, small application for streaming with cloning based on TCP/IP - reliable data streaming



Dolly
 for reliable data casting
 on all spanning trees

- star (n-ary)
- 2-ary, 3-ary
- chain (unary)

23

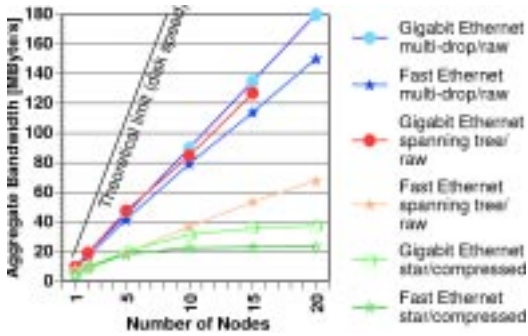
Experimental Setup

- Topologies:
 - Star
 - 3-ary spanning tree
 - Multi-drop chain
- Fast Ethernet / Gigabit Ethernet
- Compressed / Uncompressed Images

All experiments: Distribute 2 GByte to 1..15 clients

24

Scalability



25

Data Streaming Results

- Optimal configuration derived from our [model](#).
- In most cases, the [multi-drop chain](#) delivers perfect scalability (thanks to the Ethernet switch) and is better than any other spanning tree configuration.
- Complex [multi-cast protocols](#) cannot be implemented for the required throughput - but they [are not required](#) anyway.
- Resulting simple [Streaming Tool Dolly](#) transfers a 2 GB Windows NT installation image to 60 workstations in less than 5 minutes at a sustained transfer rate of 9 MB/s per node

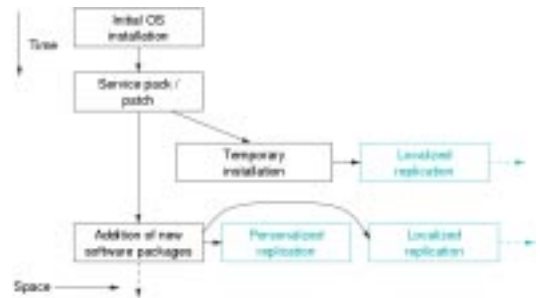
26

Evolution of Software Installations

- Software Installations are incremental and constantly changing
 - Temporal diffs: New installations, Patches, Upgrades
 - ⇒ need ability to go back
 - Spatial diffs: Localization, Personalization
 - ⇒ OS dependant, file system dependant, removal not clean
- Cloning needs no OS and file system knowledge but leads to intense image archives
- Solution: Store only changes in OS independent partition repository

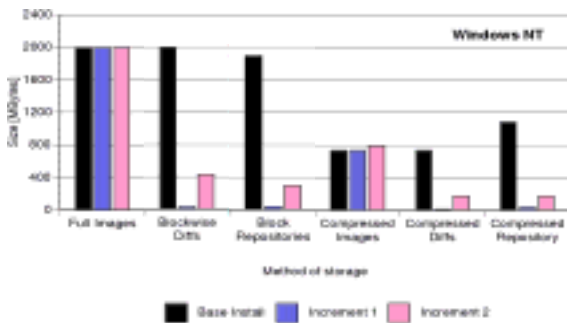
27

Characteristics of Installations



28

Incremental Images in Repository



29

Incremental Repository Results

- Problem of software maintenance in large clusters of PCs was analyzed.
- The proposed system is based on an [storing and distributing entire partition images](#).
- Investigated [block-wise differential partition repositories](#) work well and optimally compress the archive.
- Solution is [independent](#) on file system and work with all operating systems, also future versions of current OSes.

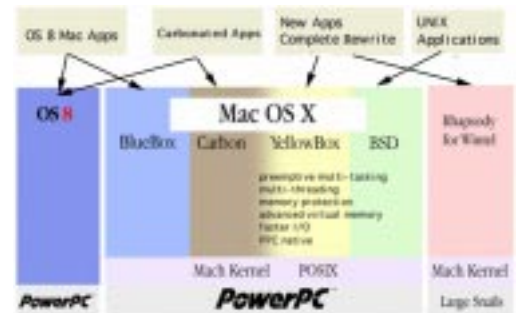
30

Alternatives to Multiboot

- Binary Translators, [Emulators](#), VMs, JIT
 - Softwindows for Macintosh, Java Virtual Machine
- Micro Kernel with [OS adaption layers](#)
 - latest attempt by Steve Jobs: Mac OS X
- Low level [virtual machines](#)
 - Disco Project, Stanford University, Prof. Mendel Rosenblum and Eduard Bagnion, SGI Cluster
 - **VMWare™**: appliance of technology on PCs, Linux and Windows NT

31

Micro Kernel with OS adaption layers



32

VMWare™: Low Level Virtual Machines



33

Conclusions

- Successful **installation and beginning of operation** of Patagonia, a universal multi-purpose Cluster for **Research, Education and Collaborative Work**.
- Facilitation of maintenance and Installation through:
 - small [service operating system](#) (Linux)
 - fast, large disks (SCSI)
 - switched network
- **Multi-boot** installations lead to [great flexibility](#)
- Cloning of entire software installations scales well
- Differential Repository optimally compresses sequences of OS installations

34

Published Papers

- F. Rauch, Ch. Kurmann, T. Stricker: **Partition Repositories for Partition Cloning - OS Independent Software Maintenance in Large Clusters of PCs**. Proceedings of the IEEE International Conference on Cluster Computing 2000, Chemnitz, Germany.
- F. Rauch, Ch. Kurmann, T. Stricker: **Partition Cast - Modelling and Optimising the Distribution of Large Data Sets in PC Clusters**. Distinguished paper published at European Conference on Parallel Computing, Euro-Par 2000, Munich, Germany.
- F. Rauch, Ch. Kurmann, B. M. Müller-Lagunez, T. Stricker: **Patagonia - A Dual Use Cluster of PCs for Computation and Education**. Proc. of the second workshop on Cluster-Computing, March 1999, Karlsruhe, Germany.

35

Q&A

CoPs (Clusters of PCs)

Project-Homepage

<http://www.cs.inf.ethz.ch/CoPs>



Described tools including source code are available for download under the GNU general public license from the above URL.



36