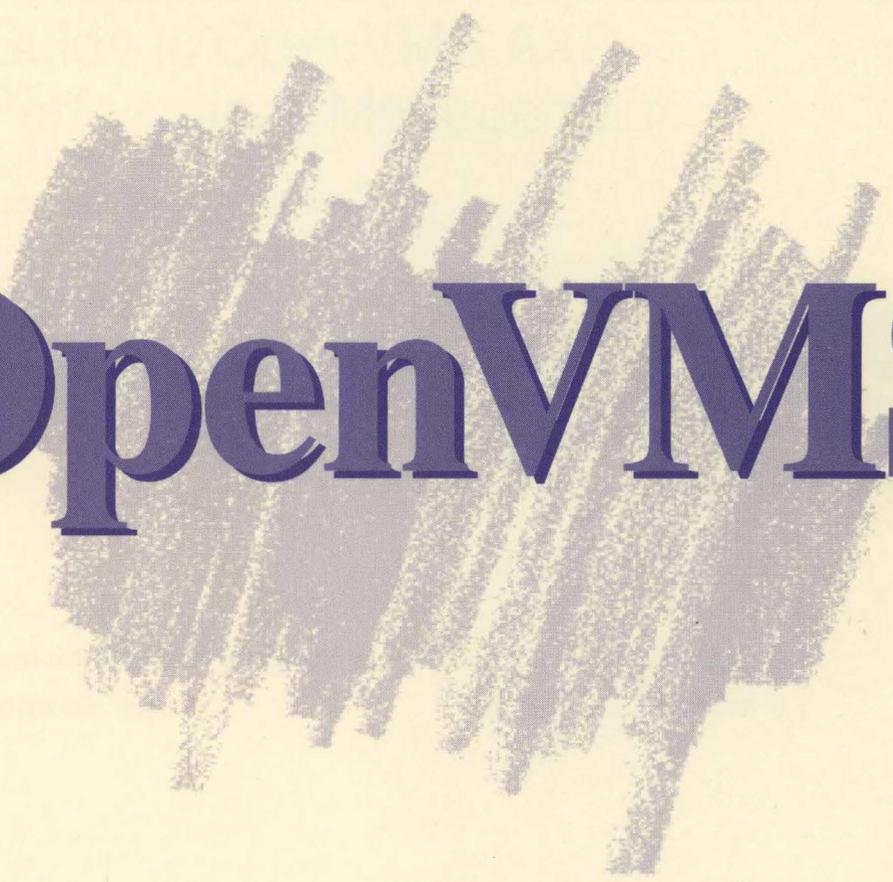

digital

Guide to OpenVMS AXP
Performance Management



OpenVMS

Part Number: AA-Q28WA-TE

Guide to OpenVMS AXP Performance Management

Order Number: AA-Q28WA-TE

March 1994

This manual is a conceptual and tutorial guide for experienced users responsible for optimizing performance on OpenVMS AXP systems.

Revision/Update Information: This is a new manual.

Software Version: OpenVMS AXP Version 6.1

**Digital Equipment Corporation
Maynard, Massachusetts**

March 1994

Digital Equipment Corporation makes no representations that the use of its products in the manner described in this publication will not infringe on existing or future patent rights, nor do the descriptions contained in this publication imply the granting of licenses to make, use, or sell equipment or software in accordance with the description.

Possession, use, or copying of the software described in this publication is authorized only pursuant to a valid written license from Digital or an authorized sublicensor.

© Digital Equipment Corporation 1994. All rights reserved.

The postpaid Reader's Comment form at the end of this document requests your critical evaluation to assist in preparing future documentation.

The following are trademarks of Digital Equipment Corporation: ACMS, Alpha AXP, AXP, Bookreader, CI, DBMS, DECdtm, DECnet, DECram, DECwindows, Digital, HSC, MSCP, OpenVMS, VAX, VAX DOCUMENT, VAXcluster, VMS, VMScluster, and the DIGITAL logo.

The following are third-party trademarks:

Internet is a registered trademark of Internet, Inc.

Motif is a registered trademark of Open Software Foundation, Inc.

All other trademarks and registered trademarks are the property of their respective holders.

ZK6374

This document is available on CD-ROM.

This document was prepared using VAX DOCUMENT Version 2.1.

Send Us Your Comments

We welcome your comments on this or any other OpenVMS manual. If you have suggestions for improving a particular section or find any errors, please indicate the title, order number, chapter, section, and page number (if available). We also welcome more general comments. Your input is valuable in improving future releases of our documentation.

You can send comments to us in the following ways:

- Internet electronic mail: OPENVMSDOC@ZKO.MTS.DEC.COM
- Fax: 603-881-0120 Attn: OpenVMS Documentation, ZK03-4/U08
- A completed Reader's Comments form (postage paid, if mailed in the United States), or a letter, via the postal service. Two Reader's Comments forms are located at the back of each printed OpenVMS manual. Please send letters and forms to:

Digital Equipment Corporation
Information Design and Consulting
OpenVMS Documentation
110 Spit Brook Road, ZK03-4/U08
Nashua, NH 03062-2698
USA

You may also use an online questionnaire to give us feedback. Print or edit the online file `SYS$HELP:OPENVMSDOC_SURVEY.TXT`. Send the completed online file by electronic mail to our Internet address, or send the completed hardcopy survey by fax or through the postal service.

Thank you.

Contents

Preface	xv
1 Performance Management	
Overview	1-1
Purpose	1-1
Definitions	1-1
Strategies and Procedures	1-2
System Manager's Role	1-3
Duties and responsibilities	1-3
Prerequisites	1-3
System utilities and tools	1-3
Why use them?	1-3
Knowing your work load	1-3
Developing a Strategy	1-4
Why have a strategy?	1-4
Three areas of system use	1-4
Managing the work load	1-4
Distributing the work load	1-5
Application code sharing	1-5
2 Investigating Complaints	
Overview	2-1
Purpose	2-1
Definitions	2-1
Analyzing Complaints	2-1
Prerequisites	2-1
Preliminary steps	2-2
Evaluating user complaints	2-2
Hardware problem?	2-2
Blocked process?	2-3
Unrealistic expectations?	2-3
3 Tuning	
Overview	3-1
Purpose	3-1
Definitions	3-1
Tuning to Improve Performance	3-2
Tuning suggestions from Digital	3-2
Prerequisites	3-2
Tools and utilities	3-3
When to use AUTOGEN	3-3

Adjusting system parameters	3-3
Using AUTOGEN feedback	3-3
Evaluating Tuning Success	3-4
Performing a test	3-4
When to stop tuning	3-4
Still not satisfied?	3-5

4 Performance Options

Overview	4-1
Purpose	4-1
Definitions	4-1
Postinstallation System Management Options	4-1
Decompressing system libraries	4-1
Disabling file system high-water marking	4-2
Setting RMS file-extend parameters	4-2
Installing frequently used images	4-2
Enabling virtual I/O caching	4-3
Reducing system disk I/O	4-3

5 Basic Memory Management Concepts

Overview	5-1
Purpose	5-1
Definitions	5-1
Memory	5-2
Pages and pagelets	5-2
Physical memory	5-2
Primary page cache	5-2
Secondary page cache	5-2
Virtual memory	5-2
How memory is configured	5-2
Process execution characteristics	5-3
Working Set Paging	5-4
Working set size	5-4
Upper limit	5-4
What is paging?	5-4
Page faults	5-4
Process Swapping	5-4
What is the swapper?	5-4
What is swapping?	5-5
Types of swapping	5-5

6 Advanced Memory Management Concepts

Overview	6-1
Purpose	6-1
Definitions	6-1
Initial Working Set Limits and Characteristics	6-2
Processes	6-2
Subprocesses and detached processes	6-2
Batch queues	6-2
Interactive and batch processing	6-3
User programs	6-3

Guidelines	6-3
AWSA	6-4
What is AWSA?	6-4
Why use AWSA?	6-4
What are AWSA parameters?	6-4
Default values	6-4
Working set regions	6-4
How does AWSA work?	6-5
Page fault rates	6-6
Voluntary decremting	6-10
Adjusting AWSA parameters	6-11
Caution	6-11
Performance management strategies for tuning AWSA	6-11
Swapper Trimming	6-12
What is swapper trimming?	6-12
First-level trimming	6-13
Second-level trimming	6-13
How are candidates chosen for second-level trimming?	6-13
How are processes chosen for swapping?	6-14
Suspended processes	6-14
Dormant process pseudoclass	6-14
Disabling second-level trimming	6-14
Swapper trimming versus voluntary decremting	6-15
Proactive Memory Reclamation from Idle Processes	6-15
Idle processes	6-15
Reclaiming memory from long-waiting processes	6-15
Reclaiming memory from periodically waking processes	6-16
Setting the FREEGOAL parameter	6-16
Sizing paging and swapping files	6-17
How is the policy enabled?	6-17
Memory Sharing	6-17
What is memory sharing?	6-17
Global pages	6-17
System overhead	6-20
Controlling the overhead	6-20
Installing shareable images	6-20
Verifying memory sharing	6-20
OpenVMS Scheduling	6-21
Scheduling	6-21
Time slicing	6-21
Process state	6-22
Process priority	6-22
Priority boosting	6-22
Scheduling real-time processes	6-22
Tuning	6-23
Processes	6-23
Subprocesses and detached processes	6-23
Batch queues	6-23

7 Evaluating System Resources

Overview	7-1
Purpose	7-1
Resource Management	7-1
What are system resources?	7-1
Tools and utilities	7-2
Prerequisites	7-2
Guidelines	7-2
Collecting and Interpreting Image-Level Accounting Data	7-3
What is image-level accounting?	7-3
Why is it useful?	7-3
Guidelines	7-3
Enabling and disabling image-level accounting	7-4
Generating a report	7-4
Collecting the data	7-4
Interpreting the data	7-5
Creating, Maintaining, and Interpreting MONITOR Summaries	7-7
Guidelines	7-7
Types of output	7-8
MONITOR modes of operation	7-8
Creating a performance information database	7-8
Saving your summary reports	7-9
Customizing your reports	7-9
Report formats	7-9
Using MONITOR in live mode	7-9
More about multifile reports	7-10
Interpreting MONITOR statistics	7-10

8 Managing System Resources

Overview	8-1
Purpose	8-1
Definition	8-1
Understanding System Responsiveness	8-1
Interacting resources	8-1
Overcommitted resources	8-1
Detecting bottlenecks	8-1
Balancing resource capacities	8-2
Evaluating Responsiveness of System Resources	8-2
Using MONITOR	8-2
Measuring system responsiveness	8-2
Improving Responsiveness of System Resources	8-3
Equitable sharing	8-3
Reducing resource consumption	8-3
Load balancing	8-3
Offloading	8-3

9 The CPU Resource

Overview	9-1
Purpose	9-1
Definition	9-1
Evaluating CPU Responsiveness	9-1
Compute queue	9-1
Estimating available CPU capacity	9-3
Types of scheduling wait states	9-3
Voluntary wait states	9-3
Involuntary wait states	9-4
Obtaining MONITOR statistics	9-6
Improving CPU Responsiveness	9-6
Prerequisites	9-6
Equitable CPU sharing	9-7
Reduction of CPU consumption by the system	9-8
CPU offloading	9-12
CPU offloading between processors on the network	9-13
CPU load balancing in a VMSccluster	9-13
Other VMSccluster load-balancing techniques	9-14
Obtaining MONITOR statistics	9-15

10 The Memory Resource

Overview	10-1
Purpose	10-1
Definitions	10-1
Understanding the Memory Resource	10-1
Similarities and differences	10-1
Working set size	10-2
Locality of reference	10-2
Obtaining working set values	10-2
Displaying working set values	10-3
Evaluating Memory Responsiveness	10-5
Memory allocation	10-5
Page faulting	10-5
Swapping and swapper trimming	10-7
Obtaining MONITOR statistics	10-8
Improving Memory Responsiveness	10-9
Equitable memory sharing	10-9
Reduction of memory consumption by the system	10-10
Memory offloading	10-13
Memory load balancing	10-14
Obtaining MONITOR statistics	10-15

11 The Disk I/O Resource

Overview	11-1
Purpose	11-1
Definition	11-1
Understanding the Disk I/O Resource	11-1
Measuring disk I/O responsiveness	11-1
Components of a disk transfer	11-2
Disk capacity and demand	11-2
Evaluating Disk I/O Responsiveness	11-3

Average disk response time	11-3
Improving Disk I/O Responsiveness	11-6
Equitable disk I/O sharing	11-6
Reduction of disk I/O consumption by the system	11-7
Disk I/O offloading	11-10
Disk I/O load balancing	11-11
Obtaining MONITOR statistics	11-13

12 Diagnosing Resource Limitations

Overview	12-1
Purpose	12-1
Definitions	12-1
Diagnostic Strategy	12-1
Getting started	12-1
Investigative procedures	12-2
Methods	12-2
Guidelines	12-2
Road map	12-2
Investigating Resource Limitations	12-2
Memory limitations	12-2
I/O limitations	12-3
CPU limitations	12-3
After the Preliminary Investigation	12-4
Isolating the problem	12-4
Observing the tuned system	12-4
Obtaining a listing of system current values	12-5

13 Isolating Memory Limitations

Overview	13-1
Purpose	13-1
Definition	13-1
Analyzing the Excessive Paging Symptom	13-1
When to investigate	13-1
What is excessive paging?	13-1
Guidelines	13-2
Excessive image activations	13-2
Characterizing hard versus soft faults	13-3
Small total working set size	13-4
Inappropriate WSDEFAULT, WSQUOTA, and WSEXTENT values	13-5
Ineffective borrowing	13-6
AWSA might be disabled	13-7
AWSA is ineffective	13-7
Analyzing the Swapping Symptom	13-9
Swapping versus paging	13-9
Detecting harmful swapping	13-10
Investigating harmful swapping	13-10
Causes of harmful swapping	13-10
Why processes consume unreasonable amounts of memory	13-11
Large, compute-bound processes	13-12
Large waiting processes	13-13
Too many competing processes	13-13
Borrowing is too generous	13-14

Swapper trimming is ineffective	13-14
Excessively large working sets	13-14
Disk thrashing occurs	13-15
System swaps rather than pages	13-17
Demand exceeds available memory	13-17
Analyzing the Limited Free Memory Symptom	13-17
Present capacity versus anticipated demand	13-17
Reallocating memory	13-17

14 Isolating I/O Limitations

Overview	14-1
Purpose	14-1
Definitions	14-1
Disk or Tape Operation Problems (Direct I/O)	14-1
Detecting direct I/O problems	14-1
Software and hardware solutions	14-2
Determining I/O rates	14-2
Device I/O rate is below capacity	14-2
Abnormally high direct I/O rate	14-3
Paging or swapping disk activity	14-4
Reduce I/O demand or add capacity	14-5
Terminal Operation Problems (Buffered I/O)	14-5
Buffered I/O problems	14-5
Detecting terminal I/O problems	14-5
High buffered I/O count	14-6
Operations count	14-6
Excessive kernel mode time	14-6

15 Isolating CPU Limitations

Overview	15-1
Purpose	15-1
Definition	15-1
Detecting CPU Limitations	15-1
Examining the compute queue	15-1
Higher priority blocking processes	15-2
Time slicing between processes	15-2
Excessive interrupt state activity	15-2
Disguised memory limitation	15-3
Operating system overhead	15-3
RMS misused	15-4
CPU at full capacity	15-4
Correcting the problem	15-4

16 Compensating for Resource Limitations

Overview	16-1
Purpose	16-1
Definition	16-1
Changing System Parameters	16-1
Prerequisites	16-1
Guidelines	16-2
Using AUTOGEN	16-2

When to use SYSGEN	16-3
Monitoring the results	16-3

17 Compensating for Memory-Limited Behavior

Overview	17-1
Purpose	17-1
Definition	17-1
Solutions for Memory-Limited Behavior	17-1
Reduce number of image activations	17-1
Increase page cache size	17-2
Decrease page cache size	17-2
Adjust working set characteristics	17-3
Tune to make borrowing more effective	17-5
Tune AWSA to respond quickly	17-6
Disable voluntary decremting	17-7
Tune voluntary decremting	17-7
Turn on voluntary decremting	17-7
Enable AWSA	17-7
Adjust swapper trimming	17-7
Convert to a system that rarely swaps	17-8
Adjust BALSETCNT	17-8
Reduce large page caches	17-9
Curtail large, compute-bound process	17-9
Suspend large, compute-bound process	17-9
Control growth of large, compute-bound processes	17-10
Enable swapping for disk ACPs (ODS-1 only)	17-10
Enable swapping for other processes	17-10
Reduce number of concurrent processes	17-10
Discourage working set loans	17-10
Increase swapper trimming memory reclamation	17-11
Reduce rate of inswapping	17-11
Induce paging to reduce swapping	17-11
Add paging files	17-11
Reduce demand or add memory	17-11

18 Compensating for I/O-Limited Behavior

Overview	18-1
Purpose	18-1
Definition	18-1
Solutions for I/O-Limited Behavior	18-1
Use virtual I/O caching	18-1
Use a RAM disk	18-4
Remove blockage due to ACP	18-4
Improve RMS caching	18-6
Adjust file system caches	18-7

19 Compensating for CPU-Limited Behavior

Overview	19-1
Purpose	19-1
Definition	19-1
Solutions for CPU-Limited Behavior	19-1
Adjust priorities	19-1
Adjust QUANTUM	19-2
Reduce demand or add CPU capacity	19-2

A Decision Trees

B MONITOR Data Items

C MONITOR Multifile Summary Report

Index

Examples

7-1	Image-Level Accounting Report	7-4
10-1	Procedure to Obtain Working Set Information	10-2
10-2	Displaying Working Set Values	10-3

Figures

3-1	Time Spent Tuning Versus Performance Improvements	3-5
5-1	OpenVMS Memory Configuration	5-3
6-1	Working Set Regions for a Process	6-5
6-2	Effect of Working Set Size on Page Fault Rate—Graph 1	6-7
6-3	Effect of Working Set Size on Page Fault Rate—Graph 2	6-8
6-4	Effect of Working Set Size on Page Fault Rate—Graph 3	6-9
6-5	An Example of Working Set Adjustment at Work	6-10
6-6	Example Without Shared Code	6-18
6-7	Example with Shared Code	6-19
A-1	Verifying the Validity of a Performance Complaint	A-1
A-2	Steps in the Preliminary Investigation Process	A-2
A-3	Investigating Excessive Paging—Phase I	A-3
A-4	Investigating Excessive Paging—Phase II	A-4
A-5	Investigating Excessive Paging—Phase III	A-5
A-6	Investigating Excessive Paging—Phase IV	A-6
A-7	Investigating Excessive Paging—Phase V	A-7
A-8	Investigating Swapping—Phase I	A-8
A-9	Investigating Swapping—Phase II	A-9
A-10	Investigating Swapping—Phase III	A-10
A-11	Investigating Limited Free Memory—Phase I	A-11
A-12	Investigating Disk I/O Limitations—Phase I	A-12
A-13	Investigating Disk I/O Limitations—Phase II	A-13

A-14	Investigating Terminal I/O Limitations—Phase I	A-14
A-15	Investigating Terminal I/O Limitations—Phase II	A-15
A-16	Investigating Specific CPU Limitations—Phase I	A-16
A-17	Investigating Specific CPU Limitations—Phase II	A-17
C-1	Prime-Time VMScluster Multifile Summary Report	C-2

Tables

6-1	Parameter MMG_CTLFLAGS Bit Settings	6-17
11-1	Components of a Typical Disk Transfer (Four- to Eight-Block Transfer Size)	11-2
B-1	Summary of Important MONITOR Data Items	B-1

Preface

This manual presents techniques for evaluating, analyzing, and optimizing performance on a system running OpenVMS AXP. Discussions address such wide-ranging concerns as the following:

- Understanding the relationship between work load and system capacity
- Learning to use performance-analysis tools
- Responding to complaints about performance degradation
- Helping the site adopt those programming practices that result in the best system performance
- Using the system features that distribute the work load for better resource utilization
- Knowing when to apply software corrections to system behavior—tuning the system to allocate resources more effectively
- Evaluating the effectiveness of a tuning operation; knowing how to recognize success and when to stop

The manual includes detailed procedures to help you evaluate resource utilization on your system and to diagnose and overcome performance problems resulting from memory limitations, I/O limitations, CPU limitations, human error, or combinations of these. The procedures feature sequential tests that use OpenVMS tools to generate performance data; the accompanying text explains how to evaluate it.

Whenever an investigation uncovers a situation that could benefit from adjusting system values, those adjustments are described in detail, and hints are provided to clarify the interrelationships of certain groups of values. When such adjustments are not the appropriate or available action, other options are defined and discussed.

Decision-tree diagrams summarize the step-by-step descriptions in the text. A decision-tree diagram consists of nodes that describe steps in your performance evaluation. These diagrams should also serve as useful reference tools for subsequent investigations of system performance.

This manual does not describe methods for capacity planning, nor does it attempt to provide details about using OpenVMS RMS features (hereafter referred to as RMS). Refer to the *Guide to OpenVMS File Applications* for that information. Likewise, the manual does not discuss DECnet for OpenVMS performance issues, since the *DECnet for OpenVMS Networking Manual* provides that information.

Intended Audience

This manual addresses system managers and other experienced users responsible for maintaining a consistently high level of system performance, for diagnosing problems on a routine basis, and for taking appropriate remedial action.

Document Structure

This manual is divided into 19 chapters and 3 appendixes, each covering a related group of performance management topics as follows:

- Chapter 1 provides a review of workload management concepts.
- Chapter 2 describes guidelines for evaluating user complaints about system performance.
- Chapter 3 includes a discussion of performance investigation and tuning strategies.
- Chapter 4 lists postinstallation operations for enhancing performance.
- Chapter 5 discusses basic OpenVMS memory management concepts.
- Chapter 6 discusses advanced OpenVMS memory management concepts.
- Chapter 7 explains how to use Digital-supplied utilities and tools to collect and analyze data on your system's hardware and software resources. Included are suggestions for reallocating certain resources should analysis indicate such a need.
- Chapter 8 describes how to evaluate system resource responsiveness.
- Chapter 9 describes how to evaluate the performance of the CPU resource.
- Chapter 10 describes how to evaluate the performance of the memory resource.
- Chapter 11 describes how to evaluate the performance of the disk I/O resource.
- Chapter 12 outlines procedures for investigating performance problems.

- Chapter 13 outlines procedures for investigating performance problems and isolating specific memory resource limitations.
- Chapter 14 outlines procedures for investigating performance problems and isolating specific disk I/O resource limitations.
- Chapter 15 outlines procedures for investigating performance problems and isolating specific CPU resource limitations.
- Chapter 16 provides general recommendations for improving performance with available resources.
- Chapter 17 provides specific recommendations for improving the performance of the memory resource.
- Chapter 18 provides specific recommendations for improving the performance of the disk I/O resource.
- Chapter 19 provides specific recommendations for improving the performance of the CPU resource.
- Appendix A lists the decision trees used in the various performance evaluations described in this manual.
- Appendix B summarizes the MONITOR data items you will find useful in evaluating your system.
- Appendix C provides an example of a MONITOR multifile summary report.

Associated Documents

For additional information on the topics covered in this manual, you can refer to the following documents:

- *OpenVMS System Manager's Manual*
- *Guide to OpenVMS File Applications*
- *OpenVMS System Management Utilities Reference Manual*

Conventions

In this manual, every use of OpenVMS AXP means the OpenVMS AXP operating system.

In this manual, every use of DECwindows and DECwindows Motif refers to DECwindows Motif for OpenVMS software.

The following conventions are also used in this manual:

- | | |
|----------------|---|
| Ctrl/ <i>x</i> | A sequence such as Ctrl/ <i>x</i> indicates that you must hold down the key labeled Ctrl while you press another key or a pointing device button. |
| PF1 <i>x</i> | A sequence such as PF1 <i>x</i> indicates that you must first press and release the key labeled PF1 and then press and release another key or a pointing device button. |

GOLD *x*

A sequence such as GOLD *x* indicates that you must first press and release the key defined as GOLD and then press and release another key. GOLD key sequences can also have a slash (/), dash (-), or underscore (_) as a delimiter in EVE commands.

The GOLD key definition is often mapped to the PF1 key on the keypad.

Return

In examples, a key name enclosed in a box indicates that you press a key on the keyboard. (In text, a key name is not enclosed in a box.)

...

Horizontal ellipsis points in examples indicate one of the following possibilities:

- Additional optional arguments in a statement have been omitted.
- The preceding item or items can be repeated one or more times.
- Additional parameters, values, or other information can be entered.

.
. .
.

Vertical ellipsis points indicate the omission of items from a code example or command format; the items are omitted because they are not important to the topic being discussed.

()

In command format descriptions, parentheses indicate that, if you choose more than one option, you must enclose the choices in parentheses.

[]

In command format descriptions, brackets indicate optional elements. You can choose one, none, or all of the options. (Brackets are not optional, however, in the syntax of a directory name in an OpenVMS file specification or in the syntax of a substring specification in an assignment statement.)

{ }

In command format descriptions, braces surround a required choice of options; you must choose one of the options listed.

boldface text	<p>Boldface text represents the introduction of a new term or the name of an argument, an attribute, or a reason (user action that triggers a callback).</p> <p>Boldface text is also used to show user input in Bookreader versions of the manual.</p>
<i>italic text</i>	<p>Italic text emphasizes important information and indicates complete titles of manuals and variables. Variables include information that varies in system messages (Internal error <i>number</i>), in command lines (<i>/PRODUCER=name</i>), and in command parameters in text (where <i>device-name</i> contains up to five alphanumeric characters).</p>
UPPERCASE TEXT	<p>Uppercase text indicates a command, the name of a routine, the name of a file, or the abbreviation for a system privilege.</p>
struct	<p>Monospace type in text identifies the following C programming language elements: keywords, the names of independently compiled external functions and files, syntax summaries, and references to variables or identifiers introduced in an example.</p> <p>A hyphen in code examples indicates that additional arguments to the request are provided on the line that follows.</p>
numbers	<p>All numbers in text are assumed to be decimal unless otherwise noted. Nondecimal radices—binary, octal, or hexadecimal—are explicitly indicated.</p>

Performance Management

Overview

This manual describes many performance management and tuning issues and problems that system managers might encounter when using an OpenVMS AXP computer system. To help you understand the scope and interrelationship of these issues, this chapter deals with the following topics:

- A review of workload management concepts
- Guidelines for developing a performance management strategy

Managing system performance involves being able to evaluate and coordinate system resources and workload demands.

Purpose

To develop a strategy for evaluating system performance.

Definitions

Performance management means optimizing your hardware and software resources for the current work load. This involves performing the following tasks:

- Acquiring a thorough knowledge of your work load and an understanding of how that work load exercises the system's resources
- Monitoring system behavior on a routine basis in order to determine when and why a given resource is nearing capacity
- Investigating reports of degraded performance from users
- Planning for changes in the system work load or hardware configuration and being prepared to make any necessary adjustments to system values
- Performing certain optional system management operations after installation

A **system resource** is a hardware or software component or subsystem under the direct control of the operating system, which is responsible for data computation or storage. The following subsystems are system resources:

- CPU
- Memory
- Disk I/O

A **throughput rate** is the amount of work accomplished in a given time interval, for example, 100 transactions per second.

Strategies and Procedures

This manual describes several strategies and procedures for evaluating performance, evaluating system resources, and diagnosing resource limitations as shown in the following list:

- Develop workload strategy (Chapter 1)
 - Managing the work load
 - Distributing the work load
 - Sharing application code
- Develop tuning strategy (Chapter 3)
 - AWSA
 - AUTOGEN
 - Proactive memory management
- Perform general system resource evaluation (Chapter 7)
 - CPU resource
 - Memory resource
 - Disk I/O resource
- Review techniques for improving system resource responsiveness (Chapter 8)
 - Equitable sharing of resources
 - Reducing resource consumption
 - Load balancing
 - Offloading
- Conduct a preliminary investigation of specific resource limitations (Chapter 12)
 - Isolating memory resource limitations
 - Isolating disk I/O resource limitations
 - Isolating CPU resource limitations
- Apply specific remedy to compensate for resource limitations (Chapter 16)
 - Compensating for memory-limited behavior
 - Compensating for I/O-limited behavior
 - Compensating for CPU-limited behavior

System Manager's Role

Duties and responsibilities

As a system manager, you must be able to do the following:

- Assume the responsibility for understanding the system's work load sufficiently to be able to recognize normal and abnormal behavior
- Predict the effects of changes in applications, operations, or usage
- Recognize typical throughput rates
- Evaluate system performance
- Perform tuning as needed

Prerequisites

Before you adjust any system parameters, you should:

- Be familiar with system tools and utilities
- Know your work load
- Develop a strategy for evaluating performance

System utilities and tools

You can observe system operation using the following tools:

- Accounting utility (ACCOUNTING)
- Authorize utility (AUTHORIZE)
- AUTOGEN command procedure
- DCL SHOW commands
- Monitor utility (MONITOR)

Why use them?

System utilities and tools allow you to do the following:

- Collect and analyze key data items
- Observe usage trends
- Predict when your system reaches its capacity
- Adjust system parameters
- Modify users' privileges and quotas

Knowing your work load

The experienced system manager can answer the following questions:

- What is the typical number of users on the system at each time of day?
- What is the typical response time for various tasks for this number of users, at each hour of operation?
- What are the peak hours of operation?
- Which jobs typically run at which time of day?

- Which commonly run jobs are intensive consumers of the CPU? Of memory? Of disk?
- Which applications involve the most image activations?
- Which parts of the system software, if any, have been modified or user written, such as device drivers?
- Are there any known system bottlenecks? Are there any anticipated ones?

Note

If you are a novice system manager, you should spend a considerable amount of time observing system operation using ACCOUNTING, MONITOR, and DCL SHOW commands.

Developing a Strategy

Why have a strategy?

A strategy is your plan for evaluating system performance.

Three areas of system use

Each installation site must develop its own strategy for optimizing system performance. Such a strategy requires knowledge about system use in the following areas:

- Managing the work load
- Distributing the work load
- Application code sharing

Managing the work load

Before you attempt to adjust any system values, always ask yourself the following questions:

- Is there a time of day when the work load peaks, that is, when it is noticeably heavier than at other times?
- Is there any way to balance the work load better? Perhaps measures can be adopted by users.
- Could any jobs be run better as batch jobs, preferably during nonpeak hours?
- Have primary and secondary hours of operation been employed with users?
- Can future applications be designed to work around any known or expected system bottlenecks? Can present applications be redesigned for the same purpose?

- Are you using all of the code-sharing features that the OpenVMS AXP system offers you?

Note

Do not adjust any system values until you are satisfied that all these issues are resolved and that your workload management strategy is correct.

Distributing the work load

Distribute the work load as evenly as possible using the following techniques:

- Run large jobs as batch jobs
 - Establish a site policy that encourages the submission of large jobs on a batch basis
 - Regulate the number of batch streams so that batch usage is high when interactive usage is low
 - Use DCL command qualifiers to run batch jobs at lower priority, adjust the working set sizes, and control the number of concurrent jobs
- Restrict system use
 - Do not permit more users to log in at one time than the system can support with adequate response time
 - Restrict the number of interactive users with the DCL command SET LOGINS/INTERACTIVE
 - Control the number of concurrent processes with the MAXPROCESSCNT system parameter
 - Control the number of remote terminals allowed to access the system at one time with the RJOBLIM system parameter
 - Restrict system use to groups of users to certain days and hours of the day
- Design applications to reduce demand on binding resources
 - Find out where system bottlenecks are
 - Plan use that minimizes demands on the bottleneck points

Application code sharing

Application code sharing provides a cost-effective means of optimizing memory utilization. To ensure optimum performance of your system, make sure that frequently used code is shared.

Use the site-specific startup procedure to install user-written programs and routines as known images that are designed for sharing and have reached production status or are in general use.

Encourage programmers to write shareable code.

Investigating Complaints

Overview

Typically, an investigation into system performance begins when you receive a complaint about a slowdown of interactive response times or about some other symptom of decreased throughput. This chapter presents the following topics:

- Guidelines for evaluating user complaints about system performance
- A discussion of performance investigation and system tuning strategies
- A checklist of system management options

Purpose

The purpose of this chapter is as follows:

- To provide a methodology for evaluating user complaints
- To describe the preliminary stages of a performance investigation
- To discuss tuning and when it is required

Definitions

A **semaphore** is a synchronization tool that is used to control exclusive access to a shared database or other resource. It ensures that only one process at a time is within the critical region of code that accesses the resource.

A process in the miscellaneous resource wait (**MWAIT**) state is blocked either by a miscellaneous resource wait or a mutual exclusion semaphore (**MUTEX**).

Analyzing Complaints

Prerequisites

Before you decide that the current complaint reflects a performance problem, you should:

- Be convinced that hardware resources are adequate
- Know the work load reasonably well

- Have been managing the work load according the guidelines in Chapter 1, Knowing your work load.

Preliminary steps

You will need some additional information as described in the following table:

Step	Action
1	Obtain the following information: <ul style="list-style-type: none"> • Number of users on the system at the time the problem occurred • Response times • Evidence of jobs hung and unable to complete
2	Compare these facts with your knowledge of the normal work load and operation of your system.
3	Follow the procedure shown in Figure A-1.
4	Did you observe the problem? Can you duplicate the problem?

Evaluating user complaints

Use Figure A-1 to verify the validity of a performance complaint.

Hardware problem?

Hardware problems are a common source of performance complaints.

Step	Action
1	Check the operator log and error log for indications of problems with specific devices.
2	Enter the DCL commands SHOW ERROR and ANALYZE/ERROR_LOG to help determine if hardware is contributing to a performance problem.
3	Review the previous day's error log as part of your morning routine.
4	Obtain a count of errors logged since yesterday.

Note

To obtain a count of errors logged, use the following DCL command:

```
$ ANALYZE/ERROR LOG/BRIEF/LOG -
_ $ /OUTPUT=DAILY.LOG/SINCE=YESTERDAY
```

Blocked process?

A process enters the miscellaneous resource wait (MWAIT) state usually because some resource, such as paging file or mailbox, is unavailable.

IF...	THEN...
the process entered the MWAIT state	do the following: <ul style="list-style-type: none"> • enter DCL command MONITOR STATES. • look for processes in the MWAIT state.
the system fails to respond while you are investigating an MWAIT condition	check the system console for error messages.
the MWAIT condition persists after you increase the capacity of the appropriate resource	investigate the possibility of a programming design error.

Unrealistic expectations?

Always bear in mind that what appears to be a performance problem at first can turn out to be a case of unrealistic expectations. For example:

- Users expect response times to remain constant, even as the system work load increases.
- An unusual set of circumstances has caused exceptionally high demand on the system all at once.

Adjusting system values will accomplish nothing in such circumstances.

Note

Whenever you can anticipate a temporary workload change that will affect your users, you should notify them through broadcasts, text, or both, in the system notices.

Overview

Generally, system performance problems are the result of poor operation, lack of understanding of the work load and its operational ramifications, lack of resources, poor application design, human error, or a combination of these factors. This chapter discusses the following topics:

- Tuning
- Performance options available to the system manager

You will rarely need to make major adjustments to system parameters.

Purpose

To determine when tuning a system is appropriate.

Definitions

AUTOGEN is a Digital-supplied command procedure that establishes initial values for all the configuration-dependent system parameters so that they match your particular configuration.

Tuning is the process of altering various system values to obtain the optimum *overall* performance possible from any given configuration and work load.

Note

Tuning does not include the acquisition and installation of additional memory or devices, although in many cases such additions (when made at the appropriate time) can vastly improve system operation and performance.

Tuning to Improve Performance

Tuning suggestions from Digital

Digital believes that tuning is rarely required for its systems for the following reasons:

- The system includes AUTOGEN.
- The system includes features that in a limited way permit it to adjust itself dynamically during operation. The system can detect the need for adjustment in the following areas:
 - Nonpaged dynamic pool
 - Working set size
 - Number of pages on the free- and modified-page lists

As a result, these areas can grow dynamically, as appropriate, during normal operation.

- Experience has shown that the most common cause of disappointment in system performance is insufficient hardware capacity.

Prerequisites

Before you undertake any action, you must recognize that the following sources of performance problems cannot be cured by adjusting system values:

- Improper operation
- Unreasonable performance expectations
- Insufficient memory for the applications attempted
- Inadequate hardware configuration for the work load, such as too slow a processor, too few buses for the devices, too few disks, and so forth
- Improper device choices for the work load, such as using disks with insufficient speed or capacity
- Hardware malfunctions
- Human error, such as poor application design or allowing one process to consume all available resources

Note

Always aim for best overall performance, that is, performance viewed over time. The work load is constantly changing on most systems. Therefore, what guarantees optimal workload performance at one time might not produce optimal performance a short time later as the work load changes.

Tools and utilities

Digital recommends that you use the AUTOGEN command procedure to manage your system parameters. See the *OpenVMS System Manager's Manual* for a detailed description of the AUTOGEN command procedure.

Use AUTHORIZE to change user account information, quotas, and privileges.

When to use AUTOGEN

Digital recommends running AUTOGEN in the following circumstances:

- During a new installation or upgrade
- Whenever your work load changes significantly
- When you add an optional (layered) software product
- When you install images with the /SHARED qualifier
- On a regular basis to monitor changes in your system's work load
- When you adjust system parameters

AUTOGEN will not fix a resource limitation.

Adjusting system parameters

When it becomes necessary to make adjustments, you normally select a very small number of values for change, based on a careful analysis of the behavior observed.

IF you want to...	THEN...
modify system parameters	use the AUTOGEN command procedure.
change entries in the UAF	use AUTHORIZE.

Using AUTOGEN feedback

AUTOGEN has special features that allow it to make automatic adjustments for you in associated parameters. Periodically running AUTOGEN in **feedback mode** can ensure that the system is optimally tuned.

The operating system keeps track of resource shortages in subsystems where resource expansion occurs. AUTOGEN in feedback mode uses this data to perform tuning.

Evaluating Tuning Success

Performing a test

Whenever you make adjustments to your system, you must spend time monitoring its behavior afterward to ensure that you obtain the desired results. Use the following procedure to evaluate how successful your tuning was:

Step	Action
1	Run a few programs that produce fixed and reproducible results at the same time you run your normal work load.
2	Measure the running times.
3	Adjust system values.
4	Run the programs again at the same time you run your normal work load under nearly identical conditions as those in step 1.
5	Measure the running times under nearly identical workload conditions.
6	Compare the results.
7	Continue to observe system behavior closely for a time after you make any changes.

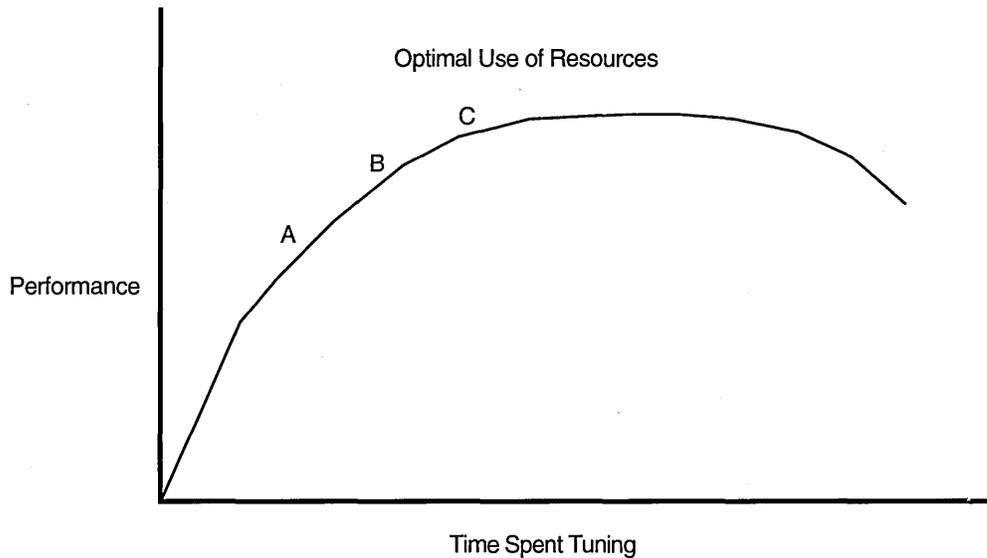
Note

This test alone does not provide conclusive proof of success. There is always the possibility that your adjustments have favored the performance of the image you are measuring—to the detriment of others.

When to stop tuning

In every effort to improve system performance, there comes a point of diminishing returns. In other words, you will find that once you obtain a certain level of improvement, you can spend a great deal of time tuning the system beyond that point and achieve only marginal improvement. Figure 3-1 illustrates this pattern.

Figure 3-1 Time Spent Tuning Versus Performance Improvements



ZK-1118-GE

As a guideline, if you make adjustments and see a marked improvement, make more adjustments and see about half as much improvement, and then fail to make more than a small improvement on your next attempt or two, you should stop and evaluate the situation. You can probably assume you have done your best and you are close to point C on the graph. In most situations, this is the point at which to stop tuning.

Still not satisfied?

If you are not satisfied with the final performance, consider increasing your capacity through the addition of hardware.

Generally, memory is the single piece of hardware needed to solve the problem. However, some situations warrant obtaining additional disks or more CPU power.

Very few situations warrant the expense of continued adjustments for minimal potential improvement once the improvement depicted at point C has been obtained.

Performance Options

Overview

Optional system management operations, normally performed after installation, often result in improved overall performance. This chapter discusses the following topics:

- Decompressing system libraries
- Disabling file system high-water marking
- Setting RMS file-extend parameters
- Installing frequently used images
- Reducing system disk I/O

Note, however, that not all options are appropriate at every site.

Purpose

To present additional ways to increase system throughput.

Definitions

High-water marking is a security feature that guarantees that users cannot read data they have not written. It is implemented by erasing the previous contents of the disk blocks allocated every time a file is created or extended.

An **I/O operation** is the process of requesting a transfer of data from a peripheral device to memory (or vice versa), the actual transfer of the data, and the processing and overlaying activity to make both of those events happen.

The **multiblock count** is the number of blocks that RMS moves in and out of the I/O buffer during each I/O operation for a sequential file.

The **multibuffer count** is the number of buffers RMS uses to perform an I/O operation.

Postinstallation System Management Options

Decompressing system libraries

Most of the OpenVMS AXP libraries are in compressed format in order to conserve disk space.

The CPU dynamically decompresses the libraries whenever they are accessed. However, the resulting performance slowdown is especially noticeable during link operations and when requesting online help.

To decompress the libraries, invoke the command procedure `SYS$UPDATE:LIBDECOMP.COM`.

Note

Decompressed object libraries take up about 25 percent more disk space than when compressed; the decompressed help libraries take up about 50 percent more disk space.

Disabling file system high-water marking

High-water marking is set by default whenever a volume is initialized. Disabling the feature depends on the following considerations:

- How often new files are created
- How often existing files are extended
- How fragmented the volume is

To disable high-water marking, specify the `/NOHIGHWATER_MARKING` qualifier when initializing the volume or do the following at any time:

1. Enter a DCL command similar to the following:

```
$ SET VOLUME/NOHIGHWATER_MARKING device-spec[:]
```
2. Dismount and remount the volume.

Setting RMS file-extend parameters

Because files extend in increments of twice the multiblock count (default is 16), system defaults now provide file extensions of only 32 blocks. Thus, when files are created or extended, increased I/O can slow performance. The problem can be overcome by doing the following:

- Specifying larger values for system file-extend parameters
- Setting the system parameter `RMS_EXTEND_SIZE`
- Specifying a larger multiblock count
- Specifying a larger multibuffer count

Installing frequently used images

When an image is used concurrently by more than one process on a routine basis, install the image with the Install utility (`INSTALL`), specifying the `/OPEN`, `/SHARED`, and `/HEADER_RESIDENT` qualifiers. You will ensure the following:

- All processes use the same physical copy of the image
- The image will be activated in the most efficient way

Enabling virtual I/O caching

Enable virtual I/O caching to reduce the number of disk I/O operations.

Reducing system disk I/O

Remove frequently accessed files from the system disk and use logical names, or where necessary, use other pointers to access them as shown in the following table:

Logical Name	File
AUDIT_SERVER	Audit server master file
QMAN\$MASTER	Job queue database master file ¹
Directory specification ²	Job queue database queue and journal files
NETPROXY	NETPROXY.DAT
OPC\$LOGFILE_NAME	Operator log files
RIGHTSLIST	RIGHTSLIST.DAT
SYS\$errorlog	ERRFMT log files
SYS\$JOURNAL	DECdtm transaction log files
SYS\$MONITOR	MONITOR log files
SYSUAF	SYSUAF.DAT
VMSMAIL_PROFILE	VMSMAIL_PROFILE.DATA

¹Mount the disk on which it resides in SYLOGICALS.

²When used with the DCL command START/QUEUE/MANAGER.

Basic Memory Management Concepts

Overview

Once you have taken the necessary steps to manage your work load, you can evaluate reports of performance problems. Before you proceed, however, it is imperative that you be well versed in the concepts of resource management.

If you lack such an understanding, you are likely to encounter unnecessary problems in your tuning attempts.

Purpose

To develop an understanding of basic memory management concepts.

Definitions

A **balance set** is the sum of all working sets currently in physical memory.

An **image** is a set of procedures and data bound together by the linker.

A **page** is either an 8 KB, 16 KB, 32 KB, or 64 KB segment of virtual address space.

A **pagelet** is a 512-byte unit of memory. One AXP pagelet is the same size as one VAX page. On an OpenVMS AXP 16 KB system, 32 AXP pagelets equal 1 AXP page.

A **process** is the basic entity that is scheduled by the system. It provides the context in which an image executes.

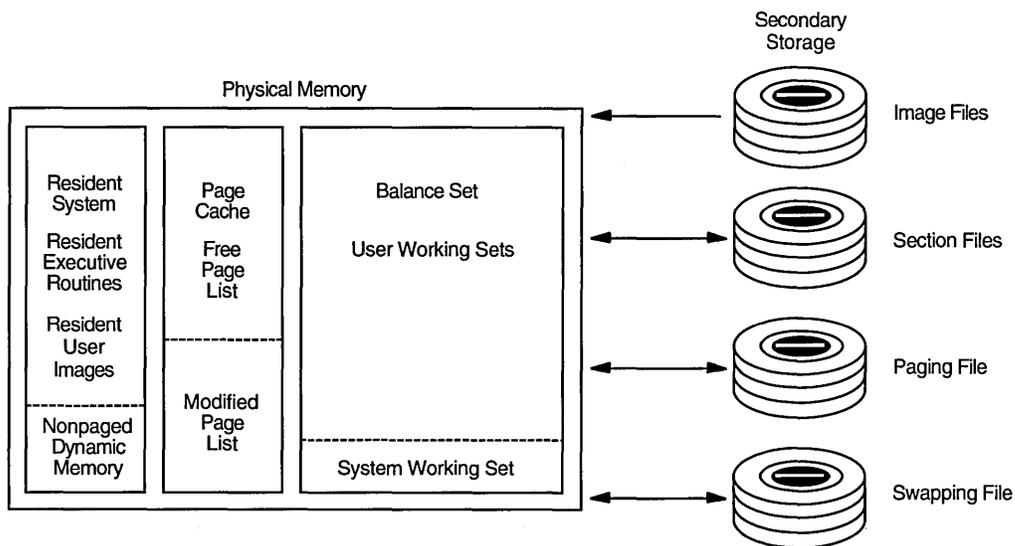
The **swapper** schedules physical memory. It keeps track of the pages in both physical memory and on the disk paging and swapping files so it can ensure that each process has a steady supply of pages for each job.

A **working set**, also called the primary page cache, is the total number of a process's pages in physical memory. It is a subset of the total number of pages allocated to a process.

Memory

Pages and pagelets	On OpenVMS AXP systems, some system parameter values are allocated in units of pages, while others are allocated in units of pagelets.
Physical memory	<p>Physical memory consists of three major parts according to use:</p> <ul style="list-style-type: none">• Primary page cache where processes execute• Secondary page cache where data is stored for movement to and from the disks• Operating system resident executive <p>Each disk has only one access path available to transfer data from and to physical memory, that is, to perform disk I/O.</p>
Primary page cache	The balance set resides in the primary page cache.
Secondary page cache	<p>The secondary page cache consists of two sections as follows:</p> <ul style="list-style-type: none">• Free-page list—Pages whose contents have not been modified• Modified-page list—Pages whose contents have been modified
Virtual memory	<p>Virtual memory is the set of storage locations in:</p> <ul style="list-style-type: none">• Physical memory• Secondary storage (disk) <p>From the programmer's point of view, the secondary storage locations appear to be locations in physical memory.</p>
How memory is configured	Figure 5-1 illustrates the configuration of memory for OpenVMS systems.

Figure 5-1 OpenVMS Memory Configuration



ZK-1009-GE

Process execution characteristics

Processes executing on a general timesharing system use CPU time and memory in the following manner:

- A process executes in physical memory until it must wait—usually for the completion of an I/O request.
- Every time a process has to wait, another process may use the CPU.
- Multiprogramming allows the system to keep more than one process in memory at one time.
- The operating system maintains an even balance in the use of memory, CPU time, and the number of processes running at once.
- Each process has an available amount of time to perform its work—called its **quantum**.

Note

If no other process is waiting to exercise its quantum, the current process can keep renewing its quantum and retain control of the CPU.

Working Set Paging

Working set size The initial size of a process's working set is defined (in pagelets) by the process's working set default quota `WSDEFAULT`.

Upper limit When ample memory is available, a process's working set upper growth limit can be expanded to its working set extent, `WSEXTENT`.

