z/OS

Resource Measurement Facility Performance Management Guide

SC33-7992-03

z/OS

Resource Measurement Facility Performance Management Guide

SC33-7992-03

Note
Before using this information and the product it supports, be sure to read the general information under "Notices" on page 219.

Fourth Edition, September 2004

This edition applies to Version 1 Release 6 of z/OS (5694-A01), z/OS.e Version 1 Release 6 (5655-G52), and to all subsequent releases and modifications until otherwise indicated in new editions.

Order publications through your IBM® representative or the IBM branch office serving your locality. Publications are not stocked at the address below.

IBM welcomes your comments. A form for readers' comments may be provided at the back of this publication, or you may address your comments to the following address:

```
IBM DEUTSCHLAND ENTWICKLUNG GMBH
eSERVER PERFORMANCE MANAGEMENT DEVELOPMENT
SCHOENAICHER STRASSE 220
71032 BOEBLINGEN
GERMANY
```

If you prefer to send comments electronically, use one of the following methods:
FAX (RMF™ Development): Your International Access Code +49+7031+16+4240
Internet: rmf@de.ibm.com

Internet
Visit our homepage at http://www.ibm.com/servers/eserver/zseries/zos/rmf/

If you would like a reply, be sure to include your name, address, telephone number, or FAX number.

Make sure to include the following in your comment or note:
- Title and order number of this book
- Page number or topic related to your comment

When you send information to IBM, you grant IBM a nonexclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

Contents

Figures

About this document

The z/OS Resource Measurement Facility (RMF) is a performance management tool that is designed to measure selected areas of system activity and present the data collected in the form of *system management facility (SMF)* records, formatted printed reports, or formatted display reports. Use RMF to report on resource usage, evaluate system performance and identify reasons for performance problems.

This document offers a practical, task-oriented approach to analyzing performance issues using RMF. Covering the most frequently asked questions, it offers guidelines and suggestions concerning performance management of a z/OS sysplex.

Who should use this document

Anyone involved in understanding, measuring, and improving the performance of z/OS systems. This document assumes that the reader has a basic understanding of RMF and how z/OS systems work. The detailed description of all RMF functions and reports is available in *z/OS RMF User's Guide* and *z/OS RMF Report Analysis*.

When to use this document

Use this document when you are trying to improve the performance of your z/OS sysplex. Reasons could include:

- Complaints are coming in from users
- Service level objectives are not being met
- Ongoing performance monitoring showing adverse trends
- Resource usage is changing
- Furthering your own understanding

Furthermore, this document gives hints how to use RMF measurements for preparing the migration from compatibility mode to goal mode.

Do not use this document when your efforts have taken you beyond the scope of RMF and MVS™ performance. For example:

- Network tuning
- Detailed internal subsystem tuning for DB2®

How this document is organized

This document contains the following chapters:

Chapter 1, "Performance Overview"
The chapter describes performance concepts common to all z/OS systems.

Chapter 2, "Diagnosing a Problem: The First Steps"
This chapter helps you analyze a performance problem to determine which resource is constrained, and thus where you should proceed with further analysis.

Chapter 3, "Analyzing Processor Activity"
Here, you find information for further analysis of CPU constraints.

Chapter 4, "Analyzing I/O Activity"
This chapter concentrates on the analysis of I/O constraints (Cache subsystem, DASD, and tape).

Chapter 5, "Analyzing Processor Storage Activity"
The analysis of processor storage constraints (central storage and paging subsystem) will be explained in this chapter.

Chapter 6, "Analyzing Sysplex Activity"
This chapter covers key indicators for workload management and coupling facility performance aspects.

Appendix A, "PR/SM LPAR Considerations"
Here, you find performance considerations for systems running in LPAR mode.

Appendix B, "The Intelligent Resource Director"
This appendix describes the functions of the Intelligence Resource Director and its reporting in RMF.

Appendix C, "Data-In-Memory"
This appendix concentrates on Data-in-Memory techniques.

Appendix D, "Other Delays"
The last appendix discusses some delays shown in Monitor III that have not been covered before.

You should not need to read this document cover-to-cover to find it useful. Feel free to turn directly to the topic of immediate interest to you.

The z/OS RMF library

This table shows the full titles and order numbers of the books in the RMF library for z/OS.

Table 1. RMF Library

Title	Order Number
Books available as Hardcopy and Softcopy	
z/OS *z/OS RMF User's Guide*	SC33-7990
z/OS *z/OS RMF Report Analysis*	SC33-7991
z/OS *z/OS RMF Performance Management Guide*	SC33-7992
z/OS *z/OS RMF Messages and Codes*	SC33-7993
z/OS *z/OS RMF Programmer's Guide*	SC33-7994
z/OS *z/OS RMF Reference Summary*	SX33-9033
Softcopy documentation as part of the *z/OS Collection* (SK3T-4269 (CD-Rom) and SK3T-4271 (DVD))	
z/OS *z/OS RMF NewsFLASH*	SC33-7995

Related information

For additional information on z/OS, see the *z/OS Information Roadmap*, SA22-7500.

Using LookAt to look up message explanations

LookAt is an online facility that allows you to look up explanations for z/OS messages and system abends.

Using LookAt to find information is faster than a conventional search because LookAt goes directly to the explanation.

LookAt can be accessed from the Internet or from a TSO command line.

You can use LookAt on the Internet at:

`www.ibm.com/servers/eserver/zseries/zos/bkserv/lookat/lookat.html`

To use LookAt as a TSO command, LookAt must be installed on your host system. You can obtain the LookAt code for TSO from the LookAt Web site by clicking on the **News and Help** link or from the *z/OS Collection*, SK3T-4269.

To find a message explanation from a TSO command line, simply enter: **lookat** *message-id* as in the following:

```
lookat erb100i
```

This results in direct access to the message explanation for message ERB100I.

To find a message explanation from the LookAt Web site, simply enter the message ID and select the release you are working with.

Summary of Changes

What's new in z/OS V1.6

Summary of Changes
for SC33-7992-04
z/OS Version 1 Release 6

This document contains information previously presented in *z/OS RMF Performance Management Guide*, SC33-7992-02, which supports the Resource Measurement Facility.

This document includes terminology, maintenance, and editorial changes. The following information describes the enhancements that are being distributed with z/OS Version 1 Release 6. All technical changes or additions to the text are indicated by a vertical line to the left of the change.

Support of the @server zSeries® Application Assist Processor

z/OS V1.6 provides the ability to run Java™ applications on a new type of processor called the @server **zSeries Application Assist Processor (zAAP)**. The zSeries Application Assist Processor is also known as an IFA (Integrated Facility for Applications). RMF uses the term IFA in the affected reports, messages, and in this and other publications.

RMF provides measurements of IFA processor activity in the following reports:

- Postprocessor:
  - CPU Activity report and its Partition Data section
  - Workload Activity report
- Monitor III:
  - CPC Capacity report
  - Enclave report
  - System information (SYSINFO) report

This new functionality is available as SPE and needs to be installed as APAR OA05731.

History of Changes

What's New in z/OS Version 1 Release 4

Note:
An appendix with z/OS product accessibility information has been added.

State Samples Breakdown in the WLMGL Report

Up to this release, state samples have been reported as a percentage of average transaction response time (response time breakdown). The response time is calculated when a transaction completes. This can result in percentages greater

than 100 when samples are included for long running transactions which have not completed in the gathering interval (see also "Understanding Workload Activity Data for IMS or CICS" on page 148).

Percentages greater than 100 in the breakdown section are now avoided by showing the state values as percentages of the total transaction samples (state samples breakdown) instead of percentages of response time.

This functionality is available as SPE and needs to be installed as APAR OW52227.

What's New in z/OS Version 1 Release 2

Online Reporting of the License Manager

A new Monitor III **CPC Capacity** report is available which provides information about defined and consumed processor capacity of all partitions in a CPC (central processor complex) which special emphasis on License Manager aspects.

This report is available as SPE and needs to be installed as APAR OW49807 (available in September/October time frame).

Reporting of Report Class Periods

Statistics on response times is a key performance metric used for groups of work having response time objectives. WLM report classes allow aggregation of performance data so that you can evaluate performance of applications independent of the definition how they are managed. Now, reporting of report classes is available with the same granularity as for service classes.

This can be seen in the Postprocessor **WLMGL** report, and in several Monitor III reports: **GROUP**, **STORS**, **SYSINFO**, **SYSRTD**, **SYSSUM**, and **SYSWKM**.

Enhanced Reporting for the Coupling Facilities

CF duplexing ensures high application availability in a Parallel Sysplex®. The performance management aspects which have to be covered for CF duplexing are provided by RMF in the Postprocessor **CF Activity** report about new peer CF connectivity. This enables you to evaluate and monitor your CF configuration, and you can apply the necessary changes to tune accommodation of new structure instances resulting from system-managed duplexing.

What's New in z/OS Version 1 Release 1

The following information describes the enhancements that are being distributed with z/OS Version 1 Release 1. It is indicated by a vertical line to the left of the text.

There is one release RMF 2.10 that can run in OS/390® 2.10 as well as in z/OS V1R1. Nevertheless, some new functions are available only when z/OS is running on a zSeries 900 server.

Intelligent Resource Director (IRD)

The Workload Manager is extended to work with PR/SM™ on zSeries 900 (z900) servers to dynamically expand resources that are available across LPARs.

An **LPAR cluster** is defined as the set of logical partitions in a single CEC that belong to the same parallel sysplex. Based on business goals, WLM can direct PR/SM to enable or disable CP capacity for an LPAR, without human intervention. This combination of WLM working with PR/SM on a z900 server is called IRD.

IRD is made up of three parts which work together to help increase your business' productivity:

- **LPAR CPU Management**

  Based on workload resource demand, the Workload Manager is able to dynamically adjust the number of logical processors and the weight of a logical partition. This allows the system to distribute the CPU resource in an LPAR cluster to partitions where the CPU demand is high.

  The dynamic adjustment of processor resources within the partitions is reflected in the Postprocessor **CPU Activity (Partition Data)** report which provides LPAR views as well as aggregated views on LPAR cluster level.

- **Dynamic Channel Path Management**

  Dynamic Channel Path Management provides the capability to dynamically assign channels to control units in order to respond to peaks in demand for I/O channel bandwidth. This is possible by allowing you to define pools of so-called managed channels that are not related to a specific control unit. With the help of the Workload Manager, channels can float between control units to best service the work according to their goals and their importance.

  All channel and I/O queuing reports have been extended to differentiate static channels from floating channels.

- Channel Subsystem Priority Queuing

  This topic is not reflected directly in any RMF report.

This powerful triad of functions can increase your productivity by putting your most business-critical work first.

IBM License Manager

The IBM License Manager is the base for a new software pricing model. It allows vendors to enable their products for licensed software management by customers and is the basic tool IBM will use to implement the **Workload License Charges** pricing model on z900 servers.

The Postprocessor **CPU Activity (Partition Data)** report will show CPU resource consumption within an LPAR in terms of millions of service units (MSUs) and the corresponding LPAR MSU defined capacity. This helps you to understand how much of the defined capacity an LPAR is consuming.

FICON® Director

RMF offers new reporting capabilities for the FICON director. Due to the different technology and implementation compared to ESCON®, the new Postprocessor **FICON Director Activity** report will provide information about director and port activities. This will assist you in analyzing performance problems and in capacity planning.

Goal Mode Enhancements

The Workload Manager has implemented some enhancements to help increase production use of goal mode. They are

- Classification by System Name
- Separation of Production from Test CICS/IMS Regions
- Address Space Storage Isolation
- Pro-active CPU Protection

This is done by new specifications (CPU protection and storage protection) in the service policy.

RMF supports these enhancements in the **WLMGL** Postprocessor report and in several online reports:

- Monitor II ARD/ARDJ report
- Monitor III SYSWKM report
- Monitor III DELAY report
- Monitor III JOB report
- Monitor III STORF report
- Monitor III ENCLAVE report - Details

Chapter 1. Performance Overview

Let's Start with an Overview

This chapter contains an overview of key performance topics. It explains the concepts and terms we will be using throughout the book:

- Performance Management Definition
- Service Level Agreements
- Capacity Planning Concepts
- Using RMF
- Components of Transaction Response Time
- Analyzing Workload Characteristics
  - Processor
  - Processor Storage
  - I/O Activity
- Processor Speed Indicators

Defining Performance Management

Performance management means monitoring and allocating data processing resources to applications according to *Service Level Agreements* (SLA) or informal objectives. It involves an ongoing cycle of measuring, planning, and modifying.

The goal of performance management is to make the best use of your current resources, to meet your current objectives without excessive tuning effort.

Setting Performance Goals

The human view of the performance of a system is often subjective, emotional and difficult to manage to. However, meeting the business needs of the users is the reason the system exists.

To match business needs with subjective perception, the concept of SLA was introduced.

The SLA is a contract that objectively describes such measurables as:

- Average transaction response time for network, I/O, CPU, or total
- The distribution of these response times (for example, 90% TSO trivial at less than 0.2 of a second)
- Transaction volumes
- System availability

A *transaction* is a business unit of work and can be a CICS® end user interaction, a batch job, etc. Ideally, the definition of a transaction is from a user's point of view.

Planning System Capacity

We can define capacity planning as:

A process of planning for sufficient computer capacity in a cost-effective manner to meet the service needs for all users.

Capacity planning involves asking the following questions:

- How much of your computer resources are being used?
  - CPU
  - Processor storage
  - I/O
  - Network
- Which workloads are consuming the resources (workload distribution)?
- What are the expected growth rates?
- When will the demands on current resources impact service levels?

Benefits of Capacity Planning

An effective capacity planning process provides:

- A mapping of business objectives (user requirements) into quantifiable *information technology (IT)* resources.
- Management oriented reporting of service, resource usage and cost. In an objective way this makes clear what is involved in providing users with good performance.
- Input for making business decisions which involve IT.
- A way to avoid surprises.

Common Mistakes in Capacity Planning

Things to avoid include:

- **Forgetting the concept of balanced systems.** Remember that it takes a combination of resources to process work: CPU, processor storage, and I/O. Increasing the capacity of one resource by itself, may not allow more work to be processed.
- **Ignoring latent demand**. Latent demand is work that is 'hidden' and waiting to be unleashed. One way to begin quantifying latent demand is to note your peak-to-average ratio (e.g. peak hour CPU busy compared to prime-shift average CPU busy). This ratio will drop as latent demand builds up so track it and compare it to your current system (see *Balanced Systems and Capacity Planning* for more details). Relieving *any* bottleneck (for example, upgrading constrained CPU, increasing network bandwidth, or improving I/O subsystem) will release latent demand.
- **Being out of touch with business forecasts**. This one is obvious, but can be difficult to avoid. The capacity planner needs to stay informed of business plans that will impact IT resource requirements (for example, departmental growth, new applications, and mergers).
- **Ignoring complexity**. Factors other than transaction growth will increase demand on IT resources. Enhancements to existing applications, new government regulations, growing databases, etc. - all can result in an increase in resource consumption. This means that "the same" transaction running today may require 10-15% more resource than it did last year.

- **Striving for total precision**. Many variables will impact your capacity plan (growth projections, resource capacity estimates, measurement precision, politics). A plan which achieves an accuracy of ± 10% is considered good. Your time is probably better spent on reasonability checks on your input than on trying to achieve total accuracy.

Performance Management versus Capacity Planning

These two disciplines have much in common. Performance management concentrates on allocating existing resources to meet service objectives. Capacity planning is a means of predicting resources needed to meet future service objectives.

Similar approaches can be used for both; here are some examples:

- Rules of thumb
- Comparison with other systems
- Parametric methods
  - Transaction profile (10 read calls, 2 update calls, 8 physical I/Os)
  - Cost of function (CICS: 15ms per physical I/O in a 3390 non-cached)
- Analytic (queuing theory) models: For example the IBM capacity planning tool CP90.
- Simulation, using a computer program that has the essential logic of the system: for example the Snap/Shot modelling system from IBM.
- Benchmarks, *Teleprocessing Network Simulator (TPNS)*.

In addition, performance and capacity work tend to require similar input data from the system. In particular, both require resource consumption by workload data (see “Analyzing Workload Characteristics” on page 9).

Using RMF

RMF issues reports about performance problems as they occur, so that your installation can take action before the problems become critical.

Your installation can use RMF to:

- Determine that your system is running smoothly
- Detect system bottlenecks caused by contention for resources
- Evaluate the service your installation provides to different groups of users
- Identify the workload delayed and the reason for the delay
- Monitor system failures, system stalls, and failures of selected applications

RMF comes with three monitors, Monitor I, II and III. Monitor III with its ability to determine the 'cause of delay' is where we start:

Monitor III provides short term data collection and online reports for continuous monitoring of system status and solving performance problems.

Monitor III is a good place to begin system tuning. It allows the system tuner to distinguish between delays for important jobs and delays for jobs that are not as important to overall system performance.

Monitor I provides long term data collection for system workload and resource utilization. The Monitor I session is continuous, and measures various areas of system activity over a long period of time.

You can get Monitor I reports directly as real-time reports for each completed interval (single-system reports only), or you let run the **Postprocessor** to create the reports, either as single-system or as sysplex reports. Many installations produce daily reports of RMF data for ongoing performance management. In this publication, sometimes a report is called a Monitor I report (for example, the Workload Activity report) although it can be created only by the Postprocessor.

Monitor II provides online measurements on demand for use in solving immediate problems.

A Monitor II session can be regarded as a snapshot session. Unlike the continuous Monitor I session, a Monitor II session generates a requested report from a single data sample. Since Monitor II is an ISPF application, you might use Monitor II and Monitor III simultaneously in split-screen mode to get different views of the performance of your system.

In addition, you can use the **Spreadsheet Reporter** for further processing the measurement data on a workstation by help of spreadsheet applications. The following chapters provide sample reports including the name of the corresponding macro.

You find a detailed description on how to create the reports and records and on how to use the macros in the *z/OS RMF User's Guide*.

There is another function in RMF to exploit the workstation for monitoring and analyzing your system, called **RMF Performance Monitoring (RMF PM)**, formerly known as PM of OS/390. You find a detailed description on how to use this function in the *z/OS RMF User's Guide*.

This book will discuss key RMF indicators from the different monitors that can be used in a daily report and in problem diagnosis.

What Sampling Cycle and Reporting Interval Should You Use?

For Monitor III: Use the default sampling cycle of 1 second, with the default of 100 seconds for the reporting interval. Adjust this interval if needed to match a problem occurrence that you are investigating.

For Monitor I: Again use the default sampling cycle of 1 second. For the reporting interval, the default of 30 minutes is fine to start with. Adjust this if you prefer, or if you need to home in on a problem.

Analyzing Transaction Response Time

To characterize the performance of a transaction you need to understand its different response time components.

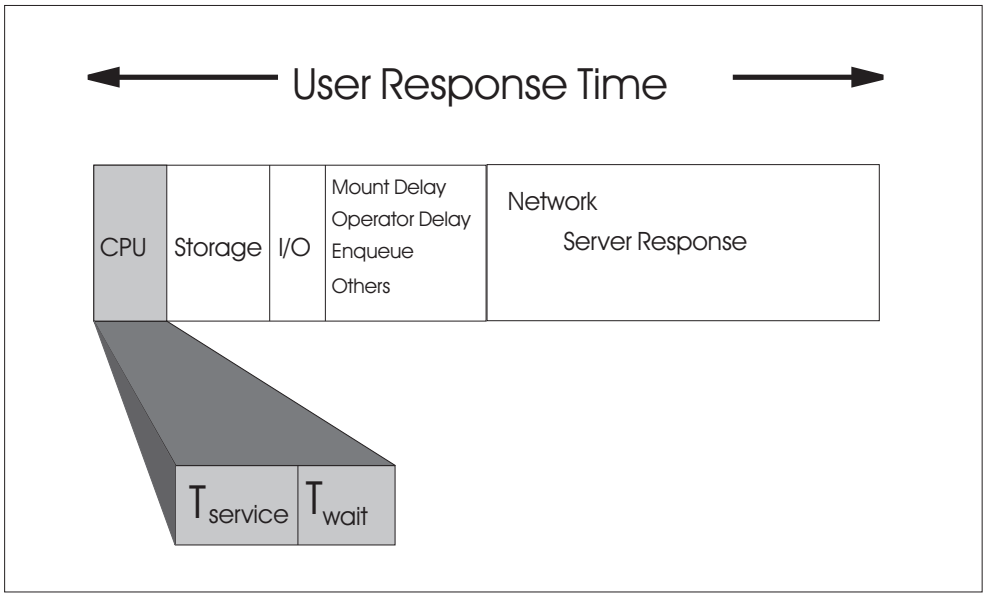


Figure 1. End-to-End Response Time Components

The reasons you should care about this are:

- For capacity planning you need to know resource consumption.
- For performance management you need to break down response time into components, to see where tuning can be done.

General Formulas

Response time is made up of service time (the time actual work is done) and waiting time (the time waiting for resource): $T_r = T_s + T_w$

- T_r is Response Time
- T_s is Service Time
- T_w is Waiting Time (that is `Queue` time or `Delay` in RMF Monitor III).

Similarly total transaction service and wait times are made up of the individual resource service and wait times.

$T_s = T_s(CPU) + T_s(I/O) + T_s(TP) + T_s(Other)$

$T_w = T_w(CPU) + T_w(I/O) + T_w(TP) + T_w(Storage) + T_w(Other)$

The Monitor I Workload Activity report shows some of this data for a group of transactions (service or report class). This applies for TSO and batch work (and potentially CICS, with interval recording).

The Monitor III Group Response Time report also shows this data, with a more useful breakdown of response time components (see Figure 2 on page 8).

GROUP Report

```
                    RMF V1R5  Group Response Time
Command ===>

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

Class: TSOPROD       Period: 1    Description: TSO Production
Primary Response Time Component:  Storage delay for local paging

                                                TRANS    --- Response Time ----
WFL    Users    Frames   Vector   EXCP   PgIn   Ended    -- Ended TRANS-(Sec) -
 %    TOT  ACT   %ACT     UTIL    Rate   Rate   Rate      WAIT   EXECUT  ACTUAL
 50     2    1      5        0     3.2    4.1     0.2     0.000   3.432   3.432

                         -AVG USG-   -------------Average Delay--------------
                   Total   PROC  DEV   PROC   DEV  STOR  SUBS  OPER   ENQ OTHER
Average Users      0.600   0.10 0.17   0.02  0.03  0.22  0.00  0.00  0.00  0.06
Response Time ACT  3.432   0.57 0.97   0.11  0.17  1.25  0.00  0.00  0.00  0.34

                     ---STOR Delay---  ---OUTR Swap Reason---  ---SUBS Delay---
                     Page  Swap  OUTR    TI    TO    LW    XS   JES   HSM   XCF
Average Users        0.19  0.03  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
Response Time ACT    1.08  0.17  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
```

Figure 2. Monitor III Group Response Time Report

Report Analysis

See `Response Time ACT` field for a breakdown of the components of response time. Think of USING as service time (T_s) and DELAY as wait time (T_w)

- `AVG USG`: 0.57 + 0.97 = 1.54 sec
- `Average Delay`: 0.11 + 0.17 + 1.25 + 0.34 = 1.87 sec

So here, if you were investigating a response time problem, `STOR` would be the best starting point.

Monitor III provides insight into the components of response. With an emphasis on causes of *delay*, it shows you the resources for which work is being delayed and the address spaces which are holding the resources.

Analyzing Workload Characteristics

Much of performance work is done, not at the individual transaction level, but at a larger *workload* level. Understanding resource requirements by workload is key to effective performance management.

The reasons you should analyze your workload are:

- To understand your system's behavior
- For setting your SLA
- For tuning
  - Where is the pain?
  - Interactions with other workloads
- For input to the capacity planning process
  - Workload growth projections
  - Processor requirements
  - Storage requirements
  - DASD requirements
  - Network requirements

Identifying Workloads

You will need to identify the different types of work in your system. Usually this is done at a service class level. To assign work to service classes you must classify the various workloads by their unique characteristics and requirements, that is:

- Response time needs
- Resource consumption (CPU, Storage, I/O)
- Priority
- Anticipated growth

Examples of workload identification at the service class level include:

- Trivial TSO (first period)
- Non-trivial TSO
- Batch
- Production CICS
- Development CICS
- IMS™
- Graphics application

Ideally, you may want to take workload differentiation one level further, matching workloads to true business functions (for example, claims processing and order fulfilment). This may require more detailed data from SMF records.

RMF reports data about different workloads grouped into categories which you have defined as *workloads* and *service classes* in your service policy.

The appropriate grouping of workloads is important.

- If you have different applications which should be managed according to the same goals, you should define the same ***service class*** for them. Applications with different goals need to be assigned to different service classes.

- If you want to get separate reporting for different applications in the same service class, you can define separate ***report classes*** for each of them. Reporting for report classes is possible with the same level of detail (report class period) as for service classes.

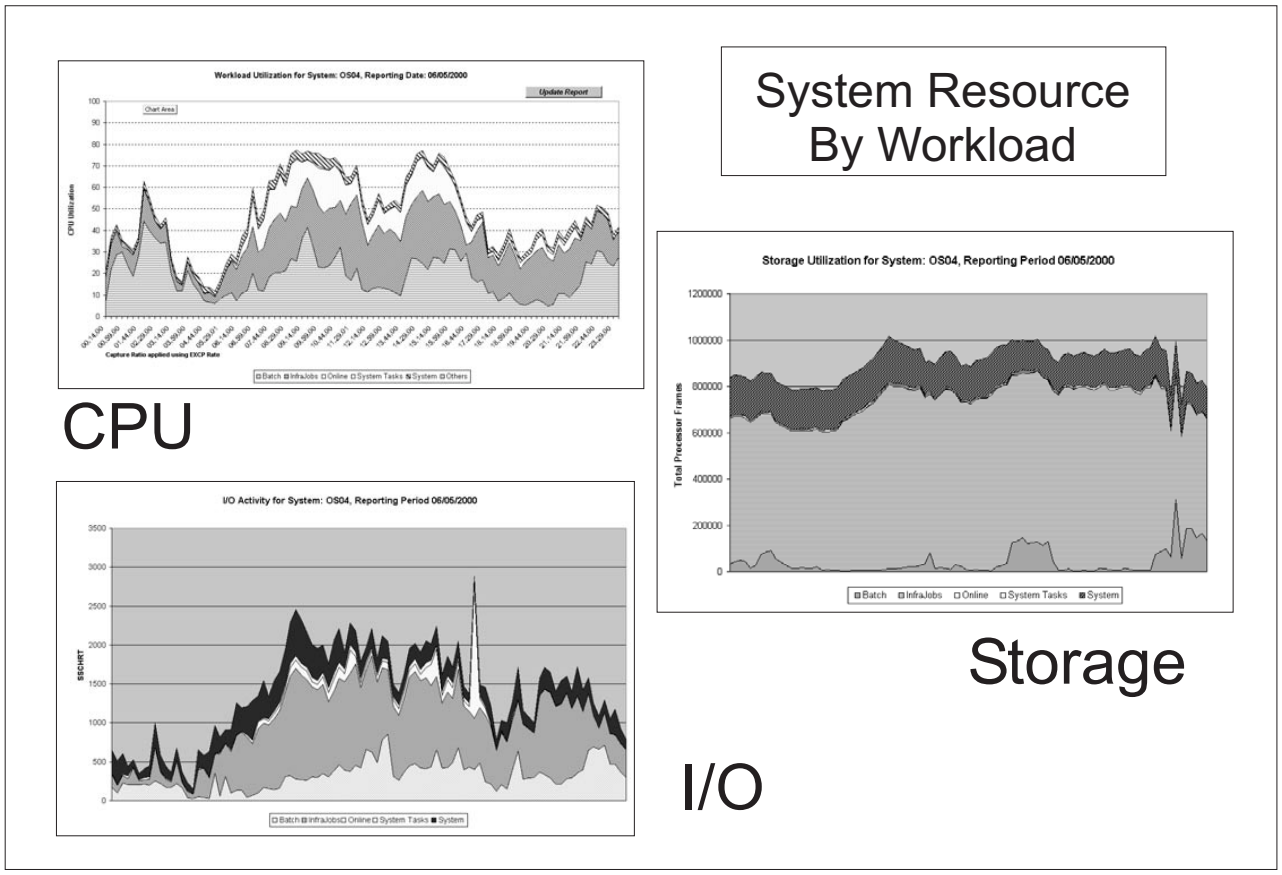


Figure 3. System Resource Summary by Workload

This figure illustrates one way to report your resource utilization by workload.

You can use the Spreadsheet Reporter to create spreadsheet files from Postprocessor data and to display them graphically. The following chapters show how to use the sample macros for creating the graphics that help you in understanding and analyzing the performance of your system.

In a similar fashion (could be graphical or numeric) you will need to record such data to:

- Measure resource consumption:
  - CPU utilization
  - Processor storage usage (*central storage* (CS) and *expanded storage* (ES))
  - I/O rates
- Understand what makes up response time:
  - Waiting for and using the resources
  - Use Monitor III Group Response Time report
- Understand the factors that influence the above

Virtual storage is another type of resource. While not as dynamic as I/O, CPU, or processor storage, virtual storage is one of the basic resources that you need to plan, even though you cannot buy it. Virtual storage can have a critical impact on delivery of service to end users: an *initial program load (IPL)* is required when not enough virtual storage is available.

RMF can measure virtual storage. It should be monitored, and the Virtual Storage report helps you verify the size of the different MVS components, as explained in "Virtual Storage Report" on page 57.

Using Monitor I Reports to Analyze Workloads

The CPU Activity report and the Workload Activity report provide the data you can use for a workload analysis.

CPU Activity Report - Part 1

```
                                    C P U   A C T I V I T Y
                                                                                PAGE    1
            z/OS V1R6               SYSTEM ID   OS04          DATE 03/14/2004          INTERVAL 14.59.996
                                    RPT VERSION V1R5 RMF     TIME 09.29.00            CYCLE 1.000 SECONDS

CPU  9672   MODEL  Z97

---CPU---  ONLINE TIME   LPAR BUSY      MVS BUSY     CPU SERIAL  I/O TOTAL         % I/O INTERRUPTS
NUM  TYPE  PERCENTAGE    TIME PERC      TIME PERC    NUMBER      INTERRUPT RATE    HANDLED VIA TPI
 0   CP    100.00        76.93          99.15        045104      320.9              0.33
 1   CP    100.00        76.85          99.13        145104      324.0              0.22
 2   CP    100.00        76.86          99.09        245104      321.6              0.22
 3   CP    100.00        76.87          98.99        345104      329.5              0.25
 4   CP    100.00        76.89          99.01        445104      325.9              0.33
 5   CP    100.00        76.87          98.88        545104      328.7              0.36
 6   CP    100.00        76.88          98.86        645104      338.4              0.46
 7   CP    100.00        76.85          98.73        745104      341.1              0.51
 8   CP    100.00        76.82          98.59        845104      335.9              0.67
CP TOTAL/AVERAGE         76.87          98.94                     2966               0.37
```

Figure 4. Monitor I CPU Activity Report - Part 1

The first part of this report provides an overview on all processors belonging to this system.

Workload Activity Report

```
                                    W O R K L O A D   A C T I V I T Y
                                                                                                    PAGE    1

      z/OS V1R6               SYSPLEX SPLEX01           DATE 06/05/2004          INTERVAL 14.59.628   MODE = GOAL
                              RPT VERSION V1R5 RMF      TIME 09.29.00

                                       POLICY ACTIVATION DATE/TIME 06/02/2004 10.43.08
                                            - WORKLOAD & SERVICE CLASS PERIODS -

 REPORT BY: POLICY=SP000010   WORKLOAD=WLONLINE   SERVICE CLASS=SCONL006   RESOURCE GROUP=*NONE
                                                  CRITICAL     =NONE
                                                  DESCRIPTION  =DB2 Production

 TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT   --DASD I/O--   ----SERVICE----   --SERVICE TIMES--   PAGE-IN RATES    ----STORAGE----
 AVG     28.23   ACTUAL                769   SSCHRT  41.4   IOC          0    TCB          581.6   SINGLE     0.0   AVG        0.00
 MPL     28.23   EXECUTION             768   RESP     2.9   CPU      42164K   SRB            0.0   BLOCK      0.0   TOTAL      0.00
 ENDED    7288   QUEUED                  0   CONN     2.4   MSO          0    RCT            0.0   SHARED     0.0   CENTRAL    0.00
 END/S    8.10   R/S AFFINITY            0   DISC     0.0   SRB          0    IIT            0.0   HSP        0.0   EXPAND     0.00
 #SWAPS      0   INELIGIBLE              0   Q+PEND   0.5   TOT      42164K   HST            0.0   HSP MISS   0.0
 EXCTD       0   CONVERSION              0   IOSQ     0.0   /SEC     46869    IFA            0.0   EXP SNGL   0.0   SHARED     0.00
 AVG ENC  0.00   STD DEV             1.918                                    APPL%  CP     64.6   EXP BLK    0.0
 REM ENC  0.00                                             ABSRPTN   1660    APPL% IFACP    0.0   EXP SHR    0.0
 MS ENC   0.00                                             TRX SERV  1660    APPL% IFA      0.0



 PER IMPORTANCE  PERF     --TRANSACTIONS--     -------------RESPONSE TIME--------------   -EX VEL%-   TOTAL   -EXE--
                 INDX     -NUMBER-     -%-     ------GOAL------  ---ACTUAL---     TOTAL   GOAL  ACT   USING%  DELAY%
 1   2            0.7         4937      68     00.00.00.300  75%    89.2%          60.5%       64.2     5.8     3.4
 2   3            0.5         1584      22     00.00.02.000  70%    97.0%                      72.0     5.8     2.3
 3   5            0.4          767      11                                                  20 51.7     2.3     2.2
TOTAL                         7288     100
```

Figure 5. Workload Activity Report - Service Class

```
                                    W O R K L O A D   A C T I V I T Y
                                                                                                    PAGE    1

      z/OS V1R6               SYSPLEX SPLEX01           DATE 06/05/2004          INTERVAL 14.59.628   MODE = GOAL
                              RPT VERSION V1R5 RMF      TIME 09.29.00

                                       POLICY ACTIVATION DATE/TIME 06/02/2004 10.43.08
                                            - WORKLOAD & SERVICE CLASS PERIODS -

                                       POLICY ACTIVATION DATE/TIME 06/02/2004 10.43.08

 ------------------------------------------------------------------------------------------------------------------- SERVICE POLICY

 REPORT BY: POLICY=SP000010
                 Default Policy

 TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT   --DASD I/O--   ----SERVICE----   --SERVICE TIMES--   PAGE-IN RATES    ----STORAGE----
 AVG    203.75   ACTUAL          16.29.483   SSCHRT  1955   IOC      17400K   TCB         5333.7   SINGLE     0.0   AVG      4711.96
 MPL    203.70   EXECUTION           3.092   RESP    13.8   CPU     386675K   SRB          411.7   BLOCK      0.0   TOTAL     959830
 ENDED   15298   QUEUED          15.35.046   CONN     3.0   MSO      19213K   RCT            5.4   SHARED     0.0   CENTRAL   403598
 END/S   17.00   R/S AFFINITY            0   DISC     9.5   SRB      29845K   IIT           44.7   HSP        0.0   EXPAND    556232
 #SWAPS   6378   INELIGIBLE         52.969   Q+PEND   0.6   TOT     453133K   HST            0.2   HSP MISS   0.0
 EXCTD    6937   CONVERSION              4   IOSQ     0.7   /SEC    503690    IFA            0.0   EXP SNGL   0.7   SHARED    584.45
 AVG ENC  0.00   STD DEV       3.48.42.725                                    APPL%  CP    644.2   EXP BLK    0.2
 REM ENC  0.00                                             ABSRPTN   2473    APPL% IFACP    0.0   EXP SHR    0.0
 MS ENC   0.00                                             TRX SERV  2472    APPL% IFA      0.0
```

Figure 6. Workload Activity Report - Service Policy

The Workload Activity report shows workload-related data on resource consumption with different levels of detail. The above examples summarizes data for one service class (SCONL006) which is DB2 Production and for the total system (*service policy*).

Analyzing Processor Characteristics

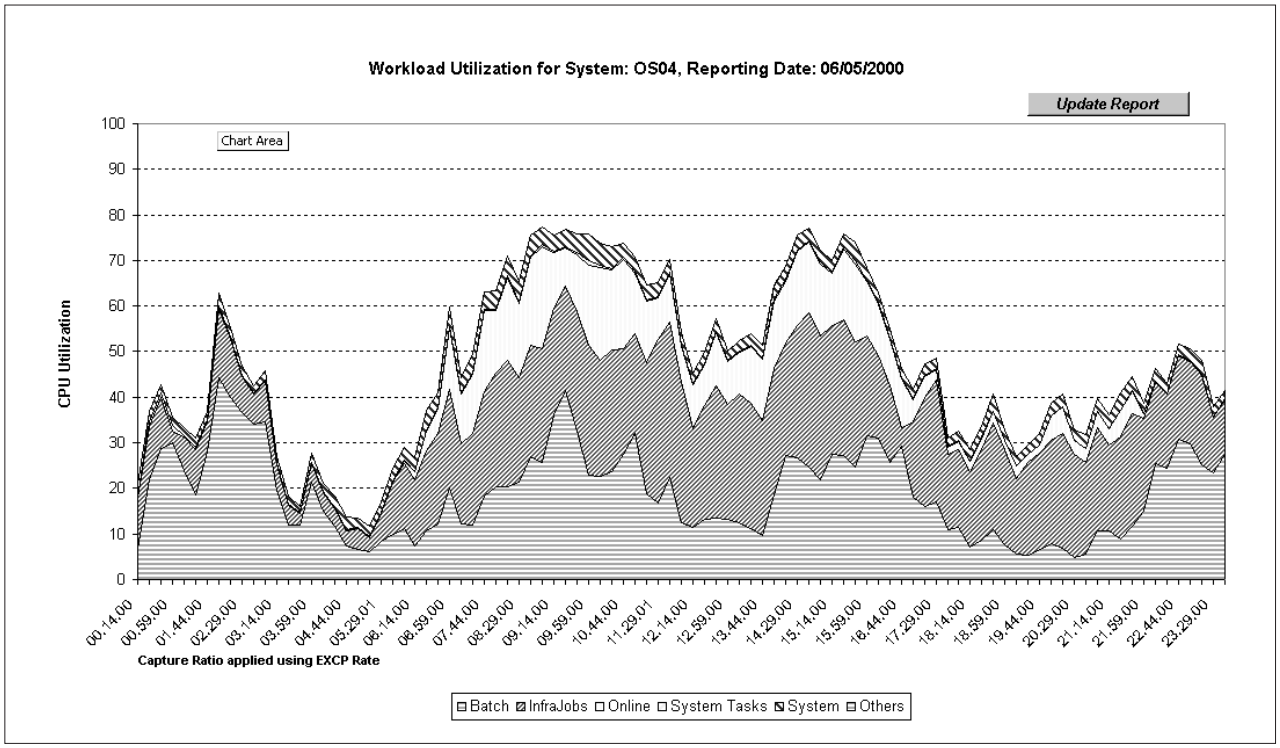


Figure 7. Processor Utilization by Workload. Spreadsheet Reporter macro RMFY9WKL.XLS (Workload Trend Report) - DayUtilization

This section will discuss how to measure CPU utilization by workload. Our approach is to convert *application time* for each service class into overall percent of CPU used for that workload. The total of all these numbers is the base for the APPL% CP value in the report.

The concept of *capture ratio (CR)* must be understood in order to do this. Simply put, the CPU time reported for the sum of all your workloads never adds up to the total CPU time used by the system. The total CPU time reported for all your workloads will typically account for 85-90% of the total CPU time used. This is not an error. It is the best level of accuracy that the reporting tools can achieve, *in a consistent, repeatable manner.*

The *uncaptured* time is sometimes, misleadingly, called system overhead. This is not so, most of it is genuine user work — it is either a political or philosophical view whether activities like paging are seen as genuine work or just overhead, it depends ...

Thus the capture ratio is:

```
      Captured CPU Time
CR =  -----------------
       Total CPU Time
```

The question then becomes: how do you account for the uncaptured, though real, CPU time? The answer is to distribute the uncaptured time among your workloads, so that the total CPU time is accounted for. Several approaches to doing this will be discussed.

How to Compute CPU Time per Workload Type

Following are the basic steps for computing CPU time by workload. Graphing CPU consumption over time (for example, by hour or by day) can help you understand your workload flow, as shown in Figure 7 as an example of a Spreadsheet Reporter macro.

1. Calculate total CPU time in seconds:

   ```
   Total CPU time = MVS Busy % * Interval * #CPs
   ```

   If your system is running in an PR/SM environment, the calculation has to be performed using LPAR BUSY and the number of logical processors.

2. Get the captured time for all address spaces.

   The Workload Activity report shows the percentage of application time. It is based on the sum of *task control block (TCB)* time, *service request block (SRB)* time, *region control task (RCT)* time, *I/O interrupt (IIT)* time, and *hiperspace service (HST)* time, and is calculated as:

   ```
              (TCB + SRB + RCT + IIT + HST - IFA)
   APPL% CP = ---------------------------------- * 100
                           Interval
   ```

   ```
   Captured time = (APPL% CP / 100) * Interval
   ```

   This results in the uncaptured CPU time:

   ```
   Uncaptured CPU time = Total CPU time - Captured time
   ```

3. Calculate the capture ratio:

   ```
         Captured CPU Time
   CR =  -----------------
          Total CPU Time
   ```

4. Get application (captured) percentage for one workload you are interested in:

   This APPL% CP value represents the percentage based on one processor. You get the percentage based on the total capacity by:

   ```
                       APPL% CP
   Total Captured % = --------
                         #CPs
   ```

5. Distribute uncaptured time based on the capture ratio:

   ```
             Total Captured %
   Total % = ----------------
                    CR
   ```

   There are more ways to distribute the uncaptured time, see “Methods of Distributing Uncaptured Time” on page 15 for a discussion of this topic.

Report Analysis

Figure 4, Figure 5, and Figure 6 give us the following values for standard CP processors:

- Total CPU time = 0.7687 × 900 × 9 = 6226.47 sec
- Captured time (POLICY) = 6.442 × 900 = 5797.8 sec
- Uncaptured time = 6226.47 - 5797.8 = 428.67 sec
- Capture ratio = 5797.8 / 6226.47 = 0.93
- DB2 (Service Class SCONL006) = 64.6 % of one processor

Figure 7 is a graphic being created with the Spreadsheet Reporter, it shows the workload utilization for different service classes during one day. The macro provides different ways to distribute the uncaptured time.

Methods of Distributing Uncaptured Time

To get a full picture of CPU consumption, the uncaptured time needs to be distributed across the workloads. There are several methods of doing this:

1. Not bad: By workload APPL time - as done in the previous calculation.
2. Better: By workload I/O rate (% of system total)
3. Best:
   a. Distribute uncaptured time by I/O rate, AND...
   b. Distribute system address space (AS) time to relevant users, e.g.
      - VTAM® time: to TSO and CICS
      - JES time: to TSO and BATCH
      - etc.
4. Probably not worth the effort: Regression analysis, etc.

How to Compute CPU Time per Transaction

There are times when additional detail of CPU utilization is needed. For capacity planning especially, you may need to know the average CPU time consumed per transaction. This is a simple calculation, dividing application time by number of transactions.

Report Analysis

Figure 5 on page 12 gives us the following values for DB2 (Service Class SCONL006):

- `ENDED TRANSACTIONS:` 7288
- `APPL% CP * INTERVAL  :` 581.4 sec (Application time)

This results in

```
                            581400
CPU time per transaction = ------- = 80 msec
                             7288
```

Analyzing Workload I/O Characteristics

It can also be useful to know the I/O rate each different workload generates. This is especially useful in capacity planning, but also for performance management - a change in workload I/O rates (or relative I/O content) can significantly change its performance.

Base for the following graphic are these OVW statements:

```
OVW(IOBATCH(SSCHRT(W.WLBATCH)))
OVW(IOINFRA(SSCHRT(W.WLINFRA)))
OVW(IOONLIN(SSCHRT(W.WLONLINE)))
OVW(IOSYSTK(SSCHRT(W.WLSYSTEM)))
OVW(IOSYSTM(SSCHRT(W.SYSTEM)))
```

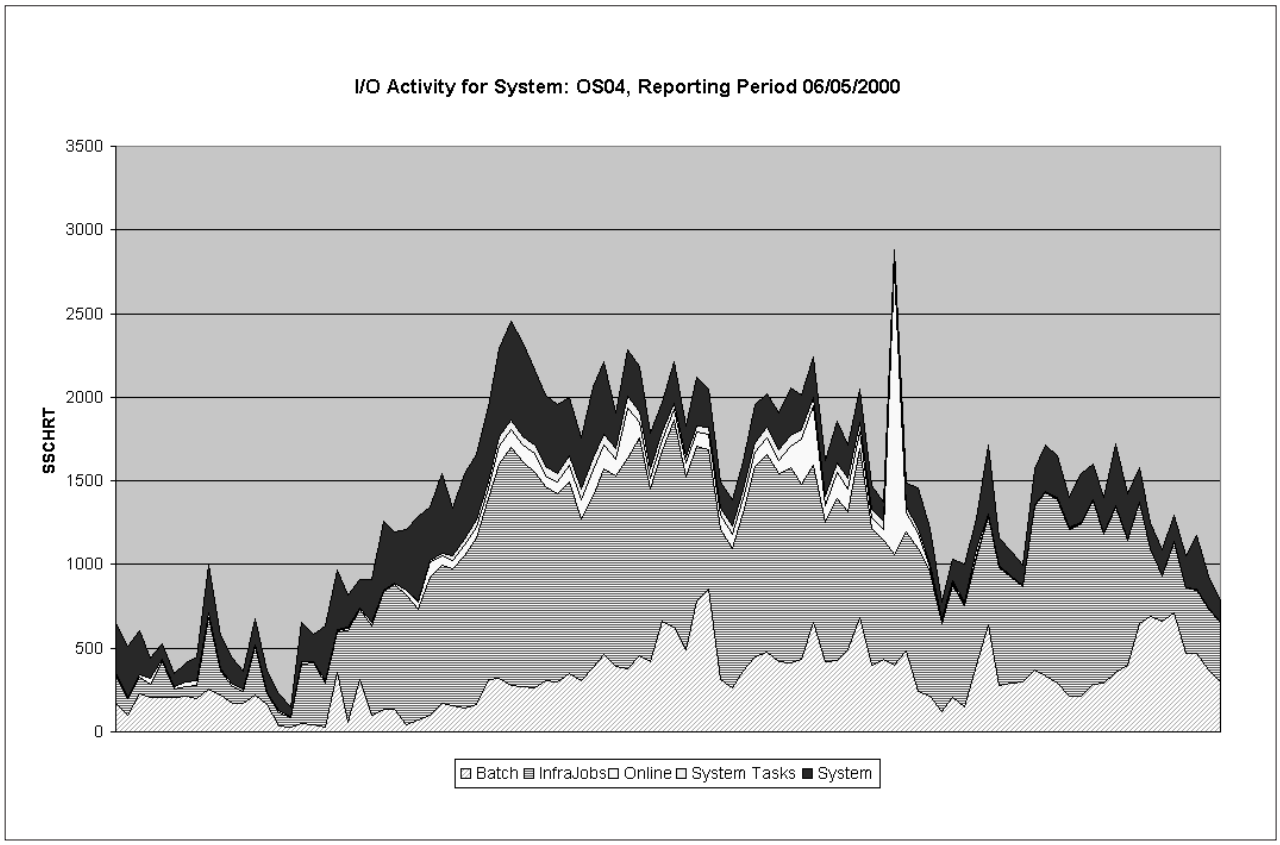


Figure 8. I/O Activity by Workload

```
REPORT BY: POLICY=SP000010   WORKLOAD=WLONLINE    SERVICE CLASS=SCONL006    RESOURCE GROUP=*NONE
                                                  CRITICAL     =NONE
                                                  DESCRIPTION  =DB2 Production

TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT   --DASD I/O--   ----SERVICE----   --SERVICE TIMES--   PAGE-IN RATES    ----STORAGE----
AVG     28.23   ACTUAL                769   SSCHRT  41.4   IOC          0    TCB          581.6   SINGLE     0.0   AVG         0.00
MPL     28.23   EXECUTION             768   RESP     2.9   CPU      42164K   SRB            0.0   BLOCK      0.0   TOTAL       0.00
ENDED    7288   QUEUED                  0   CONN     2.4   MSO          0    RCT            0.0   SHARED     0.0   CENTRAL     0.00
END/S    8.10   R/S AFFINITY            0   DISC     0.0   SRB          0    IIT            0.0   HSP        0.0   EXPAND      0.00
#SWAPS      0   INELIGIBLE              0   Q+PEND   0.5   TOT      42164K   HST            0.0   HSP MISS   0.0
EXCTD       0   CONVERSION              0   IOSQ     0.0   /SEC     46869    IFA            0.0   EXP SNGL   0.0   SHARED      0.00
AVG ENC  0.00   STD DEV             1.918                                    APPL% CP      64.6   EXP BLK    0.0
REM ENC  0.00                                             ABSRPTN   1660    APPL% IFACP    0.0   EXP SHR    0.0
MS ENC   0.00                                             TRX SERV  1660    APPL% IFA      0.0

REPORT BY: POLICY=SP000010
                Default Policy

TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT   --DASD I/O--   ----SERVICE----   --SERVICE TIMES--   PAGE-IN RATES    ----STORAGE----
AVG    203.75   ACTUAL           16.29.483   SSCHRT  1955   IOC      17400K   TCB         5333.7   SINGLE     0.0   AVG      4711.96
MPL    203.70   EXECUTION            3.092   RESP    13.8   CPU     386675K   SRB          411.7   BLOCK      0.0   TOTAL     959830
ENDED   15298   QUEUED           15.35.046   CONN     3.0   MSO      19213K   RCT            5.4   SHARED     0.0   CENTRAL   403598
END/S   17.00   R/S AFFINITY             0   DISC     9.5   SRB      29845K   IIT           44.7   HSP        0.0   EXPAND    556232
#SWAPS   6378   INELIGIBLE           52.969   Q+PEND   0.6   TOT     453133K   HST            0.2   HSP MISS   0.0
EXCTD    6937   CONVERSION               4   IOSQ     0.7   /SEC    503690    IFA            0.0   EXP SNGL   0.7   SHARED    584.45
AVG ENC  0.00   STD DEV        3.48.42.725                                    APPL% CP     644.2   EXP BLK    0.2
REM ENC  0.00                                             ABSRPTN   2473    APPL% IFACP    0.0   EXP SHR    0.0
MS ENC   0.00                                             TRX SERV  2472    APPL% IFA      0.0
```

Figure 9. Workload Activity Report - I/O Activity

The Workload Activity report shows directly the I/O activities by workload in the value SSCHRT. This is the number of start subchannels (SSCH) per second and gives the number of DASD non-paging I/Os.

Report Analysis

We see the following numbers in the sample report:

Workload	SSCH/second
DB2 Production	41.4
Policy - Total System	1955

Analyzing Processor Storage Characteristics

As with CPU and I/O, it is useful to build a clear picture of processor storage use on your system. The Workload Activity report can give you a good start on the processor storage use by workload, at the workload level.

The Spreadsheet Reporter does not contain a macro for displaying storage utilization by workload. But you can create very easily a graphic as shown in the following example. Assuming that you have the workloads `WLBATCH`, `WLINFRA`, `WLONLINE`, and `WLSYSTEM`, you can use the following OVW statements to create the required spreadsheet data:

```
OVW(TOTBATCH(STOTOT(W.WLBATCH)))
OVW(TOTINFRA(STOTOT(W.WLINFRA)))
OVW(TOTONLIN(STOTOT(W.WLONLINE)))
OVW(TOTSYSTK(STOTOT(W.WLSYSTEM)))
OVW(TOTSYSTM(STOTOT(W.SYSTEM)))
```

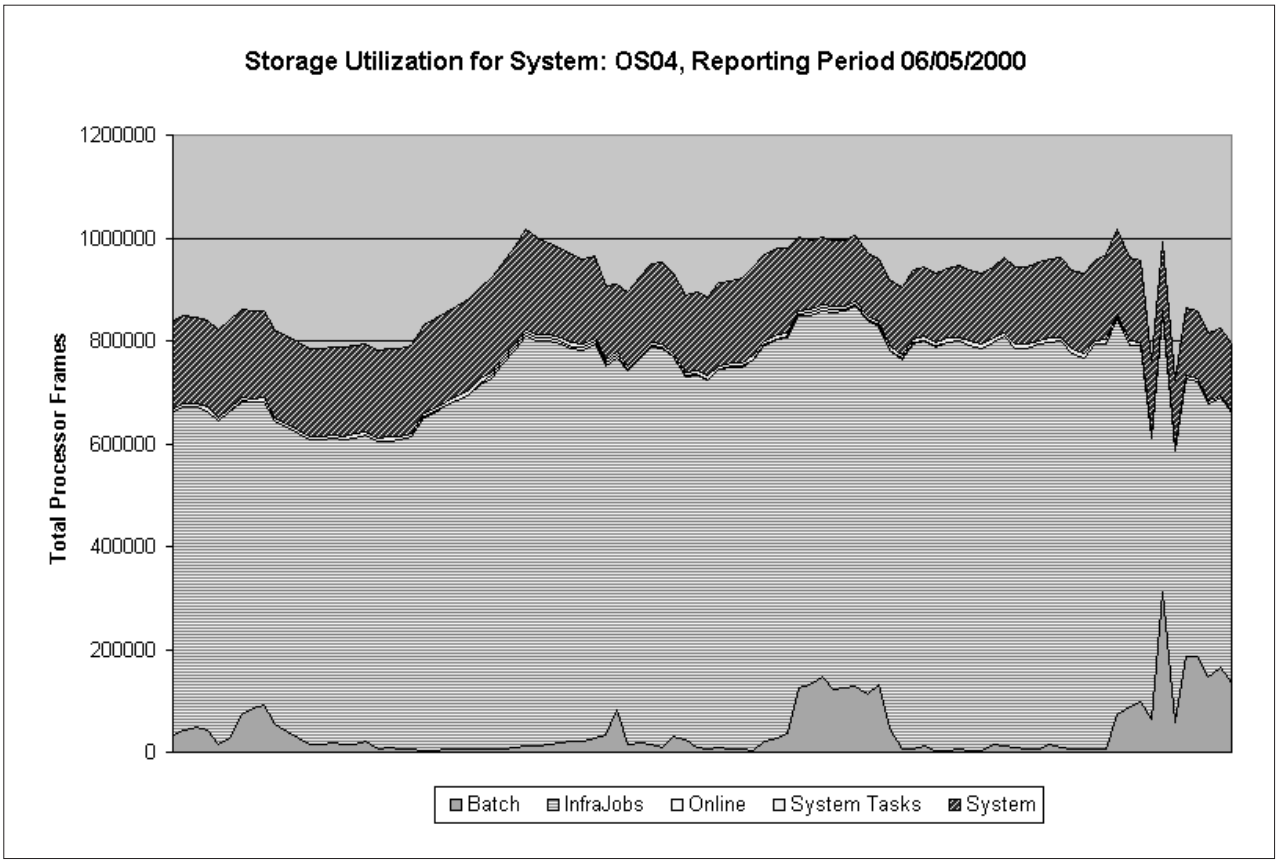


Figure 10. Processor Storage by Workload

Of course, you can display this in more detail by using other OVW conditions, for example `STOCEN` and `STOEXP` instead of `STOTOT`. You find a detailed description of all OVW conditions in chapter *"Overview and Exception Conditions"* in the *z/OS RMF User's Guide*.

Workload Activity Report

```
REPORT BY: POLICY=SP000010   WORKLOAD=WLONLINE   SERVICE CLASS=SCONL000   RESOURCE GROUP=*NONE
                                                 CRITICAL     =NONE
                                                 DESCRIPTION  =Normal 3-period TSO work

TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT   --DASD I/O--   ----SERVICE----   --SERVICE TIMES--   PAGE-IN RATES    ----STORAGE----
AVG      1.69   ACTUAL              4.821   SSCHRT  30.6   IOC     192865    TCB          51.2   SINGLE     0.0   AVG      1616.34
MPL      1.69   EXECUTION           4.812   RESP     4.9   CPU       3710K   SRB           1.0   BLOCK      0.0   TOTAL    2727.02
ENDED    1459   QUEUED                 8   CONN     3.2   MSO      84455    RCT           0.8   SHARED     0.0   CENTRAL   561.35
END/S    1.62   R/S AFFINITY           0   DISC     0.9   SRB      72522    IIT           0.4   HSP        0.0   EXPAND   2165.67
#SWAPS   1568   INELIGIBLE             0   Q+PEND   0.6   TOT       4060K   HST           0.0   HSP MISS   0.0
EXCTD       0   CONVERSION             0   IOSQ     0.2   /SEC      4513    IFA           0.0   EXP SNGL   3.9   SHARED      0.00
AVG ENC  0.00   STD DEV          2.31.597                                    APPL% CP      5.9   EXP BLK    0.0
REM ENC  0.00                                             ABSRPTN   2675    APPL% IFACP   0.0   EXP SHR    0.1
MS ENC   0.00                                             TRX SERV  2670    APPL% IFA     0.0
```

Figure 11. Workload Activity Report - Processor Storage Use

Processor Storage Use is given in the fields:

AVG # of processor frames on average per swapped-in AS in the group.

TOTAL # of processor frames on average for all AS in the group.

```
TOTAL = AVERAGE * MPL
```

This total is also broken out into CENTRAL and EXPAND(ed) storage use.

If your system is running on a zSeries processor, all fields in the RMF reports which are related to expanded storage will be empty. Expanded storage is not supported anymore in the zArchitecture.

This workload use of processor storage will account for most, but not all, of the processor storage on your system. To account for the rest, you should also look at:
- System area use of storage (see Monitor I Paging Activity report)
- Available frames (see Monitor I Paging Activity report)
- Swapped-out users (both in CS & ES - see Monitor II ASD report)

It can also be useful to know how much paging your individual workloads are doing. From the Workload Activity report, you also get paging data, from the fields:

SINGLE Page-ins per second from DASD

BLOCK Page-ins per second from DASD, for blocked pages

EXP SNGL Page-ins per second from ES

EXP BLK Page-ins per second from ES, for blocked pages

Use this paging information to help decide whether a workload's current processor storage is sufficient (see Chapter 5, "Analyzing Processor Storage Activity" for more detail on paging).

Monitor II and III also provide processor storage use (and paging) data, down to the AS level.

Processor Speed Indicators

This section presents some of the terms and formulas used in discussing CPU performance and capacity.

- **Cycle Time**

  ```
  CPU time = Path length * (Cycles / Instruction) * Cycle time
  ```

  – Path length depends on the instruction set and the compiler
  – Cycles/instruction depends on the design (e.g. microcoding of instructions)
  – Cycle time depends on the technology (for example, CPU model)

  Thus, cycle time by itself is not a complete indicator of CPU speed.

- **MIPS**

  This term depends on cycle time and number of cycles per instructions, it is only valid for comparison if the instruction sets are the same.

  In common usage today, MIPS are used as a single number to reflect relative processor capacity. As with any single-number capacity indicator, your experience may vary considerably. These numbers are often based on vendor announcements.

