WHITE PAPER

January 2000

Prepared by
Compaq EMEA HPC
solution center

Compaq Corporate
Corporation

CONTENTS

Compaq Cluster Management Utility Features

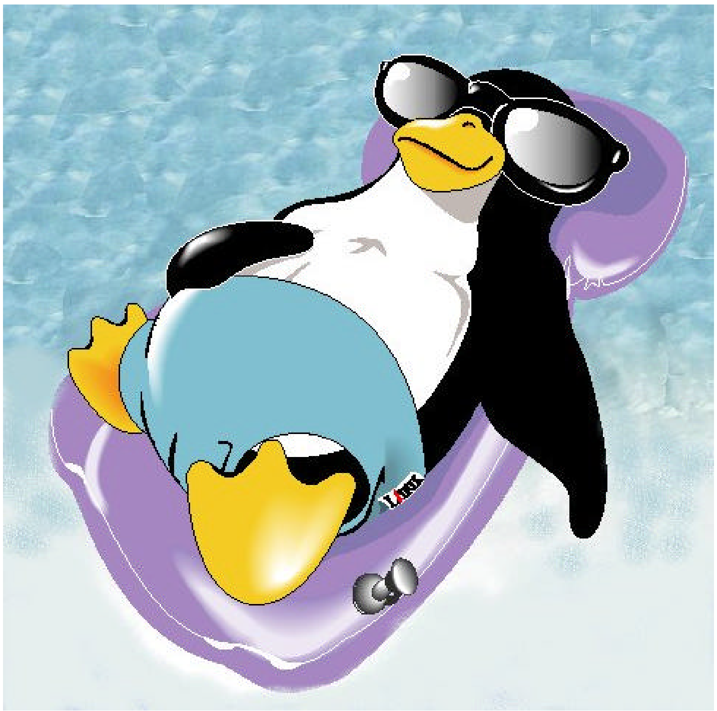
Overview

Today High Performance Computing architecture can be split into four classes: Massively Parallel Processing as with the IBM SP2 and SGI/CRAY T3E; High End SMP systems like Compaq GS320 or HP V2500; clusters of midrange computers like Compaq's HPC160/320, and clusters of workstation.

This latest architecture, also called Beowulf clusters, consists of large collections of low cost systems (running LINUX) linked to a computing infrastructure via an Ethernet network or low latency SAN. This kind of cluster can be very cost effective if the problem to solve can be split in a large number of distributed memory jobs that will run efficiently on one CPU and requires little inter-job communication.

Compaq Cluster Management Utility is a tool that will help you to manage a large collection of systems within a Beowulf cluster environment. CMU makes the management of this cluster more user friendly, efficient and error free. CMU will also help you keep your Beowulf cluster a cost-effective solution by decreasing the management functions related to cost of ownership.

This document describes The Compaq Cluster Management Utility features and usage.



Tux taking advantage of Compaq CMU.

Where to use Cluster Management Utility.

The Cluster Management Utility can be used everywhere you need to manage a number of standalone systems that are close together in a hardware and software configuration. It has been designed to manage Beowulf clusters or CPU farms, where almost all nodes are identical and identically configured. Slightly different hardware is managed by creation of groups.

CMU takes advantage of the Serial Remote Management (SRM) console of the Compaq AlphaServer or Alpha Workstation systems. The console server(s) is used to concentrate system console from a single monitoring node.

The compute nodes can run either Linux or Compaq Tru64 Unix.

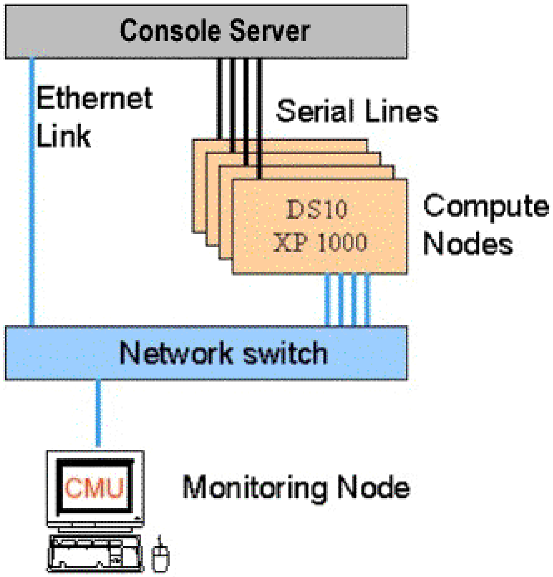


Figure 1: CMU Hardware Environment

Cluster Management Utility Features.