What is paging? The exchange of pages between physical memory and secondary storage is called **paging**. The following table lists conditions under which paging can occur:

When...	Then...
image activation begins	the process brings in the first set of pages from the image file and uses them in its own working set.
the process's demand for pages exceeds those available in the working set	some of the process's pages must be moved to the page cache to make room or the process's working set is expanded.
the page cache fills up	the swapper transfers a cluster of pages from the modified-page cache to a paging file.

Page faults A **hard** fault requires a read operation from a page or image file on disk.

A **soft** fault involves mapping to a page already in memory; this can be a global page or a page in the secondary page cache.

Process Swapping

What is the swapper? The swapper process schedules physical memory. It keeps track of the pages in both physical memory and on the disk paging and swapping files so it can ensure that each process has a steady supply of pages for each job.

What is swapping?

When a process whose working set is in memory becomes inactive, the entire working set or part of it may be removed from memory to provide space for another process's working set to be brought in for execution.

Swapping is the partial or total removal of a process's working set from memory.

Types of swapping

A process's working set can be removed from memory using either of the following techniques:

- Swapper trimming—Pages are removed from the target working set but the working set is not swapped out.
- Process swapping—All pages are swapped out of memory.

Advanced Memory Management Concepts

Overview

The operating system employs several memory management mechanisms and a memory reclamation policy to improve performance on the system. This chapter discusses the following topics:

- Automatic working set adjustment (AWSA)
- Swapper trimming
- Proactive memory reclamation policy
- Memory sharing
- Memory scheduling

The operating system uses the memory management mechanisms to implement its policy of memory reclamation.

The operating system, as provided by Digital, enables these features by default. In the majority of situations, they produce highly desirable results in optimizing system performance. However, under rare circumstances, they might contribute to performance degradation by incurring their own overhead. The following sections describe these features and provide insight into how to adjust, or even turn them off, through tuning.

Purpose

To describe how the operating system uses its advanced features to dynamically manage memory.

Definitions

The **adjustment period** is the time from the start of quantum right after an adjustment occurs until the next quantum after the time specified by the `AWSTIME` parameter elapses as shown in the following equation:

$$\text{adjustment period} = \text{QUANTUM} + \text{AWSTIME}$$

Context switching involves interrupting the activity in progress and switching to another activity. Context switching occurs as one process after another is scheduled for execution.

The **scheduler** controls both when and how long a process executes.

The **working set count** is the actual number of pages the working set requires. It consists of the process's pages plus any global pages the process uses.

Initial Working Set Limits and Characteristics

The memory management strategy depends initially on the values in effect for the working set quota (WSQUOTA) and working set extent limit (WSEXTENT).

Processes

The working set characteristics of processes are derived from the following:

- User authorization file (UAF) record created by the system manager.
- UAF record using system-assigned default values in the DEFAULT record. The AUTHORIZE command SHOW DEFAULT displays the default values.

When an interactive job runs, the values in effect might have been lowered or raised either by the corresponding qualifiers on the last SET WORKING_SET command to affect them, or by the system service \$ADJWSL.

Subprocesses and detached processes

Subprocesses and detached processes derive their working set characteristics from one of the following:

- \$CREPRC system service
- DCL command RUN

If characteristics are not specified by either of the above, then the values of the corresponding process quota and creation limit (PQL) system parameters are used as shown in the following table:

Parameter	Characteristic
PQL_DWSDEFAULT	Default WSDEFAULT
PQL_DWSQUOTA	Default WSQUOTA
PQL_DWSEXTENT	Default WSEXTENT

Batch queues

When a batch queue is created, the DCL command INITIALIZE /QUEUE establishes the default values for jobs with the /WSDEFAULT, /WSQUOTA, and /WSEXTENT qualifiers.

These qualifiers can, however, be set to defer to the user's values in the UAF record.

When a batch job runs, the values in effect might have been lowered by the corresponding qualifiers on the DCL commands `SUBMIT` or `SET QUEUE/ENTRY`.

Interactive and batch processing

For processing that involves system components, the following working set limits are suggested:

- Small (16 to 50 pages)—For editing, and for compiling and linking small programs (typical interactive processing)
- Large (64 pages or more)—For compiling and linking large programs, and for executing programs that manipulate large amounts of data in memory (typical batch processing)

User programs

Working set limits for user programs depend on the code-to-data ratio of the program and on the amount of data in the program.

Programs that manipulate mostly code and that include small or moderate amounts of data or use RMS require a small working set.

Programs that manipulate mostly data such as sort procedures, compilers, linkers, assemblers, and librarians require a large working set.

Guidelines

The following guidelines are suggested for initial working set characteristics:

- System parameters—Set `WSMAX` at the highest number of pages required by any program.
- UAF options
 - Set `WSDEFAULT` at the median number of pages required by a program that the user will run interactively.
 - Set `WSQUOTA` at the largest number of pages required by a program that the user will run interactively.
 - Set `WSEXTENT` at the largest number of pages you anticipate the process will need. Be as realistic as possible in your estimate.
- Batch queues for user-submitted jobs
 - Set `WSDEFAULT` at the median number of pages required.
 - Set `WSQUOTA` to the number of pages that will allow the jobs to complete within a reasonable amount of time.
 - Set `WSEXTENT` (using the DCL command `INITIALIZE/QUEUE` or `START/QUEUE`) to the largest number of pages required.

This arrangement effectively forces users to submit large jobs for batch processing because otherwise, the jobs will not run efficiently and interactively. To further restrict the user who

attempts to run a large job interactively, you can impose CPU time limits in the UAF.

AWSA

What is AWSA? The **automatic working set adjustment** (AWSA) feature refers to a system where processes can acquire additional working set space (physical memory) under control of the operating system. The system recognizes the amount of page faulting that is occurring for each process and factors this into the operation.

Why use AWSA? The goal of this activity is to reduce the amount of page faulting. By reviewing the need for each process to add some pages to its working set limit through the AWSA feature, the operating system can better balance the working set space allocation among processes.

What are AWSA parameters? The AWSA mechanism depends heavily on the values of the key system parameters: PFRATH, PFRATL, WSINC, WSDEC, QUANTUM, AWSTIME, AWSMIN, GROWLIM, and BORROWLIM. Normally, the default values that the system provides for these parameters correctly match the operational needs.

Note

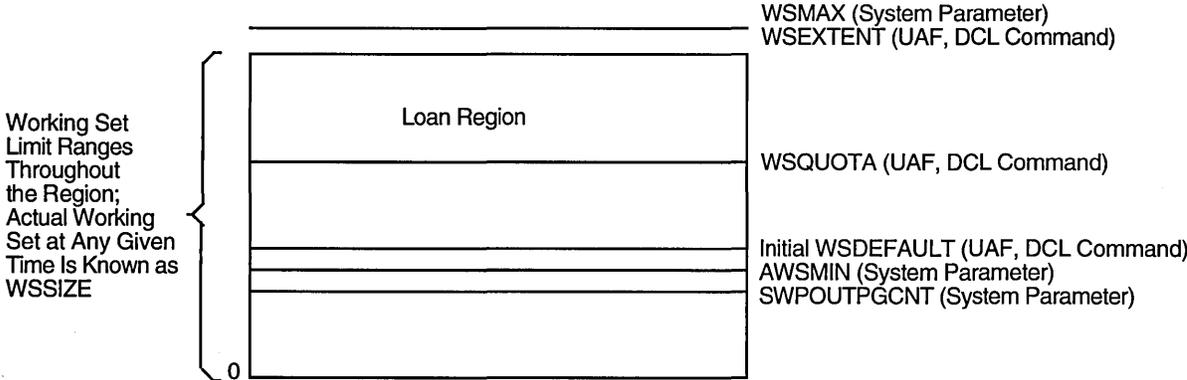
The possibility that AWSA parameters are out of balance is so slight that you should not attempt to modify any of the key parameter values without a very thorough understanding of the entire mechanism.

Default values All processes have an initial default limit of pages of physical memory defined by the system parameter WSDEFAULT.

Working set regions Any process that needs more memory is allowed to expand to the amount of a larger limit known as the **working set quota** defined by WSQUOTA.

Whenever a process's working set size increases, the growth occurs in increments according to the value of the system parameter, WSINC. Figure 6-1 illustrates these important regions.

Figure 6-1 Working Set Regions for a Process



ZK-1121-GE

How does AWSA work?

The following table summarizes how AWSA works:

Stage	Description						
1	The system samples the page faulting rate of each process during the adjustment period.						
2	At the end of each process's adjustment period, the system reviews the need for growth and does the following:						
	<table border="1"> <thead> <tr> <th>IF the page fault rate is...</th> <th>THEN the system...</th> </tr> </thead> <tbody> <tr> <td>too high compared with PFRATH</td> <td>approves an increase in the working set size of that process in the amount of system parameter WSINC up to the value of its WSQUOTA.</td> </tr> <tr> <td>too low compared with PFRATL (when PFRATL is nonzero)</td> <td>approves an decrease in the working set size of that process in the amount of system parameter WSDEC. No process will be reduced below the size defined by AWSMIN.</td> </tr> </tbody> </table>	IF the page fault rate is...	THEN the system...	too high compared with PFRATH	approves an increase in the working set size of that process in the amount of system parameter WSINC up to the value of its WSQUOTA.	too low compared with PFRATL (when PFRATL is nonzero)	approves an decrease in the working set size of that process in the amount of system parameter WSDEC. No process will be reduced below the size defined by AWSMIN.
IF the page fault rate is...	THEN the system...						
too high compared with PFRATH	approves an increase in the working set size of that process in the amount of system parameter WSINC up to the value of its WSQUOTA.						
too low compared with PFRATL (when PFRATL is nonzero)	approves an decrease in the working set size of that process in the amount of system parameter WSDEC. No process will be reduced below the size defined by AWSMIN.						

The system parameters PFRATL and PFRATH define the upper and lower limits of acceptable page faulting for all processes.

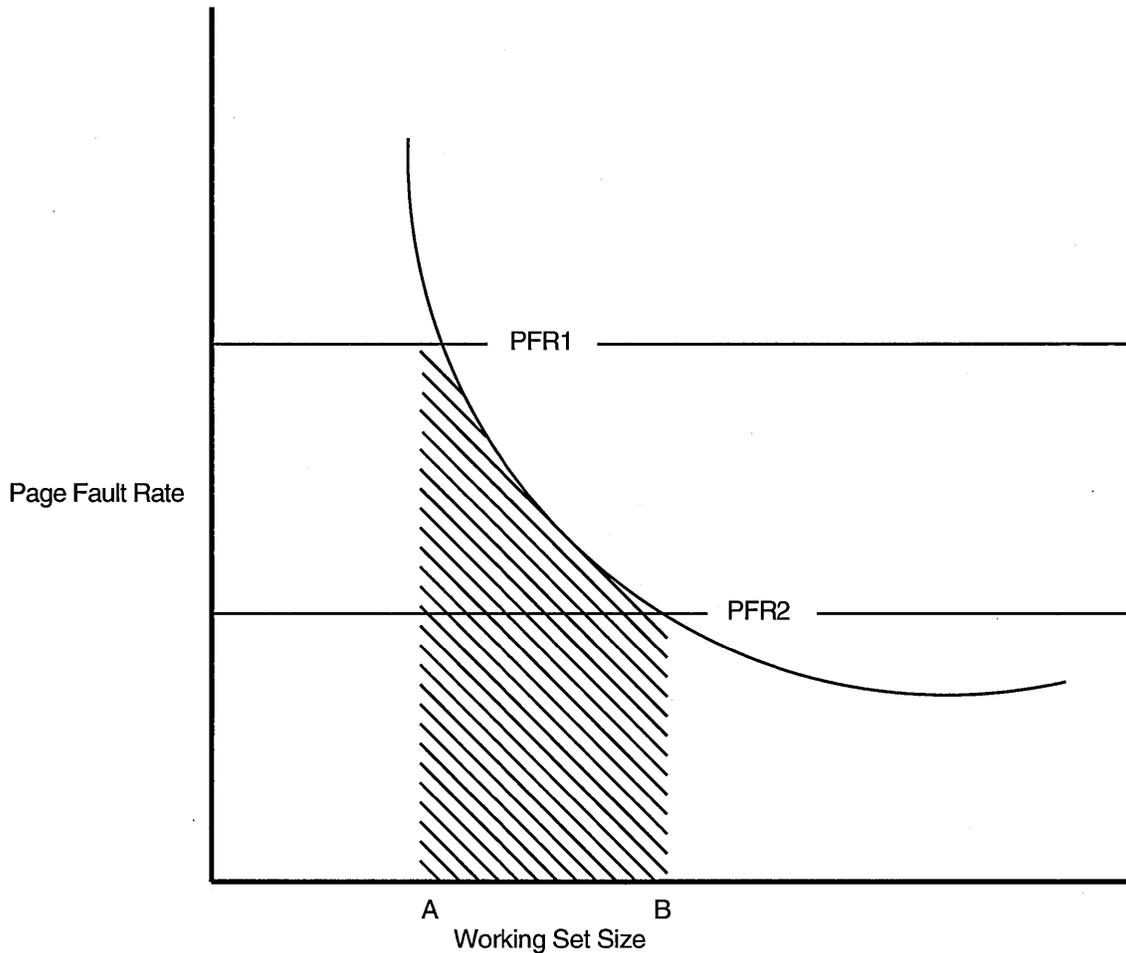
Stage	Description												
3	<p>If too many processes attempt to add pages at once, an additional mechanism is needed to stop the growth while it is occurring. However, the system only stops the growth of processes that have already had the benefit of growing beyond their quota.</p>												
	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">IF...</th> <th style="text-align: left;">THEN the system...</th> </tr> </thead> <tbody> <tr> <td>the increase in working set size puts the process above the value of WSQUOTA and thus requires a loan</td> <td>compares the availability of free memory with the value of BORROWLIM. The AWSA feature allows a process to grow above its WSQUOTA value only if there are at least as many pages of free memory as specified by BORROWLIM.</td> </tr> <tr> <td>a process page faults after its working set count exceeds WSQUOTA</td> <td>examines the value of the parameter GROWLIM before it allows the process to use more of its WSINC loan. Note that this activity is not tied into an adjustment period but is an event-driven occurrence based on page faulting.</td> </tr> <tr> <td>the number of pages on the free-page list is at least equal to or greater than GROWLIM</td> <td>continues to allow the process to add pages to its working set.</td> </tr> <tr> <td>the number of free pages is less than GROWLIM</td> <td>will not allow the process to grow; the process must give back some of its pages before it reads in new pages.</td> </tr> <tr> <td>proactive memory reclamation is enabled</td> <td>both BORROWLIM and GROWLIM are set to very small values to allow active processes maximum growth potential.</td> </tr> </tbody> </table>	IF...	THEN the system...	the increase in working set size puts the process above the value of WSQUOTA and thus requires a loan	compares the availability of free memory with the value of BORROWLIM. The AWSA feature allows a process to grow above its WSQUOTA value only if there are at least as many pages of free memory as specified by BORROWLIM.	a process page faults after its working set count exceeds WSQUOTA	examines the value of the parameter GROWLIM before it allows the process to use more of its WSINC loan. Note that this activity is not tied into an adjustment period but is an event-driven occurrence based on page faulting.	the number of pages on the free-page list is at least equal to or greater than GROWLIM	continues to allow the process to add pages to its working set.	the number of free pages is less than GROWLIM	will not allow the process to grow; the process must give back some of its pages before it reads in new pages.	proactive memory reclamation is enabled	both BORROWLIM and GROWLIM are set to very small values to allow active processes maximum growth potential.
IF...	THEN the system...												
the increase in working set size puts the process above the value of WSQUOTA and thus requires a loan	compares the availability of free memory with the value of BORROWLIM. The AWSA feature allows a process to grow above its WSQUOTA value only if there are at least as many pages of free memory as specified by BORROWLIM.												
a process page faults after its working set count exceeds WSQUOTA	examines the value of the parameter GROWLIM before it allows the process to use more of its WSINC loan. Note that this activity is not tied into an adjustment period but is an event-driven occurrence based on page faulting.												
the number of pages on the free-page list is at least equal to or greater than GROWLIM	continues to allow the process to add pages to its working set.												
the number of free pages is less than GROWLIM	will not allow the process to grow; the process must give back some of its pages before it reads in new pages.												
proactive memory reclamation is enabled	both BORROWLIM and GROWLIM are set to very small values to allow active processes maximum growth potential.												

Page fault rates

Each of the characteristic curves illustrates that as you decrease the working set size, you should expect the page fault rate to increase.

Figure 6-2 illustrates how the page fault rate and working set size are related for most processes.

Figure 6-2 Effect of Working Set Size on Page Fault Rate—Graph 1



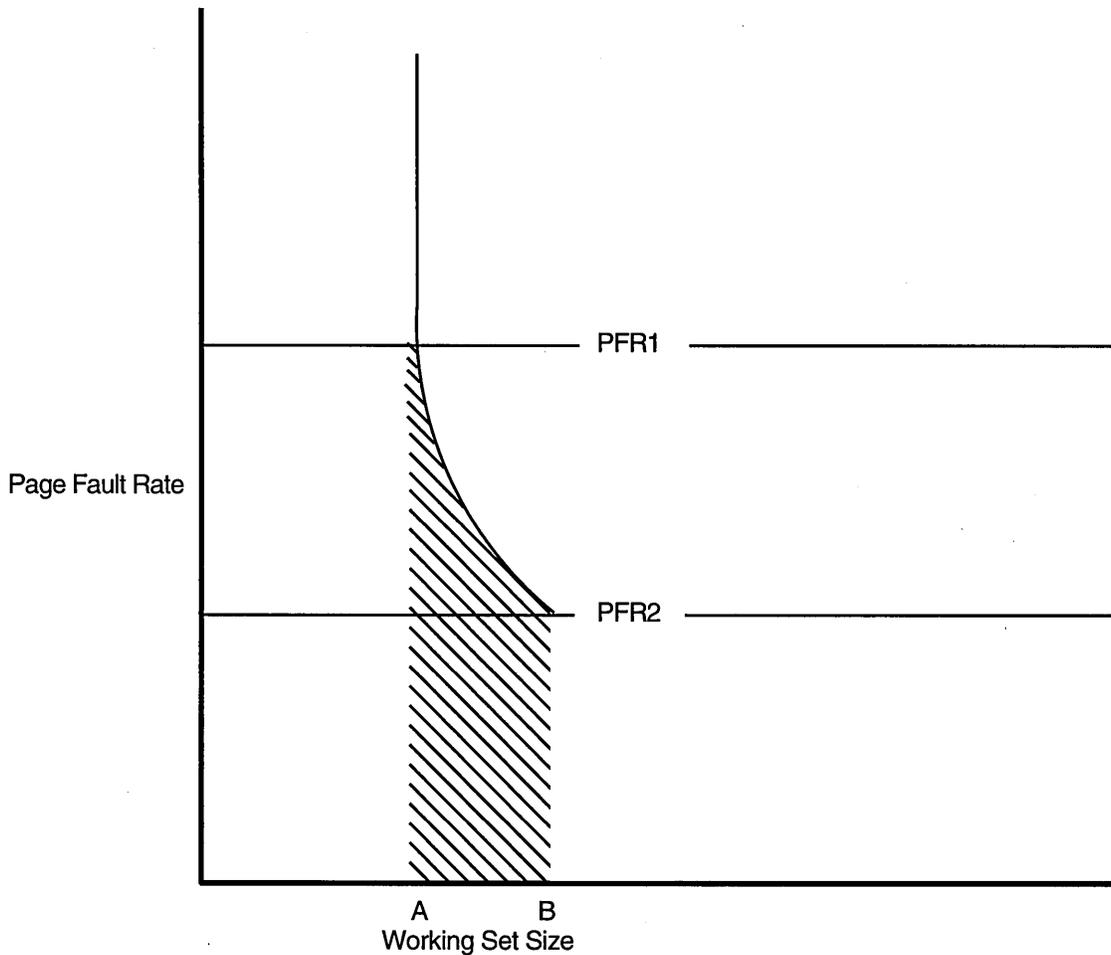
ZK-1139-GE

Not all working sets for all images exhibit the same curve as depicted in Figure 6-2.

IF...	THEN...
you establish a maximum acceptable page fault rate of PFR1	for each image there is a minimum required working set size as shown at point A in Figures 6-2, 6-3, and 6-4.
you determine that the minimum level of page faulting is defined by PFR2 for all images	for each image there is a point (shown at point B) that is the maximum size the working set needs to reach.

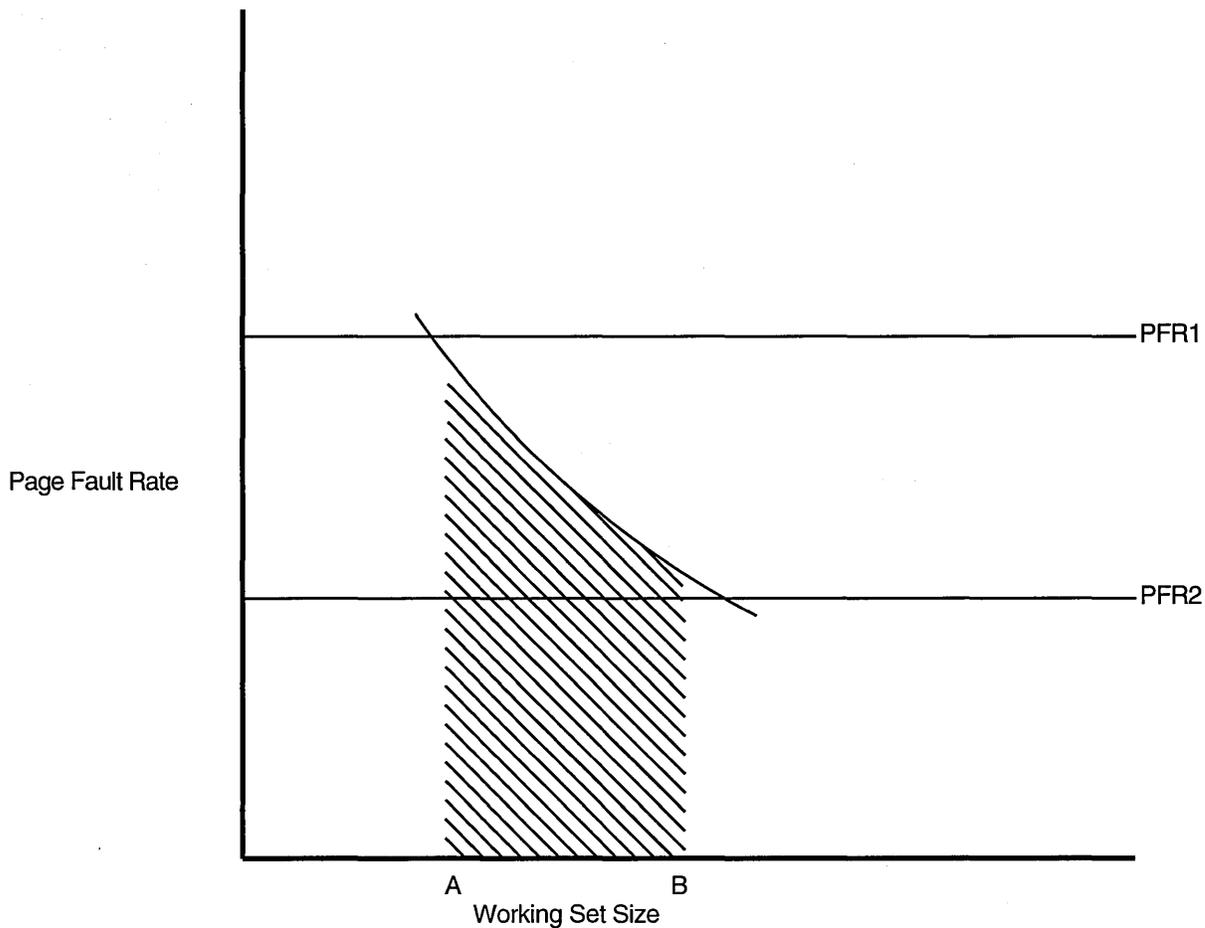
For example, for other images the working sets might behave more like the curves in Figures 6-3 or 6-4. Yet each of these characteristic curves illustrates that as you decrease the working set size, you should expect the page fault rate to increase. Note that if you establish a maximum acceptable page fault rate of PFR1, there is a minimum required working set size for each image, as shown at point A on each figure. If you determine that the minimum level of page faulting is defined by PFR2 for all images, then for each image there is a point (shown at point B) that is the maximum size the working set needs to reach.

Figure 6-3 Effect of Working Set Size on Page Fault Rate—Graph 2



ZK-1140-GE

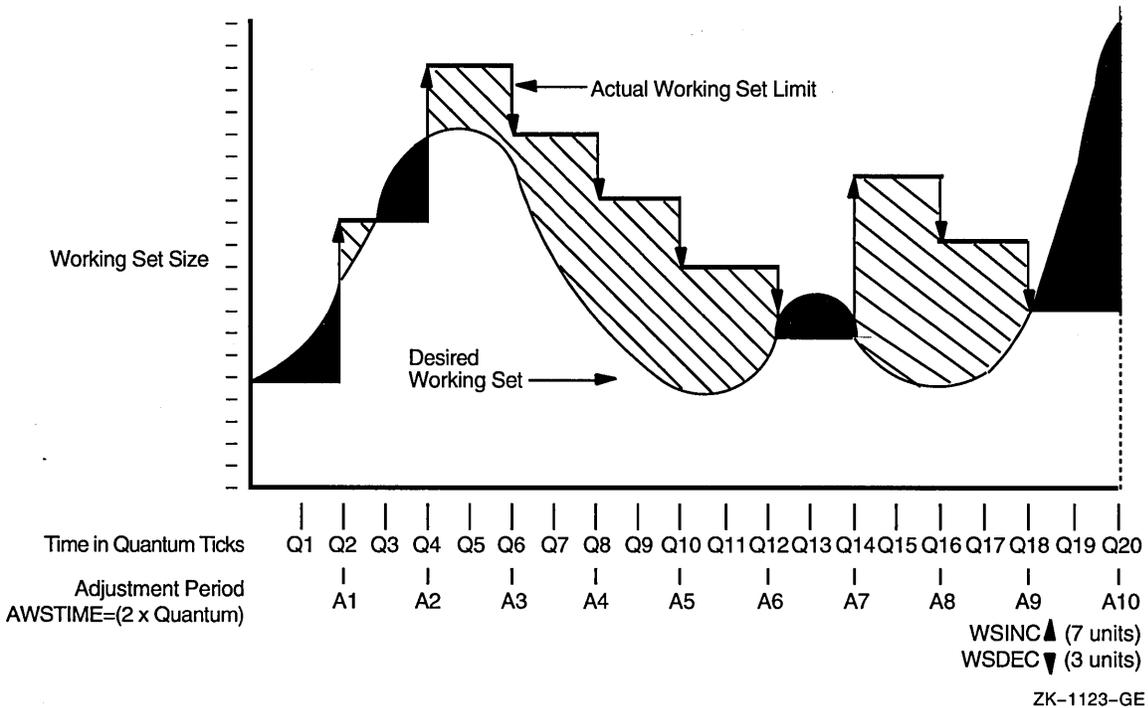
Figure 6-4 Effect of Working Set Size on Page Fault Rate—Graph 3



ZK-1141-GE

In Figure 6-5, the shaded area identifies where paging occurs. The portion between the desired working set size and the actual working set limit (shown with cross-hatching) represents unnecessary memory overhead—an obvious case where it costs memory to minimize page faulting.

Figure 6-5 An Example of Working Set Adjustment at Work



Voluntary decrementing

The parameters PFRATH and WSDEC, which control voluntary decrementing, are very sensitive to the application work load.

For the PFRATH and PFRATH parameters, it is possible to define values that appear to be reasonable page faulting limits but yield poor performance.

The problem results from the page replacement algorithm and the time spent maintaining the operation within the page faulting limits.

For example, for some values of PFRATH, you might observe that a process continuously page faults as its working set size grows and shrinks while the process attempts to keep its page fault rate within the limits imposed by PFRATH and PFRATH.

However, you might observe the same process running in approximately the same size working set, without page faulting once, with PFRATH turned off (set to zero).

Oscillation occurs when a process's working set size never stabilizes. To prevent the site from encountering an undesirable extreme of oscillation, the system turns off voluntary decrementing by initially setting parameter PFRATH equal to zero. You will achieve voluntary decrementing only if you deliberately turn it on.

Adjusting AWSA parameters

The following table summarizes adjustments to AWSA parameters:

Task	Adjustment
Enable voluntary decrementing	Set PFRATL greater than zero.
Disable borrowing	Set WSQUOTA equal to WSEXTENT.
Disable AWSA (per process)	Enter the DCL command SET WORKING_SET/NOADJUST.
Disable AWSA (systemwide)	Set WSINC to zero.

Note

If you plan to change any of these AWSA parameters, review the documentation for all of them before proceeding. You should also be able to explain why you want to change the parameters and what system behavior will occur. In other words, never make whimsical changes to the AWSA parameters on a production system.

Caution

It is possible to circumvent the AWSA feature by using the DCL command SET WORKING_SET/NOADJUST.

Use caution in disabling the AWSA feature because conditions could arise that would force the swapper to trim the process back to the value of the SWPOUTPGCNT system parameter.

Once AWSA is disabled for a process, the process cannot increase its working set size after the swapper trims the process to the SWPOUTPGCNT value. If the value of SWPOUTPGCNT is too low, the process is restricted to that working set size and will fault badly.

Performance management strategies for tuning AWSA

By developing a strategy for performance management that considers the desired automatic working set adjustment, you will know when the parameters are out of adjustment and how to direct your tuning efforts.

Sites typically choose one of the following general strategies for tuning AWSA parameters:

- Rapid response—Tune to provide a rapid response whenever the load demands greater working set sizes, allowing proactive memory reclamation to return memory from idle processes. To implement this strategy:
 - Set PFRATH low (possibly even to zero).
 - Set a low value for AWSTIME.

- Set a relatively large value for WSINC.
- Set BORROWLIM low and WSEXTENT high (even as high as WSMAX) to provide either large working set quotas or generous loans.

This is the default OpenVMS strategy where both BORROWLIM and GROWLIM are set equal to the value of FREELIM to allow maximum growth by active processes, and proactive reclamation is enabled to return memory idle processes.

- Less dynamic response—Tune for a less dynamic response that will stabilize and track moderate needs for working set growth. To implement this strategy:
 - Establish moderate values for AWSTIME, WSINC, and PFRATH. For example, set WSINC equal to approximately 10 percent of the typical value for WSDEFAULT.
 - Provide more generous working set defaults so that you do not need to set BORROWLIM so low as to ensure that loans would always be granted.

Swapper Trimming

The swapper process performs two types of memory management activities—swapping and swapper trimming. Sometimes, if process requirements so dictate, the operating system will *swap out* processes to a swapping file on disk so that the remaining processes can benefit from the use of memory without excessive page faulting.

Swapping refers to writing a process out to a reserved disk file known as a swapping file.

What is swapper trimming?

To better balance the availability of memory resources among processes, the operating system normally reclaims memory through a somewhat more complicated sequence of actions known as **swapper trimming**.

The system initiates swapper trimming whenever it detects too few pages in the free-page list.

Stage	Description
1	The system detects too few pages (below the value of FREELIM) in the free-page list.

Stage	Description						
2	The system checks whether the minimum number of pages exists in the modified-page list as defined by system parameter MPW_THRESH.						
	<table border="1"> <thead> <tr> <th>IF the minimum...</th> <th>THEN the system...</th> </tr> </thead> <tbody> <tr> <td>exists in the modified-page list</td> <td>invokes the modified-page writer to write out the modified-page list and free its pages for the free-page list.</td> </tr> <tr> <td>does NOT exist in the modified-page list to match FREEGOAL</td> <td>concludes that some of the processes should be trimmed; that is, forced to relinquish some of their pages or else be swapped out.</td> </tr> </tbody> </table>	IF the minimum...	THEN the system...	exists in the modified-page list	invokes the modified-page writer to write out the modified-page list and free its pages for the free-page list.	does NOT exist in the modified-page list to match FREEGOAL	concludes that some of the processes should be trimmed; that is, forced to relinquish some of their pages or else be swapped out.
IF the minimum...	THEN the system...						
exists in the modified-page list	invokes the modified-page writer to write out the modified-page list and free its pages for the free-page list.						
does NOT exist in the modified-page list to match FREEGOAL	concludes that some of the processes should be trimmed; that is, forced to relinquish some of their pages or else be swapped out.						

Trimming takes place at two levels (at process level and systemwide) and occurs before the system resorts to swapping.

First-level trimming

The swapper performs first-level trimming by checking for processes with outstanding loans; that is, processes that have borrowed on their working set extent. Such processes can be trimmed, at the swapper's discretion, back to their working set quota.

Second-level trimming

If first-level trimming failed to produce a sufficient number of free pages, then the swapper can trim at the second level.

With second-level trimming, the swapper refers to the systemwide trimming value SWPOUTPGCNT. The swapper selects a candidate process and then trims the process back to SWPOUTPGCNT and outswaps it.

As soon as the needed pages are acquired, the swapper stops trimming on the second level.

How are candidates chosen for second-level trimming?

Because the swapper does not want to trim pages needed by an active process, it selects the processes that are candidates for second-level trimming based on their states.

How are processes chosen for swapping?

Stage	Description
1	The swapper compares the length of real time that a process has been waiting since entering the hibernate (HIB) or local event flag wait (LEF) state with the system parameter LONGWAIT.
2	From its candidate list, the system selects the better processes for outswapping that have been idle for a time period equal to or greater than LONGWAIT.
3	By freeing up pages through outswapping, the system should allow enough processes to satisfy their CPU requirements, so that those processes that were waiting can resume execution sooner.

Suspended processes

Memory is always reclaimed from suspended processes before it is taken from any other processes. The actual algorithm used for the selection in each of these cases is complex, but those processes that are in local event flag wait or hibernate wait state are the next likeliest candidates.

Dormant process pseudoclass

After suspended (SUSP) processes, dormant processes are the most likely candidates for memory reclamation by the swapper. Two criteria define a dormant process as follows:

- The process must be a nonreal-time process whose current priority is equal to or less than the system parameter DEFPRI (default 4).
- The process must be a computable process that has not had a significant event (page fault, direct or buffered I/O, CPU time allocation) within an elapsed time period defined by the system parameter DORMANTWAIT (default 10 seconds).

Disabling second-level trimming

To disable second-level trimming, increase SWPOUTPGCNT to such a large value that second-level trimming is never permitted.

The swapper will still trim processes that are above their working set quotas back to SWPOUTPGCNT, as appropriate.

If you encounter a situation where any swapper trimming causes excessive paging, it might be preferable to eliminate second-level trimming and initiate swapping sooner. In this case, tune the swapping with the SWPOUTPGCNT parameter.

For a process with the PSWAPM privilege, you can also disable swapping and second-level trimming with the DCL command SET PROCESS/NOSWAPPING.

Swapper trimming versus voluntary decrementing

On most systems, swapper trimming is more beneficial than voluntary decrementing for the following reasons:

- Swapper trimming occurs on an as-needed basis
- Voluntary decrementing occurs on a continuous basis and affects only active, computable processes
- Voluntary decrementing can reach a detrimental condition of oscillation

The AUTOGEN command procedure, which establishes parameter values when the system is first installed, provides for swapper trimming but disables voluntary decrementing.

Proactive Memory Reclamation from Idle Processes

The memory management subsystem includes a policy that is designed to proactively reclaim memory from inactive processes when a deficit is first detected but before the memory resource is depleted.

Idle processes

The proactive memory reclamation policy acts on two types of idle processes:

- Long-waiting processes
- Periodically waking processes

Reclaiming memory from long-waiting processes

A candidate process for this policy would be in the LEF or HIB state for longer than number of seconds specified by the system parameter LONGWAIT.

First-Level Trimming

By setting FREEGOAL to a high value, memory reclamation from idle processes is triggered before a memory deficit becomes crucial and thus results in a larger pool of free pages available to active processes.

The system uses standard first-level trimming to reduce the working set size.

Second-Level Trimming

Second-level trimming with proactive memory reclamation enabled occurs, but with a significant difference.

When shrinking the working set to the value of SWPOUTPGCNT, the proactive memory reclamation policy removes pages from the working set but leaves the working set size (the limit to which pages can be added to the working set) at its current value, rather than reducing it to the value of SWPOUTPGCNT.

In this way, when the process is outswapped and eventually swapped in, it can readily fault the pages it needs without rejustifying its size through successive adjustments to the working set by AWSA.

Swapping Long-Waiting Processes

Long-waiting processes are swapped out when the size of the free-page list drops below the value of FREEGOAL.

A candidate long-waiting process is selected and outswapped no more than once every 5 seconds.

Reclaiming memory from periodically waking processes

The proactive memory reclamation policy also targets processes that do the following:

- Wake periodically
- Do minimal work
- Return to a sleep state

Watchdog Processes

Because it has a periodically waking behavior, a watchdog process is not a candidate for swapping but might be a good candidate for memory reclamation (trimming).

For this type of process, the policy tracks the relative wait-to-execution time.

How Trimming Is Performed

When the proactive memory reclamation policy is enabled, standard first- and second-level trimming are not used.

When the size of the free-page list drops below twice the value of FREEGOAL, the system initiates memory reclamation (trimming) of processes that wake periodically.

If a periodically waking process is idle 99 percent of the time and has accumulated 30 seconds of idle time, the policy trims 25 percent of the pages in the process's working set as the process reenters a wait state. Therefore, the working set remains unchanged.

Setting the FREEGOAL parameter

The system parameter FREEGOAL controls how much memory is reclaimed from idle processes.

Setting FREEGOAL to a larger value reclaims more memory; setting FREEGOAL to a smaller value reclaims less.

For information about AUTOGEN and setting system parameters, refer to the *OpenVMS System Manager's Manual*.

Sizing paging and swapping files

Because it reclaims memory from idle processes by trimming and swapping, the proactive memory reclamation policy can increase paging and swapping file use.

Use AUTOGEN in feedback mode to ensure that your paging and swapping files are appropriately sized for the potential increase.

For information about sizing paging and swapping files using AUTOGEN, refer to the *OpenVMS System Manager's Manual*.

How is the policy enabled?

Proactive memory reclamation is enabled by default.

However, by using the system parameter MMG_CTLFLAGS, you can enable or disable proactive memory reclamation long-waiting processes, periodically waking processes, or both. The system parameter MMG_CTLFLAGS is bit encoded. Table 6–1 describes how to enable or disable proactive memory reclamation.

Table 6–1 Parameter MMG_CTLFLAGS Bit Settings

Bit ¹	Meaning
<0>	If this bit is set, proactive memory reclamation is enabled for trimming periodically waking processes. Otherwise, if clear, it disables it.
<1>	If this bit is set, proactive memory reclamation is enabled by swapping out long-waiting processes. Otherwise, if clear, it disables it.

¹If MMG_CTLFLAGS equals 0, then proactive memory reclamation is disabled.

Memory Sharing**What is memory sharing?**

Memory sharing allows multiple processes to map to (and thereby gain access to) the same pages of physical memory.

Memory sharing (either code or data) is accomplished using a systemwide global page table similar in function to the system page table.

Global pages

Figures 6–6 and 6–7 illustrate how memory can be conserved through the use of global (shared) pages. The three processes (A, B, and C) run the same program, which consists of 2 pages of read-only code and 1 page of writable data.

Figure 6–6 shows the virtual-to-physical memory mapping required when each process runs a completely private copy of the program. Figure 6–7 illustrates the physical-memory gains possible and the data-structure linkage required when the

read-only portion of the program is shared by the three processes. Note that each process must still maintain a private data area to avoid corrupting the data used by the other processes.

Figure 6-6 Example Without Shared Code

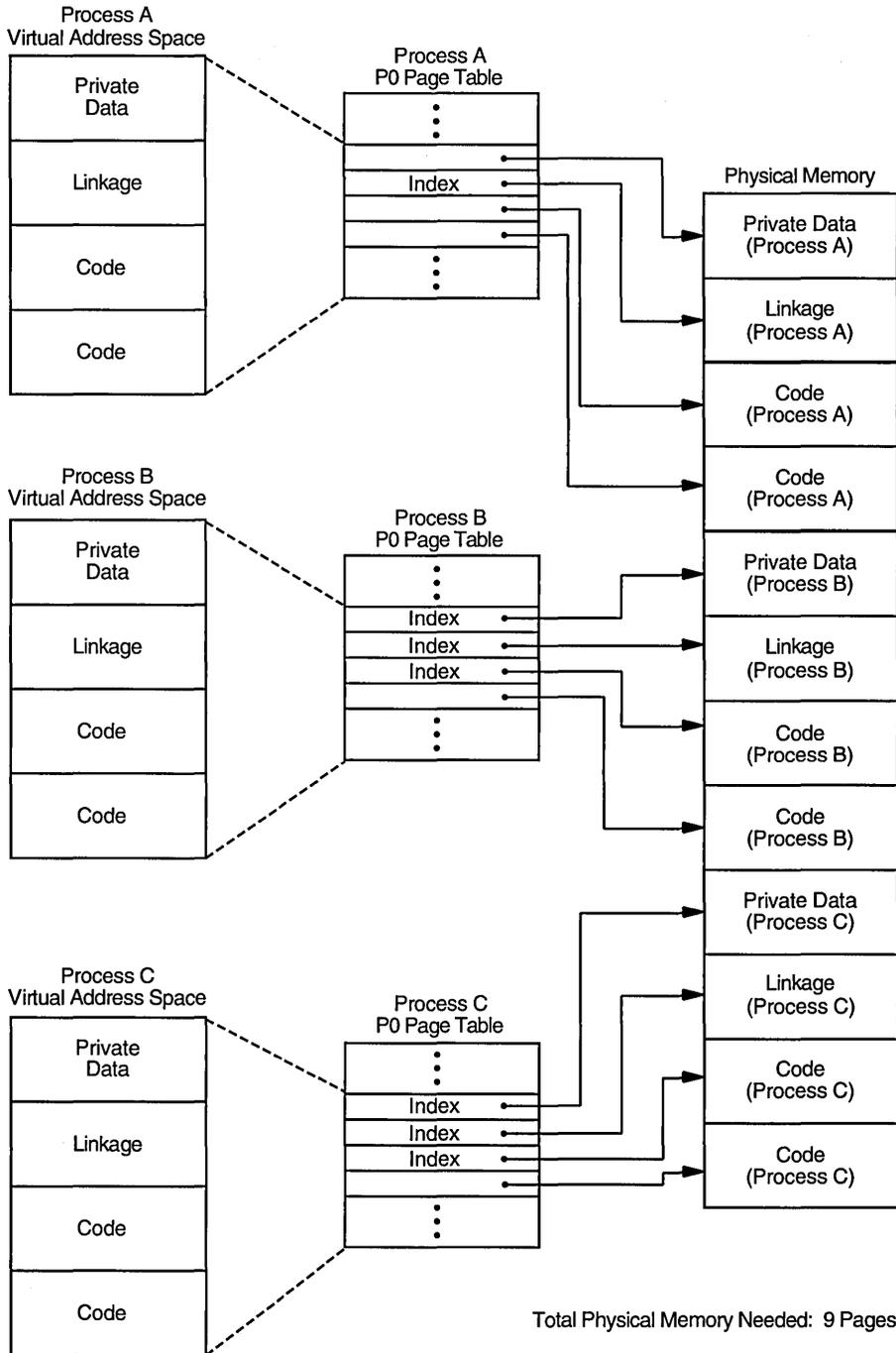
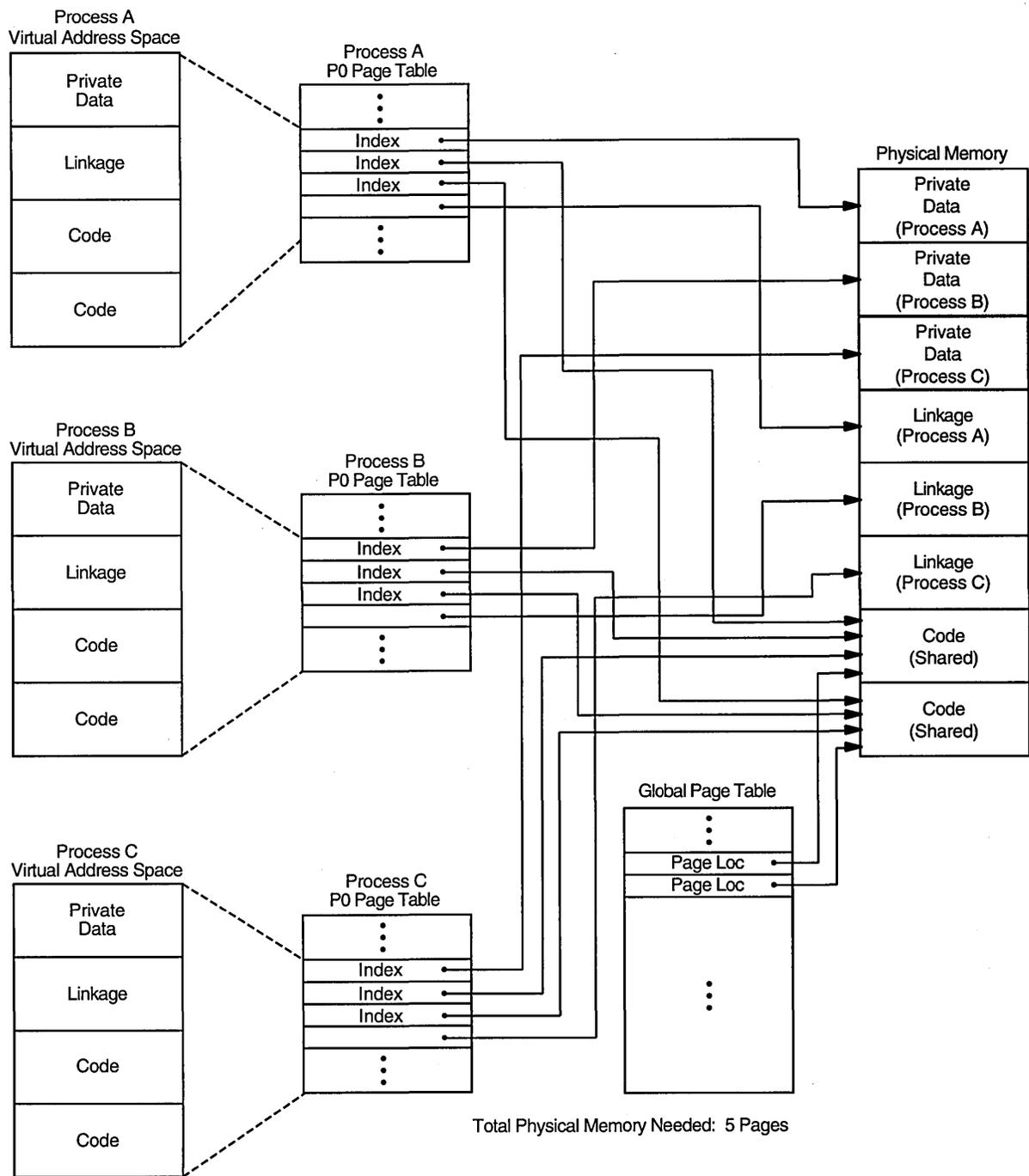


Figure 6-7 Example with Shared Code



ZK-7042A-GE

Memory can be saved by sharing code among several processes as shown in the following example:

$$savedmemory = pages\ of\ shared\ readonly\ code * sharing\ processes - 1$$

System overhead

The overhead consists of the data-structure space required for the (1) global page table entries and (2) global section table entries, both of which are needed to provide global mapping.

Each...	Requires a...	Allocated from the...
global page	8-byte global page table entry	global page table
global section	global section table entry global section descriptor	global section table paged dynamic pool

For more information about global sections, see the *OpenVMS Linker Utility Manual*.

Controlling the overhead

Two system parameters determine the maximum sizes for the two data structures in the process header as follows:

- **GBLPAGES**—Defines the size of the global page table. The system working set size as defined by **SYSMWCNT** must be increased whenever you increase **GBLPAGES**.
- **GBLSECTIONS**—Defines the size of the global section table.

Installing shareable images

Once a shareable image has been created, it can be installed as a permanently shared image. (See the *OpenVMS Linker Utility Manual* and the *OpenVMS System Manager's Manual*). Memory will only be saved, however, when there is more than one process actually mapped to the image at a time.

Also, use **AUTHORIZE** to increase the user's working set characteristics (**WSDEF**, **WSQUO**, **WSEXTENT**) wherever appropriate, to correspond to the expected use of shared code. (Note, however, that this increase does not mean that the actual memory usage will increase. Sharing of code by many users actually decreases the memory requirement.)

Verifying memory sharing

If physical memory is especially limited, investigate whether there is much concurrent image activation that results in savings. If you find there is not, there is no reason to employ code sharing. You can use the following procedure to determine if there is active sharing on image sections that have been installed as shareable:

1. Invoke the OpenVMS Install utility (**INSTALL**) and enter the **LIST/FULL** command. For example:

```
$ INSTALL
INSTALL> LIST/FULL LOGINOUT
```

INSTALL displays information in the following format:

```
DISK$AXPVMSRL4:<SYS0.SYSEX>.EXE
```

```

LOGINOUT;3      Open Hdr   Shar Priv
Entry access count   = 44
Current / Maximum shared = 3 / 5
Global section count   = 2
Privileges = CMKRNL SYSNAM TMPMBX EXQUOTA SYSPRV

```

2. Observe the values shown for the Current/Maximum shared access counts:
 - The Current value is the current count of concurrent accesses of the known image.
 - The Maximum value is the highest count of concurrent accesses of the image since it became known (installed). This number appears only if the image is installed with the /SHARED qualifier.