  By the way, *MIPS* means *Millions of Instructions Per Second* (some people say *Misleading Information about Processor Speed*). There does not exist a term *MIP* which often is used erroneously. A small processor has a speed of 1 MIPS, not 1 MIP.

- **SRM Constant — MSU**

  The SRM constant is a number derived by product engineering to normalize the speed of a CPU, so that a given amount of work would report the same service unit consumption on different processors. This removes the need to redesign your IPS performance parameters each time you upgrade your CPU.

  Service Units (SU) are typically used by the *System Resource Manager* (SRM) as a basis for resource management.

  ```
  Service Units = CPU seconds * SRM constant
  ```

  Many installations who have to charge the users do so on the basis of SUs. This is a reasonable approximation of relative capacity. But of course, *any* single-number metric for processor capacity comparison can be misleading, and LSPR numbers, based on your own workload mix, are the best to use. You get more details about this in the following section that describes ITR and LSPR.

  The SRM constants are contained in internal tables within MVS and are published in the *z/OS MVS Initialization and Tuning Guide*. The SUs are reported in many Monitor I and Monitor III reports.

  The power of a system can be characterized also in terms of MSUs (millions of service units per hour). These numbers are published in the LSPR documentation.

- **Internal Throughput Rate (ITR)**

  ITR numbers are measurements by workload type which determine the capability of a machine in terms of the number of transactions per CPU second. This is important, since processor capacity varies according to the workload mix.

```
      Number of transactions
ITR = ----------------------
            CPU time
```

  ITR is a function of:

  – CPU speed
  – Operating system
  – Transaction characteristics (workloads differ)

  ITRs are derived from the LSPR methodology. See *Large Systems Performance Reference* for details. ITR provides a reliable basis for measuring capacity and performance.

- **External Throughput Rate (ETR)**

```
      Number of transactions
ETR = ---------------------- = ITR * CPU utilization
      Elapsed time
```

  Any bottleneck, internal or external to the system, will effect the ETR. Examples include: I/O constraints, tape mount delay and paging delay. Thus it is more difficult to get a repeatable measure than with ITR.

- **Relative Processor Power (RPP)**

  RPP is the ratio of ITRs for a specific workload mix (usually 25% Batch, 25% TSO, 25% CICS, and 25% IMS) for different machines. RPP is normalized to a base machine.

  So, if a given function takes a certain amount of RPPs, you can estimate how much it would consume on a different machine by using the formula:

```
                    Used RPPs * 100
Your utilization = ------------------
                    Your machine RPPs
```

- **MFLOPS**

  MFLOPS = millions of floating point instructions per second. Used only in numerically intensive computing (scientific/technical).

In summary, various numbers may commonly be used as rough indicators of CPU capacity. Remember that most of these are very rough approximations only, and your actual capacity may vary significantly.

To get the most accurate assessment of CPU capacity for your workload, differences in workload processing characteristics must be taken into account using a methodology such as that described in the *Large Systems Performance Reference*. You find the most current version via

`http://www.ibm.com/servers/eserver/zseries/lspr/zSeries.html`

Where Do You Go from Here?

Having reviewed some of the main MVS performance concepts, the following chapters will discuss how to recognize and resolve specific performance problems.

One thought to bear in mind on your journey:

You can only do so much as a system programmer; you need help from your application designers and data base designers. A poorly written application can nullify a lot of system tuning effort. Participate in the design reviews, where you can stress that you need reasonable call structures to data bases, practical data base design etc.

Chapter 2. Diagnosing a Problem: The First Steps

Let's Start the Diagnosis

This chapter explains the first steps on your way to performance management:

- How to recognize a performance problem
- How to find that system in the sysplex that contributes most to the problem
- How to find the major cause of the problem (CPU, processor storage, I/O, etc.).

  Generally, there are two different approaches:
  - One for response time problems (primarily using Monitor III delay reports)
  - And one for system indicators showing stress (primarily using Monitor I/II resource-usage reports)
- How to continue by reading the other chapters in this book for further problem analysis and resolution

What Is a Performance Problem?

There are many views on what constitutes a performance problem. Most of them revolve around unacceptably high response times or resource usage, which we can collectively refer to as ″pain.″ The need for performance investigation and analysis is for example indicated by:

- Bad or erratic response time
  - Service level objectives are being exceeded
  - Users complaining about slow response
  - Unexpected changes in response times or resource utilizations
- Other indicators showing stress
  - Monitor III Workflow/Exceptions
  - System resource indicators (for example, paging rates, DASD response)
  - Expected throughput on the system not being attained

Ultimately, you will have to decide for yourself whether a given situation is a problem worth pursuing or not. This will be based on your own experience, knowledge of your system, and sometimes politics. We will simply assume for the following discussions that you are trying to relieve some sort of numerically quantifiable ″pain″ in your system.

Generally, a performance problem is the result of some workload not getting the resource(s) it needs to complete in a timely manner. Or, less common, the resource is obtained but is not fast enough to provide the desired response time.

There are three ways to gain the resource needed:

- Buy it.
- Create the illusion you bought it. This is known as tuning. Capability to do this implies you have been previously wasting resources. Like purchasing, there is a cost. The cost is people; while it may be higher, it is always less visible than purchase.
- Steal it (take it from a less important application). Again there is a cost. Here the cost is lower service to the application from which the resources were stolen.

If none of these options are technically or financially possible, it will be necessary to change users' (and management's) expectations.

You may have experienced the situation where you complete an extensive performance analysis only to conclude that no further tuning or stealing of resources can be done. One of the goals of this book is to assist you in determining whether you have reached this point.

Getting Started: A Top-Down Approach to Tuning

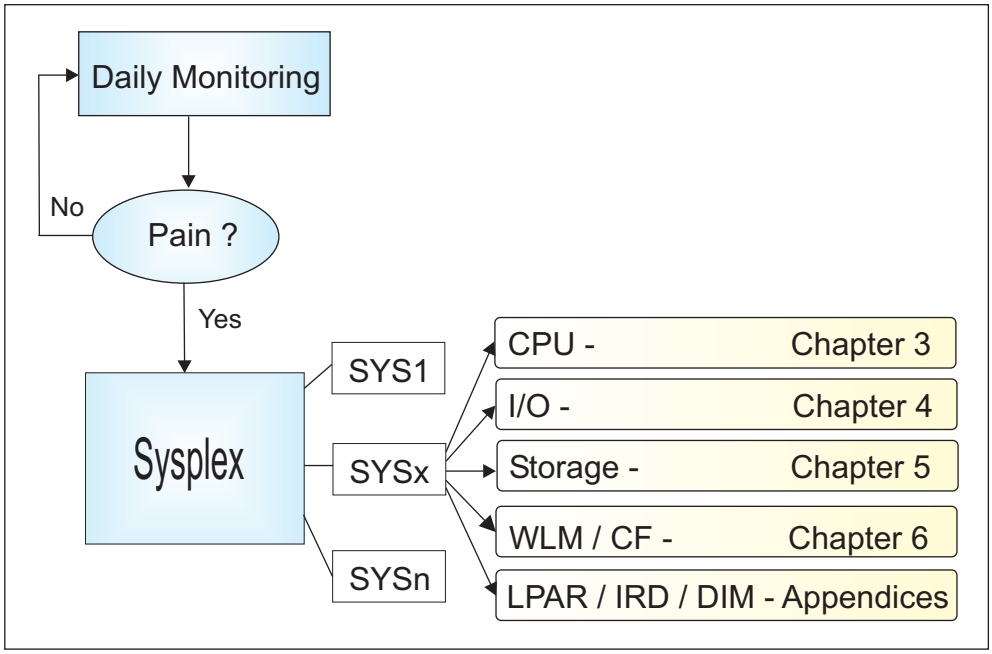


Figure 12. Simplified View of Performance Management

This book will take the following top-down approach to investigating and resolving performance problems. Figure 12 is a simplified view:

1. Through your daily performance monitoring activities recognize any of the ″pain″ indicators as being a problem on your sysplex.
2. Discover the system where you can see the most important performance problems.
3. Analyze the data to determine which component of the response time holds the most promise for relief.
4. Go to the appropriate chapter or appendix for more detailed advice.

Daily performance monitoring — what does this mean? If you ask ten people, probably you will get eleven answers. Or vice versa, there is not just one answer to the question: *How should I do daily performance monitoring?* Having in mind customers who want to be satisfied with the service that they receive from the system you are responsible for, the focus should be on observing the service level agreements. If the goals being described there can be attained — then you are providing the service that your customers are expecting. In the past, it was not a simple task of monitoring the response time and throughput objectives that are described in the service level agreements. And furthermore, it was not a simple task to adjust all parameters in the Parmlib members IEAIPSxx and IEAICSxx to reach the optimal performance of your system.

But now, it's easy and simple: you just define your goals in the service policy and let the system run. The workload manager is doing its best in managing the workload in your system for achieving all goals — your job is done automatically, be happy. STOP — if this would be true we could stop at this point of time. Of course, it's not true because there are many aspects which have to be mentioned with regard to performance management.

Starting with the sysplex view, you get a high-level overview about the performance indicators of your workloads. This will lead you to the most relevant systems in the sysplex. There, you may start with the largest cause of delay, do what you can to address it, go back and address the next largest cause of delay, and so on. Continue to do this until the service objectives are met, or until you reach a point of diminishing returns, where further effort does not pay off.

In short ...
... use the data available from RMF to ensure that your efforts are focused where they can do the most good.

Of course, with any performance problem there are many different approaches you can take to analyze the data. We have selected these two paths:

- **Response time problems**. Follow this path if you have a user or group of users complaining about response time.
- **System indicators**. Follow this path if your ongoing performance monitoring shows signs of a problem, or for a general health check. This path starts on page 41.

Where to Start Sysplex Monitoring?

When using RMF, you are interested in getting a performance overview at a glance. You have defined a service policy that contains several workloads, each of them with service classes and report classes that have goals of different types.

Monitor III and the Postprocessor offer you some reports that show how your sysplex is running at different levels of detail.

To start with:

Monitor III Sysplex Summary Report

This report shows (based on your option selection) all or only selected workloads, service and report classes, and periods with their goals and actual performance values.

When you run Monitor III in GO mode, you see the *Performance Status Line* that summarizes the key performance indicators for up to 80 time ranges, showing you the performance history of your sysplex.

To continue with:

Monitor III Response Time Distribution Report

You can use this report to see how several systems of the sysplex contribute in servicing one specific service class. This is the entry point from the sysplex to single-system reports, for example:

GROUP	Group Response Time Report
STOR	Storage Delays Report
DELAY	Delay Report
JOB	Job Delays Report
WFEX	Workflow/Exceptions Report

Monitor III Work Manager Delays Report

Here, RMF provides detailed performance data about the CICS and IMS subsystems. The Work Manager Delays report shows you for CICS and IMS several kinds of delay data about your transactions, being in the begin-to-end or in the execution phase. You get an overview of how the different CICS address spaces (for example, the AOR, TOR, or FOR) provide service to the service classes your transactions belong to.

Postprocessor Workload Activity Report

You can use this report to get performance and resource data for your service classes and workloads in the sysplex. You will find a detailed discussion including report samples in “Understanding Workload Activity Data for IMS or CICS” on page 148.

Monitor III Indicators

Sysplex Summary Report

```
                    RMF V1R5  Sysplex Summary - WTSCPLX1       Line 1 of 72
Command ===>                                                  Scroll ===> HALF

WLM Samples: 385     Systems: 3  Date: 04/30/04 Time: 09.30.00 Range: 100  Sec



                >>>>>>>>XXXXXXXXXXXXXXXXXX<<<<<<<<

Service Definition: CICSPOL              Installed at: 03/24/04, 11.57.51
     Active Policy: CICSMAJ              Activated at: 03/24/04, 11.58.16

              ------- Goals versus Actuals --------  Trans  --Avg.  Resp.  Time-
              Exec Vel  --- Response Time ---  Perf  Ended   WAIT  EXECUT  ACTUAL
Name     T  I Goal Act  ---Goal--- --Actual--  Indx  Rate    Time   Time    Time

BATCH    W          12                                0.000  0.000 0.000 0.000
BATCHLOW S  5     1  12                          0.09  0.000  0.000 0.000 0.000
CICS     W         N/A                               18.86  0.580 1.576 2.156
CICSDFLT S  2       N/A  0.300 AVG  5.079  AVG  16.9  0.111  2.970 2.109 5.079
CICSRGN  S  2    70  70                          1.00  0.000  0.000 0.000 0.000
CICUSRTX S  2       N/A  0.090 AVG  2.139  AVG  23.8  18.75  0.563 1.576 2.139
SYSTEM   W          68                                0.000  0.000 0.000 0.000
SYSSTC   S      N/A  70   N/A                          0.000  0.000 0.000 0.000
SYSTEM   S      N/A  63   N/A                          0.000  0.000 0.000 0.000
TSO      W          68                                0.537  0.006 3.460 3.466
```

Figure 13. Monitor III Sysplex Summary Report

You can select what should be part of this report by using the report options:

- *What types should be reported?* These can be everything from workload to service class period, as well as report classes.
- *What performance index should be used as threshold?* This allows to show only the data that have an index high enough for you to want to be informed about it.
- *Is any importance important?* If you want to see only high or medium important work - you can select this.
- *Are you interested in all workload groups you have defined?* Then you can display them - otherwise the report shows only active workload groups and their details.

Performance Status Line

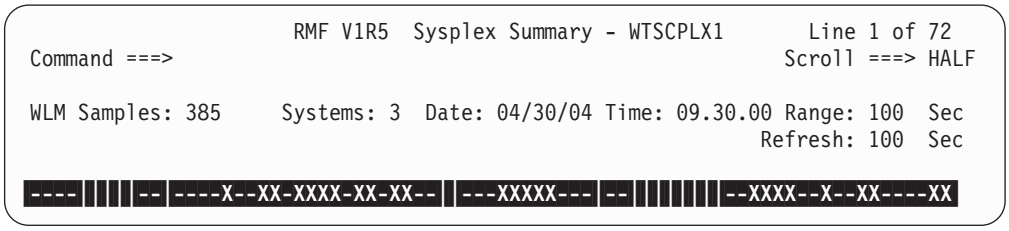


Figure 14. Monitor III Sysplex Summary Report - GO Mode

Indicator

Field: `----|||||--|----X--XX-XXXX-XX-XX-`

Description: The performance status line gives a performance indication, by its symbols and colors, for each range when the Monitor III reporter session was in GO mode:

Green All goals have been attained for this range

Yellow
Service class periods with low or medium importance have not attained the goal.

Red Service class periods with high importance have not attained the goal.

If you switch from GO mode to STOP mode, the reporter builds no reports for this time, and therefore, no indicator for the status line will be created. When continuing with GO mode, you will see one or several blank fields in the status line.

The status line is updated with each refresh of the report in GO mode. Assuming a range and refresh value of 100 seconds, you will see the *full* status line after 8000 seconds with each interval indicator shifted to the left by one position.

You can take this status line to see at a glance the performance status of your sysplex during the previous minutes or hours.

If green is the dominant color, your sysplex seems to be in good shape, as more yellow or red indicators appear, you might start investigating the sources of possible problems.

Indicator

Field: `Perf Indx`

Description: The performance index shows how well a performance goal could be achieved and is calculated from the goal and actual performance data.

Guideline: A value of **1** means that the actual value met the goal exactly, a lesser value indicates that the actual value is better than the goal, a higher value could be seen as an indication of a performance problem.

If the goal could not be attained, the lines for the service class period, service class and workload are displayed in red (for high importance) and yellow (for medium and low importance). The same color is also given to the corresponding field in the performance status line.

Problem Area: The red and yellow lines can be indicators of performance problems in the sysplex. If it is a one-time event, you might ignore it. If some lines show red continuously, further investigation is recommended.

Potential Solution: Depending on the type of the workload you want to study, you can proceed with the Response Time Distribution report to get an impression of how this service class performs on the different systems in the sysplex, or you might choose the Work Manager Delays report that shows you more about CICS and IMS delays.

Note: The calculation of the performance index is done with actual values that are averages for the reporting range (could be, for example, 100 seconds for Monitor III, or 30 minutes in the Workload Activity report).

You might distinguish between this value and the performance index that is used by the workload manager internally to manage all service classes. This index is calculated every 10 seconds and you find the values in SMF records type 99. Due to the different ranges it is obvious that a difference will be seen.

Indicator

Field: `Avg. Resp. Time`

Description: The response time information for all workloads and service classes is given in three fields:

1. `ACTUAL Time`

   The average response time for all ended transactions.

   Note that these response times are for ended transactions only. Thus, if there is a problem where transactions are completely locked out, either while queued or running, you can't see the problem on this report until the locked out transactions end.

2. `EXECUT Time`

   For CICS transactions, this includes execution time in AOR and following regions.

   For IMS transactions, this includes execution time within the MPR.

   For Batch, TSO, etc., this is the average time that transactions spent in execution.

   Note: In the Postprocessor Workload Activity report, you see this field as `EXECUTION TIME`.

3. `WAIT Time`

   This time is calculated as the difference between ACTUAL and EXECUT time, as long as ACTUAL time is the bigger value. However, for subsystem data, it can happen that EXECUT time is more than the ACTUAL time.

   For CICS transactions, this includes not only queuing in the TOR and AOR, but also processing time within the TOR.

   For IMS transactions, this includes not only queuing for the MPR, but also processing time within the CTL region.

   Otherwise, this is the average time that transactions spent waiting on a JES or APPC queue. Note that WAIT time may not always be meaningful, depending on how the customer schedules work. For example, if a customer submits jobs in hold status and leaves them until they are ready to be run, all of the held time counts as queued time.

The "server" service classes are blank in the `Avg. Resp. Time` the and `Trans Ended Rate` columns, because their "transactions" are address spaces, and response times are available only for ended transactions.

Problem Area / Potential Solution: See chapter "Understanding Workload Activity Data for IMS or CICS" on page 148.

Response Time Distribution Report

```
                        RMF V1R5  Response Time - WTSCPLX1        Line 1 of 7
Command ===>                                                   Scroll ===> HALF

WLM Samples: 385     Systems: 3  Date: 04/30/04 Time: 09.30.00 Range: 100  Sec

 Class: CICUSRTX  %   60|                                                  XXXX
Period: 1         of    |                                                  XXXX
                 TRX    |                                                  XXXX
Goal:                   |                                                  XXXX
0.090 sec avg          30|                                                  XXXX
                         |                                                  XXXX
                         |                                                  XXXX
                         |  .............................                   XXXX  Resp.
                        0|----+//---+---+---+---+---+---+---+---+---+//---- Time
                          <.054   0.063       0.090           0.126   >.135 (sec)
               --Avg. Resp. Time--   Trx     --Subsystem Data--  --Exec Data--
System   Data   WAIT  EXECUT ACTUAL   Rate    Actv  Ready Delay   Ex Vel  Delay

*ALL           0.563  1.576  2.139   18.75      1     2    79
SC47     all   0.413  1.312  1.725    8.24      1     2    71
SC49     all   0.703  1.845  2.548    5.16      1     2    85
SC50     none
SC52     all   0.660  1.692  2.352    5.35      1     2    87
SC53     none
SC54     none
```

Figure 15. Monitor III Response Time Distribution Report

This report enables you to analyze the distribution of a response time to see whether a response time goal was met and, if not, how close it came to failing. The report shows how the response time for a specific service or report class period is distributed. Two levels of detail are shown:

- A character graphic shows the distribution of response time for all systems in a sysplex which have data available in this period.

  This graphic is shown only for a periods with a response time goal.
- The details of how each system contributed to the overall response time.

Using cursor-sensitive control, you can navigate from this report to the GROUP report, which provides a detailed analysis of the response time, or to the SYSINFO report.

Indicator

Field: Response Time Distribution Graphic

Description: The horizontal axis shows response time (in seconds) with the response time goal in the middle. The middle section of the graph surrounding the goal shows the distribution of transactions that met between 60% and 150% of the goal.

Guideline: Easy — the more *green* fields you see the better your system is performing.

Problem Area: A big number of transactions on the right of the graphic in *red* indicates a problem.

Potential Solution: There can be different reasons for getting a distribution graphic like the one shown in the sample report:

- If you get this graphic more or less permanently, then probably different types of transactions are classified for the same service class. You can analyze the response time of the transactions by classifying them to different report classes, and then — as the second step — to different service classes to get more meaningful reports.
- The problem can also be either an untuned sysplex, a temporary performance problem, or an unrealistic performance goal. Then you need to investigate further.

Indicator

Field: `System`

Description: The bottom part of the report shows each system in the sysplex that provides service to the chosen service class.

Guideline: You can use this report to evaluate possible anomalies among the different systems that provide service.

Problem Area: This part of the report shows whether the any system has problems achieving the required goal.

Potential Solution: With cursor-sensitive control you can navigate from this report into the SYSINFO report for the system you are interested in. Then you can continue performance analysis as described in "Using Monitor III Reports" on page 36.

Work Manager Delays Report

```
                    RMF V1R5  Work Manager Delays - WTSCPLX1    Line 1 of 7
Command ===>                                                  Scroll ===> HALF

WLM Samples: 385     Systems: 3  Date: 04/30/04 Time: 09.30.00 Range: 100  Sec

Class:  CICUSRTX   Period: 1         Avg. Resp. time: 2.139 sec for  1805  TRX.
Goal:   0.090 sec average            Avg. Exec. time: 1.576 sec for  1558  TRX.
Actual: 2.139 sec average            Abnormally ended:                 0  TRX.

Sub  P  --------------- Response time breakdown (in %) ---------- -Switched--
Type     Tot  Act  Rdy Idle ----------------Delayed by-----------  Time (%)
                           LOCK  I/O CONV DIST SESS TIME PROD MISC LOC SYS REM

CICS B  79.0  0.5  1.8    0  2.3  1.3 67.0    0  6.1    0    0    0  67   0   0
CICS X  66.4  0.5  0.4    0  2.8 21.6    0    0 40.8    0    0  0.3  22   0   0


-----------  Address Spaces Serving this Service Class CICUSRTX  --------------
Jobname   M  ASID System Serv-Class Service Proc-Usg I/O-Usg  Veloc  Capp Quies

CICSAOR1       46   SC47    SYSSTC      100      .61       4     55     0     0
CICSAOR2       49   SC47    SYSSTC      100      3.7       3     74     0     0
CICSAOR3       50   SC49    SYSSTC      100      4.2       1     43     0     0
CICSAOR4       52   SC52    SYSSTC      100      2.1       2     79     0     0
CICSCMAS       54   SC47    SYSSTC       50       11       8     54     0     0
CICSFOR        39   SC47    CICSRGN     100       14       6     77     0     0
CICSTOR        51   SC47    CICSRGN     100       14       8     63     0     0
```

Figure 16. Monitor III Work Manager Delays Report

Indicator

Field: Response time breakdown (in %)

Description: This part of the report provides performance information for the begin-to-end phase (CICS only) and the execution phase of CICS and IMS.

Guideline: This information is the same as in the Postprocessor Workload Activity report with just the difference in the interval length.

For details, refer to “Understanding Workload Activity Data for IMS or CICS” on page 148.

Indicator

Field: `Address Spaces Serving this Service Class`

Description: You can see in this part of the report what address spaces provide service on different systems in the sysplex.

Guideline: The velocity (`Veloc`) and the processor service (`Proc-Usg`) for each address space can be seen as initial indicators to know how well the address spaces are providing service to the service classes of your transactions.

Identifying the Major Delay in a Response Time Problem

You have a response time problem? Want to know where to go next? This section will take you through several methods of analyzing a response time problem to find the major cause. It then points you to the relevant chapter for further diagnosis and problem resolution suggestions.

For a response time problem this means understanding the end-to-end response time components as completely as possible. As shown in Figure 17, most transactions can be viewed as having two major response time components:

- *internal* (for example, CPU time, paging, I/O)
- *external* (for example, network delay, server delay from other platforms)

RMF can help you analyze the internal response time components, and that is where this book will focus. The external response time components are beyond the scope of RMF, and will require other monitoring tools such as NetView®*.

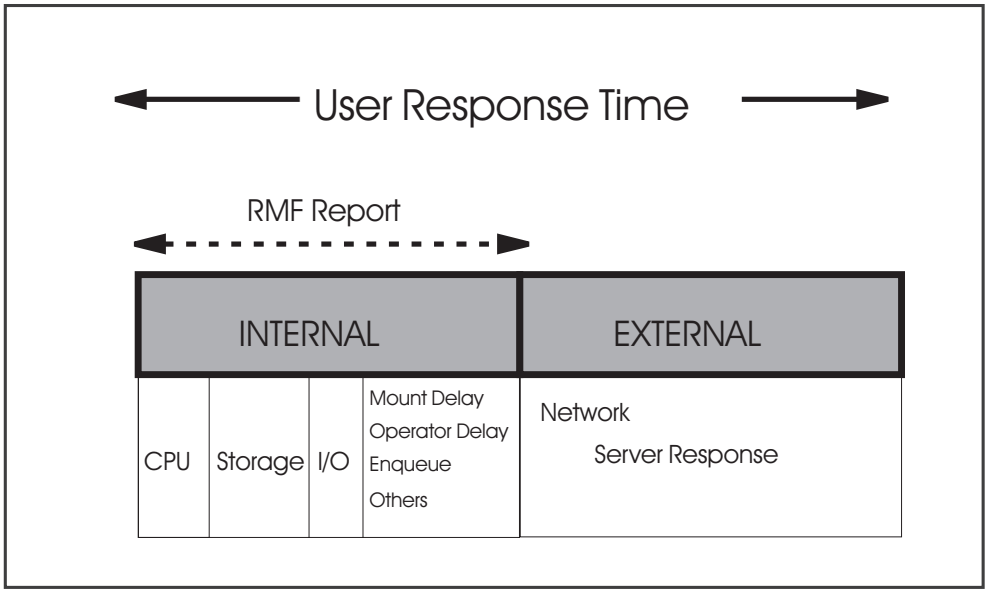


Figure 17. End-to-End Response Time Components

Understanding the relative proportions of these components allows you to see where you have the most leverage to improve response times. For example, if 90% of your response time is in I/O, why start with the CPU?

For response time problems, generally the best place to start is with Monitor III. It reports on contention for resources (both hardware and software) and any associated *delays*.

Start by setting the *range* (or interval) of the data you're looking at to match at least a portion of the time in which the problem occurred. This could be current data in GO mode, or prior data. The choice of the range value must not be too high, which would smooth the peaks, or too low, to have enough samples to provide useful data.

The range should contain at least about 100 cycles. If the cycle is one-second (which is a good value, and is the default in RMF) a 1-5 minute time for the interval will generally fit. 100 seconds is the recommended value. Displayed percentages are then easily converted into elapsed time. If you are unfamiliar with altering these values, see the *z/OS RMF User's Guide* for more details.

If your response problem is sporadic, or due to brief spikes, you may need to adjust your range, moving forward or backward in time to find the problem. If you're having trouble, set the range to 100 seconds and move through time to pinpoint a spike.

Do not base your analysis on a single occurrence, as it could be an anomaly. Check over multiple days or times of day, to verify that the problem is real.

Using Monitor III Reports

This section will discuss several reports, showing you how to identify the major cause of delay to a workload. Your starting point might be the GROUP report when you used the cursor sensitivity in the SYSRTD report.

GROUP Report

If you have a response time problem for TSO or batch, go to the Group Response Time report. Look at the Primary Response Time Component field to see the largest component of response time.

```
                    RMF V1R5  Group Response Time
Command ===>

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

Class: TSOPROD       Period: 1    Description: TSO Production
Primary Response Time Component:  Storage delay for local paging

                                                TRANS    --- Response Time ----
WFL    Users    Frames   Vector   EXCP   PgIn   Ended    -- Ended TRANS-(Sec) -
 %    TOT  ACT   %ACT     UTIL    Rate   Rate   Rate      WAIT   EXECUT  ACTUAL
 50     2    1      5        0     3.2    4.1    0.2     0.000   3.432   3.432

                         -AVG USG-   -------------Average Delay--------------
                  Total   PROC  DEV   PROC   DEV  STOR  SUBS  OPER   ENQ OTHER
Average Users     0.600   0.10 0.17   0.02  0.03  0.22  0.00  0.00  0.00  0.06
Response Time ACT 3.432   0.57 0.97   0.11  0.17  1.25  0.00  0.00  0.00  0.34

                    ---STOR Delay---  ---OUTR Swap Reason---  ---SUBS Delay---
                    Page  Swap  OUTR    TI    TO    LW    XS   JES   HSM   XCF
Average Users       0.19  0.03  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
Response Time ACT   1.08  0.17  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
```

Figure 18. Monitor III Group Response Time Report

Report Analysis

This report has two indicators for storage problems:

- Primary Response Time Component: Storage delay for local paging
- Response Time ACT: Average storage delay of 1.25 seconds

For more information, position the cursor on the field you are interested in and press ENTER. For this example, the reporting link will take you to the Storage Delay report, shown in figure 19.

STOR Report

```
                        RMF V1R5  Storage Delays                 Line 1 of 2
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

             Service  DLY    ------- % Delayed for ------   -- Working Set --
Jobname   C  Class     %     COMM  LOCL   VIO  SWAP  OUTR   Central  Expanded
BAJU      T  TSO      21        0    18     0     3     0       273         5
BHOL      T  TSO       1        1     0     0     0     0      1111         4
Report is for service class TSO only.
```

Figure 19. Monitor III Storage Delays Report

This report shows the address spaces which are affected by storage delays, in this example just for performance group two.

Report Analysis

Address space BAJU is shown with the largest delay.

DELAY Report

If you do not know which group a delayed user belongs to, use the Delay report.

```
                         RMF V1R5  Delay Report                 Line 1 of 43
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

            Service    WFL USG DLY IDL UKN ------ % Delayed for ----  Primary
Name     CX  Class  Cr  %   %   %   %   %  PRC  DEV STR SUB OPR  ENQ  Reason

BHOLEQB  B  BATCH        0   0  51   0  49    0   0   0   0   0   51  SYSDSN
JES2     S  SYSSTC       0   0   2   0  98    0   2   0   0   0    0  SYSPAG
SMF      S  SYSSTC       0   0   1   0  99    0   1   0   0   0    0  SYSPAG
BAJU     T  TSO          29   9  22  67   2    1   0  21   0   0    0  LOCL
RMFGAT   S  SYSSTC       50   1   1   0  98    1   0   0   0   0    0  BAJU
BHOL     T  TSO          78  18   5  73   4    1   3   1   0   0    0  RMFUSR
BHOLPRO2 B  BATCH        93  93   7   0   0    7   0   0   0   0    0  BHOLPRO1
```

Figure 20. Monitor III Delay Report

The report lists all delayed address spaces sorted in ascending order by workflow percentage.

Report Analysis

There are two jobs with significant delays:

- BHOLEQB with 51% enqueue delay
- BAJU with 21% storage delay

We will continue analyzing the storage problem by positioning the cursor at the BAJU address space name for further processing. This leads to the Job Delay report with initial information about possible causes, as shown in figure 21.

JOB Report

```
                          RMF V1R5  Job Delays                  Line 1 of 4
Command ===>                                                 Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

Job: BAJU         Primary delay: Paging of private area storage.

Probable causes: 1) Job may be using excessive central storage.
                 2) Paging configuration may need tuning.
                 Help panels contain more possible causes.

--------------------------- Job Storage Usage Data ----------------------------
Average Frames:   300          Working Set:    278          Fixed Frames:    43
Active Frames:     92          Aux Slots:      978          DIV Frames:       0
Idle Frames:      209          Page In Rate:  14.1          ES Move Rate:   0.0

--------------------------- Job Performance Summary ---------------------------
        Service    P WFL -Using%- DLY IDL UKN ----- % Delayed for ----- Primary
CX ASID Class   P Cr  %  PRC DEV   %   %   %  PRC  DEV STR SUB OPR ENQ  Reason
T  0027 TSO     *     29   4   5  22  67   2    1    0  21   0   0   0  LOCL
        TSO     1     33   1   1   4  67   2    0    0   4   0   0   0  SWAP
        TSO     2     35   2   4  11   0   0    1    0  10   0   0   0  LOCL
        TSO     3     13   1   0   7   0   0    0    0   7   0   0   0  LOCL
```

Figure 21. Monitor III Job Delays Report

Report Analysis

Monitor III has done basic analysis of the status and provides some possible causes. Use these as a starting point, and investigate further to confirm or rule-out these possibilities. In this example, swap trim may also be worth investigating.

In this example you would need to investigate processor storage further. See Chapter 5, “Analyzing Processor Storage Activity” for details.

WFEX Report

If your help desk uses the Workflow/Exceptions report as a continual system monitor, they would be alerted to exception conditions. These could be your first indicators to a problem. This depends on your having customized the screen to your particular needs and thresholds. See the *z/OS RMF Report Analysis* for tailoring information.

Or, you can use this screen as your first diagnosis aid if investigating a user problem.

```
                    RMF V1R5  Workflow/Exceptions               Line 1 of 4
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

--------------------------- Speed (Workflow) ---------------------------------
           Speed of 100 = Maximum, 0 = Stopped         Average CPU Util: 100 %
Name       Users Active      Speed          Name       Users Active      Speed
*SYSTEM       43      3         72          TSOPROD        2      1         50
ALL TSO        2      1         50          BATCHPRD       6      3         75
ALL STC       35      0         84
ALL BATCH      6      3         75
ALL ASCH                   Not avail
*PROC         17      2         93
*DEV           6      0         21
------------------------------ Exceptions ----------------------------------
Name       Reason        Critical val. Possible cause or action
*ECSA*     SECS% > 85     90.6 %       System ECSA usage 91 %.
*STOR      TSQAO > 0      784K bytes   SQA overflow into CSA 784K.
BHOLEQB    ENQ -SYSDSN    51.0 % delay ENQ.TEST.DAT
HSM        Not avail                   Job HSM is not running.
```

Figure 22. Monitor III Workflow/Exceptions Report

Speed, or workflow, can be used as a performance indicator either for the total system or for groups of users or resources. It is an indicator of how well a workload is able to get the system resources it wants. 100% means it gets whatever it wants whenever it wants it; 0% means it never gets what it wants. Speed is also used on a resource level (processors or devices) to reflect contention for that resource.

Speed can also be a useful indicator for exception reporting for your workloads. Make note of the workflow values reported when your system is running well. Use these as your own guideline values to set up exception reporting.

Report Analysis

The Speed of 50% for all TSO users might suggest a response time problem, if users are complaining, or if your experience has shown that a higher workflow is required to meet service objectives on your system.

However, your system may function well at a 50% speed. Use your judgment and review related indicators to determine whether or not a low speed is a problem.

DELAY versus USING

So far, we have been concentrating on workload delays. Your workload's dominant state could also be USING. Both cases will be discussed further in the relevant chapters.

Performance problem due to DELAY**:** If the largest single component of response or delay for the workload you're interested in is:

- PROC (or processor delays), go to Chapter 3, "Analyzing Processor Activity" for further analysis
- DEV (or device delays), go to Chapter 4, "Analyzing I/O Activity" for further analysis
- STOR (or storage delays), go to Chapter 5, "Analyzing Processor Storage Activity" for further analysis

If some other component (ENQ, IDL, SUBS, OPER, UKN) indicates a problem, see Appendix D, "Other Delays."

Performance problem due to USING**:** If there are no delays of any significance (at least 5-10%), look at the USING% values for PROC and DEV in the Job Delay report (other reports also have this information). Select the larger, and go to the corresponding CPU or I/O chapter for some thoughts on reducing the workload's need for that resource or speeding up the processing.

Monitoring Daily Performance: System Indicators

As discussed at the beginning of this chapter, there are different kinds of performance problems. Response time complaints or SLA exceptions are two examples. Potential problems may also be indicated by your daily performance monitoring. This section will discuss some RMF indicators which will be of use in your daily monitoring.

Here are some of the most frequently used global indicators of MVS system health, along with some *rules-of-thumb* (ROTs), on what constitutes potential trouble.

As with any ROTs, these should be looked at as a starting point only. There is no magic single number for any indicator that suggests you have a problem.

The most important thing is that your service levels are being met.

We will highlight the indicators by using RMF reports, with pointers to the appropriate chapter of this book for more details.

As you read these, remember that Monitor III will show you the impact of most of these indicators:

- *WHO* is being delayed?
- For *HOW LONG?*
- And *BY WHOM?*

Use this delay information to help you decide whether a given indicator really means trouble or not.

Summary of Major System Indicators

How do you like this approach?

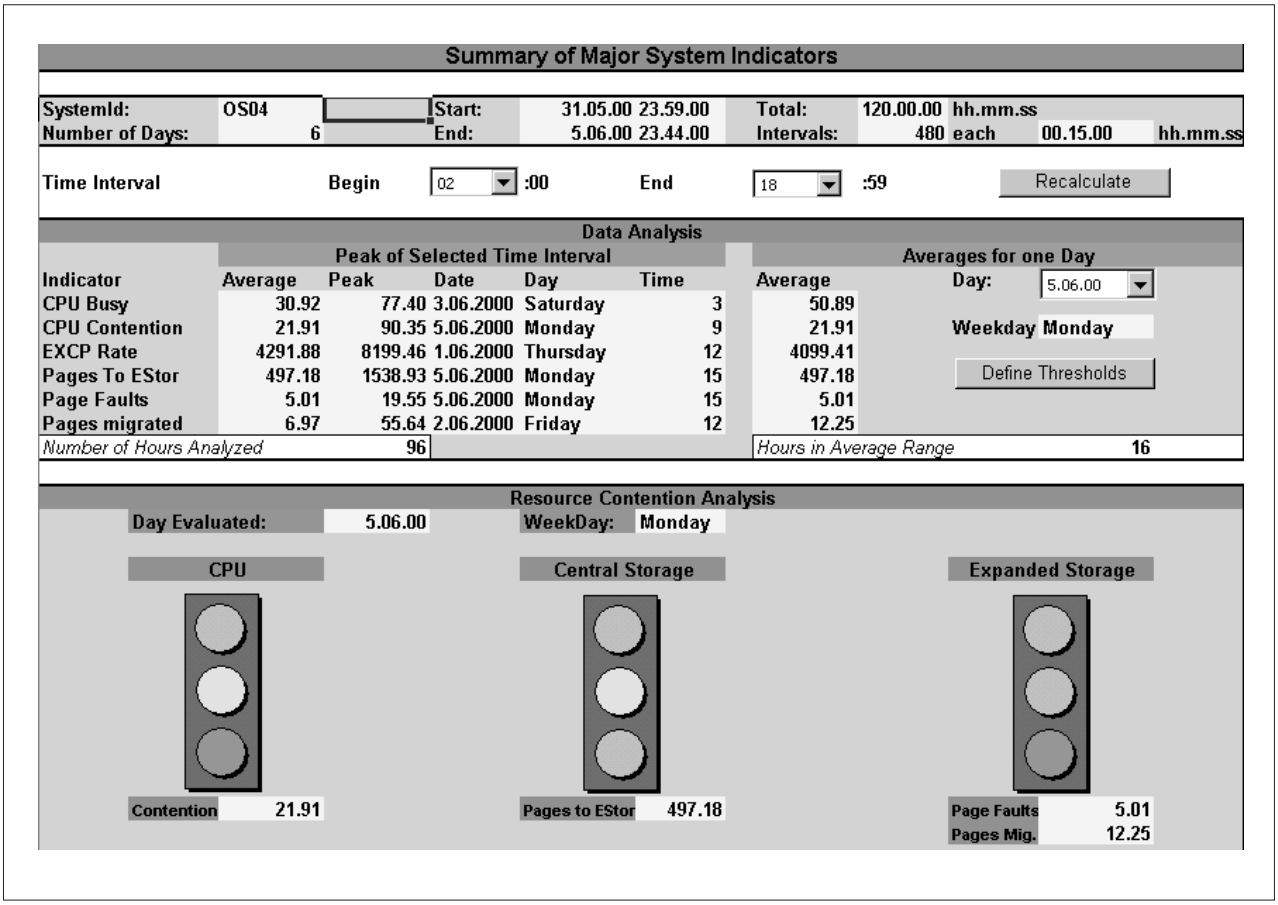


Figure 23. Summary of Major System Indicators. Spreadsheet Reporter macro RMFY9OVW.XLS (System Overview Report - Summary)

The Summary report displays key indicators representing the performance of your system. At a glance, you see whether your system experiences processor or storage contention.

A resource contention analysis is done for three indicators:

CPU Contention The percent of time with at least one job waiting for the CPU.

CS Contention The page movement rate from CS to ES is used as indicator for CS contention.

ES Contention The page fault rate and the migration rate are used as indicators for ES contention.

You can specify a time range and a day to perform the contention evaluation. The calculated values will be compared against a set of defined thresholds:

CPU Contention 20 - 80

CS Contention 200 - 800

ES Contention	10 - 50 (page fault rate)
	30 - 80 (migration rate)

You can specify the low and high thresholds which are used to set the traffic light to red or green according to your own experiences. Especially, the ES contention threshold depends on the key workload you are running in your system. If it is an online application (for example CICS or DB2), you should decrease the values to 10 - 30 (page fault rate) and to 20 - 50 (migration rate). On the other hand, a TSO system can run with much higher values, here you might specify 20 - 70 and 50 - 150.

Please keep in mind that the contention analysis should be done with data for a longer interval, the sample report covers prime shift data (8:00am to 3:59pm), therefore, the thresholds are different to other values in this publication, because those will be taken for interval reports of 15 or 30 minutes.

Using Monitor I Reports

This section discusses the use of the following Monitor I reports:

CPU Activity Report

```
                                        C P U  A C T I V I T Y
                                                                                PAGE 1

           z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004         INTERVAL 14.59.996
                                  RPT VERSION V1R5 RMF      TIME 09.29.00           CYCLE 1.000 SECONDS

CPU  9672   MODEL  Z97

---CPU---  ONLINE TIME   LPAR BUSY      MVS BUSY     CPU SERIAL  I/O TOTAL         % I/O INTERRUPTS
NUM  TYPE  PERCENTAGE    TIME PERC      TIME PERC    NUMBER      INTERRUPT RATE    HANDLED VIA TPI
 0    CP   100.00         76.93           99.15        045104       320.9              0.33
 1    CP   100.00         76.85           99.13        145104       324.0              0.22
 2    CP   100.00         76.86           99.09        245104       321.6              0.22
 3    CP   100.00         76.87           98.99        345104       329.5              0.25
 4    CP   100.00         76.89           99.01        445104       325.9              0.33
 5    CP   100.00         76.87           98.88        545104       328.7              0.36
 6    CP   100.00         76.88           98.86        645104       338.4              0.46
 7    CP   100.00         76.85           98.73        745104       341.1              0.51
 8    CP   100.00         76.82           98.59        845104       335.9              0.67
CP TOTAL/AVERAGE          76.87           98.94                      2966               0.37



SYSTEM ADDRESS SPACE ANALYSIS         SAMPLES =    898

            NUMBER OF ASIDS                          DISTRIBUTION OFQUEUE LENGTHS  (%)
TYPE       MIN    MAX     AVG       0     1     2     3     4     5     6     7-8   9-10   11-12  13-14   14+
----     ------ ------ --------   -----  ----- -----  ----- -----  ----- -----  -----  -----  -----  -----  -----
IN
READY         5     37     14.8     0.0   0.0   0.0   0.0   0.0   0.2   0.4   2.7    9.7    16.8   22.3   47.5

                                    0     1-2   3-4   5-6   7-8   9-10  11-15 16-20  21-25  26-30  31-35   35+
                                   -----  ----- -----  ----- -----  ----- -----  -----  -----  -----  -----  -----

IN           161    198    177.5    0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0    0.0    0.0    0.0  100.0
OUT
READY          0      2      0.0   98.2   1.7   0.0   0.0   0.0   0.0   0.0   0.0    0.0    0.0    0.0    0.0
OUT
WAIT         161    188    174.5    0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0    0.0    0.0    0.0  100.0
LOGICAL
OUT RDY        0      2      0.0   97.9   2.0   0.0   0.0   0.0   0.0   0.0   0.0    0.0    0.0    0.0    0.0
LOGICAL
OUT WAIT      50    105     83.4    0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0    0.0    0.0    0.0  100.0

BATCH         76     85     79.9    0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0    0.0    0.0    0.0  100.0
STC          218    232    226.9    0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0    0.0    0.0    0.0  100.0
TSO           87     97     92.4    0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0    0.0    0.0    0.0  100.0
ASCH           0      1      0.0   99.7   0.2   0.0   0.0   0.0   0.0   0.0   0.0    0.0    0.0    0.0    0.0
OMVS          35     43     36.2    0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0    0.0    0.0   62.5   37.4
```

Figure 24. Monitor I CPU Activity Report

Rules-of-Thumb

BUSY TIME PERCENTAGE

A value of over 85%, along with the total of the IN READY numbers for the columns 0,1,...N below 80%, implies contention for the CPU.

If your CPU is more than 85% busy, look at the values for the DISTRIBUTION OF QUEUE LENGTHS (%). Add up the IN READY numbers for the columns 0,1,...N, where N is the number of CPs in your system (e.g. for a system with two CPs add up columns 0+1+2). This total is the percent of time that you had enough CPs to handle all the work on the IN READY queue. The remaining time, you had work waiting for CPs. If your total for the columns 0,1,...N was 80%, then 20% of the time you had ready work waiting for a CP.

This is not to say that you cannot run an MVS system at higher utilizations, on up to 100% busy. You can. Just be aware that the busier the CPU, the longer the CPU delay for lower-priority workloads will be. The more low-priority work you have, and the fewer non-CPU bottlenecks you have (for example, responsive I/O), the busier you will be able to drive your CPU and still maintain good response for your high-priority workloads.

If your goal is to run your system to 100% busy, minimize your non-CPU bottlenecks and use the information from the Monitor III Processor Delay report to keep an eye on CPU delay for your low-priority work.

BUSY TIME PERCENTAGE

Conversely, a low CPU busy could indicate that other bottlenecks in the system are preventing work from being dispatched.

A peak-to-average ratio (for example, peak hour busy/prime shift average busy) of less than 1.3-1.4 for a commercial workload could also indicate bottlenecks preventing work from using the CPU.

This ratio is inversely correlated to IN READY users: the numbers of IN READY users go up as the peak-to-average ratio comes down. This is due to latent demand and the inability of the system to dispatch all the ready work.

OUT READY

Number of address spaces swapped out (probably TSO and batch), but ready to execute. If it is greater than 1, and TSO or batch response is an issue, look into this. This could reflect processor storage constraints, and probably a need for updating IPS and OPT parameters.

Report Analysis

- BUSY TIME PERCENTAGE is 98.94%
- Total of IN READY for 0 to 10: 13%

This might point to a problem, because at least 87% of the time one or more address spaces are delayed for CPU. The Monitor III Processor Delay report will tell you which AS were delayed, and for how long.

See Chapter 3, "Analyzing Processor Activity" and Chapter 5, "Analyzing Processor Storage Activity" for further analysis.

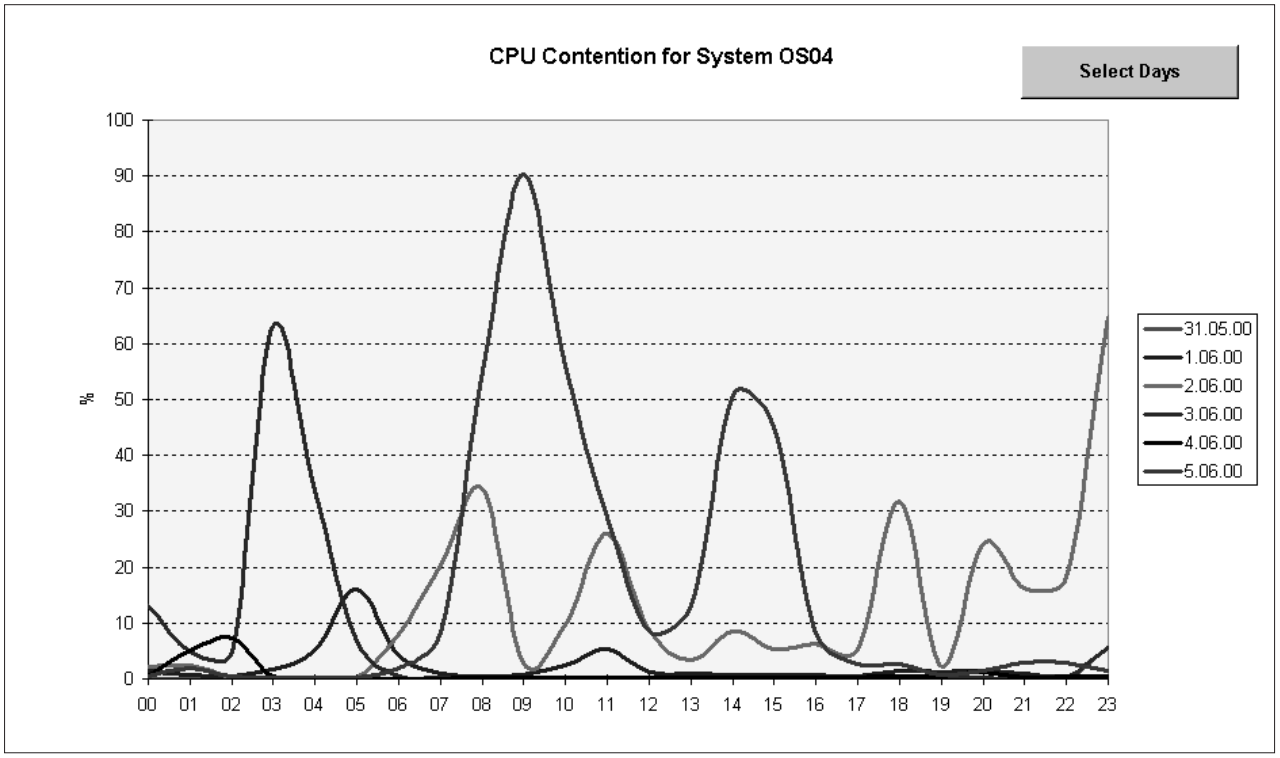


Figure 25. CPU Contention. Spreadsheet Reporter macro RMFY9OVW.XLS (System Overview Report - DaysCont)

The CPU Contention report helps you in understanding your processor capacity. It extends the *long-term* indicator in the Summary report (Figure 23 on page 42) and informs you about the contention of your system in some more detail.

Partition Data Report

```
                                    P A R T I T I O N   D A T A   R E P O R T
                                                                                                                PAGE    2
           z/OS V1R6              SYSTEM ID NP1             DATE 04/15/2004          INTERVAL 14.59.678
                                  RPT VERSION V1R5 RMF      TIME 09.30.00            CYCLE 1.000 SECONDS

MVS PARTITION NAME                    NP1                    NUMBER OF PHYSICAL PROCESSORS           9
IMAGE CAPACITY                        100                                   CP                      9
NUMBER OF CONFIGURED PARTITIONS         9                                   ICF                     0
WAIT COMPLETION                        NO
DISPATCH INTERVAL                 DYNAMIC

--------- PARTITION DATA -----------------  -- LOGICAL PARTITION PROCESSOR DATA --   -- AVERAGE PROCESSOR UTILIZATION PERCENTAGES --

                  ----MSU----  -CAPPING--  PROCESSOR-  ----DISPATCH TIME DATA----   LOGICAL PROCESSORS  --- PHYSICAL PROCESSORS ---
NAME       S   WGT  DEF    ACT  DEF   WLM%  NUM   TYPE   EFFECTIVE       TOTAL      EFFECTIVE    TOTAL  LPAR MGMT  EFFECTIVE  TOTAL

NP1        A    20  100     10  NO    62.2  1.2   CP    00.04.29.502  00.04.27.519       25.10    24.92      ****       3.33   3.30
NP2        A     1    0      1  YES    0.0    4   CP    00.00.22.680  00.00.22.083        0.75     0.73      ****       0.28   0.27
NP3        A    10    5      8  NO     3.3  1.0   CP    00.03.37.761  00.03.35.859       24.20    23.99      ****       2.69   2.67
NP4        A   300   95    155  NO     0.0  0.3   CP    01.12.08.405  01.12.06.405       80.18    80.15      ****      53.46  53.43
NP5        A   200   50     52  NO     0.0    4   CP    00.24.13.447  00.24.11.311       40.39    40.33      ****      17.95  17.92
CFC1       A   DED    0     32         0.0    1   CP    00.14.59.611  00.14.59.625       99.99    99.99      0.00      11.11  11.11
CFC2       A   DED    0      0         0.0    1   CP    00.00.00.000  00.00.00.000        0.00     0.00      0.00       0.00   0.00
*PHYSICAL*                                                             00.00.03.603                          0.04              0.04
                                                                      ------------  ------------                         ------ ------
TOTAL                                                                 01.59.51.408  01.59.46.408                          0.04      88.81  88.75

CB88       D
CB89       D
```

Figure 26. Monitor I CPU Activity Report - Partition Data Report Section

If your system is showing signs of CPU constraint (see previous report), and you are running as a *logical partition (LP)* in a PR/SM environment, check the Partition Data report to be sure that your LP is correctly configured. You would do this to allocate more CPU resource to a particular LP.

See Appendix A, “PR/SM LPAR Considerations” for further analysis.

Channel Path Activity Report

```
                                C H A N N E L   P A T H   A C T I V I T Y

           z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004         INTERVAL 14.59.996
                                  RPT VERSION V1R5 RMF      TIME 09.29.00           CYCLE 1.000 SECONDS

IODF = 01   CR-DATE: 07/10/04   CR-TIME: 21.00.01   ACT: POR           MODE: LPAR        CPMF: EXTENDED MODE

  B5 CNC_S Y    25.22  26.16                                  BD CNC_S Y    22.65  23.48
  B6 CNC_P     OFFLINE                                       BE CNC_S Y     0.00   0.01
  B7 CNC_S Y     0.01   0.01                                  BF CTC_S Y     0.40   1.48
  B8 CNC_S Y     0.00   0.00                                  C0 CNC_S Y     2.11   2.77
  B9 CNC_S Y     0.00   0.00                                  C1 CNC_S Y     0.00   0.00
  BA CNC_S Y     1.11   1.16                                  C2 CNC_S Y     0.00   0.00
  BB CNC_S Y     2.55   2.56                                  C3 CNC_P     OFFLINE
  BC CNC_S Y     4.63   9.21                                  C4 CNC_S Y     0.00   0.00

  C5 CNC_S Y     0.00   0.00                                  CD CFS   Y   ------   0.00
  C6 CNC_S Y    13.70  14.72                                  D0 CNC_S Y     2.03   2.68
  C7 CNC_S Y    33.26  33.26                                  D1 CNC_S Y     0.00   0.00
  C8 CNC_S Y     0.00   0.00                                  D2 CVC         0.00   0.00
  C9 CNC_S Y     4.40   8.84                                  D3 CNC_S Y    32.96  32.97
  CA CNC_S Y     0.00   0.00                                  D4 CNC_S Y    37.13  39.54
  CB CNC_? Y     0.00   0.00                                  D5 CNC_S Y     1.24   1.27
  CC CFS   Y   ------   0.00                                  D6 CNC_S Y     0.00   0.00
```

Figure 27. Monitor I Channel Path Activity Report

Note: In the past, it gave rules-of-thumb about channel utilization providing numbers up to 50%. For new channel types, for example, OSA-Express Gigabit Ethernet, you may see values for channel utilization close to 100% without any performance degradation. For more information, please refer to the *OSA-Express Gigabit Ethernet Performance Report*.

Tip:

Under

`http://www.ibm.com/servers/eserver/zseries/library/techpapers/`

in section **Networking & Connectivity**, you can find a document called *FICON and FICON Express Channel Performance* that helps you with understanding the performance characteristics of the different versions of FICON and FICON Express channels. This paper also explains in detail the benefits of FICON and FICON Express channels in both FC and FCV mode, discusses several channel configurations, compares ESCON and FICON technology and, last not least, recommends how to use FICON RMF information from various reports for I/O configuration performance measurement.

Report Analysis

Channel 54 has a partition utilization of 37.13 % and a total utilization of 39.54 %. This may not be a problem.

To find out which *logical control unit (LCU)* is using this channel, look in the I/O Queuing Activity report (see Figure 28 on page 49). From there, you can go on to check device response times.

See Chapter 4, “Analyzing I/O Activity” for further analysis.

I/O Queuing Activity Report

```
                                        I/O  Q U E U I N G  A C T I V I T Y

           z/OS V1R6              SYSTEM ID   SYS1          DATE 03/14/2004         INTERVAL 15.00.036
                                  RPT VERSION V1R5 RMF      TIME 16.30.00           CYCLE 1.000 SECONDS

TOTAL SAMPLES =    900   IODF = 01   CR-DATE: 05/10/2000   CR-TIME: 21.00.01     ACT: POR


      - INITIATIVE QUEUE -    ------- IOP UTILIZATION -------    -- % I/O REQUESTS RETRIED --    -------- RETRIES / SSCH ---------
 IOP     ACTIVITY    AVG Q     % IOP   I/O START    INTERRUPT          CP    DP    CU    DV            CP     DP     CU     DV
           RATE      LNGTH      BUSY      RATE         RATE       ALL  BUSY  BUSY  BUSY  BUSY     ALL  BUSY   BUSY   BUSY   BUSY

  00      1282.40     0.15     54.58      853.56       984.56     67.7  34.0  15.1  11.3   7.3    2.10  1.05   0.47   0.35   0.23
  02      1188.30     0.11     51.25      797.23       856.23     63.4  27.5   9.2  18.3   8.4    1.74  0.75   0.25   0.50   0.23

 SYS      2470.70     0.13     52.83     1650.79      1840.79     65.8  31.1  12.4  14.5   7.8    1.92  0.91   0.36   0.42   0.23


                                                                                                  DELAY
 LCU    CONTROL UNITS     DCM GROUP   CHAN     CHPID    % DP   % CU  CONTENTION      Q
                        MIN MAX DEF  PATHS     TAKEN    BUSY   BUSY     RATE       LNGTH

 0056   0100                          D4      58.953   42.50   1.22
                                      EC      57.117   43.34   0.90
        0101                          45      55.935   44.62   1.14
                                      28      67.282   28.32   1.39
                                       *      239.29   39.91   1.16      8.596      0.41

 0057   0140                          D4      61.322   42.20   0.98
                                      EC      60.601   41.87   0.82
        0141                          45      59.820   42.99   1.02
                                      28      74.294   24.96   1.15
                                       *      256.04   38.20   0.99      8.913      0.22
```

Figure 28. Monitor I I/O Queuing Activity Report

Report Analysis

Channel path D4 is connected to LCU 0056 - 0059. The corresponding devices can be seen in the Device Activity report.