The Cluster Management Utility is designed to help with the management of a large number of compute nodes. CMU comes with a Graphics User Interface (GUI) that can be tailored for your needs and for any number of nodes in the cluster. It allows access to all compute node consoles from a single screen with only a single mouse click. The CMU main window gives you access to a configuration screen where you can setup console server hardware.

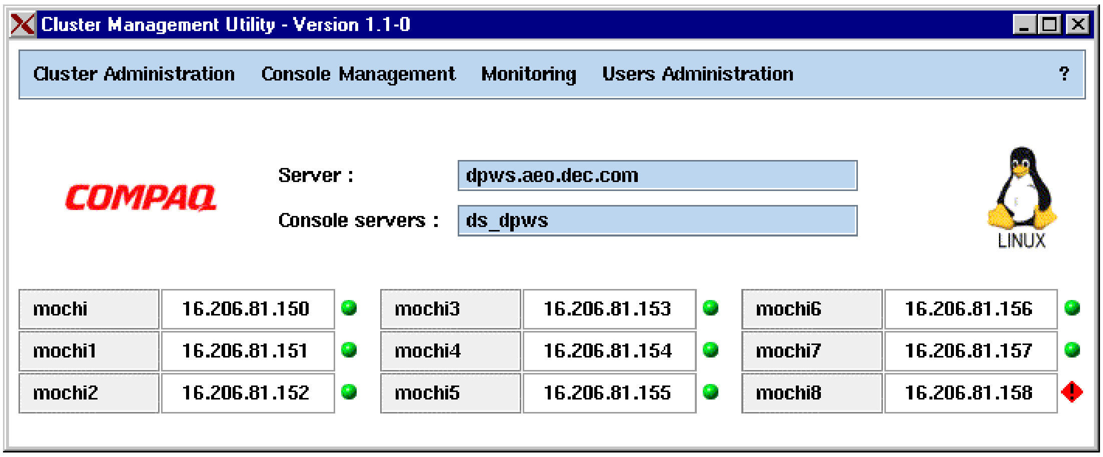
From CMU you can monitor, halt, boot or reboot any selection of nodes. You can connect to several nodes in the cluster and broadcast commands to them from a single keyboard session. CMU also helps you manage events coming from the cluster, e.g., nodes going up or down. More details on this feature are provided in the chapter titled, "Day-to-day Administration Features".

CMU also has the capability to propagate a system configuration to all the compute nodes in the cluster. CMU can clone disk partition contents from an image server to the compute nodes local disks over the network. This can be used for first time installation of compute nodes and to propagate updates to the kernel or the current system configuration. CMU takes care of the target disk partitioning in cases where the target partition differs from the initial image. CMU partitions the target disk during the cloning phase, avoiding the partitioning of each compute node during the first installation. More details about cloning is available in the chapter titled, "Cloning feature".

Day to day administration feature

CMU Graphic User Interface.

CMU helps in the day-to-day administration through its Graphical User Interface (GUI). Figure 2, displays a sample GUI showing an entry for each node declared in the cluster.



Server : dpws.aeo.dec.com

Console servers : ds_dpws

mochi	16.206.81.150	🟢	mochi3	16.206.81.153	🟢	mochi6	16.206.81.156	🟢
mochi1	16.206.81.151	🟢	mochi4	16.206.81.154	🟢	mochi7	16.206.81.157	🟢
mochi2	16.206.81.152	🟢	mochi5	16.206.81.155	🟢	mochi8	16.206.81.158	♦

Figure 2: CMU Graphic User Interface Main Window

For each entry, CMU displays the node name, its IP address and status. A green bullet indicates that the node is reachable over the network. The status is refreshed on a regular basis.