The Maximum value should be at least 3 or 4. A lower value indicates that overhead for sharing is excessive.

Note

In general, your intuition, based on knowledge of the work load, is the best guide. Remember that the overhead required to share memory is counted in bytes of memory, while the savings are counted in pages of physical memory. Thus, if you suspect that there is occasional concurrent use of an image, the investment required to make it shareable is worthwhile.

OpenVMS Scheduling

Scheduling

The scheduler uses a modified round-robin form of scheduling: processes receive a chance to execute on rotating basis, according to process priority and state.

Time slicing

Each computable process receives a time slice for execution. The time slice equals the system parameter QUANTUM, and rotating the time slices among processes is called **time slicing**. Once its quantum starts, each process executes until one of the following events occurs:

- A process of higher priority becomes computable
- The process is no longer computable because of a resource wait

- The process itself voluntarily enters a wait state
- The quantum ends

If there is no other computable (COM) process at the same priority ready to execute when the quantum ends, the current process receives another time slice.

Process state

A change in process state causes the scheduler to reexamine which process should be allowed to run.

Process priority

When required to select the next process for scheduling, the scheduler examines the priorities assigned to all the processes that are computable and selects the process with the highest priority.

Priorities are numbers from 0 to 31.

Processes assigned a priority of 16 or above receive maximum access to the CPU resource (even over system processes) whenever they are computable. These priorities, therefore, are used for real-time processes.

Priority boosting

For processes below priority 16, the scheduler can increase or decrease process priorities as shown in the following table:

Stage	Description
1	While processes run, the scheduler recognizes events such as I/O completions, the completion of an interval of time, and so forth.
2	As soon as one of the recognized events occurs and the associated process becomes computable, the priority of that process may be increased. The amount of the increase is related to the associated event. ¹
3	The scheduler examines which computable process has the highest priority and, if necessary, causes a context switch so that the highest priority process runs.
4	As soon as a process is scheduled, its priority is reduced by one to allow processes that have received a priority boost to begin to return their base priority. ²

¹For example, if the event is the completion of terminal I/O input, a large increase is given so that the process can run again sooner.

²The priority is never decreased below the base priority or increased into the real-time range.

Scheduling real-time processes

When real-time processes (those with priorities from 16 to 31) execute, the following conditions apply:

- They never receive a priority boost
- They do not experience automatic working set adjustments

- They do not experience quantum-based time slicing

The system permits real-time processes to run until either they voluntarily enter a wait state or a higher priority real-time process becomes computable.

Tuning

From a tuning standpoint, you have very few controls you can use to influence process scheduling. However, you can modify the following:

- Base priorities of processes
- Length of time for a quantum

All other aspects of process scheduling are fixed by the behavior of the scheduler and the characteristics of the work load.

Processes

A process receives a default base priority from the following:

- /PRIORITY qualifier in the UAF record
- DEFAULT record in the UAF record

A process can change its priority using the following:

- \$SETPRI system service.
- DCL command SET PROCESS/PRIORITY to reduce the priority. You need ALTPRI privilege to increase the priority of your process.

A user requires GROUP or WORLD privilege to change the priority of other processes.

Subprocesses and detached processes

A subprocess or detached process receives its base priority from the following:

- \$CREPRC system service
- DCL command RUN

If no priority is specified, the priority of the creator is used.

Batch queues

When a batch queue is created, the DCL command INITIALIZE /QUEUE/PRIORITY establishes the default priority for a job.

However, when you submit a job with the DCL command SUBMIT or change characteristics of that job with the DCL command SET QUEUE/ENTRY, you can adjust the priority with the /PRIORITY qualifier.

With either command, increases are permitted only for submitters with the OPER privilege.

Evaluating System Resources

Overview

This chapter describes command procedures that help you evaluate the performance CPU, memory, and disk I/O subsystem resources using MONITOR and to a lesser extent, other standard utilities. The following topics are discussed:

- Using ACCOUNTING to obtain image-level accounting data and provide guidelines for interpreting the data
- Interpreting MONITOR summary reports

Discussions focus on the utilization of each hardware resource by major software components and on the measurement, analysis, and possible reallocation of the hardware resources. Suggestions for corrective actions are provided in case your evaluation indicates that improvements are possible.

Purpose

To help you verify that your system is performing well and to provide information to aid you in maintaining performance at an acceptable level.

Resource Management

What are system resources?

For practical purposes, managing the performance of a system is best approached by managing its resources. The following table describes hardware and software resources:

Component	Resources
Hardware	CPU, memory, and peripherals
Software	Application programs, optional products, and operating system facilities ¹

¹For example, the Extended QIO Processor (XQP) and memory and I/O management mechanisms.

The term resources, as used in this manual, refers to the three major hardware resources—CPU, memory, and disk I/O.

Tools and utilities

You can become knowledgeable about your system's operation if you use MONITOR, ACCOUNTING, and AUTOGEN feedback on a regular basis to capture and analyze certain key data items.

You should exercise care in selecting the items you want to measure and the frequency with which you capture the data.

If you are overzealous, the consumption of system resources to collect, store, and analyze the data can distort your picture of the system's work load and capacity.

Prerequisites

It is assumed that your system can be classified as a general timesharing system. It is further assumed that you have followed the workload management techniques and installation guidelines described in Chapter 1 and Chapter 4, respectively.

The procedures outlined in this chapter differ from those in Chapters 12 and 16 in the following ways:

- They are designed to help you conduct an evaluation of your system and its resources, rather than to execute an investigation of a specific problem. If you discover problems during an evaluation, you can refer to the decision-tree diagrams in Chapters 12 and 16 for further analysis.
- For simplicity, they are less exhaustive, relying on certain rules of thumb to evaluate the major hardware resources and to point out possible deficiencies, but stopping short of pinpointing exact causes.
- They are centered on the use of MONITOR, particularly the summary reports, both standard and multifile.

Note

Some information in this chapter might not apply to certain specialized types of systems or to applications such as workstations, database management, real-time operations, transaction processing, or any in which a major software subsystem is in control of resources for other processes.

Guidelines

As you conduct your evaluations, keep the following rules in mind:

- Complete the entire evaluation. It is important to examine all the resources in order to evaluate the system as a whole. A partial examination can lead you to attempt an improvement in an area where it might have minimal effect because more serious problems exist elsewhere.

- Become as familiar as possible with the applications running on your system. Get to know what their resource requirements are. You can obtain a lot of relevant information from the ACCOUNTING image report shown in Example 7–1. User's guides associated with Digital and third-party software can also be helpful in identifying resource requirements.
- If you believe that a change in software parameters or a hardware configuration can improve performance, execute such a change cautiously, being sure to make only one change at a time. Evaluate the effectiveness of the change before deciding to make it permanent.

Note

When specific values or ranges of values for MONITOR data items are recommended, they are intended only as guidelines and will not be appropriate in all cases.

Collecting and Interpreting Image-Level Accounting Data

What is image-level accounting?

Image-level accounting is a feature of ACCOUNTING that provides statistics and information on a per-image basis.

Why is it useful?

Image-level accounting can be useful in helping you gain an understanding of resource utilization on a per-image basis.

By knowing which images are heavy consumers of resources at your site, you can better direct your efforts of controlling them and the resources they consume.

Frequently used images are typically good candidates for code sharing, whereas images that consume large quantities of various resources can be forced to run in a batch queue where the number of simultaneous processes can be controlled.

Guidelines

You should be judicious in using image-level accounting on your system. Consider the following guidelines when using ACCOUNTING:

- Enable image-level accounting only when you plan to invoke ACCOUNTING to process the information provided in the file `SY$MANAGER:ACCOUNTING.DAT`.
- Disable image-level accounting once you have collected enough data for your purposes.

While image activation data can be helpful in performance analysis, it wastes processing time and disk storage if it is collected and never used.

Enabling and disabling image-level accounting

You enable image-level record collection by issuing the DCL command SET ACCOUNTING/ENABLE=IMAGE.

Disable image-level accounting by issuing the DCL command SET ACCOUNTING/DISABLE=IMAGE.

Note

The collection of image-level accounting data consumes CPU cycles. The collected records can consume a significant amount of disk space. Remember to enable image-level accounting only for the period of time needed for the report.

Generating a report

A series of commands like the following will generate output similar to that shown in Example 7-1.

```
$ ACCOUNTING /TYPE=IMAGE /OUTPUT=BYNAM.LIS -
_$ /SUMMARY=IMAGE -
_$ /REPORT=(PROCESSOR,ELAPSED,DIRECT IO,FAULTS,RECORDS)
_$ SORT BYNAM.LIS BYNAM.ORD /KEY=(POS=16,SIZ=13,DESCEND)
.
.
.
(Edit BYNAM.ORD to relocate heading lines)
.
.
.
$ TYPE BYNAM.ORD
```

Collecting the data

Example 7-1 assumes that image-level accounting records have been collected previously.

Example 7-1 Image-Level Accounting Report

Image name	Processor Time	Elapsed Time	Direct I/O	Page Faults	Total Records
EDT	0 00:34:21.34	0 15:51:34.78	5030	132583	390
DTR32	0 00:19:30.94	0 03:17:37.48	7981	83916	12
PASCAL	0 00:15:19.42	0 01:04:19.57	38473	143107	75
MAIL	0 00:10:40.88	1 02:54:02.89	26139	106854	380
LINK	0 00:05:44.41	0 00:23:54.54	7443	57092	111
RTPAD	0 00:04:58.40	0 20:49:19.24	668	8004	72
LOGINOUT	0 00:04:53.98	0 02:01:31.81	2809	67579	893
EMACS	0 00:04:30.40	0 05:25:01.37	420	8461	1

(continued on next page)

Example 7-1 (Cont.) Image-Level Accounting Report

MACRO32	0 00:04:26.22	0 00:14:55.00	1014	34016	46
BLISS32	0 00:03:45.80	0 00:12:58.87	98	32797	8
DIRECTORY	0 00:03:26.20	0 01:22:34.47	1020	27329	275
FORTRAN	0 00:03:13.87	0 00:14:15.08	1157	28003	47
NOTES	0 00:01:39.90	0 02:06:01.95	8011	6272	32
DELETE	0 00:01:37.31	0 00:57:43.31	834	25516	332
TYPE	0 00:01:06.35	0 00:28:58.26	406	14457	173
COPY	0 00:00:57.08	0 00:11:11.40	2197	4943	42
SHOW	0 00:00:56.39	0 00:24:53.22	23	11505	166
ACC	0 00:00:54.43	0 00:03:41.46	132	2007	7
MONITOR	0 00:00:53.91	0 02:37:13.84	159	5649	40
CALENDAR	0 00:00:43.55	0 00:30:15.52	1023	3557	25
PHONE	0 00:00:40.56	0 00:54:59.39	24	1510	33
ERASE	0 00:00:37.88	0 00:03:51.04	105	9873	113
LIBRARIAN	0 00:00:35.58	0 00:03:37.98	1134	10297	62
FAL	0 00:00:34.27	0 00:20:56.63	110	4596	122
SDA	0 00:00:27.34	0 00:09:28.68	52	4797	3
SET	0 00:00:27.02	0 00:02:30.28	160	9447	206
NETSERVER	0 00:00:26.89	0 02:38:17.90	263	10164	407
CDU	0 00:00:24.32	0 00:01:57.67	13	21906	17
VMSHELP	0 00:00:12.83	0 00:05:40.96	121	1943	14
RENAME	0 00:00:09.56	0 00:00:57.44	6	3866	47
SDL	0 00:00:09.55	0 00:01:19.78	11	3158	4
SUBMIT	0 00:00:08.14	0 00:01:08.50	9	2991	28
NCP	0 00:00:07.30	0 00:02:26.20	7	1765	16
QUEMAN	0 00:00:06.44	0 00:01:38.75	201	1561	20

This example shows a report of system resource utilization for the indicated period, summarized by unique image name, in descending order of CPU utilization. Only the top 34 CPU consumers are shown. (The records could easily have been sorted differently.)

The Total Records column is a count of image terminations, requested by specifying the RECORDS report key in the ACCOUNTING command that generated the report. The /SINCE and /BEFORE qualifiers can be specified to select any time period of interest.

Interpreting the data**Image Name**

Most image names are programming languages and operating system utilities, indicating that the report was probably generated in a program-development environment.

Processor Time

Data in this column shows that no single image is by far the highest consumer of the CPU resource. It is therefore unlikely that the installation would benefit significantly by attempting to reduce CPU utilization by any one image.

Direct I/O

In the figures for direct I/O, the two top images are PASCAL and MAIL. One way to compare them is by calculating I/O operations per second. The total elapsed time spent running PASCAL is roughly 3860 seconds, while the time spent running MAIL is a little under 96843 seconds (several people used MAIL all afternoon). Calculated on a time basis, MAIL caused roughly 1/4 to 1/3 of an I/O operation per second, whereas PASCAL caused about 10 operations per second.

Note that by the same calculation, LINK caused about five I/O operations per second. It would appear that a sequence of PASCAL/LINK commands contributes somewhat to the overall I/O load. One possible approach would be to look at the RMS buffer parameters set by the main PASCAL users. You can find out who used PASCAL and LINK by entering a DCL command:

```
$ ACCOUNTING/TYPE=IMAGE/IMAGE=(PASCAL, LINK) -  
_ $ /SUMMARY=(IMAGE, USER)/REPORT=(ELAPSED, DIRECT)
```

This command selects image accounting records for the PASCAL and LINK images by image name and user name, and requests Elapsed Time and Direct I/O data. You can examine this data to determine whether the users are employing RMS buffers of appropriate sizes. Digital recommends that two fairly large buffers be used for sequential I/O, each being approximately 64 blocks in size.

Page Faults

As with direct I/O, page faults are best analyzed on a time basis. One technique is to compute faults-per-10-seconds of processor time and compare the result with the value of the SYSGEN parameter PFRATH. A little arithmetic shows that on a time basis, PASCAL is incurring more than 1555 faults per 10 seconds. Suppose that the value of PFRATH on this system is 120 (120 page faults per 10 seconds of processor time), which is considered typical in most environments. What can you conclude by comparing the two values?

Whenever a process's page fault rate exceeds the PFRATH value, memory management attempts to increase the process working set, subject to system management quotas, until the fault rate falls below PFRATH. So, if an image's fault rate is persistently greater than PFRATH, it is not obtaining all the memory it needs.

Clearly, the PASCAL image is causing many more faults per CPU second than would be considered normal for this system. You should, therefore, make an effort to examine the working set limits and working set adjustment policies for the PASCAL users. To lower the PASCAL fault rate, the process working sets must be increased—either by adjusting the appropriate UAF quotas directly or by setting up a PASCAL batch queue with generous working set values.

Total Records

These figures represent the count of activations for images run during the accounting period; in other words, they show each image's relative popularity. You can use this information to ensure that the most popular images are installed (see Chapter 4, Postinstallation System Management Options). For customer applications, you might consider linking options such as /NOSYSSHR and reassigning PSECT attributes to speed up activations (see the *OpenVMS Linker Utility Manual*).

Note that the number of LOGINOUT activations far exceeds that of all other images. This situation could result from a variety of causes, including attempts to breach security, an open terminal line, a runaway batch job, or a large number of network operations. Further ACCOUNTING commands would be necessary to determine the exact cause. At this site, it turned out that most of the activations were caused by an open terminal line. The problem was detected by an astute system manager who checked the count of LOGFAIL entries in the accounting file.

You can also use information in this field to examine the characteristics of the average image activation. That knowledge would be useful if you wanted to determine whether it would be worthwhile to set up a special batch queue.

For example, the average PASCAL image uses 51 seconds of elapsed time and the average LINK uses 13 seconds. You can therefore infer that the average PASCAL and LINK sequence takes about a minute. This information could help you persuade users of those images to run PASCAL and LINK in batch mode. If, on the other hand, the average time were only 5 seconds, batch processing would probably not be worthwhile.

Creating, Maintaining, and Interpreting MONITOR Summaries

Guidelines

Consider the following guidelines when using MONITOR:

- Before capturing data, have a specific plan for how you will analyze and apply it.
- Avoid an interval value so long that you require unnecessary disk storage for the data.
- Do not select an interval so short that you miss significant events occurring in the interim.

See the *OpenVMS System Manager's Manual* and the *OpenVMS System Management Utilities Reference Manual* for information about using MONITOR.

Types of output

MONITOR generates the following types of output:

- ASCII screen images of statistics from a running system (/DISPLAY qualifier)
- Binary recording files containing data collected from a running system (/RECORD qualifier)
- Formatted ASCII summary files of statistics extracted from binary recording files (/SUMMARY qualifier)

MONITOR modes of operation

MONITOR provides two input modes of operation for collecting data—live and playback.

Live Mode

Use live mode to collect data on a running system and to generate one or more of the following types of MONITOR output—ASCII screen images, binary recording files, or formatted ASCII summary files.

Use live mode to display data about a remote system connected to your system with DECnet for OpenVMS.

Playback Mode

Use playback mode to read a binary recording file and produce one or more of the following types of MONITOR output—ASCII screen images, binary recording files, or formatted ASCII summary files.

Creating a performance information database

As a foundation for the strategy discussed in this chapter, you must develop a database of performance information for your system by running MONITOR continuously as a background process.

The SYS\$EXAMPLES directory provides three command procedures you can use to establish the database. The following table describes the procedures:

Procedure	Description
SUBMON.COM	Starts MONITOR.COM as a detached process.
MONITOR.COM	Creates a summary file from the binary recording file of the previous boot, then begins recording for this boot. The recording interval is 10 minutes.

Procedure	Description
MONSUM.COM	Generates two VMScluster multifile summary reports: one for the previous 24 hours and one for the previous day's prime-time period (9 a.m. to 6 p.m.). These reports are mailed to the system manager, and then the procedure resubmits itself to run each day at midnight.

When MONITOR data is recorded continuously, a summary report can cover any contiguous time segment.

Saving your summary reports

The two multifile summary reports are not saved as files. To keep them, you must do either of the following:

- Extract them from your mail file
- Alter the MONSUM.COM command procedure to save them

Customizing your reports

The report you require for the evaluation procedure is one that covers a period that best represents the typical operation of your system. You might want, for example, to evaluate your system only during hours of peak activity.

To generate a summary of the appropriate time segment, edit the MONSUM.COM command procedure and change the beginning and ending times on one of the two MONITOR commands that produce the summary reports.

Report formats

The summary reports produced by MONSUM.COM are in the multifile summary format—there is one column of averages for each node in a VMScluster, as well as some overall row statistics. For noncluster systems, the row statistics can be ignored.

If you prefer to use a report in the standard summary format (which includes current, minimum, and maximum statistics), execute a MONITOR playback summary command referencing the input data file of interest as the only file in the /INPUT list. Note that a new data file is created for each system whenever it reboots. Remember to use the /BEGINNING and /ENDING qualifiers to select the desired time period.

Using MONITOR in live mode

You are encouraged to observe current system activity regularly by running MONITOR in live mode. In live mode, always begin an analysis with the MONITOR CLUSTER and MONITOR SYSTEM classes to obtain an overview of system performance.

Then, monitor other classes to examine components of particular interest.

Note

All references to MONITOR items in this chapter are assumed to be for the average statistic, unless otherwise noted.

**More about
multifile
reports**

A page or more is devoted to each MONITOR class. Each column represents one node, and is headed by the node name and beginning and ending times of the segment requested. In most cases, time segments for all nodes will be roughly the same. Differences of a few minutes are typical because data collection on the various nodes is not synchronized.

In some cases, one or more time segments will be shorter than others; in these cases, some of the requested data was not recorded (probably because the nodes were unavailable). Note that if data is unavailable for some period within the bounds of a request, that fact is not explicitly specified.

However, such a gap can occur only when the column of data uses more than one input file; and if multiple files contributed to the column, the number is shown in parentheses to the right of the node name. In cases where a time segment is missing, this number must be greater than 1. If no number appears, there is only one input data file for that column, and the column includes no missing time segments.

To summarize: if all beginning and ending times are not roughly the same or if a parenthesized number appears, some data might be unavailable and you might want to base your evaluation on a different time segment that includes more complete data. Whenever the multifile report is based on incomplete data, the Row Average statistic can be weighted unfairly in favor of one or more nodes.

**Interpreting
MONITOR
statistics**

While interpreting MONITOR statistics, keep in mind that the collection interval has no effect on the accuracy of MONITOR rates. It does, however, affect *levels* because they represent sampled data. In other words, the smaller the collection interval, the more accurate MONITOR level statistics will be. (For more information on MONITOR rates and levels, refer to the *OpenVMS System Manager's Manual*.)

Although the interval value supplied with MONITOR.COM is adequate for most purposes, it does represent a trade-off between statistical accuracy and the consumption of disk space. Thus, before you base major decisions on MONITOR level statistics, be sure to verify them by running MONITOR for a time with a much smaller collection interval while carefully observing disk space usage.

Managing System Resources

Overview

Overall responsiveness of a system depends largely on the responsiveness of its CPU, memory, and disk I/O resources. If each resource responds satisfactorily, then so will the entire system. This chapter discusses the following topics:

- Evaluating overall system responsiveness
- Improving overall system responsiveness

Purpose

To evaluate overall system responsiveness using system utilities and procedures.

Definition

A **binding resource** or **bottleneck** is an overcommitted resource that causes the others to be blocked or burdened with overhead operations.

Understanding System Responsiveness

Interacting resources

Each resource must operate efficiently by itself and it must also interact with other resources.

Overcommitted resources

An important aspect of your evaluation is to distinguish between resources that might be performing poorly because they are overcommitted and those that might do so because one or both of the following conditions has occurred:

- They are blocked by the overcommitted resource.
- They are incurring additional overhead operations caused by the overcommitted resource.

Detecting bottlenecks

Detecting bottlenecks is particularly important for analyzing interactions of the CPU with each of the other resources.

Example

For example, CPU blockage occurs when CPU capacity, though it appears sufficient to meet demand, cannot be used because the CPU must wait for disk I/O to complete or memory to be allocated.

Upgrading a nonbinding resource will do nothing to improve a bottlenecked system.

Balancing resource capacities

Because of the potential for bottlenecks, it is especially important to maintain balance among the capacities of your system's resources.

Example

For example, when upgrading to a faster CPU, consider the effect the additional CPU power will have on the other primary resources. Since the faster CPU can initiate more I/O requests per unit of time, you must ensure that the disk I/O subsystem has sufficient capacity to handle the increased traffic.

Evaluating Responsiveness of System Resources

Using MONITOR

For each resource, key MONITOR statistics help you answer such questions as the following:

- How well is the resource responding to requests for service?
- How well is the capacity of the resource meeting demand?
- Does the resource have any excess capacity, and if so, can that capacity be attributed to blockage by another, overcommitted resource?

Measuring system responsiveness

Two prime measures of resource responsiveness are as follows:

- The size of the queue of requests for service (compute queue)
- The amount of time it takes the system to respond to those requests (response time)

For each resource, you can use MONITOR summaries to examine or estimate one or both of these quantities.

Improving Responsiveness of System Resources

If the responsiveness of a poorly performing resource cannot be improved by the following methods, you should consider augmenting its capacity with additional or upgraded hardware:

- Equitable sharing
- Reducing resource consumption
- Load balancing
- Offloading

Equitable sharing

Is the resource shared equitably among processes?

Reducing resource consumption

Can the system's consumption of a resource be reduced, thereby making more of that resource available to users?

The effective amount of a resource available to users is that remaining after the operating system has used its portion.

Load balancing

How well distributed is the demand for a resource? Can overall system responsiveness be improved, either by reconfiguring hardware or by better distributing the demand for it?

Offloading

Can overall system responsiveness be improved by offloading some of the activity on a resource to other less heavily used resources types?

Example

Excess memory capacity is often used to reduce the demand on an overworked disk I/O subsystem by increasing the size of each I/O transfer, thereby reducing the total number of I/O operations.

The CPU benefits as well because it needs to do less work executing system services and device driver software.

The primary means of offloading I/O to memory is the extensive use of caches (page caches, XQP caches, virtual I/O caching, RMS blocking) to reduce the number of I/O operations.

The CPU Resource

Overview

The CPU is the central resource in your system and it is the most costly to augment. Good CPU performance is vital to that of the system as a whole because the CPU performs the two most basic system functions: it allocates and initiates the demand for all the other resource, and it provides instruction execution service to user processes. This chapter discusses the following topics:

- Evaluating CPU responsiveness
- Improving CPU responsiveness

Purpose

To evaluate the performance of the CPU resource.

Definition

A **spin lock** is a mechanism that guarantees the synchronization of processors in their manipulation of operating system databases.

Evaluating CPU Responsiveness

Compute queue

Only one process can execute on a CPU at a time. The CPU resource must be shared sequentially. Because several processes might be ready to use the CPU at any given time, the system maintains a **queue** of processes waiting for the CPU.

These processes are in the compute (COM) or compute outswapped (COMO) scheduling states.

Quantum

The system allocates the CPU resource for a period of time known as a quantum to each process that is not waiting for other resources.

During its quantum, a process can execute until any of the following events occur:

- The process is preempted by a higher priority process.

- The process voluntarily yields the CPU by requesting a wait state for some purpose (for example, to wait for the completion of a user I/O request).
- The process enters an involuntary wait state, such as when it triggers a hard page fault (one that must be satisfied by reading from disk).

CPU Response Time

A good measure of the CPU response is the average number of processes in the COM and COMO states over time—that is, the average length of the compute queue.

If the number of processes in the compute queue is close to 0, unblocked processes will rarely need to wait for the CPU.

Factors Affecting Response Time

Several factors affect how long any given process must wait to be granted its quantum of CPU time:

- Interrupt state
- Computing requirements of the processes in the compute queue
- CPU type
- Scheduling priority

The worst-case scenario involves a large compute queue of compute-bound processes. Each compute-bound process can retain the CPU for the entire quantum period.

Compute-Bound Processes

Assuming no interrupt time and a default quantum of 200 milliseconds, a group of five compute-bound processes of the same priority (one in CUR state and the others in COM state) will acquire the CPU once every second.

As the number of such processes increases, there is a proportional increase in the waiting time.

If the processes are not compute bound, they can relinquish the CPU before having consumed their quantum period, thus reducing waiting time for the CPU.

Determining Optimal Queue Length

The best way to determine a reasonable length for the compute queue at your site is to note its length during periods when all the system resources are performing adequately and when users perceive response time to be satisfactory.

Then, watch for deviations from this value and try to develop a sense for acceptable ranges.

Estimating available CPU capacity

Observe the average amount of idle time and the average number of processes in the various scheduling wait states.

While idle time is a measure of the percentage of unused CPU time, the wait states indicate the reasons that the CPU was idle and might point to utilization problems with other resources.

Overcommitted Resources

Before using idle time to estimate growth potential or as an aid to balancing the CPU resource among processes in a VMScluster, ensure that the other resources are not overcommitted, thereby causing the CPU to be underutilized.

Scheduling Wait States

Whenever a process enters a scheduling wait state—a state other than CUR (process currently using the CPU) and COM—it is said to be **blocked** from using the CPU.

Most times, a process enters a wait state as part of the normal synchronization that takes place between the CPU and the other resources.

But certain wait states can indicate problems with those other resources that could block viable processes from using the CPU.

MONITOR data on the scheduling wait states provides clues about potential problems with the memory and disk I/O resources.

Types of scheduling wait states

There are two types of scheduling wait states—**voluntary** and **involuntary**. Processes enter voluntary wait states directly; they are placed in involuntary wait states by the system.

Voluntary wait states

Processes in the local event flag wait (LEF) state are said to be voluntarily blocked from using the CPU; that is, they are temporarily requesting to wait before continuing with CPU service. Since the LEF state can indicate conditions ranging from normal waiting for terminal command input to waiting for I/O completion or locks, you can obtain no useful information about potentially harmful blockage simply by observing the number of processes in that state. You can usually assume, though, that most of them are waiting for terminal command input (at the DCL prompt).

Disk I/O Completion

Some processes might enter the LEF state because they are awaiting I/O completion on a disk or other peripheral device. If the I/O subsystem is not overloaded, this type of waiting is temporary and inconsequential. If, on the other hand, the I/O resource, particularly disk I/O, is approaching capacity, it could be causing the CPU to be seriously underutilized.

Long disk response times are the clue that certain processes are in the LEF state because they are experiencing long delays acquiring disk service. If your system exhibits unusually long disk response times, refer to Chapter 11, Evaluating Disk I/O Responsiveness, and try to correct that problem before attempting to improve CPU responsiveness.

Waiting for a Lock

Other processes in the LEF state might be waiting for a lock to be granted. This situation can arise in environments where extensive file sharing is the norm—particularly in VMSclusters. Check the ENQs Forced to Wait Rate. (This is the rate of \$ENQ lock requests forced to wait before the lock was granted.) Since the statistic gives no indication of the duration of lock waits, it does not provide direct information about lock waiting. A value significantly larger than your system’s normal value, however, can indicate that users will start to notice delays.

IF you suspect...	THEN...
that the lock waiting is caused by file sharing ¹	attempt to reduce the level of sharing, if possible.
that the lock waiting results from user or third-party application locks	attempt to influence the redesign of such applications.

¹RMS and the XQP use locks to synchronize record and file access.

Process Synchronization

Processes can also enter the LEF state or the other voluntary wait states [common event flag wait (CEF), hibernate (HIB), and suspended (SUSP)] when system services are used to synchronize applications. Such processes have temporarily abdicated use of the CPU; they do not indicate problems with other resources.

Involuntary wait states

Involuntary wait states are not requested by processes but are invoked by the system to achieve process synchronization in certain circumstances. The free page wait (FPG), page fault wait (PFW), and collided page wait (COLPG) states are associated with memory management and are discussed in Chapter 5, Secondary Page Cache. The current section is concerned with the miscellaneous resource wait (MWAIT) state.

MWAIT State

The presence of processes in the MWAIT state indicates that there might be a shortage of a systemwide resource (usually page or swapping file capacity) and that the shortage is blocking these processes from the CPU.

If you see processes in this state, do the following:

- Check the type of resource wait by examining the MONITOR PROCESSES data available in the collected recording files.
- Check the resource wait states by playing back the data files and examining each PROCESSES display. Note that a standard summary report contains only the last PROCESSES display and the multifile summary report contains no PROCESSES data.

- Issue a MONITOR command like the following:

```
$ MONITOR /INPUT=SYS$MONITOR:file-spec -
_ $ /VIEWING_TIME=1 PROCESSES
```

This command will display all the PROCESSES data available in the input file.

- Look for RWxxx scheduling states, where xxx is a three-character code indicating the depleted resource for which the process is waiting. (The codes are listed in the *OpenVMS System Management Utilities Reference Manual* under the description of the STATES class in the MONITOR section.)

Mutex wait state (indicated by the state keyword MUTEX in the MONITOR PROCESSES display) is a temporary wait state and is not discussed here.

Other Types of Resource Wait States

The most common types of resource waits are those signifying depletion of the page and swapping files as shown in the following table:

State	Description
RWSWP	Indicates a swapping file of deficient size
RWMBP ¹	Can indicate a paging file that is too small
RWAST	Indicates that the process is waiting for a resource whose availability will be signaled by delivery of an asynchronous system trap (AST)

¹Also applies to RWMPE and RWPGF.

You can determine paging and swapping file sizes and the amount of available space they contain by entering the SHOW MEMORY/FILES/FULL command.

The AUTOGEN feedback report provides detailed information about paging and swapping file use. AUTOGEN uses the data in the feedback report to resize or to recommend resizing the paging and swapping files.

**Obtaining
MONITOR
statistics**

Use the following MONITOR commands to obtain the appropriate statistic:

Command	Statistic
Compute Queue	
STATES	Number of processes in compute (COM) and compute outswapped (COMO) scheduling states
Estimating CPU Capacity	
STATES	All items
MODES	Idle time
Voluntary Wait States	
STATES	Number of processes in local event flag wait (LEF), common event flag wait (CEF), hibernate (HIB), and suspended (SUSP) states
LOCK	ENQs Forced to Wait Rate
Involuntary Wait States	
STATES	Number of processes in miscellaneous resource wait (MWAIT) state
PROCESSES	Types of resource waits (RW _{xxx})

Improving CPU Responsiveness**Prerequisites**

It is always good practice to review the methods for improving CPU responsiveness—equitable CPU sharing, reducing resource consumption by the system, CPU load balancing, and CPU offloading—to see if there are ways to recover CPU power.

Before taking action to correct CPU resource problems, do the following:

- Complete your evaluation of all the system's resources.
- Resolve any pending memory or disk I/O responsiveness problems before attempting to improve CPU responsiveness.

Equitable CPU sharing

If you have concluded that a large compute queue is affecting the responsiveness of your CPU, try to determine whether the resource is being shared on an equitable basis. Ask yourself the following questions:

- Have you assigned different base priorities to different classes of users?
- Is your system supporting one or more real-time processes?
- Are some users complaining about poor service while others have no problems?

The operating system uses a round-robin scheduling technique for all nonreal-time processes at the same scheduling priority. However, there are 31 priority levels, and as long as a higher level process is ready to use the CPU, none of the lower level processes will execute. A compute-bound process whose base priority is elevated above that of other processes can usurp the CPU. Conversely, the CPU will service processes with base priorities lower than the system default only when no other processes of default priority are ready for service.

Do not confuse inequitable sharing with the priority-boosting scheme of the operating system, which gives temporary priority boosts to processes encountering certain events, such as I/O completion. These boosts are temporary and they cannot cause inequities.

Detecting Inequitable CPU Sharing

You can detect inequitable sharing by using either of the following methods:

- Examine the CPU Time column of the MONITOR PROCESSES display in a standard summary report (not included in the multiframe summary report). A process with a CPU time accumulation much higher than that of other processes could be suspect.
- Use the MONITOR playback feature to obtain a display of the top CPU users during each collection interval. (This is the preferred method.) To view the display, enter a command of the form:

```
$ MONITOR /INPUT=SYS$MONITOR:file-spec -
_ $ /VIEWING_TIME=1 PROCESSES /TOPCPU
```

You might want to select a specific time interval using the /BEGINNING and /ENDING qualifiers if you suspect a problem. Check whether the top process changes periodically.

CPU Allocation and Processing Requirements

It can sometimes be difficult to judge whether processes are receiving appropriate amounts of CPU allocation because the allocation depends on their processing requirements.

IF...	THEN...
the MONITOR collection interval is too large to provide a sufficient level of detail	enter the command on the running system (live mode) during a representative period using the default three-second collection interval.
there is an inequity	try to obtain more information about the process and the image being run by entering the DCL command SHOW PROCESS/CONTINUOUS.

Reduction of CPU consumption by the system

Depending on the amount of service required by your system, operating system functions can consume anywhere from almost no CPU cycles to a significant amount. Any reductions you can make in services represent additional available CPU cycles. These can be used by processes in the COM state, thereby lowering the average size of the compute queue and making the CPU more responsive.

The information in this section will help you identify the system components that are using the CPU. You can then decide whether it is reasonable to reduce the involvement of those components.

Processor Modes

The principal body of information about system CPU activity is contained in the MONITOR MODES class. Its statistics represent rates of clock ticks (1-millisecond units) per second; but they can also be viewed as percentages of time spent by the CPU in each of the various processor modes.

Note that interrupt time is really kernel mode time that cannot be charged to a particular process. Therefore, it is sometimes convenient to consider these two together.

The following table lists of some of the activities that execute in each processor mode:

Mode	Activity
Interrupt ^{1,2}	CPU time spent handling interrupts from peripheral devices such as disks, tapes, printers, and terminals. The majority of system scheduling code executes in interrupt state because for most of the time spent executing that code, there is no current process.
MP Synchronization	Time spent by a processor in a multiprocessor system waiting to acquire a spin lock.
Kernel ²	Most local system functions execute in kernel mode. These include local lock requests, file system (XQP) requests, memory management, and most system services (including \$QIO).
Executive	The major consumer of executive mode time is RMS. Some optional products such as ACMS, DBMS, and Rdb also run in executive mode.
Supervisor	The command language interpreters DCL and MCR execute in this mode.
User	Most user-written code executes in this mode.
Idle	Time during which all processes are in scheduling wait states and there are no interrupts to service.

¹In a VMScluster configuration, services performed on behalf of a remote node execute in interrupt state because there is no local process to which the time can be charged. These include functions involving system communication services (SCS), such as remote lock requests and MSCP requests.

²As a general rule, the combination of interrupt time and kernel mode time should be less than 40 percent of the total CPU time used.

Although MONITOR provides no breakdown of modes into component parts, you can make inferences about how the time is distributed within a mode by examining some of the other MONITOR classes in your summary report and through your knowledge of the work load.

Interrupt Time

In VMScluster systems, interrupt time per node might be higher than in noncluster systems because of the remote services performed. However, if this time appears excessive, you should investigate the remote services and look for deviations from typical values. Enter the following commands:

- **MONITOR DLOCK**—Observe the distributed lock manager activity. Activity labeled incoming and outgoing is executed in interrupt state.

- **MONITOR SCS/ITEM=ALL**—Observe internode traffic over the computer interconnect (CI).
- **MONITOR MSCP_SERVER**—Observe the MSCP server activity.
- **SHOW DEVICE /SERVED /ALL**—Observe the MSCP server activity.

Even though VMScluster systems can be expected to consume marginally more CPU resources than noncluster systems because of this remote activity, there is no measurable loss in CPU performance when a system becomes a member of a VMScluster. VMSclusters achieve their sense of “clusterness” by making use of SCS, a very low overhead protocol. Furthermore, in a quiescent cluster with default system parameter settings, each system needs to communicate with every other system only once every five seconds.

MP Synchronization Time

MP synchronization time is a measure of the contention for spin locks in a multiprocessing (MP) system. A certain amount of time in this mode is expected for MP systems. However, MP synchronization time above roughly 8% of total processing time usually indicates a moderate to high level of paging, I/O, or locking activity. You should evaluate the usage of those resources by examining the IO, DLOCK, PAGE, and DISK statistics.

Kernel Mode Time

High kernel mode time (greater than 25%) can indicate several conditions warranting further investigation:

- A memory limitation. In this case, the MONITOR IO class should indicate a high page fault rate, a high inswap rate, or both. Refer to Chapter 10, Understanding the Memory Resource, for information on the memory resource.
- Excessive local locking. Become familiar with the locking rates (New ENQ, Converted ENQ, and DEQ) shown in the MONITOR LOCK class, and watch for deviations from the typical values. (In VMScluster environments, use the DLOCK class instead; only the local portion of each of the locking rates is executed in kernel mode.)
- A high process creation rate. Process creation is a CPU-intensive operation. Process accounting can help determine if this activity is contributing to the high level of kernel mode time.
- Excessive file system activity. The file system, also known as the XQP, performs various operations on behalf of users and RMS. These include file opens, closes, extends, deletes, and window turns (retrieval of mapping pointers). The CPU Tick Rate of the MONITOR FCP class can be viewed as a percentage of the CPU being consumed by the file system. It

is highly dependent on application file handling and can be kept to a minimum by encouraging efficient use of files, by performing periodic backups to minimize disk fragmentation, and so forth. The Erase Rate of the FCP class is the rate of erase operations performed to support the high-water marking security feature. If you do not require this feature at your installation, be sure to set your volumes to disable it. (See Chapter 4, Postinstallation System Management Options.)

Rate ¹	Description
CPU tick rate	Can be viewed as the percentage of the CPU being consumed by the file system. It is highly dependent on application file handling and can be kept to a minimum by encouraging efficient use of files, by performing periodic backups to minimize disk fragmentation, and so forth.
Erase rate ²	Is the rate of erase operations performed to support the high-water marking feature.

¹MONITOR FCP class.

²If you do not require this feature at your site, be sure to set your volumes to disable it. (See Chapter 4, Postinstallation System Management Options.)

- Excessive direct I/O rate. While direct I/O activity, particularly disk I/O, is important in an evaluation of the I/O resource, it is also important in an evaluation of the CPU resource because it can be costly in terms of CPU cycles. The direct I/O rate is included in the MONITOR IO class. The top users of direct I/O are indicated in the MONITOR PROCESSES /TOPDIO class.
- A high image activation rate. The image activation code itself does not use a significant amount of CPU time, but it can cause consumption of kernel mode time by activities like the following:
 - An excessive amount of logical name translation as file specifications are parsed.
 - Increased file system activity to locate and open the image and associated library files (this activity also generates buffered I/O operations).
 - A substantial number of page faults as the images and libraries are mapped into working sets.
 - A high demand zero fault rate (shown in the MONITOR PAGE class). This activity might be accompanied by a high global valid fault rate, a high page read I/O (hard fault) rate, or both.

A possible cause of a high image activation rate is excessive use of DCL command procedures. You should expect to see high levels of supervisor mode activity if this is the case. Frequently invoked, stable command procedures are good candidates to be rewritten as images.

- Excessive use of DECnet. Become familiar with the packet rates shown in the MONITOR DECNET class and watch for deviations from the typical values.

Executive Mode Time

High levels of executive mode time can be an indication of excessive RMS activity. File design decisions and access characteristics can have a direct impact on CPU performance. For example, consider how the design of indexed files may affect the consumption of executive mode time:

- Bucket size determines average time to search each bucket.
- Fill factor and record add rate determine rate of bucket splits.
- Index, key, and data compression saves disk space and can reduce bucket splits but requires extra CPU time.
- Use of alternate keys provides increased retrieval flexibility but requires additional disk space and additional CPU time when adding new records.

Be sure to consult the *Guide to OpenVMS File Applications* when designing an RMS application. It contains descriptions of available alternatives along with their performance implications.

CPU offloading

The following are some techniques you might use to reduce demand on the CPU:

- Decompress the system libraries. (See Chapter 4, Postinstallation System Management Options.)
- Force compute-intensive images to execute only in a batch queue, with a job limit. A good technique for enforcing such batch execution is to use the access control list (ACL) facility as follows:

```
$ SET FILE file-spec -  
_ $ /ACL = (IDENTIFIER=INTERACTIVE+NETWORK,ACCESS=NONE)
```

This command will force batch execution of the image file for which the command is entered.

- Implement off-shift timesharing or set up batch queues to spread the CPU load across the hours when the CPU would normally not be used.
- Disable code optimization. Compilers such as FORTRAN and Bliss do some code optimizing by default. However, code optimization is a CPU- and memory-intensive operation. It might be beneficial to disable optimization in environments where frequent iterative compilations are done. Such activity

is typical of an educational environment where students are learning a new language.

- Use a dedicated batch engine. It might be beneficial during prime time to set up in a VMScluster one system dedicated to batch work, thereby isolating the compute-intensive, noninteractive work from the online users. You can accomplish this by making sure that the cluster-accessible generic batch queue points only to executor batch queues defined on the batch system. If a local area terminal server is used for terminal access to the cluster, you can limit interactive access to the batch system by making that system unknown to the server.

CPU offloading between processors on the network

Users of standalone workstations on the network can take advantage of local and client/server environments when running applications. Such users can choose to run an application based on DECwindows on their workstations, resources permitting, or on a more powerful host sending the display to the workstation screen. From the point of view of the workstation user, the decision is based on disk space and acceptable response time.

Although the client/server relationship can benefit workstations, it also raises system management questions that can have an impact on performance. On which system will the files be backed up—workstation or host? Must files be copied over the network? Network-based applications can represent a significant additional load on your network depending on interconnect bandwidth, number of processors, and network traffic.

CPU load balancing in a VMScluster

You can improve responsiveness on an individual CPU in a VMScluster by shifting some of the work load to another, less used processor. You can do this by setting up generic batch queues or by assigning terminal lines to such a processor. Some terminal server products perform automatic load balancing by assigning users to the least heavily used processor.

Note

Do not attempt to load balance among CPUs in a VMScluster until you are sure that other resources are not blocking (and thus not inflating idle time artificially) on a processor that is responding poorly—and until you have already done all you can to improve responsiveness on each processor in the cluster.

Assessing Relative Load

Your principal tool in assessing the relative load on each CPU is the MODES class in the MONITOR multifile summary. Compare the Idle Time figures for all the processors. The processor with the most idle time might be a good candidate for offloading the one with the least idle time.

On a VMScluster member system where low-priority batch work is being executed, there might be little or no idle time. However, such a system can still be a good candidate for receiving more of the VMScluster work load. The interactive work load on that system might be very light so that it would have the capacity to handle more default-priority work at the expense of the low-priority work.

There are several ways to tell whether a seemingly 100% busy processor is executing mostly low-priority batch work as follows:

- Enter a MONITOR command like the following and observe the TOPCPU processes:

```
$ MONITOR /INPUT=SYS$MONITOR:file-spec -  
_ $ /VIEWING_TIME=1 PROCESSES /TOPCPU
```
- Examine your batch policies to see whether the system is favored for such work.
- Use the ACCOUNTING image report described in Chapter 7, Collecting and Interpreting Image-Level Accounting Data (or a similarly generated process accounting report) to examine the kind of work being done on the system.

Other VMScluster load-balancing techniques

The following are some techniques for VMScluster load balancing. Once you have determined the relative CPU capacities of individual member systems, you might do any of the following:

- Use a local area terminal server to distribute interactive users.
- Increase the job limit for batch queues on high-powered systems. The distributed job controller attempts to balance the number of currently executing batch jobs with the batch queue job limit, across all executor batch queues pointed to by a generic queue. You can increase the percentage of jobs that the job controller will assign to the higher powered CPU by increasing the job limit of the executor batch queues on that system.
- Design batch work loads to execute in parallel across a VMScluster. For example, a large system-build procedure could be redesigned so that all nodes in the VMScluster would participate in the compilation and link phases. Synchronization would be required between the two phases and could be accomplished with the DCL command SYNCHRONIZE.

- Reallocate lock directory activity. You might want to let the more powerful processors handle a larger portion of the distributed lock manager directory activities. This can be done by increasing the system parameter LOCKDIRWT above the default value of 1 on the more powerful machines. Note that this approach can be beneficial only in VMSclusters that support high levels of lock directory activity.

Obtaining MONITOR statistics

Use the following MONITOR commands to obtain the appropriate statistic:

Command	Statistic
Reducing CPU Consumption	
MODES	All items
Interrupt State	
IO	Direct I/O Rate, Buffered I/O Rate, Page Read I/O Rate, Page Write I/O Rate
DLOCK	All items
SCS	All items
MP Synchronization Mode	
MODES	MP Synchronization
IO	Direct I/O Rate, Buffered I/O Rate
DLOCK	All items
PAGE	All items
DISK	Operation Rate
Kernel Mode	
MODES	Kernel mode
IO	Page Fault Rate, Inswap Rate, Logical Name Translation Rate
LOCK	New ENQ Rate, Converted ENQ Rate, DEQ Rate
FCB	All items
PAGE	Demand Zero Fault Rate, Global Valid Fault Rate, Page Read I/O
DECNET	Sum of packet rates
CPU Load Balancing	
MODES	Time spent by processors in each mode

See Table B-1 for a summary of MONITOR data items.

The Memory Resource

Overview

This chapter discusses the following topics:

- Understanding the memory resource
- Evaluating memory responsiveness
- Improving memory responsiveness

Purpose

To evaluate the performance of the memory resource.

Definitions

Locality of reference is a characteristic of a program that indicates how close or far apart the references to locations in virtual memory are over time. A program with a high degree of locality does not refer to many widely scattered virtual addresses in a short period of time.

The **nonpaged pool area** is a portion of physical memory permanently allocated to the system for the storage of data structures and device drivers.

The **system working set** is an area of physical memory reserved to satisfy page faults of virtual addresses in system space.

Understanding the Memory Resource

Similarities and differences

The memory resource shares some similarities with the other resources, but it exhibits some notable differences. It is similar to CPU and disk in that it is a single resource pool that must be shared, but different in the sense that it can be separated into pieces of varying size, all of which can be allocated to processes simultaneously. A process can retain its allocation of memory until memory is demanded by other processes (page faulting), at which time the sizes of the pieces are reconfigured. In some cases, certain processes must wait longer for their allocations (swapping).

Working set size

The key to good performance of the memory subsystem is to maintain working sets of appropriate size for resident processes. As a rule, the total of all resident process working set quotas should be within the amount of free memory available on the system. When there is abundant free memory available, the borrowing mechanism of the memory management subsystem allows working sets to grow to the value specified in the user authorization file by WSEXTENT. However, you should set the WSQUOTA value so that user programs can have reasonable faulting behavior even if they can grow only to WSQUOTA. See Chapter 6, AWSA, for guidelines on estimating appropriate WSQUOTA values.

Locality of reference

Erratic code and data reference patterns by user programs can cause memory to be used inefficiently. The effectiveness of a virtual memory system is based upon good locality of reference. If an application has been designed with poor virtual address reference patterns, it can require an extremely large WSQUOTA value to perform satisfactorily. In addition, applications such as AI and CAD/CAM, which perform an inordinately large amount of dynamic memory allocation, often require very large WSQUOTA values.

Obtaining working set values

One way to obtain information about working set values on the running system (Example 10-2) is to use the procedure shown in Example 10-1. You might want to execute it several times during some representative period of loading to gain an idea of the steady-state working set requirements for your system.