DASD Activity Report

```
                                D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y

            z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004          INTERVAL 14.59.996
                                   RPT VERSION V1R5 RMF      TIME 09.29.00            CYCLE 1.000 SECONDS

TOTAL SAMPLES =  1,807   IODF = 12  CR-DATE: 03/01/04   CR-TIME: 10.23.17

                                        DEVICE   AVG  AVG   AVG  AVG  AVG   AVG   AVG  AVG    %      %      %     AVG     %      %
STORAGE  DEV  DEVICE   VOLUME PAV  LCU  ACTIVITY RESP IOSQ  DPB  CUB  DB    PEND  DISC CONN   DEV    DEV    DEV   NUMBER ANY    MT
 GROUP   NUM  TYPE     SERIAL          RATE     TIME TIME  DLY  DLY  DLY   TIME  TIME TIME   CONN   UTIL   RESV  ALLOC  ALLOC  PEND

SGGDG    0340 33903    IDP386      0056   0.139  18.4  0.0   0.0  0.0  0.0   0.6   0.7 17.1   0.24   0.25   0.0   1.3   100.0   0.0
SGGDG    0341 33903    IDP387      0056   0.001   0.9  0.0   0.0  0.0  0.0   0.3   0.0  0.6   0.00   0.00   0.0   0.0   100.0   0.0
SGGDG    0342 33903    IDP388      0056   3.311  22.4  0.0   0.0  0.0  0.2   0.6   0.2 21.5   7.12   7.18   0.0   1.0   100.0   0.0
SGGDG    0343 33903    IDP389      0056   0.001   0.9  0.0   0.0  0.0  0.0   0.3   0.0  0.6   0.00   0.00   0.0   0.0   100.0   0.0
SGCCBS   0344 33903    CCB178      0056   0.150  12.8  0.0   0.0  0.0  0.0   0.5   9.9  2.4   0.04   0.18   0.0   2.0   100.0   0.0
SGCCBS   0345 33903    CCB179      0056   3.703  23.2 11.0   0.0  0.0  0.0   0.3   4.8  7.1   2.63   4.40   0.0   3.7   100.0   0.0
SGDBSWN  0346 33903    DBSW11      0056   0.959   2.3  0.0   0.0  0.0  0.2   0.5   0.0  1.8   0.17   0.17   0.0  11.0   100.0   0.0
SGDBSWN  0347 33903    DBSW12      0056   0.003   0.9  0.0   0.0  0.0  0.0   0.3   0.0  0.6   0.00   0.00   0.0   2.0   100.0   0.0
SGSW     0348 33903    IDP302      0056   0.003   0.9  0.0   0.0  0.0  0.0   0.3   0.0  0.6   0.00   0.00   0.0   0.0   100.0   0.0
         0349 33903    PLEXP1      0056   0.002   0.7  0.0   0.0  0.0  0.0   0.2   0.1  0.4   0.00   0.00   0.0   0.0   100.0   0.0
:
:
SGSW     037E 33903    IDP303      0056   0.003   0.8  0.0   0.0  0.0  0.0   0.3   0.0  0.5   0.00   0.00   0.0   0.0   100.0   0.0
SGCRITC  037F 33903    IDP076      0056   0.003   0.8  0.0   0.0  0.0  0.0   0.2   0.0  0.6   0.00   0.00   0.0   1.0   100.0   0.0

                       LCU         0056  44.580  13.0  0.5   0.0  0.0  0.2   0.7   3.6  8.2   0.57   0.82   0.0   137   100.0   0.0
```

Figure 29. Monitor I DASD Activity Report

Rules-of-Thumb

AVG RESP TIME

Typical values by DASD device type:

- **Uncached 3380** 25 ms
- **Uncached 3390** 20 ms
- **Uncached 9345** 18 ms
- **Cached 3380** 12 ms
- **Cached 3390** 10 ms
- **Cached 9345** 9 ms

This table seems to have a slightly nostalgic look because most of these device may not be found in any installation. You might take it as a historical view about the development of DASD devices. Today, cached 3390 is the current standard, at least for all those installations which have not started with the migration of their I/O subsystem to the *Enterprise Storage Server*® (ESS) which can provide significant better I/O performance. For details, please refer to "Improving I/O Performance with the Enterprise Storage Server" on page 85.

If your response times significantly exceed these, for devices with high activity rates and/or important workloads, you should investigate further.

Note that some workloads are prone to higher response times, which may still be quite acceptable (for example, DB2 sequential prefetch, page/swap I/O).

Tape LCU AVG RESP TIME

Look at the ratio of `LCU DISC + PEND` time to `CONN` time. When the `DISC+PEND` exceeds `CONN` for extended periods of time, this is an indication of channel and/or control unit contention.

Report Analysis

Several devices 340, 342, 344, or 345 have response times that are far above typical values and need further investigation. (Which workloads are using these devices? How much are they being delayed? Are these response times typical for my system?).

See Chapter 4, "Analyzing I/O Activity" for further analysis.

RMF DASD System Summary

System Id:	OS04	Operating System:	OS/390 02.10.00	Report Range(seconds):	899.628
Reporting Date:	6.05.00	Reporting Time:	09.29.00	Report Range(hh.mm.ss):	14.59.628

Overall	Total Number of Logical Control Units	116	
	Total Number of DASDs	6588	
	Installed Capacity	18458.12	GB
	Total Activity Rate	1990.76	I/O per sec
	Access Density	0.11	
Times	Connect Time	3.11	ms
	Disconnect Time	2.63	ms
	Service Time	5.74	ms
	Pending Time	0.60	ms
	IOS Queue Time	0.55	ms
	Response Time	6.90	ms
	Response Time / Service Time	1.20	
Intensities	I/O Intensity	13732.83	ms / s
	Service Time Intensity	11426.68	ms / s
	Path Intensity	6188.40	ms / s
DASD Skew	Average Service Time Intensity	1.73	ms / s
	Highest Service Time Intensity	838.88	ms / s
	Ratio: Highest/Average (DASD Skew)	483.65	
	Typical DASD Skew	47.89	
	Typical Range (low / high)	23.70	96.77

Installed Capacity Estimation

Types	GB each	Number	Capacity
3380D	0.63	0	0
3380E	1.26	0	0
3380J	0.63	0	0
3380K	1.89	0	0
33901	0.94	0	0
33902	1.89	0	0
33903	2.83	6236	17647.88
33906	8.44	0	0
33909	8.44	96	810.24
9345	1.50	0	0
Unknown	NA	256	NA
Additional	NA	NA	0
Total		6588	18458.12

Figure 30. DASD Summary. Spreadsheet Reporter macro RMFR9DAS.XLS (DASD Activity Report - System)

In the DASD Summary report, you see at a glance the key data of your DASD subsystem: performance values as well as capacity data. If you feel that you might see some more details, you can have a look on the ten most busy DASD volumes in your system:

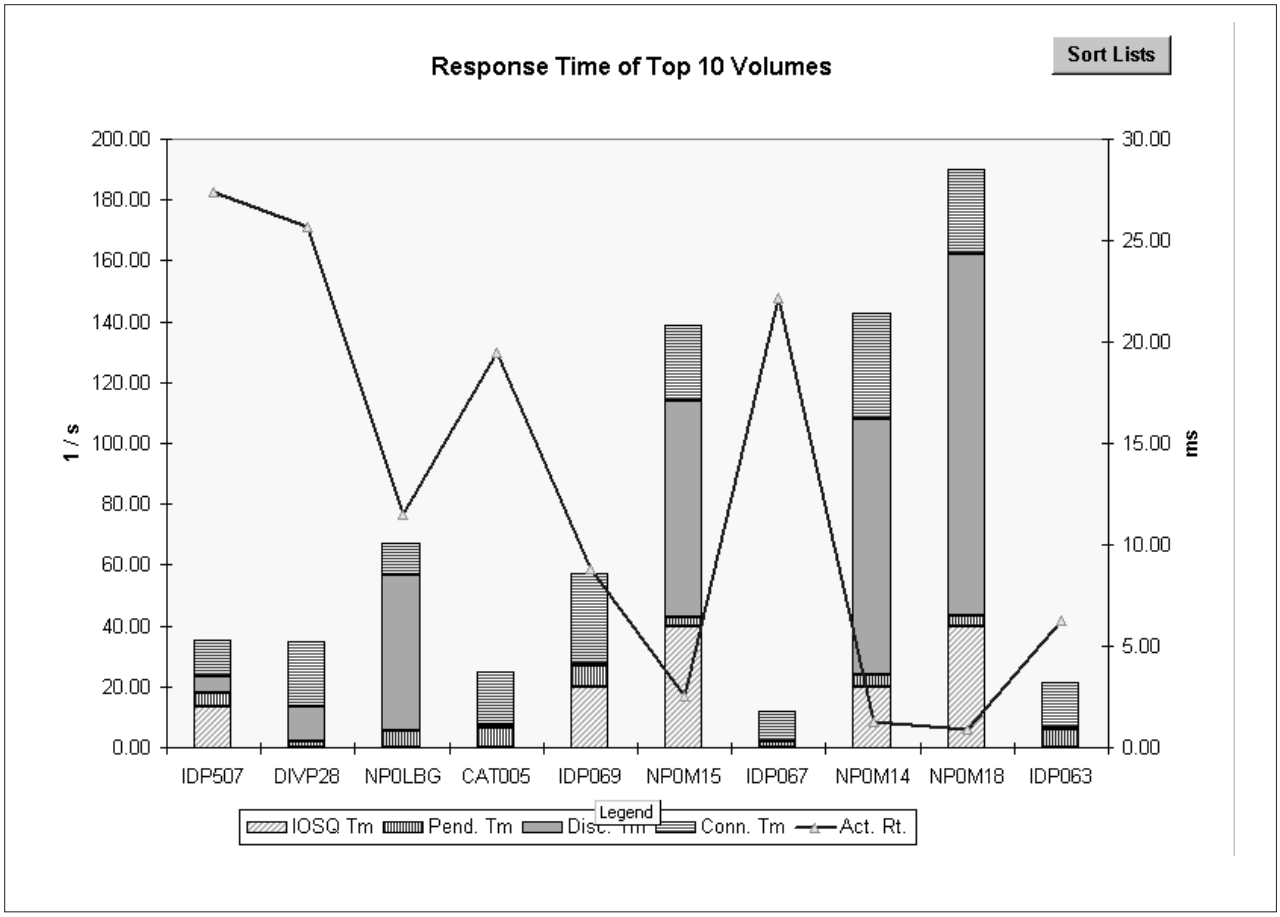


Figure 31. Response Time of Top-10 Volumes. Spreadsheet Reporter macro RMFR9DAS.XLS (DASD Activity Report - Top10RT)

Paging Activity Report - Page 3

```
                                        P A G I N G   A C T I V I T Y

          z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004         INTERVAL 14.59.996
                                 RPT VERSION V1R5 RMF      TIME 09.29.00           CYCLE 1.000 SECONDS

OPT = IEAOPT02  MODE=ESAME              S W A P   P L A C E M E N T   A C T I V I T Y

                        *------ AUX STORAGE -------* *---LOGICAL SWAP--*  *---------------- EXPANDED STORAGE ----------------*

                                  AUX                AUX STOR                                                MIGRATED                LOG SWAP
                                  STOR     AUX STOR     VIA                    LOG SWAP   EXP STOR   EXP STOR    FROM     EXP STOR   /EXP STOR
                        TOTAL     TOTAL    DIRECT    TRANSITION LOG SWAP     EFFECTIVE  DIRECT     TOTAL     EXP STOR  EFFECTIVE  EFFECTIVE

TERMINAL      CT      1,558         0        0          0      1,558      1,535         0         23          0        23       1,558
INPUT/OUTPUT  RT       1.73      0.00     0.00       0.00       1.73       1.71      0.00       0.03       0.00      0.03        1.73
WAIT           %      23.2%     0.0%     0.0%       0.0%     100.0%      98.5%     0.0%       1.5%       0.0%     100.0%     100.0%

LONG          CT      2,081         0        0          0      2,081      2,060         0         21          0        21       2,081
WAIT          RT       2.31      0.00     0.00       0.00       2.31       2.29      0.00       0.02       0.00      0.02        2.31
               %      30.9%     0.0%     0.0%       0.0%     100.0%      99.0%     0.0%       1.0%       0.0%     100.0%     100.0%

DETECTED      CT      2,918         1        0          1      2,918      2,894         0         24          1        23       2,917
WAIT          RT       3.24      0.00     0.00       0.00       3.24       3.22      0.00       0.03       0.00      0.03        3.24
               %      43.4%     0.0%     0.0%     100.0%     100.0%      99.2%     0.0%       0.8%       4.2%     95.8%     100.0%

UNILATERAL    CT          6         0        0          0          6          6         0          0          0         0           6
              RT       0.01      0.00     0.00       0.00       0.01       0.01      0.00       0.00       0.00      0.00        0.01
               %       0.1%     0.0%     0.0%       0.0%     100.0%     100.0%     0.0%       0.0%       0.0%      0.0%     100.0%

EXCHANGE ON   CT          3         0        0          0          3          3         0          0          0         0           3
RECOMMENDA-   RT       0.00      0.00     0.00       0.00       0.00       0.00      0.00       0.00       0.00      0.00        0.00
TION VALUE     %       0.0%     0.0%     0.0%       0.0%     100.0%     100.0%     0.0%       0.0%       0.0%      0.0%     100.0%

TRANSITION    CT          8         0        0          0          0          0         8          8          0         8           8
TO NON-       RT       0.01      0.00     0.00       0.00       0.00       0.00      0.01       0.01       0.00      0.01        0.01
SWAPPABLE      %       0.1%     0.0%     0.0%       0.0%       0.0%       0.0%    100.0%     100.0%       0.0%    100.0%     100.0%

OMVS          CT        153         0        0          0        153        150         0          3          0         3         153
INPUT         RT       0.17      0.00     0.00       0.00       0.17       0.17      0.00       0.00       0.00      0.00        0.17
WAIT           %       2.3%     0.0%     0.0%       0.0%     100.0%      98.0%     0.0%       2.0%       0.0%     100.0%     100.0%

TOTAL         CT      6,727         1        0          1      6,719      6,648         8         79          1        78       6,726
              RT       7.48      0.00     0.00       0.00       7.47       7.39      0.01       0.09       0.00      0.09        7.48
               %     100.0%     0.0%     0.0%     100.0%      99.9%      98.9%     0.1%       1.2%       1.3%     98.7%     100.0%


AUXILIARY STORAGE - AVERAGE PAGES PER SWAP OUT -         1       AVERAGE PAGES PER SWAP IN -       257
OCCURRENCES OF TERMINAL OUTPUT WAIT -        87
```

Figure 32. Monitor I Paging Activity Report - Page 3

Rule-of-Thumb

LOG SWAP/EXP STOR EFFECTIVE

This is the percentage of all swaps that are satisfied from processor storage. You want the TOTAL value under this column to be a high number (e.g. over 95%) to keep swap delay times low.

The percentage under TERMINAL INPUT/OUTPUT WAIT should also be high, as this is where most TSO swaps are reflected.

Report Analysis

The LOG SWAP/EXP STOR EFFECTIVE values for

- TOTAL: 100%
- TERMINAL INPUT/OUTPUT WAIT: 100%

are perfect.

See Chapter 5, "Analyzing Processor Storage Activity" for further analysis.

Page/Swap Data Set Activity Report

```
                                P A G E / S W A P  D A T A S E T  A C T I V I T Y

           z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004         INTERVAL 14.59.996
                                  RPT VERSION V1R5 RMF      TIME 09.29.00           CYCLE 1.000 SECONDS

NUMBER OF SAMPLES =     898                    PAGE DATA SET USAGE
                                              -------------------------

PAGE                                                                    %    PAGE                 V
SPACE   VOLUME  DEV  DEVICE    SLOTS  ---- SLOTS USED ---  BAD        IN   TRANS  NUMBER  PAGES  I
TYPE    SERIAL  NUM  TYPE      ALLOC    MIN    MAX   AVG   SLOTS      USE  TIME   IO REQ  XFER'D O  DATA SET NAME

PLPA    PAGP11  0378 33903     45000  16466  16466  16466      0     0.11  0.200       6       5    PAGE.OS04.APLPA

COMMON  PAGP10  0377 33903     21240   5345   5378   5355      0     0.11  0.027      38      37    PAGE.OS04.ACOMMON

LOCAL   PAGP12  0379 33903    566820  67670  70965  69140      0     2.67  0.004     711   5,697 N  PAGE.OS04.LOCAL01

LOCAL   PAGP13  037A 33903    566820  65887  69007  67190      0     3.01  0.005     771   5,711 N  PAGE.OS04.LOCAL02

LOCAL   PAGP14  037B 33903    566820  66611  69723  67928      0     3.01  0.005     821   5,625 Y  PAGE.OS04.LOCAL03

LOCAL   PAGP15  038D 33903    566820  66296  69563  67618      0     3.23  0.005     726   5,523 Y  PAGE.OS04.LOCAL04

LOCAL   PAGP16  039F 33903    566820  66390  69311  67560      0     3.12  0.005     766   5,548 Y  PAGE.OS04.LOCAL05
```

Figure 33. Monitor I Page/Swap Data Set Activity Report

Rules-of-Thumb

% IN USE

This is really the % busy for the data set. Above 30% you may start to see response times increase. However, you also need to consider other active data sets on the same volume; check the total device utilization on the device activity report.

PAGES XFER'D

The total number of pages transferred to and from a given page/swap dataset. Divide this number by the number of INTERVAL seconds to get the rate in pages/second. Above 30 pages/second for a single data set may mean it is time to add another page pack, dedicate the packs to paging (or better yet, look into more processor storage).

Report Analysis

If we check the two highest values in the report, we get:

- % IN USE: 3.23% - no problem seen
- PAGES XFER'D: 5711 - this is equivalent to 6 pages/second.

See Chapter 5, "Analyzing Processor Storage Activity" for further analysis.

Workload Activity Report

```
REPORT BY: POLICY=SPECIAL    WORKLOAD=ONLINE     SERVICE CLASS=TSO        RESOURCE GROUP=*NONE      PERIOD=1 IMPORTANCE=2

TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT   --DASD I/O--   ----SERVICE----   --SERVICE TIMES--   PAGE-IN RATES    ----STORAGE----
AVG      1.25   ACTUAL                 96   SSCHRT  10.0   IOC      30100    TCB          32.3   SINGLE   16.72   AVERAGE  124.31
MPL      0.86   EXECUTION              96   RESP    15.4   CPU      357.3K   SRB           5.1   BLOCK     6.23   TOTAL    107.44
ENDED    4,945  QUEUED                  0   CONN     2.8   MSO      203.7K   RCT           0.0   SHARED    0.00   CENTRAL   75.66
END/S     8.21  R/S AFFINITY            0   DISC     2.5   SRB      56334    IIT           0.1   HSP       0.00   EXPAND    31.78
#SWAPS   4,952  INELIGIBLE              0   Q+PEND  10.2   TOT      647.4K   HST           0.0   HSP MISS  0.00
EXCTD       0   CONVERSION              0   IOSQ     0.0   /SEC     1,075    IFA           0.0   EXP SNGL   100   SHARED     0.00
AVG ENC  0.00   STD DEV                119                                  APPL% CP      15.0   EXP BLK   0.00
REM ENC  0.00                                             ABSRPTN  1,244    APPL% IFACP    0.0   EXP SHR   0.00
MS ENC   0.00                                             TRX SERV   856    APPL% IFA      0.0

...

REPORT BY: POLICY=SPECIAL

TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT   --DASD I/O--   ---SERVICE----   --SERVICE TIMES--   PAGE-IN RATES    ----STORAGE----
AVG     60.79   ACTUAL                749   SSCHRT 368.7   IOC      1106K   TCB          509.9   SINGLE    0.88   AVERAGE  152.18
MPL     55.66   EXECUTION             652   RESP    12.4   CPU      5630K   SRB          124.9   BLOCK     0.48   TOTAL     8,471
ENDED  14,418   QUEUED                 96   CONN     2.4   MSO      9228K   RCT            0.2   SHARED    0.00   CENTRAL   4,904
END/S   23.95   R/S AFFINITY            0   DISC     2.1   SRB      1380K   IIT            1.3   HSP       0.00   EXPAND    3,568
#SWAPS   14.8K  INELIGIBLE              0   Q+PEND   7.9   TOT     17343K   HST            0.0   HSP MISS  0.00
EXCTD       0   CONVERSION              0   IOSQ     0.0   /SEC   28,819    IFA            0.0   EXP SNGL  9.50   SHARED     0.00
AVG ENC  0.00   STD DEV             5.769                                  APPL% CP     105.4   EXP BLK   0.46
REM ENC  0.00                                             ABSRPTN   517    APPL% IFACP    0.0   EXP SHR   0.00
MS ENC   0.00                                             TRX SERV  474    APPL% IFA      0.0
```

Figure 34. Workload Activity Report

Rules-of-Thumb

APPL% CP

This value tells you how much CPU time was captured for the workload in a given service class or performance group. You can use this information for several purposes:

- Compute a *capture ratio (CR)* for your system
- Input to capacity planning
- Check for a single-CP constraint

Remember that the APPL% CP is given as percent of a single engine, so you will need to divide by the number of engines if your system has more than one.

Most MVS systems today have capture ratios greater than 80%. If yours is below that, you may want to investigate further.

PAGE-IN RATES

For any workload that is sensitive to paging (for example CICS), check the SINGLE and BLOCK paging values to be sure that no paging is taking place. The amount of paging that these workloads can sustain is up to you, but the best answer is probably zero.

Report Analysis

- Calculation of the capture ratio:

```
                    APPL% CP          105.4
Capture ratio = ---------------- = ----------- = 0.71
                MVS BUSY % * #CP    73.88 * 2
```

  This value is somewhat low.

  If your system is running in an PR/SM environment, the calculation has to be performed using LPAR BUSY and the number of logical processors.

- The paging rates for TSO are

```
SINGLE  16.72 page-ins per second
BLOCK   6.23 page-ins per second
```

  This may be worth further investigation.

See Chapter 3, “Analyzing Processor Activity” and Chapter 5, “Analyzing Processor Storage Activity” for further analysis.

Virtual Storage Report

```
                                V I R T U A L   S T O R A G E   A C T I V I T Y
                                                                                PAGE 1

          z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004         INTERVAL 14.59.996
                                 RPT VERSION V1R5 RMF      TIME 09.29.00           CYCLE 1.000 SECONDS


                                          COMMON STORAGE SUMMARY
NUMBER OF SAMPLES      90

      STATIC STORAGE MAP                                          ALLOCATED CSA/SQA
  AREA        ADDRESS   SIZE            ------------ BELOW 16M --------------  ------- EXTENDED (ABOVE 16M) --------
EPVT         17600000  1674M              MIN             MAX             AVG    MIN             MAX             AVG
ECSA          ADA1000   200M         SQA  2076K 09.34.31  2100K 09.35.51  2084K  59.5M 09.35.01  59.7M 09.30.41  59.6M
EMLPA         A984000  4212K         CSA  1744K 09.29.11  1748K 09.29.01  1747K   151M 09.31.42   152M 09.40.31   152M
EFLPA         A981000    12K
EPLPA         6B7B000  62.0M         ALLOCATED CSA BY KEY
ESQA          26AD000  68.8M          0    952K 09.29.11   956K 09.29.01   952K  18.4M 09.34.51  18.4M 09.34.21  18.4M
ENUC          1000000  22.7M          1     44K 09.29.01    44K 09.29.01    44K  1168K 09.29.01  1180K 09.41.11  1170K
----- 16 MEG BOUNDARY ------          2     40K 09.29.01    40K 09.29.01    40K  5024K 09.29.01  5024K 09.29.01  5024K
NUCLEUS        FC9000   220K          3      0K 09.29.01     0K              0K    60K 09.29.01    64K 09.33.41    60K
SQA            D2F000  2664K          4     28K 09.29.01    28K 09.29.01    28K  7576K 09.31.21  7976K 09.42.31  7673K
PLPA           B23000  2096K          5      8K 09.29.01     8K 09.29.01     8K  3896K 09.29.01  3904K 09.37.21  3899K
FLPA           B22000     4K          6    180K 09.29.01   184K 09.32.11   183K  89.6M 09.30.21  90.1M 09.30.01  89.8M
MLPA           ABB000   412K          7    432K 09.29.01   432K 09.29.01   432K  24.9M 09.33.11  25.1M 09.40.21  25.0M
CSA            800000  2796K        8-F    60K 09.29.01    60K 09.29.01    60K   976K 09.29.01   984K 09.39.31   978K
PRIVATE          1000  8188K
PSA                 0     4K         SQA EXPANSION INTO CSA
                                          0K 09.29.01     0K              0K     0K 09.29.01     0K              0K

PLPA INTERMODULE SPACE -    6K IN PLPA AND  187K IN EPLPA
PLPA SPACE REDUNDANT WITH MLPA/FLPA -    2K IN PLPA AND 3053K IN EPLPA


                     ---------- BELOW 16M ----------------  ---------------- ABOVE 16M -----------
                       MIN             MAX             AVG    MIN             MAX             AVG
CSA
 FREE PAGES (BYTES)  1048K 09.29.01  1052K 09.29.11  1048K  48.3M 09.40.31  49.2M 09.31.42  48.8M
 LARGEST FREE BLOCK  1048K 09.29.01  1052K 09.29.11  1048K  48.1M 09.29.01  48.1M 09.29.01  48.1M
 ALLOCATED AREA SIZE 1744K 09.29.11  1748K 09.29.01  1747K   152M 09.29.01   152M 09.29.01   152M
SQA
 FREE PAGES (BYTES)   564K 09.35.51   588K 09.34.31   579K  9352K 09.30.41  9568K 09.35.01  9447K
 LARGEST FREE BLOCK   360K 09.29.01   360K 09.29.01   360K  8116K 09.29.01  8116K 09.29.01  8116K
 ALLOCATED AREA SIZE 2304K 09.29.01  2304K 09.29.01  2304K  68.8M 09.29.01  68.8M 09.29.01  68.8M

MAXIMUM POSSIBLE USER REGION -  7996K BELOW AND 1666M ABOVE
```

Figure 35. Monitor I Virtual Storage Activity Report - Common Storage Summary

Rules-of-Thumb

Maximum Allocated Percent

Calculate this value for the CSA, the ECSA, the SQA, and the ESQA. For example, for the CSA:

$$\text{Maximum Allocated Percent(CSA)} = \frac{\text{CSA ALLOCATED AREA SIZE}}{\text{CSA SIZE}} * 100$$

For CSA and ECSA, this value should be less than 65%.

The virtual storage that is most volatile is the SQA and CSA, both above and below the 16MB line. This is a number that is readily tracked on the Virtual Storage report on either a daily or weekly basis.

RMF calculates the size of this area as the difference between the highest and lowest address occupied by allocated storage, and includes all free blocks that lie between allocated blocks. Significant segmentation causes this number to be much larger than the amount of storage actually used.

If the size allocated for ESQA is too small or is used up, the system attempts to steal pages from the ECSA. When both areas are full, the system allocates space from the SQA and CSA below the 16MB line. SRM will attempt to reduce the demand for resources. If no storage is available, the result is a system failure.

If the size of SQA below 16M is too small, additional data will spill into the CSA below the 16MB line. If no storage is available, the result is a system failure.

- Is CSA/ECSA allocated percent high?
  - Review ECSA, CSA, ESQA, and SQA sizes
  - Increase CSA size: cleanup LPA, split address spaces (MRO for example), use Version/Release products that exploit ESA, or convert applications to exploit ESA (for example to Cobol II)
- Is SQA/ESQA allocated percent low?
  - ESQA or SQA may be too large.
- Is `SQA EXPANSION INTO CSA` low?
  - SQA size may be too large; you may be able to decrease SQA size and increase CSA size.

Using the Monitor III Common Storage report is a very good way to understand which AS has the storage.

Report Analysis

The maximum allocated percent for the CSA is:

```
1744K
----- * 100 = 62.4%
2796K
```

That concludes there is no problem in this area.

The SQA EXPANSION INTO CSA is zero, that means the SQA is large enough.

Where Do You Go from Here?

By now you will have found the major cause of your performance problem. The following chapters will go into more detail to help you further.

As you know, performance tuning is an iterative process and bear in mind:

- Take a base measurement as your starting point
- Only change one thing at a time
- Predict how your key measurements will change
- Measure your change against your expectation

Chapter 3. Analyzing Processor Activity

Let's Analyze Your Processors

This chapter discusses how to analyze a processor problem. It takes the following approach:

- What RMF values can you look at in a performance problem situation?
- What do these fields mean?
- What can you do to solve the problem?

  We make recommendations for two cases:

  – An application being *delayed* by higher priority work

  – An application *using* excessive CPU, or overall CPU under stress

As you continue through this chapter, bear in mind that tuning the processor is generally a high-effort/low-return activity. Quick fixes are rare. However, we will discuss the tuning activities that can apply.

Do You Have a Processor Problem?

There are many views on what constitutes a processor performance problem. You should conclude that you have a processor performance problem because your:

- Service level objectives are being exceeded.
- Users are complaining about slow response.
- AND, the CPU indicators (discussed in the following section) show signs of stress. Or better still, you see a processor delay directly in Monitor III.

We suggest there are two main ways to approach this problem analysis, depending on how you are alerted:

- If a user or group of users complains about response times then start with Monitor III.
- If your normal daily or weekly performance tracking procedures show a trend or problem developing, then start with Monitor I.

Table 2 shows for each monitor, which fields in which RMF reports can detect and resolve a processor performance problem.

Table 2. Processor Indicators in Monitor III and Monitor I

	REPORTS	DELAY Fields	USING Fields
Monitor III	GROUP	Avg Delay PROC	AVG USG PROC
	DELAY	%Delayed for PRC	
	WFEX	Speed *PROC	
	PROC	DLY%	USG% Appl% CP
	CPC	WLM Capping %	
Monitor I	CPU Activity		BUSY TIME PERCENTAGE
	Partition Data		PHYSICAL PROCESSOR TOTAL

Monitor III Indicators

This section lists the Monitor III indicators that may be used to alert you to a potential processor problem.

GROUP Report

```
                        RMF V1R5  Group Response Time
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

Class: BATCH         Period: 1    Description: Primary Batch
Primary Response Time Component:  Using the processor

                                                    TRANS    --- Response Time ----
WFL    Users    Frames   Vector   EXCP   PgIn   Ended    -- Ended TRANS-(Sec) -
 %    TOT  ACT   %ACT     UTIL    Rate   Rate   Rate      WAIT   EXECUT  ACTUAL
 60    30    3     25        0     2.1    0.5     0.0     830.5   152.7   983.3

                          -AVG USG-   -------------Average Delay--------------
                    Total   PROC  DEV   PROC   DEV  STOR  SUBS  OPER   ENQ OTHER
Average Users       3.270   1.89 0.10   0.74  0.06  0.01  0.00  0.00  0.51  0.00
Response Time ACT   152.7   88.3 4.67   34.6  2.80  0.47  0.00  0.00  23.8  0.00

                     ---STOR Delay---  ---OUTR Swap Reason---  ---SUBS Delay---
                     Page  Swap  OUTR    TI    TO    LW    XS   JES   HSM   XCF
Average Users        0.01  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
Response Time ACT    0.47  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
```

Figure 36. Monitor III Group Response Time Report

This report presents information for a specific service class period.

DELAY Indicator

Field: Average Delay PROC for Response Time ACT

Description: This is a component of average response time, showing the average number of seconds, transactions active during the range period in the specific service class period, were delayed by the lack of CPU. It is the CPU queue time.

Guideline: Look at the different Average Delay values, in the line Response Time ACT. If the PROC value is the largest, continue your processor investigation. Look at the different service classes (coming back to the Primary Panel and choosing option 3) and start with the service class showing the largest value.

Problem Area: This points out processor delays for the service class listed.

Potential Solution:
- See “DELAY for Processor: Guidelines for Improvement” on page 73
- See “Processor USING: Trimming Activities” on page 76

USING Indicator

Field: AVG USG PROC for Response Time ACT

Description: This is a component of average response time, namely the average number of seconds, transactions belonging to the service class were using the CPU during the range period.

Guideline: Check all the service classes and start with that one showing the average largest processor consumption.

Problem Area: This points out processor consumption.

Potential Solution:
- See "Processor USING: Trimming Activities" on page 76

DELAY Report

```
                          RMF V1R5  Delay Report                  Line 1 of 44
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

            Service     WFL USG DLY IDL UKN ----- % Delayed for -----  Primary
Name     CX Class    Cr  %   %   %   %   %  PRC  DEV STR SUB OPR  ENQ  Reason

BHOLEQB  B   BATCH       0  51   0  49   0   0    0   0   0  51    0  SYSDSN
CATALOG  S   SYSSTC      0   0   3   0  97   0    0   3   0   0    0  LOCL
BAJU     T   TSO        14   2  12  85   1   0    0  12   0   0    0  LOCL
BHOLNAM4 B   BATCH      62  31  19   0  32  18    0   1   0   0    0  BHOLPRO2
JES2     S   SYSSTC     67  13   7   0  83   1    6   0   0   0    0  SYSPAG
BHOLPRO1 B   BATCH      72  72  28   0   0  28    0   0   0   0    0  BHOLPRO2
BHOLPRO2 B   BATCH      74  74  26   0   0  26    0   0   0   0    0  BHOLPRO1
RMF      S   SYSSTC     75   3   1   0  96   1    0   0   0   0    0  BHOLPRO2
*MASTER* S   SYSTEM     92  23   2   0  76   0    2   0   0   0    0  SYSPAG
BHOL     T   TSO        97  29   1  67   3   1    0   0   0   0    0  CONSOLE
CONSOLE  S   SYSSTC    100   2   0   0  98   0    0   0   0   0    0
BHOLSTO9 B   BATCH     100   1   0  99   0   0    0   0   0   0    0
VTMLCL   S   SYSSTC    100   1   0   0  99   0    0   0   0   0    0
```

Figure 37. Monitor III Delay Report

This is a good report to look at for a quick snapshot of 'Who is delayed' for CPU and for how long. The report is sorted by workflow percentage (WFL % column) in ascending order. Check this report to see if any priority work is being delayed.

DELAY Indicator

Field: % Delayed for PRC

Description: This gives for each address space (AS), the percentage of the measured time that a transaction running in that AS was delayed by the lack of CPU.

Guideline: Look at the largest %Delayed for PRC value. Compare it with the DLY % value of the same job. If the PRC value is the main part of the delay, look at this job.

You can put the cursor on the line describing the job that may be a problem and press ENTER: a new panel is displayed (see Figure 43 on page 73) that lists the main cause of the delay and suggests some possible actions to reduce the delay.

Problem Area: This points out a delay for processing.

Potential Solution:

- See "DELAY for Processor: Guidelines for Improvement" on page 73
- See "Processor USING: Trimming Activities" on page 76

WFEX Report

```
                    RMF V1R5  Workflow/Exceptions               Line 1 of 4
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

--------------------------- Speed (Workflow) ---------------------------------
           Speed of 100 = Maximum, 0 = Stopped         Average CPU Util: 100 %
Name       Users Active      Speed          Name       Users Active      Speed
*SYSTEM       41      4         63
ALL TSO        2      0         70
ALL STC       32      1         79
ALL BATCH      7      3         59
*PROC         24      3         73
*DEV           6      1         33
*MASTER*       1      0         92
------------------------------ Exceptions ------------------------------------
Name       Reason        Critical val. Possible cause or action
*ECSA*     SECS% > 85     90.6 %       System ECSA usage 91 %.
*STOR      TSQAO > 0      784K bytes   SQA overflow into CSA 784K.
BHOLEQB    ENQ -SYSDSN    51.0 % delay ENQ.TEST.DAT
```

Figure 38. Monitor III Workflow/Exceptions Report

DELAY Indicator

Field: *PROC Speed (Workflow)

Description: This is a good global indicator for CPU performance. If equal to 100%, no work has been delayed by the CPU. If equal to 50%, this means that half of the attempts to use CPU were denied because the CPU was busy (that is, five out of ten ready AS were being delayed). This field does not deal with how busy the CPU is, but rather the amount of delay that exists.

Guideline: If this value is low (less than 40%), it may be due to CPU delay. If this value is greater than 80%, there is no CPU problem. Between these two values, it depends on the characteristics of the environment.

Problem Area: This points out a delay for processing.

Potential Solution:
- See "DELAY for Processor: Guidelines for Improvement" on page 73
- See "Processor USING: Trimming Activities" on page 76

PROC Report

```
                        RMF V1R5  Processor Delays                Line 1 of 19
Command ===>                                                   Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

            Service  DLY USG  Appl EAppl ----------- Holding Job(s) -----------
Jobname   CX Class     %   %     %     %  %  Name      %  Name      %  Name

#STONE1   B  NRPRIME   41  60  55.1  62.1  18 D10PAV1   15 SAYLEKR   12 DAVEP
SAYLEKR   B  NRPRIME   24  21  20.7  20.7  13 DB2MDIST  12 #STONE1    9 CATALOG
SAYLEKQ   B  NRPRIME   22  14  14.8  14.8  16 D10PAV1   11 SAYLEKR    8 TREVORJ
RSTSHYS0  B  NRPRIME   15  24  18.6  18.6   9 DAVEP      8 #STONE1    7 GRS
CHARLESR  B  NRPRIME   15   7   2.9   4.2   8 D10PAV1    6 SAYLEKR    5 DAVEP
LUCKYSMP  B  NRPRIME    6   8   4.1   4.1   5 DAVEP      4 STEVEB     2 DSEAGER
DFHSM     S  SYSSTC     5  17  13.0  13.0   3 D10PAV1    3 #STONE1    2 *MASTER*
APETER    B  NRPRIME    5   5   3.7   3.7   3 DB2MDIST   3 SAYLEK     2 #STONE1
MARYPATM  B  NRPRIME    5   2   2.2   3.1   2 GRS        2 DAVEP      2 #STONE1
D10PAV1   T  TSOPRIME   4  25  23.5  23.5   2 BKELLER    2 DFHSM      2 RHANSON
DAVEP     T  TSOPRIME   4  25  14.9  14.9   3 DB2MDIST   2 DSEAGER    2 LUCKYSMP
HJPB88E   B  NRPRIME    4   3   1.1   1.1   3 #STONE1    2 GRS        2 DB2MDIST
```

Figure 39. Monitor III Processor Delays Report

This report displays all AS waiting for or using the processor during the range period. The report is sorted by descending overall delay percentages: the first line is the job you need to look at first.

DELAY Indicator

Field: DLY %

Description: The percentage of time an AS is delayed because of contention for the processor during the range period. For the multitask AS, RMF reports only one delay when several tasks are delayed at the same time.

Guideline: Use the cursor sensitivity: put the cursor on the name of the AS you want to analyze (usually we start with the most delayed job). Hence the Job Delay report (as in Figure 43 on page 73) is displayed; it allows further investigation as explained later in this chapter.

Problem Area: Points out a processor delay.

Potential Solution:

- Start by identifying the holding jobs that are causing the most delay
- See "Determine CPU 'Holders'" on page 73

USING Indicator

Field: USG % and Appl% CP

Description:

- USG %: The percentage of time an AS is using the processor during the range period. This is a sampled single state value.
- Appl% CP: Percentage of CPU time as sum of TCB time, global and local SRB time, and preemptable or client SRB time consumed on behalf of this address space. This is a measured multi-state value: if the address space is using more than one processor, this value can exceed 100%.

Guideline: Start with the AS with the largest Appl% CP value, then look at its USG% value: if the USG% value is large, you have a heavy processor consumer.

Problem Area: Points out processor consumption.

Potential Solution:

- See "Processor USING: Trimming Activities" on page 76

Monitor I Indicators

This section describes the Monitor I indicators that may be used to alert you to a potential processor problem.

CPU Activity Report

```
                                                C P U  A C T I V I T Y
                                                                                                PAGE 1
           z/OS V1R6               SYSTEM ID OS04            DATE 06/05/2004         INTERVAL 14.59.996
                                   RPT VERSION V1R5 RMF      TIME 09.29.00           CYCLE 1.000 SECONDS

CPU  9672   MODEL  Z97

---CPU---  ONLINE TIME   LPAR BUSY      MVS BUSY     CPU SERIAL  I/O TOTAL        % I/O INTERRUPTS
NUM  TYPE  PERCENTAGE    TIME PERC      TIME PERC    NUMBER      INTERRUPT RATE   HANDLED VIA TPI
 0    CP   100.00         76.93          99.15         045104      320.9              0.33
 1    CP   100.00         76.85          99.13         145104      324.0              0.22
 2    CP   100.00         76.86          99.09         245104      321.6              0.22
 3    CP   100.00         76.87          98.99         345104      329.5              0.25
 4    CP   100.00         76.89          99.01         445104      325.9              0.33
 5    CP   100.00         76.87          98.88         545104      328.7              0.36
 6    CP   100.00         76.88          98.86         645104      338.4              0.46
 7    CP   100.00         76.85          98.73         745104      341.1              0.51
 8    CP   100.00         76.82          98.59         845104      335.9              0.67
CP TOTAL/AVERAGE         76.87          98.94                      2966               0.37


SYSTEM ADDRESS SPACE ANALYSIS         SAMPLES =    898

            NUMBER OF ASIDS                              DISTRIBUTION OF QUEUE LENGTHS  (%)
TYPE       MIN    MAX     AVG       0      1     2      3      4      5      6     7-8   9-10   11-12  13-14   14+
----      ------ ------ --------   -----  -----  -----  -----  -----  -----  -----  -----  -----  -----  -----  -----
IN
READY         5     37     14.8     0.0    0.0    0.0    0.0    0.0    0.2    0.4    2.7    9.7   16.8   22.3   47.5

                                    0     1-2    3-4    5-6    7-8   9-10   11-15  16-20  21-25  26-30  31-35   35+
                                   -----  -----  -----  -----  -----  -----  -----  -----  -----  -----  -----  -----

IN           161    198    177.5    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0  100.0
OUT
READY          0      2      0.0   98.2    1.7    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
OUT
WAIT         161    188    174.5    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0  100.0
LOGICAL
OUT RDY        0      2      0.0   97.9    2.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
LOGICAL
OUT WAIT      50    105     83.4    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0  100.0

BATCH         76     85     79.9    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0  100.0
STC          218    232    226.9    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0  100.0
TSO           87     97     92.4    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0  100.0
ASCH           0      1      0.0   99.7    0.2    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
OMVS          35     43     36.2    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0   62.5   37.4
```

Figure 40. Monitor I CPU Activity Report

USING Indicator

Field: BUSY TIME PERCENTAGE

Description: The percentage of the interval time that the processor was busy executing instructions; it includes the non-captured CPU-Time. The calculation is different depending on whether you are in basic mode or LPAR mode (see *z/OS RMF Report Analysis*).

Guideline: If this value is greater than 85% AND the total of IN READY values for the columns 0,1,...N is below 80% (N = the number of CPs in your system), then you have some work with CPU delay.

Note: Depending on your environment, the guidelines may be very different:
- In a batch or scientific environment, a 100% CPU busy is usually not a problem.
- In a transaction environment, user response times usually increase when CPU busy goes over 80%. The type of the application and the number of CPs available to it can influence how high this CPU busy value can go before response time suffers.

Problem Area: Points out processor consumption.

Potential Solution:
- See "Processor USING: Trimming Activities" on page 76

Note: BUSY TIME PERCENTAGE by itself is not enough as a CPU indicator. You must remember that the distribution of the arrival rate of work demanding CPU is not constant. So, chances are that even at a low utilization level, a queue will build up. The total of IN READY values for the columns 0,1,...N gives the percentage of time when no contention for CPU was detected, that is the number of AS with a ready process (TCB or SRB) was equal to or less than the number of CPs. For example, if your total is 80%, then 20% of the time (100%-80%) you have work delayed.

This is not to say that you cannot run an MVS system at higher utilizations, on up to 100% busy. You can. Just be aware that the busier the CPU, the longer the CPU delay for lower-priority workloads will be. The more low-priority work you have, and the fewer non-CPU bottlenecks you have (e.g. responsive I/O), the busier you will be able to drive your CPU and still maintain good response for your high-priority workloads.

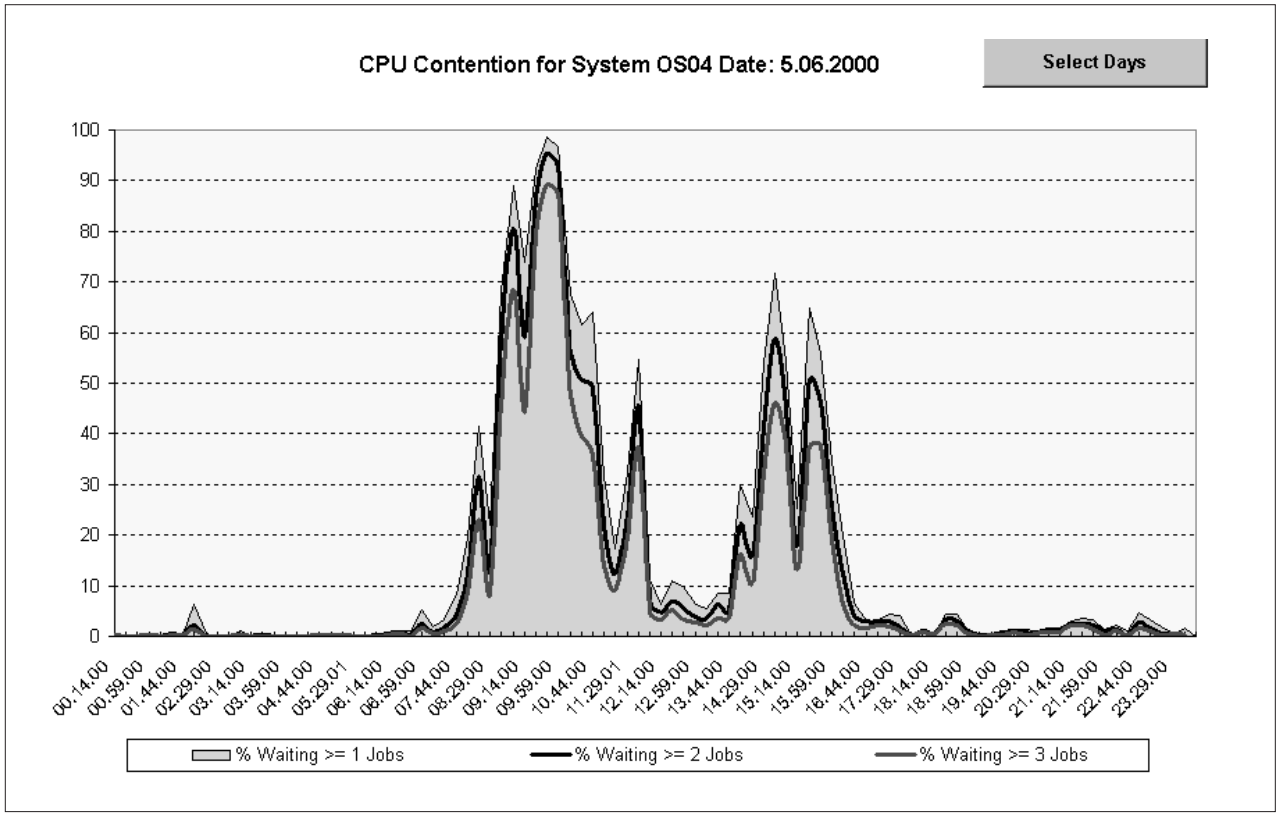


Figure 41. CPU Contention Report. Spreadsheet Reporter macro RMFY9SUM.XLS (System Overview Report - OneCpuCont)

This report provides some more detailed data on CPU contention. In addition to the contention value which is given by the percentage of time when at least one job was ready and waiting for the processor, you see here the percentages when at least two or three jobs were waiting.

Partition Data Report

```
                                    P A R T I T I O N   D A T A   R E P O R T
                                                                                                                    PAGE    2
            z/OS V1R6               SYSTEM ID NP1             DATE 04/15/2004          INTERVAL 14.59.678
                                    RPT VERSION V1R5 RMF      TIME 09.30.00            CYCLE 1.000 SECONDS

 MVS PARTITION NAME                   NP1                     NUMBER OF PHYSICAL PROCESSORS          9
 IMAGE CAPACITY                       100                                 CP                         9
 NUMBER OF CONFIGURED PARTITIONS        9                                 ICF                        0
 WAIT COMPLETION                       NO
 DISPATCH INTERVAL                DYNAMIC

--------- PARTITION DATA ----------------  -- LOGICAL PARTITION PROCESSOR DATA --   -- AVERAGE PROCESSOR UTILIZATION PERCENTAGES --

                  ----MSU----  -CAPPING--  PROCESSOR-  ----DISPATCH TIME DATA----   LOGICAL PROCESSORS  --- PHYSICAL PROCESSORS ---
NAME       S   WGT  DEF    ACT  DEF   WLM%  NUM   TYPE   EFFECTIVE       TOTAL      EFFECTIVE    TOTAL  LPAR MGMT  EFFECTIVE  TOTAL

NP1        A    20  100     10  NO    62.2  1.2   CP    00.04.29.502  00.04.27.519       25.10    24.92       ****       3.33   3.30
NP2        A     1    0      1  YES    0.0    4   CP    00.00.22.680  00.00.22.083        0.75     0.73       ****       0.28   0.27
NP3        A    10    5      8  NO     3.3  1.0   CP    00.03.37.761  00.03.35.859       24.20    23.99       ****       2.69   2.67
NP4        A   300   95    155  NO     0.0  0.3   CP    01.12.08.405  01.12.06.405       80.18    80.15       ****      53.46  53.43
NP5        A   200   50     52  NO     0.0    4   CP    00.24.13.447  00.24.11.311       40.39    40.33       ****      17.95  17.92
CFC1       A   DED    0     32         0.0    1   CP    00.14.59.611  00.14.59.625       99.99    99.99       0.00      11.11  11.11
CFC2       A   DED    0      0         0.0    1   CP    00.00.00.000  00.00.00.000        0.00     0.00       0.00       0.00   0.00
*PHYSICAL*                                                            00.00.03.603                           0.04                0.04
                                                                      ------------  ------------                          ------ ------
TOTAL                                                                 01.59.51.408  01.59.46.408                           0.04      88.81  88.75

CB88       D
CB89       D
```

Figure 42. Monitor I CPU Activity Report - Partition Data Report Section

This report provides data about configured partitions in LPAR mode. Only partitions active at the end of the duration interval are reported.

USING Indicator

Field: PHYSICAL PROCESSORS TOTAL

Description: The total amount of time the physical processor resource was assigned to a partition AND to the management of the LPAR itself. The partition identified by the name *PHYSICAL* is not a configured partition: this is a way to report uncaptured time which was used by LPAR but could not be attributed to a specific logical partition.

Guideline: If the total physical processor complex is 100% busy for one partition, this LPAR could be constrained by other LPARs.

Problem Area: Points out processor consumption.

Potential Solution: Check the partition weights, capping, and number of CPs defined, to be sure you are allocating the CP resources the way you intended.

• See Appendix A, "PR/SM LPAR Considerations"

Is Your Processor Idle due to Other Bottlenecks?

If your system shows no processor indicators under stress, that does not mean you don't have a potential processor problem: You may have some other bottleneck in the system preventing work from using the processor. If your daily processor-busy peaks start to plateau, before reaching 100%, this could be an indication of such a bottleneck. The major ones are:

- Storage
- I/O
- ENQ
- Operator
- Network

If you think this might be the case, see the other chapters for guidance on finding bottlenecks.

Remember that all CPUs wait at the same speed...

DELAY for Processor: Guidelines for Improvement

We assume in this section that you have found that your application's primary delay is due to the processor. This means that other workloads are using the processor when your application wanted it. In this section, we will discuss what can be done to decrease processor delay for a workload.

The section "Processor USING: Trimming Activities" on page 76 is worth checking when you have a delay problem as well: the delay can be caused by an AS using too much of the processor. Trimming this AS solves the delay problem.

Determine CPU 'Holders'

Bringing up the Monitor III Processor Delays report (shown in 39), use the cursor sensitivity. By positioning the cursor on the line describing the job you want to look at, and pressing ENTER, a new panel is displayed, on which the AS delaying the workload are named, as shown in Figure 43.

JOB Report

```
                         RMF V1R5  Job Delays                  Line 1 of 1
Command ===>                                                Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

Job: BHOLPRO2     Primary delay: Job is waiting to use the processor.

Probable causes: 1) Job BHOLPRO1 may be looping.
                 2) Higher priority work is using the system.
                 3) Improperly tuned dispatching priorities.

------------------------ Jobs Holding the Processor --------------------------
Job:        BHOLPRO1         Job:            BHOL         Job:        *MASTER*
Holding:          4%         Holding:          3%         Holding:          1%
PROC Using:      98%         PROC Using:       3%         PROC Using:       1%
DEV Using:        0%         DEV Using:        8%         DEV Using:        4%
--------------------------- Job Performance Summary ---------------------------
        Service       WFL -Using%- DLY IDL UKN ----- % Delayed for ------ Primary
CX ASID Class    P Cr  %  PRC DEV   %   %   %  PRC  DEV STR SUB OPR  ENQ  Reason
B  0033 BATCH    1     94  94   0    6   0   0   6    0   0   0   0    0  BHOLPRO1F
```

Figure 43. Monitor III Job Delays Report

Some probable causes of trouble are mentioned as well. This is the best starting point. See which AS are delaying your workload.

Review Dispatching Priorities

The *dispatching priority (DP)* of each address space is specified dynamically according to the goals of the respective service class. You should reconsider your service policy definition if the workload delaying your application a has higher dispatching priority.

The Monitor II ASD report shows address space dispatching priorities (DP PR).

ASD Report

```
                    RMF - ASD Address Space Data Report           Line 1 of 64
Command ===>                                                   Scroll ===> HALF

                        CPU= 67/ 55 UIC=2540 PR=   0         System= SYSF Total

13:43:53          S C  R  DP   CS   ESF   CS   TAR  X  PIN   ES  TX  SWAP   WSM
JOBNAME  SRVCLASS P L  LS PR    F        TAR   WSS  M   RT   RT  SC    RV    RV

ANTAS000 STCDEF   1 NS    F1  871
MASTER*  SYSTEM   1 NS    FF  141               0     0.0  0.0   0     0
PCAUTH   STCDEF   1 NS    74   26               0  X  0.0   15   0   175
RASP     STCDEF   1 NS    FF   48               0  X  0.0   15   0     0
TRACE    STCDEF   1 NS    76   78               0  X  0.0   15   0     0
GRS      SYSTEM   1 NS    FF 1381               0  X  0.0   15   0   175
DUMPSRV  SYSTEM   1 NS    FF   31               0     0.0   15   0   175
CONSOLE  SYSTEM   1 NS    FF   62               0  X  0.0   15   1     0
ALLOCAS  SYSTEM   1 NS    71  317               0  X  0.0   15   1     0
LLA      SYSSTC   1 NS    FC  331               0  X  0.0   24   0     0
VTAM     SYSSTC   1 NS    FD  518               0     0.0   44   0   175
NETVS    SYSSTC   1 NS    FC   77               0  X  0.0   32   1   175
NETVIEW  SYSSTC   1 NS    FC  497               0     0.0   30   0   175
SMF      SYSTEM   1 NS    FF   65               0  X  0.0   15   0   175
RMF33    SYSSTC   1 NS    FF  128               0     0.0   38   0   175
DFHSM    SYSTEM   1 NS    7D  674               0     0.0  671   1   175
AMSAQTS  SYSTEM   1 NS    74   37               0     0.0   50   1   175
VLF      SYSTEM   1 NS    78 2297               0  X  0.0   22   0   175
JES2     SYSTEM   1 NS    FE  827               0     0.0   25   0   175
SMS      SYSTEM   1 NS    79   59               0  X  0.0   24   0   175
JOEUSER  BTCHPRD  1 WT TI FF  764               0  X  0.0   15   0    18 -1000
```

Figure 44. Monitor II Address Space State Data Report

Check for CPU Delay due to Other LPARs

If you are running as an LPAR sharing CPs with other LPARs, and your MVS system is not 100% busy, you might ask: Are the logical CPs not working more because there is no demand (wait time) or because the LPAR is not getting enough CPU?

One way to check if other LPARs are constraining your system is to do the following: as described in a prior section, add up your total of IN READY values for the columns 0,1,...N (N = the number of CPs in your LPAR) from the Monitor I CPU Activity report. If this total is more than your CPU BUSY PERCENTAGE, then you have another LPAR constraining you (you have CPU delay % > CPU busy %). If this is the case, check the partition weights, capping, and number of CPs defined, to be sure you are allocating the CP resources the way you intended.

Check LPAR Balancing Using the CPC Capacity Report

You can use the Monitor III CPC Capacity report to check whether the capacity being defined for each partition is sufficient to provide the service that all applications are expecting.

```
                  RMF V1R5 CPC Capacity Report                  Line 1 of 1
Command ===>                                                 Scroll ===> HALF

Samples: 100     System: PRD1  Date: 10/07/04  Time: 10.32.00  Range: 100   Sec

 ----- 2064 Model 114 -- Partition Z2       -----     --- 4 Hour MSU  --
 CPC Capacity    410        Weight % of Max  74.2     Average          51
 Image Capacity   50        WLM Capping %     5.4     Maximum         185

 Partition    -- MSU --  Cap  Proc    Logical Util %   - Physical Util % -
              Def   Act  Def   Num    Effect   Total   LPAR  Effect  Total
 *CP                                                    1.0    10.1   11.1
  TZ1         130   122   NO   4.3      11.2    12.5    0.4     3.4    3.8
  Z1          150    89   NO   5.2       9.3     9.6    0.1     3.5    3.6
  Z2           50    58   NO   2.1      11.5    12.8    0.2     1.7    1.9
  Z3          N/A     0  YES   2.4       8.8    10.6    0.3     1.5    1.8
  PHYSICAL                                              0.1            0.1

 *ICF total                                             0.1     7.1    7.1
  CF1           0    29          1      99.9    99.9    0.0     7.1    7.1
  CF2           0     0          1       0.0     0.0    0.0     0.0    0.0
  PHYSICAL                                              0.1            0.1
```

Figure 45. Monitor III CPC Capacity Report

The report provides information about processor resource consumption of each partition, and it displays the 4-hours average consumption. If this average is higher than the defined image capacity, WLM will start capping this partition. In the above example, the average (51 MSUs/hour) is higher than the image capacity (50 MSUs/hour). Therefore, partition Z2 will get less processor resources in the following intervals until the average is below the defined capacity. This might result in processor constraints and performance problems. If this is not just a temporary but a permanent problem, one should consider about increasing the image capacity for this partition.

Processor USING: Trimming Activities

This section describes the activities that help decrease the processor consumption. In case of delay, this list of activities can be used to trim the AS that uses too much processor resource.