The server field identifies the node running the CMU software and is used as the image server of the configuration.

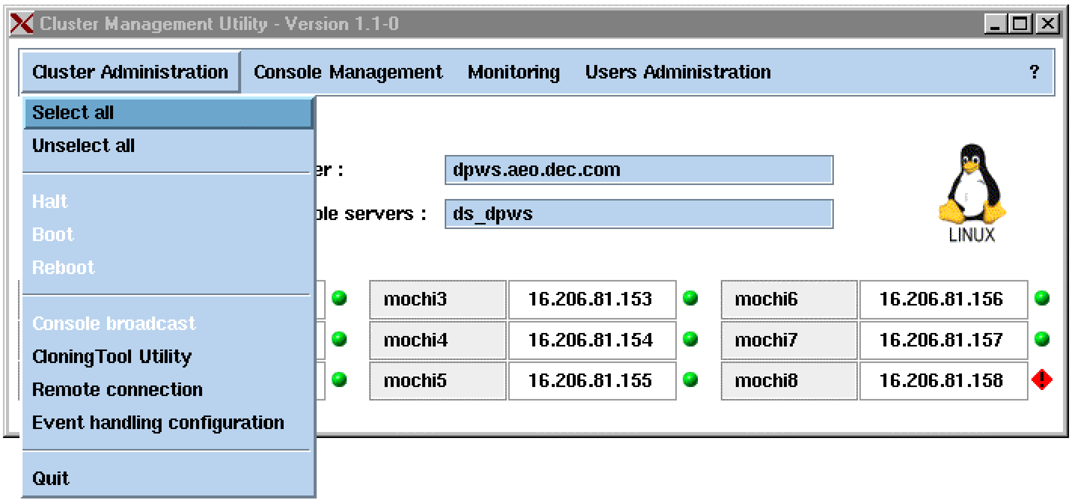
The console server field lists the console server(s) configured within the cluster. These are used to gain access to compute node consoles.

The total number of nodes and the number of columns displayed can be tailored in the configuration file to meet your specific needs.

- Selection:
  You are able to select/deselect one or several compute nodes by clicking on a node name. This is often used with selections like boot, reboot, halt or the monitoring tools.

- Short-cuts:
  A right-click on a machine name opens a telnet session to the selected node. When CMU is correctly configured, a Ctrl right-click opens a terminal session through the console server. This can be done even if the compute node is not booted.

Cluster Administration



er : dpws.aeo.dec.com

ple servers : ds_dpws

mochi3	16.206.81.153	●	mochi6	16.206.81.156	●
mochi4	16.206.81.154	●	mochi7	16.206.81.157	●
mochi5	16.206.81.155	●	mochi8	16.206.81.158	◆

Figure 3: Cluster Administration Menu

The menu (above left) allows you to halt, boot or reboot one or all nodes selected.

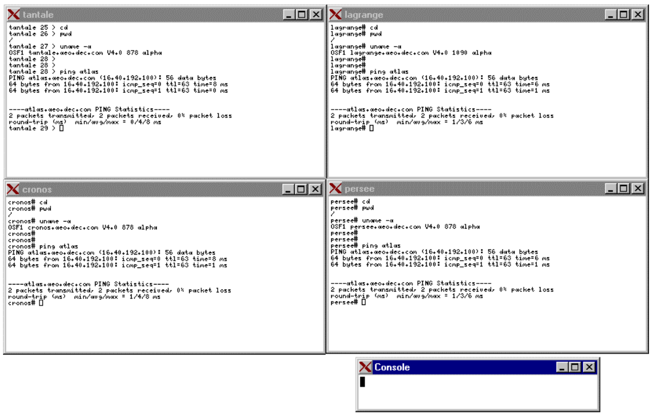
Halt: rsh command is used to halt selected nodes.

Boot: The boot SRM command is used to boot the node selection. A sub-menu defines if the node boot from network or from its local disk. It also defines how many nodes can boot at a time. When booting over the network this feature helps to avoid network congestion.