Example 10-1 Procedure to Obtain Working Set Information

```
$ !
$ ! WORKSET.COM - Command file to display working set information.
$ ! Requires 'WORLD' privilege to display other processes.
$ a = ""
$ pid = ""
$ context = ""
$ IF pl.NES. "" THEN pid = pl
$ WRITE sys$output -
"                               Working Set Information"
$ WRITE sys$output ""
$ WRITE sys$output -
"                               WS    WS    WS    WS  Pages  Page"
$ WRITE sys$output -
"Username  Processname  State  Extnt  Quota  Deflt  Size in WS faults  Image"
$ WRITE sys$output ""
$ START:
$ IF pl.EQS. "" THEN pid = F$PID(context)
$ IF pid.EQS. "" THEN EXIT
$ pid = a+pid+a
$ username = F$GETJPI('pid,"USERNAME")
```

(continued on next page)

Example 10-1 (Cont.) Procedure to Obtain Working Set Information

```

$ IF username.EQS."" THEN GOTO START
$ processname = F$GETJPI('pid,"PRCNAM")
$ imagename = F$GETJPI('pid,"IMAGNAME")
$ imagename = F$PARSE(imagename,,,"NAME")
$ state = F$GETJPI('pid,"STATE")
$ wsdefault = F$GETJPI('pid,"DFWSCNT")
$ wsquota = F$GETJPI('pid,"WSQUOTA")
$ wsextent = F$GETJPI('pid,"WSEXTENT")
$ wssize = F$GETJPI('pid,"WSSIZE")
$ globalpages = F$GETJPI('pid,"GPCCNT")
$ processpages = F$GETJPI('pid,"PPGCNT")
$ pagefaults = F$GETJPI('pid,"PAGEFLTS")
$ pages = globalpages + processpages
$ text = F$FAO("!AS!15AS!5AS!5(6SL)!7SL!AS",-
  username,processname,state,wsextent,wsquota,wsdefault,wssize,-
  pages,pagefaults," "+imagename)
$ WRITE sys$output text
$ IF p1.NES."" THEN EXIT
$ GOTO START

```

**Displaying
working set
values**

The WORKSET.COM procedure produces the following display:

Example 10-2 Displaying Working Set Values

Working Set Information										
Username	Processname	State	WS Extnt	WS Quota	WS Deflt	WS Size	Pages in WS	Page faults	Page Image	
SYSTEM	ERRFMT	HIB	1024	512	100	60	60	165	ERRFMT	
SYSTEM	CACHE_SERVER	HIB	1024	512	100	512	75	55	FILESERV	
SYSTEM	CLUSTER_SERVER	HIB	1024	512	100	60	60	218	CSP	

(continued on next page)

Example 10-2 (Cont.) Displaying Working Set Values

SYSTEM	OPCOM	LEF	2048	512	100	210	59	5764	OPCOM
SYSTEM	JOB CONTROL	HIB	1024	512	100	360	238	1459	JOBCTL
SYSTEM	CONFIGURE	HIB	1024	512	100	125	121	101	CONFIGURE
SYSTEM	SYMBIONT_0001	HIB	1024	512	100	668	57	67853	PRTSMB
DECNET	NETACP	HIB	1500	750	175	1200	812	10305	NETACP
DECNET	EVL	HIB	1024	350	175	210	33	84080	EVL
SYSTEM	REMACP	HIB	1024	350	175	60	47	74	REMACP
SYSTEM	VAXsim Monitor	HIB	1024	200	100	350	210	1583	VAXSIM
SYSTEM	DBMS MONITOR	LEF	1000	512	150	62	62	488	DBMMON
SYSTEM	TINKERBELLE	LEF	1024	350	175	325	177	1627	
SYSTEM	NULF	COM	1024	350	250	350	246	1007	FAC
HALL	CFAI	COM	2400	1024	512	662	358	567	CFAI
VTXUP	VTX_SERVER	LEF	2400	1024	512	962	696	624	VTXSRV
WEINSTEIN	Jane	LEF	2400	1024	512	662	432	13132	EDT
HURWITZ	HURWITZ	LEF	2400	1024	512	512	350	4605	
CARMODY	CARMODY	LEF	2400	1024	512	812	546	16822	MAIL
CAPARILLIO	CAPARILLIO	CUR	2400	1024	512	512	282	10839	
STRATFORD	Kathy	LEF	2400	1024	512	512	210	9852	
FREY	_VTA270:	LEF	2400	1024	512	512	163	1021	
CHRISTOPHER	_VTA271:	LEF	2400	1024	512	512	252	379	
STANLEY	_STANLEY	LEF	2048	1024	512	512	295	10369	
MINSKY	MINSKY	LEF	2400	1024	512	512	143	60316	
TESTGEN	TESTGEN	LEF	4100	1024	512	234	84	75753	
CLAYMORE	Cluster Buster	LEF	2400	1024	512	1262	932	1919	CREATOR
DINEAUX	Sally	LEF	2400	1024	512	512	330	31803	
DECNET	SERVER_0848	LEF	1024	350	175	325	183	647	NETSERVER
LUZ	Lars	LEF	2400	1024	512	1024	980	95420	TEX
DECNET	MAIL_222	LEF	1024	350	175	325	234	526	MAIL
STEVENS	STEVENS	LEF	2400	1024	512	512	221	7851	
ZEN	_VTA259:	LEF	2400	1024	512	1024	319	4267	SHOW
ZEN	_ZEN_2	LEF	2400	1024	512	512	171	3026)	

Field	Description
WS Deflt	Refers to the default working set size, which is reestablished at each image activation.
WS Size	Is the current size of the working set. When the number of pages actually allocated (Pages in WS) reaches this threshold, subsequent page faults will cause page replacement.
Pages in WS	Includes both private and global pages.
WS Extnt	Represents threshold values to which WS Size can be adjusted.
WS Quota	
Page faults	Shows the total number of faults that have occurred since process creation.

Evaluating Memory Responsiveness

Memory allocation

The key measure of responsiveness for the memory management subsystem is the amount of time required for a process to be allocated its share of memory.

Since allocation time is not measured directly, you should be concerned with the rates of the two memory management activities that extend the processing time experienced by processes in a virtual memory system—namely, *page faulting* and *swapping*. These activities not only incur overhead on the CPU and disk resources, but they also block the execution of processes during the time the system needs to allocate memory and the time the processes spend waiting for memory allocation.

Thus, your goal in evaluating the memory resource is to ensure that faulting and swapping rates are kept within reasonable bounds.

Page faulting

Whenever a process references a virtual page that is not in its working set, a page fault occurs. For process execution to continue, memory management software is called to acquire and map a physical page into the working set.

Hard and Soft Page Faults

The fault can be **hard** or **soft**. A hard fault (measured by the Page Read I/O Rate item in the MONITOR PAGE class) is one that requires a read operation from a page or image file on disk. A soft fault is one that is satisfied by mapping to a page already in memory; this can be a global page or a page in the secondary page cache. (The secondary page cache consists of the free-page and the modified-page list; the primary page cache is each process's working set.) The following categories of soft faults are measured and reported in the MONITOR PAGE class:

- **Free List Fault Rate**—The rate of page faults satisfied by reclaiming from the free-page list a page that was previously allocated to a process. An excessive rate of free-page list faults can occur when working set quotas are too small, causing excessive page replacement.
- **Modified List Fault Rate**—The rate of page faults satisfied by reclaiming a page from the modified-page list. An excessive rate of modified-page list faults can occur when working set quotas are too small.
- **Demand Zero Fault Rate**—The rate of page faults satisfied by allocating a free page and initializing its contents to zero. This type of fault is typically seen during image activation and whenever the virtual address space is expanded.

- **Global Valid Fault Rate**—The rate of page faults satisfied by mapping a shared page that is already valid (one already in another process's working set). Swapping or image activation can cause an elevated global valid fault rate.
- **Write in Progress Fault Rate**—The rate of page faults satisfied by mapping to a page that is in the process of being written back to disk. The rate for this type of fault is typically very low.

The total Page Fault Rate is equal to the sum of the hard fault rate (Page Read I/O Rate) plus the soft fault rate, which is the sum of the five categories listed above.

System Fault Rate is the rate of faults for which the referenced virtual address is in system space (hex address 80000000 and above). It is not included in the overall Page Fault Rate, and is discussed separately in reducing resource consumption by the system in the section on Improving Memory Responsiveness.

Your own judgment, based on familiarity with the data in your MONITOR summaries, is the best determinant of an acceptable Page Fault Rate for your system.

When any of the following thresholds are exceeded, you should attempt to improve memory responsiveness. (See Improving Memory Responsiveness.)

- Hard faults (Page Read I/O Rate) should be kept as low as possible, but to no more than 10% of the overall Page Fault Rate. When the hard fault rate exceeds this threshold, you can assume that the secondary page cache is not being used efficiently.
- Overall Page Fault Rate begins to become excessive when more than 1% of the CPU is devoted to soft faulting (faulting that involves no disk I/O).

While these rules do not represent absolute upper limits, rates that exceed the suggested limits are warning signs that the memory resource should either be improved by one of the four means listed in the section on Improving Memory Responsiveness, or that a memory upgrade should perhaps be considered. Note, however, that more memory will not reduce the number of page faults caused by image activation.

Secondary Page Cache

Paging problems typically occur when the secondary page cache (free-page list and modified-page list) is too small. This systemwide cache, which is sized by AUTOGEN, should be large enough to ensure that the overall fault rate is not excessive and that most faults are soft faults.

When evaluating paging activity on your system, you should check for processes in the free page wait (FPG), collided page wait (COLPG), and page fault wait (PFW) states and note departures from normal figures. The presence of processes in the FPG state almost always indicates serious memory management problems because it implies that the free-page list has been depleted.

Processes in the PFW and COLPG states are waiting for hard faults (from disk) to be satisfied. Note, however, that while hard fault waiting is undesirable, it is not as serious as swapping.

An average free-page list size that is between the values of the FREELIM and FREEGOAL system parameters is usually an indicator of deficient memory and is often accompanied by a high page fault rate. If either condition exists, or if the hard fault rate exceeds the recommended percentage, you must consider enlarging the free- and modified-page lists, if possible. Enlarging the secondary page cache could reduce hard faulting, provided such faulting is not the result of image activation.

The easiest way to increase the free page cache is to increase the value of FREEGOAL. Proactive reclamation will then attempt to recover more memory from idle processes. Typically, overall fault rates decrease when proactive reclamation is enabled because memory is more readily available to active processes.

A high rate of modified-page writing, for example, as shown in the Page Write I/O Rate field of the MONITOR PAGE display, is an indication that the modified-page list might be too small. A write rate of 1 every 2 seconds is fairly high. The modified-page list should be large enough to provide an equilibrium between the rate at which pages are added to the list versus the modified-page list fault rate without causing excessive list writing by reaching MPW_HILIMIT. If you do adjust the size of the modified-page list using MPW_HILIMIT, make sure you retain the relationship among MPW_HILIMIT, MPW_WAITLIMIT, and MPW_LOWAITLIMIT by using AUTOGEN.

If you are able to increase the size of the free-page list, you can then allocate more memory to the modified-page list. Using AUTOGEN, you can increase the modified-page list by adjusting the appropriate MPW system parameters. (See the *OpenVMS System Management Utilities Reference Manual* for a description of MPW parameters.)

Swapping and swapper trimming

Swapping, when considered in isolation, is an expensive operation. It can place a huge transfer load on the I/O subsystem instantaneously. Swapping also can place heavy demand on CPU resources. However, when used as part of the proactive memory reclamation policy, swapping results in improved—that is, reduced—memory consumption and a lower page fault rate.

Good and Bad Swapping

There is good swapping and bad swapping. The latter occurs as the last step of reactive memory reclamation when the free-page list is exhausted—that is, when it is smaller than FREELIM. However, having a significant number of outswapped processes on your system when proactive memory reclamation is enabled is not a cause for alarm. A much more reliable indicator that harmful swapping is occurring is a high inswap rate—for example, greater than one process per second.

Artificially Induced Swapping

Before attempting to improve a system with a high inswap rate, do the following:

- Check for a condition known as **artificially induced swapping**. This condition occurs when there are no available balance set slots.
- Check the BALSETCNT system parameter. Swapping might have been artificially induced because BALSETCNT is set too low.

See Chapter 17, Solutions for Memory-Limited Behavior. You can obtain information on balance slots with the DCL command SHOW MEMORY.

A possible, although unlikely, reason for a high inswap rate might be an overly large value for FREEGOAL when proactive memory reclamation is enabled. Although this policy outswaps only long-waiting processes, a very large value for FREEGOAL will cause the outswapping of many long-waiting processes over time, thus increasing the inswap rate as these processes become computable.

Obtaining MONITOR statistics

Use the following MONITOR commands to obtain the appropriate statistics:

Command	Statistic
Page Faulting	
PAGE	All items
Secondary Page Cache	
STATES	Number of processes in the free page wait (FPG), collided page wait (COLPG), and page fault wait (PFW) states

Command	Statistic
Swapping and Swapper Trimming	
STATES	Number of outswapped processes
IO	Inswap Rate

Improving Memory Responsiveness

It is always good practice to check the four methods for improving memory responsiveness to see if there are ways to free up more memory, even if no problem seems to exist currently. The easiest way to improve memory utilization significantly is to make sure that proactive memory reclamation is enabled.

Equitable memory sharing

When proactive memory reclamation is enabled, the system distributes memory among active processes in an equitable and expeditious manner. If you feel page faulting is excessive with this policy enabled, make sure processes have not reached their WSEXTENT values. Note that precise WSQUOTA values are not very important when this policy is enabled, provided that GROWLIM and BORROWLIM are set equal to FREELIM using AUTOGEN.

If proactive memory reclamation is not enabled (that is, the value of MMG_CTLFLAGS is 0), then overall system page fault behavior is highly dependent on current process WSQUOTA values. The following discussion can help you to determine if inequitable memory sharing is occurring.

Because page fault behavior is so heavily dependent on the page referencing patterns of user programs, the WSQUOTA values you assign might be satisfactory for some programs but not for others. Use the ACCOUNTING image report described in Chapter 7, *Collecting and Interpreting Image-Level Accounting Data*, to identify the programs (images) that are the heaviest faulters on your system, and then compensate by encouraging users to run such images as batch jobs on queues you have set up with large WSQUOTA values.

Inequitable Sharing

You might be able to detect inequitable sharing by looking at the Faults column of the MONITOR PROCESSES display in a standard summary report (it is not contained in the multiframe summary report). A process with a page fault accumulation much higher than that of other processes might be suspect, although it depends on how long the process has been active. A better means

of detection is to use the MONITOR playback feature to view a display of the top page faulters during each collection interval:

```
$ MONITOR /INPUT=SYS$MONITOR:file-spec -  
_ $ /VIEWING_TIME=1 PROCESSES /TOPFAULT
```

You might want to select a time interval using the /BEGINNING and /ENDING qualifiers when you suspect that a problem has occurred.

Check to see whether the top process changes periodically. If it appears that one or two processes are consistently the top faulters, you might want to obtain more information about which images they are running and consider upgrading their WSQUOTA values, using the guidelines in Chapter 6, AWSA. Sometimes a small adjustment in a WSQUOTA value can make a drastic difference in the page faulting behavior if the original value was near the *knee* of the working-set/page-fault curve. (See Figures 6-2, 6-3, and 6-4.)

If you find that the MONITOR collection interval is too large to provide sufficient detail, try entering the previous command on the running system (live mode) during a representative period, using the default 3-second collection interval. If you discover an inequity, try to obtain more information about the process and the image being run by entering the SHOW PROCESS /CONTINUOUS command.

Another way to check for inequitable sharing of memory is to use the WORKSET.COM command procedure described in Understanding the Memory Resource. Examine the various working set values and ensure that the allocation of memory, even if not evenly distributed, is appropriate.

Reduction of memory consumption by the system

The operating system uses physical memory for storage of the code and data structures it requires to support user processes. You have control over the sizes of two of the memory areas reserved for the system: the system working set and the nonpaged pool area. Both of these areas are sized by AUTOGEN. The sizes set by AUTOGEN are normally adequate but might not be optimal because AUTOGEN cannot anticipate all operational requirements.

System Working Set

Virtual addresses in the system working set can be code or data (paged pool, for example). Since the same system working set is used for all processes on the system, there is very little locality associated with it.

Therefore, the system fault rate can be expected to change slowly in relation to changes in the system working set size (as controlled by the system parameter SYSMWCNT). A rule of thumb is to try to keep the system fault rate to less than 2 per second.

Keep in mind, however, that pages allocated to the system working set by raising the value of SYSMWCNT are considered permanently allocated to the system and are therefore no longer available for process working sets.

Nonpaged Pool

AUTOGEN determines the initial size of the nonpaged pool, but automatic expansion will occur if necessary. The system expands pool as required by permanently allocating a page of memory from the free-page list. Pages allocated in this manner are not available for use by process working sets until the system is rebooted.

Adaptive Pool Management

The high-performance nonpaged pool allocator reduces the probability of system outages due to exhaustion of memory allocated for system data structures (pool). Adaptive pool management virtually eliminates the need to actively manage the allocation of pool resources. The nonpaged pool area and lookaside lists are combined into one region (defined by the system parameters NPAGEDYN and NPAGEVIR), allowing memory packets to migrate from lookaside lists to general pool and back again based on demand. As a result, the system is capable of tuning itself according to the current demand for pool, optimizing its use of these resources, and reducing the risk of running out of these resources.

On OpenVMS AXP systems, it is important to set NPAGEDYN sufficiently large for best performance. If the nonpaged area grows beyond NPAGEDYN, that area will not be included in the large data page granularity hint region (GHR). Applications will experience degraded performance when accessing the expanded nonpaged pool due to an increase in translation buffer (TB) misses.

Internal to the allocator is an array of lookaside lists that contiguously span an allocation range from 1 to 5120 bytes. These lookaside lists require no external tuning. They are automatically prepopulated during bootstrapping based on previous demand and each continuously adapts its number of packets based on changing demand during the life of the system. The result is very high performance due to a very high hit percentage on the internal lookaside lists; typically, over 99 percent.

Deallocation of nonpaged pool has always required that the caller pass the size of the packet being deallocated in either R1 or in the word starting at the eighth byte in the packet itself.

Deallocating Nonpaged Pool When deallocating nonpaged pool, the allocator requires that you pass an accurate packet size either in R1 or the word starting at the eighth byte in the packet itself. The size of the packet determines to which internal lookaside list the packet will be deallocated.

Enabling and Disabling Pool Monitoring The setting of the parameter POOLCHECK at boot time also controls which version of the pool allocator is loaded as follows:

- If POOLCHECK equals a nonzero value, a monitoring version is loaded, which contains the corruption-detecting code and statistics maintenance.

The following System Dump Analyzer (SDA) commands are also enabled:

- SHOW POOL/STATISTICS—Displays the address of the listhead, the list packet size, and the number of attempts, failures, and deallocations made to that list since bootstrapping for each of the internal lookaside lists.
- SHOW POOL/RING_BUFFER—Displays in reverse chronological order information about the last 512 requests made to nonpaged pool. It is useful in analyzing potential corruption problems.

Refer to the *OpenVMS AXP System Dump Analyzer Utility Manual* for more information about these commands.

- If POOLCHECK equals zero, a minimal version is loaded containing no corruption-detecting or statistics maintenance code.

For more information about the POOLCHECK parameter, refer to the *OpenVMS System Management Utilities Reference Manual*.

Nonpaged Pool Granularity The granularity of nonpaged pool is 64 bytes. In other words, the size of any allocated block of nonpaged pool will be an integral multiple of 64. Any code that explicitly assumes the granularity of nonpaged pool to be 16 bytes or makes use of the symbol EXE\$C_ALCGRNMSK to perform (for example) structure alignment, will have to be changed to use the symbol EXE\$M_NPAGGRNMSK, which reflects the nonpaged pool's new granularity.

Additional Consistency Checks

The system parameter SYSTEM_CHECK is used to investigate intermittent system failures by enabling a number of run-time consistency checks on system operation and recording some trace information.

Enabling SYSTEM_CHECK causes the system to behave as if the following system parameter values are set:

Parameter ¹	Value	Description
BUGCHECKFATAL	1	Crashes the system on nonfatal bugchecks

¹Note that the values of the parameters are not actually changed.

Parameter ¹	Value	Description
POOLCHECK	%X616400FF	Enables all pool checking with an allocated pool pattern of %X61616161 ('aaaa') and a deallocated pool pattern of %X64646464 ('dddd')
MULTIPROCESSING	2	Enables full synchronization checking

¹Note that the values of the parameters are not actually changed.

While SYSTEM_CHECK is enabled, the previous settings of the BUGCHECKFATAL and MULTIPROCESSING parameters are ignored.

Setting SYSTEM_CHECK causes certain image files to be loaded that are capable of the additional system monitoring. These image files are located in SYS\$LOADABLE_IMAGES and can be identified by the suffix _MON.

For more information about the SYSTEM_CHECK parameter, refer to the *OpenVMS System Management Utilities Reference Manual*.

Memory offloading

While the most common and probably most cost-effective type of offloading is that performed by shifting the CPU and disk resources onto memory, it is possible to improve memory responsiveness by offloading it onto disk. This procedure is recommended only when sufficient disk resource is available and its use is more cost effective than purchasing additional memory.

Some of the CPU offloading techniques described in CPU offloading apply also to memory. Additional techniques are as follows:

- Install images with the appropriate attributes. When an image is accessed concurrently by more than one process on a routine basis, it should be installed /SHARED so that all processes use the same physical copy of the image. The LIST/FULL command of the Install utility shows the highest number of concurrent accesses to an image installed with the /SHARED qualifier. This information can help you decide whether installing an image is worth the space.
- Favor process swapping over working set trimming for process-intensive applications. There are cases where an image creates several subprocesses that might not be used continuously during the run time. These idle processes take up a share of physical memory, so that it might be wise to swap them out. This typically occurs when users walk away from their terminals for long periods of time.

The following two techniques, used concurrently, will make the system favor swapping out inactive processes over trimming the working sets of highly active processes:

- On a per-process basis—Increase the working set quotas of the active processes, thus reducing reclamation from first-level trimming.
- On a systemwide basis—Increase the value of the system parameter SWPOUTPGCNT perhaps as high as a typical WSQUOTA. As a result, fewer pages will be trimmed, so it is more likely that swapping will occur.

After making adjustments, monitor the inswap rate closely. If it becomes excessive, lower the value of SWPOUTPGCNT.

Evaluating the Swapping File

When you increase swapping, it is important to evaluate the size of the swapping file. If the swapping file is not large enough, system performance will degrade. Use AUTOGEN feedback to size the swapping file appropriately.

Memory load balancing

You can balance the memory load by using some of the CPU load-balancing techniques for VMSclusters described in Chapter 9, Improving CPU Responsiveness, to shift user demand.

To balance the load by reconfiguring memory hardware, perform the following steps:

1. Examine the multifile summary report
2. Look at the Free List Size item of the PAGE class.

The Free List Size item gives the relative amounts of free memory available on each CPU. If a system seems to be deficient in memory and is experiencing memory management problems, perhaps the best solution is to reconfigure the VMScluster by moving some memory from a memory-rich system to a memory-poor one—provided the memory type is compatible with both CPU types.

Note

The Free List Size item is an average of levels, or snapshots. Since it is not a rate, its accuracy is dependent on the collection interval.

**Obtaining
MONITOR
statistics**

Use the following MONITOR commands to obtain the appropriate statistic:

Command	Statistic
Reducing Memory Consumption by the System	
PAGE	System Fault Rate
POOL	All items
Memory Load Balancing	
PAGE	Free List Size

See Table B-1 for a summary of MONITOR data items.

The Disk I/O Resource

Overview

This chapter discusses the following the topics:

- Understanding the disk I/O resource
- Evaluating disk I/O responsiveness
- Improving disk I/O responsiveness

Purpose

To evaluate problems associated with disk I/O responsiveness.

Definition

A **channel** is a logical path connecting a user process to a physical device unit. A user process requests that the operating system assign a channel to a channel so the process can communicate with that device.

Understanding the Disk I/O Resource

Since the major determinant of system performance is the efficient use of the CPU and since a process typically cannot proceed with its use of the CPU until a disk operation is completed, the key performance issue for disk I/O performance is the amount of time it takes to complete an operation.

Measuring disk I/O responsiveness

The principal measure of disk I/O responsiveness is the average amount of time required to execute an I/O request on a particular disk—that disk's *average response time*. It is important to keep average response times as low as possible to minimize CPU blockage. If your system exhibits unusually long disk response times, refer to Improving Disk I/O Responsiveness for suggestions on improving disk I/O performance.

To help you interpret response time as a measure of disk responsiveness, some background information about the components of a disk transfer and the notions of disk capacity and demand is provided in the following sections.

Components of a disk transfer

Table 11–1 shows a breakdown of a typical disk I/O request. CPU time is at least two orders of magnitude less than an average disk I/O elapsed time. You will find it helpful to understand the amount of time each system I/O component consumes in order to complete an I/O request.

Table 11–1 Components of a Typical Disk Transfer (Four- to Eight-Block Transfer Size)

Component	Elapsed Time (%)	Greatest Influencing Factors
I/O Preprocessing	4	Host CPU speed
Controller Delay	2	Time needed to complete controller optimizations
Seek Time	58	Optimization algorithms Disk actuator speed Relative seek range
Rotational Delay	20	Rotational speed of disk Optimization algorithms
Transfer Time	12	Controller design Data density and rotational speed
I/O Postprocessing	4	Host CPU speed

Note that the CPU time required to issue a request is only 8% of the elapsed time and that the majority of the time (for 4- to 8-block transfers) is spent performing a seek and waiting for the desired blocks to rotate under the heads. Larger transfers will spend a larger percentage of time in the transfer time stage. It is easy to see why I/O-bound systems do not improve by adding CPU power.

Disk capacity and demand

As with any resource, the disk resource can be characterized by its capacity to do work and by the demand placed upon it by consumers.

In evaluating disk capacity in a performance context, the primary concern is not the total amount of disk space available but the speed with which I/O operations can be completed. This speed is determined largely by the time it takes to access the desired data blocks (seek time and rotational delay) and by the data transfer capacity (bandwidth) of the disk drives and their controllers.

Seek Capacity

Overall seek capacity is determined by the number of drives (and hence, seek arms) available. Since most disk drives can be executing a seek operation simultaneously with those of other disk drives, the more drives available, the more parallelism you can obtain.

Data Transfer Capacity

A data transfer operation requires a data channel—the path from disk through controller, across buses, to memory. In this context, a channel consists of a single K.sdi adapter of a hierarchical storage controller (HSC). Data transfer on one channel can occur concurrently with data transfer on other channels. For this reason, it is a good idea to attempt to locate disks that have large data transfer operations on separate channels. On an HSC, seek operations can be initiated for other devices on a channel transferring data.

Demand

Demand placed on the disk resource is determined by the user work load and by the needs of the system itself. The demand on a seek arm is the number, size (distance), and arrival pattern of seek requests for that disk. Demand placed on a channel is the number, size, and arrival pattern of data transfer requests for all disks attached to that channel.

In a typical timesharing environment, 90% of all I/O transfers are smaller than 16 blocks. Thus, for the vast majority of I/O operations, data transfer speed is not the key performance determinant; rather, it is the time required to access the data (seek and rotational latency of the disk unit). For this reason, the factor that typically limits performance of the disk subsystem is the number of I/O operations it can complete per unit of time, rather than the data throughput rate. One exception to this rule is swapping I/O, which uses very large transfers. Certain applications, of course, can also perform large data transfers; MONITOR does not provide information about transfer size, so it is important for you to gain as much information as possible about the I/O requirements of applications running on your system. Knowing whether elevated response times are the result of seek/rotational delays or data transfer delays provides a starting point for making improvements.

Evaluating Disk I/O Responsiveness

Average disk response time

The principal measure of disk I/O responsiveness is the average response time of each disk. While not provided directly by MONITOR, it can be estimated using the I/O Operation Rate and I/O Request Queue Length items from the DISK class.

Note

Since, for each disk, the total activity from all nodes in the VMScluster is of primary interest, all references to disk

statistics will be to the Row Sum column of the MONITOR multifile summary instead of the Row Average.

Disk I/O Operation Rate

Disk statistics are provided in the MONITOR DISK class for mounted disks only. I/O Operation Rate is the rate of I/O operations completed on each mounted disk. It includes system I/O (paging, swapping, XQP) and user I/O. While operation rates are influenced by the hardware components of each disk and channel and depend upon transfer size, a general rule of thumb for operations of the size typically seen on timesharing systems can be stated: for most disks, an I/O rate less than 8 per second represents a light load, 15 per second is moderate, and a disk with an operation rate of 25 or more is heavily loaded. These figures are independent of host CPU configuration.

I/O Request Queue Length

The I/O Request Queue Length item is the average number of I/O requests outstanding at any time during the measurement period, including those being serviced and those waiting for service. For example, a queue length of 1.0 indicates that, on the average, there was one request in service throughout the measurement period.

Note

Although this item is an average of levels, or snapshots, its accuracy is NOT dependent on the MONITOR collection interval since it is internally collected once per second.

As useful as these two measurements are in assessing disk performance, an even better measure is that of average response time in milliseconds. It can be estimated from these two items, for each disk, by using the following formula:

$$\text{average response time} = \frac{\text{average queue length}}{\text{average I/O operation rate}} * 1000$$

Average disk response time is an important statistic because it gives you a means of ranking the relative performance of your disks with respect to each other and of comparing their observed performance against a value in the range of 25 to 40 milliseconds. Although faster response times are possible, values in this range represent the best you can reasonably expect to achieve for disks on a timesharing system with little or no contention. Situations that might increase response time include:

- Contention caused by multiple users accessing and transferring data on the same drive or channel
- Large transfer sizes

Since a certain amount of disk contention is expected in a timesharing environment, response times can be expected to be longer than the achievable values.

The response time measurement is especially useful because it indicates the perceived delay from the norm, independent of whether the delay was caused by seek-intensive or data-transfer-intensive operations. Disks with response time calculations significantly larger than achievable values are good candidates for improvements, as discussed later. However, it is worth checking their levels of activity before proceeding with any further analysis. The response time figure says nothing about how often the disk has been used during the measurement period. Improving disks that show a high response time but are used very infrequently might not noticeably improve overall system performance.

In most environments, a disk with a sustained queue length greater than 0.20 can be considered moderately busy and worthy of further analysis. You should try to determine whether activity on disks that show excessive response times and that are, at least, moderately busy, is primarily seek intensive or data transfer intensive. Such disks exhibiting moderate-to-high operation rates are most likely seek intensive, whereas those with low operation rates and large queue lengths (greater than 0.50) tend to be data transfer intensive. (An exception is a seek-intensive disk that is blocked by data transfer from another disk on the same channel; it can have a low operation rate and a large queue length but is not itself data transfer intensive). If a problem still exists after attempting to improve disk performance using the means discussed in Improving Disk I/O Responsiveness, consider upgrading your hardware resources. An upgrade to address seek-intensive disk problems usually centers on the addition of one or more spindles (disk drives), whereas data transfer problems are usually addressed with the addition of one or more data channels.

Note

All the disk measurements discussed in this chapter are averages over a relatively long period of time, such as a prime-time work shift. Significant response-time problems can exist in bursts and might not be obvious when examining long-term averages. If you suspect performance problems during a particular time, obtain a MONITOR multifile summary for that period by playing back the data files you already have, using the /BEGINNING and /ENDING qualifiers to select the period of interest. If you are not sure whether significant peaks of disk activity are occurring, check the I/O Request Queue Length MAX columns of individual summaries of each node. To pinpoint the times when peaks occurred, play back the data file of interest and watch the displays for a CUR value equal to

the MAX value already observed. The period covered by that display is the peak period.

Disk I/O Statistics for MSCP Served Disks

In VMScluster configurations, the MSCP server software is used to make locally attached and HSC disks available to other nodes. A node has **remote access** to a disk when it accesses the disk through another node using the MSCP server. A node has **direct access** when it directly accesses a locally attached or HSC disk.

In the MONITOR MSCP display, an “R” following the device name indicates that the displayed statistics represent I/O operations requested by nodes using remote access. If an “R” does not appear after the device name, the displayed statistics represent I/O operations issued by nodes using direct access. Such I/O operations can include those issued by the MSCP server on behalf of remote requests.

Improving Disk I/O Responsiveness

It is always good practice to check methods for improving disk I/O responsiveness to see if there are ways to use the available capacity more efficiently, even if no problem seems to exist currently.

Equitable disk I/O sharing

If you identify certain disks as good candidates for improvement, check for excessive use of the disk resource by one or more processes. The best way to do this is to use the MONITOR playback feature to obtain a display of the top direct I/O users during each collection interval. The direct I/O operations reported by MONITOR include all user disk I/O and any other direct I/O for other device types. In many cases, disk I/O represents the vast majority of direct I/O activity on OpenVMS AXP systems, so you can use this technique to obtain information on processes that might be supporting excessive disk I/O activity.

Enter a MONITOR command similar to the following:

```
$ MONITOR /INPUT=SYS$MONITOR:file-spec -  
_ $ /VIEWING_TIME=1 PROCESSES /TOPDIO
```

You might want to specify the /BEGINNING and /ENDING qualifiers to select a time interval that covers the problem period.

Examining Top Direct I/O Processes

If it appears that one or two processes are consistently the top direct I/O users, you might want to obtain more information about which images they are running and which files they are using. Since this information is not recorded by MONITOR, it can be obtained in any of the following ways:

- Run MONITOR in live mode. Enter DCL SHOW commands when the situation reoccurs.
- Use the ACCOUNTING image report described in Chapter 7, Collecting and Interpreting Image-Level Accounting Data.
- Survey heavy users of system resources.

Using MONITOR Live Mode

To run MONITOR in live mode, do the following:

- Choose a representative period.
- Use the default 3-second collection interval.
- When you have identified a process that consistently issues a significant number of direct I/O requests, use the DCL command SHOW PROCESS/CONTINUOUS to look for more information about the process and the image being run.
- In addition, you can use the SHOW DEVICE /FILES command to show all open files on particular disk volumes. It is important to know the file names of heavily used files to perform the offloading and load-balancing operations.

Reduction of disk I/O consumption by the system

The system uses the disk I/O subsystem for three activities: paging, swapping, and XQP operations. This kind of disk I/O is a good place to start when setting out to trim disk I/O load. All three types of system I/O can be reduced readily by offloading to memory. Swapping I/O is a particularly data-transfer-intensive operation, while the other types tend to be more seek-intensive.

Paging I/O Activity

Paging Read I/O Operations Page Read I/O Rate, also known as the **hard fault rate**, is the rate of read I/O operations necessary to satisfy page faults. Since the system attempts to cluster several pages together whenever it performs a read, the number of pages actually read will be greater than the hard fault rate. The rate of pages read is given by the Page Read Rate.

Use the following equation to compute the average transfer size (in blocks) of a page read I/O operation:

$$\text{average transfer size} = \frac{\text{page read rate}}{\text{page read IO rate}}$$

Effects on the Secondary Page Cache Most page faults are *soft* faults. Such faults require no disk I/O operation because they are satisfied by mapping to a global page or to a page in the secondary page cache (free-page list and modified-page list). For the cache to function effectively, the rate of *hard* faults—those requiring a disk I/O operation—should be less than 10% of the overall page fault rate, with the remaining 90% being soft faults. Even if the hard fault rate is less than 10%, you should try to reduce it further if it represents a significant fraction of the disk I/O load on any particular node or individual disk. (See Secondary Page Cache in Chapter 5).

Note that the number of hard faults resulting from image activation can be reduced only by curtailing the number of image activations or by exercising LINKER options such as /NOSYSSHR (to reduce image activations) and reassignment of PSECT attributes (to increase the effectiveness of page fault clustering).

Paging Write I/O Operations The Page Write I/O Rate represents the rate of disk I/O operations to write pages from the modified-page list to backing store (paging and section files). As with page read operations, page write operations are clustered. The rate of pages written is given by the Page Write Rate.

Use the following equation to compute the average transfer size (in blocks) of a page write I/O operation:

$$\text{average transfer size} = \frac{\text{page write rate}}{\text{page write IO rate}}$$

The frequency with which pages are written depends on the page modification behavior of the work load and on the size of the modified-page list. In general, a larger modified-page list must be written less often than a smaller one.

Obtaining Information About Paging Files You can obtain information on each paging file, including the disks on which they are located, with the DCL command SHOW MEMORY/FILES /FULL.

Swapping I/O Activity

Swapping I/O should be kept as low as possible. The Inswap Rate item of the I/O class lists the rate of inswap I/O operations. In typical cases, for each inswap, there can also be just as many outswap operations. Try to keep the inswap rate as low as possible—no greater than 1. This is not to say that swapping should always be eliminated. Swapping, as implemented by the proactive memory reclamation policy, is desirable to force inactive processes out of memory.

Swap I/O operations are very large data transfers; they can cause device and channel contention problems if they occur too frequently. Enter the DCL command `SHOW MEMORY/FILES /FULL` to list the swapping files in use. If you have disk I/O problems on the channels servicing the swapping files, attempt to reduce the swap rate. (Refer to Chapter 17 for information about converting to a system that rarely swaps.)

File System (XQP) I/O Activity

To determine the rate of I/O operations issued by the XQP on a nodewide basis, do the following:

- Add the Disk Read Rate and Disk Write Rate items of the FCP class for each node.
- Compare this number to the sum of the I/O Operation Rate figures for all disks on that same node.

If this number represents a significant fraction of the disk I/O on that node, attempt to make improvements by addressing one or more of the following three sources of XQP disk I/O operations: cache misses, erase operations, and fragmentation.

Examining Cache Hit and Miss Rates Check the `FILE_SYSTEM_CACHE` class for the level of activity (Attempt Rate) and Hit Percentage for each of the seven caches maintained by the XQP. The categories represent types of data maintained by the XQP on all mounted disk volumes. When an attempt to retrieve an item from a cache misses, the item must be retrieved by issuing one or more disk I/O requests. It is therefore important to supply memory caches large enough to keep the hit percentages high and disk I/O operations low.

XQP Cache Sizes Cache sizes are controlled by the ACP/XQP system parameters. Data items in the `FILE_SYSTEM_CACHE` display correspond to ACP/XQP parameters as follows:

<code>FILE_SYSTEM_CACHE</code> Item	ACP/XQP Parameters
Dir FCB	ACP_SYSACC ACP_DINDXCACHE
Dir Data	ACP_DIRCACHE
File Hdr	ACP_HDRCACHE
File ID	ACP_FIDCACHE
Extent	ACP_EXTCACHE ACP_EXTLIMIT
Quota	ACP_QUOCACHE
Bitmap	ACP_MAPCACHE

The values determined by AUTOGEN should be adequate. However, if hit percentages are low (less than 75%), you

should increase the appropriate cache sizes (using AUTOGEN), particularly when the attempt rates are high.

If you decide to change the ACP/XQP cache parameters, remember to reboot the system to make the changes effective. For more information on these parameters, refer to the *OpenVMS System Management Utilities Reference Manual*.

High-Water Marking If your system is running with the default HIGHWATER_MARKING attribute enabled on one or more disk volumes, check the Erase Rate item of the FCP class. This item represents the rate of erase I/O requests issued by the XQP to support the high-water marking feature. If you did not intend to enable this security feature, see Chapter 4, Postinstallation System Management Options, for instructions on how to disable it on a per-volume basis.

Disk Fragmentation When a disk becomes seriously fragmented, it can cause additional XQP disk I/O operations and consequent elevation of the disk read and disk write rates. You can restore contiguity for badly fragmented files by using the Backup (BACKUP) and Convert (CONVERT) utilities, the DCL command COPY /CONTIGUOUS, or the DEC File Optimizer for OpenVMS, an optional software product. It is a good performance management practice to do the following:

- Perform image backups of all disks periodically, using the output disk as the new copy. BACKUP consolidates allocated space on the new copy, eliminating fragmentation.
- Test individual files for fragmentation by entering the DCL command DUMP /HEADER to obtain the number of file extents. The fewer the extents, the lower the level of fragmentation.
- Pay particular attention to heavily used indexed files, especially those from which records are frequently deleted.
- Use the Convert utility (CONVERT) to reorganize the index file structure.

Disk I/O offloading

This section describes techniques for offloading disk I/O onto other resources, most notably memory.

- Increase the size of the secondary page cache and XQP caches.
- Install frequently used images to save memory and decrease the number of I/O operations required during image activation. (See Chapter 4, Postinstallation System Management Options.)
- Decompress library files (especially HELP files) to decrease the number of I/O operations and reduce the CPU time required for library operations. Users will experience faster response to DCL HELP commands. (See Chapter 4, Postinstallation System Management Options.)

- Use global data buffers (if your system has sufficient memory) for the following system files: VMSMAIL_PROFILE.DATA, SYSUAF.DAT, and RIGHTSLIST.DAT.
- Tune applications to reduce the number of I/O requests by improving their buffering strategies. However, you should make sure that you have adequate working sets and memory to support the increased buffering. This approach will decrease the number of accesses to the volume at the expense of additional memory requirements to run the application. The following are suggestions of particular interest to application programmers:
 - Read or write more data per I/O operation.
 - * For sequential files, increase the multiblock count to move more data per I/O operation while maintaining proper process working set sizes.
 - * Turn on deferred write for sequential access to indexed and relative files; an I/O operation then occurs only when a bucket is full, not on each \$PUT. For example, without deferred write enabled, 10 \$PUTs to a bucket that holds 10 records require 10 I/O operations. With deferred write enabled, the 10 \$PUTs require only a single I/O operation.
 - Enable read ahead/write behind for sequential files. This provides for the effective use of the buffers by allowing overlap of I/O and buffer processing.
 - Given ample memory on your system, you might consider having a deeper index tree structure with smaller buckets, particularly with shared files. This approach sometimes reduces the amount of search time required for buckets and can also reduce contention for buckets in high-contention index file applications.
 - For indexed files, you might try to cache the entire index structure in memory by manipulating the number and size of buckets.
 - If it is not possible to cache the entire index structure, you might be able to reduce the index depth by increasing the bucket size. This will reduce the number of I/O operations required for index information at the expense of increased CPU time required to scan the larger buckets.

Disk I/O load balancing

The objective of disk I/O load balancing is to minimize the amount of contention for use by the following:

- Disk heads available to perform seek operations
- Channels available to perform data transfer operations

You can accomplish that objective by moving files from one disk to another or by reconfiguring the assignment of disks to specific channels.

Contention causes increased response time and, ultimately, increased blocking of the CPU. In many systems, contention (and therefore response time) for some disks is relatively high, while for others, response time is near the achievable values for disks with no contention. By moving some of the activity on disks with high response times to those with low response times, you will probably achieve better overall response.

Moving Disks to Different Channels

Use the guidelines in Chapter 11, *Evaluating Disk I/O Responsiveness*, to identify disks with excessively high response times that are at least moderately busy and attempt to characterize them as mainly seek intensive or data transfer intensive. Then use the following techniques to attempt to balance the load—moving files from one disk to another or by moving an entire disk to a different channel:

- Separate data-transfer-intensive activity and seek-intensive activity onto separate disks.
- Reconfigure the assignment of disks to separate channels.
- Distribute seek-intensive activity evenly among the disks available for that purpose.
- Distribute data-transfer-intensive activity evenly among the disks available for that purpose (on separate channels where possible).

Note

On an HSC controller, it is important to know which disks are attached to which K.sdi data channels. You can obtain this information from the HSC console.

Moving Files to Other Disks

To move files from one disk to another, you must know, in general, what each disk is used for and, in particular, which files are ones for which large transfers are issued. You can obtain a list of open files on a disk volume by entering the DCL command `SHOW DEVICE/FILES`. However, since the system does not maintain transfer-size information, your knowledge of the applications running on your system must be your guide.

Load Balancing System Files

The following are suggestions for load balancing system files:

- Use search lists to move read-only files, such as images, to different disks. This technique is not well suited for write operations to the target device because the write will take place to the first volume/directory for which you have write access.
- Define volume sets to distribute access to files requiring read and write access. This technique is particularly helpful for applications that perform many file create and delete operations, because the file system will allocate a new file on the volume with the greatest amount of free space.
- Move paging and swapping activity off the system disk by creating, on a less heavily utilized disk, secondary page and swapping files that are significantly larger than the primary ones on the system disk. This technique is particularly important for a shared system disk in a VMScluster, which tends to be very busy.
- Move frequently accessed files off the system disk. Use logical names or, where necessary, other pointers to access them. (See Chapter 4, Postinstallation System Management Options, for a list of frequently accessed system files.) This technique is particularly effective for a shared system disk in a VMScluster.

**Obtaining
MONITOR
statistics**

Use the following MONITOR statistics to obtain the appropriate information:

Command	Statistic
Average Disk Response Time	
DISK	I/O Operation Rate I/O Request Queue Length
Paging I/O Activity	
PAGE	Page Fault Rate, Page Read Rate, Page Read I/O Rate, Page Write Rate, Page Write I/O Rate
Swapping I/O Activity	
I/O	Inswap Rate

Command	Statistic
File System (XQP) I/O Activity	
FCP	Disk Read Rate, Disk Write Rate, Erase Rate
FILE_SYSTEM_CACHE	All items

See Table B-1 for a summary of MONITOR data items.

Diagnosing Resource Limitations

Overview

This chapter describes how to track down system resources that can limit performance. When you suspect that your system performance is suffering from a limited resource, you can begin to investigate which resource is most likely responsible. In a correctly behaving system that becomes fully loaded, one of the three resources—memory, I/O, or CPU—becomes the limiting resource. Which resource assumes that role depends on the kind of load your system is supporting.

Purpose

To describe how to conduct a preliminary investigation to detect resource limitations.

Definitions

CPU limitations can become the binding resource when the work load places extensive demand on it. Perhaps all the work becomes heavily computational, or there is some condition that gives unfair advantages to certain users.

I/O limitations occur when the number or speed of devices is insufficient. You will also find an I/O limitation when application design errors either place inappropriate demand on particular devices or do not employ sufficiently large blocking factors or numbers of buffers.

Memory limitations are manifestations of such diverse problems as too little physical memory for the work attempted, inappropriate use of the memory management features, improper assignments of memory resources to users, and so forth.

Diagnostic Strategy

Getting started

If you are uncertain where to begin, your preliminary investigation can proceed by checking for the possibility of memory limitations, then I/O limitations, and finally a CPU limitation.

Investigative procedures	The investigative procedures are summarized in Figure A-2. Note that the diagram includes command recommendations to help you obtain required information. The recommended commands appear in parentheses below the description of the information required.
Methods	The procedures use the process of elimination to determine the source of performance problems. There are fairly simple tests you can use to rule out certain classes of problems.
Guidelines	Use the following guidelines when conducting your preliminary investigation: <ul style="list-style-type: none">• You must be able to observe the undesirable behavior while you are running these tests. You can determine nothing with these methods unless your system is exhibiting the problem.• Be aware that it is possible to have overlapping limitations; that is, you could find a memory limitation and an I/O limitation occurring simultaneously.• You should be able to detect all major limitations for further resolution using the methods outlined in this section, repeating them as necessary.• Your final investigations might lead you to conclude that the real source of the problem is human error, possibly misuse of the resources by one or more users.
Road map	Use Figure A-2 to conduct your preliminary investigation.

Investigating Resource Limitations

Memory limitations	You can rule out memory limitations if you use the DCL commands MONITOR IO or MONITOR SYSTEM as shown in the following table:
---------------------------	---

IF you...	THEN you...
observe <ul style="list-style-type: none"> • a substantial amount of free memory¹ • little or no paging² • little or no swapping³ 	can rule out memory limitations.
observe significant inswapping, little free memory, or significant paging	should investigate memory limitations further.

¹See the entries for Free List Size and Modified List Size.
²See the Page Fault Rate.
³See the Inswap Rate.

I/O limitations

To determine if you can rule out an I/O limitation, enter the DCL command `MONITOR IO` or `MONITOR SYSTEM` and observe the rates for direct I/O and buffered I/O.

IF...	THEN you...
your system is not performing any direct I/O	do not have a disk I/O limitation.
you observe that there is no buffered I/O	do not have a terminal I/O limitation.
either or both operations are occurring	cannot rule out the possibility of an I/O limitation.

CPU limitations

To determine if there is a CPU limitation, use the DCL command `MONITOR STATES`.

You might also use the DCL command `MONITOR MODES` to observe the amount of user mode time. The `MONITOR MODES` display also reveals the amount of idle time, which is sometimes called the null time.

IF...	THEN...
many of your processes are in the computable state	you can conclude there is a CPU limitation.
many of your processes are in the computable outswapped state	be sure to address the issue of a memory limitation first.

IF...	THEN...
the user mode time is high	it is likely there is a limitation occurring around the CPU utilization.
there is almost no idle time	it is fair to conclude that the CPU is being heavily used.

A final indicator of a CPU limitation that the MONITOR MODES display provides is the amount of kernel mode time. A high percentage of time in kernel mode can indicate excessive consumption of the CPU resource by the operating system. This problem is more likely the result of a memory limitation but could indicate a CPU limitation as well. If you decide to investigate the CPU limitation further, proceed through the steps in Chapter 15.

After the Preliminary Investigation

Isolating the problem

When you have completed your preliminary investigation, you are ready to do the following:

- Isolate the cause of the observed behavior.
- Conclude, in general terms, what remedies are available to you.
- Apply one or more of the specific corrective procedures outlined in this chapter or in Chapter 16.

Observing the tuned system

Once you take the appropriate remedial action, you must monitor the effectiveness of the changes and, if you do not obtain sufficient improvement, try again. In some cases, you will need to repeat the same steps, but either increase or decrease the magnitude of the changes you made. In other cases, you will proceed further in the investigation and uncover some other underlying cause of the problem and take corrective steps.

The diagrams and text do not attempt to depict this looping. Rather, repetition is always implied, pending the outcome of the changes. Therefore, tuning is frequently an iterative process. The approach to tuning presented by this chapter and Chapter 16 assumes that multiple causes of performance problems are uncovered by repeating the steps shown until you achieve satisfactory performance.

Note

Effective tuning requires that you can observe the undesirable performance behavior while you test.