Check for Loops

PROC Report

```
                    RMF V1R5  Processor Delays               Line 1 of 71
Command ===>                                               Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

             Service  DLY  USG   Appl EAppl ----------- Holding Job(s) -----------
Jobname   CX Class     %    %      %     %  %  Name      %  Name      %  Name

BOE1HCP$  B  BATCH     10   11   9.8   9.8   9 BJPI       8 BTFL       4 CATALOG
BTFL      T  TSO  2     9   91  83.6  83.7   8 BJPI       3 BOE1MGR$   3 BOE1HCL$
BOE1TBL$  B  BATCH      9   28  26.7  26.7   8 BTFL       7 BJPI       5 JES3
BOE1FIR$  B  BATCH      9   21  20.5  20.5   8 BTFL       7 BJPI       7 JES3
BOE1HCL$  B  BATCH      7   19  15.3  15.3   7 BJPI       7 BTFL       4 BOE1MGR$
BOE1DEV$  B  BATCH      7   11   9.7   9.7   6 BJPI       5 BTFL       4 TESTRTX
BOE1MGR$  B  BATCH      6   15  12.7  12.7   6 BJPI       5 BTFL       3 BOE1HCP$
BJPI      T  TSO        4   96  87.9  87.9   4 BTFL       3 JES3       2 XXVTAM
BOE1HIT$  B  BATCH      3   13  10.4  10.4   3 BTFL       3 BJPI       2 CATALOG
BOE1HOM$  B  BATCH      3    5   3.0   3.0   3 VIBTS      3 BJPI       2 BOE1DEV$
JES3      S  SYSSTC     2   21  22.0  22.0   2 BJPI       1 BTFL       1 *MASTER*
BRTV      T  TSO        2    6   4.9   4.9   2 BTFL       2 BJPI       1 BOE1HCL$
BRMC      T  TSO        2    5   5.0   5.0   2 TESTRTX    1 CATALOG    1 BJPI
XXVTAM    S  SYSSTC     1   22  15.8  15.8   1 JES3       1 BOE1TBL$   1 AKS
TESTRTX   S  SYSSTC     1   13  12.2  12.2   1 ELF        1 GRS        1 BTFL
VIBTS     S  SYSSTC     1    9   5.5   5.5   1 *MASTER*   1 BTFL       1 TESTRTX
BHWG      T  TSO        1    7   4.4   4.4   1 MAT        1 SNALINK    1 BOE1
BOE1      T  TSO        1    4   2.7   2.7   1 OMEGA      1 BOE1FIR$   1 GRS
BPFX      T  TSO        1    4   2.1   2.1   1 BTFL       1 BJPI       1 TCPIP
BOE1GCR$  B  BATCH      1    4   2.7   2.7   1 BHWG       1 BJPI       1 XXVTAM
```

Figure 46. Monitor III Processor Delays Report

Detect loops by inspecting DLY % and USG % in this report for individual address spaces. If the sum of both is 100%, you could be in an enable loop.

You can check this by moving the cursor to jobname BTFL, and press ENTER. This links to the Job Delays report.

Job Report

```
                          RMF V1R5  Job Delays                  Line 1 of 1
Command ===>                                                 Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

Job: BTFL         Primary delay: Job is waiting to use the processor.

Probable causes: 1) Job BJPI may be looping.
                 2) Higher priority work is using the system.
                 3) Improperly tuned dispatching priorities.

------------------------- Jobs Holding the Processor --------------------------
Job:            BJPI         Job:        BOE1MGR$         Job:        BOE1HCL$
Holding:          8%         Holding:          3%         Holding:          3%
PROC Using:      96%         PROC Using:      15%         PROC Using:      19%
DEV Using:        0%         DEV Using:        5%         DEV Using:        8%
-------------------------- Job Performance Summary ---------------------------
        Service       WFL -Using%- DLY IDL UKN ----- % Delayed for ------ Primary
CX ASID Class    P Cr  %  PRC DEV   %   %   %  PRC  DEV STR SUB OPR  ENQ  Reason
T  0148 TSO      2     91   91  0    9   0   0   9    0   0   0   0    0  BJPI
```

Figure 47. Monitor III Job Delays Report

A good indication of a loop is the fact that no other delays than processor delay are shown.

This is also detected by Monitor III which points to another job that might be looping.

Check with the owner of this address space (in this example, it is a TSO user), before deciding how to proceed - don't cancel it too quickly!

Check for Slip Traps

WFEX Report

```
                    RMF V1R5  Workflow/Exceptions               Line 1 of 12
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

--------------------------- Speed (Workflow) --------------------------------
           Speed of 100 = Maximum, 0 = Stopped         Average CPU Util:  81 %
Name       Users Active      Speed          Name       Users Active      Speed
*SYSTEM      505     13         54
ALL TSO      433     10         55
ALL BATCH      2      0         42
ALL STC       70      2         55
ALL ASCH                 Not avail

------------------------------ Exceptions -----------------------------------
Name       Reason        Critical val. Possible cause or action
* SLIP *   SLIP PER TRAP               SLIP ID=SR01 is active.
BEVK       Rate < 2.0    1.220 /sec    Tx rate is 1.220 /s.
BSHR       STOR-COMM     23.1 % delay
CSAHOG     JCSA% > 15    18.3 %        Job CSA usage 18 %, system 57 %.
POK063     DAR > 20      23.22 /sec    I/O rate is 23.22 /s on volume POK063.
DATAPK     Not avail                   Volume DATAPK is not mounted.
```

Figure 48. Monitor III Workflow/Exceptions Report

Use the workflow/exceptions report to watch for signs of a SLIP PER active trap. *Program Event Recording (PER)* is a function of S/390® architecture that generates program interrupts and uses the CPU unproductively (when not used to analyze a problem).

PER is started by a SLIP trap and has significant overhead. The WFEX report shows any active SLIP traps using PER. Check to see if the trap is still needed.

The `* SLIP *` exception line, in yellow, is always shown as the first exception. This line cannot be suppressed when the SLIP is active.

Reduce I/O

Rules of Thumb

- 1000 I/Os per second consume about 7.6% of a 9672 Z87 system.
- Every 160 DASD demand pages/sec uses about 1.0% of a 9672 Z87 system.

You can avoid I/O by:
- Using *Data-In-Memory (DIM)* techniques (refer to Appendix C, "Data-In-Memory").
- Increasing block sizes
- Exploring SAM/E chain scheduling possibilities
- Moving I/O bound activities to other shifts
- Decreasing the paging

Avoiding I/O may help in saving CPU. Consequently you improve your performance (response and throughput); however chances are that your CPU is going to be busier, productively, because you decrease I/O delay.

Decrease Monitor Overhead

- Check for obsolete operations clists.
- Close down duplicate/redundant monitoring.
- Reduce scope and/or frequency of other monitors.
- Control and decrease SMF records generation.
- Increase RMF cycle, control the use of RMFMON and RMF commands.

Tune System Address Spaces

The system AS that usually consume the most processor are:

- Console AS, due to commands and message traffic in MVS console. Consider message filtering to reduce traffic.
- Catalog AS (CAS). Decrease the use of Stepcat and Jobcat statements in JCL (particularly effective in a batch and TSO environment).
- VTAM AS. Consult a TP specialist.
- Master Scheduler AS, mainly due to paging activities in SRB mode. See Chapter 5, “Analyzing Processor Storage Activity” for ways to reduce paging.
- JES AS. Decrease batch and some TSO activity (drastic action for small payback).

Tracing

GTF traces can be used to see the use of supervisor code in your TCB time (SVCs and PCs).

Sometimes you may see a lot of GETMAIN, EXCP, or CLOSE activities in your GTF trace. This can be a clue for application redesign.

Calculate Your Capture Ratio

If your prime shift capture ratio is below about 80%, check the following:

- How much CPU am I spending with swapping (non-captured)?
- How much CPU am I spending with paging (some non-captured)?

Redesign Application

Sometimes the only answer may be to redesign an application completely or partly, or change the user’s expectations.

Install More CPU Power

When you have done all you can, and the problem does not go away, you can still resort to buying more CPU power.

Summary

To analyze processor problems you should:

- Determine if this is a *DELAY* due to another user or a *high USING* processor consumption.
- Look at the Monitor III Job Delay report first in case of *DELAY*.
- Reduce all the processor overhead that you can: it will usually be a very small amount.

Chapter 4. Analyzing I/O Activity

Let's Analyze Your I/O Subsystem

This chapter discusses how to analyze an I/O problem. We review:

- How to perform an overall health check of your I/O subsystem
- How to improve I/O performance with the Enterprise Storage Server
- How to analyze and solve I/O problems related to
  - Cache
  - DASD
  - Tape

We start by reviewing some of the key concepts, terms, and functions that will be used throughout the chapter.

Do You have an I/O Problem?

You can conclude that you have an I/O problem if:

- Service level objectives are being missed
- Users are complaining about response times
- AND, the I/O indicators discussed in this chapter show signs of stress. Or better still, you see high `DEV DLY` or `USG` for an important workload directly in Monitor III reports (for example, `SYSINFO`, `DELAY`, `DEV`, or `DEVR`).

Here are the primary RMF indicators you can use to assess your I/O subsystem. After listing these indicators, we will go on to discuss each one in more detail, with guidelines on values that may indicate contention and suggestions for further action.

RMF I/O INDICATORS

DASD LCU level

I/O rate

Response Time

- Disconnect
- Connect
- Queue
- Pend

Cache Statistics

- Read hits
- DASD fast write hits
- Cache Fast write hits
- Staging
- DFW retries
- Bypass
- Inhibit
- Off
- ASYNC

DASD Device level

I/O rate

Response Time

- Disconnect
- Connect
- Queue
- Pend

Utilization

Cache Statistics

- Read hits
- DASD fast write hits
- Cache Fast write hits
- Staging
- DFW retries
- Bypass
- Inhibit
- Off
- ASYNC

Tape

I/O rate

Response Time

- Disconnect
- Connect
- Queue
- Pend

Never start any tuning effort simply because one indicator seems to be a problem. Always check other related indicators to build a good understanding of your I/O subsystem (and overall system). Check to see which workloads are delayed, and by how much. Do all this to make sure you have the big picture well in hand, and only then begin to plan your tuning efforts.

As you analyze I/O problems, remember the following:

- **Manage to service level agreements**. No I/O response time, or other indicator, is inherently good or bad. The basic premise of performance management is that you want good I/O response times for important applications. What is ″good″ response time? The answer, of course, is that it depends. One way you can determine good response is to measure the response times of the volumes and

logical control units (LCUs) when your applications are running well, meeting the formal or informal service level agreements.

- **Keep the balanced systems concept in mind**. Before embarking on I/O subsystem analysis, review the CPU and processor storage resources. Keeping the big picture in mind will help you make better decisions on matters that affect overall system throughput, and on trade-offs such as using additional processor storage to reduce I/O rates.
- **Proceed in a top-down manner**. This implies three things:
  - If you think you see a problem, start with the basic facts. Which workload has the problem? What are the I/O rates and response times? Are service levels really being impacted? A ″bad″ response time may not be worth spending time on, if the I/O rate is low, or the workload has low priority.
  - Analyze your overall subsystem before trying to tune for specific DASD volumes or tape subsystems.
  - Analyze any response-time problem to see which component is greatest (disconnect, queue, etc.), and address that component first.
- **Optimize DASD data placement**. The best response time is achieved if the I/O never happens. This may be accomplished by a smarter program that avoids the I/O, by buffering, or by data-in-memory techniques where the I/O is resolved somewhere in processor storage. The next best response time is achieved by resolving the I/O directly in a cache control unit, a cache hit. The worst response time occurs when a DASD I/O must be resolved directly from a DASD device. See Appendix C, “Data-In-Memory.”

I/O Subsystem Health Check

This section describes how to perform an overall health check of your I/O subsystem. Use this section to help you establish a baseline measurement of I/O subsystem performance, set up your performance monitoring process, and provide a base of knowledge from which to begin performance tuning.

Understand your Business Priorities

All devices and workloads do not have the same priority. Before starting any tuning effort, be sure you understand the important business applications and their formal or informal service objectives. Know which volumes serve the important workloads.

Review CPU and Processor Storage

Alleviating an I/O bottleneck will mean that more work will be delivered to the CPU for processing. Keep the 'balanced system' concept in mind, making sure that CPU and processor storage resources are available. Review processor storage to see how much might be available for increased I/O elimination using data-in-memory techniques.

Establish a Baseline Measurement

Establishing a set of measurements for when your system is running well, provides a good base for later comparison. This will tell you what numbers are good for YOUR environment, which is better than simply following any rule of thumb guideline.

Measure the I/O rate, response times, and the cache activity at the LCU and device level. Disconnect time is inversely proportional to cache effectiveness and is the key measurement.

Concentrate on the DASD LCUs and individual devices with the highest load or most critical data, and account for all I/Os being resolved either in CACHE or directly to DASD (this gives you the cache hit ratio). Then explode these two categories CACHE and DASD into their components, to gain further insights into how the cache controller is performing and what options may be available to manage the performance of the data.

Using the Spreadsheet Report, you can display these measurements by shift and by hour for further evaluation.

Using the guidelines and recommendations given in this chapter, analyze these measurements to determine overall DASD subsystem health and to identify areas of further opportunity. See *Cache Performance Management* for further details and examples.

Improving I/O Performance with the Enterprise Storage Server

Introduction

The Enterprise Storage Server (ESS) is the latest IBM storage product to be developed using IBM's Seascape® architecture. It provides all the open system functions of the Versatile Storage Server™, and it provides all the functions of the 3990 Storage Control too - and a lot more. This chapter describes only features and functions that are relevant in a z/OS environment.

For more information about ESS as well as for z/OS as for open systems, please refer to *IBM Enterprise Storage Server*.

In ESS, there are exciting new features that significantly enhance performance. These features, which together with z/OS software deliver a new world to the S/390 storage environment, are probably the biggest change since disk caching was introduced.

The Enterprise Storage server is the natural successor to the IBM 3990. It provides nearly all the functions that were available on the 3990, including peer-to-peer remote copy (PPRC), extended remote copy (XRC), and concurrent copy. For the many customers who have installed the RAMAC® Virtual Array (RVA), with its revolutionary log structure files (LSF) architecture, ESS is protecting their investment in this technology, too.

Many problems with I/O performance being described in this publication will disappear when you migrate from 3390 (or still 3380) to the Enterprise Storage Server, and many *traditional* recommendations about I/O tuning will become obsolete.

RMF has incorporated a number changes to accommodate and report on the advanced ESS functions related to performance, for example, the number of parallel access volumes or information about RAID ranks in the cache subsystem.

Architecture

The Enterprise Storage Server is a high performance, high availability, high capacity storage subsystem. It contains two 4-way RISC processors with 6 GB of cache and 384 MB of non-volatile storage to protect from data loss. It has a maximum capacity of over 11 TB and will be connected to S/390 through up to 32 ESCON channels. The first two models of the Enterprise Storage Server that became available are:

- The IBM 2105-E20 Enterprise Storage Server

  This model supports the full complement of 128 disks in two 2105 cages.
- The IBM 2105-E10 Enterprise Storage Server

  This model has a limited capacity in terms of disk arrays: only 64 disks can be installed in the base rack. These occupy a single 2105 cage.

Ongoing and frequent ESS performance upgrades can be expected as a direct result of its underlying Seascape architecture. The first example of such upgrades, the ESS Models F10 and F20, incorporates 64-bit RISC processing, higher effective host adapter utilization, more PCI buses, larger cache, and other improvements at the component level. The maximum sequential throughput capability of the ESS Model

F20 is approximately double that of the ESS Model E20. The maximal cache size in the new models is 16 GB, and they have 128 drives in the main cabinet and 256 drives in the extension.

Disks can be installed in the cages in groups of 8. These are called *disk 8-packs*. Each group of 8 disks is configured as a RAID Rank (of either 6 Data + Parity + Spare, or 7 Data + Parity) or JBOD (Just a Bunch Of Disks) with no Parity.

For each 8-pack, you have the choice of three different disks for use:

- 9.1 GB

  Use this disk size for the highest performance RAID ranks.
- 18.2 GB

  Use this disk size for the high performance and capacity - this is the optimum choice for most applications.
- 36.4 GB

  Use this disk size for the high capacity and standard performance.

These guidelines are applicable for most workloads, however, disk performance is most applicable to cache-unfriendly operations that rely on disk performance characteristics for application performance.

A RAID rank is formatted as a set of Logical Volumes (LV). The number of LVs in a rank depends on the capacity of the disks in the array. The LVs are striped across all the data and parity disks in the array.

A non-RAID rank, also called a JBOD rank, is very different, each disk in the group of 8 is a rank in itself. So there are 8 ranks in a JBOD group. Each JBOD disk can be defined as one or more LVs. It is not RAID protected and, should a disk fail, all data on it will be lost.

JBOD or RAID 5? A group of JBODs will normally be used in a situation where you need high random write performance and you do not need RAID protection. With the improvement of ESS, the known disadvantages of the RAID 5 write penalty have been eliminated. It is strongly recommended for virtually all customer environments to use RAID 5, which delivers very high performance, the best data protection and fault tolerance.

z/OS Parallel Sysplex I/O Management

In the z/OS Parallel Sysplex, the Workload Manager (WLM) controls where work is run and optimizes the throughput and performance of the total system. Until now, WLM management of the I/O has been limited. With ESS, there are some exciting new functions that allow WLM to control I/O across the sysplex. These functions include parallel access to both single system and shared volumes and the ability to prioritize the I/O based upon the WLM goals. The combination of these features can significantly improve performance in a wide variety of workload environments.

ESS Performance Features

This chapter covers the performance features of the ESS including PAV, multiple allegiance, and I/O priority queuing.

Concurrent Access Features

z/OS retains device queuing features that were developed to ensure effective serial access to physical devices. These features were developed at a time when physical

devices rather than emulated devices were the norm and long before RAID solutions were developed. They ensured that only one channel program could be active to a disk at any time. This ensured that there was no possibility of interference between channel programs.

The traditional queuing method is illustrated in Figure 49 which shows that Application B is processing a request for volume 2000. If applications A and C issue I/O requests to the same volume, they will get a device busy or UCB busy response.

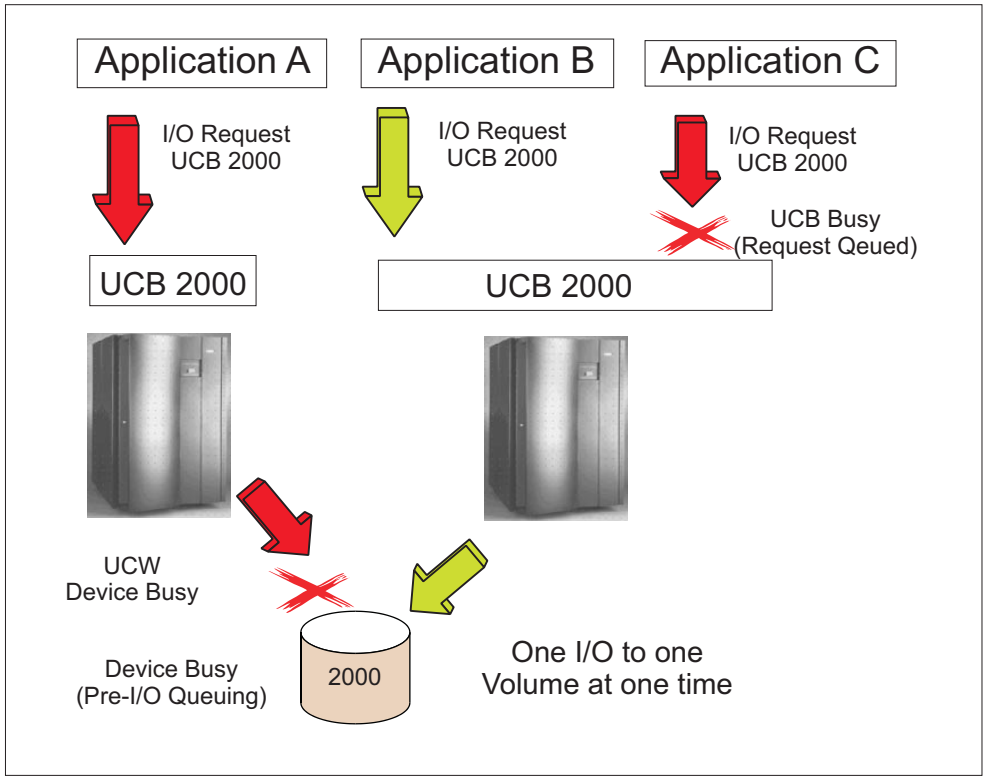


Figure 49. z/OS Traditional Device Serialization

z/OS interacting with the ESS relaxes some of these restrictions. Two features are introduced to allow this.

- Multiple allegiance
- Parallel Access Volumes (PAV)

Although there may be a large number of concurrent operations active to a particular logical volume, the ESS ensures that no I/O operations that have the potential to conflict over access to the same data will be scheduled together. Effectively no data can be accessed by one channel program that has the potential to be altered by another active program. Channel programs that are deemed to be incompatible with an active program are queued within the ESS.

Multiple Allegiance

S/390 device architecture has defined that a state of implicit allegiance exists between a device and the channel path group that is accessing it. This allegiance is created in the control unit between the device and a channel path group when an I/O operation is accepted by the device. The allegiance causes the control unit to

guarantee access, no busy status presented, to the device for the remainder of the channel program over the set of paths associated with the allegiance. This concept has been expanded to support the ESS with a concept of multiple allegiance.

ESS's concurrent operations capability supports concurrent accesses to or from the same volume from multiple channel path groups, system images. The ESS's multiple allegiance support allows different hosts to have concurrent implicit allegiances provided that there is no possibility that any of the channel programs can alter any data that another channel program might read or write.

Multiple allegiance requires no additional software or host support, other than to support the ESS. It is not externalized to the operating system or operator. Multiple allegiance will reduce contention reported as PEND time.

Resources that will benefit most from multiple allegiance are:

- Volumes that:
  - Have many concurrent read operations
  - Have a high read to write ratio
- Data sets that:
  - Have a high read to write ratio
  - Have multiple extents on one volume
  - Are concurrently shared by many users

Parallel Access Volumes

z/OS systems queue I/O activity on a unit control block (UCB) that represents the physical device. High parallel I/O activity can adversely effect performance, traditionally because high accesses usually correlated to high levels of mechanical motion and in subsystems with large caches and RAID arrays because the volumes were treated as a single resource, serially reused. This contention is worst for large volumes with numerous small data sets. The symptom displayed is extended IOSQ time. The operating system can not attempt to start more than one I/O operation at a time to the device.

The ESS's concurrent operations capabilities also support concurrent data transfer operations to or from the same volume from the same system. A volume accessed in this way is called a Parallel Access Volume (PAV).

Figure 50 illustrates multiple allegiance and PAV allowing concurrent I/O processing.

PAV exploitation requires both software enablement and an optional feature on your ESS. PAV support must be installed on each ESS. It enables the issuing of multiple channel programs to a volume from a single system, and allows simultaneous access to the logical volume by multiple users or jobs. Reads, as well as writes to different extents, can be satisfied simultaneously. The domain of an I/O consists of the specified extents to which the I/O operation applies. Writes to the same domain still have to be serialized to maintain data integrity.

Support is implemented by defining multiple UCBs for volumes. The UCBs are of two types.

Base address
: This is the actual unit address of the volume. There is only one base address for any volume.

Alias address

Alias addresses are mapped back to a base device address. I/O scheduled for an alias is physically performed against the base by the ESS. No physical disk space is associated with an alias address, however, they do occupy storage within z/OS. Alias UCBs are stored above the 16MB line.

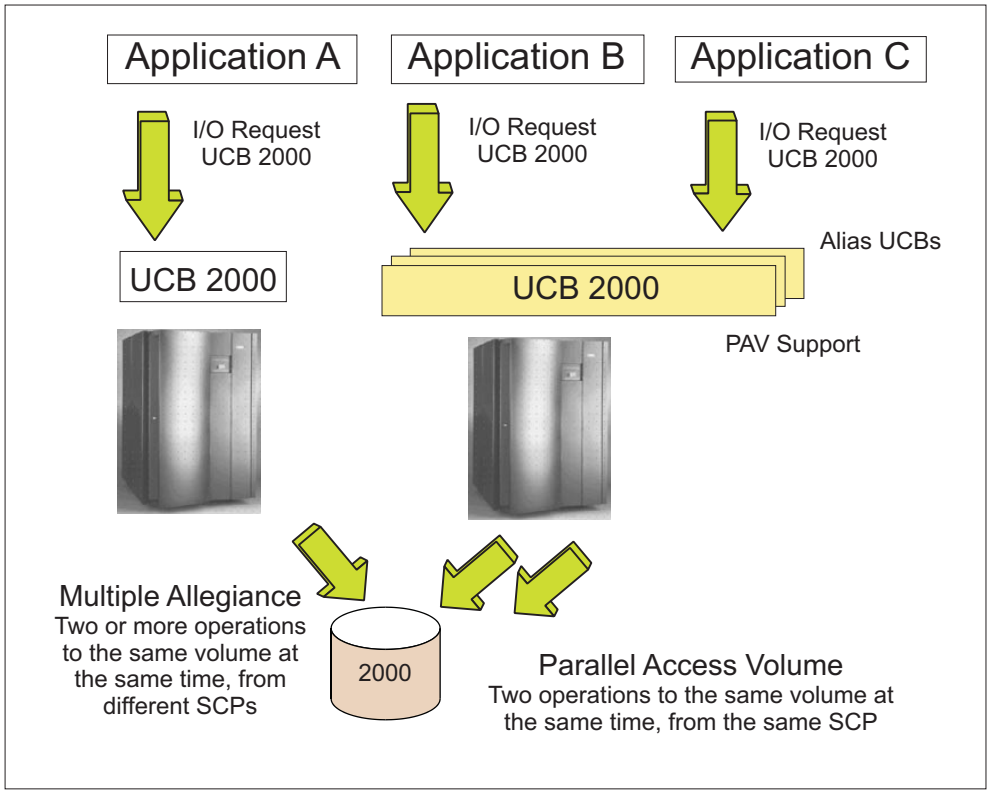


Figure 50. Device Queuing in a Parallel Access Volume Environment

The workloads that are most likely to benefit from the PAV function being available include:

- Volumes that have many concurrently open data sets, for example, volumes in a work pool
- Volumes that have high read to write ratio per extent
- Volumes reporting high IOSQ times

Candidate data types are:

- High read to write ratio
- Many extents on one volume
- Concurrently shared by many readers
- Accessed using media manager or allocated as VSAM extended format

PAVs can be assigned to base UCBs either manually or automatically by WLM. PAVs assigned manually are called static, while those movable by WLM are called dynamic. If WLM is used to manage I/O priorities, then you can use WLM dynamic PAV management as well. Otherwise, WLM can only use dynamic PAVs to balance device utilizations, not to directly help work achieve its goal.

I/O Priority Queueing

If I/Os cannot run in parallel, for example, due to extent conflicts, the ESS will internally queue I/Os. This reduces operating system overheads incurred by having to post "device busy" and redriving channel programs. The ESS will queue I/Os in the order in which they are received. This helps to reduce problems that occur when one processor can respond to interrupts faster than a sharing one, and so monopolize a device.

You also have the option to enable priority queuing of I/Os to the ESS. WLM sets a priority bit in the CCW when running in goal mode. Priority queuing is within a sysplex and queuing is at a volume level. The ESS queues I/O requests in the order specified by WLM. I/O may be queued in the following situations:

- An extent conflict exists for a write operation.
- To allow servicing of a cache miss, device will reconnect when data has been staged to cache.
- A reserve request is issued and other accesses are current with a different path group ID.

Cache Performance

Below are the CACHE and DASD components which you should measure. They are obtained from the Monitor I Cache Subsystem Activity report. In addition, the Monitor I Device Activity report data should be collected for the same period. You will also need to know which devices are used by your important workloads: you may already know this, or you can get this information from Monitor III DEVR or DEV reports.

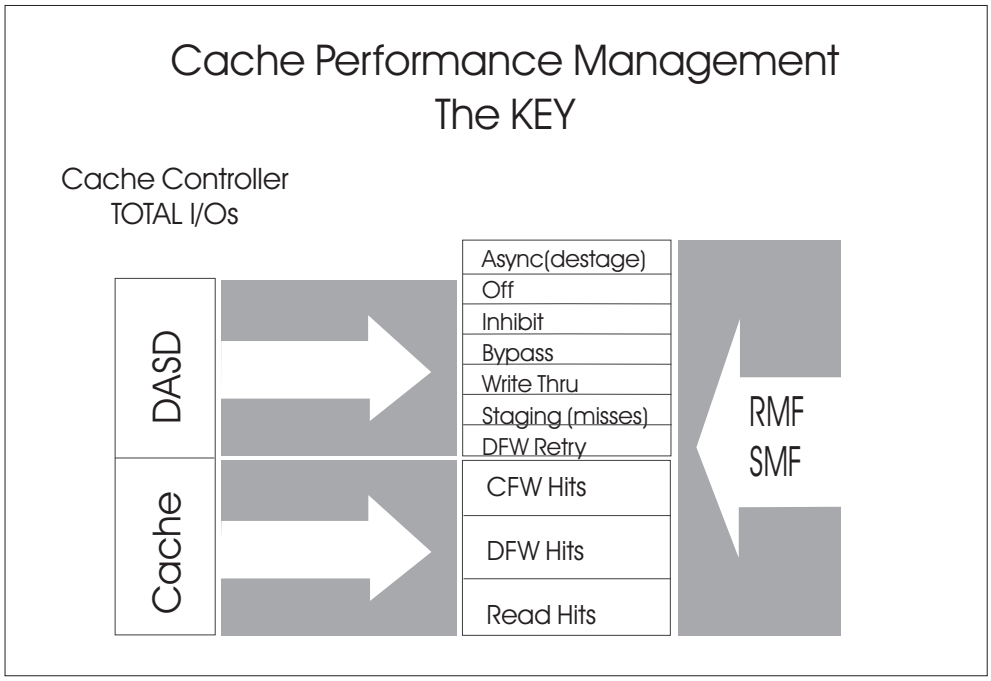


Figure 51. Cache Performance Management: The Key

CACHE

RHIT Normal and sequential read hits

DFWHIT DASD fast write (DFW) hits

CFWHIT Cache fast write (CFW) write and read hits

DASD

STAGE Read misses and DFW misses. The record was not found in the cache. It is read or written, and that record with the rest of the track is staged into the cache. In the ESS, stages no longer occur when data for a DFW request is absent from cache. Instead, this is handled as a write ″promote″ (and counts as a hit).

DFWRETRY DASD fast write retry. The record is in cache and is about to result in a cache hit, but non-volatile storage (NVS) is full. The operation is not really retried but written directly to DASD.

THRU Write 'THRU' to DASD. These are writes to devices behind storage controls that are not enabled for DFW.

BYPASS The bypass mode was specified in the define extent. The I/O is sent directly to DASD, the track is not promoted and the record is invalidated in cache. The ESS does not actually bypass the cache, but ensures that data which specifies bypass mode is destaged quickly.

INHIBIT The inhibit mode was specified in the define extent and the record was not in the cache. Examples of inhibit include DFDSS for reads and DFSMS dynamic cache management (DCM or DCME) ″maybe cache″ and ″never cache″ storage classes. The I/O will retrieve the record in cache if it is there; if it is not in the cache, the I/O is sent directly to DASD, but no staging of the rest of the track occurs. If inhibit mode was specified and the result was a cache hit, the I/O would fall into one of the cache hit categories.

OFF Device is turned off for caching.

ASYNC Asynchronous destage of records to DASD. This includes anticipatory destage to write new or updated data from cache to DASD if NVS or cache is full. The unit is in tracks per second. All the other values given above are in I/Os per second. Async is not actually part of the I/O rate. It is the consequence of a write hit which must eventually be written to DASD. Counting it in the I/O rate would cause the I/O to be accounted for twice, once on the write hit and again when the track was destaged. Async can offer insight into the DFW.

It should be noted that the causes of bypass and inhibit listed above may not be all inclusive. In particular, you should consult your software vendor to understand which caching mode, if any, they use. The default mode is normal or sequential. However, bypass and inhibit mode can explicitly be set in software on an I/O basis. For example, it is possible for some I/Os to a data set to use inhibit and some normal caching mode, even within the execution of a particular application program. The same is true for DB2. Within DB2, one plan may access a table space in normal caching mode and another plan may optimize to access in prefetch mode, utilizing the bypass mode to accomplish this.

Cache Management with ESS

The ESS provides performance improvements over those provided by raw disk by using caching. Caching algorithms are executed by the storage directors and

determine what data occupies cache. Attached hosts can offer information about their data access intentions that the storage directors will use to help select the best algorithm to use for cache management. The following caching features are provided.

Read Caching

Read hits occur when all of the data requested for a data access is located in cache. The ESS improves the performance of read caching by using algorithms to store in cache tracks that have the greatest probability of being accessed by a read operation.

An I/O operation that results in a read hit will not disconnect from the channel and data will be immediately available to the application. If the requested data is not located in cache a write miss occurs and data is read from disk, returned to the program, and loaded to cache. This is called staging. While records are being read from disk the channel program is disconnected allowing other applications to access the channel.

There are three types of staging:

Record staging
Only those records accessed by the channel program are staged into cache

Partial track staging
The required records and the rest of the track are staged to cache. This is the default mode of cache operation

Full track staging
Based on the prediction of sequential I/O processing. This can be predicted either the previous behavior of the application or by the application signalling in the I/O request that sequential access will be used.

Data transferred using inhibit cache load or bypass cache attributes will be loaded into cache but eligible for accelerated destaging.

The ESS offers new capabilities to support the optimization of sequential performance; improved and more sensitive pre-fetching algorithms, and new channel commands that improve overheads and provide increased bandwidth.

Write Caching

Two forms of write caching are supported: DASD Fast Write (DFW) and Cache Fast Write (CFW). Cache fast write is a z/OS function that is intended for data that does not need to be written to disk. CFW is used only when explicitly requested by the application and should only be selected for transient data, data that is not required beyond the end of the current job step and that can more easily be recreated from scratch than recovered from a checkpoint. An example of this type of data is intermediate work files, such as used by a sort program.

DFW is the more usual form of write caching. With DFW the application is told that an I/O operation is complete once data has successfully been written to cache and Non-Volatile Storage (NVS). Data integrity and availability is maintained by retaining two copies of the data until it is hardened to disk, one copy in cache on one cluster and the second in NVS of the other cluster. NVS is protected by battery backup. Normal access to the data is from the copy retained in cache.

Destaging of data that is backed up in NVS from cache to disk is based on a Least Recently Used (LRU) algorithm. Data may be retained in cache after being written

to disk based on the cache activity. Destaging from NVS and cache is anticipatory and threshold based. The intention is to always have NVS and cache resources available to accept new data.

Tracks at the top of the cache LRU list are checked for updates in NVS that have not been destaged to disk. The ESS schedules tracks at the top of the NVS LRU for destaging, so that they can be allocated without a delay during destaging.

Write Performance

Caching benefits write performance as almost all writes are at cache speeds. DFW minimizes any potential penalty of RAID 5 generation of parity. This performance benefit is clearly demonstrated by the published results of ESS performance tests. In addition, write performance is enhanced by striping. The ESS automatically stripes logical volumes across all the drives in the RAID array. This provides automatic load balancing across the disk in the array, and an elimination of hot spots. This design should reduce the amount of effort that storage administrators spend, hand placing data, at the same time offering performance improvements.

The ESS RAID 5 implementation gives a minimal RAID 5 write penalty for sequential writes. When writing a sequential ’stripe’ across the disks in an array, the ESS generates the parity only once for all the disks. This is sometimes called a RAID 3-like implementation, and it provides high performance in sequential operations.

Cache Subsystem Activity Report

The report provides cache statistics on a subsystem basis as well as on a detailed device-level basis. You may start the analysis of your cache subsystem with the Summary report.

Subsystem Summary

The report allows you a top-down approach to analyze the storage subsystems in your configuration because you can see at a glance the most important data. Looking at this report, the storage subsystems causing problems can be easily identified and analyzed in a second Postprocessor run requesting more details.

C A C H E   S U B S Y S T E M   S U M M A R Y

PAGE 1

z/OS V1R6 SYSTEM ID OS04 DATE 06/05/2004 INTERVAL 14.59.996

RPT VERSION V1R5 RMF TIME 09.30.00

SSID	CU-ID	TYPE	CACHE	NVS	I/O	OFF	--CACHE HIT RATE-			--------DASD I/O RATE--------					ASYNC	TOTAL	READ	WRITE	%
					RATE	RATE	READ	DFW	CFW	STAGE	DFWBP	ICL	BYP	OTHER	RATE	H/R	H/R	H/R	READ
0600	0600	3990-006	256	64	3.1	0.0	1.0	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.000	1.000	1.000	30.8
06C0	06E0	3990-006	512	32	38.6	0.0	16.5	22.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7	0.999	0.997	1.000	42.9
0A80	0ABE	3990-006	512	32	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	N/A	N/A	N/A	N/A
6801	4000	2105-F20	8192	192	55.3	0.0	0.0	55.2	0.0	0.0	165.5	0.0	0.0	0.0	486.2	1.000	1.000	1.000	0.1
6802	4100	2105-F20	8192	192	49.5	0.0	0.2	49.3	0.0	0.0	145.1	0.0	0.0	0.0	500.8	1.000	1.000	1.000	0.4
6803	4200	2105-F20	8192	192	56.9	0.0	0.2	56.7	0.0	0.0	169.9	0.0	0.0	0.0	489.1	1.000	1.000	1.000	0.4
6804	4300	2105-F20	8192	192	50.7	0.0	0.2	50.5	0.0	0.0	149.0	0.0	0.0	0.0	465.6	1.000	1.000	1.000	0.4
6805	4400	2105-F20	8192	192	53.0	0.0	0.2	52.8	0.0	0.0	158.3	0.0	0.0	0.0	467.6	1.000	1.000	1.000	0.4
6806	4500	2105-F20	8192	192	42.0	0.0	0.3	41.8	0.0	0.0	123.6	0.0	0.0	0.0	434.6	1.000	1.000	1.000	0.6
6807	4600	2105-F20	8192	192	50.3	0.0	0.2	50.0	0.0	0.0	150.2	0.0	0.0	0.0	454.1	1.000	1.000	1.000	0.4

Figure 52. Cache Subsystem Activy Report - Summary Report

Top-20 Device Lists

The report consists of two top-20 lists of devices, sorted in descending order by DASD I/O rate and by total I/O rate. These two lists allow you to identify the volumes with the highest I/O rates to the lower interface of a subsystem as well as

the volumes with the highest I/O rates in total. Solving a possible problem, one of the listed devices would probably be of most benefit to the overall subsystem.

```
                                C A C H E   S U B S Y S T E M   S U M M A R Y
                                                                                                                                PAGE    2
          z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004          INTERVAL 14.59.996
                                 RPT VERSION V1R5 RMF      TIME 09.30.00

*** DEVICE LIST BY DASD I/O RATE ***

VOLUME     DEV  SSID     %    I/O   ---CACHE HIT RATE--   ----------DASD I/O RATE----------  ASYNC   TOTAL    READ   WRITE       %
SERIAL     NUM         I/O   RATE    READ    DFW    CFW   STAGE  DFWBP    ICL    BYP  OTHER   RATE     H/R     H/R     H/R    READ

PRD440    077E  00B1  25.7   53.1    21.3   16.7    0.0    13.9    0.0    1.3    0.0    0.0    1.6   0.714   0.604   1.000    67.9
PPDS14    0220  00CC   6.9   29.1     1.0   15.7    0.0     0.1    0.0   12.4    0.0    0.0    2.0   0.573   0.943   1.000     6.2
PRD437    0214  00CC   4.5   19.0    12.5    2.0    0.0     4.4    0.0    0.0    0.0    0.0    2.0   0.765   0.738   0.998    89.2
PPD026    0B76  00F1  27.4   71.4    65.9    1.1    0.0     4.3    0.0    0.0    0.0    0.0    0.5   0.939   0.939   0.975    98.4
PRD554    06B9  00FE  59.1   32.4    28.9    0.0    0.0     3.5    0.0    0.0    0.0    0.0    0.0   0.893   0.893   1.000    99.9
PRD339    0231  00CC   2.4   10.3     7.2    0.0    0.0     3.1    0.0    0.0    0.0    0.0    0.0   0.698   0.698     N/A     0.0
 ...

*** DEVICE LIST BY TOTAL I/O RATE ***

VOLUME     DEV  SSID     %    I/O   ---CACHE HIT RATE--   ----------DASD I/O RATE----------  ASYNC   TOTAL    READ   WRITE       %

SERIAL     NUM         I/O   RATE    READ    DFW    CFW   STAGE  DFWBP    ICL    BYP  OTHER   RATE     H/R     H/R     H/R    READ

PPD026    0B76  00F1  27.4   71.4    65.9    1.1    0.0     4.3    0.0    0.0    0.0    0.0    0.5   0.939   0.939   0.975    98.4
PRD440    077E  00B1  25.7   53.1    21.3   16.7    0.0    13.9    0.0    1.3    0.0    0.0    1.6   0.714   0.604   1.000    67.9
PRD327    0200  00CC  11.8   49.8     3.0   46.8    0.0     0.1    0.0    0.0    0.0    0.0    6.3   0.998   0.973   1.000     6.2
PRD343    0515  00E4  17.5   48.9    24.8   24.0    0.0     0.1    0.0    0.0    0.0    0.0    5.5   0.998   0.996   0.999    50.8
PBV321    022C  00CC  11.2   47.3    47.1    0.2    0.0     0.0    0.0    0.0    0.0    0.0    0.1   1.000   1.000   1.000    99.7
PRD307    0507  00E4  15.3   42.8     4.1   38.2    0.0     0.5    0.0    0.0    0.0    0.0   13.0   0.988   0.891   1.000    10.8
 ...
```

Figure 53. Cache Subsystem Activy Report - Top-20 Reports

Based on these reports, you can decide whether further investigation is necessary. The next step could be the creation of reports which show some more details. This can be done on subsystem level as well as on device level.

Subsystem-level Reporting

This generates a report with three sections.

- Cache Subsystem Status
- Cache Subsystem Overview
- Cache Subsystem Device Overview

The subsystem-level report gives an overall view of the storage control, that is the amount of cache storage and non-volatile storage installed, as well as the current status of the cache. In addition, the performance analyst finds the number of I/O requests sent to the control unit and their resolution in the cache (*hits*). The report is completed by a list of all volumes attached to the subsystem, showing their specific utilization of the cache.

Device-level Reporting

This generates, in addition to the above described report, a report with two sections:

- Cache Device Status
- Cache Device Activity

The device-level report provides detailed information for each single device attached to the selected control unit. The report is intended to analyze the cache usage in detail on the basis of the information about the applications that access these volumes.

Based on the data in the Summary report (Figure 52), one might decide that the subsystem SSID=00B1 needs further investigation.

Cache Subsystem Overview

```
                                C A C H E   S U B S Y S T E M   A C T I V I T Y
                                                                                                                                PAGE    1
           z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004          INTERVAL 14.59.996
                                  RPT VERSION V1R5 RMF      TIME 09.30.00

SUBSYSTEM  3990-06     CU-ID  074C     SSID 00B1     CDATE  06/05/2004     CTIME 09.30.03      CINT  15.00
TYPE-MODEL 9396-001

------------------------------------------------------------------------------------------------------------------------------------
                                                  CACHE SUBSYSTEM STATUS
------------------------------------------------------------------------------------------------------------------------------------

SUBSYSTEM STORAGE           NON-VOLATILE STORAGE       STATUS

CONFIGURED     1024.0M      CONFIGURED       32.0M     CACHING                 - ACTIVE
AVAILABLE      1019.9M      PINNED            0.0      NON-VOLATILE STORAGE    - ACTIVE
PINNED            0.0                                 CACHE FAST WRITE        - ACTIVE
OFFLINE           0.0                                 IML DEVICE AVAILABLE    - YES

------------------------------------------------------------------------------------------------------------------------------------
                                                  CACHE SUBSYSTEM OVERVIEW
------------------------------------------------------------------------------------------------------------------------------------

TOTAL I/O   186363   CACHE I/O   181922   CACHE OFFLINE        0
TOTAL H/R    0.865   CACHE H/R    0.886

CACHE I/O    -------------READ I/O REQUESTS-------------  ----------------------WRITE I/O REQUESTS----------------------          %
REQUESTS       COUNT     RATE      HITS     RATE    H/R    COUNT     RATE      FAST     RATE      HITS     RATE     H/R      READ

NORMAL        116003    128.9     95313    105.9  0.822    15150     16.8     15150     16.8     15134     16.8   0.999      88.4
SEQUENTIAL     16655     18.5     16617     18.5  0.998    34114     37.9     34114     37.9     34114     37.9   1.000      32.8
CFW DATA           0      0.0         0      0.0    N/A        0      0.0         0      0.0         0      0.0     N/A       N/A

TOTAL         132658    147.4    111930    124.4  0.844    49264     54.7     49264     54.7     49248     54.7   1.000      72.9

----------------------CACHE MISSES-----------------------          ------------MISC------------          ------NON-CACHE I/O-----
REQUESTS       READ    RATE   WRITE    RATE  TRACKS    RATE                          COUNT   RATE                        COUNT   RATE
                                                                   DFW BYPASS         289    0.3          ICL           4435    4.9
NORMAL        20690    23.0      16     0.0   12993    14.4       CFW BYPASS           0    0.0          BYPASS           6    0.0
SEQUENTIAL       38     0.0       0     0.0     870     1.0       DFW INHIBIT          0    0.0          TOTAL         4441    4.9
CFW DATA          0     0.0       0     0.0                       ASYNC (TRKS)      7239    8.0

TOTAL         20744    RATE    23.0

---CKD STATISTICS---      ---RECORD CACHING---

WRITE           268      READ MISSES    14502
WRITE HITS      268      WRITE PROM     12938
```

Figure 54. Cache Subsystem Activy Report - Status and Overview

This section provides, at a glance, all relevant data for one subsystem. It distinguishes between I/O requests to be handled by the cache (*cachable I/Os*) and non-cache I/O requests.

Cachable I/O Requests

These requests are shown in three categories:

NORMAL
Cache will be managed by *least-recently-used (LRU)* algorithm for making cache space available.

SEQUENTIAL
Tracks following the track assigned in the current CCW chain are promoted, they will be transferred from DASD to cache in anticipation of a near-term requirement.

CFW DATA
WRITE and READ-AFTER-WRITE requests are processed in cache. The data might not be written to DASD. Because CFW does not use the NVS, the application is responsible for restoring the data after a cache or subsystem failure.

Non-Cache I/O Requests

These requests that switched off cache processing, are shown in two categories:

ICL Inhibit cache load — number of I/O requests that inhibited load of data into cache, and the data was not found in the cache.

Note: If the data was in the cache, it has been counted as cache hit. Therefore, this is actually the number of ICL misses.

BYPASS
Number of I/O requests that explicitly bypassed the cache, no matter whether the data is in the cache or not.

I/O Requests

You find the following numbers in the sample report:

Cachable I/O Requests

```
TOTAL READS     132658
READ HITS       111930  →     READ HIT RATE (H/R)   0.844

TOTAL WRITES     49264
WRITE HITS       49248  →     WRITE HIT RATE (H/R)  1.000

TOTAL CACHE I/O 181922
CACHE HITS      161178  →     CACHE HIT RATE (H/R)  0.886
```

Non-Cache I/O Requests

```
ICL               4435
BYPASS               6
TOTAL             4441
```

Total I/O Requests

The sum of these two groups is shown as total number of I/O requests for cached devices of this subsystem:

```
CACHE I/O       181922
NON-CACHE I/O     4441
TOTAL I/O       186363
CACHE HITS      161178  →     TOTAL HIT RATE (H/R)  0.865
```

While hit percentage by itself (CACHE H/R) does not drive the cache performance management process, hit percentage should be derived by dividing the hits by all I/Os (TOTAL H/R).

Hit percents derived by dividing the hits by cachable I/Os are misleading. Cachable I/Os refers to those I/Os that are eligible for caching. At first glance, the distinction between cachable I/Os and all I/Os seems rather insignificant since most I/Os are eligible for caching. However, some software (e.g. DFSMS Dynamic Cache Management and DB2) will selectively disable I/O operations for caching, using inhibit and bypass modes of operations. This could result in high cachable

hit percentages but lower hit percentages. In this case the disconnect time would be higher than what you would expect for a high cachable hit percent, but the disconnect time would be reasonable for a lower hit percent. Where all I/Os are eligible for caching, the hit percent and cachable hit percent will be equal.

You can also get from the report the relationship between READ and WRITE requests as %READ. Sometimes, performance experts work with the READ/WRITE ratio — this value is not shown in the report because of arithmetic problems in case of very less (or zero) WRITE requests, but can be calculated easily, otherwise.

READ/WRITE Ratio

```
TOTAL CACHE I/O      181922
TOTAL CACHE READS    132658  →     % READ        72.9
TOTAL CACHE WRITES    49264  →     R/W ratio     2.69
```

The R/W ratio can also be calculated as

```
R/W ratio = 72.9 / (100 - 72.9) = 2.69
```

Cache Subsystem Device Overview

```
                                C A C H E   S U B S Y S T E M   A C T I V I T Y
                                                                                                                                PAGE    2
           z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004          INTERVAL 14.59.996
                                  RPT VERSION V1R5 RMF      TIME 09.30.00

SUBSYSTEM  2105-01     CU-ID  4000     SSID 6801     CDATE  06/05/2004     CTIME 09.30.03      CINT  15.00
TYPE-MODEL 2105-F20

------------------------------------------------------------------------------------------------------------------------------------
                                                CACHE SUBSYSTEM DEVICE OVERVIEW
------------------------------------------------------------------------------------------------------------------------------------

VOLUME     DEV  RRID     %    I/O   ---CACHE HIT RATE--   ----------DASD I/O RATE----------    ASYNC   TOTAL    READ   WRITE       %
SERIAL     NUM         I/O   RATE    READ    DFW    CFW   STAGE  DFWBP    ICL    BYP  OTHER     RATE     H/R     H/R     H/R    READ

*ALL                100.0   55.3     0.0   55.2    0.0     0.0  165.5    0.0    0.0    0.0    486.2   1.000   1.000   1.000     0.1
*CACHE-OFF            0.0    0.0
*CACHE              100.0   55.3     0.0   55.2    0.0     0.0  165.5    0.0    0.0    0.0    486.2   1.000   1.000   1.000     0.1
SR4B00    4000 0000   0.0    0.0     0.0    0.0    0.0     0.0    0.0    0.0    0.0    0.0      0.0     N/A     N/A     N/A     N/A
SR4B01    4001 0000  17.5    9.7     0.0    9.7    0.0     0.0   29.0    0.0    0.0    0.0     39.1   1.000     N/A   1.000     0.0
SR4B02    4002 0000  17.9    9.9     0.0    9.9    0.0     0.0   29.7    0.0    0.0    0.0     36.3   1.000     N/A   1.000     0.0
SR4B03    4003 0000  10.3    5.7     0.0    5.6    0.0     0.0   16.9    0.0    0.0    0.0     16.8   1.000   1.000   1.000     0.9
SR4B04    4004 0000  17.7    9.8     0.0    9.8    0.0     0.0   29.4    0.0    0.0    0.0     39.2   1.000     N/A   1.000     0.0
SR4B05    4005 0000  18.1   10.0     0.0   10.0    0.0     0.0   30.0    0.0    0.0    0.0     33.8   1.000     N/A   1.000     0.0
SR4B06    4006 0000  18.4   10.2     0.0   10.2    0.0     0.0   30.5    0.0    0.0    0.0     34.6   1.000     N/A   1.000     0.0
```

Figure 55. Cache Subsystem Activy Report - Device Overview

This section of the report provides an overview for one control unit with all attached devices. You can use this report for investigating the cache effectiveness of each volume.

During this analysis, you should use the values `I/O RATE` and `% I/O` as indicators to decide whether a volume has an I/O activity that makes it worth for further performance investigation.

If you come to the conclusion that further analysis of a specific volume might be necessary, you should request the report on device level. It has the same structure as the Subsystem Overview report, but provides the details for each volume.

RAID Rank Activity

```
                                C A C H E   S U B S Y S T E M   A C T I V I T Y
                                                                                                                                PAGE    3
           z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004          INTERVAL 14.59.996
                                  RPT VERSION V1R5 RMF      TIME 09.30.00

SUBSYSTEM  2105-01     CU-ID  4000     SSID 6801     CDATE  06/05/2004     CTIME 09.30.03      CINT  15.00
TYPE-MODEL 2105-F20
------------------------------------------------------------------------------------------------------------------------------------
                                                  RAID RANK ACTIVITY
------------------------------------------------------------------------------------------------------------------------------------

ID    RAID      DA   HDD    -------- READ REQ -------     ------- WRITE REQ --------   ---------- HIGHEST UTILIZED VOLUMES ----------
      TYPE                 RATE  AVG MB  MB/S  RTIME     RATE  AVG MB  MB/S  RTIME

*ALL                 14     265   0.055  14.6    150      243   0.111  27.0    249
0000  RAID-5    01     7     257   0.055  14.2    150      125   0.089  11.1    266   SR4B06  SR4B05  SR4B02  SR4B04  SR4B01  SR4B03
0001  RAID-5    01     7       7   0.055   0.4    148      118   0.134  15.9    232
```

Figure 56. Cache Subsystem Activy Report - RAID Rank Activity

This section of the report provides an overview about the activities of all RAID ranks belonging to the storage server. For each rank, you get the list of those six logical volumes which have the highest utilization, this might help you in deciding whether the mapping between RAID ranks and logical volumes is well balanced, or whether some changes should be performed.

Cache Device Activity

```
                                    C A C H E   D E V I C E   A C T I V I T Y
                                                                                                                    PAGE   62
           z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004          INTERVAL 14.59.996
                                  RPT VERSION V1R5 RMF      TIME 09.30.00

SUBSYSTEM  3990-06     CU-ID  074C     SSID 00B1     CDATE  06/05/2004     CTIME 09.30.03      CINT  15.00
TYPE-MODEL 9396-001
VOLSER     PRD440      NUM    077E     RRID N/A
-------------------------------------------------------------------------------------------------------------------------------
                                                  CACHE DEVICE STATUS
-------------------------------------------------------------------------------------------------------------------------------

CACHE STATUS                                    DUPLEX PAIR STATUS

CACHING             - ACTIVE                    DUPLEX PAIR        - NOT ESTABLISHED
DASD FAST WRITE     - ACTIVE                    STATUS             - N/A
PINNED DATA         - NONE                      DUAL COPY VOLUME   - N/A

-------------------------------------------------------------------------------------------------------------------------------
                                                  CACHE DEVICE ACTIVITY
-------------------------------------------------------------------------------------------------------------------------------

TOTAL I/O    47833   CACHE I/O    46696   CACHE OFFLINE   N/A
TOTAL H/R    0.714   CACHE H/R    0.731

CACHE I/O    -------------READ I/O REQUESTS-------------  ----------------------WRITE I/O REQUESTS----------------------         %
REQUESTS       COUNT     RATE      HITS     RATE     H/R    COUNT     RATE      FAST     RATE      HITS     RATE     H/R      READ

NORMAL         31692     35.2     19131     21.3   0.604      487      0.5       487      0.5       487      0.5   1.000      98.5
SEQUENTIAL         0      0.0         0      0.0     N/A    14517     16.1     14517     16.1     14517     16.1   1.000       0.0
CFW DATA           0      0.0         0      0.0     N/A        0      0.0         0      0.0         0      0.0     N/A       N/A

TOTAL          31692     35.2     19131     21.3   0.604    15004     16.7     15004     16.7     15004     16.7   1.000      67.9

----------------------CACHE MISSES-----------------------          ------------MISC------------          ------NON-CACHE I/O-----
REQUESTS      READ    RATE   WRITE    RATE  TRACKS    RATE                          COUNT   RATE                       COUNT   RATE
                                                                  DFW BYPASS          27    0.0          ICL           1137    1.3
NORMAL       12561    14.0       0     0.0    7020     7.8        CFW BYPASS           0    0.0          BYPASS           0    0.0
SEQUENTIAL       0     0.0       0     0.0       0     0.0        DFW INHIBIT          0    0.0          TOTAL         1137    1.3
CFW DATA         0     0.0       0     0.0                        ASYNC (TRKS)     1396    1.6

TOTAL        12561    RATE    14.0

---CKD STATISTICS---     ---RECORD CACHING---

WRITE              0     READ MISSES     8825
WRITE HITS         0     WRITE PROM      6466
```

Figure 57. Cache Subsystem Activy Report - Device Report

This report shows the statistics for volume PRD440. For example, you might look at this data if users are complaining about the response time of a specific application, and Monitor III reports flag PRD440 as a major cause of delay.

If you prefer to see data for more than one interval, you can create an Overview report with the selection of those data you are interested in. Let us assume that you want to get the following data for volume PRD440 (with SSID=00B1 and address 077E):

Exception	Meaning
CADRT	Total READ I/O rate
CADWT	Total WRITE I/O rate
CADRHRT	Total READ hit ratio
CADWHRT	Total WRITE hit ratio
CADMT	Total staging rate
CADDFWB	Total DFW bypass rate
CADNCICL	Non-cache ICL rate

You find a listing of all conditions in the *z/OS RMF User's Guide*.

You get the Overview report by specifying the following Postprocessor options:

```
OVERVIEW(REPORT)
OVW(READRATE(CADRT(SSID(00B1),DEVN(077E))))
OVW(WRTRATE(CADWT(SSID(00B1),DEVN(077E))))
OVW(READHR(CADRHRT(SSID(00B1),DEVN(077E))))
OVW(WRTHR(CADWHRT(SSID(00B1),DEVN(077E))))
OVW(STGRT(CADMT(SSID(00B1),DEVN(077E))))
OVW(DFWBPRT(CADDFWB(SSID(00B1),DEVN(077E))))
OVW(ICLRT(CADNCICL(SSID(00B1),DEVN(077E))))
```

This creates the following report:

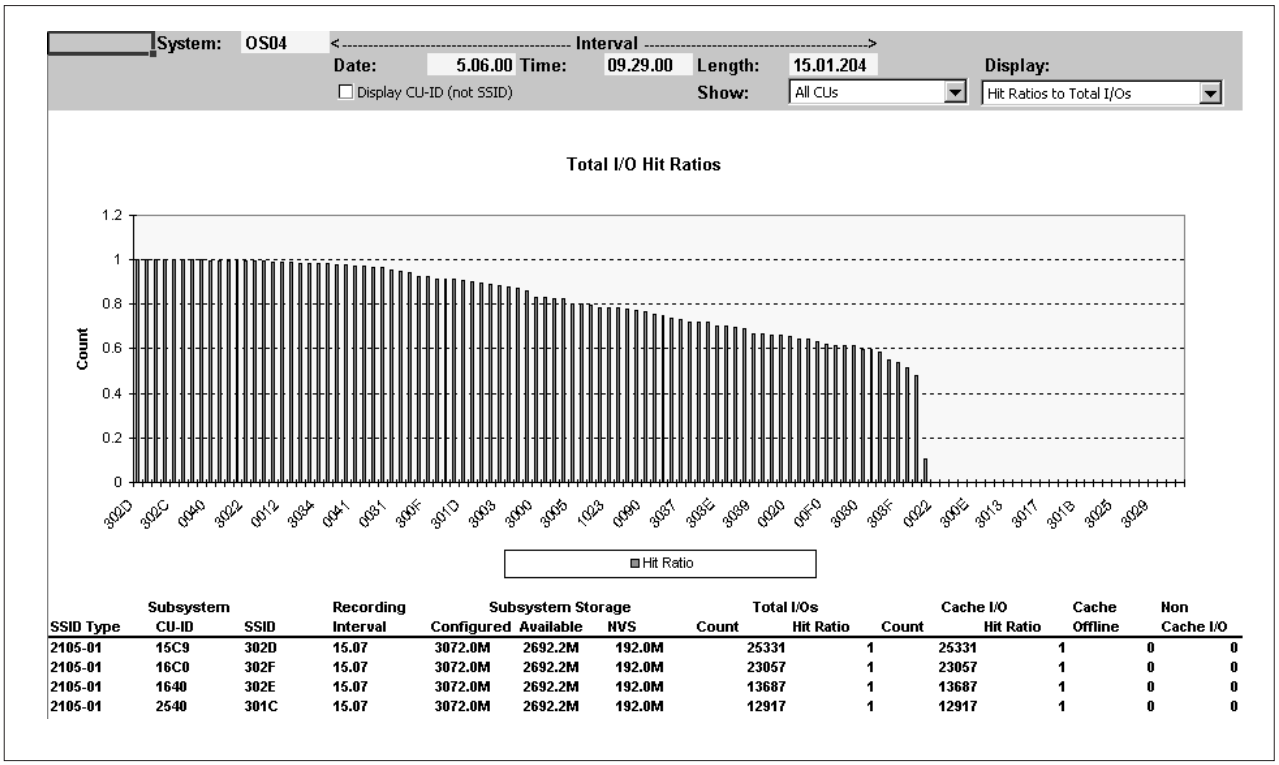
```
                                    R M F   O V E R V I E W   R E P O R T
                                                                                                    PAGE  001
           z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004         INTERVAL 14.59.996
                                  RPT VERSION V1R5 RMF      TIME 09.30.00            CYCLE 1.000 SECONDS

NUMBER OF INTERVALS 5            TOTAL LENGTH OF INTERVALS 01.14.57

DATE    TIME    INT    READRATE   WRTRATE    READHR     WRTHR     STGRT   DFWBPRT     ICLRT
MM/DD HH.MM.SS MM.SS
06/05 10.44.00 15.00     35.213    16.671     0.604     1.000    13.957     0.030     1.263
06/05 10.59.00 14.59      2.799    12.444     0.514     1.000     1.359     0.089     0.021
06/05 11.14.00 15.00      2.769     0.159     0.318     1.000     1.889     0.000     0.063
06/05 11.29.00 14.59      3.692     0.225     0.377     1.000     2.301     0.001     0.034
06/05 11.44.00 14.59      0.535     0.261     0.696     0.996     0.164     0.003     0.022
```

Figure 58. Overview Report for Cached DASD Device

In this report, you see that the volume was very active just in one interval, this means that further investigation is required to highlight the critical volumes of this Cache subsystem.