Reboot: rsh command is used to reboot nodes defined by the selection.

Console broadcast

CMU provides the ability to pass commands to several compute nodes within a single activity session. All the characters typed in the master console will be broadcast to all xterm sessions of your selection. You can also iconify some xterm sessions if their number is too high for your monitor size.



```
tantale 25 > cd
tantale 26 > pwd
/
tantale 27 > uname -a
OSF1 tantale.aeo.dec.com V4.0 878 alpha
tantale 28 >
tantale 28 >
tantale 28 > ping atlas
PING atlas.aeo.dec.com (16.40.192.100): 56 data bytes
64 bytes from 16.40.192.100: icmp_seq=0 ttl=63 time=8 ms
64 bytes from 16.40.192.100: icmp_seq=1 ttl=63 time=0 ms


----atlas.aeo.dec.com PING Statistics----
2 packets transmitted, 2 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 0/4/8 ms
tantale 29 > 
```

```
lagrange# cd
lagrange# pwd
/
lagrange# uname -a
OSF1 lagrange.aeo.dec.com V4.0 1090 alpha
lagrange#
lagrange#
lagrange# ping atlas
PING atlas.aeo.dec.com (16.40.192.100): 56 data bytes
64 bytes from 16.40.192.100: icmp_seq=0 ttl=63 time=6 ms
64 bytes from 16.40.192.100: icmp_seq=1 ttl=63 time=1 ms


----atlas.aeo.dec.com PING Statistics----
2 packets transmitted, 2 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 1/3/6 ms
lagrange# 
```

```
cronos# cd
cronos# pwd
/
cronos# uname -a
OSF1 cronos.aeo.dec.com V4.0 878 alpha
cronos#
cronos#
cronos# ping atlas
PING atlas.aeo.dec.com (16.40.192.100): 56 data bytes
64 bytes from 16.40.192.100: icmp_seq=0 ttl=63 time=8 ms
64 bytes from 16.40.192.100: icmp_seq=1 ttl=63 time=1 ms


----atlas.aeo.dec.com PING Statistics----
2 packets transmitted, 2 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 1/4/8 ms
cronos# 
```

```
persee# cd
persee# pwd
/
persee# uname -a
OSF1 persee.aeo.dec.com V4.0 878 alpha
persee#
persee#
persee# ping atlas
PING atlas.aeo.dec.com (16.40.192.100): 56 data bytes
64 bytes from 16.40.192.100: icmp_seq=0 ttl=63 time=6 ms
64 bytes from 16.40.192.100: icmp_seq=1 ttl=63 time=1 ms


----atlas.aeo.dec.com PING Statistics----
2 packets transmitted, 2 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 1/3/6 ms
persee# 
```

Figure 4: Console Broadcast

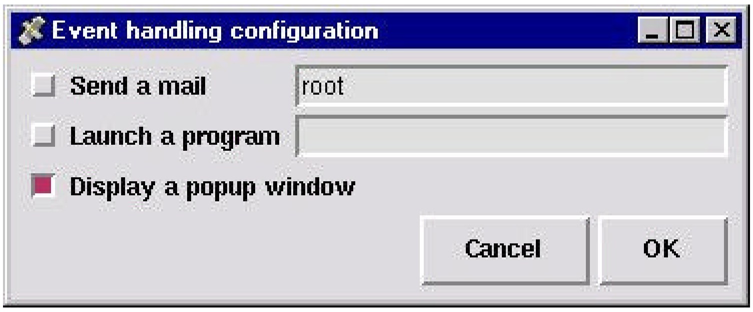
You can use direct network connection or console connection for broadcasting send commands to several nodes even if the nodes are not presently booted.

When the master console is closed, all related xterm sessions within the same broadcast will also be closed.

Event handling configuration

CMU monitors compute nodes by polling their TCP/IP echo port. When CMU detects a change in node status, an event is generated. This menu defines how CMU handles the event and how it should inform the system manager.
CMU can take three actions:

- Display a popup window
- Send a mail to system manager
- Launch a custom program or script



□ Send a mail: root

□ Launch a program:

☑ Display a popup window

Figure 5: Event Handling Configuration Window

Console Management.