**Obtaining
a listing of
system current
values**

You will find it especially helpful to keep a listing of the current values of all your system parameters nearby as you conduct the following investigations. Running SYSGEN and specifying a file name is one method for obtaining this listing. (See the *OpenVMS System Manager's Manual*.)

```
$ RUN SYS$SYSTEM:SYSGEN
SYSGEN> SET/OUTPUT=filename
SYSGEN> SHOW/ALL
SYSGEN> SHOW/SPECIAL
SYSGEN> EXIT
$ PRINT/DELETE filename
```

Isolating Memory Limitations

Overview

The key to successful performance management of an AXP system is to keep the memory management activity to a minimum. You will find that memory limitations cause paging, swapping, or both, precisely the activities you want to minimize. It requires skillful balancing of the memory management mechanism to reduce one without incurring too much of the other.

Purpose To determine if the memory resource is limiting performance.

Definition **Disk thrashing** means excessive reading and writing to disk that accomplishes little.

Analyzing the Excessive Paging Symptom

When to investigate Whenever you detect paging or swapping on a system with degraded performance, you should investigate a memory limitation. If you observe a lack of free memory but no serious paging or swapping, the system might be just at the point where it will begin to experience excessive paging or swapping if demand grows any more.

In this case, you have a bit of advance warning, and you might want to examine some preventive measures.

What is excessive paging? There are no universally applicable scales that rank page faulting rates from moderate to excessive.

Although the only good page faulting rate is zero page faults per second, you need to think in terms of the maximum tolerable rate of page faulting for your system.

Guidelines

- Define the maximum tolerable page fault rate. You should view any higher page fault rate as excessive.
- Paging always consumes system resources (CPU and I/O), therefore, its harmfulness depends entirely on the availability of the resources consumed.
- In judging what page faulting rate is the maximum tolerable rate for your system, you must consider your configuration and the type of paging that is occurring.
 For example, on a system with slow disks, what might otherwise seem to be a low rate of paging to the disk could actually represent intolerable paging because of the response time through the slow disk. This is especially true if the percentage of page faults from the disk is high relative to the total number of faults.
- You can judge page fault rates only in the context of your own configuration.
- The statistics must be examined in the context of both the overall faulting and the apparent system performance. The system manager who knows the configuration can best evaluate the impact of page faulting.

Once you have determined that the rate of paging is excessive, you need to determine the cause. As Figure A-3 shows, you can begin by looking at the number of image activations that have been occurring.

Excessive image activations

Use ACCOUNTING to examine the total number of images started.

IF...	THEN...
image-level accounting is enabled and the value is in the low-to-normal range for typical operations at your site	the problem lies elsewhere.
image-level accounting is NOT enabled	check the display produced by the MONITOR PAGE command for demand zero faults.
50% of all page faults are demand zero faults	you have evidence that confirms the possibility that image activations are too frequent.

Additional Considerations

If image activations are excessive, do the following:

- Enable image-level accounting (if it is not enabled) at this time and collect enough data to confirm the conclusion about the high percentage of demand zero faults.

- Determine how to reduce the number of image activations by reviewing the guidelines for application design in Chapter 16. The problem of paging induced by image activations is unlikely to respond to any attempt at system tuning. The appropriate action involves application design changes.

Characterizing hard versus soft faults

You should characterize your page faulting. *Soft paging* refers to paging from the page cache in main memory. Paging from disk is *hard paging*, and it is the less desirable of the two.

Although soft paging is undesirable when it is excessive, it is normally much less costly to overall system performance than disk paging, simply because it is faster.

System Page Faulting

All the system tuning solutions for excessive paging involve a reallocation of the memory resource, and nothing more. Consider the following suggestions:

- You should not reduce the size of the operating system's working set and offer that memory to the process working sets or the page cache because it is much more costly to performance when the system incurs page faults than when other processes experience either hard or soft page faults.
- You should always strive to keep the system page fault rate below 2 faults per second. (You can observe the system fault rate with the MONITOR PAGE command.)
- Rather than reducing the system's working set and risking the possibility of introducing system page faulting, you should consider purchasing more memory first.

Page Cache Is Too Small

In situations of excessive paging not due to image activations, you should determine what kinds of faults and faulting rates exist. Use the MONITOR PAGE command and your knowledge of your work load. If you are experiencing a high hard fault rate (represented by Page Read I/O Rate), evaluate the overall faulting rate (represented by Page Fault Rate). If the overall faulting rate is low while the hard fault rate is high, the page cache is ineffective; that is, the size of the free-page list, the modified-page list, or both, is too small. You need to increase the size of the cache. This relatively rare problem occurs when a system has been mistuned; for example, perhaps AUTOGEN was bypassed.

Before deciding to acquire more memory, try increasing the values of MPW_LOLIMIT, MPW_THRESH, FREEGOAL, and FREELIM. (See Chapter 17, Solutions for Memory-Limited Behavior.) You might also try reducing the system parameter BALSETCNT or reducing the working set characteristics. However, if these changes result immediately in the following problems when the cache is too large and the working sets are too small (and lowering the cache parameter values a bit does not bring them

into balance), you have no other tuning options. You must reduce demand or acquire more memory. (See Solutions for Memory-Limited Behavior, Reduce demand or add memory.)

Saturated System Disk

If you have the combination of a high hard fault rate with high faulting overall, it is quite possible the load is too high on your system, which means that the system disk activity is saturated and you must reduce the page faulting to disk.

However, first perform the checks described in Chapter 17, Solutions for Memory-Limited Behavior, for small working set sizes. This action will rule out or correct the possibility that the combination of heavy overall faulting with heavy hard faulting is due to too large a page cache while too many processes attempt to work with small working sets. The solution will require you to reduce the cache size and increase the WSQUOTA values.

If this investigation fails to produce results, you can conclude that the system disk is saturated. Therefore, you should consider the following:

- Adding another paging file on another disk
- Reducing demand
- Adding more memory

Since adding more memory is less costly than acquiring a disk, it is usually preferable, unless you have an underutilized disk drive available.

Page Cache Is Too Large

If you find that your faults are mostly of the soft variety, check to see if the overall faulting rate is high. If so, you might have the relatively rare problem of an unnecessarily large page cache. As a guideline, you should expect the size of your page cache to be one order of magnitude less than the total memory consumed by the balance set under load conditions.

The only way to create a page cache that is too large is by seriously mistuning a system. (Perhaps AUTOGEN was bypassed.) Chapter 17, Decrease page cache size, describes how to reduce the size of the page cache through the MPW_LOLIMIT, MPW_THRESH, FREEGOAL, and FREELIM system parameters.

Small total working set size

If your page cache size is appropriate, you need to investigate the likelihood that excessive paging is induced when a number of processes attempt to run with working set sizes that are too small for them. If the total memory for the balance set is too small, one of the following three possibilities (or a combination thereof) is at work:

- The working set size might be inappropriate because:
 - The working sets have been set too small with the WSDEFAULT and WSQUOTA characteristics in the UAF.
 - The effective working set quota has been lowered by DCL commands or system services that were invoked as the process ran.
 - The processes are not succeeding in borrowing working set space (in the loan region).
- Perhaps the automatic working set adjustment feature (AWSA) has been turned off or is for some reason not as effective as it could be.
- Swapper trimming might be reducing the working set sizes too vigorously.

Figures A-4, A-5, and A-6 summarize the procedures for isolating the cause of working set sizes that are too small.

**Inappropriate
WSDEFAULT,
WSQUOTA, and
WSEXTENT
values**

Begin to narrow down the possible causes of unusually small total working set sizes by looking first at your system's allocation of working set sizes. To gain some insight into the work load and which processes have too little memory, do the following:

- Enter the MONITOR PROCESSES/TOPFAULT command to learn which processes are faulting because their working set sizes are too small.
- Use the SHOW PROCESS/CONTINUOUS command to learn what the top faulting processes are doing and how much memory they are using.
- Look at the memory consumed by the other larger processes by entering the SHOW SYSTEM and MONITOR PROCESSES commands.

Perhaps you can conclude that one large process (or several) does not need as much memory as it is using. If you reduced its WSQUOTA or WSEXTENT values, or both, the other processes could use the memory the large process currently takes.

Learning About the Process

To form any firm conclusions at this point, you need to learn more about the process's behavior as its working set size grows and shrinks. Use the MONITOR PROCESSES command and the lexical function F\$GETJPI for this purpose.

To look at the current values as the process executes, follow these steps:

1. Note the process identification number (PID) on the MONITOR PROCESSES display.
2. Ensure that you have the WORLD privilege.

3. For each heavily faulting process you want to investigate, request these items:

- Working set quota size
- Process page count
- Global page count
- Working set extent

Obtaining Process Information

To request the items, use the system service SYS\$GETJPI or the lexical function F\$GETJPI. When using F\$GETJPI, specify the process ID (PID) in quotation marks and a keyword (GPGCNT, PPGCNT, WSEXTENT, WSQUOTA, or WSSIZE) denoting the type of process information to be returned as shown in the following example:

```
$ WSQUOTA = F$GETJPI("pid","WSQUOTA")
$ SHOW SYMBOL WSQUOTA
$ WSSIZE = F$GETJPI("pid","WSSIZE")
$ SHOW SYMBOL WSSIZE
$ PPGCNT = F$GETJPI("pid","PPGCNT")
$ SHOW SYMBOL PPGCNT
$ GPGCNT = F$GETJPI("pid","GPGCNT")
$ SHOW SYMBOL GPGCNT
$ WSEXTENT = F$GETJPI("pid","WSEXTENT")
$ SHOW SYMBOL WSEXTENT
```

Suggestion: Write a program or command procedure that requests the PID and then formats and displays the resulting data.

The lexical function item PPGCNT represents the process page count, while GPGCNT represents the global page count. You need these values to determine how full the working set list is. The sum of PPGCNT plus GPGCNT is the actual amount of memory in use and should always be less than or equal to the value of WSSIZE. By sampling the actual amount of memory in use while processes execute, you can begin to evaluate just how appropriate the values of WSQUOTA and WSEXTENT are.

If the values of WSQUOTA and WSEXTENT are either unnecessarily restricted or too large in a few obvious cases, they need to be adjusted; proceed next to the discussion of adjusting working sets in Chapter 17, Solutions for Memory-Limited Behavior.

Ineffective borrowing

If you observe that few of the processes are able to take advantage of loans, then borrowing is ineffective. Chapter 17, Solutions for Memory-Limited Behavior, discusses how to make the necessary adjustments so that borrowing is more effective.

AWSA might be disabled

You need to investigate the status of automatic working set adjustment (AWSA) by checking the value of the system parameter WSINC. If you find WSINC is greater than zero, you know that automatic working set adjustment is turned on. (More precisely, the part of automatic working set adjustment that permits working set sizes to grow is turned on). However, at the same time, you should also check whether WSDEC, PFRATL, or both, are zero. While setting WSINC=0 turns the full automatic working set adjustment mechanism off, setting PFRATL=0 when WSINC is greater than zero will disable just that part of automatic working set adjustment that provides the voluntary decrements in the working set sizes. (For example, in Figure 6-5, if PFRATL and WSDEC equaled zero, the actual working set limit line would have leveled off at Q4 and would not have changed until Q18.)

If automatic working set adjustment is disabled, processes are unable to increase their working set sizes. You will observe that although processes have WSQUOTA values greater than their WSDEFAULT values, those processes that are currently active (doing some computing) do not show a working set size count above their WSDEFAULT values. At the same time, your system is experiencing heavy page faulting. You should enable automatic working set adjustment, by setting WSINC greater than zero, so that working set growth is possible. See the discussion about enabling AWSA in Chapter 17, Solutions for Memory-Limited Behavior.

AWSA is ineffective

If AWSA is turned on, there are four ways that it could be performing less than optimally, and you must evaluate them:

- AWSA might not be responding quickly enough to increased demand. That is, when page faulting increases significantly, working set sizes are not increased quickly enough to sufficiently large values.
- AWSA with voluntary decrementing enabled might be causing the working set sizes to oscillate.
- AWSA with voluntary decrementing enabled might be shrinking the working sets too quickly, thereby inducing unnecessary paging.
- AWSA might not be decrementing the working set sizes where possible because voluntary decrementing is disabled.

AWSA Is Not Responsive to Increased Demand

If you use the SHOW PROCESS/CONTINUOUS command for those processes that MONITOR PROCESSES/TOPFAULT shows are the heaviest page faulters, you might find that the automatic working set adjustment is not increasing their working set sizes quickly enough in response to their faulting. If the default values of WSINC, PFRATH, or AWSTIME have been changed, you should restore them to their original values and consider

adjusting the WSDEF and WSQUO values of the offending process.

AWSA with Voluntary Decrementing Enabled Causes Oscillations

It is possible for the voluntary decrementing feature of automatic working set adjustment to cause processes to go into a form of oscillation where the working set sizes never stabilize, but keep growing and shrinking while accompanied by page faulting. When you observe this situation, through the SHOW PROCESS/CONTINUOUS display, you should disable voluntary decrementing by setting PFRATL=0. See Chapter 17, Disable voluntary decrementing.

AWSA Shrinks Working Sets Too Quickly

From the SHOW PROCESS/CONTINUOUS display, you can also determine if the voluntary decrementing feature of automatic working set adjustment is shrinking the working sets too quickly. In that event, you should consider decreasing WSDEC and decreasing PFRATL. See Chapter 17, Tune voluntary decrementing.

AWSA Needs Voluntary Decrementing Enabled

You might observe the case of one or more processes that rapidly achieve a very large working set count and then maintain that size over some period of time. However, you know or suspect that those processes should not require that much memory continuously. Although those processes are not page faulting, other processes are. You should check whether voluntary decrementing is turned off (PFRATL=0 and optionally WSDEC=0). See Figure A-6. It might be that, for your work load, voluntary decrementing would bring about improvement since it is time based, not load based. You could enable voluntary decrementing according to the suggestions in Chapter 17, Turn on voluntary decrementing, to see if any improvement is forthcoming.

If you decide to take this step, keep in mind that it is the exception rather than the rule. You could make conditions worse rather than better. Be certain to monitor your system very carefully to ensure that you do not induce working set size oscillations in your overall work load, as described previously. If no improvement is obtained, you should turn off voluntary decrementing. Probably your premise that the working set size could be reduced was incorrect. Also, if oscillations do result that do not seem to stabilize with a little time, you should turn voluntary decrementing off again. You must explore, instead, ways to schedule those processes so that they are least disruptive to the work load.

Swapper Trimming Is Too Vigorous

Perhaps there are valid reasons why at your site WSINC has been set to zero to turn off automatic working set adjustment. For example, the applications might be well understood, and the memory requirements for each image might be so predictable that the value for WSDEFAULT can be accurately set. Furthermore, it is possible that if automatic working set adjustment is enabled at your site, you are satisfied that your system is using appropriate values for WSQUOTA, WSEXTENT, PFRATH, BORROWLIM, and GROWLIM. In these situations, perhaps swapper trimming is to blame for the excessive paging. In particular, perhaps trimming on the second level is too severe.

Figure A-7 illustrates the investigation for paging problems induced by swapper trimming. Again, you must determine the top faulting processes and evaluate what is happening and how much memory is consumed by these processes. Use the MONITOR PROCESSES/TOPFAULT and MONITOR PROCESSES commands. By selecting the top faulting processes and scrutinizing their behavior with the SHOW PROCESS /CONTINUOUS command, you can determine if there are many active processes that seem to display working set sizes with the following values:

- Their WSQUOTA values
- The systemwide value set by the system parameter SWPOUTPGCNT

Either finding indicates that swapper trimming is too severe.

If such is the case, consider increasing the system parameter SWPOUTPGCNT while evaluating the need to increase the system parameter LONGWAIT. The swapper uses LONGWAIT to detect those processes that are truly idle. If LONGWAIT specifies too brief a time, the swapper can swap temporarily idle processes that would otherwise have become computable again soon. See Chapter 17, Adjust swapper trimming. For computable processes, the same condition can occur if DORMANTWAIT is set too low.

Analyzing the Swapping Symptom

Swapping versus paging

Experience with systems has shown that swapping of active processes is less desirable than modest paging because swapping involves disk accesses (true of only hard page faults). Swapping requires each process and its context to be written out to disk, an event that is normally slower than the average paging operation since it involves more blocks. There is additional system overhead for swapping caused by stopping and starting processes. In using

the disk resource heavily, the swapper might cause additional entries in the queue on its disk, thus delaying other processes that need access to that disk.

Detecting harmful swapping

Harmful swapping manifests itself in heavy consumption of the CPU resource and the disk, to the detriment of other processes. You should use the following tests to check for any symptoms that indicate swapping is harmful:

- Enter the DCL command `MONITOR IO` and examine the `inswap` rate. If the rate is zero, you have no swapping and you need not pursue this series of tests any further.
- Enter the DCL command `MONITOR STATES`. If you observe few processes in the `COMO` state, swapping is not affecting CPU operations.

If your swapping passes these three tests, you can conclude that swapping is not so harmful on your system that you should eliminate it.

Investigating harmful swapping

Indications of harmful swapper activity, such as heavy disk or CPU consumption, warrant attention. (Figures A-8, A-9, and A-10 summarize the investigation for swapping.)

Limiting Swapping

You might consider converting your system to one that only pages and rarely if ever swaps, particularly if your system is a small configuration. You accomplish this by performing the following tasks:

- Lowering the system parameter `SWPOUTPGCNT`
- Setting the system parameter `BALSETCNT` equal to a value that is two less than the value of the system parameter `MAXPROCESSCNT`
- Adding more memory

Reducing Process Working Set

Optionally, you could decide to reduce the process working set quotas (in the UAF). See Chapter 17, *Solutions for Memory-Limited Behavior*.

Even if you tune your system so that it rarely swaps, you still need a swapping file on your system. However, the space requirement for the swapping file is reduced. If disk space is at a premium, you can adjust your swapping file space requirement to 75 percent of its previous value with the `AUTOGEN` command procedure. (See the *OpenVMS System Manager's Manual*.)

Causes of harmful swapping

If you find that your system is showing symptoms of harmful swapping and that performance has degraded, no free balance slots and insufficient free memory for all working sets are two possible causes.

No Free Balance Slots

If there are no free balance slots, use the DCL command `SHOW MEMORY` to check the number of free balance slots. If the number available is small and you know there is still adequate free memory (which you can also check with `SHOW MEMORY`), then you should be able to alleviate the swapping by increasing the system parameter `BALSETCNT`.

Insufficient Free Memory for All the Working Sets

If there are free balance slots but the total of the working set sizes exceeds available memory, you can safely conclude that there is not enough free memory to support all the working sets at once. This condition can result from one or more of the following factors:

- Improper partitioning of memory due to a page cache that is too large
- Situations where some users use unreasonably large amounts of memory
- Demand that is simply too high for capacity

Large Page Cache To determine if the page cache is too large, do the following:

1. Use the `SHOW MEMORY` display to determine the total usable memory (the total physical memory less the memory used by the operating system).
2. Add the values for the two system parameters `FREEGOAL` and `MPW_THRESH` to determine how much memory is allocated to the page cache. If the page cache size is more than 15 percent of the total usable memory, the page cache might be too large.

Only when a system has been seriously mistuned should you find that the page cache is too large. (Perhaps `AUTOGEN` was bypassed.) Chapter 17, *Solutions for Memory-Limited Behavior*, describes how to reduce the size of the page cache through the `MPW_LOLIMIT`, `MPW_THRESH`, `FREEGOAL`, and `FREELIM` system parameters.

If you determine that the page cache is not too large, or having reduced its size, you find that there is still insufficient free memory for all the working sets, you need to investigate other potential causes for the problem. These causes are described in the next sections.

Why processes consume unreasonable amounts of memory

Swapping can be induced whenever one or a small number of processes devour memory at the expense of other processes. You can find out if a few users are using large amounts of memory by examining the display produced by the `MONITOR PROCESSES` command.

**Large,
compute-bound
processes**

At this point, you should be particularly alert for the situation where one or more very large, compute-bound processes at low priority consume memory at the expense of a number of smaller processes. Typically, the smaller processes might be trying to perform some terminal I/O, such as editing. When memory becomes tight, the large process that is compute bound is less likely to be selected for outswapping than any process that is in the local event flag wait state. Consequently, in this situation, the operating system will select processes running the editor for outswapping as soon as they start to wait for I/O. As a result, the editing processes will experience symptomatically poor response times due to frequent outswapping. The SHOW SYSTEM command provides a valuable tool for checking the priority and state of the large process.

Note the process identification number from the MONITOR PROCESSES display and ensure that you have the WORLD privilege. Then, for each large process you want to investigate, use the lexical function F\$GETJPI as described in Inappropriate WSDEFAULT, WSQUOTA, and WSEXTENT values, to request the working set quota, size, process page count, global page count, and working set extent.

If you find that any of the processes are above their working set quotas, you might want to decrease DORMANTWAIT and monitor performance for a time. If decreasing DORMANTWAIT proves ineffective, you can enter the DCL command SET PROCESS /SUSPEND as described in Chapter 17, Solutions for Memory-Limited Behavior, to suspend the large, compute-bound process that is over WSQUOTA. This action offers a rapid means of restoring other process activities. (Once the process is suspended, the swapper can trim the process to its SWPOUTPGCNT value.) As soon as SHOW PROCESS/CONTINUOUS reveals that the process has been trimmed, you can safely resume it. If the AWSA is set correctly, the problem should not recur since the process will be unable to grow beyond its quota while memory is scarce.

However, you must determine the underlying cause of the problem (for example, the working set quota might be too large for the process) and take corrective action. For example, you could lower WSQUOTA and increase WSEXTENT. Borrowing will then be reclaimed by the swapper. If the large, compute-bound process is not above its working set quota, suspending the process might provide temporary relief, but as soon as you allow the process to resume, it can start to devour memory again. Thus, the most satisfactory corrective action is the permanent solution discussed in Chapter 17, Solutions for Memory-Limited Behavior.

Large waiting processes

While using the `SHOW SYSTEM` command to look for large processes that are compute bound, you might, instead, observe that one or more large processes are hibernating or are in some other wait state. Possibly, swapping has been disabled for these processes. You could use the `SHOW PROCESS/CONTINUOUS` command for each process to determine if any inactive process escapes outswapping. As a next step, you could invoke the System Dump Analyzer (SDA) with the DCL command `ANALYZE /SYSTEM` to see if the process status line produced by the SDA command `SHOW PROCESS` reveals the process status of `PSWAPM`.

If you find a process that is not allowed to swap, yet apparently consumes a large amount of memory when it is inactive, you might conclude that swapping should be enabled for it. Enabling swapping would give other processes a more equitable chance of using memory when memory is scarce and the large process is inactive. You should discuss your conclusions with the owner of the process to determine if there are valid reasons why the process must not be swapped. (For example, most real-time processes should not be swapped.) If the owner of the process agrees to enable the process for swapping, use the DCL command `SET PROCESS/SWAPPING` (which requires the `PSWAPM` privilege). See the discussion of enabling swapping for all other processes in Chapter 17, *Solutions for Memory-Limited Behavior*.

If the offending process is a disk ACP (ODS-1 only), you need to set the system parameter `ACP_SWAPFLGS` appropriately and reboot the system. See the discussion about enabling swapping for disk ACPs in Chapter 17, *Solutions for Memory-Limited Behavior*.

Too many competing processes

If the data you collected with the `F$GETJPI` lexical function reveals that the working set counts (the actual memory consumed by the processes) are not particularly large, you might simply have too many processes attempting to run concurrently for the memory available. If they are and the problem persists, you might find that performance improves if you reduce the system parameter `MAXPROCESSCNT`, which specifies the number of processes that can run concurrently. See the discussion about reducing the number of concurrent processes in Chapter 17, *Solutions for Memory-Limited Behavior*.

However, if `MAXPROCESSCNT` already represents the number of users who must be guaranteed access to your system at once, reducing `MAXPROCESSCNT` is not a viable alternative. Instead, you must explore other ways to reduce demand (redesign your application, for example) or add memory. See the discussion about reducing demand or adding memory in Chapter 17, *Solutions for Memory-Limited Behavior*.

Borrowing is too generous

For the processes that seem to use the most memory, use the `SHOW PROCESS/CONTINUOUS` command to check if the processes are operating in the `WSEXTENT` region; that is, their working set sizes range between the values of `WSQUOTA` and `WSEXTENT`. If not, it might be beneficial to increase the values of `BORROWLIM`, `GROWLIM`, or both. Increasing both `BORROWLIM` and `GROWLIM` discourages loans when memory is scarce. By judiciously increasing these values, you will curtail the rate of loans to processes with the largest working sets, particularly during the times when the work load peaks. See the discussion about discouraging working set loans in Chapter 17, *Solutions for Memory-Limited Behavior*.

Swapper trimming is ineffective

If memory is insufficient to support all the working set sizes of active processes, ineffective swapper trimming might be the cause.

In this case, the value of `SWPOUTPGCNT` might be too large. You should compare the value of `SWPOUTPGCNT` to the actual working set counts you observe. If you decide to reduce `SWPOUTPGCNT`, be aware that you will increase the amount of memory reclaimed every time second-level trimming is initiated. Still, this is the parameter that most effectively converts a system from a swapping system to a paging one and vice versa. As you lower the value of `SWPOUTPGCNT`, you run the risk of introducing excessive paging. If this situation occurs and you cannot achieve a satisfactory balance between swapping and paging, you must reduce demand or add memory. See the discussion about reducing demand or adding memory in Chapter 17, *Solutions for Memory-Limited Behavior*.

Excessively large working sets

If you conclude that `SWPOUTPGCNT` is not too large, then you have already determined that the working sets are fairly large but not above quota and that few processes are computable. You will probably discover that one or more of the following conditions exist:

- The working set quotas are too large in some cases
- The parameter `WSINC` is too large or `PFRATH` is too low
- Too many working sets are locked in memory and cannot be outswapped

The first two conditions can be determined from information you have collected. However, if you suspect that too many users have used the `DCL` command `SET PROCESS/NOSWAPPING` to prevent their processes from being outswapped (even when not computable), you need to invoke the `F$GETJPI` lexical function for suspicious processes. (Suspicious processes are those that remain in the local event flag wait state for some time while the system is swapping heavily. You can observe that condition with the `SHOW SYSTEM` command.) If the flag `PSWAPM` in

the status field (STS) is on, the process cannot be swapped. (The documentation for the system service \$GETJPI specifies the status flags. See the *OpenVMS System Services Reference Manual*).

As an alternative, you can use the ANALYZE/SYSTEM command to invoke SDA to enter the SHOW PROCESS command for the suspicious processes. Those that cannot be swapped will include the designation PSWAPM in the status line at the top of the display.

If you determine that one or more processes should be allowed to swap, you should seek agreement and cooperation from the users. (If agreement is reached but users do not follow through, you could remove the users' PSWAPM or SETPRV privileges with the /PRIVILEGES qualifier of AUTHORIZE.) See the discussion about enabling swapping for all other processes in Chapter 17, Solutions for Memory-Limited Behavior.

Disk thrashing occurs

If you find that a large number of processes are computable at this point in your investigation, you should ensure that disk thrashing is not initiated by the outswapping of processes while they are computing. Disk thrashing, in this case, is the outswapping of processes rapidly followed by the inswapping of the same processes.

Processes in the COMO state on the MONITOR STATES display are normally those that have finished waiting for a local event flag and are ready to be inswapped. On a system without swapping, they are new processes. However, you might find computable outswapped processes that were swapped out while they were computable. Such undesirable swapping is harmful if it occurs too frequently.

A particular work load problem must exist to provoke this situation. Suppose a number of compute-bound processes attempt to run concurrently. The processes will not be changing states while they compute. Moreover, since they are computing, they escape second-level swapper trimming to the SWPOUTPGCNT value. This condition can result in memory becoming scarce, which then could force the processes to begin swapping in and out among themselves. Whenever an outswapped process becomes computable, the scheduler is awakened to begin rescheduling. A process that is outswapped while it is computable also prompts immediate rescheduling. Thus, if the processes cannot gain enough processing time from the CPU before being outswapped and, if they are outswapped while they are computable, thrashing occurs.

Outswapped Processes at Base Priority

If you enter the `SHOW SYSTEM` command and note that many of the computable outswapped processes are at their base priority, you should check to be sure that the processes are not being swapped out while they are computable. (The fact that the processes are at their base priority implies they have been attempting to run for some time. Moreover, a number of `COMO` processes at base priority strongly suggests that there is contention for memory among computable processes.)

Low I/O Rates

You can enter the `SHOW PROCESS/CONTINUOUS` command for the `COM` processes and observe whether they fail to enter the `LEF` state before they enter the `COMO` state. Alternatively, you might observe whether their direct and buffered I/O rates remain low. Low I/O rates also imply that the processes have seldom gone into a local event flag wait state.

If you observe either indication that processes are being outswapped while computable, it is probable that too many highly computational processes are attempting to run concurrently or that `DORMANTWAIT` is set too low. However, you should rule out the possible effects of too many batch jobs running at the same time, before you attempt to adjust the rate at which processes are inswapped.

Concurrent Batch Jobs

Enter the `DCL` command `SHOW SYSTEM/BATCH` to determine the number of batch jobs running concurrently and the amount of memory they consume. If you conclude that the number of concurrent batch jobs could be affecting performance, you can reduce the demand they create by modifying the batch queues with the `/JOB_LIMIT` qualifier. Include this qualifier on the `DCL` command you use to establish the batch queue (`INITIALIZE/QUEUE` or `START/QUEUE`).

If you have ruled out any possible memory contention from large concurrent batch jobs, you can conclude that the solution involves correcting the frequency at which the system outswaps then inswaps the computable processes. Assuming the system parameter `QUANTUM` represents a suitable value for all other work loads on the system, you can draw the second conclusion. If you find the current priorities of the compute-bound processes are less than or equal to `DEFPRI`, you should consider increasing the special parameter `SWPRATE` so that inswapping of compute-bound processes occurs less frequently. In that way, the computing processes will have a greater amount of time to run before they are outswapped to bring in the `COMO` processes. See the discussion about reducing the rate of inswapping in Chapter 17, *Solutions for Memory-Limited Behavior*.

System swaps rather than pages

If you have found a large number of computable processes that are not at their base priority and if their working sets are fairly large yet not above their working set quotas, you should investigate whether any real paging is occurring. Even when there is no real paging, there can be paging induced by swapping activity. You can identify paging due to swapping whenever a high percentage of all the paging is due to global valid page faults. Use the display produced by the MONITOR PAGE command to evaluate the page faulting.

If you conclude that most of the paging is due to swapper activity, your system performance can improve if you induce some real paging by decreasing the working set sizes, an action that can reduce swapping. To induce paging, you might also reduce the automatic working set adjustment growth by lowering WSINC or increasing PFRATH. See the discussion about reducing paging to induce swapping in Chapter 17, Solutions for Memory-Limited Behavior.

Demand exceeds available memory

If you reach this point in the investigation and still experience swapping in combination with degraded performance, you have ruled out all the appropriate ways for tuning the system to reduce swapping. The problem is that the available memory cannot meet demand.

Analyzing the Limited Free Memory Symptom

Present capacity versus anticipated demand

If your system seems to run low on free memory at times, it is a warning that you are likely to encounter paging or swapping problems. You should carefully investigate your capacity and anticipated demand.

IF you...	THEN...
see little future growth demand	you are unlikely to experience a problem in the near future.
see that your future growth demand will soon exceed your capacity	it is time to review all possible options.
conclude that the only suitable option is to order memory	the only suitable option is to order memory.

Reallocating memory

Before you decide to order more memory, you might want to look at how you have allocated memory. See Figure A-11. Perhaps you could benefit by adjusting physical memory utilization so that

the page cache is larger and there is less disk paging. To make this adjustment, you might have to relinquish some of the total working set space.

If working set space has been too generously configured in your system, you have found an important adjustment you can make before problems arise. Chapter 17, *Solutions for Memory-Limited Behavior*, describes how to decrease working set quotas and working set extents.

Isolating I/O Limitations

Overview

At this point, you have observed either a direct I/O rate or a buffered I/O rate and need to determine if there could be an I/O limitation causing degraded system performance. Direct I/O is generated by disks and tapes. Buffered I/O can be produced by a number of devices, including terminals, line printers, the console disk drive, and communications devices. This chapter discusses the following topics:

- Performance problems relating to buffered I/O
- Performance problems relating to direct I/O

Purpose

To determine if the disk I/O resource is limiting performance.

Definitions

Buffered I/O is an input/output operation, such as terminal or mailbox I/O, in which an intermediate buffer from the system buffer pool is used instead of a process-specified buffer.

Direct I/O is an input/output operation in which the system locks the pages containing the associated buffer in physical memory for the duration of the I/O operation. The I/O transfer takes place directly from the process buffer.

Disk or Tape Operation Problems (Direct I/O)

Detecting direct I/O problems

Direct I/O problems for disks or tapes reveal themselves in long delay times for I/O completions. The easiest way to confirm a direct I/O problem is to detect a particular device with a queue of pending requests. A queue indicates contention for a device or controller. For disks, the MONITOR command `MONITOR DISK/ITEM=QUEUE_LENGTH` provides this information.

Since direct I/O refers to direct memory access (DMA) transfers that require relatively little CPU intervention, the performance degradation implies one or both of the following device-related conditions:

- The device is not fast enough
- The aggregate demand on the device is so high that some requests are blocked while others are being serviced

Software and hardware solutions

For a disk or tape I/O limitation that degrades performance, the only relatively low-cost solution available through tuning the software uses memory to increase the sizes of the caches and buffers used in processing the I/O operations, thereby decreasing the number of device accesses. The other possible solutions involve purchasing additional hardware, which is much more costly.

Determining I/O rates

When you enter the MONITOR IO command and observe evidence of direct I/O, you will probably be able to determine whether the rate is normal for your site. A direct I/O rate for the entire system that is either higher or lower than what you consider normal warrants investigation. See Figures A-12 and A-13.

You should proceed in this section only if you deem the operation rates of disk or tape devices to be significant among the possible sources of direct I/O on your system. If necessary, rule out any other possible devices as the primary source of the direct I/O with the lexical function F\$GETDVI.

Compare the I/O rates derived in this manner or observed on the display produced by the MONITOR DISK command with the rated capacity of the device. (If you do not know the rated capacity, you should find it in literature published for the device, such as a peripherals handbook or a marketing specifications sheet.)

Device I/O rate is below capacity

Sometimes you might detect a lower direct I/O rate for a device than you would expect. This condition implies that either very large data transfers are not completing rapidly (probably in conjunction with a memory limitation centered around paging and swapping problems) or that some other devices are blocking the disks or tapes.

If you have already investigated the memory limitation and taken all possible steps to alleviate it (which is the recommended step before investigating an I/O problem), then you should try to determine the source of the blockage.

A blockage in the I/O subsystem suggests that I/O requests are queueing up because of a bottleneck. For disks, you can determine that this condition is present with the MONITOR DISK/ITEM=QUEUE_LENGTH command.

When you find a queue on a particular device, you cannot necessarily conclude that the device is the bottleneck. At this point, simply note all devices with queues for later reference. (You will need to determine which processes are issuing the I/O operations for the devices with queues.)

As the next step, you should rule out the possibility of a lockout situation induced by an ancillary control process (ACP). (Note that this condition arises only if you have ODS-1 disks.) If the system attempts to use a single ACP for both slow and fast devices, I/O blockages can occur when the ACP attempts to service a slow device. This situation can occur only if you have mounted a device with the /PROCESSOR qualifier.

Abnormally high direct I/O rate

An abnormally high direct I/O rate for any device, in conjunction with degraded system performance, suggests that I/O demand for that device exceeds its capacity. First, you need to find out where the I/O operations are occurring. Enter the MONITOR PROCESSES/TOPDIO command. From this display, you can determine which processes are heavy users of I/O and, in particular, which processes are succeeding in completing their I/O operations—not which processes are waiting.

Next, you must determine which of the devices used by the processes that are the heaviest users of the direct I/O resource also have the highest operations counts so that you can finally identify the bottleneck area. Here, you must know your work load sufficiently well to know the devices the various processes use. If you note that these devices are among the ones you found queued up, you have now found the bottleneck points.

Once you have identified the device that is saturated, you need to determine the types of I/O activities it experiences. Perhaps some of them are being mishandled and could be corrected or adjusted. Possibilities are file system caching, RMS buffering, use of explicit QIOs in user programs, and paging or swapping. After you eliminate these possibilities, you might conclude that the device is simply unable to handle the load.

File System Caching Is Suboptimal

To evaluate the effectiveness of caching, observe the display produced by the MONITOR FILE_SYSTEM_CACHE command. If cache hits are 70 percent or greater, caching activity is normal. A lower percentage, combined with a large number of attempts, indicates that caching is less than optimally effective.

You should be certain that your applications are designed to minimize the opening and closing of files. You should also verify that the file allocation and extent sizes are appropriate. Use the DCL command DIRECTORY/SIZE=ALL to display the space used by the files and the space allocated to them. If the proportion of space used to space allocated seems close to 90 percent, no changes are necessary. However, significantly lower utilization

should prompt you to set more accurate values, either explicitly or by changing the defaults, particularly on critical files. You use the RMS_EXTEND_SIZE system parameter to define the default file extents on a systemwide basis. The DCL command SET RMS_DEFAULT/EXTEND_QUANTITY permits you to define file extents on a per-process basis (or on a systemwide basis if you also specify the /SYSTEM qualifier). For more information, see the *Guide to OpenVMS File Applications*.

If these are standard practices at your site, then you should see Chapter 18, Solutions for I/O-Limited Behavior, for a discussion of how to adjust the following ACP system parameters: ACP_HDRCACHE, ACP_MAPCACHE, and ACP_DIRCACHE.

RMS Errors Induce I/O Problem

Misuse of RMS can cause direct I/O limitations. If users are blocked on the disks because of multiblock counts that are unnecessarily large, instruct the users to reduce the size of their disk transfers by lowering the multiblock count with the DCL command SET RMS_DEFAULT/BLOCK_COUNT. See Chapter 18, Solutions for I/O-Limited Behavior, for a discussion about how to improve RMS caching.

If this course is partially effective but the problem is widespread, you could decide to take action on a systemwide basis. You can alter one or more of the system parameters in the RMS_DFMB group with AUTOGEN, or you can include the appropriate SET RMS_DEFAULT command in the systemwide login command procedure. See the *Guide to OpenVMS File Applications*.

Explicit QIO Usage Is Too High

Next, you need to determine if any process using a device is executing a program that employs explicit specification of QIOs rather than RMS. If you enter the MONITOR PROCESSES /TOPDIO command, you can identify the user processes worth investigating. It is possible that the user-written program is not designed properly. It might be necessary to enable virtual I/O caching. I/O requests using the function modifier IO\$_READVBLK can read from the virtual I/O cache.

Paging or swapping disk activity

If you do not detect processes running programs with explicit user-written QIOs, you should suspect that the operating system is generating disk activity due to paging or swapping activity, or both. The paging or swapping might be quite appropriate and not introduce any memory management problem. However, some aspect of the configuration is allowing this paging or swapping activity to block other I/O activity, introducing an I/O limitation. Enter the MONITOR IO command to inspect the Page Read I/O Rate and Page Write I/O Rate (for paging activity) and the Inswap Rate (for swapping activity). Note that because system I/O activity to the disk is not reflected in the direct I/O count MONITOR provides, MONITOR IO is the correct tool to use here.

If you find indications of substantial paging or swapping (or both) at this point in the investigation, you need to consider whether the paging and swapping files are located on the best choice of device, controller, or bus in the configuration. You should also consider whether introducing secondary files and separating the files would be beneficial. A later section discusses relocating the files to bring about performance improvements.

Reduce I/O demand or add capacity

The only low-cost solutions that remain require reductions in demand. You could try to shift the work load so that less demand is placed simultaneously on the direct I/O devices. Instead, you might reconfigure the magnetic tapes and disks on separate buses to reduce demand on the bus. (If there are no other available buses configured on the system, you might want to acquire buses so that you can take this action.)

If none of the above solutions improved performance, you might need to add capacity. You probably need to acquire disks with higher transfer rates rather than simply add more disks. However, if you have been employing magnetic tapes extensively, you might want to investigate ways of shifting your applications to use disks more effectively. Chapter 18, Solutions for I/O-Limited Behavior, provides a number of suggestions for reducing demand or adding capacity.

Terminal Operation Problems (Buffered I/O)

Buffered I/O problems

Terminal operation, when improperly handled, can present a serious drain on system resources. However, the resource that is consumed is the CPU, not I/O. Terminal operation is actually a case for CPU limitation investigation but is included here because it may initially appear to be an I/O problem.

Detecting terminal I/O problems

You will first suspect a terminal I/O problem when you detect a high buffered I/O rate on the display for the MONITOR IO command. See Figure A-14. Next, you should enter the MONITOR STATES command to check if processes are in the COM state. This condition, in combination with a high buffered I/O rate, suggests that the CPU is constricted by terminal I/O demands. If you do not observe processes in the computable state, you should conclude that while there is substantial buffered I/O occurring, the system is handling it well. In that case, the problem lies elsewhere. Proceed to Chapter 15 to investigate other forms of CPU limitation.

High buffered I/O count

If you do observe processes in the COM state, you must verify that the high buffered I/O count is actually due to terminals and not to communications devices, line printers, graphics devices, devices or instrumentation not provided by Digital, or devices that emulate terminals. You must examine the operations counts for all such devices with the lexical function F\$GETDVI. See Disk or Tape Operation Problems (Direct I/O) for a discussion about determining direct I/O rates. A high operations count for any device other than a terminal device indicates that you should explore the possibility that the other device is consuming the CPU resource.

Operations count

If you find that the operations count for terminals is a high percentage of the total buffered I/O count, you can conclude that terminal I/O is degrading system performance. To further investigate this problem, enter the MONITOR MODES command. From this display, you should expect to find much time spent either in interrupt state or in kernel mode. Too much time in interrupt state suggests that too many characters are being transmitted in a few very large QIOs. Too much time in kernel mode could indicate that too many small QIOs are occurring.

Excessive kernel mode time

If the MONITOR MODES display shows much time spent in kernel mode, perhaps the sheer number of QIOs involved is burdening the CPU. See Figure A-15. You should explore whether the application can be redesigned to group the large number of QIOs into smaller numbers of QIOs that transfer more characters at a time. Such a design change could alleviate the condition, particularly if burst output devices are in use. It is also possible that some adjustment in the work load is feasible, which would balance the demand.

If neither of these approaches is possible, you need to reduce demand or increase the capacity of the CPU.

Isolating CPU Limitations

Overview

This chapter discusses the following topics:

- Blocking due to preemption
- Blocking due to process priority
- Blocking due to memory limitation
- Blocking due to system overhead

Purpose

To determine if the CPU resource is the limiting resource.

Definition

A **blocked** process is a process waiting for an event to occur (a specific semaphore signaled) before continuing execution.

Detecting CPU Limitations

Examining the compute queue

The surest way to determine whether a CPU limitation could be degrading performance is to check for a state queue with the MONITOR STATES command. See Figure A-16. If any processes appear to be in the COM or COMO state, a CPU limitation might be at work. However, if no processes are in the COM or COMO state, you need not investigate the CPU limitation any further.

If processes are in the COM or COMO state, they are being denied access to the CPU. One or more of the following conditions is occurring:

- Processes are blocked by the execution of another process at higher priority.
- Processes are time slicing with other processes at the same priority.
- Processes are blocked by excessive activity in interrupt state.
- Processes are blocked by some other resource. (Note that this last possibility means the limitation is not a CPU limitation but is instead a memory or I/O limitation.)

Higher priority blocking processes

If you suspect the system is performing suboptimally because processes are blocked by a process running at higher priority, do the following:

1. Gain access to an account that is already running.
2. Ensure you have the ALTPRI privilege.
3. Set your priority to 15 with the DCL command SET PROCESS /PRIORITY=15.
4. Enter the DCL command MONITOR PROCESSES/TOPCPU to check for a high-priority lockout.
5. Enter the DCL command SHOW PROCESS/CONTINUOUS to examine the current and base priorities of those processes that you found were top users of the CPU resource. You can now conclude whether any process is responsible for blocking lower priority processes.
6. Restore the priority of the process you used for the investigation. Otherwise, you might find that process causes its own system performance problem.

If you find that this condition exists, your option is to adjust the process priorities. See Chapter 19, Solutions for CPU-Limited Behavior, for a discussion of how to change the process priorities assigned in the UAF, define priorities in the login command procedure, or change the priorities of processes while they execute.

Time slicing between processes

Once you rule out the possibility of preemption by higher priority processes, you need to determine if there is a serious problem with time slicing between processes at the same priority. Using the list of top CPU users, compare the priorities and assess how many processes are operating at the same one. Refer to Chapter 19, Solutions for CPU-Limited Behavior, if you conclude that the priorities are inappropriate.

However, if you decide that the priorities are correct and will not benefit from such adjustments, you are confronted with a situation that will not respond to any form of system tuning. Again, the only appropriate solution here is to adjust the work load to decrease the demand or add CPU capacity. See Chapter 19, Solutions for CPU-Limited Behavior.

Excessive interrupt state activity

If you discover that blocking is not due to contention with other processes at the same or higher priorities, you need to find out if there is too much activity in interrupt state. In other words, is the rate of interrupts so excessive that it is preventing processes from using the CPU?

You can determine how much time is spent in interrupt state from the MONITOR MODES display. If the percentage of time in interrupt state is less than 10 percent, you could view this as moderate. However, if you observe percentages of 20 percent or more, you should consider this time excessive. (The higher the percentage, the more effort you should dedicate to solving this resource drain.)

If the interrupt time is excessive, you need to explore which devices cause significant numbers of interrupts on your system and how you might reduce the interrupt rate.

The decisions you make will depend on the source of heavy interrupts. Perhaps they are due to communications devices or special hardware used in real-time applications. Whatever the source, you need to find ways to reduce the number of interrupts so that the CPU can handle work from other processes. Otherwise, the solution might require you to adjust the work load or acquire CPU capacity. See Chapter 19, Solutions for CPU-Limited Behavior.

Disguised memory limitation

Once you have either ruled out or resolved the types of CPU limitation blocks, you need to determine which other resource limitation produces the block. Your next check should be for the amount of idle time. See Figure A-17. Use the MONITOR MODES command. If there is any idle time, another resource is the problem and you might be able to tune for a solution. If you reexamine the MONITOR STATES display, you will likely observe a number of processes in the COMO state. You can conclude that this condition reflects a memory limitation, not a CPU limitation. Follow the procedures described in Chapter 13 to find the cause of the blockage, and then take the corrective action recommended in Chapter 16.

Operating system overhead

If the MONITOR MODES display indicates that there is no idle time, your CPU is 100 percent busy. You will find that processes are in the COM state on the MONITOR STATES display. You must answer one more question. Is the CPU being used for real work or for nonessential operating system functions? If you detect there is operating system overhead, you might be able to reduce it.

You must analyze the MONITOR MODES display carefully. If your system exhibits excessive kernel mode activity, it is possible that the operating system is incurring overhead in the areas of memory management, I/O handling, or scheduling. You should investigate the memory limitation and I/O limitation (Chapters 13 and 14), if you have not already done so.

Once you rule out the possibility of improving memory management or I/O handling, the problem of excessive kernel mode activity might be due to scheduling overhead. However, you can do practically nothing to tune the scheduling function. There is only one case that might respond to tuning. The clock-based rescheduling that can occur at quantum end is costlier than the typical rescheduling that is event driven by process state. Explore whether the value of the system parameter QUANTUM is too low and can be increased to bring about a performance improvement by reducing the frequency of this clock-based rescheduling. (See Chapter 19, Solutions for CPU-Limited Behavior.) If not, your only other recourse is to adjust the work load or acquire CPU capacity.

RMS misused

If the MONITOR MODES display indicates that a great deal of time is spent in executive mode, it is possible that RMS is being misused. If you suspect this problem, proceed to the steps described in Chapter 14, Disk or Tape Operation Problems (Direct I/O), for RMS induced I/O limitations, making any changes that seem indicated. You should also consult the *Guide to OpenVMS File Applications*.

CPU at full capacity

If at this point in your investigation the MONITOR MODES display indicates that most of the time is spent in supervisor mode or user mode, you are confronted with a situation where the CPU is performing real work and the demand exceeds the capacity. You must either make adjustments in the work load to reduce demand (by more efficient coding of applications, for example) or you must add CPU capacity. See the appropriate section in Chapter 19, Solutions for CPU-Limited Behavior.

Correcting the problem

At this point, you should:

- Know what particular resource is limited.
- Know which section of Chapters 16, 17, Chapter 18, or 19 suggests one or more possible remedies.

If not, you could be making an error interpreting the output of one or more of the suggested tools.

- Repeat the work you did for this chapter and then, if necessary, consult your Digital software specialist.

After you perform the recommended corrective actions in Chapter 16, you should repeat the steps in this chapter to observe the effects of the changes. As you repeat the steps, watch for new problems introduced by the corrective actions or previously undetected problems. Your goal should be to complete the steps in this chapter without uncovering a serious symptom or problem.

Compensating for Resource Limitations

Overview

This chapter describes corrective procedures for each of the various categories of resource limitations described in Chapter 12. Wherever the corrective procedure suggests changing the value of one or more system parameters, the description explains briefly whether the parameter should be increased, decreased, or given a specific value. Relationships between parameters are identified and explained, if necessary. However, to avoid duplicating information available in the *OpenVMS System Management Utilities Reference Manual*, complete explanations of parameters are not included.

Purpose

To describe how to adjust system parameter values.

Definition

A **dynamic** parameter can be changed while the system is running by changing the active value in memory.

Changing System Parameters

Prerequisites

You should review descriptions of system parameters, as necessary, before changing the parameters.

Before you make any changes to your system parameters, however, make a copy of the existing version of the file that is in the SYSGEN work area, using a technique such as the following:

```
$ RUN SYS$SYSTEM:SYSGEN
SYSGEN> WRITE SYS$SYSTEM:file-spec
SYSGEN> EXIT
```

You might want to use a date as part of the file name you specify for **file-spec** to readily identify the file later.