Using the Spreadsheet Reporter, you can create reports with key performance data you are interested in.

If you would like to see all SSIDs at glance, you can create the Cache Subsystem Report which displays all important cache characteristics. To get all details about cache rates for a longer period of time, you would use the Cache Trend Report.



SSID Type	Subsystem CU-ID	SSID	Recording Interval	Subsystem Storage Configured	Available	NVS	Total I/Os Count	Hit Ratio	Cache I/O Count	Hit Ratio	Cache Offline	Non Cache I/O
2105-01	15C9	302D	15.07	3072.0M	2692.2M	192.0M	25331	1	25331	1	0	0
2105-01	16C0	302F	15.07	3072.0M	2692.2M	192.0M	23057	1	23057	1	0	0
2105-01	1640	302E	15.07	3072.0M	2692.2M	192.0M	13687	1	13687	1	0	0
2105-01	2540	301C	15.07	3072.0M	2692.2M	192.0M	12917	1	12917	1	0	0

Figure 59. Cache Hits Overview Report. Spreadsheet Reporter macro RMFR9CAC.XLS (Cache Subsystem Report - SSIDOvw)

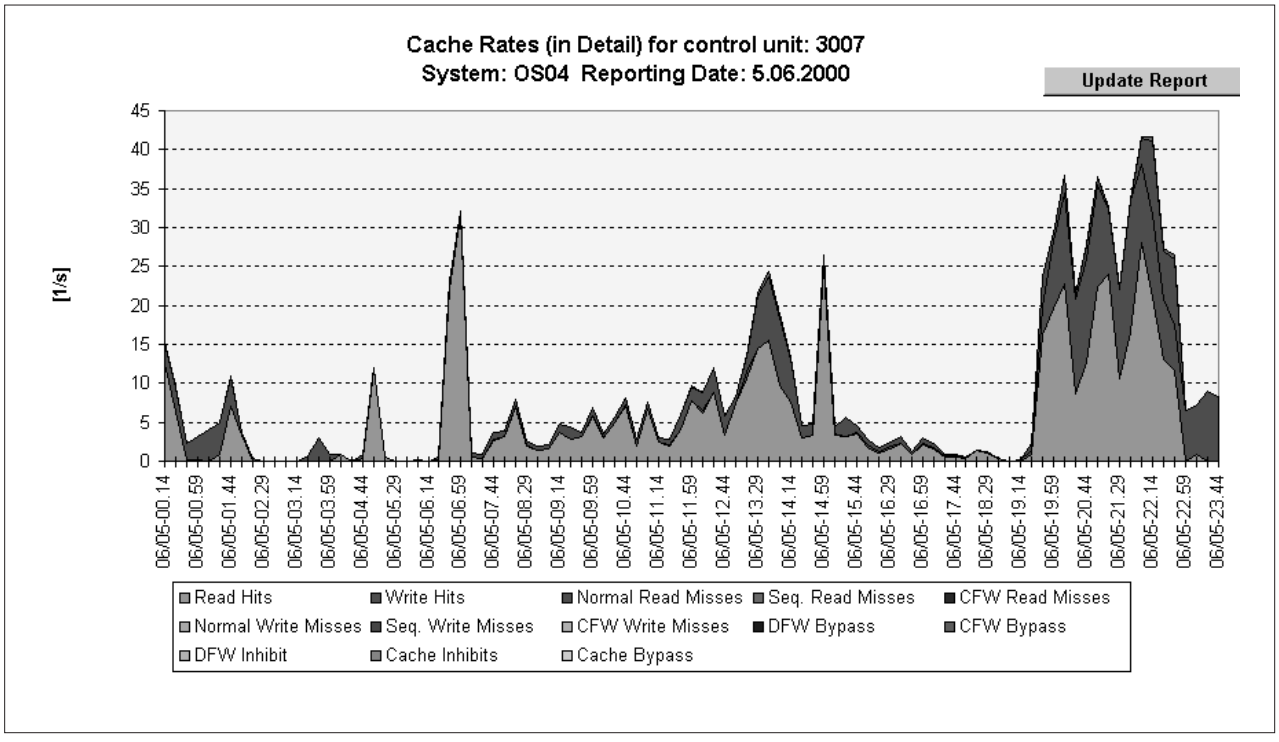


Figure 60. Cache Trend Report. Spreadsheet Reporter macro RMFX9CAC.XLS (Cache Trend Report - Summary)

Notes:

1. Cache controller information is collected by individual device addresses without segregation by issuing system. Therefore, the Monitor I Data Gatherer automatically collects the total for each device behind the storage controller. There is no way to differentiate the I/Os by each sharing system. Thus, data gathering is only required to be activated on one of the sharing systems.
2. The report uses cache storage subsystem ID (SSID) numbers to identify control units. The Monitor I DASD Activity report uses logical control unit (LCU) numbers to report on these same control units. LCU and SSID often will not be the same value, even when referring to the same control unit. You may need to compare the address range of devices on a given controller to match LCU to SSID.
3. Also, I/O rates shown in the Cache Subsystem Activity report may not exactly match I/O rates given in Shared Device Activity report (the sum of device activity from all sharing systems). This is because the storage controller counts each locate in a multiple locate CCW chain, whereas RMF only counts one I/O to start the CCW. An example of a CCW with multiple locates is DB2 list prefetch. Conversely, control unit commands, such as standalone RESERVEs, would get counted for the DASD Activity report but not for the Cache Subsystem Activity report.

Indicator

Field: Use the following fields:

- `TOTAL H/R - CACHE H/R` in the Subsystem Overview report and Device Activity report.
- `TOTAL H/R - READ H/R - WRITE H/R` in the Device Overview report

Guideline: Check whether all volumes (pay special attention to those that do a lot of I/O) are experiencing a good hit ratio (80% or higher). If this is the case, further analysis is not necessary. Note that the occurrence of a low hit ratio for one or more devices cannot be interpreted as a sure sign of the need for more cache (other analytical tools, such as the IBM Cache Analysis Aid, will be required for more thorough cache sizing). The low hit ratio may be caused by data sets which do not cache well, bypass cache, etc.

Problem Area: Low cache hit ratios will mean higher DISC times for that device. Also, devices with low cache hit ratios may use cache storage that other devices could have used more profitably.

Potential Solution: If you have DFSMS/MVS® and its *Dynamic Cache Management Extended (DCME)* capability, DCME will dynamically manage cache for those data sets you specify (via storage class). Thus no manual cache tuning should be required on your part.

If your data is not SMS-managed, you may wish to identify volumes which are not achieving a good cache hit ratio, and are doing a significant amount of staging, and turn them off for caching. This will improve the use of cache for volumes that benefit more from cache. A trial and error approach will be needed here; you will need to make a judgment call on whether to cache devices that are benefitting marginally from cache (maybe 30-80 % hit ratios).

Using Monitor III Cache Reporting

Monitor III offers two reports which can assist you in monitoring online the performance of your Cache subsystem:

CACHSUM Cache Summary report

CACHDET Cache Detail report

Both reports provide similar information as you can find in the Postprocessor Cache reports. Therefore, the performance indicators are the same, and you might refer to the description of the Postprocessor reports for further evaluation.

Cache Summary Report

The report provides an overview about the activities in the cache subsystem for all SSIDs. You might take this as starting point when analyzing I/O performance to get a first impression about the I/O processing.

If you feel that further analysis is required, you may continue with the Cache Detail report.

```
                    RMF V1R5  Cache Summary    - SYSPLEX        Line 1 of 21
Command ===> _                                                 Scroll ===> HALF

Samples: 120     Systems: 5    Date: 04/15/04  Time: 08.25.00  Range: 120   Sec
                             CDate: 04/15/04 CTime: 08.24.55 CRange: 120   Sec

SSID  CUID Type-Mod  Size  I/O   Hit  Hit   -- Miss ---  Read  Seq  Async  Off
                           Rate   %   Rate  Total Stage   %   Rate Rate   Rate

0010  071D 9396-001 1024M  99.8 98.8  98.6    1.2   1.2  99.2   1.6   0.1   0.0
0011  0520 9396-001 1024M   5.4  100   5.4    0.0   0.0   100   0.0   0.0   0.0
0030  0269 3990-006  256M  90.6 99.8  90.4    0.2   0.2   100   0.0   0.0   0.0
0040  0401 9393-002 1536M 142.0  100 141.9    0.0   0.0  97.7   0.0   0.3   0.0
0041  0460 9393-002 1536M   1.1  100   1.1    0.0   0.0  68.2   0.0   0.2   0.0
0042  05AA 9393-002 1536M   0.9  100   0.9    0.0   0.0  70.4   0.0   0.1   0.0
0043  05C6 9393-002 1536M  42.9 99.5  42.7    0.2   0.2  72.2   0.6   1.8   0.0
0044  0200 3990-006   64M 129.7 99.5 129.0    0.7   0.7  98.2   0.2   0.0   0.0
0060  0627 9393-002 1536M   9.2 29.3   2.7    0.0   0.0  40.3   0.0   0.4   0.0
0061  0654 9393-002 1536M   4.3 99.8   4.3    0.0   0.0  64.6   0.3   0.4   0.0
0062  06B0 9393-002 1536M   7.4 30.9   2.3    0.1   0.1  17.5   0.0   0.6   0.0
0063  06CE 9393-002 1536M   0.3  100   0.3    0.0   0.0  88.9   0.0   0.0   0.0
```

Figure 61. CACHSUM Report

Cache Detail Report

The CACHDET report provides detailed information about the activities of one cache subsystem.

```
                    RMF V1R5  Cache Detail     - SYSPLEX       Line 1 of 199
Command ===> _                                                Scroll ===> HALF

Samples: 120     Systems: 5    Date: 04/15/04  Time: 08.25.00  Range: 120   Sec
                              CDate: 04/15/04 CTime: 08.24.55 CRange: 120   Sec

Volume /Num SSID   I/O  I/O    Hit - Cache Hit Rate - - DASD I/O -  Seq   Async
                    %   Rate    %   Read   DFW   CFW  Total Stage   Rate  Rate

*ALL               100  42.9  99.5  30.8  11.9   0.0    0.2   0.2    0.6    1.8
*NOCAC             0.0   0.0   0.0   0.0   0.0   0.0    0.0   0.0    0.0    0.0
*CACHE             100  42.9  99.5  30.8  11.9   0.0    0.2   0.2    0.6    1.8
SYSPXC 05CB 0043  27.9  12.0   100  10.4   1.6   0.0    0.0   0.0    0.0    0.0
SYSSM3 05C3 0043  11.6   5.0   100   3.9   1.1   0.0    0.0   0.0    0.3    0.4
SYSAXC 05CA 0043  11.2   4.8   100   3.2   1.6   0.0    0.0   0.0    0.0    0.0
SYSSM6 05C6 0043  10.8   4.6  98.9   3.7   0.9   0.0    0.1   0.1    0.0    0.3
SYSSMS 05C0 0043   9.7   4.2  97.6   0.4   3.7   0.0    0.1   0.1    0.0    0.0
SYSSM8 05C8 0043   6.8   2.9   100   1.6   1.4   0.0    0.0   0.0    0.1    0.1
SYSSM5 05C5 0043   5.4   2.3   100   2.3   0.1   0.0    0.0   0.0    0.0    0.2
SYSSM2 05C2 0043   3.7   1.6   100   0.6   1.0   0.0    0.0   0.0    0.0    0.5
SYSSMB 05DA 0043   3.6   1.6   100   1.4   0.1   0.0    0.0   0.0    0.1    0.0
SYSSMA 05CD 0043   3.6   1.5  99.5   1.5   0.0   0.0    0.0   0.0    0.1    0.0
SYSSM7 05C7 0043   3.3   1.4  98.2   1.2   0.2   0.0    0.0   0.0    0.0    0.1
SYSOPE 05CC 0043   1.6   0.7  97.5   0.6   0.1   0.0    0.0   0.0    0.0    0.0
SYSSM4 05C4 0043   0.6   0.3  96.9   0.0   0.2   0.0    0.0   0.0    0.0    0.0
SYSSM9 05C9 0043   0.0   0.0   0.0   0.0   0.0   0.0    0.0   0.0    0.0    0.0
SYSSM1 05C1 0043   0.0   0.0   0.0   0.0   0.0   0.0    0.0   0.0    0.0    0.0
SYSSMC 05DB 0043   0.0   0.0   0.0   0.0   0.0   0.0    0.0   0.0    0.0    0.0
MVSWR1 05FF 0043   0.0   0.0   0.0   0.0   0.0   0.0    0.0   0.0    0.0    0.0
```

Figure 62. CACHDET Report

DASD Indicators and Guidelines

Response Time Components

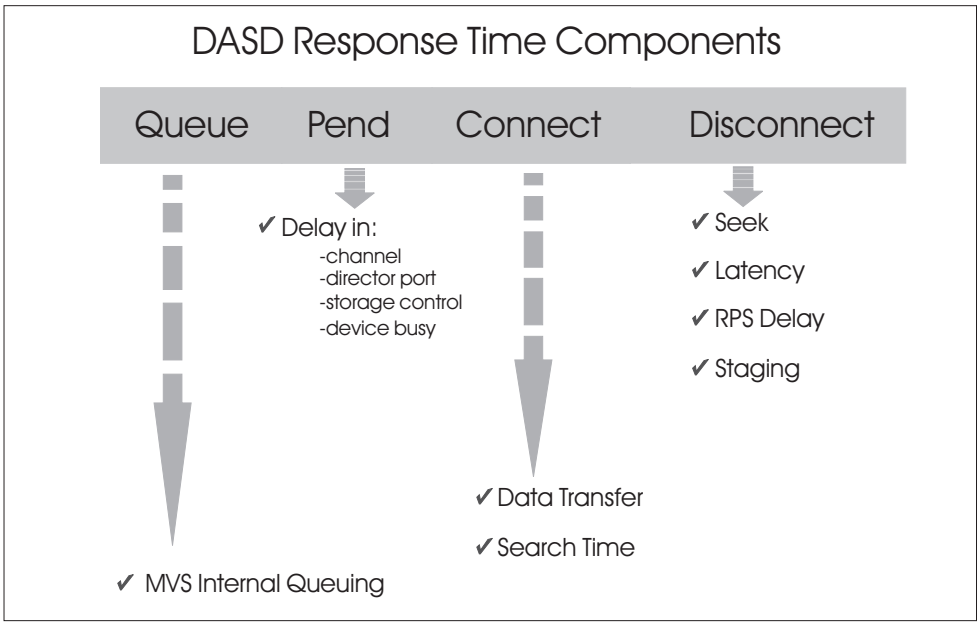


Figure 63. DASD Response Time Components

Response time is a key measure of how well your I/O is performing, and a key indicator of what can be done to speed up the I/O. It is important to understand the components of this response time, and what they mean. DASD response time is the sum of queue, pend, connect and disconnect time. The following lists the response time components and what causes them:

Connect The part of the I/O during which data is actually transferred, protocol, search and data transfer time.

Disconnect Time that an I/O request spends freed from the channel. This is the time that the I/O positions for the data that has been requested. It includes

SEEK and SET SECTOR

Moving the device to the requested cylinder and track

Latency

Waiting for the record to rotate under the head

Rotational Position Sensing (RPS)

Rotational delay, as the device waits to reconnect to the channel. The term "RPS delay" traditionally means waiting for an extra disk rotation because a transfer could not occur when the data was under the head. Delays of this type are now obsolete with the ESS.

Pend Time that the I/O is delayed in the path to the device. Pend time may be attributable to the channel, control unit, or wait for director (director port delay time), although it is often caused by shared DASD.

IOS Queue Represents the average time that an I/O waits because the device is already in use by another task on this system, signified by the device's UCBBUSY bit being on.

For most customers, reporting is based on one or both of the following measurements:

Service Time Connect time plus disconnect time

Response Time Connect time plus disconnect time plus pend time plus IOSQ time

Spreadsheet Reports

In Chapter 2, "Diagnosing a Problem: The First Steps," on page 23, you could see the DASD Summary report (Figure 30 on page 51) to be used for first steps in understanding the DASD performance of your system. If you feel that further investigation is required, you can create additional reports.

You might start with a second Summary report that provides data on the most heavily used LCUs and DASD devices:

RMF DASD System Summary

Sort Lists

System Id:	OS04	Operating System:	OS/390 02.10.00	Report Range(seconds):	899.628
Reporting Date:	6.05.00	Reporting Time:	09.29.00	Report Range(hh.mm.ss):	14.59.628

System Summary

#of LCUs	# of DASD	I/O Intens.	ST Intens.	Path Int.	Act. Rt.	Resp. Tm	Serv. Tm	IOSQ Tm	Pend. Tm	Disc. Tm	Conn. Tm	RT/ST
116	6588	13732.83	11426.68	6188.40	1990.76	6.90	5.74	0.55	0.60	2.63	3.11	1.20

LCU Summary

	LCU	I/O Intens.	ST Intens.	Path Int.	Act. Rt.	Resp. Tm	Serv. Tm	IOSQ Tm	Pend. Tm	Disc. Tm	Conn. Tm	RT/ST
Top 5	0057	1536.20	844.91	665.69	256.03	6.10	3.30	2.00	0.80	0.70	2.60	1.06
	005D	1127.20	1014.48	601.17	187.87	5.70	5.40	0.00	0.30	2.20	3.20	1.85
	012B	1067.62	843.42	261.57	53.38	19.30	15.80	3.00	0.50	10.90	4.90	1.22
	0056	957.13	669.99	622.14	239.28	3.90	2.80	0.00	1.10	0.20	2.60	1.39
	011D	949.83	846.21	155.43	86.35	10.60	9.80	0.00	0.80	8.00	1.80	1.08

sorted by I/O intensity

Device Summary Top 10

LCU	VolSer	I/O Intens.	ST Intens.	Path Int.	Act. Rt.	Resp. Tm	Serv. Tm	IOSQ Tm	Pend. Tm	Disc. Tm	Conn. Tm	RT/ST
0057	IDP507	913.50	475.02	328.86	182.70	5.30	2.60	2.00	0.70	0.80	1.80	2.04
005D	DIVP28	856.00	838.88	547.84	171.20	5.20	4.90	0.00	0.30	1.70	3.20	1.06
011D	NP0LBG	766.73	713.06	122.68	76.67	10.10	9.30	0.00	0.80	7.70	1.60	1.09
0056	CAT005	519.98	350.98	337.98	129.99	3.70	2.70	0.00	1.00	0.10	2.60	1.37
0057	IDP069	470.02	270.26	264.39	58.75	8.60	4.60	3.00	1.00	0.10	4.50	1.87
012B	NP0M15	341.60	245.95	63.20	17.08	20.80	14.40	6.00	0.40	10.70	3.70	1.44
00AC	IDP067	295.41	221.56	221.56	147.71	1.80	1.50	0.00	0.30	0.00	1.50	1.20
012B	NP0M14	179.95	152.53	44.56	8.57	21.40	17.80	3.00	0.60	12.60	5.20	1.20
012B	NP0M18	171.33	129.98	24.81	5.91	28.50	22.00	6.00	0.50	17.80	4.20	1.30
0056	IDP063	166.50	95.74	91.58	41.63	3.20	2.30	0.00	0.90	0.10	2.20	1.39

sorted by I/O intensity

Averages

I/O Intens. ST Intens. Path Int. Act. Rt. Resp. Tm Serv. Tm IOSQ Tm Pend. Tm Disc. Tm Conn. Tm RT/ST

Figure 64. DASD Summary Report. Spreadsheet Reporter macro RMFR9DAS.XLS (DASD Activity Report - Summary)

If you have seen some average data for most active volumes, you might be interested in getting some more details:

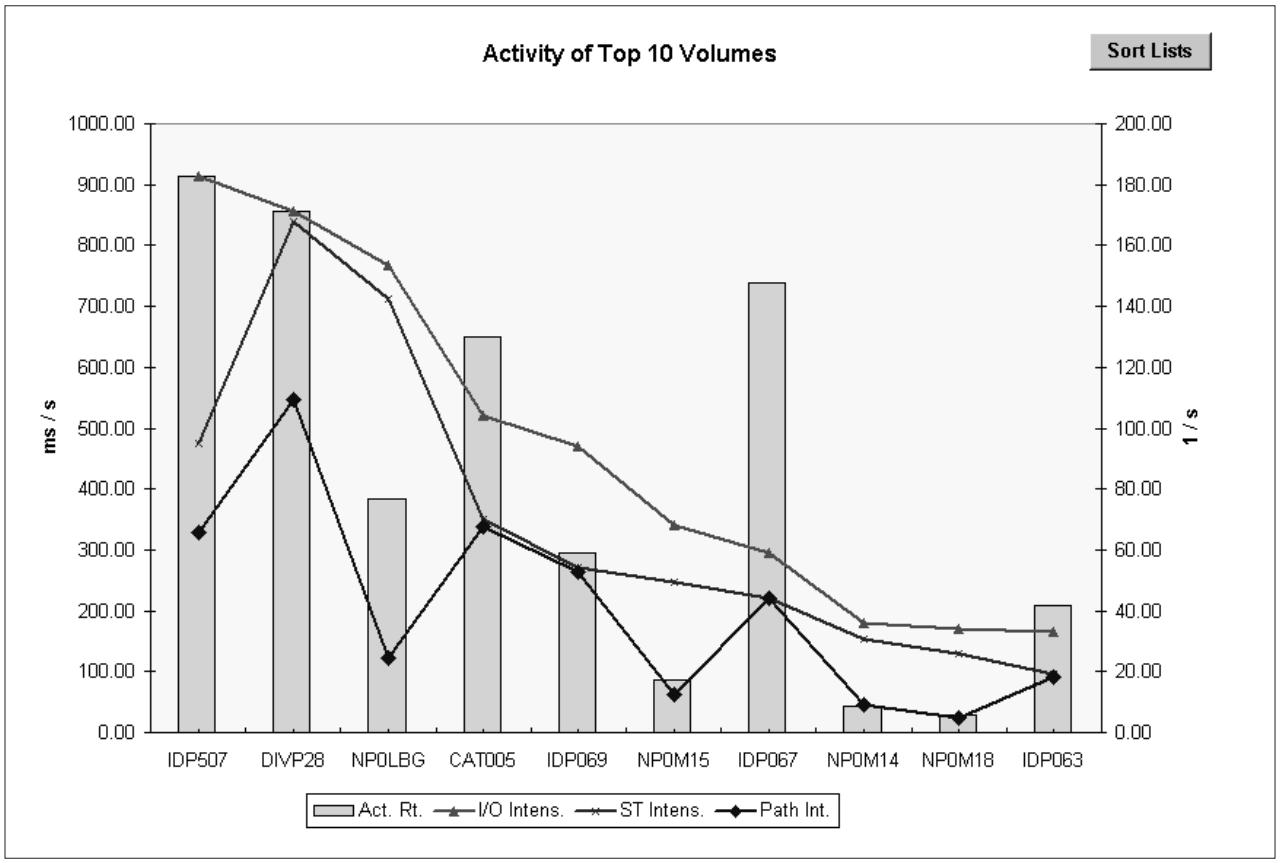


Figure 65. Activity of Top-10 Volumes. Spreadsheet Reporter macro RMFR9DAS.XLS (DASD Activity Report - Top10Act)

The report shows four values: in addition to the activity rate, you get the values for the I/O Intensity, the Service Time Intensity, and the Path Intensity.

I/O Intensity is the product of response time multiplied with the activity rate. The unit of measurement is ms/s.

It allows you to examine how many milliseconds per second applications waited for the device. The value can exceed 1000, which is an indicator that the device experiences a very heavy load and requires a long time to satisfy the requests. The value also includes the IOS Queue Time. If the activity rate is very low (e.g. less than 1 SIO/s) and the IOS Queue Time is high, the value is not very meaningful. You should use this measurement for devices with a considerable load (more than 2 IOs/s).

I/O Intensity is not a common name in the literature. Other references may use the same measurement with a different name, for example
Response Time Volume
DASD MPL (typically divided by 1000)

If you are interested in getting some details about the response times of the top-10 volumes, you can get this report:

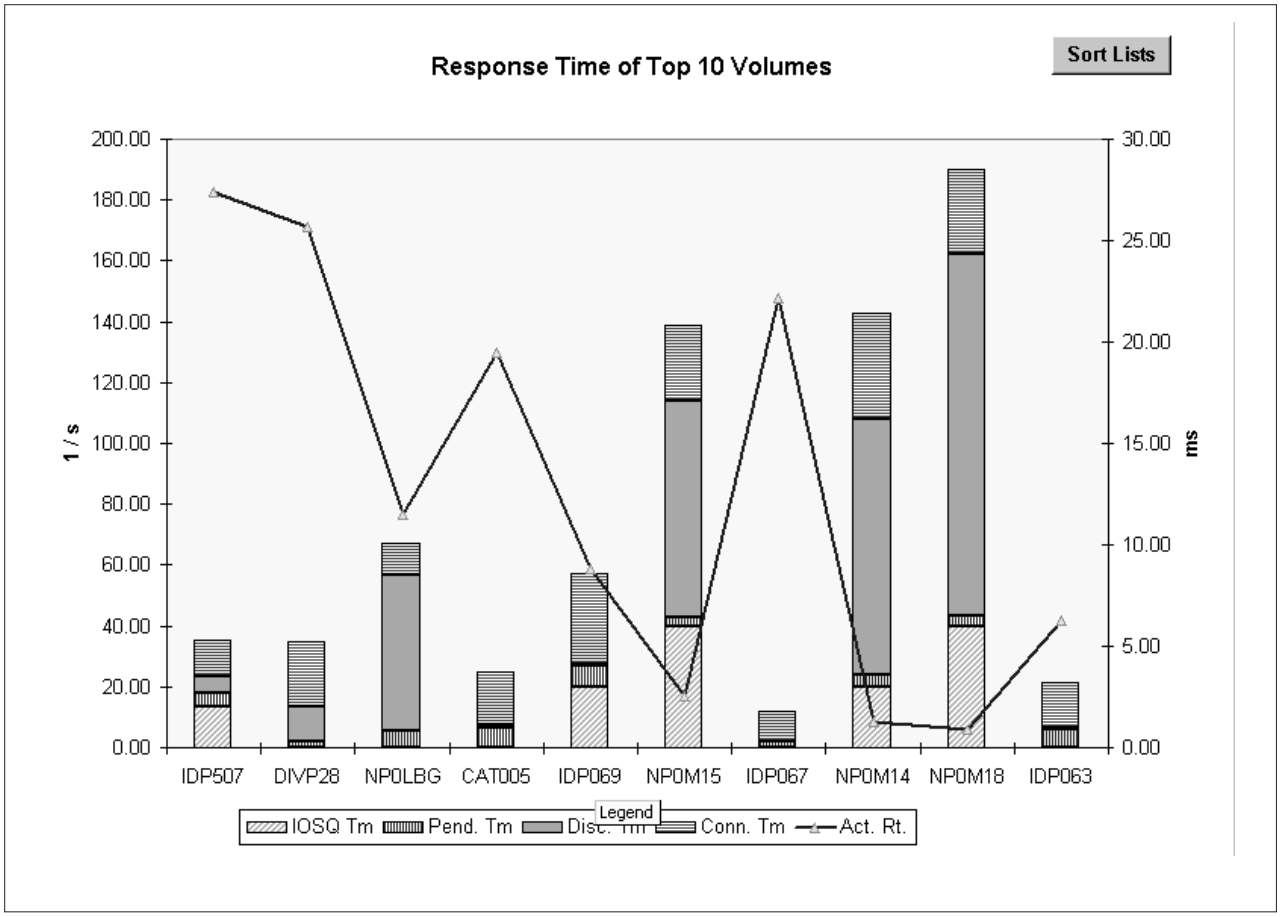


Figure 66. Response Time of Top-10 Volumes. Spreadsheet Reporter macro RMFR9DAS.XLS (DASD Activity Report - Top10Rt)

General DASD Guidelines

Following are some guidelines for expected values for typical DASD volumes. There are two practical ways to monitor DASD activities with RMF:

- The Monitor I DASD Activity report
- The Monitor III DEV and DEVR reports

The most common procedure is to analyze the Monitor I DASD Activity report.

Note: With any RMF reports, the I/O information reported is from one MVS system only. If you are sharing I/O across multiple systems, you will need to review RMF reports from each of the sharing systems in order to see the complete I/O activity to the LCUs and devices (this is not true for the Cache RMF Reporter, however; since CRR data is obtained from the control unit, it does include I/O activity from all sharing systems). Contention from a sharing system will generally be seen as increased pend and disconnect times.

DASD Activity Report

```
                                   D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y
                                                                                                                   PAGE    1
            z/OS V1R6               SYSTEM ID OS04             DATE 06/05/2004          INTERVAL 14.59.996
                                    RPT VERSION V1R5 RMF       TIME 09.30.00             CYCLE 1.000 SECONDS

TOTAL SAMPLES =  1,800    IODF = 12  CR-DATE: 03/01/04   CR-TIME: 10.23.17

                                         DEVICE   AVG  AVG   AVG  AVG  AVG   AVG  AVG  AVG    %      %      %    AVG    %      %
STORAGE  DEV  DEVICE    VOLUME PAV  LCU  ACTIVITY RESP IOSQ  DPB  CUB  DB    PEND DISC CONN   DEV    DEV    DEV  NUMBER ANY    MT
 GROUP   NUM  TYPE      SERIAL                 RATE     TIME TIME  DLY  DLY  DLY   TIME TIME TIME   CONN   UTIL   RESV ALLOC  ALLOC  PEND

         0800 3390      FRE800      0045   0.000   0.0  0.0   0.0  0.0  0.0   0.0  0.0  0.0   0.00   0.00   0.0   0.0  100.0   0.0
         0801 3390      DATA24      0045   2.092  18.1  0.0   0.0  0.0  0.0   0.2 14.9  3.0   0.63   3.75   0.0   1.3  100.0   0.0
         0802 3390      DATA25      0045   0.376  14.5  0.0   0.0  0.0  0.0   0.2 10.0  4.3   0.16   0.59   0.1   7.9  100.0   0.0
         0803 3390      SLB002      0045   0.003  11.5  0.0   0.0  0.0  0.0   0.2  9.0  2.3   0.00   0.00   0.0   1.7  100.0   0.0
         0810 3390      DATA26      0045   0.112  15.0  0.0   0.0  0.0  0.0   0.2 12.2  2.6   0.03   0.17   0.1   8.5  100.0   0.0
         0811 3390      DATA33      0045   0.051  15.1  0.0   0.0  0.0  0.0   0.2 12.5  2.4   0.01   0.13   0.1   0.0  100.0   0.0
         0812 3390      DATA34      0045   0.057  16.3  0.0   0.0  0.0  0.0   0.2 13.4  2.7   0.02   0.09   0.1   2.6  100.0   0.0
         0813 3390      DATA35      0045   0.046  15.9  0.0   0.0  0.0  0.0   0.2 13.4  2.3   0.01   0.07   0.0   8.0  100.0   0.0

                         LCU        0045   2.736  17.3  0.0   0.0  0.0  0.0   0.2 14.0  3.1   0.11   0.60   0.0  30.1  100.0   0.0
```

Figure 67. Monitor I DASD Activity Report

In this report, the fields of most interest are:

```
DEVICE ACTIVITY RATE
AVG RESP TIME
AVG IOSQ TIME
AVG PEND TIME
AVG DISC TIME
AVG CONN TIME
% DEV UTIL
```

Indicator

Field: AVG RESP TIME

Description: The average response time, in milliseconds, for an I/O to this device or LCU.

Related Fields: DEVICE ACTIVITY RATE, AVG PEND TIME, AVG DISC TIME, AVG CONN TIME, AVG IOSQ TIME

Guideline: For most situations, this measurement is the best way to quickly determine how well a device is performing.

AVG RESPONSE TIME can be considered as the overall measure of the health of a device's operation.

It is the sum of the AVG IOSQ TIME, AVG PEND TIME, AVG DISCONNECT TIME, and AVG CONNECT TIME. If response time is too high, then you can look further at these response time components and take remedial action.

What is too high? If your service level objectives are being met, then your response time is OK. If not, analyze what job is being delayed, and for how long (see prior section). Check the I/O rate to the device: if the I/O rate is insignificant, why bother tuning for response time?

Problem Area: I/O response time is a component of user response time; often it is the dominant component. High I/O response can delay user jobs and increase response times to unacceptable levels.

Potential Solution: Determine the dominant response time component, and see "Improving Your DASD Performance" on page 125 for suggestions.

Indicator

Field: AVG IOSQ TIME

Description: This is the time measured when a request is being delayed in the MVS system.

Related Fields: DEVICE ACTIVITY RATE, AVG RESP TIME

Guideline: If IOSQ is greater than 1/2 times *service time* (where service time is DISC + CONN) then it warrants a closer look. In practice, there should be little queue time (less than 5 ms).

Problem Area: High IOSQ times contribute to high AVG RESP TIME. Combined with high I/O rates for important workloads, this could become a user response time problem.

Potential Solution: If your problem is high IOSQ time then you have the following options. Traditionally IOSQ problems are usually resolved by data movement either by separating active data sets or by moving active data to faster storage, for example, a coupling facility structure.

On an ESS subsystem, you have additional options.

- If PAV is not enabled for the device, enable it.
- If you are using static PAVs, assign more aliases to the device.
- If you are using dynamic PAV, then increase the number of PAVs associated in the pool for the subsystem.
- Check to ensure that all PAVs that should be bound to the device are online and operational. You can use the DEVSERV QP and DS QP,xxxx,UNBOX commands to do this.

Indicator

Field: AVG PEND TIME

Description: This value represents the average time (ms) an I/O request waits in the hardware (channel subsystem).

Related Fields: AVG RESP TIME, DEVICE ACTIVITY RATE

Guideline: A high PEND time suggests that the channel subsystem is having trouble initiating the I/O operation. There is a blockage somewhere in the path to the device. That might be due to:

- AVG DPB DELAY. Delay due to ESCON director ports being busy.
- AVG CUB DELAY. Delay due to the DASD control unit being busy, due to I/O from another sharing MVS system.
- AVG DB DELAY. Delay due to the device being busy, due to I/O from another sharing MVS system.
- Channel path wait. Whatever PEND time is not accounted for by the above three measures is due to delay for channel paths. This is the measure of channel delay that matters - not channel busy. If you think your channels are too busy, track this component of response time, for volumes that serve important work, to see if it is really a problem.

If you think shared DASD causes the problem, look at the DASD Activity report from the other system, taken at the same time. PEND time should generally be less than 4 ms.

Problem Area: High PEND times contribute to high AVG RESP TIME. Combined with high I/O rates for important workload, this could become a user response time problem.

Potential Solution: High PEND times are usually caused by shared DASD contention or high channel path utilization.

On an ESS subsystem, multiple allegiance should reduce PEND times. If problems with PEND times exist you have the following options.

- Change the mix of data on the volume to reduce contention. If you can identify one data set that is contributing to most of the problem this may be eligible to be moved to a custom volume or moved into storage.
- Check channel utilization, especially if not using eight-path path groups. Changes have been made to CCWs to reduce channel overheads for the ESS. This will tend to lower channel utilization and increase throughput, however, using multiple allegiance and PAV successfully will cause increases in channel utilization.

Indicator

Field: AVG DISC TIME

Description: For cache controllers, DISC includes the time waiting while staging completes for the previous I/O to this particular device or until one of the four lower interfaces frees up from either transferring, staging or destaging data for other devices.

DISC can also be caused by a device which looks for the correct location on the disk to read or write the data.

Related Fields: AVG RESP TIME, DEVICE ACTIVITY RATE

Guideline: Generally, DISC time should be less than 17 ms. If the data is cached with reasonable hit ratios there should be little DISC time (3-5 ms).

Once your DISC time gets above 12 ms (about the amount of time you would expect for latency and typical seek times), you are probably experiencing RPS misses due to path contention.

Problem Area: High DISC times contribute to high AVG RESP TIME. Combined with significant I/O rates for important workloads, this could become a user response time problem.

Potential Solution: If the major cause of delay is disconnect time then you will need to do some further research to find the cause. You might use the *ESS Expert* for creating more detailed reports. The most probable cause of high disconnect time is having to read or write data from disks rather than cache. You need to check for the following conditions.

- High DISK to CACHE transfer rate, check the ESS Expert Disk to Cache Transport report and use the drill-down functions of the ESS Expert to identify the logical volumes that are experiencing response time problems.
- High disk utilization, check the ESS Expert Disk Utilization report, if the problem is limited to one physical disk, or array the best solution is to move data to balance performance across the subsystem. The other option is to move data to another subsystem, or to another storage medium, for example data in memory, or to change the way that the application uses this data.
- Low CACHE hit ratio — seen in the Cache Subsystem Device Overview report. If you are suffering from poor cache hit ratios there is little that you can do, you should look at the ESS as a whole, using ESS Expert and check that activity is balanced across both clusters of the ESS.
- NVS full condition, if the NVS is overcommitted you will see an increase in the values reported in the DFW BYPASS field of the Cache Subsystem Overview report. If this is a persistent performance indicator you may want to spread activity across more ESS subsystems.

.

Indicator

Field: AVG CONN TIME

Description: This is time as measured by the channel subsystem during which the device is actually connected to the CPU through the path (channel, control unit, DASD) and transferring data. This time is considered good in most cases, because it is transferring data.

Related Fields: AVG RESP TIME, DEVICE ACTIVITY RATE

Guideline: Connect time is a function of block size and channel speed. Typical connect times often fall in the 2-6 ms range.

Problem Area: There are times when doing large searches in catalogs, VTOC, and PDS directories causes the connect time to become excessive. This results in connect time being considered ″expensive″ time in that no real data is transferred during this activity. Also during connect time, all parts related to an I/O operation (the entire path) must be available for use by that I/O. So no other user can do I/O to or from a device while that device is in connect mode.

One word of caution here - chained scheduling and seldom ending channel programs may generate high connect time which is not indicative of a performance problem.

Potential Solution: If you see devices with poor performance and high connect times then the cause is probably application related, high connect times are associated with large data transfers, either due to the use of large block sizes or to activities like DB2 prefetch that schedule large I/O transfers. If the application is not reporting poor response and other users of the volume are not impacted, then no further work is required. High connect time is an indication of large amounts of data being transferred. An application, such as DB2, transfers large I/Os for maximum efficiency.

Indicator

Field: % DEV UTIL

Description: This is the percentage of time the UCB is ″busy″ (see IOSQ time). This includes both the time when the device was involved in an I/O operation (CONN and DISC) as well as any time it was reserved but not involved in an I/O operation.

Related Field: AVG IOSQ TIME

Guideline: A device overload condition is not always easy to detect. The device utilization as reported by RMF is not a measure of goodness or badness. If there is only one active data set on the device, and presuming that this data set cannot be made resident in processor storage, there is nothing wrong with a very high device utilization.

However, if multiple online users are all trying to get to data sets on the same device, they can be delayed due to the device being busy (from the other users). This will show up as increased IOSQ time.

If the utilization rises above 35%, response times may start to suffer, due to increased queuing for the device. If you have online response-sensitive data on a non-cached volume, you may want to investigate if utilization rises above the 25-30% range.

Problem Area: High device utilization can lead to increased IOSQ time, if users are delayed while the device is in use by others.

Potential Solution: Potential ways to decrease device utilization include:

- Caching the device (if not already cached). With a good hit ratio service time may be reduced significantly, which will reduce device utilization dramatically.
- Distribute heavily used data sets onto different volumes.
- Remove poor cache users. Turn off caching for data sets or volumes that do not achieve good hit ratios. As discussed before, this frees up cache for users that benefit more.

Indicator

Field: Multiply DEVICE ACTIVITY RATE by AVG CONN TIME

Description: Utilization (% busy) of the paths to DASD devices.

Related Fields: AVG PEND TIME, AVG DISC TIME

Guideline: The maximum allowable utilization is not a fixed number, but it can be higher as the overall hit ratio increases. In an average 4-path system (DLSE), cached or non-cached, a general guideline is not to exceed 50%, if the goal is to maintain good response for interactive users.

If the workload using these paths is not response-time sensitive, you may not care about high path utilization. For example, if you have only four jobs accessing four paths, then there is no delay and high path utilization is desired.

Since the storage path utilization is not directly reported, an estimate can be made using the RMF reported LCU I/O rate and the average LCU connect time. This amounts to a simple calculation, which you should execute for each system that accesses devices attached to the storage control unit:

1. Multiply I/O rate by average connect time
2. Divide the result by four (a 3990 has four storage paths)
3. Divide by 10 to convert to a percentage (because I/O rate is given in I/O per second, while connect time is reported in millisecond time units)

Problem Area: Higher utilizations will be an issue if they are causing delay to important work. This delay will show up as increased pend and disconnect times for volumes sharing these paths.

Potential Solution: If your storage path utilization is above 50% and causing delay to important work, a potential solution is to balance activity across control units by moving data sets.

Indicator

Field: AVG DISC TIME, ASYNC and STAGE rates

Description: These fields give you an indication of "lower interface" contention. The lower interfaces are the connections between the DASD devices and the DASD control unit.

Related Fields: LCU cache hit ratio

Guideline: Direct measure of lower interface utilization is not available from any reporting tool, but normally you should not have to worry about the lower interfaces. If the utilization of the storage paths and the hit ratios are within acceptable limits, the lower interface utilization should be fine.

The lower interfaces can be overloaded on cache controllers, however, if the hit ratio for the storage control unit falls below about 70% and a high subsystem throughput is imposed.

Problem Area: Problems here would be reflected in increased DISC times. If you suspect a problem, check your STAGE and ASYNC rates from the Cache RMF Reporter. Correlate both STAGE and ASYNC rates to your DISC times (for example, plot DISC time vs. STAGE + ASYNC rates).

Potential Solution: If you see a significant correlation between STAGE + ASYNC and DISC, investigate cache tuning or workload movement to reduce cache overload.

DASD Performance in the Sysplex

The Shared Device Activity report is a sysplex report that is available for DASD and tape devices. In this context we will discuss the DASD version of this report.

Note: The report requires matching device numbers (the physical device must have the same device number on all systems), or self-defining devices to give meaningful results.

Shared DASD Activity Report

```
                    S H A R E D   D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y
                                                                                    PAGE  1
           z/OS V1R6              SYSPLEX RMFPLEX1          DATE 04/10/2004          INTERVAL 014.59.946
                                  RPT VERSION V1R5 RMF      TIME 16.30.00            CYCLE 1.000 SECONDS

   TOTAL SAMPLES(AVG) =  900.0 (MAX) =  900.0 (MIN) =  900.0

                       SMF            DEVICE   AVG  AVG   AVG  AVG  AVG   AVG   AVG  AVG    %      %     %    AVG
DEV  DEVICE  VOLUME PAV SYS  IODF LCU  ACTIVITY RESP IOSQ  DPB  CUB  DB    PEND  DISC CONN   DEV    DEV   DEV  NUMBER
NUM  TYPE    SERIAL     ID   SUFF      RATE     TIME TIME  DLY  DLY  DLY   TIME  TIME TIME   CONN   UTIL  RESV ALLOC

06AA 3380E   DPVOL1     *ALL           14.589   3.5  0.0   0.0   0.0  0.2   0.4   0.6  2.5   3.63   4.45   0.1   651
                        MVS1  01  0043   4.550   4.8  0.0   0.0   0.0  0.0   0.2   0.5  4.1   1.86   2.10   0.0   205
                        MVS2  02  0143   5.489   3.1  0.0   0.0   0.0  0.3   0.8   0.6  1.7   0.91   1.24   0.1   210
                        MVS3  03  0043   4.550   3.8  1.0   0.0   0.0  0.2   0.3   0.6  1.9   0.86   1.11   0.0   236

077A 3380E   ERBDAT     *ALL           15.186   9.3  0.5   0.0   0.0  0.2   1.2   5.2  2.4   3.62  11.48   0.4  1078
077A                    MVS1  01  0043   4.069   5.2  2.0   0.0   0.0  0.8   1.1   0.4  1.7   0.68   0.84   0.0   343
0872                    MVS2  02  0143   6.438  14.8  0.5   0.0   0.0  0.0   1.0  10.2  3.1   2.00   8.58   0.0   357
077A                    MVS3  03  0052   4.679   8.0  2.0   0.0   0.0  0.1   1.6   2.4  2.0   0.94   2.06   0.4   378
```

Figure 68. Shared DASD Activity Report

This example reports about a sysplex consisting of three systems (MVS1, MVS2, and MVS3). Only two devices are shown.

The second device does not have the same device number on all three systems.

Indicator

Field: *ALL

Description: The system line shows the device activity contributed by all systems in the sysplex.

Guideline: The report gives you an overall performance picture of DASD devices that are shared between MVS systems in a sysplex.

For each shared DASD device the report contains one line for each system that has access to it.

The summary line allows you to identify a bottleneck caused by device delay in the sysplex. Furthermore, it allows you to see each systems share in the bottleneck.

The summary device activity rate and the device utilization show the total load on the device. The single-system values show the share of each system.

Analyzing Specific DASD Problems

This section will focus on how to analyze specific DASD problems. It is assumed that you have a good overall knowledge of your DASD subsystem (as discussed in "I/O Subsystem Health Check" on page 84) before proceeding.

One approach to investigating I/O problems can be summarized:
1. Is there a problem?
2. Who's got it?
3. Do I care?
4. What can be done about it?

Is there a problem? The Monitor III SYSINFO report shows whether there is any device delay. It also measures the size of the problem by the number of users affected.

SYSINFO Report

```
                    RMF V1R5  System Information              Line 1 of 27
Command ===>                                                 Scroll ===> HALF

Samples: 100     System: PRD1  Date: 07/28/04  Time: 10.32.00  Range: 100   Sec

Partition:    MVS1    9672 Model RX4          Appl%:      63  Policy: STANDARD
CPs Online:      4    Avg CPU Util%:  73      EAppl%:     65  Date:   07/28/04
IFAs Online:     0    Avg MVS Util%:  84      Appl% IFA:   0  Time:   14.05.07

Group     T WFL --Users--  RESP TRANS -AVG USG-   -Average Number Delayed For -
             %   TOT  ACT  Time  /SEC PROC  DEV   PROC  DEV STOR SUBS OPER  ENQ

*SYSTEM      34  664   26       13.95  5.1  5.0    1.9  4.1  7.0  2.6  2.0  2.0
*TSO         50  534    8       13.95  2.6  2.1    0.4  1.5  2.0  0.8  0.0  0.0
*BATCH       26   11   10        0.00  1.5  1.4    1.4  1.7  0.5  1.8  1.0  2.0
*STC         27  115    8        0.00  1.1  1.5    0.1  1.0  4.5  0.1  1.0  0.0
*ASCH             3    0        0.00  0.0  0.0    0.0  0.0  0.0  0.0  0.0  0.0
*OMVS             2    0        0.00  0.0  0.0    0.0  0.0  0.0  0.0  0.0  0.0
*ENCLAVE      5    4  N/A         N/A  0.2  N/A    3.7  N/A  0.0  N/A  N/A  N/A
PRIMEBAT  W  26   11   10  46.0  0.06  1.5  1.4    1.4  1.7  0.5  1.8  1.0  2.0
NRPRIME   S  26   11   10  46.0  0.06  1.5  1.4    1.4  1.7  0.5  1.8  1.0  2.0
          1  23    9    9  27.9  0.06  0.9  1.4    0.8  1.6  0.5  1.8  1.0  2.0
          2  29    0    0  54.2  0.02  0.1  0.0    0.2  0.0  0.0  0.0  0.0  0.0
          3  59    1    1  .000  0.00  0.6  0.0    0.4  0.0  0.0  0.0  0.0  0.0
PRIMETSO  W  50  527    8  .759 13.98  2.6  2.1    0.4  1.5  2.0  0.8  0.0  0.0
TSOPRIME  S  50  527    8  .759 13.98  2.6  2.1    0.4  1.5  2.0  0.8  0.0  0.0
          1  48  526    8  .403 13.98  2.1  1.9    0.3  1.3  2.0  0.8  0.0  0.0
          2  75    1    1  30.6  0.08  0.3  0.2    0.1  0.1  0.0  0.0  0.0  0.0
          3  75    0    0   126  0.02  0.1  0.0    0.0  0.0  0.0  0.0  0.0  0.0
```

Figure 69. Monitor III System Information Report

Who's got it? Look at the Delay report shown in 70. Who (which address space) has a device delay problem? This is more important than the number of users affected. Remember, I/O service is not democratic. One important job or transaction being delayed can be much more significant than several other transactions or jobs. The Delay report shows where delays exist by specific resource for specific users. This aids in the assessment of performance impact: "Who will be helped? How much?"

DELAY Report

```
                        RMF V1R5  Delay Report                  Line 21 of 49
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

            Service     WFL USG DLY IDL UKN ----- % Delayed for -----  Primary
Name     CX Class    Cr  %   %   %   %   %  PRC  DEV STR SUB OPR  ENQ  Reason

MISTYDFS B  NRPRIME       0   0 100   0   0   0    0   0   0 100    0  Message
BHOLEQB  B  NRPRIME       0   0  49   0   0   0    0   0   0   0   49  SYSDSN
BHOLWTO1 B  NRPRIME       0   0  33   0   1   0    0   0   0  33    0  Message
BRHI     T  TSOPRIME      0   0   4   0   0   0    0   4   0   0    0  COMM
CONSOLE  S  SYSSTC        0   0   3   0  97   0    0   3   0   0    0  LOCL
VTMLCL   S  SYSSTC        0   0   2   0  98   0    0   2   0   0    0  LOCL
STARTING T  TSOPRIME      0   0   2   0   7   0    0   2   0   0    0  COMM
CATALOG  S  SYSSTC        0   0   1   0  99   0    0   1   0   0    0  LOCL
INIT     S  SYSSTC        0   0   1  49   0   0    0   1   0   0    0  LOCL
BHOLDEV3 B  NRPRIME      20  20  80   0   0   1   79   0   0   0    0  BBVOL1
BTEUDASD B  NRPRIME      22  22  77   0   1   1   76   0   0   0    0  BBVOL1
BAJU     T  TSOPRIME     24  24  75   0   1   0   75   0   0   0    0  BBVOL1
KLSPRINT B  NRPRIME      25  25  75   0   0   0   75   0   0   0    0  BBVOL1
BHIM     T  TSOPRIME     31  11  25  64   0   2   18   5   0   0    0  BBVOL1
BHOLPRO1 B  NRPRIME      52  13  12   0   0  12    0   0   0   0    0  BHOLPRO2
```

Figure 70. Monitor III Delay Report

Do I care? Who is being delayed, and how much, often determines whether any action is necessary. Very important work requires immediate attention. Less important work? Small delays? Well, maybe you can look at it tomorrow.

What can be done about it? To identify the solution to an I/O problem, you must identify where the I/O pain is.

Remember, Monitor III can direct you to any device that is causing delays. This point of view is important; just because a device might have a high activity rate or a high response time, it MAY NOT be causing significant delays, or it may be delaying work you do not care about.

The reports that are useful here include:

DEV Report

```
                        RMF V1R5  Device Delays                  Line 1 of 9
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

                     DLY USG CON  ------------ Main Delay Volume(s) ------------
Jobname  C DMN  PG    %   %   %    %  VOLSER   %  VOLSER   %  VOLSER   %  VOLSER

BHOLDEV3 B  NRPRIME  79  15   4   72 BBVOL1
BTEUDASD B  NRPRIME  76  20   4   75 BBVOL1
KLSPRINT B  NRPRIME  75  23   4   75 BBVOL1
BAJU     T  TSOPRIME 75  19   1   71 BBVOL1
BHIM     T  TSOPRIME 18   6   1   16 BBVOL1    1 430D13
BHOL     T  TSOPRIME  4  11   6    9 BBVOL1
JES2     S  SYSSTC    4   3   0    6 SYSPAG
BHOLNAM4 B  NRPRIME   1   0   0    1 BBVOL1
```

Figure 71. Monitor III Device Delays Report

The Device Delays report shows jobs in your system that are delayed due to device contention.

Jobnames (address space names) are listed in descending order in delay percentages. For each jobname, the names of up to 4 device volumes are listed indicating which volumes are causing the delay. If a job is being delayed, you can immediately see which volume is causing the most delay.

DEVR Report

```
                RMF V1R5  Device Resource Delays              Line 1 of 16
Command ===>                                                 Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

Volume S/   Act  Resp  ACT CON DSC  PND %,   DEV/CU              Service USG DLY
  /Num PAV  Rate Time   %   %   %   Reasons  Type      Jobname  C Class    %   %

BBVOL1   41.8 .099   88  21  66  PND   1  3380K     KLSPRINT B  NRPRIME  6  92
  0222                                    3990-2    BHOLDEV3 B  NRPRIME  2  92
                                                    BTEUDASD B  NRPRIME  5  91
                                                    BAJU     T  TSOPRIME 3  91
                                                    BHIM     T  NRPRIME  1  22
                                                    BHOL     T  TSOPRIME 3  10
                                                    RMFGAT   S  SYSSTC   0   1
                                                    BHOLNAM4 B  NRPRIME  0   1
                                                    BHOLPRO1 B  NRPRIME  0   1
SYSPAG   10.9 .028   19   6  13  PND   0  3380K     BVAU     T  TSOPRIME 6  14
  0225                                    3990-2    JES2     S  SYSSTC   0   7
430D13    1.1 .032    4   2   2  PND   0  3380K     BENK     T  TSOPRIME 0   1
  0220                                    3990-2    BHIM     T  TSOPRIME 0   1
                                                    BHOL     T  TSOPRIME 2   0
SYSCAT    0.3 .017    0   0   0  PND   0  3380K     INIT     S  SYSSTC   0   1
  0221                                    3990-2
```

Figure 72. Monitor III Device Resource Delays Report

The DEVR report is used to evaluate the performance of each volume. It differs from the DEV report in that it lists a volume and all of the jobs that are being delayed for that volume. Thus, if you find a volume that RMF suggests is performing poorly, you can use Monitor III to determine which jobs it is affecting and how badly it is affecting any particular job.

You can get similar information from the Monitor I DASD Activity report.

Report Analysis

- In this case, from the SYSINFO report (shown in 69) we know that 1.7 jobs in service class NRPRIME are delayed because of a device.
- From the Delay report (shown in 70) we know the jobs in NRPRIME are waiting for volume BBVOL1.
- From the DEVR report (shown in 72) we identify that BBVOL1 is a 3380K that is not cached.

Replacing this disk by a faster device would be appropriate.

Performance Analysis on Data Set Level

The following sample reports do not show specific problems but can illustrate how performance analysis can be done on data set level.

If you need further analysis for a specific volume that you see in the DEVR report, you can navigate through cursor-sensitive control to some reports which provide information on data set level.

```
                    RMF V1R5  Device Resource Delays           Line 1 of 19
Command ===> _                                                Scroll ===> HALF

Samples: 76      System: D3    Date: 05/03/04  Time: 12.39.00  Range: 120   Sec

Volume    Act  Resp  ACT CON DSC  PND %,   DEV/CU              Service  USG DLY
  /Num S  Rate Time   %   %   %   Reasons  Type     Jobname  C Class     %   %

COUPLB S   2.4 .115   21   9   0  PND  12  33903    XCFAS    S SYSSTC    16   7
  9191                              DB    8  3990-3
                                    DPB   3
COUPLP S   0.8 .224   18   1   0  PND  17  33903    XCFAS    S SYSSTC    22   5
  91CD                              DB   15  3990-3
SGT103 S   0.7 .056    3   0   1  PND   2  33903    U015074  T TSOSLOW    1   1
  9182                              DPB   1  3990-3   IXGLOGR  S SYSTEM     1   1
                                                      CATALOG  S SYSTEM     0   1
                                                      U015041  T TSOSLOW    0   1
CMNAF8 S   0.5 .105    5   0   0  PND   5  33903    CATALOG  S SYSTEM     0   4
  BAF8                              DB    5  3990-6   RMFGAT   S SYSSTC     1   1
HDCAT  S   2.3 .009    0   0   0  PND   0  33903    CATALOG  S SYSTEM     1   1
  910E                                     3990-3   *MASTER* S SYSTEM     1   0
```

Figure 73. Monitor III - Device Resource Delays Report

If you select volume SGT103 and click on Enter, you get this report:

DSNV Report

```
                    RMF V1R5  Data Set Delays - Volume          Line 1 of 5
Command ===> _                                                Scroll ===> HALF

Samples: 76      System: D3    Date: 05/03/04  Time: 12.39.00  Range: 120   Sec

------------------------- Volume SGT103 Device Data -------------------------
Number:    9182        Active:       3%        Pending:    2%     Average Users
Device:    33903       Connect:      0%        Delay DB:   0%        Delayed
Shared:    Yes         Disconnect:   1%        Delay CU:   0%          0.1
                                               Delay DP:   1%
-------------- Data Set Name ---------------   Jobname   ASID  DUSG% DDLY%
IXGLOGR.ATR.UTCPLXHD.RM.DATA.A0000010.DATA     IXGLOGR   0021      1     1
U015074.CALENDAR.DATA                          U015074   0110      0     1
U015041.COURSEIN.DATA                          U015041   0125      0     1
SYS1.VVDS.VSGT103                              CATALOG   0051      0     1
SYS1.VTOCIX.VSGT103                            U015074   0110      1     0
```

Figure 74. Monitor III - Data Set Delays by Volume

In addition to general information about this volume, you see a list with all data sets which are currently used. Selecting a specific job, you navigate to the next report.