This menu helps to configure the console server(s) within the cluster. You can add or remove a console server and its associated console server port for all appropriate compute nodes.

CMU supports 8, 16, 24, or 32 port console servers (DECserver 90M, DECserver 700 or DECserver 900).

You configure a console server by entering its TCP/IP hostname and address.

CMU can also connect you to any console server port even if it is not already assigned to a compute node.

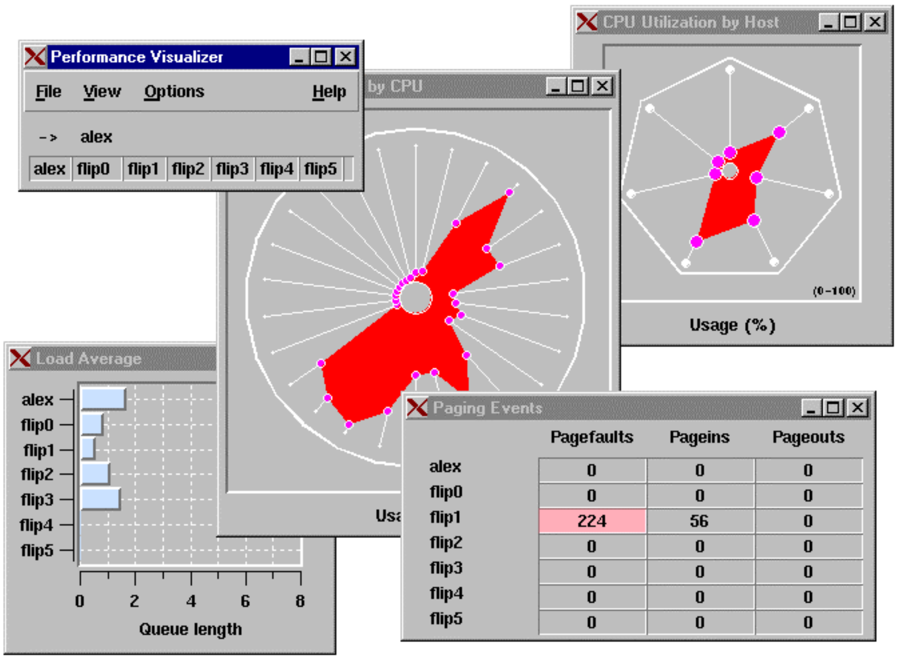
Only one telnet connection at a time is available to a console server port. Through this menu you can also activate disconnection of a port already in use.

Monitoring tools.

Monitoring tools can be launched from the CMU monitoring menu. The monitoring program is launched on all selected nodes, with the result displayed wherever CMU is running.

Monitoring tools available from CMU are:

For Linux cluster:	**For Tru64 UNIX cluster:**
xosview	PVIS
xsysinfo	top
xcpustate	
top	



	Pagefaults	Pageins	Pageouts
alex	0	0	0
flip0	0	0	0
flip1	224	56	0
flip2	0	0	0
flip3	0	0	0
flip4	0	0	0
flip5	0	0	0

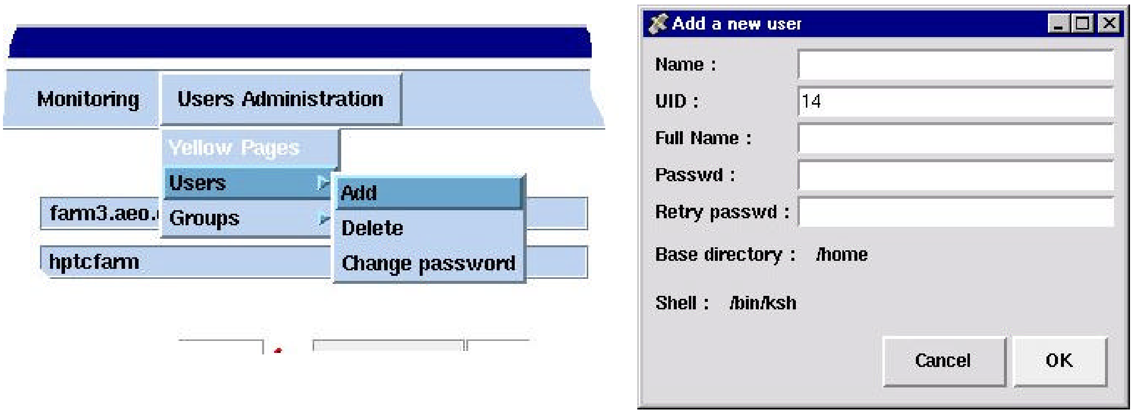
Figure 6: Performance Visualizer (PVIS)