By creating a copy of the present values, you can always return to those values at some later time. Generally you use the following technique, specifying your parameter file for **file-spec**:

```
$ RUN SYS$SYSTEM:SYSGEN
SYSGEN> USE SYS$SYSTEM:file-spec
SYSGEN> WRITE ACTIVE
SYSGEN> EXIT
```

However, if some of the parameters you changed were not dynamic, to restore them from the copied file, you must instead use the SYSGEN command WRITE CURRENT, and then reboot the system.

Guidelines

If you are planning to change a system parameter and you are uncertain of an ultimate target value and also of the sensitivity of the specific parameter to changes, you should err on the conservative side in making initial changes. As a guideline, you might make a 10 percent change in the value first so that you can observe its effects on the system.

You should change only a few parameters at a time.

IF...	THEN you ...
you see little or no effect	should try doubling or halving the original value of the parameter depending on whether you are increasing or decreasing it.
this magnitude of change had no effect	should restore the parameter to its original value with the parameter file you saved before starting.
you cannot affect your system performance with changes of this magnitude	probably have not selected the right parameter for change.

Whenever your changes are unsuccessful, make it a practice to restore the parameters to their previous values before you continue tuning. Otherwise, it can be difficult to determine which changes produce currently observed effects.

Using AUTOGEN

In most cases, you will want to use AUTOGEN to change system parameters since AUTOGEN adjusts related parameters automatically. (For a discussion of AUTOGEN, refer to the *OpenVMS System Manager's Manual*.) In the few instances where it is appropriate to change a parameter in the special parameter group, further explanation of the parameter is given in this chapter since special parameters are otherwise undocumented.

**When to use
SYSGEN**

If your tuning changes involve system parameters that are dynamic, plan to test the changes on a temporary basis first. This is the only instance where the use of SYSGEN is warranted for making tuning changes.

Once you are satisfied that the changes are working well, you should invoke AUTOGEN with the REBOOT parameter to make the changes permanent.

**Monitoring the
results**

After you change system values or parameters, you must monitor the results, as described in Chapter 3, Evaluating Tuning Success. You have two purposes for monitoring:

- You must ensure that the changes are not introducing new problems.
- You must evaluate the degree of success achieved.

You might want to return to the appropriate procedures in Chapters 12, 13, 14, and 15 as you evaluate your success after tuning and decide whether to pursue additional tuning efforts. However, always keep in mind that there is a point of diminishing returns in every tuning effort (see Chapter 3, When to stop tuning).

Compensating for Memory-Limited Behavior

Overview

This chapter describes corrective procedures for memory resource limitations described in Chapter 12. The following sections describe procedures to remedy specific conditions that you might have detected as the result of the investigation described in Chapter 13.

Purpose

To provide specific remedies for memory-limited performance.

Definition

An **ancillary control process** (ACP) acts an interface between user software and the I/O driver. The ACP supplements functions performed by the driver such as file and directory management.

Solutions for Memory-Limited Behavior

Reduce number of image activations

There are several ways to reduce the number of image activations. You and the programming staff should explore them all and apply those you deem feasible and likely to produce the greatest results.

Programs Versus Command Procedures

Excessive image activations can result from running large command procedures frequently, since all DCL commands (except those performed within the command interpreter) require an image activation. If command procedures are introducing the problem, consider writing programs to replace them.

Code Sharing

When code is actively shared, the cost of image startups decreases. Perhaps your installation has failed to design applications that share code. You should examine ways to employ code sharing wherever suitable. See the appropriate sections in Chapter 1, Developing a Strategy, and Chapter 6, Memory Sharing.

You will not see the number of image activations drop when you begin to use code sharing, but you should see an improvement in performance. The effect of code sharing is to shift the type of faults at image activation from hard faults to soft faults, a shift that results in performance improvement.

Designing Applications for Native Mode

Yet another source of excessive image activations is migration of programs from other operating systems without any design changes. For example, programs that employ the chaining technique on another operating system will not use memory efficiently on an OpenVMS AXP system if you simply recompile them and ignore design differences. When converting applications to run on an OpenVMS AXP system, always consider the benefits of designing and coding each application for native-mode operation.

Increase page cache size

You can enlarge the page cache by simply increasing the four page cache parameters: FREEGOAL, FREELIM, MPW_THRESH, and MPW_LOLIMIT. It is not necessary to remove balance slots or to reduce the working set size of any of the processes.

You should first increase the number of pages on the free-page list by augmenting FREELIM and FREEGOAL. Aim to provide at least one page on the free-page list for every process. FREEGOAL must always be greater than FREELIM. Generally, a good target size for FREEGOAL is three times FREELIM. If you feel your work load warrants it, you can increase the modified-page list size by increasing MPW_THRESH and MPW_LOLIMIT. Generally, MPW_LOLIMIT should be less than 10 percent of physical memory.

As another option, you could decide to reduce the number of balance slots, as described in Solutions for Memory-Limited Behavior.

Decrease page cache size

You decrease the size of the page cache by reducing the values for the system parameters MPW_LOLIMIT, MPW_THRESH, FREEGOAL, and FREELIM, maintaining the ratios suggested in the section describing page caches in Solutions for Memory-Limited Behavior.

In general, acceptable performance can be obtained by a page cache size that is one order of magnitude less than the available space for it and the working sets.

Adjust working set characteristics

If you have concluded that the working set quota or working set extent characteristics are incorrect in some cases, the corrective action depends on how the values were established. You must know whether the values affect a process, subprocess, detached process, or batch job.

Furthermore, if you need to fix a situation that currently exists, you must evaluate the severity of the problem. In some cases, you might have to stop images or processes or ask users to log out to permit your changes to become effective. You would take such drastic action only if the problem creates intolerable conditions that demand immediate action.

In addition to making specific changes in the working set quota and working set extent values, you should also address the need to modify the values of the system parameters BORROWLIM and GROWLIM. See the discussion of these changes and tuning to make borrowing more effective in Solutions for Memory-Limited Behavior.

Whenever you increase the values for working set extents, you should compare your planned values to the system parameter WSMAX, which specifies (on a systemwide basis) the maximum size that the working sets can achieve. It will do no good to specify any working set extent that exceeds WSMAX, since the working set can never actually achieve a count above the value of WSMAX. If you specify such a value, you should also increase WSMAX.

Establish Values for Ancillary Control Processes (ODS-1 Only)

This section will be of interest only if you are using ODS-1 disks.

Before studying the considerations for adjusting working set sizes for processes in general, consider the special case of the ACP. (Note that you will be using an ACP for disks only if you have ODS-1 disks.) The default size of the working set (and in this case, the working set quota, too) for all ACPs is determined by the system parameter ACP_WORKSET. If ACP_WORKSET is zero, the system calculates the working set size for you. If you want to provide a specific value for the working set default, you just specify the desired size in pages with AUTOGEN. (If your system uses multiple ACPs, remember that ACP_WORKSET is a systemwide parameter; any value you choose must apply equally well to all ACPs.)

If you decide to reduce ACP_WORKSET (with the intent of inducing modest paging in the ACP), use the SHOW SYSTEM command to determine how much physical memory the ACP currently uses. Then simply calculate the value that is 90 percent of the ACP's current usage. Set the system parameter ACP_WORKSET to the reduced value you calculate. However, to make the change effective for all ACPs on the system, not just the ones created after the change, you must reboot the system.

Once you reduce the size of ACP_WORKSET, observe the process with the SHOW SYSTEM command to verify that the paging you have induced in the ACP process is moderate. Your goal should be to keep the total number of page faults for the ACP below 20 percent of the direct I/O count for the ACP.

Establish Values for Other Processes

The following discussion applies to all processes other than ACPs. If the values were established for processes based on the defaults in the UAF, you should seek out the user, describe the intended change, and ask the user to enter the DCL command SET WORKING_SET/EXTENT or SET WORKING_SET/QUOTA, as appropriate.

If you observe satisfactory improvement from the new values, you must decide if the benefit would apply whenever the process runs or just during some specific activities. For specific cases, the user should enter the SET WORKING_SET command when needed. For a more consistent change, you would need to modify the UAF.

Modify Working Set Values

To modify values in the UAF, you invoke AUTHORIZE and use the SHOW and MODIFY commands to modify the values /WSQUOTA and /WSEXTENT for one or more users. If the SHOW command reveals that the values are the same as the defaults, probably the defaults have been applied. You should change all the assigned values in the existing records in the UAF, as appropriate. Then you should also modify the DEFAULT record in the UAF so that new accounts will receive the desired values.

If the working set characteristic values were adjusted by the process through the DCL command SET WORKING_SET or by a system service, you must convince the owner of the process that the values were incorrect and should be revised.

If the values were adjusted with the SET WORKING_SET command, the user can simply enter the command again, with revised values. However, if values were established by system services and the process is currently running and causing excessive paging, either the user must stop the image with Ctrl/Y or you must stop the process with the DCL command STOP. (Changing values set by system services typically requires code changes in the programs before they are run again.)

Establish Values for Detached Processes or Subprocesses

If the problem is introduced by a detached process or subprocess, you must also determine how the values became effective. If the values were established by the RUN command, they can be changed only if the user stops the detached process or subprocess (if it is running) and thereafter always starts it with a revised RUN command. (The user can stop the detached process or subprocess with the DCL command STOP.)

If the values were introduced by a system service, it is also necessary to stop the running detached process or subprocess, but code changes will be necessary as well.

If, however, the values were established by default, you might want to revise the values of the system parameters PQL_DWSEXTENT, PQL_DWSQUOTA, or both, particularly if the problem appears to be widespread. If the problem is not widespread, you can request users to use specific values that are less than or equal to their UAF defaults.

Unprivileged users cannot request values that will exceed their authorized values in the UAF. If such an increase is warranted, change the UAF records.

Establish Values for Batch Jobs

If the problem is introduced by a batch job, you must determine the source of the working set values.

If the values are those established for the queue when it was initialized, you cannot change them for this job while it is running. You must reinitialize the queue if you determine the changes would be beneficial for all future batch jobs. To reinitialize a batch queue, you must first stop it with the DCL command STOP/QUEUE, then restart it with the DCL command START/QUEUE. If the new working set values produce good results, you should ask the user to submit the job with the appropriate values in the future.

If the working set characteristics are obtained by default from the user's UAF, you might consider assigning values to the batch queues or creating additional batch queues. If you prefer to have values assigned from the UAF but have discovered instances where the best values are not in effect, before you change the UAF records, you need to determine if the changes would be beneficial at all times or only when the user submits certain jobs. It is generally better to ask the users to tailor each submission than to change UAF values that affect all the user's activities or batch queue characteristics that affect all batch jobs.

Tune to make borrowing more effective

If you have found few processes are taking advantage of loans, you should consider making the following adjustments:

- Decrease PFRATH.
- Decrease BORROWLIM, GROWLIM, or both.
- Increase the process limit WSEXTENT.

In decreasing PFRATH, you will increase the rate at which processes increase their working sets with AWSA. (See Chapter 6, AWSA, for a complete description of AWSA and its parameters. See the discussion of tuning AWSA for quick response in Solutions for Memory-Limited Behavior for guidelines regarding initial settings of the parameters.)

When you decrease BORROWLIM or GROWLIM, consider how much working set space you would like all processes to be able to obtain, according to the guidelines presented in Chapter 6, AWSA. As a rough guideline, you could target a BORROWLIM value from one-third to one-half of available memory and a GROWLIM value from one-sixth to one-fourth of available memory.

Be generous in establishing values for the working set extents, since the memory is only used when needed. As a general practice, set the working set extent value to the largest value you expect will be needed. The section in Solutions for Memory-Limited Behavior describes the various ways you adjust the working set extent characteristic. (You might also need to increase WSMAX.)

Tune AWSA to respond quickly

You might want to increase the response from AWSA to paging so that AWSA rapidly establishes a working set size that keeps paging to a reasonable rate for your configuration and work load. To do so, you need to reduce PFRATH, increase WSINC, or both.

Think of PFRATH as the target maximum paging rate for any process in the system. PFRATH should always be greater than PFRATL. As a rule, values of PFRATH larger than 32 (which specifies a desired maximum rate of 3 page faults per second of CPU time) is unreasonable.

The system parameter WSINC defines the number of pagelets by which the working set limit increases when AWSA determines that it needs to expand. The maximum practical value for this parameter is therefore the difference between WSMAX (which is the maximum size increase that any working set can experience) and MINWSCNT (which is the minimum working set size). In practical terms, however, to avoid wasting memory, it makes sense to set WSINC smaller than this difference. A fairly good rule of thumb is to set WSINC to match an approximation of a typical user's WSDEFAULT value. Such a value allows the processes to increase fairly rapidly, while limiting the potential maximum waste to the amount needed to minimally support one user.

If you are not fully satisfied with the results produced by tuning WSINC and PFRATH, you could decrease AWSTIME. However, do not decrease the value of AWSTIME below the value of QUANTUM. Your goal should be to achieve a value for AWSTIME that follows the overall trend in the total size of all the working sets. If you establish too small a value for AWSTIME, AWSA could be responding to too many frequent drastic working set size changes and not to the overall trend the changes describe.

Disable voluntary decrementing

If you find that some of the working set sizes are oscillating continuously while the processes should be in a stable state of memory demand, it is possible that voluntary (time-based) decrementing is forcing paging. To avoid this, set PFRATL to zero. This will effectively turn off voluntary decrementing. As a result, your system will rely solely on load-based memory reclamation (swapper trimming or outswapping).

Optionally, you might want to set WSDEC to zero. If you do, it will be more obvious to you or others at some future time that voluntary decrementing is turned off on the system. However, setting only WSDEC to zero does not disable the checking that automatic working set adjustment performs for voluntary decrementing.

Tune voluntary decrementing

It might be that some time-based working set trimming is desirable to reclaim memory that is not really needed (to avoid taking needed memory away from other processes, for example). However, the parameters are set so high that too much paging occurs. In this case, you should decrease WSDEC or PFRATL. Setting just PFRATL to zero or setting both WSDEC and PFRATL to zero turns off time-based decrementing. However, if you choose to maintain some voluntary decrementing, remember that to avoid fixed oscillation, WSDEC should be smaller than WSINC. In addition, WSINC and WSDEC should be relatively prime (that is, WSINC and WSDEC should have no common factors). A good starting value for WSDEC would be an order of magnitude smaller than a typical user's WSDEFAULT value.

Turn on voluntary decrementing

Sometimes time-based shrinking is completely turned off when it should be turned on. To enable voluntary decrementing, do the following:

- Set WSDEC and PFRATL greater than zero.
- Observe the guidelines for tuning voluntary decrementing in Solutions for Memory-Limited Behavior.

Enable AWSA

To turn on the part of automatic working set adjustment that permits processes to increase their working set sizes, you must set WSINC to a value greater than zero. The default parameter settings established by AUTOGEN at system installation are good starting values for most work loads and configurations.

Adjust swapper trimming

When you determine that a paging problem is caused by excessive swapper trimming, SWPOUTPGCNT is too small. There are two approaches you can use. The first is to increase SWPOUTPGCNT to a value that is large enough for a typical process on the system to use as its working set size. This approach effectively causes the swapper to swap the processes at this value rather than reduce them to a size that forces them to page heavily.

The second approach completely disables second-level swapper trimming by setting SWPOUTPGCNT to a value equal to the largest value for WSQUOTA for any process on the system. This has the effect of shifting the bulk of the memory management to outswapping, with no second-level swapper trimming.

In conjunction with swapper trimming, the system uses the system parameter LONGWAIT to control how much time must pass before a process is considered idle. The swapper considers idle processes to be better candidates for memory reclamation than active processes. The ideal value for LONGWAIT is the length of time that accurately distinguishes an idle or abandoned process from one that is momentarily inactive. Typically, this value is in the range of 3 to 20 seconds. You would increase LONGWAIT to force the swapper to give processes a longer time to remain idle before they become eligible for swapping or trimming. This approach will prove most productive when the work load is mixed and includes interactive processes. If the work load is composed primarily of nonreal-time processes, you might consider increasing DORMANTWAIT.

Convert to a system that rarely swaps

To severely reduce the swapping activity on a system, you can set the system parameter BALSETCNT equal to a value that is two less than the value of the system parameter MAXPROCESSCNT, thus allowing the maximum number of processes to operate concurrently. At the same time, you should set the system parameter SWPOUTPGCNT to a minimum value.

As a secondary action, you would reduce the working set quotas, following the recommendations for adjusting working set characteristics in Solutions for Memory-Limited Behavior.

These actions produce a system that primarily pages.

Adjust BALSETCNT

You might want to use the BALSETCNT system parameter as a tuning control for paging or swapping. Reducing BALSETCNT can reduce paging, while increasing BALSETCNT can decrease swapping. BALSETCNT is a parameter that affects a number of other parameters, so you should be conservative in changing it.

Reduce BALSETCNT to Reduce Paging

If you reduce the number of balance set slots by decreasing the parameter BALSETCNT, you can reduce the demand for memory by limiting the number of processes that compete for memory at a given time.

From the output provided by the DCL command SHOW MEMORY under a very heavy work load, you know the number of balance slots available and in use. If balance slots are available under a heavy load, it is safe to reduce the value of BALSETCNT by that amount. However, if no balance slots are available and you reduce BALSETCNT, you are likely to force swapping to occur while the system is loaded.

Increase BALSETCNT to Decrease Swapping Problems

If active swapping is being caused by a lack of balance slots when there is available memory, the first step is to increase BALSETCNT. The easiest thing to do is to set BALSETCNT equal to a value that is two less than the system parameter MAXPROCESSCNT. This guarantees that a balance slot will be available for any process that can be created. Swapping will then be forced only when memory is insufficient to meet the demand.

Rather than immediately setting BALSETCNT equal to a value that is two less than MAXPROCESSCNT (which is the largest useful value), you might try a more gradual approach. Divide the remaining free memory (as displayed by the SHOW MEMORY command) by the size of a typical working set, and then increase BALSETCNT by this number.

Reduce large page caches

If active swapping is caused by a lack of free memory, which in turn is caused by unnecessarily large page caches, as a first step, reduce the size of the caches by lowering FREELIM and FREEGOAL or MPW_LOLIMIT and MPW_THRESH. (Remember that MPW_HILIMIT relates to the maximum size of the modified-page list rather than the target minimum size.)

Good starting ratios for these parameters are given in the discussion of increasing page cache size in Solutions for Memory-Limited Behavior. Keep in mind that the problem of overly large caches is caused by mistuning in the first place. The AUTOGEN command procedure will not generate page cache values that are excessively large.

Curtail large, compute-bound process

Before suspending a large, low-priority, compute-bound process, Digital strongly recommends that you curtail its memory allocation. If the process has not had a significant event for 10 seconds or more (page fault, direct or buffered I/O, CPU time allocation), you can decrease DORMANTWAIT to make the process a more likely outswap candidate.

Suspend large, compute-bound process

When you decide to suspend a large, compute-bound process, be sure that it is not sharing files with other processes. Otherwise, the large, compute-bound process might have a shared file locked when you suspend it. If this should happen, you will soon observe that other processes become stalled. You must resume the large, compute-bound process as soon as possible with the DCL command SET PROCESS/RESUME if you are unable to achieve the benefit that suspending offers. In this case, refer to Solutions for Memory-Limited Behavior for appropriate corrective action.

Control growth of large, compute-bound processes

When it becomes clear that a large, compute-bound process gains control of more memory than is appropriate, you might find it helpful to lower the process's working set quota. Take this action if you conclude that this process should be the one to suffer the penalty of page faulting, rather than forcing the other processes to be outswapped too frequently. Solutions for Memory-Limited Behavior describes how to make adjustments to working set quotas.

Enable swapping for disk ACPs (ODS-1 only)

If a disk ACP has been set up so that it will not be outswapped and you determine that the system would perform better if it were, you must use AUTOGEN to modify the system parameter ACP_SWAPFLGS and then reboot the system. The *OpenVMS System Management Utilities Reference Manual* describes how to specify the flag value for ACP_SWAPFLGS that will permit swapping of the ACP.

Enable swapping for other processes

If you determine that users have been disabling swapping for their processes and that the effect of locking one or more processes in memory has been damaging to overall performance, you must explore several options.

If there are no valid reasons to disable swapping for one or more of the processes, you must convince the users to stop the practice. If they will not cooperate, you can remove privileges so they cannot disable swapping. Use AUTHORIZE to change privileges. (The PSWAPM privilege is required to issue the SET PROCESS/NOSWAPPING command.)

However, if the users have valid reasons for disabling swapping, you should carefully examine what jobs are running concurrently when the performance degrades. It is possible that rescheduling a few of the jobs will be sufficient to improve overall performance. See the discussion about adding page files in Solutions for Memory-Limited Behavior.

Reduce number of concurrent processes

You can reduce the number of concurrent processes by lowering the value of MAXPROCESSCNT. A change in that value has implications for the largest number of system parameters. Therefore, you should change the value of MAXPROCESSCNT in conservative steps.

Discourage working set loans

If working sets are too large because processes are using their loan regions (above WSQUOTA), you can curtail loaning by increasing GROWLIM and BORROWLIM. (To completely disable borrowing, just set GROWLIM and BORROWLIM equal to the special system parameter PHYSICAL_MEMORY, which is the upper bound on the amount of physical memory that the system will configure when the system is booted.)

You might also consider reducing the WSEXTENT size for some processes in the UAF file. If you go so far as to set the WSEXTENT values equal to the WSQUOTA values, you completely disable borrowing for those processes.

Increase swapper trimming memory reclamation

If you lower the value of SWPOUTPGCNT, you increase the amount of memory reclaimed every time second-level trimming is initiated. However, this is the parameter that most effectively converts a system from a swapping system to a paging one and vice versa. As you lower the value of SWPOUTPGCNT, you run the risk of introducing severe paging.

Reduce rate of inswapping

If you increase the special system parameter SWPRATE, you will reduce the frequency at which outswapped processes are inswapped. SWPRATE is the minimum real time between inswaps of compute-bound processes. For this calculation, any process whose current priority is less than or equal to the system parameter DEFPRI is considered to be compute bound.

Induce paging to reduce swapping

To induce paging on a system that swaps excessively, you need to lower the working set quotas, as described in Solutions for Memory-Limited Behavior. In addition, you should increase the value of PFRATH and you might also reduce the value of WSINC. With these modifications, you will slow down the responsiveness of AWSA to paging. The processes will not acquire additional working set space as readily.

It might be worthwhile to check the number of concurrent jobs in the batch queues. Use the DCL command SHOW SYSTEM /BATCH to examine the number and size of the batch jobs. If you observe many concurrent batch jobs, you might decide to enter the DCL commands STOP/QUEUE and START/QUEUE/JOB_LIMIT to impose a restriction on the number.

Add paging files

If the system disk is saturated by paging, as described in Chapter 13, Analyzing the Excessive Paging Symptom, you might want to consider adding one or more paging files, on separate disks, to share the activity. This option is more attractive when you have space available on a disk that is currently underutilized. Use the SYSGEN commands CREATE and INSTALL to add paging files on other disks. (See the *OpenVMS System Management Utilities Reference Manual*.)

The discussion of AUTOGEN in the *OpenVMS System Manager's Manual* includes additional considerations and requirements for modifying the size and location of the paging file.

Reduce demand or add memory

At this point, when all the tuning options have been exhausted, there are only two options: reduce the demand for memory by modifying the work load or add memory to the system.

Reduce Demand

Chapter 1, *Developing a Strategy*, describes a number of options (including workload management) that you can explore to shift the demand on your system so that it is reduced at peak times.

Add Memory

If you conclude you need to add memory, your next concern is to determine how much memory. You should add as much memory as you can afford. If you need to establish the amount more scientifically, you could try the following empirical technique:

- Determine or estimate a paging rate you believe would represent a tolerable level of paging on the system. (You should make allowances for global valid faults if many applications share memory by deducting the global valid fault rate from the total page fault rate.)
- Turn off swapper trimming (set `SWPOUTPGCNT` to the maximum value found for `WSQUOTA`).
- Give the processes large enough working set quotas so that you achieve the tolerable level of paging on the system while it is under load.

The amount of memory required by the processes that are outswapped represents an approximation of the amount of memory your system would need to obtain the desired performance under load conditions.

Once you add memory to your system, be sure to invoke `AUTOGEN` so that new parameter values can be assigned on the basis of the increased physical memory size.

Compensating for I/O-Limited Behavior

Overview

This chapter describes corrective procedures for I/O resource limitations described in Chapter 12.

Purpose

To provide specific remedies for I/O-limited performance.

Definition

A **cache** is a block of memory used to minimize the physical transfer of data between physical memory and secondary storage devices.

Solutions for I/O-Limited Behavior

All the tuning solutions for performance problems based on I/O limitations involve using memory to relieve the I/O subsystem. The three most accessible mechanisms are the virtual I/O cache, the ACP caches, and RMS buffering.

Use virtual I/O caching

The virtual I/O cache is a clusterwide, write-through, file-oriented, disk cache that can reduce the number of disk I/O operations and increase performance. The purpose of the virtual I/O cache is to increase system throughput by reducing file I/O response times with minimum overhead. The virtual I/O cache operates transparently of system management and application software, and maintains system reliability while it significantly improves virtual disk I/O read performance.

How Does the Cache Work?

The virtual I/O cache can store data files and image files. For example, ODS-2 disk file data blocks are copied to the virtual I/O cache the first time they are accessed. Any subsequent read requests of the same data blocks are satisfied from the virtual I/O cache (hits) eliminating any physical disk I/O operations (misses) that would have occurred.

Depending on your system work load, you should see increased application throughput, increased interactive responsiveness, and reduced I/O load.

Note

Applications that initiate single read and write requests will not benefit from virtual I/O caching as the data is never reread from the cache. Applications that rely on implicit I/O delays might abort or yield unpredictable results.

Several policies govern how the cache manipulates data as follows:

- Write-through—All write I/O requests are written to the cache as well as to the disk.
- Least Recently Used (LRU)—If the cache is full, the least recently used data in the cache is replaced.
- Cached data maintained across file close—Data remains in the cache after a file is closed.
- Allocate on read and write requests—Cache blocks are allocated for read and write requests.

Displaying Virtual I/O Cache Statistics

Use the DCL command `SHOW MEMORY/CACHE/FULL` to display statistics about the virtual I/O cache as shown in the following example:

```

System Memory Resources on 15-MAR-1994 12:35:47 PM
Virtual I/O Cache
Total Size (Kbytes)      ①          ⑥ Read IO Count
Free Kbytes              ②          ⑦ Read Hit Count
Kbytes in Use           ③          ⑧ Read Hit Rate
Write IO Bypassing Cache ④          ⑨ Write IO Count
Files retained           ⑤          ⑩ Read IO Bypassing Cache
    
```

- ① Total size—Total number of kilobytes owned by the virtual I/O cache.
- ② Free Kbytes—Current amount of memory owned by the virtual I/O cache that does not contain valid file data.
- ③ Kbytes in use—Current amount of memory owned by the virtual I/O cache that contains valid file data.
- ④ Write I/O bypassing the cache—Total number of file write I/O requests that did not involve the cache. File I/O requests bypass the cache if one of the following conditions is true: (1) the request was made using I/O function modifiers or (2) the request size exceeds 35 blocks.
- ⑤ Files retained—Current number of closed files that still have valid file data in the virtual I/O cache.

- ⑥ Read I/O count—Total number of file read I/O requests processed by the virtual I/O cache since system startup.
- ⑦ Read hit count—Total number of file read I/O requests satisfied by the virtual I/O cache since system startup.
- ⑧ Read hit rate—Percentage of read hits (read hit count compared with total read I/O count). Note that a 35% hit ratio represents the break-even point for CPU cost.
- ⑨ Write I/O count—Total number of file write I/O requests processed by the virtual I/O cache since system startup.
- ⑩ Read I/O bypassing the cache—Total number of file read I/O requests that did not involve the cache. File I/O requests bypass the cache if one of the following conditions is true: (1) the request was made using I/O function modifiers or (2) the request size exceeds 35 blocks.

Enabling Virtual I/O Caching

By default, virtual I/O caching is enabled. Use the system parameter, `VCC_FLAGS`, to enable or disable caching. Change the value of the `VCC_FLAGS` parameter in `MODPARAMS.DAT` as follows:

- `VCC_FLAGS = 0` to disable caching
- `VCC_FLAGS = 1` to enable caching

Once you have updated `MODPARAMS.DAT` to change the value of `VCC_FLAGS`, you must run `AUTOGEN` and reboot the node or nodes on which you have enabled or disabled caching. Caching is automatically enabled or disabled during system initialization. No further user action is required.

Determining If Virtual I/O Caching Is Enabled

Check the system parameter `VCC_FLAGS` to see if virtual I/O caching is enabled by using `SYSGEN` as shown in the following example:

```
$ RUN SYS$SYSTEM:SYSGEN
SYSGEN> SHOW VCC_FLAGS
```

A value of 0 indicates that caching is disabled; the value 1 indicates caching is enabled.

Adjusting the Virtual I/O Cache Size

The size of the virtual I/O cache is controlled by the system parameter `VCC_MAXSIZE`. The amount of memory specified by this parameter is statically allocated at system initialization and remains owned by the virtual I/O cache.

To increase or decrease the size of the cache, modify `VCC_MAXSIZE` and reboot the system.

Virtual I/O Cache and VMScluster Configurations

The cache works on all supported configurations from single-node systems to large mixed-interconnect VMScluster systems. The virtual I/O cache is nodal, that is, the cache is local to each VMScluster member. Any base system can support virtual I/O caching; a VMScluster license is not required to use the caching feature.

Note

If any member of a VMScluster does not have caching enabled, then no caching can occur on any node in the VMScluster (including the nodes that have caching enabled). This condition remains in effect until the node or nodes that have caching disabled either enable caching or leave the cluster.

The lock manager controls cache coherency. The cache is flushed when a node leaves the VMScluster. Files opened on two or more nodes with write access on one or more nodes are not cached.

Use a RAM disk

Using a RAM disk such as the optional software product, DECram for OpenVMS, can also enhance disk I/O performance by replacing disk I/O with main memory access. A **RAM disk** is a virtual disk device that resides in physical memory. The operating system can read from and write to the RAM disk using standard disk I/O operations.

Two types of applications benefit from using RAM disks as follows:

- Applications that frequently use system images
- Modular applications that use temporary, transient files

Note that the contents of a RAM disk do not survive a reboot.

Remove blockage due to ACP

Of the four sources of bottlenecks, the ACP lockout problem is the easiest to detect and solve. Moreover, it responds to software tuning.

The solution for an ACP lockout caused by a slow disk sharing an ACP with one or more fast disks requires that you dismount the slow device with the DCL command DISMOUNT, then enter the DCL command MOUNT/PROCESSOR=UNIQUE to assign a private ACP to the slow device. Note that you will be using an ACP for disks only if you have ODS-1 disks. However, be aware that each ACP has its own working set and caches. Thus, creating multiple ACPs requires the use of additional memory.

Also, there are situations that might share some of the symptoms of an ACP lockout that will not respond to adding an ACP. For example, when substantial I/O activity is directed to the same device so that the activity in effect saturates the device, adding an ACP for another device without taking steps to redirect or redistribute some of the I/O activity to the other device yields no improvement.

Blockage Due to a Device, Controller, or Bus (ODS-1 Only)

When you are confronted with the situation where users are blocked by a bottleneck on a device, a controller, or a bus, your first step should be to evaluate whether you can take any action that would make less demand on the bottleneck point.

Reduce Demand on the Device That Is the Bottleneck If the bottleneck is a particular device, you might try any of the following suggestions, as appropriate. The suggestions begin with areas that are of interest from a tuning standpoint and progress to application design areas.

One of the first things you should determine is whether the problem device is used for paging or swapping files and if this activity is contributing to the I/O limitation. If so, you need to consider ways to shift the I/O demand. Possibilities include moving either the swapping or paging file (or both, if appropriate) to another disk. However, if the bottleneck device is the system disk, you cannot move the entire paging file to another disk; a minimum paging file is required on the system disk. See the discussion of AUTOGEN in the *OpenVMS System Manager's Manual* for additional information and suggestions.

Another way to reduce demand on a disk device is to redistribute the directories over one or more additional disks, if possible. As described earlier in this section, you might decide to allocate memory to multiple ACPs (ODS-1 only) to permit redistributing some of the disk activity to other disks. Solutions for I/O-Limited Behavior discusses RMS caching and some of the implications of using RMS to alleviate the I/O on the device. Also consider that, if the disks have been in use for some time, the files might be fragmented. You should run the Backup utility to eliminate the fragmentation. (See the *OpenVMS System Manager's Manual*.) If this approach is highly successful, institute a more regular policy for running backups of the disks.

As a next step, you should try to schedule work that heavily accesses the device over a wider span of time or with a different mix of jobs so that the demand on the device is substantially reduced at peak times. Moving files to other existing devices to achieve a more even distribution of the demand on all the devices is one possible method. Modifications to the applications might also help distribute demand over several devices. Greater changes might be necessary if the file organization is not optimal for the application (for example, perhaps the application employs

a sequential disk file organization when an indexed sequential organization would be preferable).

Reduce Demand on the Controller That Is the Bottleneck When a controller is the bottleneck, examine the activity at the slowest device on the controller and its relationship to the other devices. You might find it helpful to group the slower devices together on the same controller (when you have more than one available).

Reduce Demand on the Bus That Is the Bottleneck Another suggestion is to place controllers on separate buses. Again, you want to segregate the slower speed units from the faster units.

When a bus becomes the bottleneck, the only solution is to acquire another bus so that some of the load can be redistributed over both buses.

Enlarge Hardware Capacity

If there seem to be few appropriate or productive ways to shift the demand away from the bottleneck point using available hardware, you might have to acquire additional hardware. Adding capacity can refer to either supplementing the hardware with another similar piece or replacing the item with one that is larger, faster, or both.

Try to avoid a few of the more common mistakes. It is easy to conclude that more disks of the same type will permit better load distribution, when the truth is that providing another controller for the disks you already have might bring much better results. Likewise, rather than acquiring more disks of the same type, the real solution might be replacing one or more existing disks with a disk that has a faster transfer rate. Another mistake to avoid is acquiring disks that immediately overburden the controller or bus you place them on.

To make the correct choice, you must know whether your problem is due to limitations in space and placement or to speed limitations. If you need speed improvement, be sure you know whether it is needed at the device or the controller. You must invest the effort to understand the I/O subsystem and the distribution of the I/O work load across it before you can expect to make the correct choices and configure them optimally. You should try to understand at all times just how close to capacity each part of your I/O subsystem is.

Improve RMS caching

The *Guide to OpenVMS File Applications* is your primary reference for information on tuning RMS files and applications. RMS reduces the load on the I/O subsystems through buffering. Both the size of the buffers and the number of buffers are important in this reduction. In trying to determine reasonable values for buffer sizes and buffer counts, you should look for the optimal balance between minimal RMS I/O (using sufficiently large buffers) and minimal memory management I/O. Note that, if you define RMS buffers that are too large, you can more than

fill the process's entire working set with these buffers, ultimately inducing more process paging.

Adjust file system caches

The considerations for tuning disk file system caches are similar to those for tuning RMS buffers. Again, the issue is the minimizing of I/O. A disk file system maintains caches of various file system data structures such as file headers and directories. These caches are allocated from paged pool when the volume is mounted for ODS-2 volumes (default). (For an ODS-1 ACP, they are part of the ACP working set.) File system operations that only read data from the volume (as opposed to those that write) can be satisfied without performing a disk read if the desired data items are in the file system caches. It is important to seek an appropriate balance point that matches the work load.

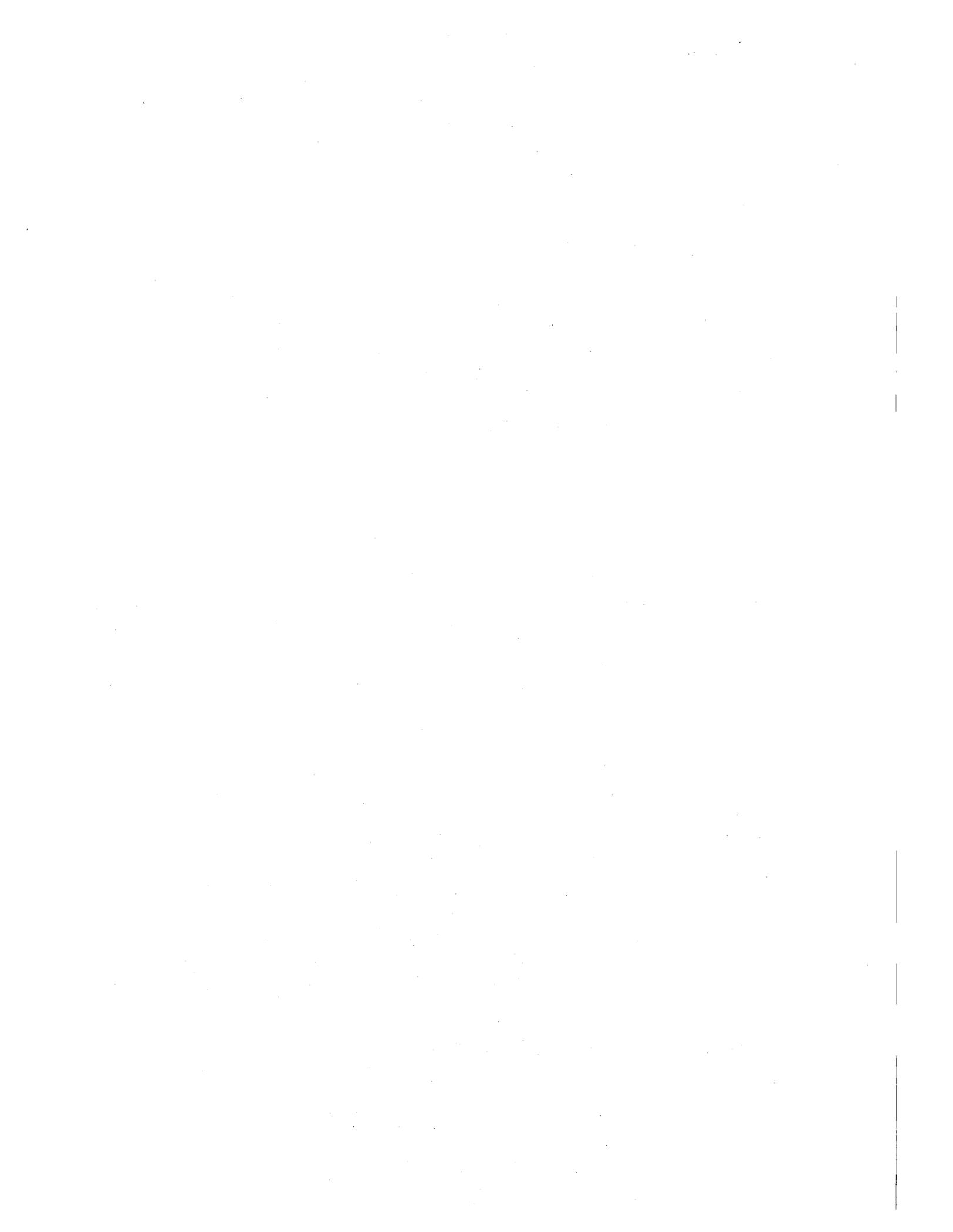
To evaluate file system caching activity, do the following:

1. Enter the MONITOR FILE_SYSTEM_CACHE command.
2. Examine the data items displayed. (For detailed descriptions of these items, refer to the *OpenVMS System Management Utilities Reference Manual*.)
3. Invoke SYSGEN and modify, if necessary, appropriate ACP system parameters.

Data items in the FILE_SYSTEM_CACHE display correspond to ACP parameters as follows:

FILE_SYSTEM_CACHE Item	ACP/XQP Parameters
Dir FCB	ACP_SYSACC ACP_DINDXCACHE
Dir Data	ACP_DIRCACHE
File Hdr	ACP_HDRCACHE
File ID	ACP_FIDCACHE
Extent	ACP_EXTCACHE ACP_EXTLIMIT
Quota	ACP_QUOCACHE
Bitmap	ACP_MAPCACHE

When you change the ACP cache parameters, remember to reboot the system to make the changes effective.



Compensating for CPU-Limited Behavior

Overview

This chapter describes corrective procedures for CPU resource limitations described in Chapter 12.

Purpose To provide specific remedies for CPU-limited performance.

Definition **Compute-bound** refers to slow system response due to the number of computations.

Solutions for CPU-Limited Behavior

There are only two ways to apply software tuning controls to alleviate performance problems related to CPU limitations:

- Specify explicit priorities (for jobs or processes).
- Modify the system parameter QUANTUM.

The other options, reducing demand or adding CPU capacity, are really not tuning solutions.

Adjust priorities

When a given process or class of processes receives inadequate CPU service, the surest technique for improving the situation is to raise the priority of the associated processes. To avoid undesirable side effects that can result when a process's base priority is raised permanently, it is often better to simply change the application code to raise the priority only temporarily. You should adopt this practice for critical pieces of work.

Priorities are established for processes through the UAF value. Users with appropriate privileges (ALTPRI, GROUP, or WORLD) can modify their own priority or those of other processes with the DCL command SET PROCESS/PRIORITY. Process priorities can also be set and modified during execution with the system service \$SETPRI. See Chapter 6, OpenVMS Scheduling.

Priorities are assigned to subprocesses and detached processes with the DCL command `RUN/PRIORITY` or with the `$CREPRC` system service at process creation. The appropriately privileged subprocess or detached process can modify its priority while running with the `$SETPRI` system service.

Batch queues are assigned priorities when they are initialized (`INITIALIZE/QUEUE/PRIORITY`) or started (`START/QUEUE/PRIORITY`). While you can adjust the priorities on a batch queue by stopping the queue and restarting it (`STOP/QUEUE` and `START/QUEUE/PRIORITY`), the only way to adjust the priority on a process while it runs is through the system service `$SETPRI`.

Adjust QUANTUM

By reducing `QUANTUM`, you can reduce the maximum delay a process will ever experience waiting for the CPU. The trade-off here is that, as `QUANTUM` is decreased, the rate of time-based context switching will increase, and therefore the percentage of the CPU used to support CPU scheduling will also increase. When this overhead becomes excessive, performance will suffer.

Note

In general, do not adjust `QUANTUM` unless you know exactly what you expect to accomplish and are aware of all the ramifications of your decision.

Reduce demand or add CPU capacity

You need to explore ways to schedule the work load so that there are fewer compute-bound processes running concurrently. Chapter 1, *Developing a Strategy*, includes a number of suggestions for accomplishing this goal.

You might find it possible to redesign some applications with improved algorithms to perform the same work with less processing. When the programs selected for redesign are those that run frequently, the reduction in CPU demand can be significant.

You also want to control the concurrent demand for terminal I/O.

Types of CPU Capacity

If you find that none of the previous suggestions or workload management techniques satisfactorily resolve the CPU limitation, you need to add capacity. It is most important to determine which type of CPU capacity you need, since there are two different types that apply to very different needs.

Work loads that consist of independent jobs and data structures lend themselves to operation on multiple CPUs. If your work load has such characteristics, you could add a processor to gain CPU capacity. The processor you choose might be of the same speed or faster, but it could also be slower. It takes over some portion of

the work of the first processor. (Separating the parts of the work load in optimal fashion is not necessarily a trivial task.)

Other work loads must run in a single-stream environment since many pieces of work depend heavily on the completion of some previous piece of work. These work loads demand that CPU capacity be increased by increasing the CPU speed with a faster model of processor. Typically, the faster processor performs the work of the old processor, which is replaced rather than supplemented.

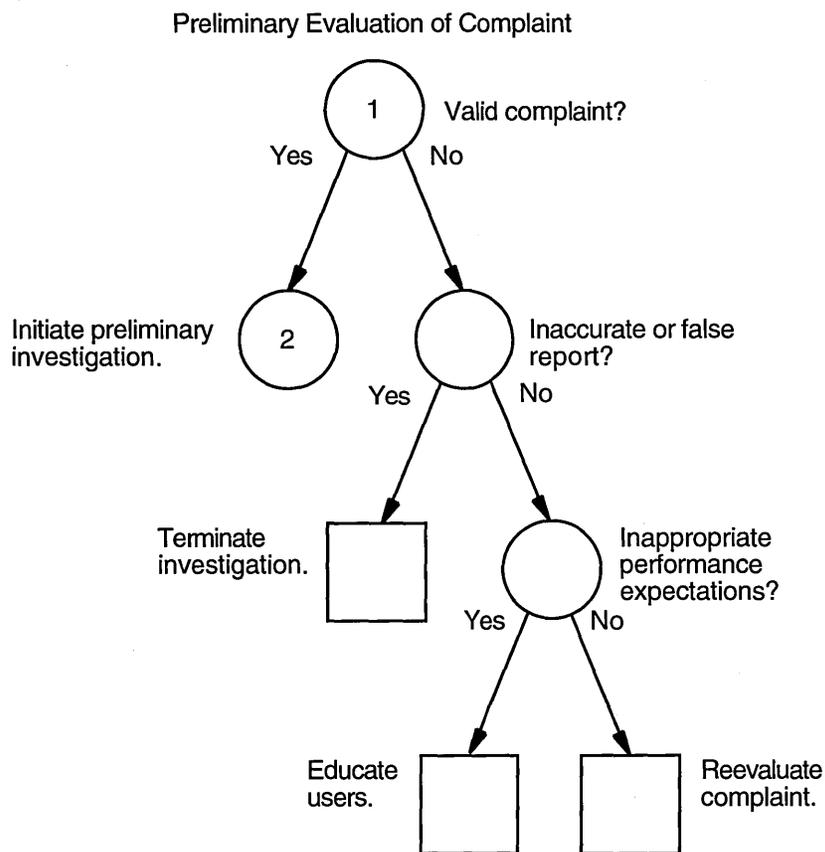
To make the correct choice, you must analyze the interrelationships of the jobs and the data structures.



Decision Trees

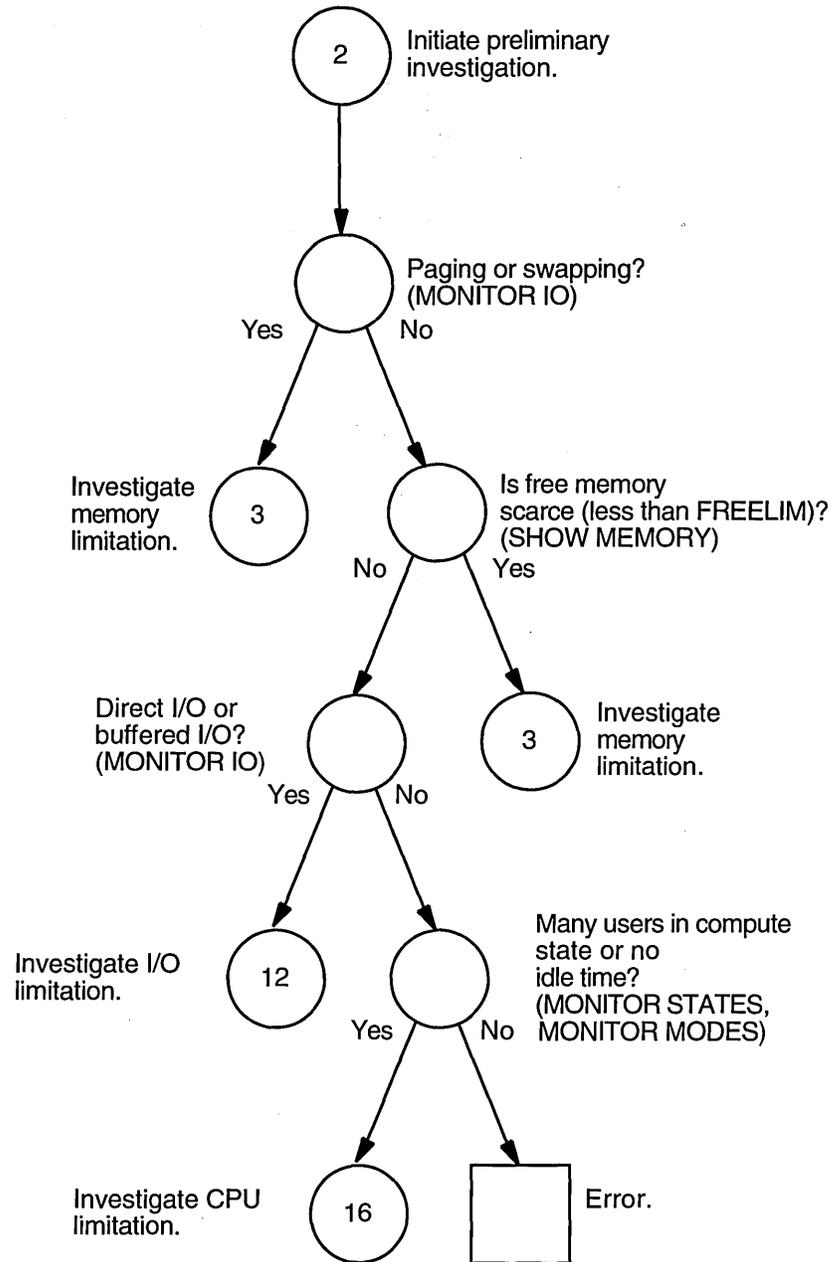
This appendix lists decision trees you can use to conduct the evaluations described in this manual. A decision tree consists of nodes that describe steps in your performance evaluation. Numbered nodes indicate that you should proceed to the next diagram that contains that number.