The Device Resource Delays report (DEVR) provides USG and DLY values for jobs that are using devices or are waiting for them. This data is gathered in a multistate fashion, this means that there may be several wait records for the same job for the same device. The reporter changes to ″pseudo multistate″, this can result in one USG counter and one DLY counter in parallel within a cycle, but does not take multiple wait records into account.

Data gathering for the Data Set Delays reports (DSND, DSNJ, and DSNV) is different. Here, several wait records referring to the same device are not treated as being the same and counted only once because they may refer to different data set names, and have to be counted individually.

As a result, the sum of the USG and DLY percentage values in these reports can be different to the USG and DLY percentage values in the DEVR report. Therefore, the three reports contain the headings DUSG% and DDLY% instead of USG% and DLY% to indicate a potential difference to the related values in the DEVR report.

It might happen that **-- N/A --** is provided instead of a data set name. This happens if only those I/O instructions have been detected for which no data set information is provided by the SMS subsystem, like
- I/Os to system data sets (like paging or spooling)
- I/Os to any data set which was opened prior to SMS subsystem initialization
- I/Os like SENSE or RELEASE
- System I/Os not done by an access method

DSNJ Report

```
                    RMF V1R5  Data Set Delays - Job            Line 1 of 2
Command ===> _                                                Scroll ===> HALF

Samples: 76      System: D3    Date: 05/03/04  Time: 12.39.00  Range: 120   Sec

Jobname: U015074             EXCP Rate      0.1      Connect:   0%

ASID  -------------- Data Set Name ---------------  Volume  Num   DUSG% DDLY%
0110  U015074.CALENDAR.DATA                        SGT103  9182      0     1
      SYS1.VTOCIX.VSGT103                          SGT103  9182      1     0
```

Figure 75. Monitor III - Data Set Delays by Job

This report shows all active data sets for a specific job. If you now are interested to know whether some other job is working with this data set, again, you can navigate to the corresponding report.

DSND Report

```
                    RMF V1R5  Data Set Delays                  Line 1 of 1
Command ===> _                                               Scroll ===> HALF

Samples: 76      System: D3    Date: 05/03/04  Time: 12.39.00  Range: 120   Sec

Input Data Set Name: SYS1.VTOCIX.VSGT103

-------------- Data Set Name --------------- Volume  Jobname   ASID DUSG% DDLY%
SYS1.VTOCIX.VSGT103                          SGT103  U015074   0110     1     0
```

Figure 76. Monitor III - Data Set Delays by Data Set

In this sample, no other job than U015074 is using the data set.

Improving Your DASD Performance

This section will suggest actions you may want to consider, to improve the performance of your DASD.

- Obviously, the best response time is achieved if the DASD I/O never happens. Eliminate I/O using buffering or data-in-memory techniques wherever possible (see Appendix C, “Data-In-Memory”).
- Next best is resolving the I/O in cache.
- Last is an I/O to DASD itself.

Also remember ways to increase the ″parallelism″ of your I/O processing.

- This could include the Data Striping capability of DFSMS, where sequential data is distributed across multiple I/O devices to allow concurrent I/O.
- As a last resort, parallelism could also mean manually splitting data across volumes, or duplicating data on multiple volumes.

In addition, here are actions to consider to address the following symptoms:

High DISC time

- Cache the volume, if it is a suitable cache candidate.
- Un-cache the volume, if it is a poor cache candidate.
- Tune cache.
  - Best to use DCME to dynamically manage cache for you.
  - If necessary, manually tune cache by turning off poor cache users.
- Review block size. Increase where applicable.
- Add paths (if not already 4-path).
- Use faster device.
- Move or copy data sets to other volumes to reduce contention.
- Reduce arm movement via data set placement (seek analysis required).

High IOSQ time

- Cache the volume, if it is a suitable cache candidate.
- Tune cache.
  - Best to use DCME to dynamically manage cache for you.
  - If necessary, manually tune cache by turning off poor cache users.
- Decrease DISC, CONN, PEND times. Decreasing the other response time components will also decrease IOSQ time proportionally.
- Use I/O priority queuing (IOQ=PRTY in IEAIPSxx).
- Move or copy data sets to other volumes to reduce contention.
- Run requests sequentially to reduce contention.

High CONN time

- Cache the volume, if it is a suitable cache candidate.
- Use indexed VTOCs.
- In a cached environment, use faster channels.
- Use PDSEs if you suspect directory search time is a problem.

High PEND time

- Address the biggest subcomponent of PEND time (see page 112).
  - Use faster channels, or add more channels (if not already 4-path).
  - Move data sets to avoid contention from sharing systems.

Tape Indicators and Guidelines

This section discusses guidelines for analyzing and resolving tape performance issues for the 3480 and 3490 families of tape devices.

From a performance perspective, you can think of tape as being like DASD with long connect times. All of the things we used to do for single path DASD are true for tape: short strings, many control units and channels. If the subsystem is configured with one channel per control unit function, the only reason to add channels or cross communication of control units is availability, not performance.

Identifying Tape-bound Jobs

An indication of a tape-bound job is when a large proportion of the total time for the job is spent in tape activity and a small amount in CPU (or other) activity. The job termination messages, `NUMBER OF MOUNTS` and `AVG MOUNT TIME` from the Magnetic Tape Device Activity report can help determine if a job is tape-bound.

Look at the job termination messages for the ratio of CPU time to the total elapsed time as indication for tape-bound jobs. You can also use the Monitor III Delay report, to verify that no other (non-I/O) significant delays exist for the job in question and that tape delay is dominant.

Once a job is identified as tape-bound then further evaluation and measurements may be justified.

RMF Measurements

Options for Tape Monitoring

The options required to measure and monitor the tape subsystems may not be the default options used in most user environments. The reason is that the jobs and applications using tape are normally batch tasks, more concerned with throughput than response time.

Some RMF options that must be enabled to have meaningful RMF data for tape subsystem measurements are discussed. These options may be modified in the ERBRMF00 parmlib member or may be modified online for current sessions:

Monitor I Session Options

INTERVAL	This value needs to be small (preferably no more than 15 minutes) if the details of the tape subsystem performance are to be evaluated.
DEVICE	`TAPE` must be one of the device types specified.
IOQ	To request I/O queuing activity for tape LCUs, `IOQ(TAPE)` must be specified.
CHAN	This option specifies that the channel activity is to be measured. The channel utilization may be of interest in evaluating multiple control units on the same channel, or for evaluating the requirement for additional channel paths.

Magnetic Tape Device Activity Report:

```
                              M A G N E T I C   T A P E   D E V I C E   A C T I V I T Y
                                                                                                  PAGE    1
          z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004         INTERVAL 14.59.996
                                 RPT VERSION V1R5 RMF      TIME 09.30.00            CYCLE 1.000 SECONDS

TOTAL SAMPLES =    900   IODF = A3   CR-DATE: 03/01/04   CR-TIME: 07.42.20     ACT: POR

                                  DEVICE   AVG  AVG   AVG  AVG  AVG   AVG  AVG  AVG    %      %     %     NUMBER   AVG      TIME
      DEV  DEVICE   VOLUME  LCU  ACTIVITY RESP IOSQ  DPB  CUB  DB    PEND DISC CONN   DEV    DEV   DEV   OF       MOUNT    DEVICE
      NUM  TYPE     SERIAL       RATE     TIME TIME  DLY  DLY  DLY   TIME TIME TIME   CONN   UTIL  RESV  MOUNTS   TIME     ALLOC

      05B0 3490              0046   0.027  3155  0.0   0.0  0.2  0.0  1411 1736  8.0   0.02   4.65  0.0       1        20     1:47
      05B1 3490              0046   0.000   0.0  0.0   0.0  0.0  0.0   0.0  0.0  0.0   0.00   0.00  0.0       0         0        0
      05B2 3490              0046   0.023   860  0.0   0.0  0.1  0.0   362  494  4.0   0.01   1.16  0.0       1        30       55
      05B3 3490              0046   0.023  2158 95.0   0.0  0.2  0.0   891 1169  2.9   0.01   2.74  0.0       1        29     1:31
      05B4 3490              0046   0.032  2241  0.0   0.0  0.1  0.0  1365  874  2.9   0.01   2.82  0.0       1        32     1:41
      05B5 3490     H3158    0046   0.000   0.0  0.0   0.0  0.0  0.0   0.0  0.0  0.0   0.00   0.00  0.0       0         0        0
      05B6 3490              0046   0.000   0.0  0.0   0.0  0.0  0.0   0.0  0.0  0.0   0.00   0.00  0.0       0         0        0
      05B7 3490              0046   0.975  81.1  0.0   0.0  0.0  0.0  53.9 20.8  6.4   0.62   2.66  0.0       1        59     3:21
      05B8 3490              0046   0.109   388  0.0   0.0  0.0  0.0   230  154  4.8   0.05   1.72  0.0       1        50     2:00
      05B9 3490              0046   1.993  23.8  0.0   0.0  0.0  0.0   9.0  9.3  5.5   1.09   2.94  0.0       1        33     2:05
      05BA 3490              0046   0.023  1461  0.0   0.0  0.3  0.0   621  833  7.3   0.02   1.96  0.0       1        46     1:36
      05BB 3490              0046   0.023  3990  0.0   0.0  0.2  0.0  1453 2529  8.1   0.02   5.92  0.0       1      1:09     2:39
      05BC 3490              0046 OFFLINE
      05BD 3490              0046 OFFLINE
      05BE 3490              0046 OFFLINE
      05BF 3490              0046 OFFLINE

                      LCU    0046   3.229   162  1.0   0.0  0.0  0.0  78.9 76.6  5.7   0.15   2.21  0.0       9        41     1:28
```

Figure 77. Monitor I Magnetic Tape Device Activity Report

Response Time Components

Response time is a key measure of how well your I/O is performing, and a key indicator of what can be done to speed up the I/O. It is important to understand the components of this response and what they mean. Tape response time is the sum of queue, pend, connect and disconnect time (see page 110). The following lists the response time components and what causes them:

Queue Time
AVG IOSQ TIME is a measure of an I/O request waiting in software for the device to become available to service the request. For tape this usually means a REWIND/UNLOAD is in progress.

Pend Time
AVG PEND TIME reflects the time an I/O spends waiting to get a path to the control unit. This is the time spent waiting for an available channel path.

Disconnect Time
AVG DISC TIME is one of the more significant measures of a heavily loaded tape subsystem. Disconnect time is that time spent waiting for the buffer in the control unit to become available to transfer data; it includes the tape rewind time. One example of a disconnected I/O operation is a buffer-full condition.

Connect Time
AVG CONN TIME is predominantly time spent transferring data between the control unit and the processor. Connect time is directly related to data block size, and to the amount of data transferred for each I/O.

General Tape Indicators

Indicator

Field: AVG DISC TIME (LCU), AVG PEND TIME (LCU), AVG CONN TIME (LCU)

Description: Ratio of DISC + PEND time to CONN time. Lower ratios indicate less contention for channels and control units.

Related Fields: DEVICE ACTIVITY RATE. If the I/O rate to the controller is negligible, don't bother with any further analysis.

Guideline: Percent channel busy is not a good indicator of tape performance. Instead look at the ratio of LCU DISC + PEND time to CONN time. If the ratio exceeds 2 to 1 for extended periods of time (with an I/O rate high enough to be of interest), this is an indication of channel and/or control unit contention.

Problem Area: This is an indicator that response times are increasing (increased DISC and PEND) due to an overloaded control unit or channels.

Potential Solution: If you have significant DISC time and little PEND time, add more or faster control units. If you have significant DISC **and** PEND times, add more/faster control units and more/faster channels. As a guideline, you probably want to configure no more than four active drives per control unit. Beyond four **concurrently active** drives behind a single control unit function, data throughput for the additional drives drops off dramatically.

Indicator

Field: AVG CONN TIME (LCU)

Description: Predominantly, data transfer time between the control unit and the CPU.

Related Fields: AVG RESP TIME, DEVICE ACTIVITY RATE

Guideline: CONN times in the 20-40 ms range are typical. If your CONN time is significantly lower (for example, 4-5 ms), you may have smaller than optimal block sizes, resulting in more I/Os required to transfer the data.

Problem Area: This could mean delay to the application, if a high number of I/Os are being generated to transfer the data. Decreasing the I/O rate by increasing the amount of data transferred per I/O will speed up the application run-time.

Potential Solution: Increase block size and/or the number of host buffers where applicable.

Improving Your Tape Performance

This section will suggest actions you may want to consider to improve tape performance.

The key to good performance is to maximize data transfer, minimize control unit communication and command processing, and minimize PEND and DISC times. This can be accomplished by using large block sizes, allocating no more than four concurrently active drives per control unit and not cross-coupling control units.

As a checklist, review the following:

- Review block size. Increase where applicable.
- Allocate data sets to DASD. TMM (tape mount management) is a method to accomplish this and significantly reduce tape mount requirements.
- Increase host buffers for priority jobs. Increase buffers from default of 5, to as high as 20 (but no more than 20).
- Use VIO for temporary data sets.
- Use faster channels.
- Add channels (if not already one channel per control unit function).
- Reduce mount delay.
  - use ACLs/ICLs (cartridge loaders) for nonspecific mounts
  - use an automated tape library for specific mounts
  - review physical logistics for manual mount efficiencies (e.g. are tapes stored near the operators, and near the drives?)
- Add control unit functions (probably want no more than 4 concurrently active drives per control unit).
- Use enhanced-capacity cartridges and tape drives.
- Restructure for more parallelism (for example, to allow application to use multiple drives).
- Use the *Improved Data Recording Capability (IDRC)* to reduce the amount of data transfer required.

Summary

Typically, I/O processing is the largest single component of transaction response time. So, while I/O performance management can be complicated and time consuming, the return in terms of response time and throughput can be significant.

Look for ways to avoid I/O altogether: Through optimized use of buffers, newer data-in-memory techniques, and eliminating unnecessary I/O.

Cache the I/O that will benefit from cache, and keep cache "tuned" to the extent that poor cache users are not allowed to get in the way of better cache users.

As with all performance management you need to:

- Understand the overall system picture (CPU, processor storage) and the overall I/O subsystem picture before tuning.
- Measure the delay to important workloads.
- Review "related fields" before tuning. For example, if response time for a device seems high, also check the I/O rate, type of workload using the device, and % DLY to users, before deciding what (if any) action to take.
- Determine where your leverage for improvement exists. Start by seeing which component of response time (DISC, CONN, etc.) is dominant, and take further action accordingly.

Chapter 5. Analyzing Processor Storage Activity

Let's Analyze Your Storage

This chapter discusses how to analyze a processor storage problem. Topics include:

- What are the indicators in RMF to review, to verify there is a problem?
- Some examples of common problems that RMF can help you resolve.
- Potential tuning actions.

Simply put, resolving processor storage problems generally means prioritizing which workloads get access to processor storage, finding ways to move fewer pages, or speeding up the movement of pages.

Do You Have a Processor Storage Problem?

Here are the primary RMF indicators you may wish to use in assessing your processor storage configuration. After listing these indicators, we will then discuss each one in more detail, with guidelines for each value that may indicate contention and suggestions for further action.

We will refer to the components of storage as *central storage (CS)* and *auxiliary storage (AUX)*.

Never start any tuning effort simply because one indicator seems to be a problem. Always check other related indicators to build a good understanding of your overall processor storage situation (and overall system situation, for that matter) and only then begin to plan your tuning efforts. For example, if your UIC is low, also look at demand paging from auxiliary storage and storage delays in Monitor III.

Table 3 shows for each monitor, which fields in which report can most often help detect and solve a processor storage problem.

Table 3. Processor Storage Indicators in Monitor III and Monitor I

MONITOR	REPORT	FIELDS
Monitor III	SYSINFO	`Average Number Delayed for STOR`
	DELAY	`%Delayed for STR`
	STOR	`DLY%`

Monitor I	Paging Activity	`PAGE`
		`MOVEMENT WITHIN CENTRAL STORAGE`
		`PAGE-IN EVENTS`
		`AVG HIGH UIC`
		`TOTAL FRAMES AVAILABLE`

	Page/Swap Activity	`SLOTS ALLOC`
		`SLOTS USED`
		`% IN USE`

	DASD Activity	`AVG RESP TIME`

Monitor III Indicators

This section reviews the most frequently used processor storage indicators in the Monitor III reports. We will start by looking at SYSINFO and DELAY reports, which can lead you to the STOR report.

SYSINFO Report

```
                        RMF V1R5  System Information              Line 1 of 23
Command ===>                                                   Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

   Partition:    MVS1    2064 Model 110          Appl%:      63  Policy: STANDARD
   CPs Online:      4    Avg CPU Util%:  73      EAppl%:     65  Date:   05/18/04
   IFAs Online:     0    Avg MVS Util%:  84      Appl% IFA:   0  Time:   14.05.07

Group     T WFL --Users--  RESP TRANS -AVG USG-  -Average Number Delayed For -
             %   TOT  ACT  Time  /SEC PROC  DEV   PROC  DEV STOR SUBS OPER  ENQ

*SYSTEM      34  664   26       13.95  5.1  5.0    1.9  4.1  7.0  2.6  2.0  2.0
*TSO         50  534    8       13.95  2.6  2.1    0.4  1.5  2.0  0.8  0.0  0.0
*BATCH       26   11   10        0.00  1.5  1.4    1.4  1.7  0.5  1.8  1.0  2.0
*STC         27  115    8        0.00  1.1  1.5    0.1  1.0  4.5  0.1  1.0  0.0
*ASCH             3    0        0.00  0.0  0.0    0.0  0.0  0.0  0.0  0.0  0.0
*OMVS             2    0        0.00  0.0  0.0    0.0  0.0  0.0  0.0  0.0  0.0
*ENCLAVE      5   4  N/A         N/A  0.2  N/A    3.7  N/A  0.0  N/A  N/A  N/A
PRIMEAPP  W       3    0  .000  0.00  0.0  0.0    0.0  0.0  0.0  0.0  0.0  0.0
APPPRIME  S       3    0  .000  0.00  0.0  0.0    0.0  0.0  0.0  0.0  0.0  0.0
          1       2    0  .000  0.00  0.0  0.0    0.0  0.0  0.0  0.0  0.0  0.0
          2       1    0  .000  0.00  0.0  0.0    0.0  0.0  0.0  0.0  0.0  0.0
PRIMEBAT  W  26   11   10  46.0  0.06  1.5  1.4    1.4  1.7  0.5  1.8  1.0  2.0
NRPRIME   S  26   11   10  46.0  0.06  1.5  1.4    1.4  1.7  0.5  1.8  1.0  2.0
          1  23    9    9  27.9  0.06  0.9  1.4    0.8  1.6  0.5  1.8  1.0  2.0
          2  29    0    0  54.2  0.02  0.1  0.0    0.2  0.0  0.0  0.0  0.0  0.0
          3  59    1    1  .000  0.00  0.6  0.0    0.4  0.0  0.0  0.0  0.0  0.0
```

Figure 78. Monitor III System Information Report

Indicator

Field: Average Number Delayed For STOR

Description: This shows you the average number of AS delayed for reasons related to processor storage.

Guideline: Look at the value for the user community you are trying to help (for example, a specific service class, all TSO, or overall SYSTEM). If the STOR delay is larger than the other delays listed, you probably have some leverage here. If the STOR delay is zero or near zero, then your leverage is elsewhere (for example, I/O or CPU).

Problem Area: This points out delays for paging, swapping, VIO, and out-ready.

Potential Solution: See the Storage Delays report to find out which AS are delayed, and the type of storage delay they are having. You can get there by positioning the cursor on the storage value you are interested in, and pressing ENTER.

DELAY Report

```
                          RMF V1R5  Delay Report                  Line 1 of 43
Command ===>                                                   Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

            Service     WFL USG DLY IDL UKN ----- % Delayed for -----  Primary
Name     CX Class    Cr  %   %   %   %   %  PRC  DEV STR SUB OPR  ENQ  Reason

BHOLEQB  B  BATCH         0   0  51   0  49   0    0   0   0   0   51  SYSDSN
JES2     S  SYSSTC        0   0   2   0  98   0    2   0   0   0    0  SYSPAG
SMF      S  SYSSTC        0   0   1   0  99   0    1   0   0   0    0  SYSPAG
BAJU     T  TSO          29   9  22  67   2   1    0  21   0   0    0  LOCL
RMFGAT   S  SYSSTC       50   1   1   0  98   1    0   0   0   0    0  BAJU
BHOL     T  TSO          78  18   5  73   4   1    3   1   0   0    0  RMFUSR
BHOLPRO2 B  BATCH        93  93   7   0   0   7    0   0   0   0    0  BHOLPRO1
BHOLPRO1 B  BATCH        95  95   5   0   0   5    0   0   0   0    0  BHOLPRO2
*MASTER* S  SYSTEM       96  24   1   0  76   0    1   0   0   0    0  SYSPAG
CATALOG  S  SYSSTC      100   2   0   0  98   0    0   0   0   0    0
BHOLSTOA B  BATCH       100   1   0  99   0   0    0   0   0   0    0
PCAUTH   S  SYSSTC            0   0   0 100   0    0   0   0   0    0
```

Figure 79. Monitor III Delay Report

Indicator

Field: % Delayed for STR

Description: This will give you the percentage of time the job was delayed for processor storage related reasons.

Guideline: See if the STR delay is larger than the others, and is significant (over 10%). If so, go to the Storage Delays panel (enter STOR at the command line) and see the guidelines described there.

Problem Area: Delays due to paging, swapping, VIO, and out-ready.

Potential Solution: See the "STOR Report" on page 138 section of this chapter.

STOR Report

```
                          RMF V1R5  Storage Delays                 Line 1 of 44
Command ===>                                                    Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

              Service  DLY    ------- % Delayed for ------   -- Working Set --
Jobname   C   Class     %     COMM  LOCL   VIO  SWAP  OUTR   Central  Expanded
BAJU      T   TSO      21        0    18     0     3     0       273         5
BHOL      T   TSO       1        1     0     0     0     0      1111         4
*MASTER*  S   SYSTEM    0        0     0     0     0     0        73         7
PCAUTH    S   SYSSTC    0        0     0     0     0     0        30         3
RASP      S   SYSSTC    0        0     0     0     0     0        60         3
TRACE     S   SYSSTC    0        0     0     0     0     0        35         2
XCFAS     S   SYSSTC    0        0     0     0     0     0       173        71
GRS       S   SYSSTC    0        0     0     0     0     0        68        66
SMXC      S   SYSSTC    0        0     0     0     0     0        20         2
SYSBMAS   S   SYSSTC    0        0     0     0     0     0        21         1
DUMPSRV   S   SYSSTC    0        0     0     0     0     0        33         6
CONSOLE   S   SYSSTC    0        0     0     0     0     0        74         6
ALLOCAS   S   SYSSTC    0        0     0     0     0     0        26         3
SMF       S   SYSSTC    0        0     0     0     0     0        47         4
LLA       S   SYSSTC    0        0     0     0     0     0       213         4
BHOLSTO9  B   BATCH     0        0     0     0     0     0         0         0
```

Figure 80. Monitor III Storage Delays Report

Indicator

Field: Check DLY % first, then find biggest component (e.g. COMM, LOCL...).

Description: This report shows AS in descending order of delay.

Guideline: See if the DLY % field is significant (over 10%) before proceeding further. The DLY % fields in Monitor III give you a good estimate of how much you can reduce response time by eliminating delay. 10% delay implies that if you are able to eliminate that delay entirely, you could improve response time by about 10%.

Problem Area: Paging, swapping, VIO, etc.

Potential Solution:

- COMM
  - See "Prioritize Access to Processor Storage" on page 145 - consider storage isolation using CVSS and CPGRT parameters from IEAIPSxx
  - See "Decrease Storage Demand" on page 145
  - See "Auxiliary Storage Tuning" on page 145
- LOCL
  - See "Prioritize Access to Processor Storage" on page 145 - storage isolation and criteria age
  - See "Decrease Storage Demand" on page 145
  - See "Auxiliary Storage Tuning" on page 145
  - If swappable, consider swap set size tuning
- VIO
  - See "Prioritize Access to Processor Storage" on page 145 - storage isolation and criteria age
  - See "Decrease Storage Demand" on page 145
  - See "Auxiliary Storage Tuning" on page 145
- SWAP
  - See "Prioritize Access to Processor Storage" on page 145 - storage isolation and criteria age
  - See "Auxiliary Storage Tuning" on page 145
  - Swap set size tuning
- OUTR
  - Check RTO setting
  - See "Decrease Storage Demand" on page 145

Monitor I Indicators

This section will review the most frequently used processor storage indicators in the Monitor I reports.

Paging Activity Report - Page 1

```
                                                   P A G I N G   A C T I V I T Y
                                                                                                    PAGE 1

            z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004          INTERVAL 14.59.996
                                   RPT VERSION V1R5 RMF      TIME 09.29.00            CYCLE 1.000 SECONDS

OPT = IEAOPT02  MODE=ESAME                   CENTRAL STORAGE PAGING RATES - IN PAGES PER SECOND
----------------------------------------------------------------------------------------------------------------------------------

                            PAGE IN                                    PAGE OUT
                 ------------------------------------        ----------------------------
                                NON          %                                         %
                         NON    SWAP,        OF                                        OF
                         SWAP,  NON    TOTAL TOTL                    NON     TOTAL     TOTL
CATEGORY         SWAP    BLOCK  BLOCK  RATE  SUM             SWAP    SWAP    RATE      SUM

PAGEABLE SYSTEM
AREAS (NON VIO)
     LPA                  0.00   0.01    0.01    0

     CSA                  0.00   0.04    0.04    0                   0.00    0.00     0
                         ------ ------  ------ ----                 ------  ------   ----
     SUM                  0.00   0.05    0.05    0                   0.00    0.00     0

ADDRESS SPACES
     HIPERSPACE           0.06           0.06    1                   0.00    0.00     0

     VIO                  0.00           0.00    0                   0.00    0.00     0

     NON VIO      0.28    7.01   2.01    9.30   99          0.00     0.01    0.01   100
                 ------  ------ ------  ------ ----       ------   ------  ------   ----
     SUM          0.28    7.07   2.01    9.37  100          0.00     0.01    0.01   100

TOTAL SYSTEM
     HIPERSPACE           0.06           0.06    1                   0.00    0.00     0

     VIO                  0.00           0.00    0                   0.00    0.00     0

     NON VIO      0.28    7.01   2.06    9.35   99          0.00     0.01    0.01   100
                 ------  ------ ------  ------ ----       ------   ------  ------   ----
     SUM          0.28    7.07   2.06    9.41  100          0.00     0.01    0.01   100

     SHARED                      0.00    0.00                        0.00    0.00

 PAGE MOVEMENT WITHIN CENTRAL STORAGE      20.84            PAGE MOVEMENT TIME %       0.0
 AVERAGE NUMBER OF PAGES PER BLOCK          5.6
 BLOCKS PER SECOND                          1.26
 PAGE-IN EVENTS (PAGE FAULT RATE)           3.31
```

Figure 81. Monitor I Paging Activity Report - Page 1

Indicator

Field: PAGE MOVEMENT WITHIN CENTRAL STORAGE

Description: Rate of page movement above and below 16MB in CS. For example, to keep long term fixed pages out of reconfigurable storage.

Guideline: This is rarely a problem. If you see large numbers here (maybe over 100/sec), it may be worth looking at. Or, if PAGE MOVEMENT TIME % reports a high value, this may indicate a problem.

Problem Area: Some CPU time is required to move these pages. Generally not enough to worry about.

Potential Solution: Review the setting of your RSU parameter for reconfigurable storage. Reduce if possible (but not at the expense of your disaster recovery planning).

Indicator

Field: PAGE-IN EVENTS (PAGE FAULT RATE)

Description: This is the rate of demand paging in from auxiliary storage. The AS waits while these pages are obtained.

Guideline: No overall system guideline for this number would make much sense. More important than the single-number here, is the amount of paging delay to important work (see Monitor III Storage Delays report). You may wish to track this one over time, or turn to this for specific MPL tuning.

Problem Area: Problems here would show up as paging delay to important work.

Potential Solution: Paging problems are probably best addressed by analyzing which workloads are paging, then addressing the problem for those workloads specifically (see Monitor III Storage Delays report). Use Monitor III to get to the AS level, or the Workload Activity report to get to the service class. There, the following solutions could apply:

- Isolate storage
- Adjust swap set size
- Adjust the criteria age value
- As a last resort, see "Auxiliary Storage Tuning" on page 145 for ways to speed up the paging.

For overall system paging reduction, most of the items in "General Storage Recommendations" on page 145 could apply.

Paging Activity Report - Page 2

```
                                        P A G I N G   A C T I V I T Y
                                                                                                        PAGE 2

          z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004         INTERVAL 14.59.996
                                 RPT VERSION V1R5 RMF      TIME 09.29.00           CYCLE 1.000 SECONDS

OPT = IEAOPT02   MODE = ESAME           CENTRAL STORAGE MOVEMENT RATES - IN PAGES PER SECOND
-------------------------------------------------------------------------------------------------------------------------------

   HIGH UIC (AVG) = 2498.3   (MAX) = 2540   (MIN) = 2133


                     WRITTEN TO       READ FROM        *--- CENTRAL STORAGE FRAME COUNTS ----*
                   CENTRAL STOR    CENTRAL STOR             MIN           MAX           AVG

   HIPERSPACE   RT          0.02            0.00             67           140           112
   PAGES

   VIO          RT          2.93            1.96             18         1,832           169
   PAGES


                                                  FRAME AND SLOT COUNTS
-------------------------------------------------------------------------------------------------------------------------------

                      CENTRAL STORAGE             EXPANDED STORAGE            LOCAL PAGE DATA SET SLOT COUNTS
                    ---------------------------  ---------------------------  ---------------------------------------------
   (90 SAMPLES)          MIN       MAX       AVG        MIN       MAX       AVG                        MIN         MAX         AVG

   AVAILABLE               1     6,697     1,765                                  AVAILABLE SLOTS  2,485,531   2,501,203   2,494,724

   SQA                16,860    17,348    17,070                                  VIO SLOTS               30          30          30
   LPA                 3,902     4,134     3,961
   CSA                22,983    24,966    24,101                                  NON-VIO SLOTS      332,867     348,539     339,345
   LSQA               32,169    34,632    33,304
   REGIONS+SWA       434,175   440,857   438,221                                  BAD SLOTS                0           0           0

   TOTAL FRAMES      524,288   524,288   524,288                                  TOTAL SLOTS      2,834,100   2,834,100   2,834,100

                         FIXED FRAMES                                                SHARED FRAMES AND SLOTS
                    ---------------------------                                  ---------------------------------------------
   NUCLEUS             5,862     5,862     5,862                                  CENTRAL STORAGE        264         643         424
   SQA                15,632    15,777    15,685
   LPA                    88        89        88                                  FIXED TOTAL              0           0           0
   CSA                11,108    11,203    11,119                                  FIXED BELOW 16 M         0           0           0
   LSQA               21,916    23,626    22,930                                  AUXILIARY SLOTS      1,651       1,701       1,655
   REGIONS+SWA        21,201    22,181    21,629
   BELOW 16 MEG          843       916       880                                  TOTAL                1,915       2,344       2,079
   BETWEEN 16M-2G        N/A       N/A       N/A

   TOTAL FRAMES       76,292    78,574    77,314
```

Figure 82. Monitor I Paging Activity Report - Page 2

Indicator

Field: HIGH UIC (AVG)

Description: Amount of time a page remains in central storage without being referenced. This indicates stress on central storage and ranges from a low of 0 to a high of 2540.

Guideline: Small values might indicate storage constraints, however, some systems have run perfectly well at low UICs.

Problem Area: Problems here would show up as delays caused by demand paging or swapping, or increased CPU overhead due to page movement. Also throughput of the system could drop, due to MPL adjustment.

Potential Solution: See "Decrease Storage Demand" on page 145.

Page/Swap Data Set Activity Report

```
                                P A G E / S W A P   D A T A S E T   A C T I V I T Y
                                                                                PAGE 1

           z/OS V1R6              SYSTEM ID OS04            DATE 06/05/2004         INTERVAL 14.59.996
                                  RPT VERSION V1R5 RMF      TIME 09.29.00           CYCLE 1.000 SECONDS

NUMBER OF SAMPLES =     898                      PAGE DATA SET USAGE
                                                -------------------------

PAGE                                                                          %   PAGE                 V
SPACE   VOLUME  DEV  DEVICE   SLOTS  ---- SLOTS USED ---  BAD       IN  TRANS  NUMBER  PAGES  I
TYPE    SERIAL  NUM  TYPE     ALLOC    MIN    MAX   AVG  SLOTS     USE  TIME   IO REQ  XFER'D O  DATA SET NAME

PLPA    PAGP11  0378 33903     45000  16466  16466  16466      0    0.11  0.200       6       5    PAGE.OS04.APLPA

COMMON  PAGP10  0377 33903     21240   5345   5378   5355      0    0.11  0.027      38      37    PAGE.OS04.ACOMMON

LOCAL   PAGP12  0379 33903    566820  67670  70965  69140      0    2.67  0.004     711   5,697 N  PAGE.OS04.LOCAL01

LOCAL   PAGP13  037A 33903    566820  65887  69007  67190      0    3.01  0.005     771   5,711 N  PAGE.OS04.LOCAL02

LOCAL   PAGP14  037B 33903    566820  66611  69723  67928      0    3.01  0.005     821   5,625 Y  PAGE.OS04.LOCAL03

LOCAL   PAGP15  038D 33903    566820  66296  69563  67618      0    3.23  0.005     726   5,523 Y  PAGE.OS04.LOCAL04

LOCAL   PAGP16  039F 33903    566820  66390  69311  67560      0    3.12  0.005     766   5,548 Y  PAGE.OS04.LOCAL05
```

Figure 83. Monitor I Page/Swap Data Set Activity Report

Note: Starting with OS/390 2.10, swap data sets are no longer supported. Therefore, the report contains information only about page data sets.

Indicator

Field: SLOTS ALLOC and AVG SLOTS USED

Description: This tells you how much of your auxiliary storage space is used.

Guideline: Keep the percent of space used to around 30% (always below 50%) of the space allocated.

Indicator

Field: `% IN USE`

Description: This is the busy percentage for the data set.

Guideline: For page data sets, you may see response time increase if this number rises above 30%. If you have swap data sets defined, you may see response time increases if they are above 15%.

Problem Area: Increased time to satisfy a page request from DASD.

Potential Solution: See "Auxiliary Storage Tuning" on page 145

DASD Activity Report

```
                                     D I R E C T   A C C E S S   D E V I C E   A C T I V I T Y
                                                                                                                    PAGE 1

             z/OS V1R6               SYSTEM ID OS04             DATE 06/05/2004          INTERVAL 14.59.996
                                     RPT VERSION V1R5 RMF       TIME 09.29.00            CYCLE 1.000 SECONDS

TOTAL SAMPLES =  1,807   IODF = 12  CR-DATE: 03/01/04   CR-TIME: 10.23.17

                                         DEVICE   AVG   AVG   AVG  AVG  AVG    AVG   AVG  AVG     %      %     %    AVG    %      %
STORAGE  DEV  DEVICE    VOLUME PAV  LCU  ACTIVITY RESP  IOSQ  DPB  CUB  DB     PEND  DISC CONN    DEV    DEV   DEV  NUMBER ANY    MT
 GROUP   NUM  TYPE      SERIAL           RATE     TIME  TIME  DLY  DLY  DLY    TIME  TIME TIME    CONN   UTIL  RESV ALLOC  ALLOC  PEND

         0377 33903     PAGP10      005F   0.044  20.2   0.0  0.0  0.0  0.0    1.9 16.5  1.8   0.01   0.08   0.0  1.0   100.0   0.0
         0378 33903     PAGP11      005F   0.011  24.9   0.0  0.0  0.0  0.0    0.2 22.7  2.0   0.00   0.03   0.0  1.0   100.0   0.0
         0379 33903     PAGP12      005F   0.805  32.1   0.0  0.0  0.0  0.0    1.7 17.9 12.5   1.01   2.45   0.0  1.0   100.0   0.0
         037A 33903     PAGP13      005F   0.863  31.8   0.0  0.0  0.0  0.0    1.4 18.2 12.2   1.05   2.62   0.0  1.0   100.0   0.0
         037B 33903     PAGP14      005F   0.915  30.6   0.0  0.0  0.0  0.0    1.4 17.9 11.3   1.04   2.68   0.0  1.0   100.0   0.0
```

Figure 84. Monitor I DASD Activity Report

Indicator

Only DASD volumes used for paging and swapping are of interest in this section. See Chapter 4, "Analyzing I/O Activity" for analysis of other I/O.

Field: `AVG RESP TIME`

Description: This will show overall response time (ms) for these volumes.

Related Field: `DEVICE ACTIVITY RATE`, Monitor III page/swap delay % values.

Guideline: Typical response times for page/swap volumes are in the 30-40 ms range. If your response times are significantly higher you may want to investigate.

Problem Area: Problems here would add page or swap delay to your important workloads.

Potential Solution: See "Auxiliary Storage Tuning" on page 145

General Storage Recommendations

This section gives actions you may want to consider to improve your processor storage performance. Before you start, bear in mind the following;

- The first priority of tuning is to make the best use of your resources.
- After this, prioritizing workload is even more critical as you will be stealing from one workload to give to another.

Decrease Storage Demand

- Reduce AS buffer sizes and use hiperspace buffering. Consider:
  - VSAM
  - DB2 hiperpools
  - PDSE with low MSR values specified
- Reduce MPL
  - Activate RCCPTRT values in IEAOPTxx
  - Decrease number of initiators
- Decrease the number of active AS by trimming CICS AS, IMS MPRs. Maybe reducing MAXUSERS keyword at IPL.
- Use IEFUSI to limit the use of AS private area (region).
- Decrease other work (for example, monitors).

Increase Storage

- Take central storage from another partition, or install more.

Auxiliary Storage Tuning

- Reduce paging (see above).
- Dedicate volumes to page data sets.
- Distribute page data sets among channel paths and control units.
- Limit use of VIO=YES.
- Add more LOCALs. Make the sum of all the page space 2 to 4 times the number of slots used.
- Use faster devices.

Prioritize Access to Processor Storage

- Make better use of the PLPA by removing unused modules and verify that concatenated libraries are needed (LPALSTxx).
- Decrease the page fixing level. RMF tells you the cause of the problem at PRIV-FF in the ARD Monitor II report.
- Adjust criteria age values via IEAOPTxx, this can be used to direct pages from CS to AUX.
- Use the storage protection specification in the service policy.

Tune DASD

- Tuning 'normal' DASD I/O reduces storage requirements. Transactions will complete faster, needing less virtual and processor storage.

Summary

To analyze storage problems you should:

- View the system as a whole and not rely on individual guidelines.

As we have said before the main thing is: Are your SLA targets being met? If not, and storage is the problem then:

- Ensure the storage is being used fully and start looking at prioritizing access.
- Use the recommendation checklist with the *z/OS MVS Initialization and Tuning Reference*.
- Only change one value at a time and monitor its effect before making any other change.

Chapter 6. Analyzing Sysplex Activity

Let's Analyze Your Sysplex

This chapter discusses some special performance aspects in a sysplex:

- How to understand CICS and IMS workload reports
- Analyzing coupling facility activities

Understanding Workload Activity Data for IMS or CICS

This chapter provides an explanation of some Workload Activity report data for work managers. RMF provides data for the subsystem work managers that support workload management. In z/OS, these are IMS and CICS. You find these values either in the Monitor III Work Manager Delays report or in the Postprocessor Workload Activity report. The following examples will describe the Workload Activity report.

This section provides some sample reports for CICS and IMS, followed by some potential problem data and explanations. Based on the explanations, you may decide to alter your service class definitions. In some cases, there may be some actions that you can take. In those cases, you can follow the suggestion. In other cases, the explanations are provided only to help you better understand the data.

Interpreting the Workload Activity Report

CICS Service Class Accessing DBCTL

Figure 85 shows an example of the work manager state section for a service class representing CICS transactions accessing Database Control (DBCTL).

```
REPORT BY: POLICY=HPTSPOL1   WORKLOAD=PRODWKLD   SERVICE CLASS=CICSHR   RESOURCE GROUP=*NONE      PERIOD=1 IMPORTANCE=1

TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT
AVG      0.00   ACTUAL                114
MPL      0.00   EXECUTION              78
ENDED     216   QUEUED                 36
END/S    0.24   R/S AFFINITY            0
#SWAPS      0   INELIGIBLE              0
EXCTD     216   CONVERSION              0
AVG ENC  0.00   STD DEV               270
REM ENC  0.00
MS ENC   0.00

           RESP  ------------------------------- STATE SAMPLES BREAKDOWN (%) -------------------------------  ------STATE------
SUB    P   TIME  --ACTIVE-- READY IDLE  ----------------------------WAITING FOR----------------------------  SWITCHED SAMPL(%)
TYPE        (%)   SUB  APPL             CONV PROD                                                           LOCAL SYSPL REMOT
CICS  BTE  93.4  10.9   0.0   0.0  0.0  89.2  0.0                                                            89.2   0.0   0.0
CICS  EXE  67.0  19.7   0.0  10.6  0.0   0.0 69.7                                                             0.0   0.0   0.0
```

Figure 85. Hotel Reservations Service Class

The fields in this report describe the CICSHR service class. CICS transactions have two phases:

- The *begin-to-end phase* (`CICS BTE`) takes place in the first CICS region to begin processing a transaction. Usually this is the terminal owning region (TOR). The TOR is responsible for starting and ending the transaction.
  - The **ENDED** field shows that 216 hotel reservation transactions completed.
  - The **ACTUAL** time shows that the 216 transactions completed in an average transaction time of 0.114 seconds.
- The *execution phase* (`CICS EXE`) can take place in an application owning region (AOR) and a file owning region (FOR). In this example, the 216 transactions were routed by a TOR to an AOR.
  - The **EXCTD** field shows that the AORs completed 216 transactions in the interval.
  - The **EXECUTION** time shows that on average it took 0.078 seconds for the AORs to execute the 216 transactions. The **EXECUTION** time applies only to the **EXCTD** transactions.

While executing these transactions, the CICS subsystem records the states the transactions are experiencing. RMF reports the states in the STATE SAMPLES BREAKDOWN (%) section of the report. Because there is a CICS BTE and a CICS EXE field, you can assume that the time spent in the TOR represents the BTE phase, and the time spent in the AOR represents the EXE phase. There is one EXE phase summarizing all the time spent in one or more AORs. Figure 86 shows the actual response time break down of the CICSHR service class.

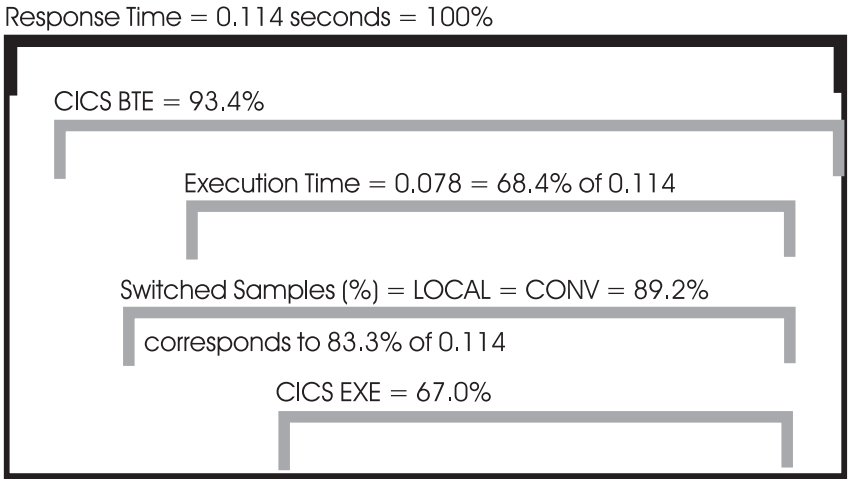


Figure 86. Response Time Breakdown of CICSHR accessing DBCTL

The CICS BTE field shows that the TORs have information covering 93.4% of the response time. RMF does not have information covering 100% of the 0.114 seconds response time, because it takes some time for the system to recognize and assign incoming work to a service class before it can collect information about it.

For most of the 93.4% of the time, the transactions did not run in the TOR, but had been routed locally to an AOR on the same MVS image. You can see this by the SWITCHED SAMPL(%) LOCAL field, which is 89.2% of the total state samples. This value accounts for 83.3% of the response time, because 100% of the total state samples correspond to 93.4% of the response time (89.2 x 93.4 / 100 = 83.3%). This value of 89.2% is close, if not equal, to the WAITING FOR CONV field, which indicates that there is no delay in the TOR once the AOR has returned the transactions.

The total execution time is some percentage of the total response time. It is the EXECUTION transaction time (0.078), divided by ACTUAL transaction time (0.114), which is 68.4%. The CICS execution phase (CICS EXE field) covers 67% of the response time. Some of that time the work is active in the AOR, sometimes it is waiting behind another task in the region, but 69.7% of the total state samples in the PROD field (which corresponds to 69.7 x 67 / 100 = 46.7% of the response time) were found outside of the CICS subsystem, waiting for another product to provide some service to these transactions. Based on the configuration of the system, the transactions are accessing DBCTL.

The LOCL, SYSP and REMT state percentages appear in the WAITING FOR section if greater than zero and show the percentages of the total state samples the service class was delayed in these states when CICS was waiting to establish a session. The STATE SWITCHED SAMPL(%) fields LOCAL, SYSPL, and REMOT show the percentage of the state samples in which transactions were routed via MRO, MRO/XCP, or VTAM connections.

CICS Service Class Accessing DBCTL with IMS V5

Figure 87 shows an example of a Workload Activity report for a CICS service class representing point-of-sale transactions (CICSPS). This example builds on the example in Figure 85 on page 148. It shows how the RMF report changes when IMS is upgraded to Version 5.

```
 REPORT BY: POLICY=WSTPOL02   WORKLOAD=CICS       SERVICE CLASS=CICSPS     RESOURCE GROUP=*NONE      PERIOD=1 IMPORTANCE=1

 TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT
 AVG      0.00   ACTUAL                503
 MPL      0.00   EXECUTION             399
 ENDED     579   QUEUED                104
 END/S    0.95   R/S AFFINITY            0
 #SWAPS      0   INELIGIBLE              0
 EXCTD     559   CONVERSION              0
 AVG ENC  0.00   STD DEV               847
 REM ENC  0.00
 MS ENC   0.00

           RESP  ------------------------------- STATE SAMPLES BREAKDOWN (%) -------------------------------  ------STATE------
SUB    P   TIME  --ACTIVE-- READY IDLE  ----------------------------WAITING FOR----------------------------  SWITCHED SAMPL(%)
TYPE        (%)   SUB  APPL             CONV PROD  I/O LOCK                                                  LOCAL SYSPL REMOT
CICS  BTE  96.6   2.5   0.0   0.1  0.0  97.4  0.0  0.0  0.0                                                   97.4   0.0   0.0
CICS  EXE  29.2  48.6   0.0  14.0  0.0   0.0 23.6 13.7  0.0                                                    0.0   0.0   0.0
IMS   EXE  43.7  92.2   0.0   0.0  0.0   0.0  0.0  0.0  7.8                                                    0.0   0.0   0.0
```

Figure 87. Point-of-Sale Service Class

The ENDED field shows that 579 point-of-sale transactions completed. The ACTUAL time shows that those transactions completed in an average response time of 0.503 seconds.

The EXCTD field shows that the AORs completed 559 transactions in the interval. This is less than the ENDED value because a TOR may not always route all transactions to an AOR. Those non-routed transaction are not counted as EXCTD. The EXECUTION field shows that on average it took the AORs 0.399 seconds to execute the 559 transactions.

While executing these transactions, the CICS subsystem reports the states the transactions are experiencing. Figure 88 shows the response time breakdown of the CICSPS service class.

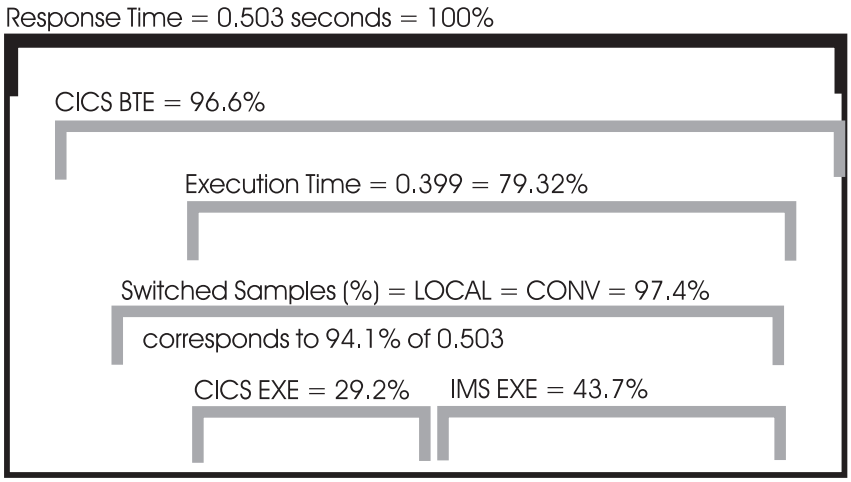


Figure 88. Response Time Breakdown of CICSPS accessing DBCTL with IMS V5

The CICS BTE field shows that the TORs have information covering 96.6% of the response time. Most of that time, the transactions were in fact not being run in the

TOR, but had been routed locally to an AOR on the same MVS image. You can see this by the SWITCHED SAMPL(%) LOCAL field, which is 97.4%. This value accounts for 94.1% of the response time (97.4 x 96.6 / 100 = 94.1%)

The EXECUTION transaction time (0.399 seconds) divided by the ACTUAL transaction time (0.503 seconds) is 79.32%. There are two execution phases shown: a CICS EXE, and an IMS EXE. The CICS execution phase covers only 29.2% of the response time. That is lower than in the previous example, because now IMS is providing information. The IMS EXE row shows the 43.7% of the response time for which it is responsible.

CICS Accessing Data via FORs

Figure 89 shows an example of a Workload Activity report for some CICS user transactions which have data accessed via file owning regions.

```
REPORT BY: POLICY=EQUALIMP   WORKLOAD=CICS       SERVICE CLASS=CICUSRTX     RESOURCE GROUP=*NONE      PERIOD=1 IMPORTANCE=1

 TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT
 AVG      0.00   ACTUAL                 93
 MPL      0.00   EXECUTION              69
 ENDED   14119   QUEUED                 24
 END/S   47.17   R/S AFFINITY            0
 #SWAPS      0   INELIGIBLE              0
 EXCTD   13067   CONVERSION              0
 AVG ENC  0.00   STD DEV                 63
 REM ENC  0.00
 MS ENC   0.00


           RESP  ------------------------------- STATE SAMPLES BREAKDOWN (%) -------------------------------  ------STATE------
SUB    P   TIME  --ACTIVE-- READY IDLE  ------------------------------WAITING FOR------------------------------  SWITCHED SAMPL(%)
TYPE        (%)   SUB  APPL              I/O CONV LOCK                                                          LOCAL SYSPL REMOT
CICS  BTE  90.5   3.0   0.0   3.9  0.2   4.0 88.8  0.0                                                          92.4   0.0   0.0
CICS  EXE  67.2   7.0   0.0   3.9  0.0  89.0  0.0  0.1                                                          60.9   0.0   0.0
```

Figure 89. CICS User Transactions Service Class

In this example, 14,119 transactions completed with a response time of 0.093 seconds. Only 13,067 transactions were executed by AORs in the interval. Others ran completely in the TOR.

The CICS execution phase contains information about the transactions while they were running in either the AOR or FOR. This shows that 89.0% of the state samples for these transactions were waiting for I/O completion. This accounts for 59.8% of the response time (67.2 x 89.0 / 100 = 59.8%). It is not possible to describe whether this time was for I/O initiated by the FOR, or for I/O initiated within the AOR. However, the SWITCHED SAMPL(%) LOCAL value in the CICS EXE line says that for 60.9% of the state samples (which is 40.9% of the response time, that is, 60.9 x 67.2 / 100), one of those regions was waiting for the other region to complete some processing for the transaction before the original region could proceed.

Problem: Very Large Response Time Percentage

Figure 90 on page 152 shows an example of a work manager state section for the CICSPROD service class. In column RESP TIME (%), both the CICS EXE and the CICS BTE rows show inflated percentages: 78.8K and 140.

```
REPORT BY: POLICY=HPTSPOL1   WORKLOAD=PRODWKLD   SERVICE CLASS=CICSPROD     RESOURCE GROUP=*NONE      PERIOD=1 IMPORTANCE=1

 TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT
 AVG      0.00   ACTUAL                111
 MPL      0.00   EXECUTION             123
 ENDED    1648   QUEUED                  0
 END/S    1.83   R/S AFFINITY            0
 #SWAPS      0   INELIGIBLE              0
 EXCTD    1009   CONVERSION              0
 AVG ENC  0.00   STD DEV               351
 REM ENC  0.00
 MS ENC   0.00

           RESP  ------------------------------- STATE SAMPLES BREAKDOWN (%) -------------------------------  ------STATE------
SUB    P   TIME  --ACTIVE-- READY IDLE  ----------------------------WAITING FOR----------------------------  SWITCHED SAMPL(%)
TYPE        (%)   SUB  APPL             MISC PROD CONV  I/O                                                 LOCAL SYSPL REMOT
CICS  BTE  78.8K  0.2   0.0   0.3  2.5  96.7  0.0  0.3  0.0                                                   0.3   0.0   0.0
CICS  EXE   140  65.6   0.0   2.2  0.0   0.0 32.4  0.0  0.1                                                   0.0   0.0   0.0
```

Figure 90. Response Time Percentages Greater than 100

Possible Explanations

Long-running transactions: The report above shows how long-running transactions can inflate the value for RESP TIME (%). While the following example does not explain the exact values in the figure, it explains why this is possible.

Suppose 100 transactions have ended within 1 second, and one transaction has been running for 5 minutes and is still executing when the RMF interval expires. The ACTUAL transaction time shows an average response time of 1 second, and RMF shows the breakdown into the states recorded by CICS or IMS. The subsystem, however, recorded a total of 6 minutes and 40 seconds (5 minutes plus 100 seconds) worth of data. That is an average of 4 seconds worth of data for each completed transaction — which is 4 times the 1 second response time. The state samples breakdown, however, shows information representing 100% of the state samples.

Also, when the one long-running transaction completes, it could easily distort the average response time during that interval. The RMF standard deviation and distribution of response times emphasizes when this occurs.

The long-running transactions could be either routed or non-routed transactions. Routed transactions are transactions that are routed from a TOR to any AOR. Long-running routed transactions could result in many samples of WAITING FOR CONV (waiting for a conversation) in the CICS BTE phase, as well as states recorded from the AOR in the execution phase.

Long-running non-routed transactions execute completely in a TOR, and have no CICS EXE phase data, and could inflate any of the state data for the CICS BTE phase.

Never-ending transactions: Never-ending transactions differ from long-running transactions in that they persist for the life of a region. For CICS, these could include the IBM reserved transactions such as CSNC and CSSY, or customer defined transactions. Never ending transactions are reported similarly to long-running transactions explained in "Long-running transactions." However, for never-ending CICS transactions, RMF might report high percentages in the IDLE, WAITING FOR TIME, or the WAITING FOR MISC (miscellaneous) fields.

Conversational transactions: Conversational transactions are considered long-running transactions. CICS marks the state of a conversational transaction as IDLE when the transaction is waiting for terminal input. Terminal input often includes long end-user response time, so you might see percentages close to 100% in the IDLE state for completed transactions.

Service class includes dissimilar work: A service class that mixes customer and IBM transactions, short and long or never-ending transactions, routed and non-routed transactions, or conversational and non-conversational transactions can expect to have RMF reports showing that the total states sampled account for more than the average response time. This could be true for both IMS and CICS, and can be expected if the service class is the subsystem default service class. The default service class is defined in the classification rules. It is the service class to which all work in a subsystem is assigned that is not assigned to any other service class.

Possible Actions

Group similar work into service classes: Make sure your service classes represent a group of similar work. This could require creating additional service classes. For the sake of simplicity, you can have only a small number of service classes for CICS or IMS work. If there are transactions for which you want the RMF state samples breakdown data, consider including them in their own service class.

Do nothing: For service classes representing dissimilar work such as the subsystem default service class, understand that the response time percentage could include long-running or never-ending transactions. RMF data for such service classes may not make immediate sense.

Problem: Response Time is Zero

Figure 91 shows an example of a work manager state section for the CICSLONG service class. The RESP TIME (%) fields shows a 0.0 value.

```
REPORT BY: POLICY=HPTSPOL1   WORKLOAD=PRODWKLD   SERVICE CLASS=CICSLONG    RESOURCE GROUP=*NONE      PERIOD=1 IMPORTANCE=1

                                                        CICS Long Running Internal Trxs

 TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT.SS.TTT
 AVG      0.00   ACTUAL                        0
 MPL      0.00   EXECUTION                     0
 ENDED       0   QUEUED                        0
 END/S    0.00   R/S AFFINITY                  0
 #SWAPS      0   INELIGIBLE                    0
 EXCTD       0   CONVERSION                    0
 AVG ENC  0.00   STD DEV                       0
 REM ENC  0.00
 MS ENC   0.00

           RESP  ------------------------------- STATE SAMPLES BREAKDOWN (%) ------------------------------  ------STATE------
SUB    P   TIME  --ACTIVE-- READY IDLE  ----------------------------WAITING FOR----------------------------  SWITCHED SAMPL(%)
TYPE        (%)   SUB  APPL             CONV  I/O PROD DIST REMT LOCK                                       LOCAL SYSPL REMOT
CICS  BTE   0.0  70.8   0.0   1.4  0.7  11.2  9.2  0.3  5.3  1.2  0.0                                        11.2   0.0   0.0
CICS  EXE   0.0  43.2   0.0   0.2  0.1  31.8 10.4  8.7  0.0  2.9  2.8                                         0.0   0.0   0.0
```

Figure 91. Response Time Percentages all Zero

Possible Explanations

No transactions completed: While a long-running or never-ending transaction is being processed, RMF stores the service class state samples in SMF 72.3 records. But when no transactions have completed, the average response time is 0. However, the calculations for the state samples breakdown will result in values greater than 0.

RMF did not receive data from all systems in the sysplex: The Postprocessor may have been given SMF records from only a subset of the systems running in the sysplex. The report may represent only a single MVS image. If that MVS image has no TOR, its AORs receive CICS transactions routed from another MVS image or from outside the sysplex. Since the response time for the transactions is reported by the TOR, there is no transaction response time for the work, nor are there any ended transactions, on this MVS image.

Possible Actions

Do nothing: You may have created this service class to prevent the state samples of long-running transactions from distorting data for your production work.

Combine all SMF records for the sysplex: When a single MVS image that does not have TORs is combined with another MVS image that does have TORs and therefore does report response times, the states and response time from the first image will be combined by RMF with the states and response time from the second.

Problem: More Executed Transactions Than Ended Transactions

In an IMS TM environment, the control region counts the number of ended transactions, and provides the actual response time information. The message processing program (MPP) counts the number of executed transactions, and provides the execution time. IMS provides only EXE phase information. There is no BTE phase information for IMS work.