In the case of PVIS, the Performance Visualizer is launched on the set of nodes defined by the selection.

User enabled management

CMU offers a GUI to deal with user enabled management. You can add, delete, or change user passwords and manage groups from CMU menu.

Users and groups can be modified on the node running CMU. If the Network Information Service (NIS) is configured to propagate your user's name, password or group definition to all your clustered nodes, CMU offers a way to update your NIS database with a single mouse click.



Name :

UID : 14

Full Name :

Passwd :

Retry passwd :

Base directory : /home

Shell : /bin/ksh

Figure 7: Users Administration Menu

Cloning feature

Overview and terminology

CMU offers the ability to clone system disk partitions of one node to other cluster nodes. This will avoid the painful and time-consuming task of system installation or configuration for each node in the cluster.

Group definition
A cluster is often split into groups of compute nodes for multiple reasons: all the nodes may not have the same hardware, may be used for different tasks or be owned by different parts of the IT organization. CMU requests that within a group, all the nodes have exactly the same hardware configuration, since the same system image will be duplicated to all group members. A set of physical nodes can have several system images to play with. In this case you need to create, as many logical groups of images that you may need to deal with. There is a unique system image per group, but a node can be part of several groups.

Clone image
After a complete system installation of one node of a group, CMU builds a compressed image of this master disk. This image can be done from a remote node within the group or from a disk on the system running CMU. The partitioning information of this master disk and the compressed contents of some of its partitions are stored within a "clone image". This image is ready to be propagated to other group members using CMU.

Image server
All the "clone images" are stored on the image server. This is the node that is running CMU. The image server stores the clone image on one of its local file systems. If a large number of groups need to be stored on an image server, the correct amount of storage should be available. A "clone image" sits in a directory entry named by group name. It contains the target system disk partitioning and compressed image of the following partitions: / (root), /usr and /var.

Cloning chronology

Disk image preparation

The first step of the cloning mechanism is to prepare a "clone image" from a master disk. We already mentioned that CMU could do this from a disk that is physically located on the image server or on a remote node.

To build a compressed image of a master disk, it must not be used by the operating system in the same time frame. That means that CMU cannot build a compressed image from a booted system disk.

For a local disk, CMU mounts the master disk in the file system space of the image server and builds the "clone image" directly.

For a remote disk, the node needs to be booted through the network to free the local disk of any operating system activity (compute nodes often have only one physical disk to boot on). To do that, CMU takes advantage of the Bootp protocol. The kernel is downloaded using TFTP protocol and the / (root) of the operating system is NFS mounted on a given space in the image server.

When the node is correctly booted the CMU "backup" script is started. The local disk is mounted, and a compressed image is built and sent along with the partitioning information to the image server through a TCP/IP socket protocol.

Image propagation

When a clone image is ready, the next step is to propagate it to all group members.

Network reboot

Since the system disk of the compute nodes needs to be modified, the running operating system at this time should not use the local system disk. So, the nodes need to boot over the network to free the local disk of the operating system tasks. All the nodes boot a specially tailored image using Bootp and FTP and NFS mounted from the same image server space.

This image automatically starts a set of actions defined in the CMU "cloning" script.

Partition check

When booted, the nodes mount their local system disk and check its partition against the clone image partitioning. If descriptions differ, the startup script changes the local disk partitioning and rebuilds file systems on it.

Nodes Upload

When the partitioning is correct, each node waits for the compressed image to be downloaded. This state is signaled to CMU running on the image server, indicating that the node is ready for upload.