Figure A-1 Verifying the Validity of a Performance Complaint



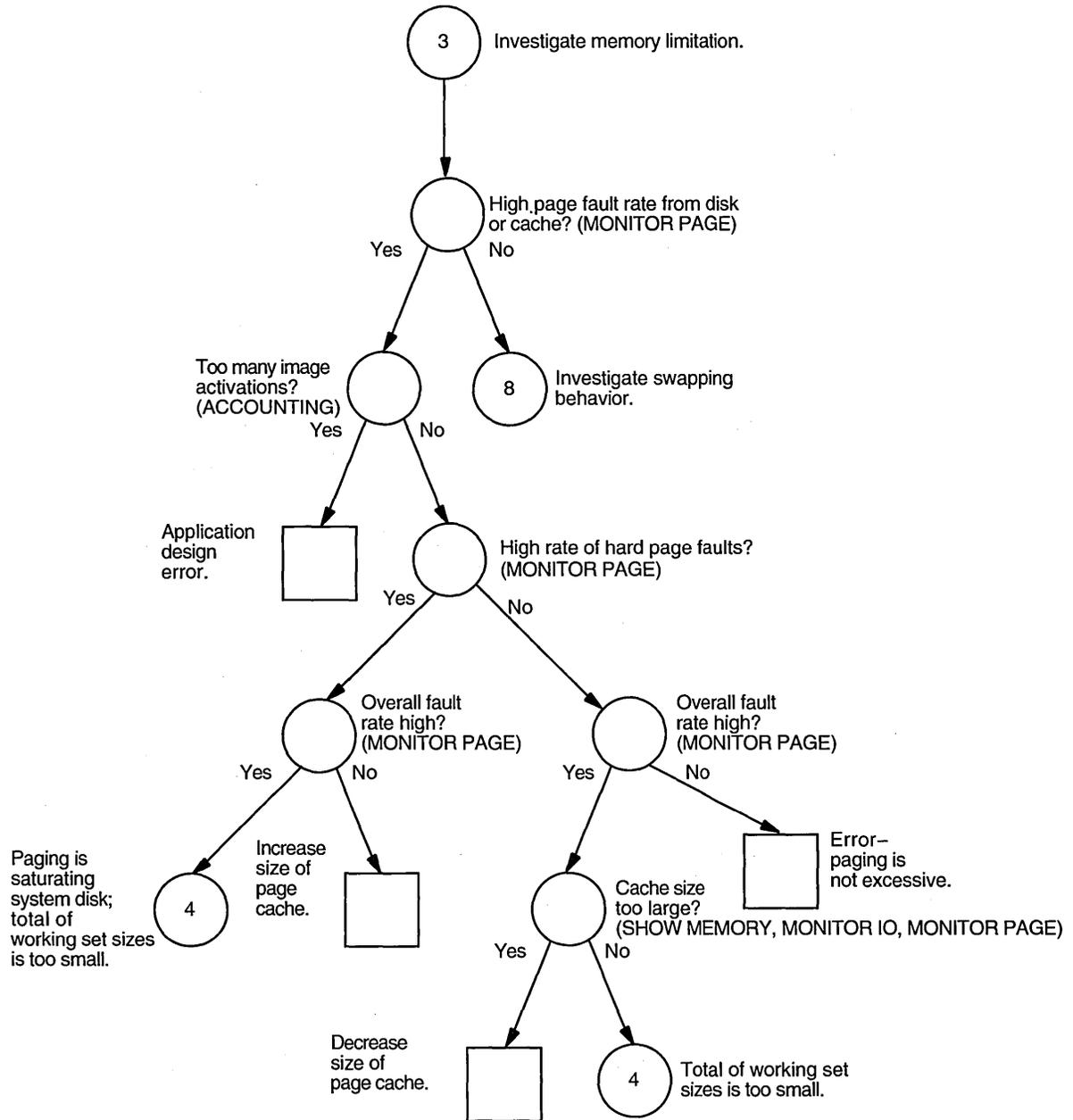
ZK-1131-GE

Figure A-2 Steps in the Preliminary Investigation Process



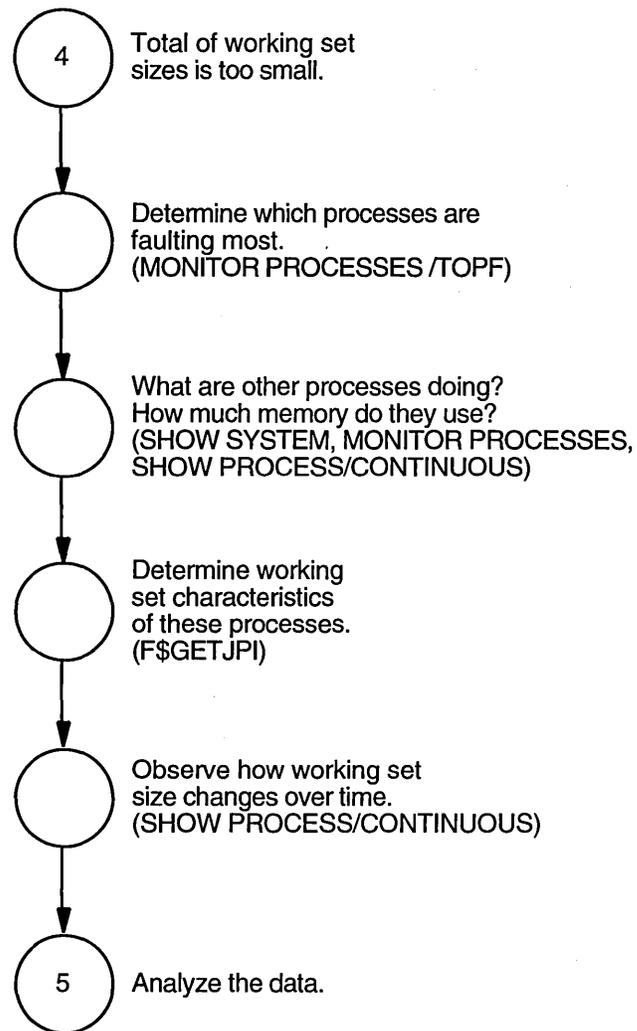
ZK-1132-GE

Figure A-3 Investigating Excessive Paging—Phase I



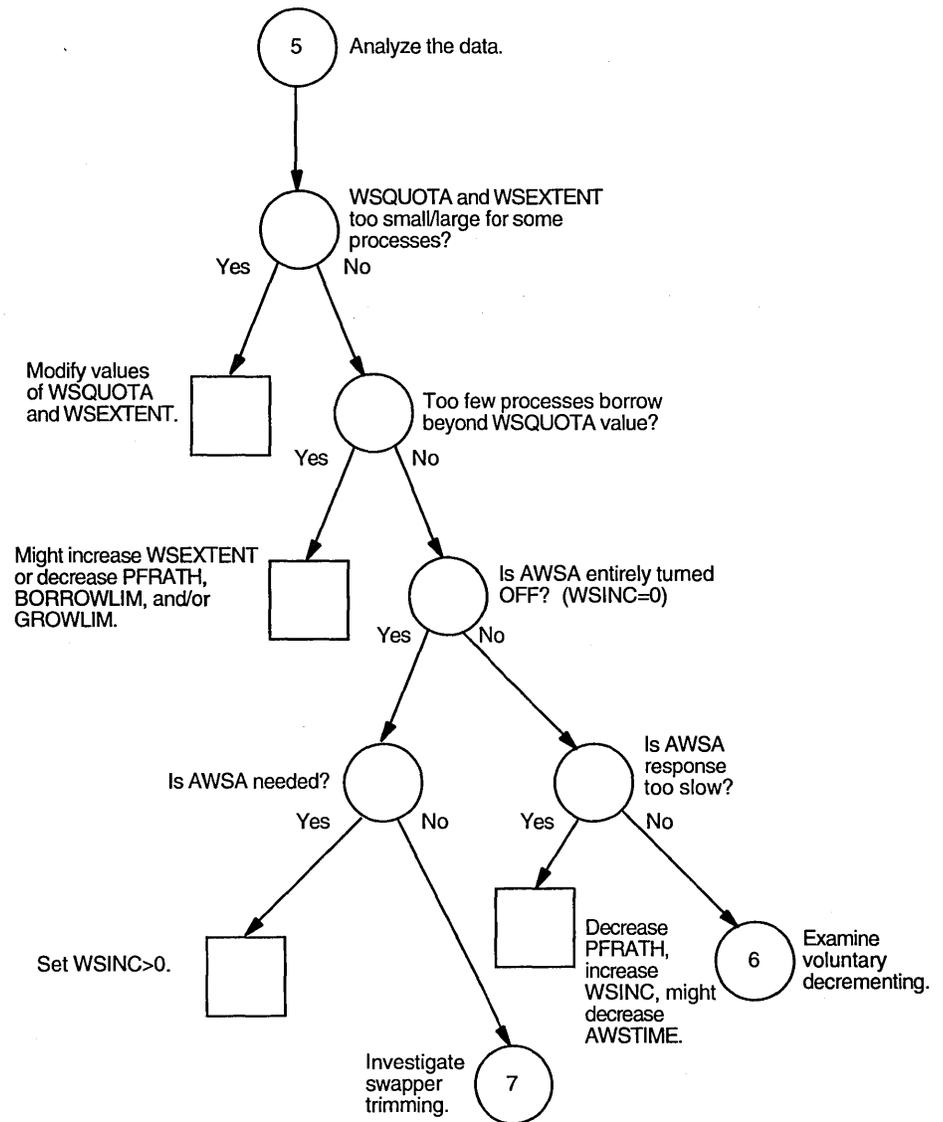
ZK-1133-GE

Figure A-4 Investigating Excessive Paging—Phase II



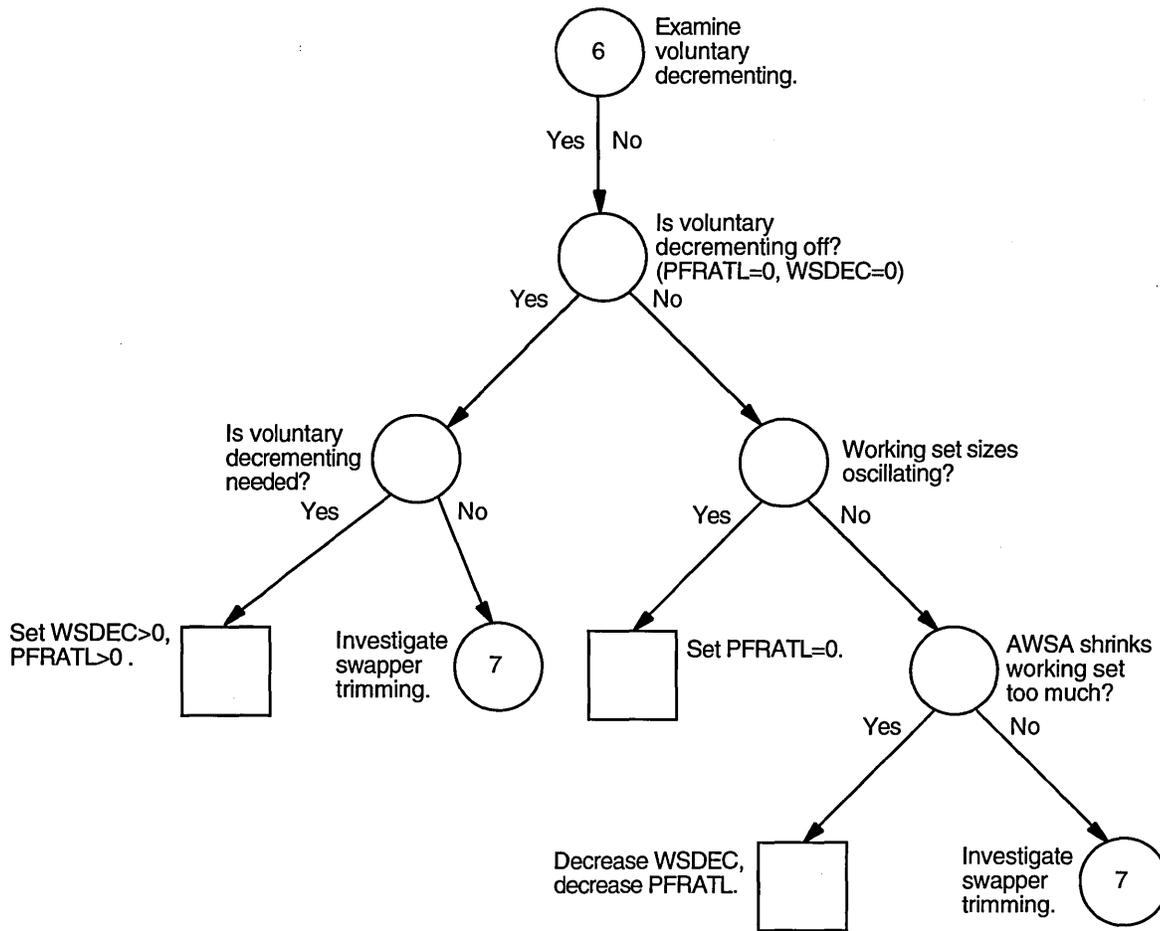
ZK-1134-GE

Figure A-5 Investigating Excessive Paging—Phase III



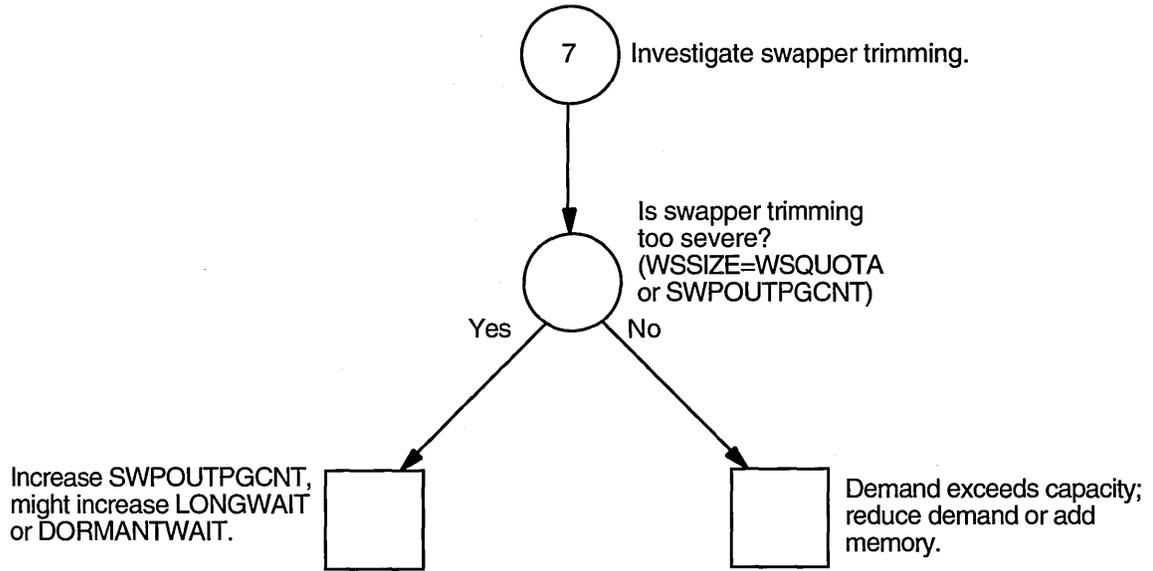
ZK-1135-GE

Figure A-6 Investigating Excessive Paging—Phase IV



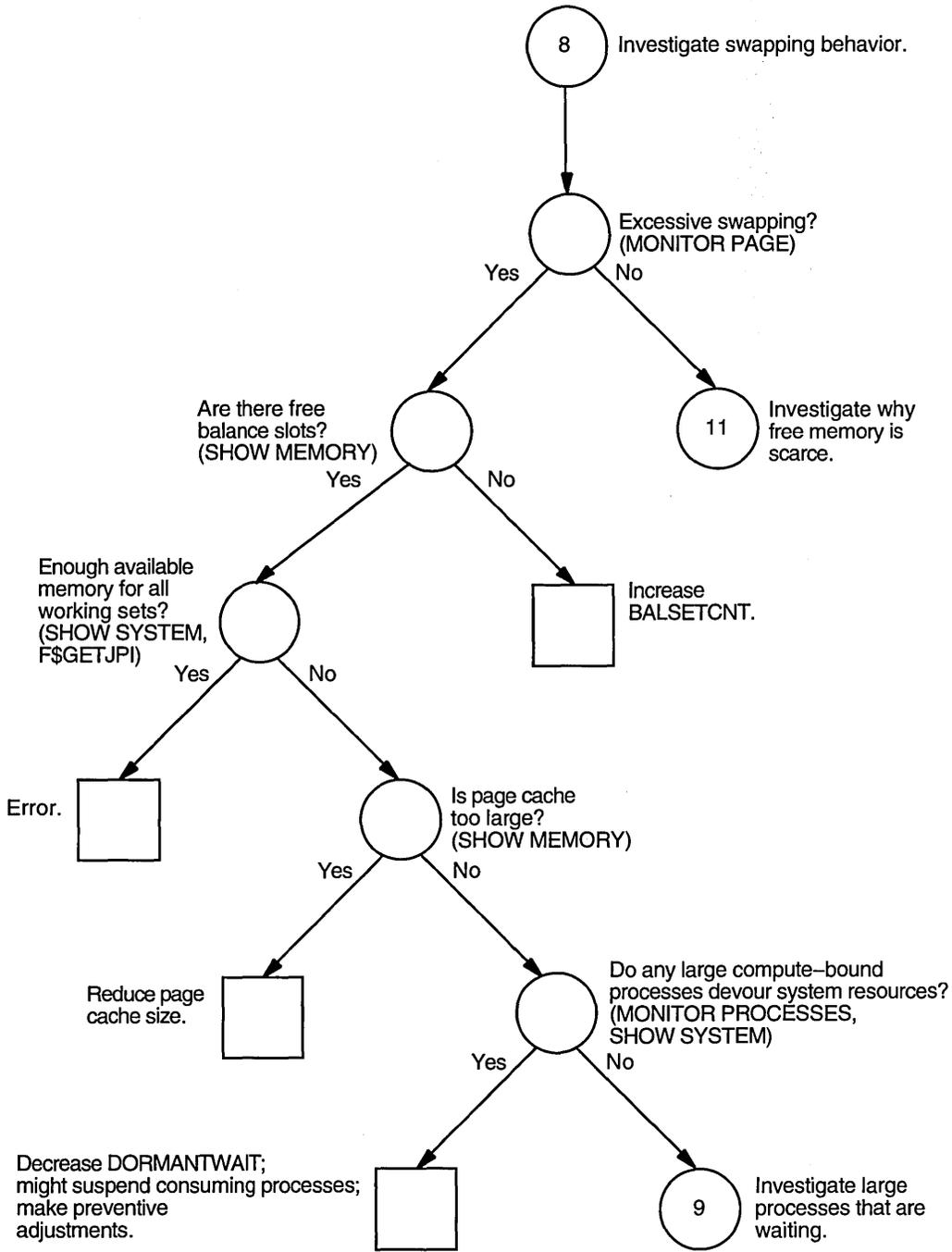
ZK-1136-GE

Figure A-7 Investigating Excessive Paging—Phase V



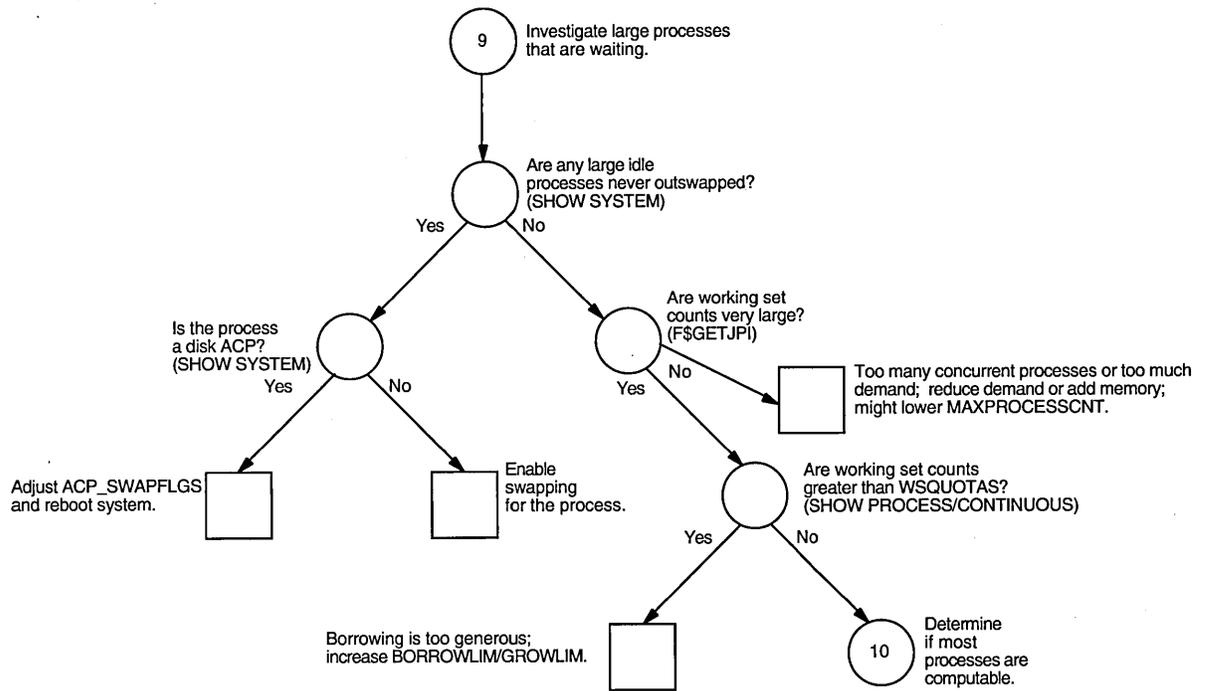
ZK-1137-GE

Figure A-8 Investigating Swapping—Phase I



ZK-1138-GE

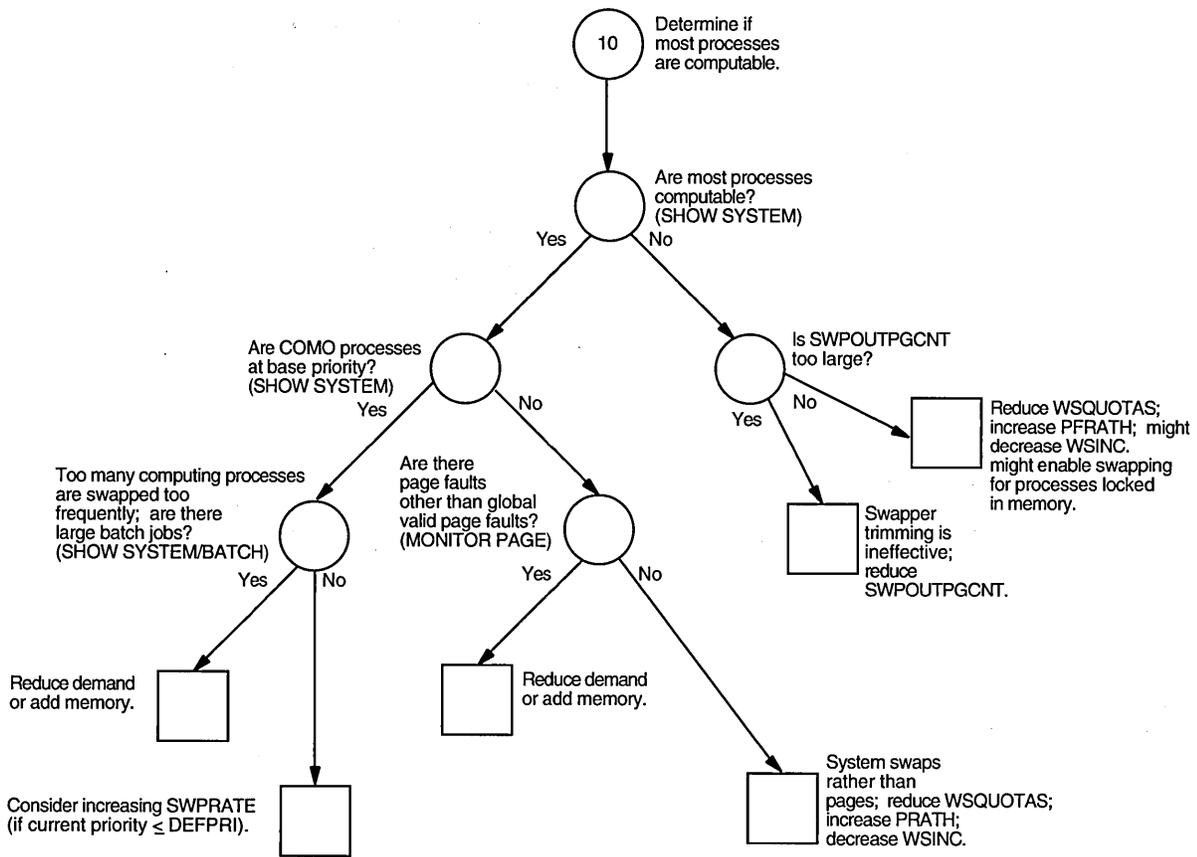
Figure A-9 Investigating Swapping—Phase II



ZK-1143-GE

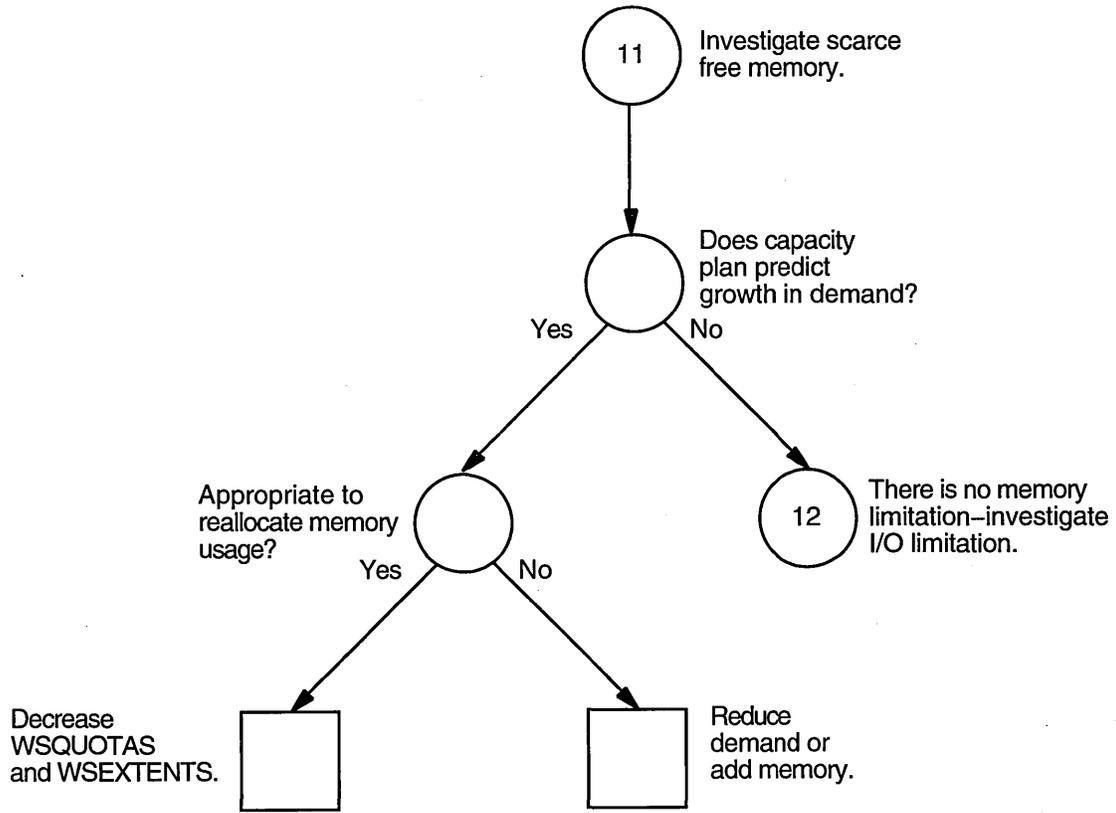
Decision Trees

Figure A-10 Investigating Swapping—Phase III



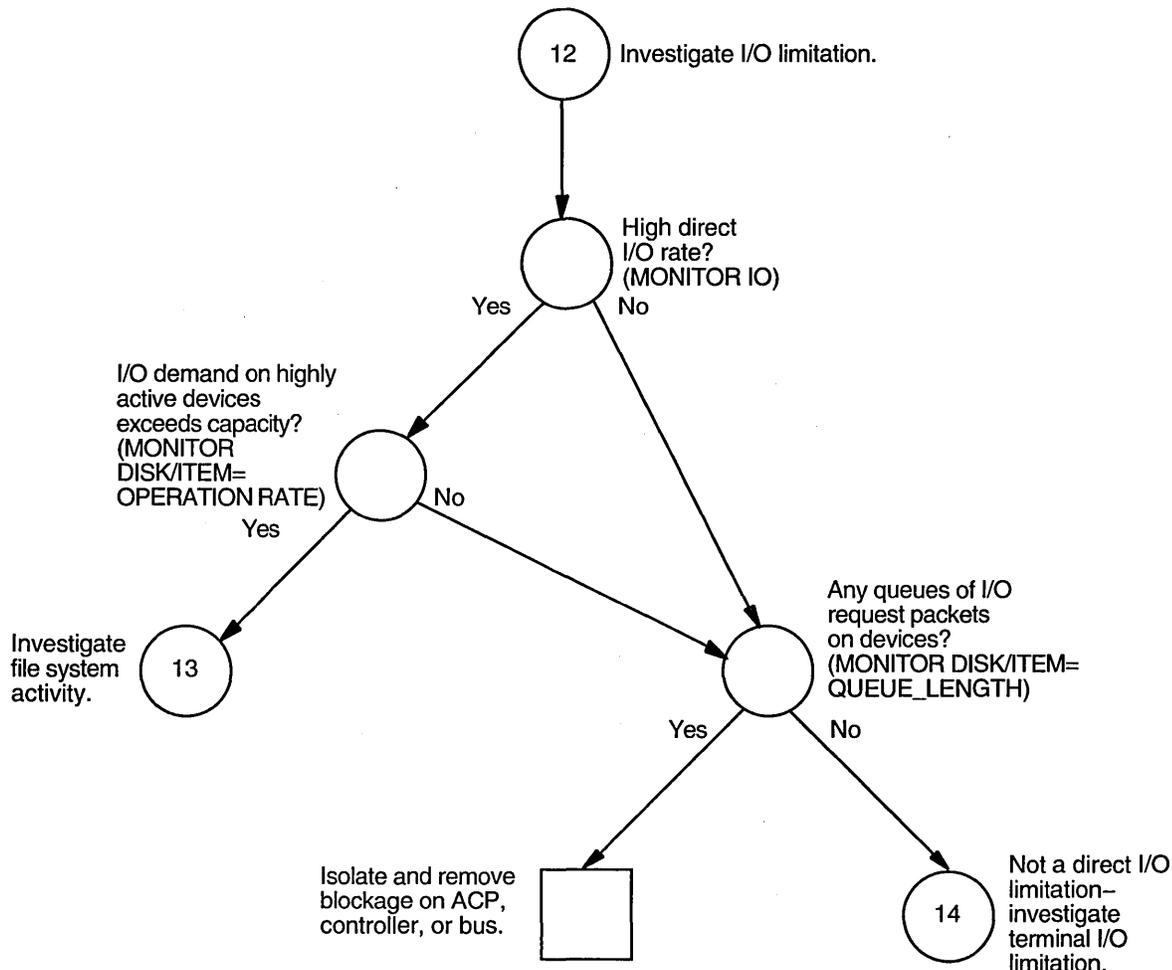
ZK-1144-GE

Figure A-11 Investigating Limited Free Memory—Phase I



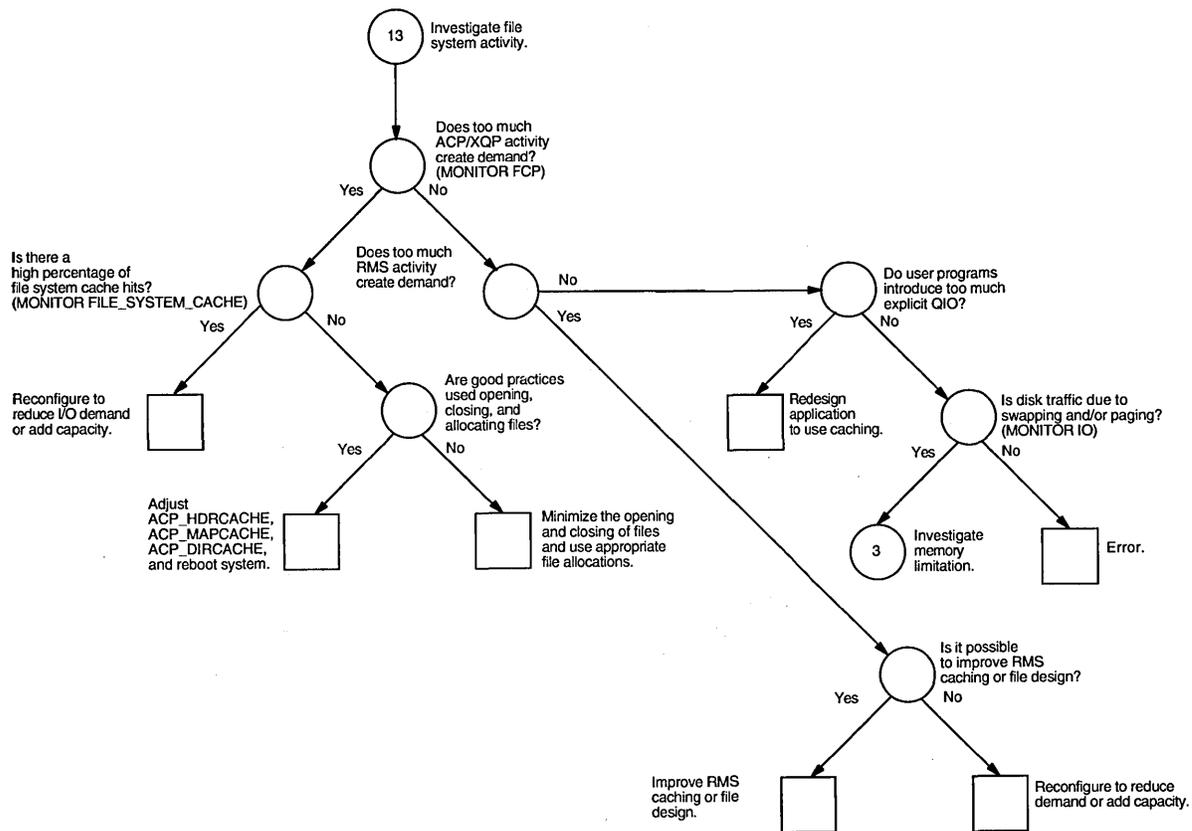
ZK-1142-GE

Figure A-12 Investigating Disk I/O Limitations—Phase I



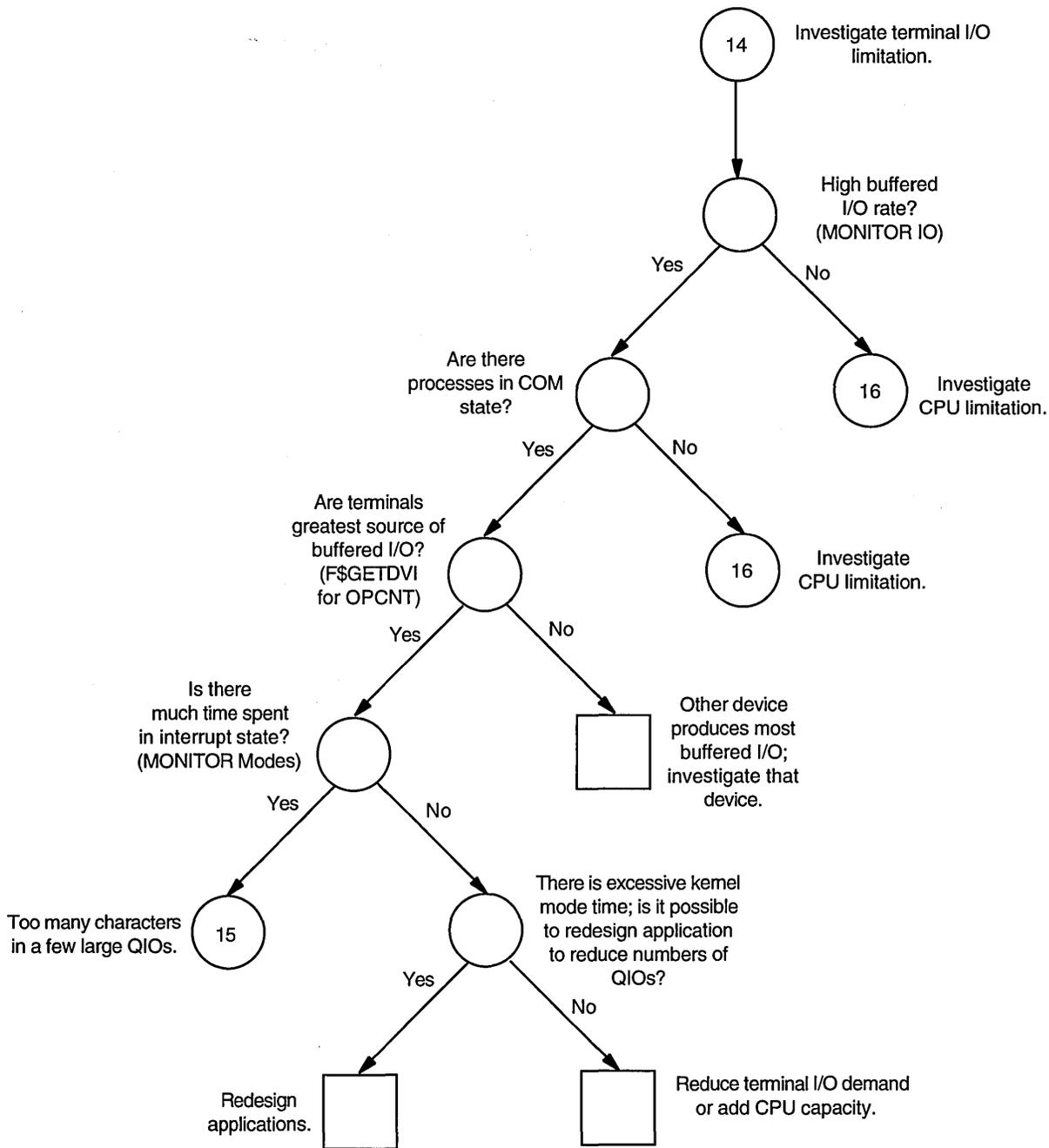
ZK-1145-GE

Figure A-13 Investigating Disk I/O Limitations—Phase II



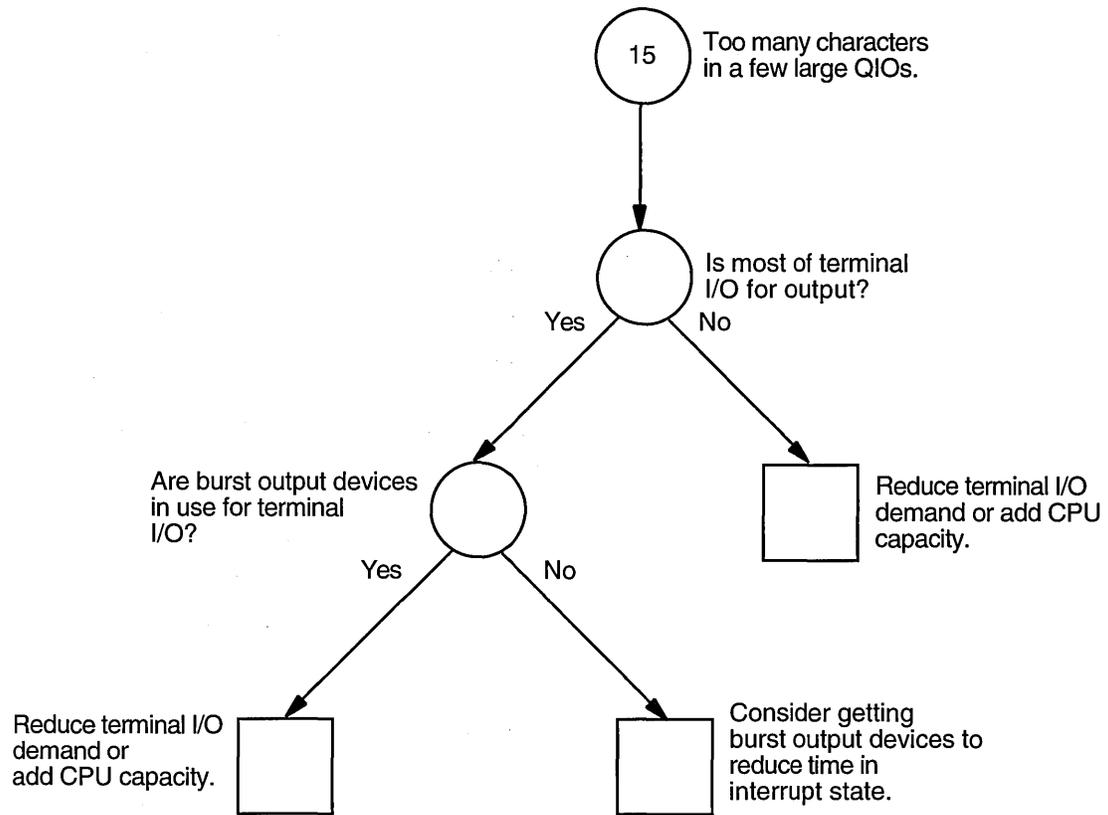
ZK-1150-GE

Figure A-14 Investigating Terminal I/O Limitations—Phase I



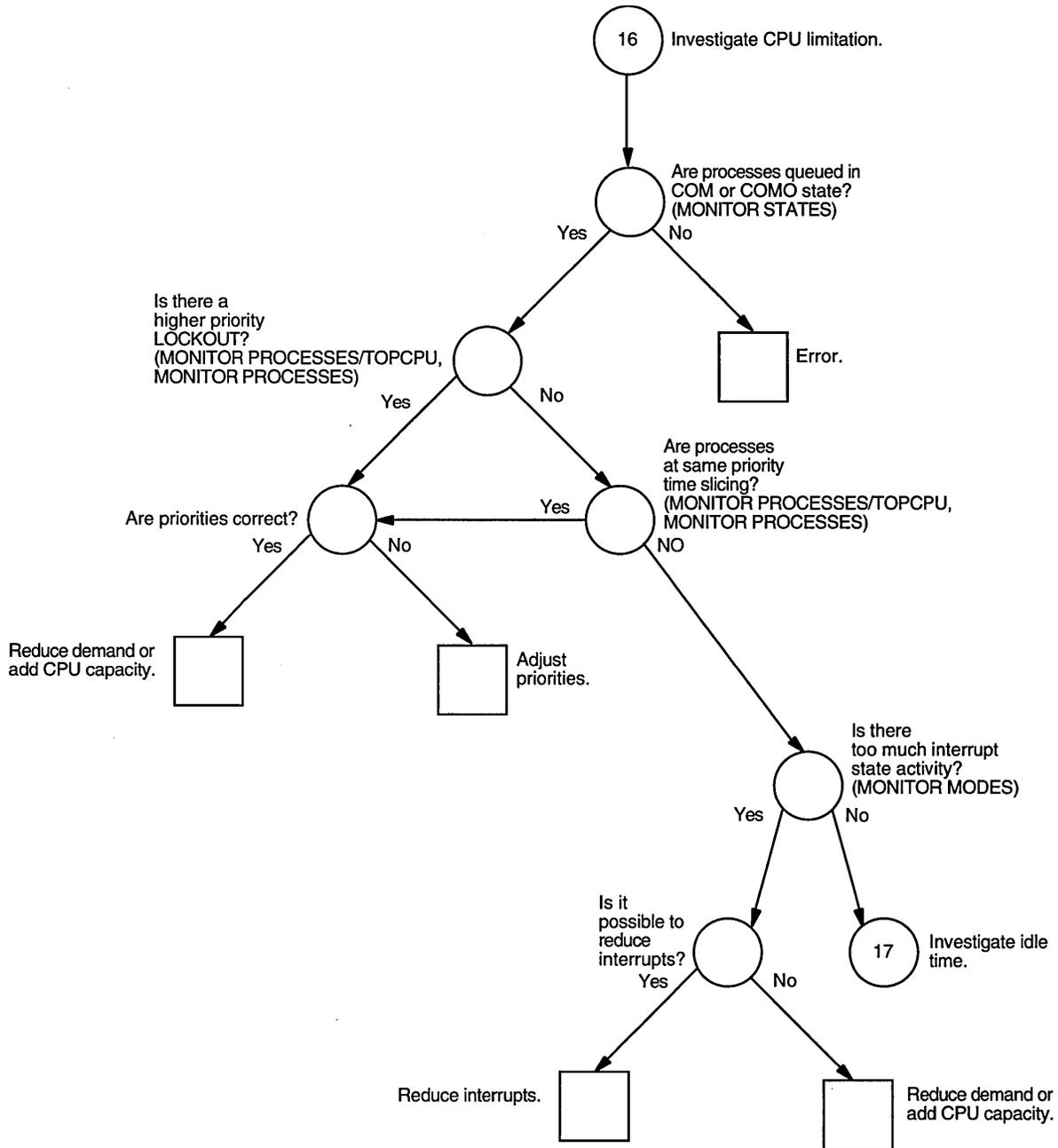
ZK-7044A-GE

Figure A-15 Investigating Terminal I/O Limitations—Phase II



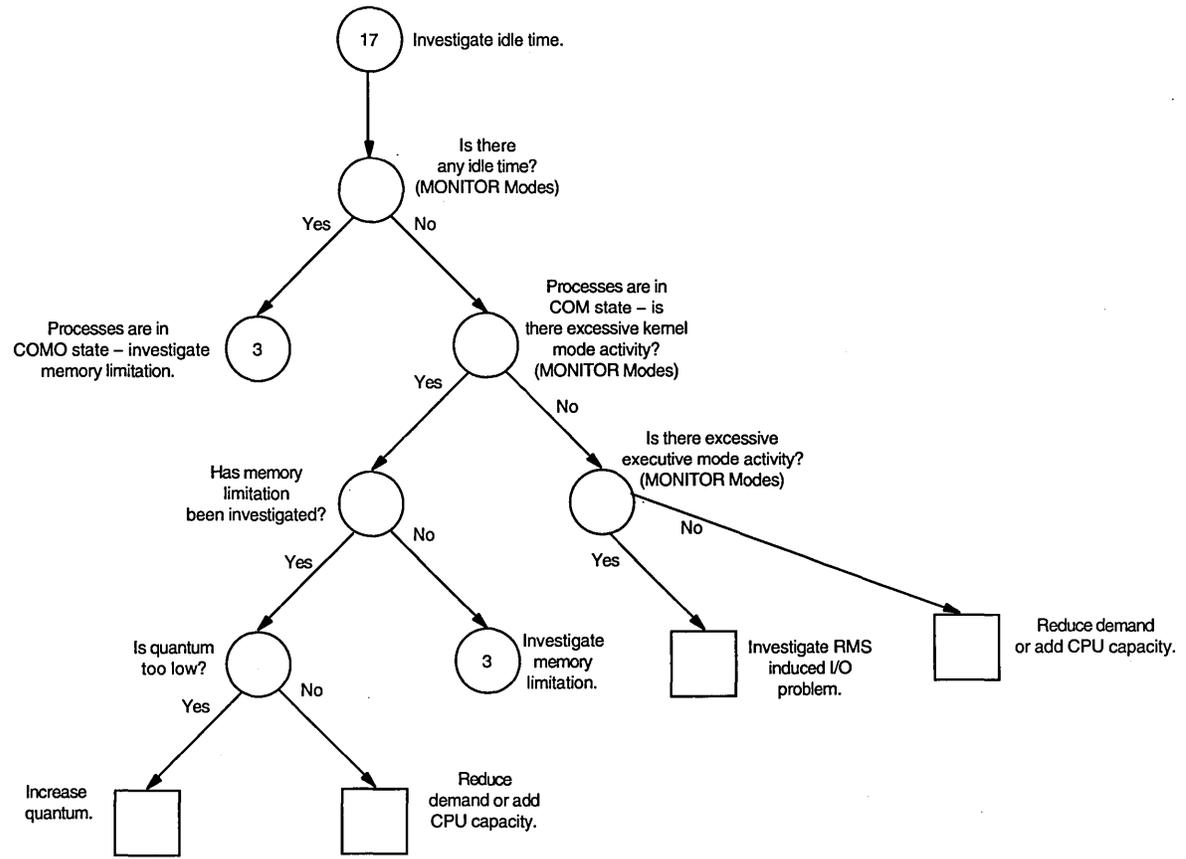
ZK-7043A-GE

Figure A-16 Investigating Specific CPU Limitations—Phase I



ZK-7045A-GE

Figure A-17 Investigating Specific CPU Limitations—Phase II



ZK-1148-GE

B

MONITOR Data Items

Table B-1 provides a quick reference to the MONITOR data items that you will probably need to check most often in evaluating your resources.