Figure 92 shows an example of a work manager data section for the IMSTRX service class. The EXCTD field shows that 5312 transactions have finished executing while the ENDED field shows that only 5212 have been ended by the CTL region.

```
REPORT BY: POLICY=WSTPOL01   WORKLOAD=IMS       SERVICE CLASS=IMSTRX

 TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT
 AVG      0.00   ACTUAL                133
 MPL      0.00   EXECUTION             107
 ENDED    5212   QUEUED                 26
 END/S    5.79   R/S AFFINITY            0
 #SWAPS      0   INELIGIBLE              0
 EXCTD    5312   CONVERSION              0
 AVG ENC  0.00   STD DEV               512
 REM ENC  0.00
 MS ENC   0.00

           RESP  ------------------------------- STATE SAMPLES BREAKDOWN (%) -------------------------------  ------STATE------
SUB    P   TIME  --ACTIVE-- READY IDLE  ----------------------------WAITING FOR----------------------------  SWITCHED SAMPL(%)
TYPE        (%)   SUB  APPL            LOCK                                                                  LOCAL SYSPL REMOT
IMS   EXE  82.4  98.1   0.0   0.0  0.0   1.9                                                                    0.0   0.0   0.0
```

Figure 92. Executed Transactions greater than Ended Transactions

Possible Explanations

IMS program-to-program switches: When an IMS transaction makes a program-to-program switch, the switched work can be considered either a new transaction, or a continuation of the originating transaction. The system checks the classification rules for the service class to be associated with the switched work.

If the service class is the same as the originating transaction, it is processed in the MPP, and it is counted as an executed transaction (EXCTD). When the response is sent back to the network, RMF shows two executed transactions: the originating

and the switched work, but only a single ended transaction (ENDED). Here, the number of executed transactions exceeds the number of ended transactions.

If the service class is different from the original transaction, then RMF counts it as a new transaction with its own EXCTD and ENDED. Here, the executed transaction value and ended value agree.

RMF processed only part of the entire sysplex data: It is possible to invoke the Postprocessor, giving it SMF records from just part of a sysplex — for example a single MVS image. Suppose that MVS image has no TOR, but its AORs just receive CICS transactions routed from another MVS image. Since the transaction ends in the TOR, but is executed in the AOR, there will be no ended transactions for the CICS service class from this MVS image.

Snapshot data is not always complete: The Workload Activity report contains "snapshot data", which is data collected over a defined interval. For IMS, in a given interval, several transactions may have already executed in an MPP, but the control region may not have reported their completion yet. Similarly for CICS, the AOR could have finished executing a transaction, but the TOR may not have completed processing and reporting the completion. Thus, RMF may show more executed transactions than completions.

Possible Actions

Classify IMS transactions uniquely: When an IMS transaction makes a program-to-program switch, the system checks the classification rules for the service class to be associated with the transaction. You can classify the switched transactions to a service class different from the service class of the originating transaction. RMF then counts a new transaction, with its own execution and end.

This way, the execution and ended values agree, and the RMF data is consistent with the data reported in the IMS Performance Analysis and Reporting System (IMSPARS). IMSPARS includes transaction timings, analysis reports, and detailed reports tracing individual transaction and database change activity. Classifying the program-to-program switched transaction to a unique service class might be useful for asynchronous program-to-program switches.

Combine all SMF records for the sysplex: When a single MVS image that does not have TORs is combined with another MVS image that does, the executed transactions from the first image will be combined by RMF with the ended transactions from the second so that the sysplex-wide report is consistent.

Problem: Execution Time is Greater Than Response Time

Figure 93 on page 156 shows an example of a work manager state section for the CICSPROD service class. In the example, there is a response time of .091 seconds, but an execution time of .113 seconds. The example also shows 1731 ENDED transactions, yet the EXCTD field shows that only 1086 have been executed.

```
REPORT BY: POLICY=HPTSPOL1   WORKLOAD=PRODWKLD   SERVICE CLASS=CICSPROD    RESOURCE GROUP=*NONE      PERIOD=1 IMPORTANCE=1

 TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT
 AVG      0.00   ACTUAL                 91
 MPL      0.00   EXECUTION             113
 ENDED    1731   QUEUED                  0
 END/S    1.92   R/S AFFINITY            0
 #SWAPS      0   INELIGIBLE              0
 EXCTD    1086   CONVERSION              0
 AVG ENC  0.00   STD DEV                92
 REM ENC  0.00
 MS ENC   0.00
```

Figure 93. Execution Time Greater than Response Time

Possible Explanation

Mixing routed and non-routed CICS transactions: The AORs may have recorded states which account for more time than the average response time of all the transactions. The non-routed transactions do not show up in the EXE phase. In addition, most non-routed transactions end very quickly, and decrease the actual response time. The response time (ACTUAL field) shows 0.091 seconds as the average of all 1731 transactions, while the AORs can only describe the execution of the 1086 transactions they participated in.

Possible Actions

Classify routed and non-routed transactions to different service classes: This would keep the numbers consistent with the expectation.

Problem: Large SWITCH Percentage in CICS Execution Phase

Figure 94 shows a work manager state data section for a CICSPROD service class. The LOCAL value in the SWITCHED SAMPL(%) section shows a value of 7092.

```
REPORT BY: POLICY=HPTSPOL1   WORKLOAD=PRODWKLD   SERVICE CLASS=CICSPROD   RESOURCE GROUP=*NONE      PERIOD=1 IMPORTANCE=1

 TRANSACTIONS    TRANS.-TIME  HHH.MM.SS.TTT
 AVG      0.00   ACTUAL                150
 MPL      0.00   EXECUTION             134
 ENDED    3599   QUEUED                 16
 END/S    4.00   R/S AFFINITY            0
 #SWAPS      0   INELIGIBLE              0
 EXCTD    2961   CONVERSION              0
 AVG ENC  0.00   STD DEV               446
 REM ENC  0.00
 MS ENC   0.00

           RESP  ------------------------------- STATE SAMPLES BREAKDOWN (%) -------------------------------  ------STATE------
SUB    P   TIME  --ACTIVE-- READY IDLE  ----------------------------WAITING FOR----------------------------  SWITCHED SAMPL(%)
TYPE        (%)   SUB  APPL             MISC PROD CONV  I/O                                                 LOCAL SYSPL REMOT
CICS  BTE  26.8K  0.3   0.0   0.4  2.5  96.3  0.0  0.6  0.0                                                   0.6   0.0   0.0
CICS  EXE  93.7  41.2   0.0   6.0  0.0   0.0 52.7  0.0  0.1                                                  7092   0.0   0.0
```

Figure 94. Large SWITCH Percentage in a CICS Execution Environment

Possible Explanations

Distributed transaction processing: If a program initiates distributed transaction processing to multiple back-end sessions, there can be many AORs all associated with the original transaction. Each of these multiple back-end regions can indicate they are switching control back to the front-end region (SWITCH to another region on the LOCAL MVS image, or to a region on another MVS image in the sysplex). Thus, with a one-to-many mapping like this, there are many samples of the

execution phase of requests switched — long enough to exceed 100% of the response time of other work completing in the service class.

Distributed program link (DPL): The distributed program link function from CICS/ESA® 3.3 builds on the distributed transaction functions available in CICS by enabling a CICS program (the client program) to call another program (the server program) in another CICS region.

While the server program is running, the client program will reflect that it is switched to another CICS region.

Possible Actions

- None -

Problem: Decreased Number of Ended Transaction with Increased Response Times

The Workload Activity report shows increased response times, and a decrease in the number of ended transactions over a few days.

Possible Explanation

Conversion from ISC link to MRO: When two CICS regions are connected via a VTAM inter-system communication (ISC) link, they behave differently than when they are connected via multi-region (MRO) option. One key difference is that, with ISC, both the TOR and the AOR are receiving a request from VTAM, so each believes it is starting and ending a given transaction.

So for a given user request routed from the TOR via ISC to an AOR, there would be 2 completed transactions. Let us assume they have response times of 1 second and 0.75 seconds, resulting in an average of 0.875 seconds.

When the TOR routes via MRO, the TOR will describe a single completed transaction taking 1 second, and the AOR will report its 0.75 seconds as execution time. Therefore, converting from an ISC link to an MRO connection, for the same workload, as shown in this example, could result in half the number of ended transactions and an increase in the response time reported by RMF. The difference could be much more significant if the AOR to FOR link was converted.

Migration of an AOR to CICS 4.1: CICS 4.1 has extended the information it passes between regions involved in processing a given transaction. If all the regions are at the 4.1 level, this allows RMF to report the number of ended transactions as just those that have completed in the TOR. But if there is a mixture of release levels involved, this is not guaranteed.

Let us assume that the TOR and FOR have been upgraded to release 4.1, but the AOR has not yet been upgraded. When the TOR receives a transaction, it will determine the service class. Then when it calls the AOR, it will pass that service class information.

But since the AOR is not yet at a 4.1 level, the AOR has no code that passes this service class along to the FOR. So when the FOR is invoked, it believes it is starting a new transaction and must classify the request to a service class. When the FOR finishes processing, it states that it has completed the transaction. Later the TOR also states that it has completed the transaction. This results in RMF showing multiple completions for a given end-user transaction (one completion

from the TOR, and one completion for each invocation of the FOR), with an average response time less than the value the TOR alone is reporting.

When the AOR is migrated to CICS 4.1, it will pass the service class to the FOR. The FOR recognizes that it is participating in executing some portion of the original end-user transaction. When the FOR returns to the AOR, and the AOR returns to the TOR, there will be just one ended transaction recorded for RMF. Its response time will be the time from reception in the TOR until the TOR has completed.

So RMF data will show a reduced number of transactions executed and a larger response time after the AOR is migrated to CICS 4.1 — even when the end-user request is processed in exactly the same elapsed time.

Possible Action

Increase CICS transaction goals: Do this prior to your conversion to an MRO connection, or prior to migrating your AOR to CICS 4.1, if the FOR transactions are classified to the same service class as your end-user transactions.

Analyzing Coupling Facility Activity

Using the Postprocessor Report

This chapter gives guidelines for some indicators when monitoring the coupling facility.

Usage Summary Section

```
                                 C O U P L I N G   F A C I L I T Y   A C T I V I T Y
                                                                                                                     PAGE   1
       z/OS V1R6              SYSPLEX PLEXPERF          DATE 04/11/2004          INTERVAL 030.00.000
                              RPT VERSION V1R5 RMF      TIME 13.39.00           CYCLE 1.000 SECONDS

------------------------------------------------------------------------------------------------------------------------------
COUPLING FACILITY NAME = CF5
TOTAL SAMPLES(AVG) =  1787  (MAX) =  1797  (MIN) =  1777
------------------------------------------------------------------------------------------------------------------------------
                                            COUPLING  FACILITY  USAGE  SUMMARY
------------------------------------------------------------------------------------------------------------------------------
STRUCTURE SUMMARY
------------------------------------------------------------------------------------------------------------------------------
                                                  % OF                % OF      AVG    LST/DIR  DATA      LOCK     DIR REC/
       STRUCTURE                        ALLOC     CF            #     ALL       REQ/   ENTRIES  ELEMENTS  ENTRIES  DIR REC
TYPE   NAME              STATUS CHG     SIZE      STORAGE     REQ     REQ       SEC    TOT/CUR  TOT/CUR   TOT/CUR  XI'S

   LIST   DSNDB1G_SCA       ACTIVE         16M         0.8     3395      0.1      1.89      26K       52K      N/A       N/A
                                                                                          92       158      N/A       N/A
          IXCPLEX_PATH1     ACTIVE         59M         2.9   257609      8.6    143.12      14K       14K      N/A       N/A
                                                                                           1        48      N/A       N/A

...

   LOCK   DSNDB1G_LOCK1     ACTIVE         31M         1.5    64373      2.2     35.76     114K        0     4194K      N/A
                                                                                          41         0     2730       N/A
          IRLMLOCK1         ACTIVE         31M         1.5     1676K    56.0    930.89     114K        0     4194K      N/A
                                                                                           6         0      256       N/A

   CACHE  DSNDB1G_GBP0      ACTIVE         21M         1.0     3289      0.1      1.83      22K     4321      N/A         0
                             PRIM                                                         58        35      N/A         0
          DSNDB1G_GBP4      ACTIVE         21M         1.0     3289      0.1      1.83      22K     4321      N/A         0
                             SEC                                                          58        35      N/A         0

...
                                      -------       -----   -------   ------   -------
             STRUCTURE TOTALS           263M        12.9     3077K      100    1709.2

   STORAGE SUMMARY
   ---------------------------------------------------------------------------------------------------------------------------


                                              ALLOC     % OF CF     ------- DUMP SPACE -------
                                              SIZE      STORAGE     % IN USE   MAX % REQUESTED

   TOTAL CF STORAGE USED BY STRUCTURES         263M       12.9
   TOTAL CF DUMP STORAGE                         6M        0.3          0.0              0.0
   TOTAL CF STORAGE AVAILABLE                    2G       86.8
                                             -------
   TOTAL CF STORAGE SIZE                         2G


                                              ALLOC     % ALLOCATED
                                              SIZE

   TOTAL CONTROL STORAGE DEFINED                 2G       13.2
   TOTAL DATA STORAGE DEFINED                    0K        0.0

   PROCESSOR SUMMARY
   ---------------------------------------------------------------------------------------------------------------------------

   COUPLING FACILITY        9672       MODEL C01      CFLEVEL 3
   AVERAGE CF UTILIZATION  (% BUSY)          7.8      LOGICAL PROCESSORS:   DEFINED   6   EFFECTIVE    6.0
```

Figure 95. Coupling Facility Activity Report - Usage Summary

The Usage Summary part of the report gives summary data for the structures, the storage and the processors.

Indicator

Fields:

```
LST/DIR   DATA      LOCK
ENTRIES   ELEMENTS  ENTRIES
TOT/CUR   TOT/CUR   TOT/CUR
```

Description: The first column reports on list entries defined and in use OR directory entries defined and in use, dependent on the type of structure. Similarly, the second column reports on the data elements defined and in use in the structure. The third column reports on lock entries.

The information presented on this report can be used to validate the size of various types of structures. In most cases, some knowledge of the structure is required to interpret the data.

- Cache structures

  The CUR in-use values will indicate if the structure is overlarge (if it never fills up) or misproportioned in terms of entry to element ratio (one never fills up but the other does). It will not help in knowing if the structure is too small (i.e. 100% full is normal). To know if the structure is too small you have to look at reclaiming and perhaps hit ratios.

- List structures

  Some list structures, like the JES2 checkpoint structure, are very static in size if not content/static data. For these structures, the TOT can be only slightly larger than CUR. Other functions, like the MVS logger, off-load data as the structure fills up. The threshold which triggers the Offload is based on parameters associated with the function. The ratio of CUR to TOT should not be higher than this threshold. Other list structures, such as the XCF structure and the VTAM Generic Resource structure, must handle large peaks of data. For these structures, the TOT should be much larger than the CUR in-use to prevent backups of signals during spikes.

- Lock structures

  Lock structures, like the IRLM lock structures, are divided in two parts: The actual lock table contains the number of entries in the LOCK ENTRIES TOT column. The number of LOCK ENTRIES CUR is a sampled value of the number in use at the end of the interval. The best way to evaluate the size of this table is to look at the FALSE CONTENTION data for the lock structure on the Coupling Facility Structure Activity report, as described later in this document.

  The second part of the lock structure contains RECORD DATA entries. The number of these entries is reported in the LIST ENTRIES column. If there are not enough of these entries, you will not be able to obtain locks, and transactions will start failing at that point.

  IRLM provides no external means of apportioning the structure storage between record data and locks; you can only obtain adequate entries for both parts by judiciously choosing the size of the lock structure. The easiest way to do this is to use the lookup tables in INFOSYS Q662530.

Indicator

Field: DIR RECLAIMS - DIR RECLAIMS XI'S

Description: All cache structures have directory entries and may or may not have data entries (for example, IMS OSAM and VSAM cache structures have only directory entries; while RACF and DB2 cache structures have both directory and data entries). N/A, for not applicable, will be displayed in the directory reclaim column for list and lock structures.

A cache structure can be over committed by the data base managers. This occurs when the total number of unique entities cached by the data base managers exceeds the number of directory entries in the cache structure. Whenever a shortage of directory entries occurs, the coupling facility will reclaim in-use directory entries associated with unchanged data. These reclaimed items will be used to satisfy new requests for directory entries by the data base manager. (Directory entries are used by the data base manager to ensure data validity). For the coupling facility to reclaim a directory entry, all users of the data item represented by the directory entry must be notified their copy of the data item is invalid. As a consequence, when the data base manager needs access to the now invalidated data item, the item must be re-read from DASD and registered with the coupling facility. When there are insufficient directory entries, the directory reclaim activity can lead to "thrashing". (The situation is analogous to real storage shortages and page stealing in MVS.) Directory reclaim activity can result in the following:

- Increased read I/O activity to the data base to re-acquire a referenced data item.
- Increased CPU utilization caused by registered interest in data items having to be re-acquired.
- Elongated transaction response times whenever a read miss occurs in the local buffer pool.

Directory reclaim activity can be managed by increasing the number of directory entries for a particular structure. This can be accomplished by:

- Increasing the size of the structure. For directory only structures, only the number of directory entries is affected (i.e. IMS OSAM and VSAM structures). For structures with data elements and directory entries, both will increase in the ratio specified by the structure user (i.e. RACF structures).
- Changing the proportion of the structure space used for directory entries and data elements. This action is dependent on the structure users' implementation. Some cache structure exploiters allow the installation to specify the ratio of directory entries to data entries, which it internally maps to a ratio of directory entries to data elements (a data entry may be composed of multiple data elements). DB2 provides this capability. On the other hand, RACF is hard coded to organize the structure in a 1:1 ratio of directory entries to data elements.

  If the cache structure directory to data element ratio is installation specifiable AND there are ample data elements, then one can increase the number of directory entries at the expense of data elements without a performance impact, and without increasing the structure size. Determining the impact of decreasing the number of data items is at best an inexact science. Unless the Structure Summary report indicates a consistent difference between the total and current number of data elements, it is difficult to estimate the impact of changing the

directory entry to data element ratio. The CUR value reported by RMF is not a peak value or average value during the interval, it is the number in use at the end of the interval.

The usual causes for directory reclaim activity are:

1. Initial structure size is insufficient to support the amount of local buffering.
2. Increasing the size of the local buffer pools without an accompanying increase in structure size.
3. Increasing the number of users of the structure without an accompanying increase in structure size.

The advice here is to monitor trends. If directory reclaims are increasing and response time is impacted, take appropriate action.

The second value in this column is the number of reclaims that caused an XI (see page 169). A high value is an indicator for a performance problem in this structure.

Indicator

Field: STRUCTURE SUMMARY - TOTAL CONTROL/DATA STORAGE DEFINED

Description: The amount of coupling facility storage that is allowed to be occupied by control information (CONTROL STORAGE) or data (DATA STORAGE).

Guideline: Each structure plus the dump area is allocated some control storage and some data storage. The coupling facility defines an area called control storage; structure control information is restricted to that area. The remaining storage is called data storage and is used for structure data. If the data storage area becomes full, structure data can then be allocated from the control storage area. If TOTAL DATA STORAGE DEFINED is zero, it means control information can reside anywhere on the coupling facility and there are no allocation restrictions.

If % ALLOC field for control storage shows a percentage approaching 100, it means the control storage is close to being completely allocated even though the CF SPACE AVAILABLE field may still show an amount of total free space. Possible actions include:

- Changing structure preference lists in the coupling facility policy specification to direct some structures away from this facility
- Adding another coupling facility to the sysplex

Indicator

Fields: AVERAGE CF UTILIZATION (%BUSY) LOGICAL PROCESSORS DEFINED / EFFECTIVE

Description: These fields report on the average CPU utilization of the CPs supporting the coupling facility partition, and in addition, on the number of logical processors assigned to the partition and the effective number of logical processors active during the RMF interval.

For example, if a CEC contains six CPs, and a test coupling facility LPAR has two logical CPs, but is capped at 5% of the CEC, then the LOGICAL PROCESSORS DEFINED will be 2 and the EFFECTIVE LOGICAL PROCESSORS number will be 0.3 (if the partition is uncapped the EFFECTIVE LOGICAL PROCESSORS number will be 2).

CFCC does not enter a wait state when there are no commands to be processed. CFCC continues to search for work until LPAR management removes processor resources or work is found (the ″active wait″ phenomenon).

Note: The above statement is true for LPARs providing real coupling facility functionality. It is NOT true for LPARs providing integrated coupling facility (ICMF) function. LPAR and CFCC have been enhanced to allow an ICMF LPAR with nothing to do to give up its CP resources. In this environment, use the Monitor I Partition Data Report, rather than the Coupling Facility Usage Summary report, to determine the utilization of the ICMF LPAR.

Guideline: These fields will assist you in diagnosing the following types of problems which have occurred in customer environments.

- ″High″ coupling facility processor utilization and excessively high service times when only one CP out of six in the CEC was assigned to the coupling facility LPAR.
- ″High″ coupling facility processor utilization and excessively high service times caused by two LPARs in the same CEC with active CFs. (Production coupling facility weight of 80, test coupling facility LPAR weight of 50, all CPs shared). The test coupling facility consumed 5/13 of the available CP resources even though it was unused.

Coupling facility processor utilization does influence structure access service times and thus the responsiveness and CPU utilization of the coupling facility exploiter.

Please, see “Factors Affecting Coupling Facility Performance” on page 176 for additional information.

Structure Activity Section

```
                              C O U P L I N G   F A C I L I T Y   A C T I V I T Y
                                                                                                                    PAGE   3
      z/OS V1R6              SYSPLEX PLEXPERF          DATE 04/11/2004          INTERVAL 030.00.000
                             RPT VERSION V1R5 RMF      TIME 13.39.00           CYCLE 1.000 SECONDS

-------------------------------------------------------------------------------------------------------------------------------
 COUPLING FACILITY NAME = CF5
-------------------------------------------------------------------------------------------------------------------------------
                                          COUPLING  FACILITY  STRUCTURE  ACTIVITY
-------------------------------------------------------------------------------------------------------------------------------

   STRUCTURE NAME = COUPLE_CKPT1      TYPE = LIST             STATUS = ACTIVE
            # REQ    -------------- REQUESTS -------------   -------------- DELAYED REQUESTS -------------
   SYSTEM   TOTAL             #     % OF  -SERV TIME(MIC)-    REASON     #    % OF  ----AVG TIME(MIC) -----   EXTERNAL REQUEST
   NAME     AVG/SEC           REQ    ALL     AVG   STD_DEV                 REQ    REQ    /DEL     STD_DEV   /ALL   CONTENTIONS

   J80       8463    SYNC   2927     5.8   306.6    127.8     NO SCH 1240   22.4  508.4      597.7  113.9   REQ TOTAL     8339
             4.70    ASYNC  5535    11.0  1502.1   1263.2     PR WT     0    0.0    0.0        0.0    0.0   REQ DEFERRED   123
                     CHNGD     1     0.0  INCLUDED IN ASYNC   PR CMP    0    0.0    0.0        0.0    0.0
                                                              DUMP      0    0.0    0.0        0.0    0.0

...

  -------------------------------------------------------------------------------------------------------------------------------
   TOTAL    50370    SYNC     17K   34.3   342.2    124.5     NO SCH 8797   26.6   1047       1245  278.4   REQ TOTAL       50K
            27.98    ASYNC    33K   65.5  557551G     0.0     PR WT     0    0.0    0.0        0.0    0.0   REQ DEFERRED   497
                     CHNGD    69     0.1                     PR CMP    0    0.0    0.0        0.0    0.0
                                                              DUMP      0    0.0    0.0        0.0    0.0
...
   STRUCTURE NAME = IRLMLOCK1         TYPE = LOCK             STATUS = ACTIVE
            # REQ    -------------- REQUESTS -------------   -------------- DELAYED REQUESTS -------------
   SYSTEM   TOTAL             #     % OF  -SERV TIME(MIC)-    REASON     #    % OF  ---- AVG TIME(MIC) -----   EXTERNAL REQUEST
   NAME     AVG/SEC           REQ    ALL     AVG   STD_DEV                 REQ    REQ    /DEL     STD_DEV   /ALL   CONTENTIONS

...

   JF0        128K   SYNC    128K    7.7   221.9     37.5     NO SCH    0    0.0    0.0        0.0    0.0   REQ TOTAL      128K
            71.26    ASYNC     0     0.0     0.0      0.0     PR WT     0    0.0    0.0        0.0    0.0   REQ DEFERRED    80
                     CHNGD     0     0.0  INCLUDED IN ASYNC   PR CMP    0    0.0    0.0        0.0    0.0   -CONT           80
                                                              DUMP      0    0.0    0.0        0.0    0.0   -FALSE CONT     25
...

  -------------------------------------------------------------------------------------------------------------------------------
   TOTAL     1676K   SYNC   1676K    100   198.1     44.7     NO SCH    0    0.0    0.0        0.0    0.0   REQ TOTAL     1674K
            930.9    ASYNC     0     0.0     0.0      0.0     PR WT     0    0.0    0.0        0.0    0.0   REQ DEFERRED   906
                     CHNGD     0     0.0                     PR CMP    0    0.0    0.0        0.0    0.0   -CONT          906
                                                              DUMP      0    0.0    0.0        0.0    0.0   -FALSE CONT    225
...
   STRUCTURE NAME = DSNDB1G_GBP3      TYPE = CACHE            STATUS = ACTIVE PRIMARY
            # REQ    -------------- REQUESTS -------------   -------------- DELAYED REQUESTS -------------
   SYSTEM   TOTAL             #     % OF  -SERV TIME(MIC)-    REASON     #    % OF  ---- AVG TIME(MIC) -----
   NAME     AVG/SEC           REQ    ALL     AVG   STD_DEV                 REQ    REQ    /DEL     STD_DEV   /ALL
...

   JA0       1032    SYNC   1025    19.1   448.8     84.3     NO SCH    7    100   2060       1671   2060
             0.57    ASYNC     0     0.0  4508.9   1409.1     PR WT     0    0.0    0.0        0.0    0.0
                     CHNGD     7     0.1  INCLUDED IN ASYNC   PR CMP    0    0.0    0.0        0.0    0.0
                                                              DUMP      0    0.0    0.0        0.0    0.0

...

  -------------------------------------------------------------------------------------------------------------------------------
   TOTAL     5356    SYNC   5342    100   432.8     89.0     NO SCH   14    100   1510       1374   1510   -- DATA ACCESS ---
             2.98    ASYNC     0     0.0  3685.2   1634.2     PR WT     0    0.0    0.0        0.0    0.0   READS         1331
                     CHNGD    14     0.3                     PR CMP    0    0.0    0.0        0.0    0.0   WRITES        1709
                                                              DUMP      0    0.0    0.0        0.0    0.0   CASTOUTS        44
                                                                                                           XI'S          1296
```

Figure 96. Coupling Facility Activity Report - Structure Activity

This section of the report provides detailed information on the frequency of structure operations and the associated service/queue times.

Indicator

Fields:

```
             ------------- DELAYED REQUESTS -------------
              REASON    #   % OF  ---- AVG TIME(MIC) -----
                      REQ    REQ   /DEL    STD_DEV   /ALL
```

Coupling Facility exploiters make structure manipulating requests through the APIs provided by Cross-system extended services (XES). In some instances, the API requests can specify that the operation be performed synchronously or asynchronously relative to requesting task. XES will execute the API request through some number of commands to the Coupling Facility. The coupling facility commands are in turn executed synchronously or asynchronously, based on the type of operation and the API specification. RMF reports on the operations to the coupling facility and the structures within. The only data on API requests is reported in CONTENTION data for lock and serialized list structures.

For an operation to be initiated against a coupling facility structure, one of the subchannels associated with the Coupling Facility must be available. When no subchannel is available, XES must either queue the operation for later scheduling, or ″spin″ waiting for a subchannel to become available. XES takes one action or the other, depending on the type of structure and operation.

Most SYNC operations which cannot be started due to unavailable subchannels are changed to ASYNC operations (RMF reports them as CHNGD). Certain SYNC operations are not changed from SYNC to ASYNC when no subchannels are available. Most notably, lock requests are not changed to ASYNC. For these operations, the processor ″spins″ until a subchannel becomes available.

RMF reports on ASYNC operations to a structure which are delayed due to subchannel unavailability. RMF does NOT report on SYNCH operations to a structure which are delayed (″spinning″) due to subchannel unavailability. The amount of delay for SYNC operations must estimated from the subchannel activity report.

Subchannel unavailable condition will manifest itself in one of the following ways:

1. A non-zero percentage of CHNGD requests relative to the total number of ASYNC requests.
2. A non-zero percentage of delayed ASYNC requests.

Depending on the structure either one or both of the above phenomena may be observed, not necessarily both.

In evaluating response times for ASYNC requests, the following should be noted:

- ASYNC command processing is performed primarily by the I/O processor (IOP or SAP).
- ASYNC requests generally take longer (elapsed time) than SYNC requests.
- The reported service time (not shown in the above example, but present on the ASYNC line) does not include delay time. To determine the average elapsed time for an AYSNC operation add the average service time to the average amortized delay time (/ALL column) for all ASYNC operations during the interval.

- ASYNC delay times (/DEL column) can be an indication of inadequate processing capacity in the coupling facility as well as insufficient number of subchannels.

In evaluating response times for SYNC requests, the following should be noted:

- SYNC command processing is performed by the CP.
- The reported service time (not shown in the above example, but present on the SYNC line) does not include the delay spent ″spinning″ awaiting subchannel busy if OW12415 is installed.
- A SYNC request delayed due to unavailable subchannel is using CPU resources.

Subchannel unavailability has the following negative consequences:

- ASYNC requests will be queued, resulting in delayed access to a coupling facility structure and corresponding diminished responsiveness by the exploiter.
- Non-immediate SYNC requests will be transformed to ASYNC requests, resulting in delayed access to a coupling facility structure and corresponding diminished responsiveness by the exploiter.
- Immediate SYNC requests will ″spin″ waiting for an available subchannel, resulting in increased CPU usage, delayed access to a coupling facility structure and corresponding diminished responsiveness by the exploiter.

Normally, none of these effects are desirable. The following activities can minimize the detrimental effects:

- Make certain CPU resource is sufficient in the coupling facility LPAR.
- Redistribute structures among the CFs to minimize the delay.
- Add additional coupling facility links between the MVS processor and the coupling facility. Each path will contribute two subchannels.
- Influence the coupling facility activity rates (downward) of coupling facility exploiters if possible. For example, you can reduce the number of XCF signals by reducing lock contention or tuning XCF to eliminate signals related the expansion of transport classes sizes.

Indicator

Fields:

EXTERNAL REQUEST CONTENTIONS	EXTERNAL REQUEST CONTENTIONS
REQ TOTAL	REQ TOTAL
REQ DEFERRED	REQ DEFERRED
	-CONT
	-FALSE CONT

Description: These fields provide information on serialized list contention.

A serialized list structure is a list structure which has an associated lock structure within. For example, JES2 and VTAM structures are serialized list structures; the XCF structure is simply a list structure. A serialized list structure can provide the structure user with increased control (exclusive usage) over the entire structure, a portion of the entries or even a single list entry. The use of the lock structure to control access is completely under the control of the exploiter.

There are a number of operations against a serialized list structure Some operations against a serialized list structure will succeed if the designated lock entry is not held by any other user. Additionally, these requests will be queued for FIFO processing by XES whenever the command cannot be processed due to lock contention. (Specifically these are UNCONDITIONAL SET and NOTHELD requests). XES will queue the request in a sysplex-wide FIFO queue until the lock is available. Whenever the lock is released, the first queued request will be restarted on the appropriate processor. XES is providing multi-system FIFO queuing on behalf of the users to the serialized list structure.

RMF will report on the number of accesses to the serialized list structure and the number of accesses deferred due to lock contention. This value is reported from the perspective of the indicated connector, the value is not a ″global″ value across all connectors to the structure. Note: A serialized list structure will not incur false contention due to the manner in which XES manages the contention.

Having XES queue and manage deferred requests requires some amount of CPU usage and XCF signalling usage. Excess contention is most likely a reflection on the usage of the structure itself by the exploiter or poor tuning specifications by the installation. For example, consider JES2s use of a checkpoint structure. Each JES2 member of the MAS will acquire and hold the checkpoint for some period of time (HOLD) and restrain itself from attempting to acquire control for some period of time upon release (DORMANCY). If all members of the MAS specify short dormancy values, then each member (except the current owner) will attempt to acquire the checkpoint lock. XES will happily manage this queue of requestors in a FIFO manner at some cost. It may be worth the effort to adjust these values so members of the MAS receive the necessary service without the overhead of XES managing the queue of requestors.

The average time a serialized list request with lock contention is queued by XES is not tracked and thus not reported by RMF. APAR OW11789 corrected a reporting problem which added the queue time into the AYSNC service time. This resulted in inflated and widely varying service times being reported.

Indicator

Fields:

```
-- DATA ACCESS ---
READS
WRITES
CASTOUTS
XI'S
```

Description: These fields provide information on cache structure accesses.

Data access counts for a cache structure have been added to the report. Prior to describing what READS, WRITES, etc. counts actually count, the following should be noted:

- The information is acquired from counters in the coupling facility and is global in nature. The information cannot be broken down to give an individual connection contributions to this global count. Thus, the information is reported in the ″TOTAL section″ of the RMF report.
- Depending on the cache structure implementation one or more of these counter values may be zero.

- Most importantly, only the cache structure user (i.e. data base manager) knows how efficiently the local buffers and cache structure are being used. If the data base manager does not provide usage information, then tuning is a best guess proposition.

The coupling facility maintains a number of structure level counters accessible to RMF. RMF reports on a subset of these counters (to be described momentarily) and stores all the counters in the SMF74 subtype 4 records.

Additional information on these counters can be obtained in *z/OS MVS Programming: Sysplex Services Guide* in the section labeled IXLCACHE REQUEST=READ_STGSTATS.

RMF provides four cache structure usage counters.

READS
This is a count of the number of times the coupling facility returned data on a read request by any connector.

Directory only caches will always have a zero value reported since the structure contains no ″data″.

WRITES
This is a count of the number of times a connector placed changed or unchanged data into the coupling facility structure. (Changed/unchanged is an attribute assigned to the data when written by the connector. From a performance/capacity view point, the importance of the attribute is: changed data cannot be reclaimed from the structure should directory or data elements become scarce.)

Just as directory only cache structures always have READS counts of zero, they have WRITES counts of zero as the structure contains no ″data″.

The condition to be concerned with is a large number of WRITEs and a small number of READS. This condition may indicate:

- Insufficient structure space allocated, and data entries (and perhaps directory entries) are being discarded by the coupling facility space management routines.
- Inappropriate allocation of the ratio of directory entry to data elements is causing the data entries to be discarded by the coupling facility space management routines.

Directory entry reclaim activity is reported by RMF as discussed earlier; however, data element reclaim activity is not reported.

CASTOUTS
This is a count of the number of times a connector retrieved a changed data entry, wrote the data to DASD (discarded it, whatever), and caused the changed attribute to be reset to unchanged. This process is known as casting out changed data from the structure.

This counter is of interest for store-in cache structures (i.e. DB2 global buffer pool structures) in determining the volume of changed data being removed from the structure. This counter is not an indicator of the number times castout processing was performed during the RMF interval.

A large amount of castout activity on a single structure may warrant additional cache structures and redirecting locally buffered data to different cache structure. Castout processing by the connectors must keep pace with

the rate at which changed data is placed in the structure. When all directory or data elements are associated with changed data, no new data items can be registered or written to the structure. This condition is not desirable and will adversely effect the data base manager and/or the user of the data base manager.

XI'S This is the number of times a data item residing in a local buffer pool was marked invalid by the coupling facility during the RMF interval. To the cache structure user, this means the data item must be re-acquired from DASD or perhaps the coupling facility structure, and interest in the item must be re-registered in the coupling facility structure.

There are several ″XI counts″ obtained from the coupling facility which are consolidated into this value. They are:

- XI for Directory Reclaim
- XI for Write
- XI for Name Invalidation
- XI for Complement Invalidation
- XI for Local Cache Vector Entry Replacement

Consult section IXLCACHE REQUEST=READ_STGSTATS in the *z/OS MVS Programming: Sysplex Services Guide* for additional information on these XI categories.

It is important to understand what the XI'S counter really counts. This is most easily explained by example. Suppose there are 5 connectors (i.e. #1, #2, #3, #4, and #5). to a cache structure. Further suppose data item ″A″ is locally cached in connectors #1, #2, #4, and #5 and registered in the coupling facility structure. Now suppose connector number #5, changes data item ″A″ and issues the command to the coupling facility to have all other copies invalidated (i.e. XI for complement invalidation). In response to the command, the coupling facility would cause the copies of data item ″A″ in the local buffer pools of connectors #1, #2 and #4 to be marked invalid. At the completion of the XI operation, the XI'S count value would be incremented by 3!

XI'S count values are seen for directory, store-in and store-thru caches. This count reflects both the amount of data sharing among the users of the cache and the amount of write/update activity against the data bases.

Subchannel Activity Section

```
                              C O U P L I N G   F A C I L I T Y   A C T I V I T Y
                                                                                                                                  PAGE   6
    z/OS V1R6              SYSPLEX PLEXPERF          DATE 04/11/2004          INTERVAL 030.00.000
                           RPT VERSION V1R5 RMF      TIME 13.39.00           CYCLE 1.000 SECONDS

------------------------------------------------------------------------------------------------------------------------------------------
 COUPLING FACILITY NAME = CF1
------------------------------------------------------------------------------------------------------------------------------------------
                                                    SUBCHANNEL  ACTIVITY
------------------------------------------------------------------------------------------------------------------------------------------
           # REQ                                  ---------- REQUESTS ----------   --------------  DELAYED  REQUESTS  --------------
 SYSTEM    TOTAL    --  CF  LINKS  --  PTH                    #   -SERVICE TIME(MIC)-                #    % OF   -----  AVG  TIME(MIC)  ----
 NAME      AVG/SEC  TYPE   GEN   USE  BUSY                   REQ     AVG    STD_DEV                  REQ    REQ    /DEL     STD_DEV     /ALL

*W04         30740  ICP      3    3     0   SYNC       957    26.8      95.5   LIST/CACHE     2   0.0  174.0        50.9    348.0
              51.2  SUBCH   42   21         ASYNC    22791   736.7      2811   LOCK           0   0.0    0.0         0.0      0.0
                                            CHANGED      0  INCLUDED IN ASYNC  TOTAL          2   0.0
                                            UNSUCC       0     0.0       0.0
*W05          9793  CFS      2    2   354   SYNC       790    43.3     116.5   LIST/CACHE  1537  19.1   1797        5626    2761K
               9.6  SUBCH    4    4         ASYNC     8100   937.8      3543   LOCK           4   0.5   69.5        15.6    278.0
                                            CHANGED      1  INCLUDED IN ASYNC  TOTAL       1541  17.3
                                            UNSUCC       0     0.0       0.0
```

Figure 97. Coupling Facility Activity Report - Subchannel Activity

The following fields can be used as a quick way of determining which systems are generating the most activity for a given facility, which in turn indicates where to focus tuning or load balancing efforts.

Indicator
Field: CF LINKS

Description: The subchannel configuration is described by these fields:

TYPE This column describes the channel path type.

GEN USE The number of subchannels that are defined, and that MVS is currently using for coupling facility requests.

The description of the channel path type can help in better analyzing the performance of the different channel path types. There are differences between two types of coupling facility channels:

CFS /CFR - CBS /CBR - ICS / ICR
Two subchannels per path.

CFP - CBP - ICP
Seven subchannels per path.

Indicator

Field: PTH BUSY

Description: Path busy - the number of times a coupling facility request was rejected because all paths to the coupling facility were busy.

Guideline: A high count combined with lengthy service times for requests indicates a capacity constraint in the coupling facility. If coupling facility channels are being shared among PR/SM partitions, the contention could be coming from a remote partition.

Identifying path contention: There can be path contention even when the path busy count is low. In fact, in a non-PR/SM environment where the subchannels are properly configured, # DELAYED REQUESTS, not PTH BUSY, is the indicator for path contention. If PTH BUSY is low but # DELAYED REQUESTS is high, it means MVS is delaying the coupling facility requests and in effect gating the workload before it reaches the physical paths.

PR/SM environment only: If coupling facility channels are being shared among PR/SM partitions, PTH BUSY behaves differently. You potentially have many MVS subchannels mapped to only a few coupling facility command buffers. You could have a case where the subchannels were properly configured (or even underconfigured), subchannel busy is low, but path busy is high. This means the contention is due to activity from a remote partition.

If a coupling facility capacity constraint is suspected, the first action to consider is adding more paths. If the coupling facility is already fully configured for paths, you need to determine whether some structures can be off-loaded to another coupling facility or whether another coupling facility should be added to the configuration. Use the Coupling Facility Usage Summary reports to balance workloads across coupling facilities based on activity rates and storage usage.

In a PR/SM environment, you may need to increase or adjust the configuration of shared coupling facility channels to reduce path contention.

Indicator Fields:

```
               --------------- DELAYED REQUESTS -----------
                      #     % OF   ------AVG TIME(MIC)-----
                      REQ    REQ    /DEL   STD_DEV   /ALL

               LIST/CACHE
               LOCK
               TOTAL
```

Description: Field TOTAL # DELAYED REQUESTS is the same as BUSY COUNT - SCH that was shown in previous versions of this report. It is the number of times an immediate request was delayed because subchannel resources were not available.

Immediate requests (such as locking operations) could not be completed because all the subchannels were busy. These requests are not queued. They are processed before the non-immediate requests which are reported in the Structure Activity report under QUEUED requests. To get a complete picture of subchannel activity, look at both fields.

Guideline: If this count is high, you should ensure that sufficient subchannels are defined.

See PTH BUSY above for suggested actions to relieve a coupling facility path constraint.

In a data sharing environment, lock structure requests account for a large percentage of coupling facility accesses. By estimating the amortized delay for SYNC operations on lock structure accesses, one can determine the time delay due to subchannel unavailable on lock structure operations and the impact on CPU busy.

Note: Requests which are delayed because a structure was being dumped were removed from this report because they are not related to subchannel contention.

CF to CF Activity Section

```
                              C O U P L I N G   F A C I L I T Y   A C T I V I T Y
                                                                                                                                                  PAGE   6
      z/OS V1R6              SYSPLEX PLEXPERF          DATE 04/11/2004          INTERVAL 030.00.000
                             RPT VERSION V1R5 RMF      TIME 13.39.00           CYCLE 1.000 SECONDS

   ------------------------------------------------------------------------------------------------------------------------------------------------
    COUPLING FACILITY NAME = CF1
   ------------------------------------------------------------------------------------------------------------------------------------------------
                                      CF TO CF  ACTIVITY
-------------------------------------------------------------------------------------------------------------------------------------------------
         # REQ                      ---------- REQUESTS ------------   ------------- DELAYED REQUESTS --------------
PEER      TOTAL  -- CF LINKS --              #   -SERVICE TIME(MIC)-              #    % OF  ---- AVG TIME(MIC) ---
CF       AVG/SEC TYPE       USE             REQ     AVG    STD_DEV               REQ   REQ   /DEL  STD_DEV    /ALL

CF2       86830  CBP          1       SYNC     86830    53.0      33.8   SYNC         19   0.2%   14.5     2.7     0.0
           86.2  ICP          1
```

Figure 98. Coupling Facility Activity Report - CF to CF Activity

This section of the Coupling Facility Activity report provides information about the activities between coupling facilities.

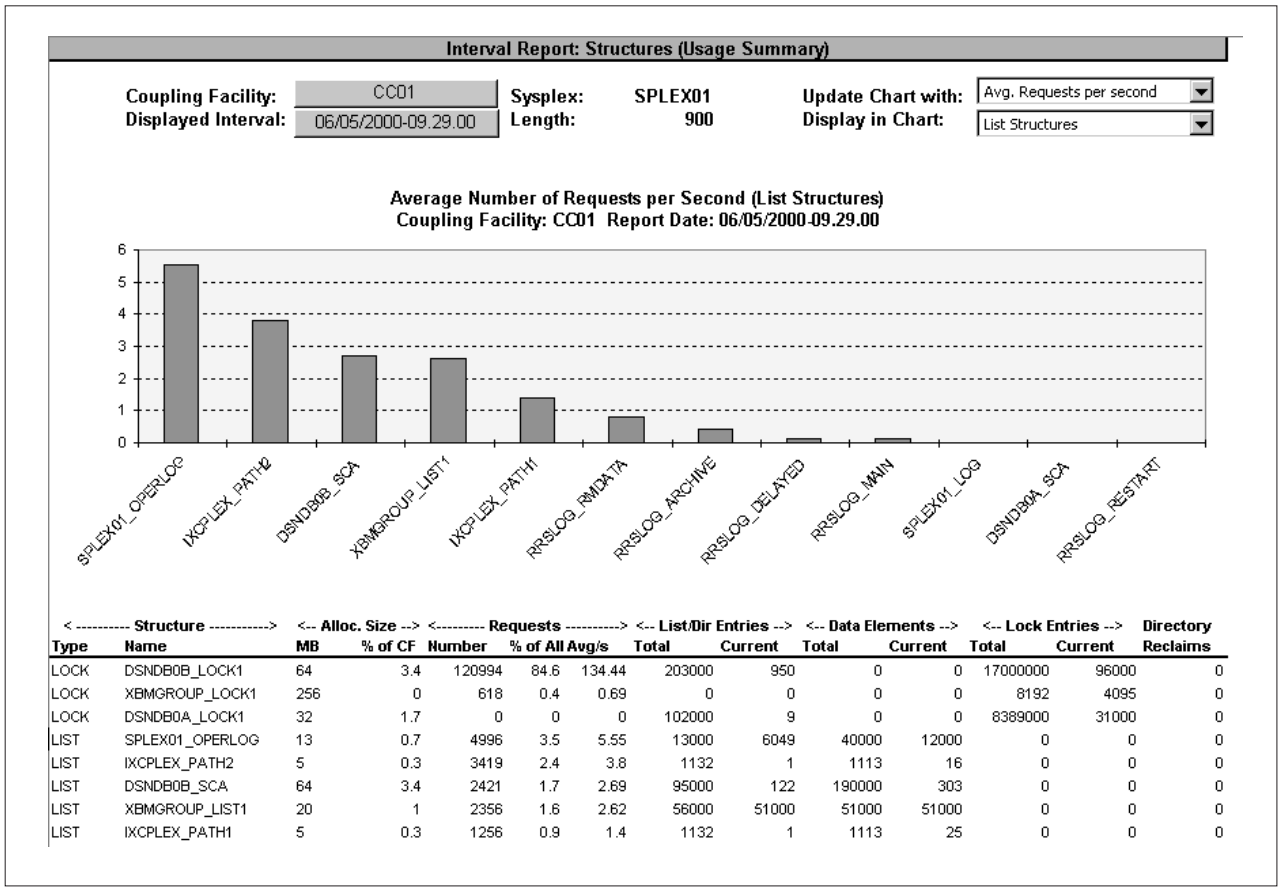
Indicator Fields:

```
---------------- DELAYED REQUESTS -----------
       #     % OF   ------AVG TIME(MIC)-----
       REQ    REQ    /DEL   STD_DEV   /ALL
```

Description: Field # DELAYED REQUESTS is the number of signals of all types which have experienced a delay in being sent from the subject CF to this remote CF. It is an contention indicator for the peer-to-peer communication between both coupling facilities.

Spreadsheet Report

If you are interested to display the data of the Coupling Facility Activity report graphically, you can use the Spreadsheet Reporter with an interval report as input. One of the several charts that you can get is the following one:



< --------- Structure ---------->		<-- Alloc. Size -->		<--------- Requests ---------->			<-- List/Dir Entries -->		<-- Data Elements -->		<-- Lock Entries -->		Directory
Type	Name	MB	% of CF	Number	% of All	Avg/s	Total	Current	Total	Current	Total	Current	Reclaims
LOCK	DSNDB0B_LOCK1	64	3.4	120994	84.6	134.44	203000	950	0	0	17000000	96000	0
LOCK	XBMGROUP_LOCK1	256	0	618	0.4	0.69	0	0	0	0	8192	4095	0
LOCK	DSNDB0A_LOCK1	32	1.7	0	0	0	102000	9	0	0	8389000	31000	0
LIST	SPLEX01_OPERLOG	13	0.7	4996	3.5	5.55	13000	6049	40000	12000	0	0	0
LIST	IXCPLEX_PATH2	5	0.3	3419	2.4	3.8	1132	1	1113	16	0	0	0
LIST	DSNDB0B_SCA	64	3.4	2421	1.7	2.69	95000	122	190000	303	0	0	0
LIST	XBMGROUP_LIST1	20	1	2356	1.6	2.62	56000	51000	51000	51000	0	0	0
LIST	IXCPLEX_PATH1	5	0.3	1256	0.9	1.4	1132	1	1113	25	0	0	0

Figure 99. Coupling Facility Structure Activity. Spreadsheet Reporter macro RMFR9CF.XLS (CF Report)

The graphic shows the total number of requests for structure ISGLOCK, and you can modify the graphic by selecting a specific type of data in the drop-down list.

Using the Monitor III Online Reports

Three new Monitor III reports enable you to analyze bottlenecks or problems in the coupling facility area that might result in a performance degradation in the parallel

sysplex. To avoid critical situations for production data bases and transaction processing systems, customers can see the immediate state of the coupling facility and the structures, and they have all relevant data at a glance in a granularity they can choose. In addition, they can see the results of tuning actions in this area, for example after having removed a coupling structure from an overloaded coupling facility to another one.

The Coupling Facility Overview Report - CFOVER

The Coupling Facility Overview report (CFOVER) gives you information about all coupling facilities which are connected to the sysplex.

```
                  RMF V1R5  CF Overview      - UTCPLXHD          Line 1 of 2
Command ===>                                                  Scroll ===> HALF

Samples: 120     Systems: 3    Date: 05/04/04  Time: 11.27.00  Range: 120   Sec

---- Coupling Facility -----    ---- Processor -----   Request    -- Storage --
Name      Type   Model Level    Util% Defined Effect   Rate       Size    Avail

CF5B      9674     C01     5    12.4      3     3.0      232.3    1008M    679M
CF6B      9674     C01     5    44.2      3     3.0       4572    1008M    651M
```

Figure 100. CFOVER Report

The Coupling Facility Activity Report - CFACT

The Coupling Facility Activity report (CFACT) gives you information about the activities in each structure.

```
                  RMF V1R5  CF Activity      - UTCPLXHD      Line 13 of 76
Command ===>                                                Scroll ===> HALF

Samples: 120     Systems: 3    Date: 05/04/04  Time: 11.27.00  Range: 120   Sec

CF: ALL          Type  ST System      --- Sync ---     --------- Async --------
                                        Rate   Avg       Rate   Avg   Chng   Del
Structure Name                                Serv             Serv   %      %

IRRXCF00_B001    CACHE    *ALL          0.1    372       0.0      0   0.0   0.0
                          D0            0.0      0       0.0      0   0.0   0.0
                          D1            0.0      0       0.0      0   0.0   0.0
                          D2            0.1    372       0.0      0   0.0   0.0
IRRXCF00_P001    CACHE    *ALL          0.5    378       0.0      0   0.0   0.0
                          D0            0.2    353       0.0      0   0.0   0.0
                          D1            0.2    401       0.0      0   0.0   0.0
                          D2            0.2    378       0.0      0   0.0   0.0
ISGLOCK          LOCK     *ALL         14.5    266       0.0      0   0.0   0.0
                          D0            1.1    235       0.0      0   0.0   0.0
                          D1            7.2    293       0.0      0   0.0   0.0
                          D2            6.1    241       0.0      0   0.0   0.0
ISTGENERIC       LIST     *ALL          1.7    363       0+    8393  50.0   150
                          D0            0.8    382       0.0      0   0.0   0.0
                          D1            0.4    353       0+    8393  50.0   150
                          D2            0.5    341       0.0      0   0.0   0.0
IXCPLEX_PATH1    LIST     *ALL          0.0      0     248.6   2553   0.0  26.0
                          D0            0.0      0     129.4   2107   0.0  39.2
                          D1            0.0      0      26.6   3981   0.0  34.0
                          D2            0.0      0      92.6   2767   0.0   5.2
IXCPLEX_PATH2    LIST     *ALL          0.0      0     182.5   3092   0.0   3.3
                          D0            0.0      0      82.6   2730   0.0   2.9
                          D1            0.0      0      25.5   4833   0.0  10.7
```

Figure 101. CFACT Report

The Coupling Facility Systems Report - CFSYS

The Coupling Facility Systems report (CFSYS) gives you information about the distribution of coupling facility requests among the systems and about the activities in the subchannels and paths attached to the coupling facilities in the sysplex.

```
                  RMF V1R5  CF Systems       - UTCPLXHD         Line 1 of 6
Command ===>                                                Scroll ===> HALF

Samples: 120     Systems: 3    Date: 05/04/04  Time: 11.27.00  Range: 120   Sec

CF Name   System      Subch   -- Paths --   -- Sync ---   ------- Async -------
                      Delay   Avail Delay    Rate  Avg     Rate  Avg  Chng  Del
                        %               %           Serv           Serv   %    %

CF5B      D0            0.0     2     2.6     5.1   344    91.0  2890  0.1  7.6
          D1            0.2     2     7.3    10.6   315    33.8  4903  0.4 22.1
          D2            0.0     2     3.3     9.3   300    82.2  3042  0.0  7.2
CF6B      D0            5.1     1     2.2    2131   242   252.2  1974 31.0 39.3
          D1            1.1     1    34.1     0.9   548    30.9  3884  0.0 42.6
          D2            0.0     2     0.7    1895   267   144.8  2865  2.8  7.7
```

Figure 102. CFSYS Report

Factors Affecting Coupling Facility Performance

This section is intended to assist one in detecting and assessing processing delays which will impact applications utilizing the coupling facility. Processing delays may manifest themselves as increased in CPU time, increased in elapsed time due to waiting, or some combination of both. There are several approaches to achieving better application or system performance:

- Perform ″work″ as quick as possible.
- Don’t create additional ″work″.
- Don’t do unnecessary work.

Perform ″Work″ as Quick as Possible

Conditions elongating coupling facility request response (service + delay) times include the following:

- Insufficient processor power allocated to the coupling facility LPAR.

  The most common cause of this condition is the sharing of CPs among coupling facility LPARs. The changes to RMF and CFCC will allow the installation to accurately determine the amount of available processing power being used.

  Elongation of SYNC service times against directory only cache structures is a good indicator of this condition. Alternatively, use the lock structure SYNC service times after removing the no subchannel delay component.

- Insufficient number of paths between the CECs containing multiple MVS images and a coupling facility

  RMF does not report on the length of the delay due to path busy conditions, only the number of requests encountering the condition are reported. Requests encountering path delay will have elongated service times.

  The most common condition to cause path busy condition is EMIFing the sender ISC links among the MVS LPARs. The current rule of thumb is that the number of requests delayed for path busy (as reported on the Subchannel Activity report in the PTH field under the BUSY COUNTS column) should be less than 10% of total requests (on the same report under the # REQ TOTAL column).

  Note: In the case of EMIFed ISC links, be certain to take into account all the MVS LPARS utilizing a common physical connection(s).

  The solution to reducing the incidence of path busy conditions is to add ISC links between the processors or refrain from using EMIF, which will probably require additional links as well.

- Insufficient number of subchannels between the MVS image and the coupling facility

  RMF reports on delay times and service times at the structure level and the coupling facility level (as previously discussed). Consequences of insufficient subchannels are:

  – Increased MVS CPU consumption when the SYNC operation cannot be converted to an ASYNC operation.
  – Increased service time for the SYNC operation which cannot be converted to an ASYNC operation.
  – Requests being converted from SYNC to ASYNC.
  – Increased queue time for ASYNC requests.

  The effect of the last two items is difficult to quantify. The effect depends on what the exploiter can do while a request is queued (i.e. was the delay anticipated and the request made asynchronously). Either of these two events will result in an elapsed time increase for some function. Whether this affects only one ″user″ or many users is dependent on the exploiter and its use of the

structure. For example, if all cache structure castout operations are delayed, then it is possible for the cache to become full of changed data and is not a desirable condition.

There are two indicators to be monitored.

1. Increased SYNC operation service times (particularly against the lock structure) and the amount of ″spin time″ associated with the service time.
2. The percentage of operations delayed.

Use the same rules and values previously discussed.

The only way to increase the number of subchannels is to add more physical paths (i.e. ISC links). There are two subchannels per ISC link. Additional ISC links will help alleviate path busy conditions in an EMIF environment as well.

- Insufficient IOP/SAP capacity on the CEC MVS image and the coupling facility

  The IOP/SAP on the CEC housing the MVS image handles IO to DASD, CTC traffic and ASYNC operations. As the number of ASYNC operations increases, the workload on the IOP/SAP increases. If the IOP/SAP becomes overloaded, the service time for ASYNC operations will elongate. ASYNC operations can be quite long and have a large standard deviation, thus changes are somewhat difficult to detect. Perhaps the easiest indicator to track is the Avg Q Length on the I/O Queuing Activity Report; if this value gets large it is an indicator that ASYNC requests as well as I/O requests are being delayed in IOP/SAP processing.
- Hardware problems causing ISC link degradation.

  Monitor the hardware incidents in LOGREC and relevant console messages. Get the problem fixed.

Don’t Create Additional ″Work″

XES and the coupling facility must work within the constraints imposed by the people that setup the parallel sysplex. When structures are not of sufficient size, XES, the coupling facility, and the exploiter will cope as best they can.

For example if a directory only cache structure is being utilized and insufficient structure space is defined, directory entries will be reclaimed according to a least recently used algorithm. This will result in locally cached data items being marked invalid. Whenever these locally cached, but invalid, items are referenced, they will have to be read from DASD and registered with the coupling facility. In essence, unnecessary work has been created due to an improper definition. CPU cycles are required to perform the I/O and to register the data item with the coupling facility. Furthermore, the entity accessing the data item was delayed (elapsed time increased) while these two events occurred.

RMF now provides information allowing the installation to detect, and perhaps alleviate, these conditions on a structure by structure basis.

- Lock Structures

  Lock structure request service times should be short. Lock contention should be avoided, as contention must be resolved with the assistance of XES lock services requiring time and CPU cycles. Lock contention also generates additional XCF signalling. False lock contention should be avoided, it is expensive to resolve and is avoidable. RMF reports on the number of lock requests, the number delayed due to contention and the number delayed due to false contention.

  False lock contention can be alleviated by increasing the size of the lock table. To reduce true lock contention may require a change to the application, the data base, or the workload mix during the period of time the conflict occurs. One

should attempt to manage true contention to less than 1.0 percent and false contention to less than 0.1 percent of the requests.

- Cache Structures

  Cache structures are not meant to replace local caches, they are intended to augment their use by:

  1. Providing a mechanism for local buffer coherency
  2. Providing additional high speed caching for frequently changed data.

  Use of the cache structure should not negatively impact the use of local cache.

  Prevent invalidation of a local cache entity unless the entity has been changed. This means avoid doing directory entry reclaims. This activity may be avoided by specifying a larger structure. RMF now reports on this activity to assist the installation in verifying the cache structure contains sufficient directory entries to keep track of locally cached data.

  If the cache structure user keeps data in the structure, attempt to keep frequently referenced data in the structure. Caching data in the coupling facility is only a performance boost when the data resides in the structure. It is important to have the structure sufficiently large and of the correct entry/data item configuration to support coupling facility caching of data without hindering performance.