CMU then transfers the selected compressed image to the first node. This is done using TCP/IP socket.

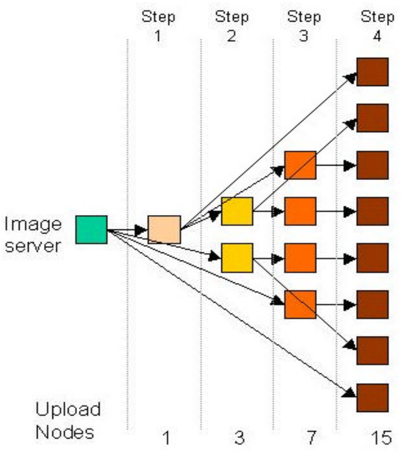


Figure 8: Image Propagation

When the transfer is completed, the compressed image is saved to the local disk partitions in / (root) /usr and /var.

The node then asks the image server if there are any successors waiting for upload. If there is, it starts to transfer the image to a group member, while the image server uploads a third one.

This is named the *tree propagation algorithm*. Once a node has received a completed image, it tries to download it to another node within the group. CMU manages the group node list waiting for an image. This mechanism speeds up the propagation process and takes advantage of the available network bandwidth.

The propagation time is proportional to the log number of nodes to be cloned and is not limited by the image server capacity since it downloads one node at a time.

This mechanism requires a large network bandwidth and is particularly designed to take advantage of switched network configurations.

Image Decompression and adjustment

Compressed image decompression

When the CMU list of nodes to be cloned runs out, it answers "no-node" to any further neighbor requests. This is the signal for compute nodes to uncompress the image on their local disk. So the image decompression will take place on all nodes in parallel. This is once more designed to speed up the cloning process.

Configuration adjustment

After this step, each node within the group has an identical system disk image on its local system disk. To be able to reboot on the local disk, some nodes specific files need to be adjusted node by node (TCP/IP node address and hostname for example).

CMU adjustment is limited to the TCP/IP name and address of the primary network interface. The TCP/IP Hostname and address set for the compute nodes are the ones defined by the Bootp configuration file in of the image server. (e.g., /etc/bootptab).

If there is need for other specific adjustments, partitions that are not part of the cloning mechanism can be used to store node specific information and an "adjust" script needs to be customized to setup this information correctly on a per node basis.

Local disk reboot

The cloning process is now terminated. Each node within the group has a system disk identical to the master image stored on the image server, except the node specific information. The nodes are now ready to be rebooted on their local disk.

Cloning Menu

CMU offers a command line interface to manage the cloning mechanism. This menu runs on the image server.

The network boot request and final local reboot are done manually and can be done efficiently by using console broadcast or the boot CMU graphic user interface menu.

Group Management:
The first menu level deals with group function. You can list, create, delete or select a group.

Clients management
When a group is selected, you reach a second level of menu from where you can list, add or delete nodes within the group.

Backup
Using the backup option you create a compressed image of a local disk or a remote node. In the case of a remote node, this command will ask you which node within the group will be used for the clone image preparation. For a local disk, you just have to specify the local device.

Cloning
The cloning command will clone either all or a subset of the group nodes.

Restrictions

Each node within a group needs to have exactly the same hardware configuration.

Concerning Linux: The very first install of a node requires a graphics card installed in the system. Since the compute nodes of a cluster usually do not have a graphic card, the user needs to adjust the configuration and check if the master disk can boot without its graphic adapter.

If the partitioning changes between two cloning operations, target disks are repartitioned according to the master disk partitions. In this case, the node specific information located on previous partitions can be lost during local disk repartitioning.

Notice

The information in this publication is subject to change without notice.

COMPAQ COMPUTER CORPORATION SHALL NOT BE LIABLE FOR TECHNICAL OR EDITORIAL ERRORS OR OMISSIONS CONTAINED HEREIN, NOR FOR INCIDENTAL OR CONSEQUENTIAL DAMAGES RESULTING FROM THE FURNISHING, PERFORMANCE, OR USE OF THIS MATERIAL.