Table B-1 Summary of Important MONITOR Data Items

Item	Class	Description ¹
Compute Queue (COM + COMO)	STATES	Good measure of CPU responsiveness in most environments. Typically, the larger the compute queue, the longer the response time.
Idle Time	MODES	Good measure of available CPU cycles, but only when processes are not unduly blocked because of insufficient memory or an overloaded disk I/O subsystem.
Inswap Rate	IO	Rate used to detect memory management problems. Should be as low as possible, no greater than 1 per second.
Interrupt State Time + Kernel Mode Time	MODES	Time representing service performed by the system. Normally, should not exceed 40% in most environments.
MP Synchronization Time	MODES	Time spent by a processor waiting to acquire a spin lock in a multiprocessing system. A value greater than 8% might indicate moderate-to-high levels of paging, I/O, or locking activity.
Executive Mode Time	MODES	Time representing service performed by RMS and some database products. Its value will depend on how much you use these facilities.
Page Fault Rate	PAGE	Overall page fault rate (excluding system faults). Paging might demand further attention when it exceeds 600 faults per second.
Page Read I/O Rate	PAGE	The hard fault rate. Should be kept below 10% of overall rate for efficient use of secondary page cache.

¹The values and ranges of values shown are *averages*. They are intended only as general guidelines and will not be appropriate in all cases.

(continued on next page)

Table B-1 (Cont.) Summary of Important MONITOR Data Items

Item	Class	Description ¹
System Fault Rate	PAGE	Rate should be kept to minimum, no more than 2 faults per second per VUP.
Response Time (ms) (computed)	DISK	Expected value is 25-40 milliseconds for RA-series disks with no contention and small transfers. Individual disks will exceed that value by an amount dependent on the level of contention and the average data transfer size.
I/O Operation Rate	DISK	Overall I/O operation rate. The following are normal load ranges for RA-series disks in a typical timesharing environment, where the vast majority of data transfers are small: 1 to 8—lightly loaded 9 to 15—light to moderate 16 to 25—moderate to heavy More than 25—heavily loaded
Page Read I/O Rate + Page Write I/O Rate + Inswap Rate (times 2) + Disk Read Rate + Disk Write Rate	PAGE PAGE IO FCP FCP	System I/O operation rate. The sum of these items represents the portion of the overall rate initiated directly by the system.
Cache Hit Percentages	FILE_ SYSTEM_ CACHE	XQP cache hit percentages should be kept as high as possible, no lower than 75% for the active caches.

¹The values and ranges of values shown are *averages*. They are intended only as general guidelines and will not be appropriate in all cases.

MONITOR Multifile Summary Report

Figure C-1, a typical VMScluster prime-time multifile summary report, provides an extended context for the data items in Table B-1.

MONITOR Multifile Summary Report

Figure C-1 Prime-Time VMScluster Multifile Summary Report

```

+-----+
| AVE |
+-----+
OpenVMS Monitor Utility
PROCESS STATES
MULTI-FILE SUMMARY

Node: CURLEY (2) LARRY MOE STOOG (3)
From: 15-DEC-1994 12:44 15-DEC-1994 09:01 15-DEC-1994 09:00 15-DEC-1994 09:01
To: 15-DEC-1994 18:09 15-DEC-1994 18:02 15-DEC-1994 18:01 15-DEC-1994 18:03

```

	Row Sum	Row Average	Row Minimum	Row Maximum
Collided Page Wait	0.00	0.00	0.00	0.00
Mutex & Misc Resource Wait	0.00	0.00	0.00	0.02
Common Event Flag Wait	0.00	0.00	0.00	0.00
Page Fault Wait	0.03	0.09	0.13	0.35
Local Event Flag Wait	16.36	21.81	26.01	18.08
Local Evt Flg (Outswapped)	0.00	0.00	0.00	0.00
Hibernate	16.36	17.50	20.33	14.44
Hibernate (Outswapped)	0.00	0.00	0.00	0.00
Suspended	0.00	0.00	0.04	0.00
Suspended (Outswapped)	0.00	0.00	0.00	0.00
Free Page Wait	0.00	0.00	0.00	0.00
Compute	3.43	2.14	1.50	1.15
Compute (Outswapped)	0.00	0.00	0.00	0.00
Current Process	0.96	1.00	1.00	0.95

ZK-7048A-GE

```

+-----+
| AVE |
+-----+
OpenVMS Monitor Utility
TIME IN PROCESSOR MODES
MULTI-FILE SUMMARY

Node: CURLEY (2) LARRY MOE STOOG (3)
From: 15-DEC-1994 12:44 15-DEC-1994 09:01 15-DEC-1994 09:00 15-DEC-1994 09:01
To: 15-DEC-1994 18:09 15-DEC-1994 18:02 15-DEC-1994 18:01 15-DEC-1994 18:03

```

	Row Sum	Row Average	Row Minimum	Row Maximum
Interrupt State	5.21	10.56	5.40	2.02
MP Synchronization	0.00	0.00	0.00	0.00
Kernel Mode	11.52	16.20	12.22	4.92
Executive Mode	2.17	4.53	4.28	1.48
Supervisor Mode	1.06	0.97	4.60	0.70
User Mode	78.51	10.23	7.98	6.47
Compatibility Mode	0.00	0.00	0.00	0.00
Idle Time	1.49	57.47	65.49	84.37

ZK-7054A-GE

```

+-----+
| AVE |
+-----+
OpenVMS Monitor Utility
PAGE MANAGEMENT STATISTICS
MULTI-FILE SUMMARY

Node: CURLEY (2) LARRY MOE STOOG (3)
From: 15-DEC-1994 12:44 15-DEC-1994 09:01 15-DEC-1994 09:00 15-DEC-1994 09:01
To: 15-DEC-1994 18:09 15-DEC-1994 18:02 15-DEC-1994 18:01 15-DEC-1994 18:03

```

	Row Sum	Row Average	Row Minimum	Row Maximum
Page Fault Rate	20.93	32.17	34.56	40.32
Page Read Rate	7.36	16.47	25.02	26.16
Page Read I/O Rate	0.79	1.98	4.07	1.41
Page Write Rate	2.05	5.14	6.25	3.06
Page Write I/O Rate	0.03	0.09	0.21	0.07
Free List Fault Rate	5.03	7.89	6.55	8.80
Modified List Fault Rate	5.42	6.38	4.68	7.24
Demand Zero Fault Rate	4.84	8.00	9.96	14.99
Global Valid Fault Rate	4.76	7.77	9.08	7.70
Wrt In Progress Fault Rate	0.01	0.02	0.02	0.00
System Fault Rate	0.45	2.16	0.15	0.58
Free List Size	2915.60	4888.03	1459.72	106504.46
Modified List Size	178.60	241.53	166.81	345.26

ZK-7055A-GE

(continued on next page)

Figure C-1 (Cont.) Prime-Time VMScluster Multifile Summary Report

OpenVMS Monitor Utility I/O SYSTEM STATISTICS MULTI-FILE SUMMARY												
	CURLEY (2)		LARRY		MOE		STOOG (3)					
	From:	To:	From:	To:	From:	To:	From:	To:				
	15-DEC-1994	Row Sum										
	12:44	18:09	09:01	18:02	09:00	18:01	09:01	18:03	Row Average			
									Row Minimum			
									Row Maximum			
Direct I/O Rate	7.14		9.12		6.68		6.78		29.7	7.4	6.68	9.12
Buffered I/O Rate	6.90		11.74		15.98		16.74		51.3	12.8	6.90	16.74
Mailbox Write Rate	0.24		0.45		0.42		0.41		1.5	0.3	0.24	0.45
Split Transfer Rate	0.30		0.44		0.63		0.56		1.9	0.4	0.30	0.63
Log Name Translation Rate	4.61		6.81		8.06		11.47		30.9	7.7	4.61	11.47
File Open Rate	0.68		0.47		0.47		1.20		2.8	0.7	0.47	1.20
Page Fault Rate	20.93		32.17		34.56		40.32		128.0	32.0	20.93	40.32
Page Read Rate	7.36		16.47		25.02		26.16		75.0	18.7	7.36	26.16
Page Read I/O Rate	0.79		1.98		4.07		1.41		8.2	2.0	0.79	4.07
Page Write Rate	2.05		5.14		6.25		3.06		16.5	4.1	2.05	6.25
Page Write I/O Rate	0.03		0.09		0.21		0.07		0.4	0.1	0.03	0.21
Inswap Rate	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
Free List Size	2914.86		4887.46		1459.75		106504.46		115766.5	28941.6	1459.75	106504.46
Modified List Size	178.63		241.46		166.09		345.26		931.4	232.8	166.09	345.26

ZK-7058A-GE

OpenVMS Monitor Utility FILE PRIMITIVE STATISTICS MULTI-FILE SUMMARY												
	CURLEY (2)		LARRY		MOE		STOOG (3)					
	From:	To:	From:	To:	From:	To:	From:	To:				
	15-DEC-1994	Row Sum										
	12:44	18:09	09:01	18:02	09:00	18:01	09:01	18:03	Row Average			
									Row Minimum			
									Row Maximum			
FCP Call Rate	1.94		1.76		1.85		3.58		9.1	2.2	1.76	3.58
Allocation Rate	0.06		0.13		0.10		0.18		0.4	0.1	0.06	0.18
Create Rate	0.04		0.10		0.10		0.13		0.3	0.0	0.04	0.13
Disk Read Rate	0.82		1.65		1.30		1.22		5.0	1.2	0.82	1.65
Disk Write Rate	0.40		0.63		0.52		0.83		2.3	0.5	0.40	0.83
Volume Lock Wait Rate	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
CPU Tick Rate	2.65		2.42		2.43		1.10		8.6	2.1	1.10	2.65
File Sys Page Fault Rate	0.07		0.07		0.07		0.07		0.3	0.0	0.07	0.07
Window Turn Rate	0.30		0.44		0.63		0.56		1.9	0.4	0.30	0.63
File Lookup Rate	0.48		0.76		0.78		1.47		3.5	0.8	0.48	1.47
File Open Rate	0.68		0.47		0.47		1.20		2.8	0.7	0.47	1.20
Erase Rate	0.01		0.05		0.03		0.09		0.2	0.0	0.01	0.09

ZK-7057A-GE

OpenVMS Monitor Utility LOCK MANAGEMENT STATISTICS MULTI-FILE SUMMARY												
	CURLEY (2)		LARRY		MOE		STOOG (3)					
	From:	To:	From:	To:	From:	To:	From:	To:				
	15-DEC-1994	Row Sum										
	12:44	18:09	09:01	18:02	09:00	18:01	09:01	18:03	Row Average			
									Row Minimum			
									Row Maximum			
New ENQ Rate	6.01		18.04		11.26		13.55		48.8	12.2	6.01	18.04
Converted ENQ Rate	10.01		7.17		6.64		18.65		42.4	10.6	6.64	18.65
DEQ Rate	5.89		17.76		11.04		13.35		48.0	12.0	5.89	17.76
Blocking AST Rate	0.06		0.17		0.09		0.10		0.4	0.1	0.06	0.17
ENQs Forced To Wait Rate	0.08		0.25		0.11		0.13		0.5	0.1	0.08	0.25
ENQs Not Queued Rate	0.03		0.09		0.05		0.02		0.2	0.0	0.02	0.09
Deadlock Search Rate	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
Deadlock Find Rate	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
Total Locks	565.79		1604.27		1098.14		1251.62		4519.8	1129.9	565.79	1604.27
Total Resources	608.53		1015.50		922.68		1074.02		3620.7	905.1	608.53	1074.02

ZK-7058A-GE

(continued on next page)

MONITOR Multifile Summary Report

Figure C-1 (Cont.) Prime-Time VMScluster Multifile Summary Report

OpenVMS Monitor Utility
 DECNET STATISTICS
 MULTI-FILE SUMMARY

Node:		CURLEY (2)		LARRY		MOE		STOOG (3)		Row	Row	Row	Row
From:		15-DEC-1994 12:44		15-DEC-1994 09:01		15-DEC-1994 09:00		15-DEC-1994 09:01		Sum	Average	Minimum	Maximum
To:		15-DEC-1994 18:09		15-DEC-1994 18:02		15-DEC-1994 18:01		15-DEC-1994 18:03					
Arriving Local Packet Rate		1.06		1.82		1.80		1.88		6.5	1.6	1.06	1.88
Departing Local Packet Rate		1.43		1.71		1.66		1.79		6.6	1.6	1.43	1.79
Arriving Trans Packet Rate		0.00		0.33		0.00		0.00		0.3	0.0	0.00	0.33
Trans Congestion Loss Rate		0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
Receiver Buff Failure Rate		0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00

ZK-7059A-GE

OpenVMS Monitor Utility
 FILE SYSTEM CACHING STATISTICS
 MULTI-FILE SUMMARY

Node:		CURLEY (2)		LARRY		MOE		STOOG (3)		Row	Row	Row	Row
From:		15-DEC-1994 12:44		15-DEC-1994 09:01		15-DEC-1994 09:00		15-DEC-1994 09:01		Sum	Average	Minimum	Maximum
To:		15-DEC-1994 18:09		15-DEC-1994 18:02		15-DEC-1994 18:01		15-DEC-1994 18:03					
Dir FCB	(Hit %)	94.50		89.39		90.76		96.95		371.6	92.9	89.39	96.95
	(Attempt Rate)	0.49		0.80		0.83		1.53		3.6	0.9	0.49	1.53
Dir Data	(Hit %)	79.48		60.87		64.44		87.13		291.9	72.9	60.87	87.13
	(Attempt Rate)	1.36		3.12		2.37		3.46		10.3	2.5	1.36	3.46
File Hdr	(Hit %)	70.35		74.12		76.04		75.61		296.1	74.0	70.35	76.04
	(Attempt Rate)	1.85		1.69		1.88		2.82		8.2	2.0	1.69	2.82
File ID	(Hit %)	99.02		97.87		98.46		97.77		393.1	98.2	97.77	99.02
	(Attempt Rate)	0.04		0.11		0.09		0.12		0.3	0.0	0.04	0.12
Extent	(Hit %)	99.12		97.39		95.41		96.89		388.8	97.2	95.41	99.12
	(Attempt Rate)	0.15		0.50		0.39		0.59		1.6	0.4	0.15	0.59
Quota	(Hit %)	98.66		99.62		100.00		99.95		398.2	99.5	98.66	100.00
	(Attempt Rate)	0.03		0.06		0.03		0.16		0.2	0.0	0.03	0.16
Bitmap	(Hit %)	7.31		23.14		28.77		3.84		63.0	15.7	3.84	28.77
	(Attempt Rate)	0.00		0.02		0.04		0.13		0.2	0.0	0.00	0.13

ZK-7060A-GE

OpenVMS Monitor Utility
 DISK I/O STATISTICS
 MULTI-FILE SUMMARY

I/O Operation Rate

Node:		CURLEY (2)		LARRY		MOE		STOOG (3)		Row	Row	Row	Row
From:		15-DEC-1994 12:44		15-DEC-1994 09:01		15-DEC-1994 09:00		15-DEC-1994 09:01		Sum	Average	Minimum	Maximum
To:		15-DEC-1994 18:09		15-DEC-1994 18:02		15-DEC-1994 18:01		15-DEC-1994 18:03					
\$111SDUA2:	TSDPERF	1.54		0.87		0.94		1.15		4.5	1.1	0.87	1.54
\$111SDUA3:	DUMPDISK	0.02		0.01		0.01		0.08		0.1	0.0	0.01	0.08
\$111SDUA4:	PAGESWAPDISK	0.10		0.66		1.53		0.00		2.3	0.5	0.00	1.53
\$111SDUA5:	BPMDISK	0.16		0.05		1.11		0.15		1.4	0.3	0.05	1.11
\$111SDUA6:	QUALD	0.10		2.30		1.03		2.66		6.1	1.5	0.10	2.66
\$111SDUA7:	SQMCLUSTERV4	2.12		3.38		4.90		2.87		13.2	3.3	2.12	4.90
\$111SDUA11:	TIMEDISK	0.25		3.01		0.05		1.14		4.4	1.1	0.05	3.01
\$111SDUA12:	QMISDB	0.11		0.32		0.40		0.13		0.9	0.2	0.11	0.40
\$111SDUA13:	TSDPERF1	0.47		0.06		1.37		1.33		3.2	0.8	0.06	1.37
\$111SDUA18:	TEAMS LIBRARY	0.00		0.08		0.00		0.00		0.0	0.0	0.00	0.08
\$111SDJA8:	ORLEAN	0.02		0.00		0.05		0.00		0.0	0.0	0.00	0.05
MOESDRA5:	USER01	0.00		0.00		0.03		0.00		0.0	0.0	0.00	0.03
MOESDMA1:	UVMSQAR	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
\$111SDJA1:	MPI\$DATA	1.33		0.00		0.00		0.00		1.3	0.3	0.00	1.33
HSC007SDUA0:	SYSTEMDISK	0.00		0.00		0.00		0.03		0.0	0.0	0.00	0.03

ZK-7061A-GE

(continued on next page)

Figure C-1 (Cont.) Prime-Time VMScluster Multifile Summary Report

OpenVMS Monitor Utility
 DISK I/O STATISTICS
 MULTI-FILE SUMMARY

I/O Request Queue Length

Node:		CURLEY (2)		LARRY		MOE		STOOG (3)		Row	Row	Row	Row
From:		15-DEC-1994	12:44	15-DEC-1994	09:01	15-DEC-1994	09:00	15-DEC-1994	09:01	Sum	Average	Minimum	Maximum
To:		15-DEC-1994	18:09	15-DEC-1994	18:02	15-DEC-1994	18:01	15-DEC-1994	18:03				
\$111SDUA2:	TSDPERF	0.06		0.03		0.03		0.05		0.1	0.0	0.03	0.06
\$111SDUA3:	DUMPDISK	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
\$111SDUA4:	PAGESWAPDISK	0.00		0.02		0.05		0.00		0.0	0.0	0.00	0.05
\$111SDUA5:	BPMDISK	0.00		0.00		0.04		0.00		0.0	0.0	0.00	0.04
\$111SDUA6:	QQUALD	0.00		0.07		0.03		0.09		0.2	0.0	0.00	0.09
\$111SDUA7:	SQMCLUSTERV4	0.10		0.15		0.24		0.13		0.6	0.1	0.10	0.24
\$111SDUA11:	TIMEDISK	0.01		0.10		0.00		0.04		0.1	0.0	0.00	0.10
\$111SDUA12:	QMISDB	0.00		0.01		0.01		0.00		0.0	0.0	0.00	0.01
\$111SDUA13:	TSDPERF1	0.03		0.00		0.03		0.04		0.1	0.0	0.00	0.04
\$111SDUA18:	TEAMS LIBRARY	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
\$111SDJAB:	ORLEAN	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
MOESDRAS:	USER01	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
MOESDMA1:	UVMSQAR	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
\$111SDJAL:	MPISDATA	0.03		0.00		0.00		0.00		0.0	0.0	0.00	0.03
HSC007SDUA0:	SYSTEMDISK	0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00

ZK-7049A-GE

OpenVMS Monitor Utility
 DISTRIBUTED LOCK MANAGEMENT STATISTICS
 MULTI-FILE SUMMARY

Node:		CURLEY (2)		LARRY		MOE		STOOG (3)		Row	Row	Row	Row
From:		15-DEC-1994	12:44	15-DEC-1994	09:01	15-DEC-1994	09:00	15-DEC-1994	09:01	Sum	Average	Minimum	Maximum
To:		15-DEC-1994	18:09	15-DEC-1994	18:02	15-DEC-1994	18:01	15-DEC-1994	18:03				
New ENQ Rate	(Local)	2.78		8.03		5.55		5.71		22.0	5.5	2.78	8.03
	(Incoming)	0.33		8.23		2.17		2.02		12.7	3.1	0.33	8.23
	(Outgoing)	2.89		1.76		3.53		5.80		14.0	3.5	1.76	5.80
Converted ENQ Rate	(Local)	2.11		4.12		4.79		4.27		15.3	3.8	2.11	4.79
	(Incoming)	6.71		2.31		0.54		1.35		10.9	2.7	0.54	6.71
	(Outgoing)	1.17		0.72		1.31		13.03		16.2	4.0	0.72	13.03
DEQ Rate	(Local)	2.78		8.00		5.54		5.71		22.0	5.5	2.78	8.00
	(Incoming)	0.27		8.06		2.06		1.95		12.3	3.0	0.27	8.06
	(Outgoing)	2.82		1.69		3.43		5.68		13.6	3.4	1.69	5.68
Blocking AST Rate	(Local)	0.01		0.06		0.03		0.04		0.1	0.0	0.01	0.06
	(Incoming)	0.05		0.02		0.05		0.02		0.1	0.0	0.02	0.05
	(Outgoing)	0.00		0.09		0.01		0.03		0.1	0.0	0.00	0.09
Dir Functn Rate	(Incoming)	3.24		2.21		2.26		1.49		9.2	2.3	1.49	3.24
	(Outgoing)	1.52		2.01		1.98		3.87		9.4	2.3	1.52	3.87
Deadlock Message Rate		0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00

ZK-7050A-GE

OpenVMS Monitor Utility
 SCS STATISTICS
 MULTI-FILE SUMMARY

Kbytes Map Rate

Node:		CURLEY (2)		LARRY		MOE		STOOG (3)		Row	Row	Row	Row
From:		15-DEC-1994	12:44	15-DEC-1994	09:01	15-DEC-1994	09:00	15-DEC-1994	09:01	Sum	Average	Minimum	Maximum
To:		15-DEC-1994	18:09	15-DEC-1994	18:02	15-DEC-1994	18:01	15-DEC-1994	18:03				
CURLEY		0.00		0.00		0.15		0.00		0.1	0.0	0.00	0.15
VANITY		29.50		29.00		28.15		35.89		122.5	30.6	28.15	35.89
MOE		0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
STOOG		0.00		0.00		0.08		0.00		0.0	0.0	0.00	0.08
LARRY		0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
DECEIT		20.85		0.00		0.00		0.18		21.0	5.2	0.00	20.85
HSC003		0.00		0.00		0.00		0.00		0.0	0.0	0.00	0.00
HSC007		0.00		0.00		0.00		2.43		2.4	0.6	0.00	2.43

ZK-7051A-GE

(continued on next page)

MONITOR Multifile Summary Report

Figure C-1 (Cont.) Prime-Time VMScluster Multifile Summary Report

OpenVMS Monitor Utility SYSTEM STATISTICS MULTI-FILE SUMMARY										
Node:		CURLEY (2)		LARRY		MOE		STOOG (3)		
From:	To:	15-DEC-1994 12:44	15-DEC-1994 18:09	15-DEC-1994 09:01	15-DEC-1994 18:02	15-DEC-1994 09:00	15-DEC-1994 18:01	15-DEC-1994 09:01	15-DEC-1994 18:03	
		Row Sum	Row Average	Row Minimum	Row Maximum					
Interrupt State		5.21	10.56	5.40	2.02	23.2	5.8	2.02	10.56	
MP Synchronization		0.00	0.00	0.00	0.00	0.0	0.0	0.00	0.00	
Kernel Mode		11.52	16.20	12.22	4.92	44.8	11.2	4.92	16.20	
Executive Mode		2.17	4.53	4.28	1.48	12.4	3.1	1.48	4.53	
Supervisor Mode		1.06	0.97	4.60	0.70	7.3	1.8	0.70	4.60	
User Mode		78.51	10.23	7.98	6.47	103.2	25.8	6.47	78.51	
Compatibility Mode		0.00	0.00	0.00	0.00	0.0	0.0	0.00	0.00	
Idle Time		1.49	57.47	65.49	84.37	208.8	52.2	1.49	84.37	
Process Count		37.13	42.42	49.20	34.84	163.6	40.9	34.84	49.20	
Page Fault Rate		20.93	32.17	34.56	40.32	128.0	32.0	20.93	40.32	
Page Read I/O Rate		0.79	1.98	4.07	1.41	8.2	2.0	0.79	4.07	
Free List Size		2950.86	4908.18	1489.87	106503.10	115852.0	28963.0	1489.87	106503.10	
Modified List Size		215.00	275.42	180.22	346.08	1016.7	254.1	180.22	346.08	
Direct I/O Rate		7.14	9.12	6.68	6.78	29.7	7.4	6.68	9.12	
Buffered I/O Rate		6.90	11.74	15.98	16.74	51.3	12.8	6.90	16.74	

ZK-7052A-GE

OpenVMS Monitor Utility MSCP SERVER STATISTICS MULTI-FILE SUMMARY										
Node:		CURLEY (2)		LARRY		MOE		STOOG (3)		
From:	To:	15-DEC-1994 13:09	15-DEC-1994 16:09	15-DEC-1994 13:02	15-DEC-1994 16:02	15-DEC-1994 13:06	15-DEC-1994 16:06	15-DEC-1994 13:05	15-DEC-1994 16:05	
		Row Sum	Row Average	Row Minimum	Row Maximum					
Server I/O Request Rate		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Read Request Rate		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Write Request Rate		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Extra Fragment Rate		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Fragmented Request Rate		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Buffer Wait Rate		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Request Size Rates (Blocks)										
1		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
2-3		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
4-7		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
8-15		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
16-31		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
32-63		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
64+		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

ZK-7053A-GE

Index

A

Accounting utility (ACCOUNTING), 7-2
collecting data, 7-4
image-level, 7-3
 disabling, 7-4
 enabling, 7-4
interpreting data, 7-5
report generation, 7-4
ACP (ancillary control process)
 establishing values for, 17-3
 for ODS-1 disks, 17-3
 parameters, 14-4
 ACP_DIRCACHE parameter, 14-4
 ACP_HDRCACHE parameter, 14-4
 ACP_MAPCACHE parameter, 14-4
 removing blockage, 18-4
Adjustment period
 definition of, 6-1
ALTPRI privilege, 15-2
Ancillary control process
 See ACP
Application code sharing, 1-5
Authorize utility (AUTHORIZE), 3-3
 priorities
 adjusting, 19-1
 quotas
 modifying, 17-4
AUTOGEN command procedure, 3-3
 changing system parameters, 16-2
 feedback mode, 3-3, 7-2
 feedback report, 9-5
 swapper trimming, 6-15
Automatic working set adjustment
 See AWSA
AWSA (automatic working set adjustment),
 6-4
 AUTOGEN, 17-7
 enabling, 17-7
 investigating status, 13-7
 page faulting, 6-4
 parameters, 6-4
 adjusting, 6-11
 swapper trimming, 6-15
 tuning, 6-11

AWSA (automatic working set adjustment)
 (cont'd)
 tuning to respond to increased demand,
 17-6
 voluntary decrementing, 6-10, 6-15

B

Backing store
 paging files, 11-8
 section files, 11-8
Backup utility (BACKUP)
 restoring contiguity on fragmented disks,
 11-10
Balance set
 definition of, 5-1
Balance slots, 13-11
BALSETCNT parameter
 adjusting, 13-3, 17-8
 artificially induced swapping, 10-8
 increasing, 17-9
 reducing, 17-8
Batch jobs
 establishing values for, 17-5
Batch processing
 working set limits, 6-3
Batch queues
 creating, 6-23
 jobs
 base priority, 6-23
 working set characteristics, 6-2
Borrowing
 analyzing problems, 13-6
 deciding when too generous, 13-14
 tuning to make more effective, 17-5
Buffered I/O
 definition of, 14-1
 in relation to terminal operation problems,
 14-5
BUGCHECKFATAL parameter, 10-12

C

Caches
 file system
 adjusting, 18-7
 primary page, 5-1
 secondary page, 10-6

- Caches (cont'd)
 - virtual I/O, 18-1
- Collided page (COLPG) wait state, 9-4
- Common event flag (CEF) wait state, 9-4
- Complaints
 - analyzing, 2-1
 - evaluating, 2-2
 - hardware, 2-2
 - log files, 2-2
 - verifying, 2-2
- Compute-bound processes
 - controlling growth, 17-10
 - curtailing, 17-9
 - suspending, 17-9
- Compute queue, 9-1
 - optimal length, 9-2
- Context switching
 - definition of, 6-1
- Convert utility (CONVERT)
 - restoring contiguity on fragmented disks, 11-10
- CPU (central processing unit)
 - adding capacity, 15-4, 19-2
 - determining when capacity is reached, 15-4
 - time spent in supervisor mode, 15-4
- CPU limitations
 - isolating, 15-1
- CPU resource
 - affected by swapping, 10-7
 - assessing relative load, 9-14
 - capacity, 9-3
 - types of, 19-2
 - compute queue, 9-1
 - optimal length, 9-2
 - equitable CPU sharing, 9-7
 - inequitable CPU sharing, 9-7
 - load balancing
 - VMScluster systems, 9-14
 - load balancing in a VMScluster, 9-13
 - offloading, 9-12
 - on a network, 9-13
 - processor mode, 9-8
 - reducing resource consumption, 9-8
 - response time, 9-2

D

- DEC File Optimizer for OpenVMS
 - restoring contiguity on fragmented disks, 11-10
- Demand Zero Fault Rate, 10-5
- Detached processes
 - base priority, 6-23
 - establishing values for, 17-4
 - working set characteristics, 6-2

- Direct I/O, 7-6, 11-6
 - definition of, 14-1
- Disk
 - activity
 - due to paging or swapping, 14-4
 - average response time, 11-3
 - direct access, 11-6
 - fragmentation
 - correcting, 11-10
 - effect on system performance, 11-10
 - MSCP served, 11-6
 - remote access, 11-6
 - thrashing
 - investigating, 13-15
 - transfer
 - components, 11-2
- Disk I/O resource
 - disk capacity and demand, 11-2
 - data transfer capacity, 11-3
 - demand by users and the system, 11-3
 - seek capacity, 11-2
 - equitable sharing, 11-6
 - evaluating responsiveness, 11-3
 - factors limiting performance, 11-3
 - function, 11-1
 - improving responsiveness, 11-6
 - load balancing, 11-11
 - offloading, 11-10
 - RAM disks, 18-4
 - virtual I/O cache, 18-1
 - reducing consumption, 11-7
- Documentation comments, sending to Digital, iii
- Dormant processes, 6-14
- DORMANTWAIT parameter, 17-9

E

- Equitable sharing
 - of CPU resource, 9-7
 - of disk I/O resource, 11-6
 - of memory resource, 10-9
- Error log file, 2-2
- Executive mode, 9-8
 - RMS, 9-12

F

- Faults
 - See Pages
 - hard
 - rate, 11-7
- Feedback mode
 - See AUTOGEN
- Feedback on documentation, sending to Digital, iii

File-extend parameters
 setting, 4-2
File system
 ACP parameters, 18-7
 cache
 ACP/XQP parameters, 11-9
 adjusting, 18-7
 hit rate, 11-9
 miss rate, 11-9
 caching, 14-3
 high-water marking, 4-2
 I/O activity, 11-9
First-level trimming, 6-13
 See also Memory management
FREEGOAL parameter, 13-3
 page faulting, 10-7
 setting, 6-16
FREELIM parameter, 13-3
 page faulting, 10-7
Free List Fault Rate, 10-5
Free page (FPG) wait state, 9-4
Free-page list, 5-2
 evaluating, 10-6
 limited free memory
 analyzing, 13-17

G

GBLPAGES parameter, 6-20
GBLSECTIONS parameter, 6-20
Global page, 6-17
 table, 6-20
 entry, 6-20
Global section
 descriptor, 6-20
 table, 6-20
 entry, 6-20
Global Valid Fault Rate, 10-5
Granularity hint region
 See Pool management

H

Hard fault rate, 11-7
Hardware
 when to enlarge capacity, 18-6
Help libraries
 decompressing, 4-1
High-water marking
 definition of, 4-1
 disabling, 4-2

I

I/O-bound processes, 11-2
I/O limitations
 adding capacity, 14-5
 device I/O rate below capacity, 14-2
 direct I/O rate abnormally high, 14-3
 for disk and tape operations, 14-1
 isolating, 14-1
 reducing demand, 14-5
I/O rates
 determining, 14-2
Idle mode, 9-8
Idle processes, 6-15
Image activations
 reducing, 17-1
Image-level accounting, 7-3
 collecting data, 7-4
 disabling, 7-4
 enabling, 7-4
 interpreting data, 7-5
Images
 definition of, 5-1
 installing, 4-2, 6-20
 known
 installing, 1-5
Install utility (INSTALL), 6-20
Inswapping
 reducing rate, 17-11
Interactive processing
 working set limits, 6-3
Interrupt state, 9-8, 9-9
 excessive activity, 15-2
 VMScluster systems, 9-9
 remote nodes, 9-8

K

Kernel mode, 9-8, 9-10
 excessive time, 14-6
Known images
 installing, 1-5

L

LIBDECOMP.COM command procedure, 4-2
Limited free memory
 analyzing, 13-17
Linkage sections, 6-17
Live mode, 11-7
 See Monitor utility
Load balancing, 8-3
 of CPU resource in a VMScluster, 9-13
 of disk I/O resource, 11-11
 of memory resource, 10-5, 10-9, 10-14

Locality of reference, 10-2
 definition of, 10-1
Log files, 2-2
Long-waiting processes, 6-15

M

Mass Storage Control Protocol
 See MSCP

MAXPROCESSCNT parameter, 1-5

Memory

 adding, 17-12
 physical, 5-2
 secondary storage, 5-2
 virtual, 5-2

Memory availability

 analyzing limits, 13-17
 competition for, 13-13
 recognizing when demand exceeds, 13-17

Memory consumption

 by large compute-bound processes, 13-12
 investigating, 13-11

Memory limitations

 compensating for, 17-1
 disguised, 15-3
 free memory
 analyzing, 13-17
 reducing image activations, 17-1

Memory management

 See also Pool management

 first-level trimming, 6-13
 memory sharing, 6-17
 paging, 5-4
 physical memory, 5-2
 policy
 proactive reclamation, 10-7, 11-8
 reactive reclamation, 10-8
 primary page cache, 5-2
 proactive reclamation, 6-15
 disabling, 6-17
 enabling, 6-17
 first-level trimming, 6-15
 idle processes, 6-15
 long-waiting processes, 6-15
 periodically waking processes, 6-16
 second-level trimming, 6-15
 setting FREEGOAL, 6-16
 sizing paging files, 6-17
 sizing swapping files, 6-17
 secondary page cache, 5-2
 second-level trimming, 6-13
 disabling, 6-14
 swapper trimming, 6-12
 swapping, 5-5, 6-14
 types of, 5-5
 virtual memory, 5-2

Memory resource

 equitable sharing, 10-9
 evaluating responsiveness, 10-5
 function, 10-1
 improving responsiveness, 10-9
 load balancing, 10-14
 offloading, 10-13
 proactive reclamation, 10-7, 10-9
 reducing consumption, 10-10

Memory sharing, 6-17

 global pages, 6-17
 linkage sections, 6-17
 overhead, 6-20
 controlling, 6-20
 verifying, 6-20

Miscellaneous (MWAIT) resource wait state,
 9-4

MMG_CTLFLAGS parameter, 6-17

Modified List Fault Rate, 10-5

Modified-page list, 5-2

 evaluating, 10-6

MONITOR.COM command procedure, 7-8

Monitor utility (MONITOR), 7-2

 data items

 summary, B-1

 direct I/O, 11-6

 live mode, 7-9, 11-7

 modes

 live, 7-8

 playback, 7-8

 MONITOR DECNET data

 kernel mode, 9-10

 MONITOR DISK data

 evaluating MSCP served disk, 11-6
 responsiveness of disk I/O subsystem,
 11-3

 MONITOR DLOCK data

 interrupt state, 9-9

 MONITOR FCP data

 file system I/O activity, 11-9

 MONITOR FILE_SYSTEM_CACHE data,
 11-9

 file system I/O activity, 11-9

 MONITOR IO data

 kernel mode, 9-10
 swapping and swapper trimming,
 10-7

 MONITOR LOCK data

 kernel mode, 9-10

 MONITOR MODES data

 CPU load balancing in a VMScluster,
 9-13

 executive mode, 9-8, 9-12

 idle time, 9-8

 interrupt state, 9-8, 9-9

 kernel mode, 9-8, 9-10

 MP synchronization mode, 9-8, 9-10

Monitor utility (MONITOR)
 MONITOR MODES data (cont'd)
 supervisor mode, 9-8
 user mode, 9-8
 MONITOR PAGE data
 disk I/O consumption, 11-7
 kernel mode, 9-10
 memory consumption, 10-10
 page fault, 10-5
 MONITOR POOL data
 memory consumption, 10-10
 MONITOR SCS data
 interrupt state, 9-9
 MONITOR STATES data
 secondary page cache, 10-6
 swapping and swapper trimming,
 10-7
 multifile reports, 7-7
 output
 types of, 7-8
 playback mode, 9-7, 11-6
 summary reports, 7-7
 MONSUM.COM command procedure, 7-8
 MP synchronization mode, 9-8, 9-10
 MPW_LOLIMIT parameter, 13-3
 MPW_THRESH parameter, 13-3
 MSCP
 definition of, 11-6
 served disks, 11-6
 MSCP protocol, 9-8
 Multiblock count, 4-1
 Multibuffer count, 4-1
 MULTIPROCESSING parameter, 10-12

N

Nonpaged pool
 See pool management
 NPAGEDYN parameter, 10-11
 NPAGEVIR parameter, 10-11

O

Offloading, 8-3
 of CPU resource, 9-12
 of disk I/O resource, 11-10
 of memory resource, 10-13
 Operator log file, 2-2

P

Page caches
 primary, 5-2
 secondary, 5-2
 free-page list, 5-2
 modified-page list, 5-2
 size

Page caches
 size (cont'd)
 adjusting related SYSGEN parameters,
 17-2
 decreasing, 17-2, 17-9
 increasing, 17-2
 Paged dynamic pool
 See Pool management
 Page fault (PFW) wait state, 9-4
 Page faulting
 function of secondary page cache, 11-7
 hard, 11-7
 soft, 11-7
 Pagelets
 definition of, 5-1
 Pages
 definition of, 5-1
 faults, 6-4
 acceptable hard fault rate, 10-6
 acceptable soft fault rate, 10-6
 hard, 5-4, 10-5
 soft, 5-4, 10-5
 system, 13-3
 pagelets, 5-1
 Paging, 5-4
 average transfer size, 11-7
 backing store
 paging files, 11-8
 section files, 11-8
 files, 5-2, 6-17
 adding, 17-11
 I/O
 read, 11-7
 write, 11-8
 rates
 Demand Zero Fault, 10-5
 Free List Fault, 10-5
 Global Valid Fault, 10-5
 Modified List Fault, 10-5
 System Fault, 10-6
 Write in Progress, 10-5
 symptoms
 for disks, 14-4
 Performance
 diagnostic strategy
 overview, 12-1
 information database, 7-8
 management
 definition of, 1-1
 information database, 7-8
 strategy, 1-4
 utilities, 1-3
 work load, 1-3
 Periodically waking processes, 6-16
 watchdog, 6-16

- Physical memory, 5-2
- Playback mode
 - See Monitor utility
- POOLCHECK parameter, 10-12
- Pool management
 - adaptive, 10-11
 - allocator, 10-11
 - consistency checking, 10-12
 - corruption detection
 - disabling, 10-12
 - enabling, 10-12
 - deallocating memory, 10-11
 - granularity, 10-12
 - hint region, 10-11
 - lookaside lists, 10-11
 - nonpaged pool, 10-11
 - paged dynamic, 6-20
 - pool monitoring
 - disabling, 10-12
 - enabling, 10-12
 - system parameters, 10-11
- PQL_DWSDEFAULT parameter, 6-2
- PQL_DWSEXTENT parameter, 6-2
- PQL_DWSQUOTA parameter, 6-2
- Primary page cache, 5-1, 5-2
- Priority
 - base, 6-23
 - boosting, 9-7
- Proactive memory reclamation, 6-15
 - See also Memory management
 - disabling, 6-17
 - enabling, 6-17
 - first-level trimming, 6-15
 - idle processes, 6-15
 - long-waiting processes, 6-15
 - periodically waking processes, 6-16
 - watchdog, 6-16
 - second-level trimming, 6-15
 - setting FREEGOAL, 6-16
 - sizing paging files, 6-17
 - sizing swapping files, 6-17
- Processes
 - adjusting priorities, 19-1
 - base priority, 6-23
 - blocked by higher-priority process, 15-2
 - compute-bound, 9-2, 17-9
 - curtailing, 17-9
 - dormant, 6-14
 - I/O-bound, 11-2
 - idle, 6-15
 - long-waiting, 6-15
 - periodically waking, 6-16
 - priority, 6-22, 15-2
 - boosting, 6-22
 - real-time, 6-22
 - reducing delay waiting for CPU, 19-2
 - states, 6-22

- Processes
 - states (cont'd)
 - collided page (COLPG), 9-4
 - common event flag (CEF), 9-4
 - compute (COM), 9-1
 - compute outswapped (COMO), 9-1
 - free page (FPG), 9-4
 - hibernate (HIB), 6-14
 - local event flag (LEF), 6-14
 - MWAIT, 2-3
 - page fault (PFW), 9-4
 - suspended, 6-14
 - synchronization, 9-4
 - time slicing, 15-2
 - working set limits, 6-2
- Process header
 - system, 6-20
- Processing
 - batch, 6-3
 - interactive, 6-3
- Processor modes
 - executive, 9-8
 - idle, 9-8
 - interrupt, 9-8
 - kernel, 9-8
 - MP synchronization, 9-8
 - supervisor, 9-8
 - user, 9-8
- Process scheduling
 - See Scheduling

Q

- QUANTUM parameter, 6-21
 - increasing, 19-2

R

- RAM disks, 18-4
- Real-time processes, 6-22
- Resident executive, 5-2
- Resource consumption, 8-3
- Resource limitations
 - compensating for, 16-1
 - diagnosing, 12-1
- RJOBLIM parameter, 1-5
- RMS
 - buffers, 7-6, 18-6
 - consumption of executive mode processing
 - time, 9-12
 - improving caching, 18-6
 - misuse, 15-4
 - performance implications of file design,
 - 9-12
- RMS_EXTEND_SIZE parameter, 4-2
- Round-robin scheduling, 6-21, 9-7

RWAST wait state, 9-5
RWMBP wait state, 9-5
RWMPE wait state, 9-5
RWPGF wait state, 9-5
RWSWP wait state, 9-5

S

Scheduler

definition of, 6-1

Scheduling

real-time processes, 6-22

round-robin, 6-21, 9-7

states

compute (COM), 9-1

compute outswapped (COMO), 9-1

tuning, 6-23

wait states, 9-3

types of, 9-3

SCS (system communication services), 9-8

Secondary page cache, 10-5, 10-6

evaluating, 10-6, 11-7

free-page list, 5-2

modified-page list, 5-2

Secondary storage, 5-2

Second-level trimming, 6-13

See also Memory management

Section files, 5-2

Semaphores

definition of, 2-1

MUTEX, 2-1

Shareable images

installing, 4-2, 6-20

Spin locks, 9-10

definition of, 9-1

SUBMON.COM command procedure, 7-8

Subprocesses

base priority, 6-23

establishing values for, 17-4

working set characteristics, 6-2

Supervisor mode, 9-8

Suspended processes, 6-14

Swapper, 5-4, 6-12

definition of, 5-1

trimming, 5-5, 6-12

adjusting, 17-7

alternative to swapping, 10-8

analyzing when ineffective, 13-14

disabling, 6-14

dormant processes, 6-14

first-level, 6-13

investigating, 13-9

memory reclamation, 17-11

second-level, 6-13

suspended processes, 6-14

Swapping, 5-4, 5-5

artificially induced, 10-8

converting to system that rarely swaps,
17-8

effect on CPU resource, 10-7

effect on disk subsystem, 10-7

enabling for disk ACPs, 17-10

files, 5-2, 6-17

I/O activity, 11-8

inducing paging to reduce, 17-11

symptoms

analyzing, 13-9

diagnosing, 13-10

for disks, 14-4

for large waiting process, 13-13

harmful, 13-10

SWPOUTPGCNT parameter

swapping and swapper trimming, 10-8,
10-14

SYSGEN parameters

adjusting page cache size, 17-2

changing, 16-2

SYSTEMWCNT parameter, 6-20

adjusting to curtail page thrashing, 10-10

System

disk, 4-3

fault rate, 10-6

libraries

decompressing, 4-1

parameters

adjusting, 3-3

changing, 16-2

process header, 6-20

resource

definition of, 1-1

System parameters

See also AWSA

SYSTEM_CHECK parameter, 10-12

T

Terminal operations

improper handling, 14-5

in relation to CPU limitation, 14-5

in relation to I/O limitation, 14-5

Throughput rate

definition of, 1-2

Time slicing, 6-21

between processes, 15-2

Translation buffer, 10-11

Tuning

definition of, 3-1

evaluating, 3-4

prerequisites, 3-2

suggestions, 3-2

when to stop, 3-4

U

- User mode, 9-8
- User programs
 - working set limits, 6-3

V

- VCC_FLAGS parameter, 18-3
- VCC_MAXSIZE parameter, 18-3
- Virtual I/O caches, 18-1
 - adjusting size, 18-3
 - caching policies, 18-2
 - disabling, 18-3
 - enabling, 18-3
 - objects, 18-1
 - parameters
 - VCC_FLAGS, 18-3
 - VCC_MAXSIZE, 18-3
 - statistics, 18-2
 - VMScluster configurations, 18-4
- Voluntary decremting, 6-15
 - disabling, 17-7
 - oscillations, 13-8
 - tuning, 17-7
 - turning on, 17-7

W

- Wait state
 - scheduling
 - types of, 9-3
- Wait states
 - HIB, 6-14
 - LEF, 6-14
 - MWAIT, 2-3, 9-4
 - resource
 - RWAST, 9-5
 - RWMBP, 9-5
 - RWMPE, 9-5
 - RWPGF, 9-5
 - RWSWP, 9-5
 - scheduling
 - involuntary, 9-4
 - voluntary, 9-3
- Working set
 - adjusting, 17-3
 - with AUTHORIZE, 6-20
 - analyzing problems, 13-4
 - characteristics, 6-2
 - count
 - definition of, 6-2
 - definition of, 5-1
 - determining when too large, 13-14
 - discouraging loans when memory is scarce, 17-10
 - information

- Working set
 - information (cont'd)
 - displaying, 10-3
 - obtaining, 10-2
 - limits
 - default, 6-4
 - guidelines, 6-3
 - initial, 6-2
 - paging, 5-4
 - regions, 6-4
 - size, 5-4
 - specifying values, 13-5
- Work load
 - distributing, 1-5
 - knowing, 1-3
 - managing, 1-4
- WORKSET.COM command procedure
 - using to obtain working set information, 10-2
- Write in Progress Fault Rate, 10-5
- WSMAX parameter, 17-3

NOTES

NOTES

NOTES

NOTES

NOTES

3

NOTES

How to Order Additional Documentation

Technical Support

If you need help deciding which documentation best meets your needs, call 800-DIGITAL (800-344-4825) and press 2 for technical assistance.

Electronic Orders

If you wish to place an order through your account at the Electronic Store, dial 800-234-1998, using a modem set to 2400- or 9600-baud. You must be using a VT terminal or terminal emulator set at 8 bits, no parity. If you need assistance using the Electronic Store, call 800-DIGITAL (800-344-4825) and ask for an Electronic Store specialist.

Telephone and Direct Mail Orders

From	Call	Write
U.S.A.	DECdirect Phone: 800-DIGITAL (800-344-4825) Fax: (603) 884-5597	Digital Equipment Corporation P.O. Box CS2008 Nashua, NH 03061
Puerto Rico	Phone: (809) 781-0505 Fax: (809) 749-8377	Digital Equipment Caribbean, Inc. 3 Digital Plaza, 1st Street Suite 200 Metro Office Park San Juan, Puerto Rico 00920
Canada	Phone: 800-267-6215 Fax: (613) 592-1946	Digital Equipment of Canada Ltd. 100 Herzberg Road Kanata, Ontario, Canada K2K 2A6 Attn: DECdirect Sales
International	—————	Local Digital subsidiary or approved distributor
Internal Orders ¹ (for software documentation)	DTN: 264-3030 (603) 884-3030 Fax: (603) 884-3960	U.S. Software Supply Business Digital Equipment Corporation 10 Cotton Road Nashua, NH 03063-1260
Internal Orders (for hardware documentation)	DTN: 264-3030 (603) 884-3030 Fax: (603) 884-3960	U.S. Software Supply Business Digital Equipment Corporation 10 Cotton Road Nashua, NH 03063-1260

¹Call to request an Internal Software Order Form (EN-01740-07).

Reader's Comments

Guide to OpenVMS AXP
Performance Management
AA-Q28WA-TE

Your comments and suggestions help us improve the quality of our publications.

Thank you for your assistance.

I rate this manual's:	Excellent	Good	Fair	Poor
Accuracy (product works as manual says)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Completeness (enough information)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Clarity (easy to understand)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Organization (structure of subject matter)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Figures (useful)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Examples (useful)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Index (ability to find topic)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Page layout (easy to find information)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

I would like to see more/less _____

What I like best about this manual is _____

What I like least about this manual is _____

I found the following errors in this manual:

Page	Description
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

Additional comments or suggestions to improve this manual:

For software manuals, please indicate which version of the software you are using: _____

Name/Title _____ Dept. _____

Company _____ Date _____

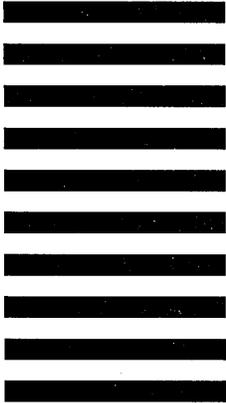
Mailing Address _____

Phone _____

--- Do Not Tear - Fold Here and Tape ---



No Postage
Necessary
if Mailed
in the
United States



BUSINESS REPLY MAIL
FIRST CLASS PERMIT NO. 33 MAYNARD MASS.

POSTAGE WILL BE PAID BY ADDRESSEE

DIGITAL EQUIPMENT CORPORATION
OpenVMS Documentation
110 SPIT BROOK ROAD ZKO3-4/U08
NASHUA, NH 03062-2642



--- Do Not Tear - Fold Here ---