  The RMF changes discussed previously will assist the installation in tracking anomalies or changes in data access patterns. RMF data is not sufficient in determining optimal read hit ratios.

Don’t Do Unnecessary Work

This activity is difficult to assess. The servers using coupling facility structures for message passing, locking, caching, state saving, etc. normally do very little coupling facility accesses unless prompted by some user of the server. Normally applications are causing the servers to perform some task involving coupling facility accesses.

In extensive data sharing environments, the data base manager and the lock manager together will account for over half of accesses to the coupling facility. The applications using the data base manager are normally high profile and critical to the business. Thus the first two activities are most often pursued.

When all coupling facility tuning efforts have failed to meet the applications response time objectives, be it CPU time and/or elapsed time, examination of the application is in order. Unfortunately, this involves real work among different organizations. The causes of contention, excessive locking, cache disruption, etc. must be determined and reduced. This reduction will involve some type of application change. This is obviously not a popular option and is presented last as it is most begrudgingly done.

Appendix A. PR/SM LPAR Considerations

Considerations in PR/SM Environments

This appendix presents performance considerations for systems running in LPAR mode, including:

- Interpretation of RMF reports
- Tuning of an LPAR configuration

Understanding the Partition Data Report

Using the Postprocessor report option REPORTS(CPU), you get the CPU Activity report. This report is divided into these sections:

CPU ACTIVITY — Information about processors and address spaces

PARTITION DATA REPORT — Information about all configured partitions

LPAR CLUSTER REPORT — Information about LPAR clusters

Although this is an inconsistency, the first section is typically referred to as CPU Activity report, too. The third section is new, it is available with the Intelligent Resource Director (IRD) of the zSeries processors and is decribed in Appendix B, "The Intelligent Resource Director."

```
                                    C P U   A C T I V I T Y
                                                                                PAGE    1
           z/OS V1R6              SYSTEM ID NP1            DATE 04/15/2004         INTERVAL 14.59.678
                                  RPT VERSION V1R5 RMF     TIME 09.30.00           CYCLE 1.000 SECONDS

CPU  2064   MODEL  109

---CPU---  ONLINE TIME   LPAR BUSY      MVS BUSY      CPU SERIAL  I/O TOTAL        % I/O INTERRUPTS
NUM  TYPE  PERCENTAGE    TIME PERC      TIME PERC     NUMBER      INTERRUPT RATE   HANDLED VIA TPI
 2          19.33         16.77          100.0        211551       0.00              0.00
 3         100.00         26.49          99.88        311551       7.98              2.48
CP TOTAL/AVERAGE         24.92          99.90                     7.98              2.48
```

Figure 103. CPU Activity Report

The first part of the CPU Activity report provides information about the processors assigned to that partition which is gathering the data.

Report Analysis

- ONLINE TIME PERCENTAGE

  The total online time percentage for all processors is 119.33% which is equivalent to 1.2 processors. This value is shown in the Partition Data report in the field PROCESSORS NUM.

- LPAR BUSY TIME PERC

  The average utilization of the processors is 24.92%, this value can be found in the Partition Data report in the field LOGICAL PROCESSORS TOTAL for partition NP1. The LPAR BUSY time is (with Wait Completion = NO) the dispatch time of the processors that are assigned to the partition, and the percentage is based on their online time.

- MVS BUSY TIME PERC

  This field shows the MVS view of the CPU utilization. The MVS BUSY time is the difference between online time and wait time. If the MVS system is busy when the partition is losing control, it will stay in busy mode until the partition will be dispatched again, then the current task can continue, and wait state will be reached later. Therefore, the MVS BUSY time can be higher than the LPAR BUSY time, and the difference between both values is an indicator for CPU constraints in the system.

The Partition Data report shows measurement data for all configured partitions. The line *PHYSICAL* is for reporting purposes only; it does not reflect a real

partition.

PHYSICAL PROCESSORS TOTAL								
DISPATCH TIME DATA TOTAL (Partition 1)			...	DISPATCH TIME DATA TOTAL (Partition n)			...	*PHYSICAL*
DISPATCH TIME DATA EFFECTIVE			...	DISPATCH TIME DATA EFFECTIVE			...	
MVS Captured Time	MVS Uncaptured Time	Partition LPAR Mgmt Time	...	MVS Captured Time	MVS Uncaptured Time	Partition LPAR Mgmt Time	...	LPAR Time not attributed

```
                                P A R T I T I O N   D A T A   R E P O R T
                                                                                                  PAGE    2
          z/OS V1R6              SYSTEM ID NP1            DATE 04/15/2004          INTERVAL 14.59.678
                                 RPT VERSION V1R5 RMF     TIME 09.30.00            CYCLE 1.000 SECONDS


 MVS PARTITION NAME                    NP1                    NUMBER OF PHYSICAL PROCESSORS          9
 IMAGE CAPACITY                        100                                    CP                     9
 NUMBER OF CONFIGURED PARTITIONS         9                                    ICF                    0
 WAIT COMPLETION                        NO
 DISPATCH INTERVAL                 DYNAMIC

--------- PARTITION DATA -----------------  -- LOGICAL PARTITION PROCESSOR DATA --   -- AVERAGE PROCESSOR UTILIZATION PERCENTAGES --

                  ----MSU----  -CAPPING--  PROCESSOR-  ----DISPATCH TIME DATA----   LOGICAL PROCESSORS  --- PHYSICAL PROCESSORS ---
NAME       S   WGT  DEF    ACT  DEF   WLM%  NUM   TYPE   EFFECTIVE       TOTAL      EFFECTIVE    TOTAL  LPAR MGMT  EFFECTIVE  TOTAL

NP1        A    20  100     10  NO    62.2  1.2   CP    00.04.29.502  00.04.27.519      25.10    24.92       ****       3.33   3.30
NP2        A     1    0      1  YES    0.0    4   CP    00.00.22.680  00.00.22.083       0.75     0.73       ****       0.28   0.27
NP3        A    10    5      8  NO     3.3  1.0   CP    00.03.37.761  00.03.35.859      24.20    23.99       ****       2.69   2.67
NP4        A   300   95    155  NO     0.0  0.3   CP    01.12.08.405  01.12.06.405      80.18    80.15       ****      53.46  53.43
NP5        A   200   50     52  NO     0.0    4   CP    00.24.13.447  00.24.11.311      40.39    40.33       ****      17.95  17.92
CFC1       A   DED    0     32         0.0    1   CP    00.14.59.611  00.14.59.625      99.99    99.99       0.00      11.11  11.11
CFC2       A   DED    0      0         0.0    1   CP    00.00.00.000  00.00.00.000       0.00     0.00       0.00       0.00   0.00
*PHYSICAL*                                                             00.00.03.603                          0.04                 0.04
                                                                      ------------  ------------                ------     ------ ------
TOTAL                                                                 01.59.51.408  01.59.46.408                0.04      88.81  88.75

CB88       D
CB89       D
```

Figure 104. Partition Data Report

Report Analysis

- PROCESSOR NUM

  The number of online processors for partition NP1 is 1.2, this reflects the online time percentages of 19.33% + 100% = 119.33% in the CPU report.

- LOGICAL PROCESSORS - TOTAL DISPATCH TIME

  ```
  DISPATCH TIME                   4.27.519 =  268 sec
  ONLINE TIME         119.33% of 14.59.678 = 1074 sec
  LP TOTAL UTIL             (268/1074)*100 = 24.9%
  ```

  The value of 24.9% is shown as LPAR BUSY TIME PERC in the CPU report.

Report Analysis

- PHYSICAL PROCESSORS - TOTAL DISPATCH TIME

```
DISPATCH TIME                  4.27.519 =  268 sec
INTERVAL TIME (9 CPSs)     9 * 14.59.678 = 8097 sec
LP TOTAL UTIL             (268/8097)*100 =  3.3%
```

  Partition NP1 has a processor utilization of 3.3% of the total 2064-109 system.

- PHYSICAL PROCESSORS - LPAR MGMT

  Each partition's CPU consumption for LPAR management is calculated as the difference between total and effective dispatch time.

  It is possible that the total dispatch time is smaller than the effective dispatch time. This situation occurs when partitions get ″overruns″ in their dispatch intervals caused by machine delays. The most typical form of this is caused by an MVS partition trying to talk to a coupling facility but getting significant delays or time-outs. It is sometimes symptomatic of recovery problems on the machine. In this case, field LPAR MGMT is filled with '****'.

- *PHYSICAL*

  The *Physical Management Time* is collected and reported by RMF in this line. The partition named *PHYSICAL* does not exist, this line is created for reporting purposes.

Report Analysis

- LOGICAL PROCESSORS - UTILIZATION CF PARTITION

  PARTITION CFC1 99.99% UTILIZATION

  CF partitions are actually always busy 100% of the time. For purposes of reporting CF utilization, the CPU time is accumulated in two buckets: *busy* and *idle*. *Idle* is the CPU time spent looking for work (the polling loop), *busy* is all other time (time spent processing a command or background work). From these times, the CF Utilization reported in the Coupling Facility Activity report is 100*busy/(busy+idle).

Defining Logical Processors

Obviously, there is a cost to operate in LPAR Mode. One of the key aspects for defining an LPAR configuration is the number of logical partitions.

Defining more logical partitions than are required to handle a partition's workload affects LPAR management time. Define as few LPARs as are needed to support your business needs.

Rule-of-Thumb
Number of Logical Processors

Number of Logical Processors ≤ 2 * Number of Physical Processors

The following example demonstrates the impact of the definition of logical processors on the total utilization of a system. In this sample, the processor is an 8-way system.

For a system with 8 physical processors, 9 partitions have been defined with a total of 70 logical processors which is far away from the rule-of-thumb that has been mentioned above.

LPAR #	LPAR Name	# LP	Weights
1	PROD1	8	80
2	PROD2	8	50
3	PROD3	8	40
4	PROD4	6	12
5	PROD5	8	50
6	TEST	8	7
7	PROD6	8	106
8	PROD7	8	40
9	PROD8	8	5

With this configuration, the LPAR management time percentage was 22.83% which is nearly the capacity of 2 processors in this 8-way system. Due to significant performance problems, the configuration has been changed with defining 20 LPs for the same number of partitions, which is close to the 2:1 rule. The weighting factor was taken to estimate the number of LPs for each partition.

LPAR #	LPAR Name	# LP	Weights
1	PROD1	4	80
2	PROD2	2	50
3	PROD3	1	40
4	PROD4	2	12
5	PROD5	3	50
6	TEST	1	7
7	PROD6	4	126
8	PROD7	2	40
9	PROD8	1	5

By changing the number of LPs from 70 to 20, the LPAR management time percentage could be reduced by 20% of the total capacity (from 22.83% to 2.87%) for an 8-way processor which is equivalent to the capacity of 1.6 processors.

Intelligent Resource Director

Considerations about the number of logical processors are relevant and required only for partitions which are not under control of WLM LPAR management. This is a function of the Intelligent Resource Director which is available on z900 servers, and it is described in the following chapter.

Appendix B. The Intelligent Resource Director

Learning about the Intelligent Resource Director

This appendix describes the functions of the Intelligent Resource Director and its reporting in RMF, including:

- Dynamic Channel Path Management
- Channel Subsystem Priority Queuing
- LPAR CPU Management

In addition, the reporting of the IBM License Manager is described.

Overview

The Intelligent Resource Director (IRD) is a functional enhancement exclusive to IBM's zSeries family. It is provided in z900 servers and is an extension of IBM's industry leading clustering architecture, the Parallel Sysplex. IRD uses LPAR clusters (see "LPAR Cluster" on page 189 for a description) that will efficiently balance processor and I/O resources between multiple applications based on quality of service goals defined by the customer. These enhancements will ensure that the unpredictable needs of new e-transaction processing workloads can be managed dynamically according to business requirements.

The current implementation addresses three separate but mutually supportive functions:

- Dynamic Channel Path Management (DCM)

  This feature enables customers to have channel paths that dynamically and automatically move to those I/O devices that have a need for additional bandwidth. The benefits are enhanced by the use of goal mode and clustered LPARS.
- Channel Subsystem Priority Queuing

  Channel subsystem priority queuing on the z900 allows the priority queuing of I/O requests within the channel subsystem and the specification of relative priority among LPARs.
- LPAR CPU Management

  WLM dynamically adjusts the number of logical processors online to an LPAR and the processor weight based on the WLM policy. The ability to move the CPU resources across an LPAR cluster provides processing power to where it is most needed based on WLM goal mode policy.

Dynamic Channel Path Management

Introduction to Dynamic Channel Path Management

There is no typical workload. The requirements for processor capacity, I/O capacity and other resources vary throughout the day, week, month and year. Traditional methods of monitoring and tuning result in configurations that attempt to address the majority of your workload characteristics, leaving resources under or over utilized at various times. Even a well tuned DASD subsystem can quickly become inappropriately configured as data is migrated to and from different subsystems. A better approach would involve tuning decisions being made more responsively, using the capabilities of the operating system to self-monitor and self-adjust.

DCM provides the ability to have the system automatically manage the number of paths available to disk subsystems. By making additional paths available where they are needed, this increases the effectiveness of your installed channels, and potentially reduces the number of channels required to deliver a given level of service. DCM also provides availability benefits by attempting to ensure that the paths it adds to a control unit have as few points of failure in common with existing paths as possible, and configuration management benefits by removing unused paths from over-configured control units. Where paths can be shared by Multiple Image Facility (MIF), DCM will co-ordinate its activities within an LPAR cluster.

Where several channels are attached from a central processor complex (CPC) to a switch, the channels can be considered as a resource pool for accessing any of the control units attached to the same switch. To achieve this without DCM would require deactivating paths, performing a dynamic I/O reconfiguration and activating new paths. DCM achieves the equivalent process automatically, using those same mechanisms.

Channels managed by DCM will be referred to here as *managed* channels.

With HCD, it has to be defined how many channel paths of a control unit are to be managed. At least one channel path must be defined as static.

Value of Dynamic Channel Path Management

- **Improved system performance**

  Improved system performance is achieved by automatic path balancing and service policy. Reassignment is done in real-time or on WLM adjustment interval. A customer may reassign once a year, WLM does it every few seconds.
- **More effective utilization of installed hardware**

  Channels will be automatically balanced, providing opportunities to use fewer I/O paths to service the same workload.
- **Simplified I/O definition**

  The connection between managed channels and control units does not have to be explicitly defined, but at least one static path is required for each control unit.
- **Reduced skills required to manage z/OS**

  Managed channels and control units are automatically monitored, balanced, tuned and reconfigured.

Reporting of Dynamic Channel Path Management

All RMF reports about channel activities and I/O queuing activities provide information about DCM. The following examples show Monitor III reports.

Channel Activity Report

```
                    RMF V1R5  Channel Path Activity           Line 1 of 12
Command ===>                                                Scroll ===> HALF

Samples: 100     System: SYS1  Date: 07/22/04  Time: 15.55.00  Range: 120   Sec

 Channel Path        Utilization(%)     Read(MB/Sec)  Write(MB/Sec)
ID No   Type SHR   Part  Total    Bus    Part  Total    Part  Total

    8  *FC_SM     21.76  38.97  40.20   10.23  23.87   11.02  18.39
   12  *FCV_M     13.69  26.86  33.39    8.98  16.43    6.55  12.01
    1  *CNC_M     17.23  34.45
80      FC     Y  22.00  43.90  44.01   11.34  24.40   10.66  19.50
81      FC_SM  Y  21.76  38.97  40.20   10.23  23.87   11.02  18.39
82      FCV    Y  14.70  28.50  31.17    8.22  17.34    6.48  11.16
83      FCV_M  Y  13.69  26.86  33.39    8.98  16.43    6.55  12.01
85      CTC_S  Y   0.10   3.95
89      CNC_S      0.75   5.82
8C      CNC    Y   1.79  13.50
8D      CNC_S  Y   1.87  13.78
DA      CNC_M  Y  17.23  34.45
```

Figure 105. CHANNEL Report

For all channels that are managed by DCM, additional information is available. These channels are not assigned to a specific control unit permanently, but belong

to a pool of channels. Based on workload requirements in the system, these channels will be assigned dynamically by DCM. On top of the report, there is a consolidated data section for managed channel paths displaying the total number of channel paths for each type and the average activity data. The character **M** as suffix of the acronym for the channel path type is an indicator that the channel is managed by DCM.

I/O Queuing Activity Report

```
                          RMF V1R5  I/O Queuing Activity        Line 1 of 54
Command ===>                                                  Scroll ===> HALF

Samples:    100 System: MVS1  Date: 07/28/04  Time: 08.53.20     Range:  100 Sec

                       DCM Group        Cont  Del Q  CHPID   %DP    %CU
Path DCM  CTL Units  MIN MAX DEF LCU   Rate  Lngth  Taken   Busy   Busy

DA        400A                   0150                0.89   0.00   0.00
BA    Y   400A                   0150                0.68  19.00   0.00
E3        400B                   0150                0.86   0.00   0.00
E5    Y   400B                   0150                0.78  18.26   0.00
                       2   2   4  0150                1.46  18.63   0.00
                                  0150   0.0   0.00   3.21   9.41   0.00
DA        440A                   0151                0.01   0.00   0.00
BA        440A                   0151                0.02  33.33   0.00
E3        440B                   0151                0.02   0.00   0.00
E5        440B                   0151                0.03  25.00   0.00
                                  0151   0.0   0.00   0.07  20.00   0.00
```

Figure 106. IOQUEUE Report

The values in columns MIN MAX report the minimum and maximum number of DCM managed channels for one LCU (in this interval). DEF is the maximum number of managed channels for each LCU as it has been defined with HCD.

The line with these values is available only for LCUs with DCM managed channels. It also contains the I/O activity rate, director port contention, and control unit contention of all DCM managed channels. These values may include also measurements of managed channels which were partially online.

Channel Subsystem Priority Queuing

Introduction to Channel Subsystem Priority Queuing

Channel subsystem (CSS) priority queuing is a new function available on z900 processors either in basic or LPAR mode that allows the operating system to specify a priority value when starting an I/O request. When there is contention causing queuing in the channel subsystem, the request will be prioritized by this value.

z/OS will set the priority based on a goal mode WLM policy. This will complement the goal mode priority management that sets I/O priority for IOS UCB queues, and for queuing in the 2105 ESS DASD subsystem.

CSS priority queuing uses different priorities calculated in a different way than the I/O priorities used for UCB and control unit queuing.

Value of Channel Subsystem Priority Queuing

- **Improved performance**

  I/O from work that is not meeting its goals may be prioritized ahead of I/O from work that is meeting its goals, providing the Workload Manager with an additional method of adjusting I/O performance. Channel subsystem priority queuing is complementary to UCB priority queuing and control unit priority queuing, each addressing a different mechanism that may affect I/O performance.
- **More effective use of providing I/O bandwidth**

  Because interference from other I/O is reduced, the total I/O bandwidth required to meet the goals of high importance work may possibly be reduced. Previously, bandwidth had to be acquired to service the whole workload in order to meet the performance requirements of important work.
- **Reduced skills required to manage z/OS**

  Monitoring and tuning requirements are reduced because of the self tuning abilities of the channel subsystem.

Reporting of Channel Subsystem Priority Queuing

There is no specific reporting of this function in RMF, but you may see a reduced number of I/O delay samples in the Workload Activity report.

LPAR CPU Management

Introduction to LPAR CPU Management

LPAR CPU management allows WLM working in goal mode to manage the processor weighting and number of logical processors across an LPAR cluster. CPU resources are automatically moved toward LPARs with the most need, by adjusting the partition's weight. The sum of the weights for the participants of an LPAR cluster is viewed as a pooled resource which can be apportioned among the participants to meet the goal mode policy. The installation can place limits on the processor weight range.

WLM will also manage the available processors by varying unneeded CPs offline (more logical CPs implies more parallelism, and less weight per CP). LPAR overheads will be decreased by reducing the number of logical processors.

LPAR Cluster

An LPAR cluster is a group of logical partitions that are resident on the same physical server and in the same sysplex.

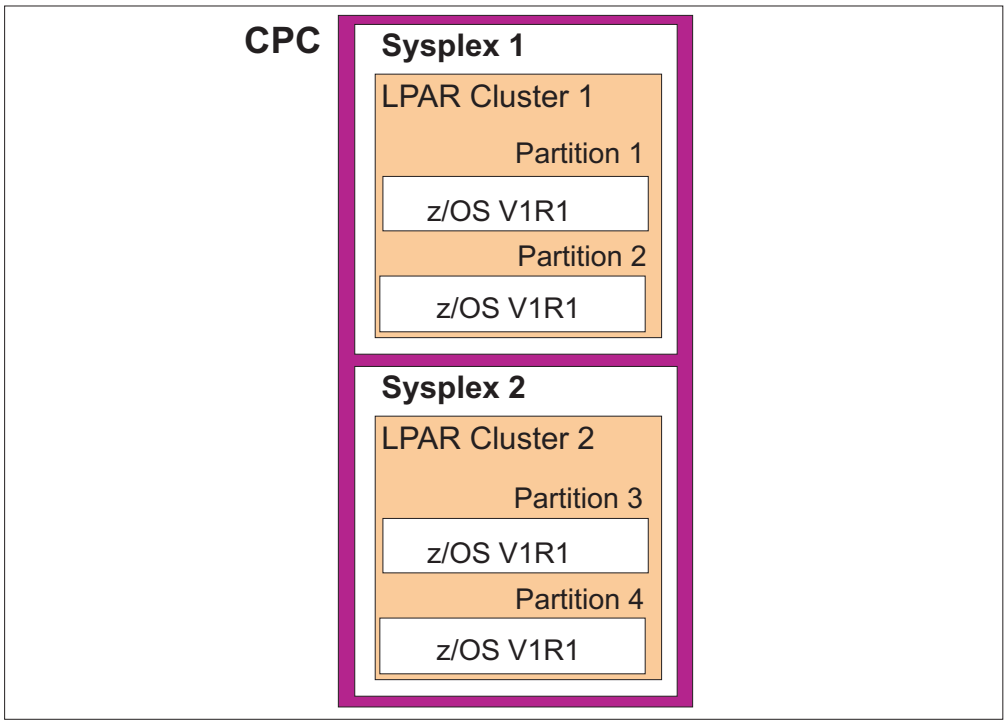


Figure 107. LPAR Cluster - Example 1

The next figure is a more complex configuration, with four LPAR clusters grouped into two sysplexes across two CPCs:

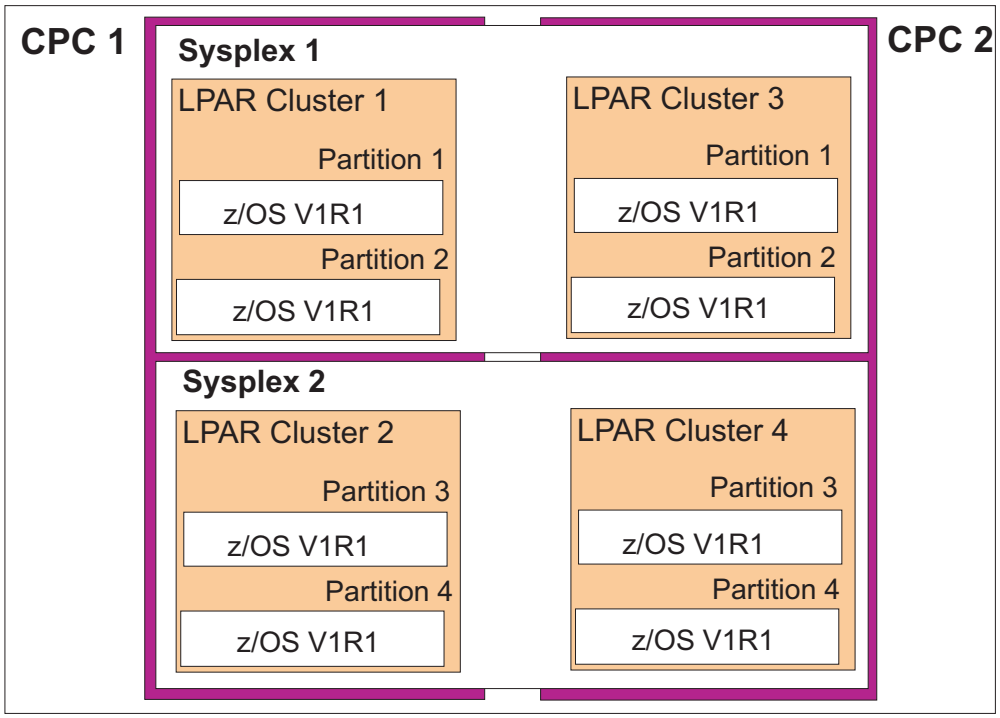


Figure 108. LPAR Cluster - Example 2

Value of LPAR CPU Management

- **Logical CPs are performing at the fastest uniprocessor speed available.**

  This results in the number of logical CPs tuned to the number of physical CPs of service being delivered by the LPARs current weight. If the LPAR is getting 4 equivalent physical CPs of service and has 8 logical CPs online to z/OS, then each logical CP only gets half of an equivalent physical CP. For example, if a CP delivers 200 MIPS, half of it will deliver 100 MIPS. This occurs because each logical CP gets fewer time slices.
- **Reduces LPAR overhead.**

  There is an LPAR overhead for managing a logical CP. The higher the number of logical CPs in relation to the number of physical CPs, the higher the LPAR overhead. This is because LPAR has to do more processing to manage the number of logical CPs which exceeds the number of physical CPs.
- **Gives z/OS more control over how CP resources are distributed.**

  z/OS is able to change the assigned CP resources (LPAR weights) and place them where they are required for the work. Making both of these adjustments is simple from an operator perspective, but what is difficult is identifying when the changes are required and whether the changes have had a positive effect.

LPAR CPU management

- identifies what changes are needed and when
- projects the likely results on both the work it is trying to help and the work that it will be taking the resources from
- performs the changes
- analyzes the results to ensure the changes have been effective.

LPAR CPU Management Decision Controls

There are two parts of LPAR CPU management:

- WLM LPAR Weight Management
- WLM LPAR Vary CPU Management

WLM LPAR Weight Management

WLM LPAR weight management involves extending the receiver/donor logic for CPU delays to CP resource being used by other LPARs in the LPAR cluster. Until now, WLM could only move CP resource from one service class period to another on the same system by adjusting the dispatching priority of the receiver and/or the donor. When using WLM LPAR weight management, WLM can now alter the weights of LPARs in the LPAR cluster, effectively moving CP resource from one LPAR to another.

In the past, an operator was required to change LPAR current weights manually. However, an operator would first have to determine that there was a CPU delay and then choose which other LPAR the weight units should be taken from. This approach was not practical.

WLM LPAR Vary CPU Management

WLM LPAR vary CPU management does not look at service class periods when making adjustments. It ensures the number of online logical CPs is approximately equal to the number of physical CPs required to do the work. It is more like a tuning action for the CP resources available to the system. Prior to LPAR CPU management, there was no function in WLM to tune the available CP resources in relation to the physical CP resources being used by the system. Once again if this

function is to be performed manually, an operator first needs to identify this is a problem before adjusting the number of logical CPs.

LPAR CPU Management Controls

The IBM z900 CPC and CPU management introduce new terms for LPAR weights. The following are used when discussing LPAR weights.

- Initial processing weight

  This weight becomes the LPAR's actual weight immediately following an IPL.
- Minimum and maximum processing weight

  This is the minimum and maximum weight that WLM LPAR weight management will assign as an LPAR's actual weight.
- Actual weight

  An LPAR's weight at a point in time when the systems is running. This is the weight that WLM LPAR weight management and WLM LPAR vary CPU management use when performing their primary functions.
- Enable work load management must be checked to enable LPAR CPU management.

  Recommended usage is to use large (2 or 3 digits) values for CPU weights to allow WLM better granuality to work within.

Reporting of LPAR CPU Management

The Postprocessor CPU Activity report has been extended by a new section called LPAR Cluster report.

LPAR Cluster Report

```
                                L P A R   C L U S T E R   R E P O R T
                                                                                                  PAGE    3
          z/OS V1R6              SYSTEM ID NP1             DATE 04/15/2004         INTERVAL 14.59.678
                                 RPT VERSION V1R5 RMF      TIME 09.30.00           CYCLE 1.000 SECONDS

                              ------ WEIGHTING STATISTICS ------   ---- PROCESSOR STATISTICS ----   ------ STORAGE STATISTICS -----

                              --- DEFINED ---  ---- ACTUAL -----   ---- NUMBER ---   -- TOTAL% --   --- CENTRAL ---  -- EXPANDED --
CLUSTER   PARTITION  SYSTEM    INIT  MIN  MAX  AVG  MIN %  MAX %   DEFINED  ACTUAL   LBUSY  PBUSY
SVPLEX1   NP1        NP1         20   20   40   20    100    0.0         4     1.2   24.92   3.30         1024           N/A
          NP2        NP2          1              1                      4       4    0.73   0.27         1024           N/A
          NP3        NP3         10   10   50   10    100    0.0         4     1.0   23.99   2.67         1024           N/A
-------------------------  ------------------------------------  ------------------------------  ------------------------------
                     TOTAL      31                                      12           49.64   6.24         3072           N/A

SVPLEX2   NP4        NP4        300  100  500  300    0.0    0.0         6     4.2   80.15  53.43         1024           N/A
          NP5        NP5        200            200                      4       4   40.33  17.92         1024           N/A
-------------------------  ------------------------------------  ------------------------------  ------------------------------
                     TOTAL     500                                      10           120.5  71.36         2048           N/A
```

Figure 109. LPAR Cluster Report

This new report, as well as the Partition Data report (also part of the CPU Activity report), provides information about LPAR CPU management.

Report Analysis

- `CLUSTER PARTITION`

  The three partitions NP1, NP2, and NP3 belong to the LPAR cluster SVPLEX1, but only NP1 and NP3 are under control of LPAR CPU management.

- `DEFINED INIT MIN MAX`

  NP1 has an initial weight of 20 which also is the defined minimum, and it has a defined maximum weight of 40.

- `ACTUAL AVG`

  The average actual weight in the interval for NP1 is 20.

- `ACTUAL MIN % MAX %`

  These values provide the percentage of the interval when the partition's weight was within 10% of the defined minimum or maximum weight. NP1 was within 10% of the specified minimum (20) for the entire interval (MIN%=100).

- `NUMBER DEFINED ACTUAL`

  NP1 has been defined with 4 logical processors. In the current interval, WLM LPAR CPU management has assigned 1.2 processors to this partition in average. This value can be seen also in the Partition Data report (PROCESSOR NUM) and reflects the total ONLINE TIME PERCENTAGE in the CPU Activity report.

- `TOTAL% LBUSY PBUSY`

  These values give the utilization of the logical processors assigned to the partition (based on its online time) and of all physical processors of the CPC (based on the interval time). They are also shown in the Partition Data report.

- `STORAGE STATISTICS`

  The amount of central and expanded storage assigned to each partition is given in these columns.

Partition Data Report

```
                                    P A R T I T I O N   D A T A   R E P O R T
                                                                                                    PAGE    2
            z/OS V1R6              SYSTEM ID NP1             DATE 04/15/2004         INTERVAL 14.59.678
                                   RPT VERSION V1R5 RMF      TIME 09.30.00           CYCLE 1.000 SECONDS

MVS PARTITION NAME                      NP1                    NUMBER OF PHYSICAL PROCESSORS           9
IMAGE CAPACITY                          100                                    CP                       9
NUMBER OF CONFIGURED PARTITIONS           9                                    ICF                      0
WAIT COMPLETION                          NO
DISPATCH INTERVAL                   DYNAMIC

--------- PARTITION DATA -----------------  -- LOGICAL PARTITION PROCESSOR DATA --   -- AVERAGE PROCESSOR UTILIZATION PERCENTAGES --

                  ----MSU----  -CAPPING--  PROCESSOR-  ----DISPATCH TIME DATA----   LOGICAL PROCESSORS  --- PHYSICAL PROCESSORS ---
NAME       S   WGT  DEF    ACT  DEF   WLM%  NUM   TYPE   EFFECTIVE       TOTAL      EFFECTIVE    TOTAL  LPAR MGMT  EFFECTIVE  TOTAL

NP1        A    20  100     10  NO    62.2  1.2   CP    00.04.29.502  00.04.27.519       25.10    24.92      ****       3.33   3.30
NP2        A     1    0      1  YES    0.0    4   CP    00.00.22.680  00.00.22.083        0.75     0.73      ****       0.28   0.27
NP3        A    10    5      8  NO     3.3  1.0   CP    00.03.37.761  00.03.35.859       24.20    23.99      ****       2.69   2.67
NP4        A   300   95    155  NO     0.0  0.3   CP    01.12.08.405  01.12.06.405       80.18    80.15      ****      53.46  53.43
NP5        A   200   50     52  NO     0.0    4   CP    00.24.13.447  00.24.11.311       40.39    40.33      ****      17.95  17.92
CFC1       A   DED    0     32         0.0    1   CP    00.14.59.611  00.14.59.625       99.99    99.99      0.00      11.11  11.11
CFC2       A   DED    0      0         0.0    1   CP    00.00.00.000  00.00.00.000        0.00     0.00      0.00       0.00   0.00
*PHYSICAL*                                                            00.00.03.603                           0.04              0.04
                                                                      ------------  ------------                          ------ ------
TOTAL                                                                 01.59.51.408  01.59.46.408                0.04      88.81  88.75

CB88       D
CB89       D
```

Figure 110. Partition Data Report

The Partition Data report contains in the fields WGT and PROCESSOR NUM the actual values for weighting and number of logical processors, either as dynamically assigned by LPAR CPU management (for example, NP1 or NP3) or as statically defined by the customer (for example, NP2).

IBM License Manager

The IBM License Manager is the base for a new software pricing model. It allows vendors to enable their products for licensed software management by customers and is the basic tool IBM will use to implement the Workload License Charges pricing model on z900 servers.

When using the License Manager, the customer can specify a capacity limit as part of the definition of an LPAR partition.

Report Analysis

The following values in the report are related to the License Manager:

- IMAGE CAPACITY

  A new value IMAGE CAPACITY in this report gives information about the CPU capacity which is available to the partition. It is measured in MSUs (millions of CPU service units) per hour. There are these alternatives:

  – The partition's defined capacity set via the Hardware Management Console, if any. The report shows that NP1 has a defined capacity of 100 MSUs.

  – The capacity based on logical processors defined for the partition, if the partition is uncapped and has no defined capacity.

    Example:

    ```
    MSU of CPC             290
    # physical processors    9
    # logical processors     4
    Image capacity         (4/9) * 290 = 129
    ```

  – The capacity at the partition's weight, if the partition is capped via the Hardware Management Console.

- MSU DEF ACT

  For each partition, the defined and actual MSU values are given. For the partition which is gathering the data, here partition NP1, DEF MSU (100) is identical to IMAGE CAPACITY while the actual MSU consumption in the interval is 10 MSUs (calculated as MSUs per hour). A value of zero in field DEF indicates there is no defined capacity set for this partition.

  ACT MSU contains the actual consumption of service units in MSUs per hour. This value does not correlate to the value of service units given in the Workload Activity report.

  – Service Units in Workload Activity Report

    In the Service Policy page you find for each processor the value SU/SEC. This value is based on the number of logical processors assigned to the partition. If you have 4 logical processors in a z900 server, then this value relates to a 2064-104 (with 148 MSUs).

    ```
    1 LP    = 10289 SU/SEC
    4 LPs   = 41156 SU/SEC = 148 MSU/H
    ```

  – Service Units in Partition Data Report

    Service unit consumption in a License Manager context is based on the number of physical processors in the CPC. In a 9-way z900 server 2064-109 with 290 MSUs, this value is taken:

    ```
    1 CP    =  8964 SU/SEC
    9 CPs   = 80676 SU/SEC = 290 MSU/H
    ```

- CAPPING WLM%

  The concept of the License Manager is to allow a partition a limited resource consumption - the defined capacity limit. If the partition is using more service units than defined, WLM will cap this partition. This is not done directly on the current consumption but it is based on a long-term average. So peaks exceeding the defined limit are possible. The field CAPPING WLM% shows the percentage of the reporting interval when a partition has been capped by WLM.

Appendix C. Data-In-Memory

Exploiting Your Storage

This appendix concentrates on Data-In-Memory techniques. The key topics are:

- Benefits of these techniques
- Evaluation before the implementation
- Different functions or products that implement DIM

The chapter concludes with a roadmap on how to proceed with your applications.

Introduction

Data-In-Memory (DIM) techniques enable individual jobs (or a group of chained jobs) to run faster by reducing the I/O component of the elapsed time for the job. The reduction is achieved by reading the data from (or in some cases writing to) a buffer in processor storage instead of tape or DASD. There are several benefits of these techniques:

- I/Os are reduced, and therefore their component of elapsed time is reduced in proportion to the reduction in I/Os.
- The remaining I/Os are done to a less busy I/O subsystem, which gives the remaining I/Os a faster response.
- The CPU is no longer used for driving I/Os.
- The CPU is no longer used for pre-processing CLISTs or LOAD modules.
- Some CPU cycles are saved as a result of less contention and a shorter transaction residency time.

On the other hand, sufficient storage and CPU resources need to be available to sustain the new load:

- Data is moved to central storage: if there is not enough storage, the paging activity may rise and become a new bottleneck.
- More work will be available sooner: the processor must be able to execute this new load without stopping the old.
- There is some processor overhead associated with the movement of pages between central and expanded storage.

The net of the CPU savings and the CPU overhead is generally fairly small, usually less than 5%. However for most transactions, waiting for I/O is the largest single component of the response time: any significant reduction in I/O rates can make a large difference in the response time.

To evaluate the benefits of the DIM techniques implementation, the following steps must be performed:

1. Understand the current processor storage situation, assessing the current storage requirements.
2. Perform a study of DIM exploiters, assessing the additional storage requirements for DIM exploitation.
3. Reduce DASD response time, assessing the degree to which remaining I/Os can be expedited and determining the new steps to do this.
4. Implement all the changes.
5. Monitor the results.

In steps 1 and 5, RMF is a good starting tool. Some other tools can help with the assessment of benefits; refer to the documentation related to the different functions or products mentioned below.

Current Implementations

The following sections briefly describe the different functions or products that implement the DIM techniques, in a batch and in an OLTP environment.

Batch LSR Subsystem

The *Batch LSR Subsystem (BLSR)* extends the benefits of VSAM *Local Shared Resources (LSR)* to applications written to use VSAM *Non-Shared Resources (NSR)*, typically programs written in high-level languages. VSAM LSR allows a program processing a VSAM data set to buffer frequently-used data in processor storage, eliminating read I/Os and reducing the elapsed time of an individual job. VSAM NSR does not provide this kind of buffering.

BLSR is implemented as an MVS address space (in particular, a subsystem). Specified VSAM NSR data set OPENs are routed to the BLSR subsystem, which converts them to VSAM LSR OPENs. It also allocates the buffer pools specified either as portions of its own virtual storage or as hiperspaces that it owns. Hence, there is no effect on the job's virtual storage usage.

Eligibility

The Batch LSR Subsystem supports the following VSAM data set organizations and alternate indexes (some restrictions may apply):

- Key Sequenced (KSDS)
- Relative Record (RRDS)
- Entry Sequenced (ESDS)

DFSORT™ Hipersorting

Many large sorts (particularly the larger ones in the batch window) require work space to hold intermediate data. Typically this is held on DASD (although tape and VIO are also supported). As much as 20% of a large sort's elapsed time might be spent performing I/O to work data sets. DFSORT hipersorting allows part or all of the sort work space to be allocated in a hiperspace residing in expanded storage. This approach leads to a reduction or elimination of sort work I/O and a consequent reduction in elapsed time for the sort.

Hipersorting uses the standard type of hiperspace so there may be an increase in paging and page data set usage because of overcommittal of expanded storage.

Eligibility

Any sort performing I/O to work data sets is eligible for hipersorting. The amount of expanded storage available when the sort begins determines the degree to which hipersorting takes place. Storage availability is measured in terms of the system impact of giving expanded storage to DFSORT rather than the number of free expanded storage frames.

DFSORT does not use hipersorting for sorts using tape work space. These sorts should use DASD work space (or all hiperspace) as tape is not recommended for sort work space.

Implementation

Hipersorting can be enabled at the system level as a default. To do this, code the HIPRMAX keyword on the ICEMAC macro as a system-level option. It can also be enabled at sort invocation by coding the HIPRMAX run-time option.

Hiperbatch

Frequently, some data sets are read sequentially several times during a batch window. These data sets may be written during the batch window and then subsequently read. Typically, installations have found that these multiply-accessed data sets produce performance bottlenecks because of DASD contention. Traditional methods of reducing this effect have been to:

- Duplicate the data set over several DASD volumes or even controllers.
- Schedule each of the jobs and jobsteps so that they access the data at different times.
- Copy the data set (if it is small) to a VIO data set (using paging I/O instead of traditional I/O).

Hiperbatch™ provides a set of services than can provide greater benefits than these techniques and without many of their associated problems. It uses the Data Lookaside Facility (DLF) component of MVS to keep copies of portions of sequentially processed data sets in hiperspaces in expanded storage.

Eligibility

Hiperbatch supports access methods VSAM Non-Shared Resources and QSAM.

A VSAM data set (KSDS, ESDS, or RRDS) is supported if:
- The data set is not accessed as VSAM LSR or GSR.
- The data set is not accessed with shareoptions 3 or 4.
- The control interval size is 4 KB or a multiple thereof.
- The data set is not a catalog.
- The data set is not accessed with CI processing.

A data set processed by QSAM is supported if:
- It resides on DASD.
- It is a physical sequential data set.
- It is not a partitioned data set or a member thereof.

Implementation

Implementing hiperbatch consists of two activities:
1. Finding the right data sets to buffer
2. Turning on hiperbatch for these candidates

Having implemented hiperbatch, its effectiveness and use of resources has to be monitored and tuned (particularly in its usage of expanded storage).

VIO to Expanded Storage

VIO to expanded storage allows temporary data sets to be buffered in expanded storage for the life of the job. This technique eliminates all the I/O to the data set and can reduce the elapsed time of a job significantly. It is an extension to traditional VIO, which uses page data sets to hold temporary data sets. With DFSMS, implementation can be such that the application developer is unaware that the temporary data sets are not on DASD but are using VIO to expanded storage. In this and many other ways, DFSMS enables easier and more effective implementation and control of VIO to expanded storage.

Eligibility

All access methods are eligible for VIO with the exception of VSAM, ISAM, and PDSE.

Implementation

Implementation should be performed cautiously as VIO to expanded storage can consume large amounts of storage if not properly managed. It consists of two main elements:

- Enabling VIO for the system.
- Making individual data sets eligible.

VSAM Large Buffer Pools

VSAM Local Shared Resources (VSAM LSR) allows users accessing VSAM data sets to be buffered using pools of buffers in processor storage. This approach avoids the repeated reading of the same data from DASD and hence reduces the I/O component of elapsed time.

Virtual Lookaside Facility

The *Virtual Lookaside Facility (VLF)* is a component of z/OS MVS designed to improve performance by retrieving frequently-used objects from virtual storage rather than performing I/O operations from DASD.

VLF can save I/Os in several ways; for example:
- Keeping modules in virtual storage.
- Keeping the directories of frequently accessed PDSs in virtual storage.
- Keeping processed CLISTS in virtual storage.

More information on VLF can be found in the *z/OS MVS Initialization and Tuning Guide*.

DB2 Buffer Pool

DB2 Version 3 extends the bufferspace to hiperspace, providing a new facility to keep a large amount of DB2 data in the expanded storage.

This ability to allocate the majority of DB2's bufferpool requirements in expanded storage means that more central storage may now be made available to other subsystems while maintaining the same bufferpool requirements in DB2.

Furthermore, in a hiperpool configuration, DB2 itself manages the page movement to and from expanded storage using the hardware feature *Asynchronous Data Mover Facility (ADMF)*, resulting in the elimination of the MVS-managed paging.

Configurations currently constrained by high MVS paging rates may find hiperpools provide an effective means of reducing MVS paging activity in the system. MVS paging avoidance or elimination in DB2 may translate into system wide response time improvements and throughput improvements from MVS paging reductions in other subsystems as well.

CICS Shared Data Tables

A shared data table is an in-storage copy of some or all of the records from a data set. CICS automatically stores the records in the data table when the corresponding file is opened. Data table records are stored in a z/OS MVS data space.

In actual measurements, the use of CICS shared data tables has shown improvements of over 50% in response time, with up to 95% improvement in the file access times alone. CPU times are also reduced by eliminating the need to process interrupts, and manage the VSAM interface.

Virtual Fetch

Virtual fetch (VF) is an MVS function that sends storage pages to expanded storage when the migration age is greater than the criteria age. This is specified as an IEAOPTxx parameter: the ESCTVF parameter specifies the criteria age at which a virtual page will be sent to expanded storage.

Methodology Roadmap

The following roadmap provides an example on how to start a DIM study in a batch environment.

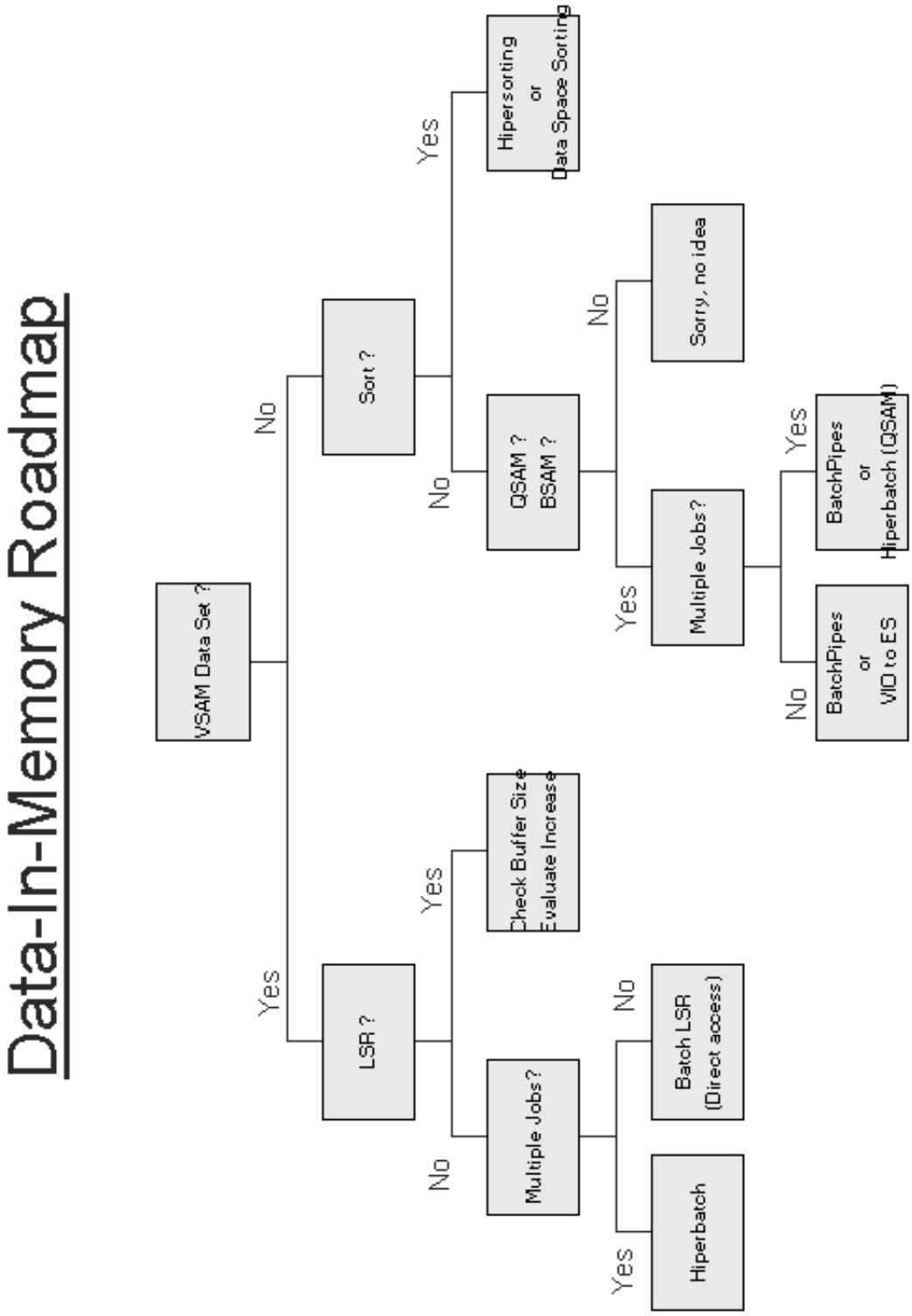


Figure 111. Data-In-Memory Roadmap

Appendix D. Other Delays

Some Other Delays

This last appendix discusses the delays shown by Monitor III that have not been covered yet:

- Enqueue delays
- HSM delays
- JES delays
- OPER delays
- Unknown delays

Enqueue Delays

Here is an approach to investigating ENQ delays using Monitor III.

ENQ Report

```
                          RMF V1R5  ENQ Delays                    Line 1 of 2
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

         DLY  -------------- Resource Waiting --------------  ---- Holding ----
Jobname   %    %  STAT Major/Minor Names (Scope)               %  Name/SYS STAT

BHOLEQB   49   49  EW  SYSDSN          (SYS)                   49 BTEUPRT
                       SYS3.BB.DATA
```

Figure 112. Monitor III Enqueue Delays Report

Look at the ENQ report for the user or user group.

Use the cursor to select the resource name with most delay and press ENTER to get the ENQR report:

- Use this report to view all the contentions.
- What is the major name contributing to most of the delay?
- Find the largest "holding" user for this resource, then go through the dialog with this jobname.

The HELP function (PF1) on the ENQJ report provides description of some different MAJOR/MINOR names. Also check the *Debugging Handbook - Volume 1* for a description of the MAJOR/MINOR names.

If the name does not exist there, it may be a local name. Knowing the resource or the application provides a greater understanding of this problem.

Major Names

SYSDSN Enqueue, look for the following:

- Submitted job needs a data set that the TSO user has
- Jobs actually have to run sequentially

SYSIGGV2 / SYSCTLG Enqueue, look for the following:

- Large LISTCATs running concurrently with a DELETE or HDELETE
- Massive deletes occurring from another address space
- Allocating and deleting but not using the ISPF log data set

Try to find out what the largest "holder" of SYSIGGV2 was doing at the time of the problem. If this is not possible, and the user is a TSO user, he may have been doing massive deletes or a LISTCAT.

If the problem occurred while exiting from ISPF, the delete of the Log data set could have caused the catalog access. If so, consider making the primary and secondary allocation of the log data set zero. This causes the log data set never to be allocated when entering ISPF.

SYSVTOC Enqueue, look for the following:

Try to determine who is allocating space and why it is taking so long. Possibly a job or user is frequently acquiring and freeing the space. If this is the case, acquire it once and do not free it. If the space is for a temporary data set, try using VIO instead.

SYSZRACF Enqueue

SYSZRACF is enqueued exclusively when updating the RACF data set. Find out why it is taking so long and reduce the time.

SYSIEFSD Enqueue

SYSIEFSD has several minor names: CHNGDEVS, DDRTPUR, ALLOCTP, DDRDA, Q4, Q6, QlO, RPL, STCQUE, VARYDEV, and TSOQUE.

If the minor name is CHNGDEVS, look for a large DEVICE USING or DEVICE DELAY for the HOLDING address space. For example DEVICE DELAY may be near 100%. Obtain the dominant volid involved. If the time duration spans several intervals look at the volids for each interval within the problem time range. Browse the SYSLOG for mount or vary activity at or about this time. Determine if anything can be done to prevent this in the future.

HSM Delays

HSM Report

```
                          RMF V1R5  HSM Delays                    Line 1 of 1
Command ===>                                                  Scroll ===> HALF

Samples: 100     System: PRD1  Date: 04/07/04  Time: 10.32.00  Range: 100   Sec

          DLY  ---------------------- Main Delay Reason -----------------------
Jobname    %    %  F-Code Explanation

DLOSHIN    7    7    3   Dataset recall from auxiliary storage.
```

Figure 113. Monitor III HSM Delays Report

Here are some general problems to look for:

- HSM is down
- HSM is partially down (HSM device not responding)
- HSM backed up because:
  - Not enough primary space
  - Not enough level one space
  - HSM doing housekeeping

Display the HSM address space with the DELAYJ command. Is the HSM data normal? Ask yourself why, when HSM data is unusual. You can compare this data to times when HSM delays were considered acceptable. Typical things to look for are:

- PROC TCB + SRB at 0.0% (HSM may be dead)
- PROC DELAY very high (HSM may have a poor dispatching queue position)
- DEVICE DELAY or DEVICE USING excessive (see DEVR report and determine why)
- A non-typical volume is in heavy use (modules being fetched?)
- Excessive ENQ DELAY time because of CATALOG or VTOC contention

Do a HLIST PVOL TERMINAL for volumes under HSM or ISMF control. The AGE column under SPACE-MGMT is the number of days a data set on this volume must be inactive before it is eligible for the type of space management indicated under TYPE. Different volumes may have different ages. MIN AGE shows the inactive age of the data set that most recently migrated from the volume.

Pick the one of the following that seems to be most probable:

- Something is broken (HSM in loop, HSM in infinite wait, control unit down, HSM address space terminated, STOR LOCL is 100%).
- HSM address space is running slowly because of contention with others (high PROC DELAY. high STOR LOCL, high DEVICE DELAY, high ENQ DELAY).
- HSM is in house cleaning mode (HSM is reorganizing data sets to reclaim space. There will be a high amount of PROC TCB + SRB and DEVICE USING for the amount of work in progress).
- There is a shortage of primary or level-one space (for example migration age less than 7 days).

- There is excessive traffic (migration age greater than 7 but many recall requests outstanding).

JES Delays

Here are some general problems to look for:

- JES is down
- JES is partially down (JES device not responding)
- JES backed up

Check the SYSLOG for more information. Look for excessive amount of I/O in the JES address space.

OPER Delays

In RMF terms, the operator is another resource hence can appear as a delay.

The Monitor III Delay report has a field of `OPER%`. This means the job is delayed by a mount request or is waiting for a reply to a WTOR message.

Where practicable, batch jobs or applications should not issue WTORs. The use of WTORs is probably historical and the time spent in recoding may well be recouped by future run time improvements.

Unknown Delays

The Monitor III Delay report has a field of `UKN%`. RMF considers jobs that are not delayed for a monitored resource, or not in an idling state to be in an unknown state.

Examples of unknown state delays are:

- AS waiting for I/O other than DASD or tape
- Idle address spaces which use an unmonitored mechanism for determining when they are active. Most STCs show as unknown when they are idle.
- AS waiting for a request from another AS to be satisfied.

Appendix E. Accessibility

Accessibility features help a user who has a physical disability, such as restricted mobility or limited vision, to use software products successfully. The major accessibility features in z/OS™ enable users to:

- Use assistive technologies such as screen readers and screen magnifier software
- Operate specific or equivalent features using only the keyboard
- Customize display attributes such as color, contrast, and font size

Using assistive technologies

Assistive technology products, such as screen readers, function with the user interfaces found in z/OS. Consult the assistive technology documentation for specific information when using such products to access z/OS interfaces.

Keyboard navigation of the user interface

Users can access z/OS user interfaces using TSO/E or ISPF. Refer to *z/OS TSO/E Primer*, *z/OS TSO/E User's Guide*, and *z/OS ISPF User's Guide Volume I* for information about accessing TSO/E and ISPF interfaces. These guides describe how to use TSO/E and ISPF, including the use of keyboard shortcuts or function keys (PF keys). Each guide includes the default settings for the PF keys and explains how to modify their functions.

z/OS information

z/OS information is accessible using screen readers with the BookServer/Library Server versions of z/OS books in the Internet library at:

`http://www.ibm.com/servers/eserver/zseries/zos/bkserv/`

One exception is command syntax that is published in railroad track format; screen-readable copies of z/OS books with that syntax information are separately available in HTML zipped file form upon request to mhvrcfs@us.ibm.com.

Glossary

A

AS. address space

auxiliary storage (AUX). All addressable storage, other than main storage, that can be accessed by means of an I/O channel; for example storage on direct access devices.

B

balanced systems. To avoid bottlenecks, the system resources (CP, I/O, storage) need to be balanced.

basic mode. A central processor mode that does not use logical partitioning. Contrast with *logically partitioned (LPAR) mode*.

bottleneck. A system resource that is unable to process work at the rate it comes in, thus creating a queue.

BLSR. Batch LSR Subsystem. See also *LSR*

C

cache fast write. A storage control capability in which the data is written directly to cache without using nonvolatile storage. This 3990 Model 3 Storage Control extended function should be used for data of a temporary nature, or data which is readily recreated, such as the sort work files created by DFSORT. Contrast with *DASD fast write*.

cache hit. Finding a record already in the storage (cache) of the DASD control unit (3990).

capture ratio. The ratio of reported CPU time to total used CPU time.

captured storage. See shared page group.

central processor (CP). The part of the computer that contains the sequencing and processing facilities for instruction execution, initial program load, and other machine operations.

central processor complex (CPC). A physical collection of hardware that consists of central storage, one or more central processors, timers, and channels.

CFWHIT. Cache fast write and read hits

channel path. The channel path is the physical interface that connect control units and devices to the CPU.

CICS. Customer Information Control System

compatibility mode. The implicit state of an MVS system when no workload manager service policies are in effect. Contrast with *goal mode*.

contention. Two or more incompatible requests for the same resource. For example, contention occurs if a user requests a resource and specifies exclusive use, and another user requests the same resource, but specifies shared use.

CP. Central processor Synonymous with *CP (central processor)*.

criteria. Performance criteria set in the WFEX report options. You can set criteria for all report classes (PROC, SYSTEM, TSO, and so on).

CRR. Cache RMF Reporter

CS. Central storage

CPU speed. Measurement of how much work your CPU can do in a certain amount of time.

Customer Information Control System (CICS). An IBM licensed program that enables transactions entered at remote terminals to be processed concurrently by user-written application programs. It includes facilities for building, using, and maintaining data bases.

cycle. The time at the end of which one sample is taken. Varies between 50 ms and 9999 ms. See also *sample*

D

DASD fast write. An extended function of the 3990 Model 3 Storage Control in which data is written concurrently to cache and nonvolatile storage and automatically scheduled for destaging to DASD. Both copies are retained in the storage control until the data is completely written to the DASD, providing data integrity equivalent to writing directly to the DASD. Use of DASD fast write for system-managed data sets is controlled by storage class attributes to improve performance. Contrast with *cache fast write*.

data sample. See *sample*

DCM. See *Dynamic Channel Path Management*

DCME. Dynamic cache management extended

delay. The delay of an address space represents a job that needs one or more resources but that must wait because it is contending for the resource(s) with other users in the system.

DFWHIT. DASD fast write hit

DFWRETRY. DASD fast write retry

DIM. Data-In-Memory

direct access storage device (DASD). A device in which the access time is effectively independent of the location of the data.

DLY. Delay

DP. Dispatching priority

dynamic channel path management. Dynamic channel path management provides the capability to dynamically assign channels to control units in order to respond to peaks in demand for I/O channel bandwidth. This is possible by allowing you to define pools of so-called floating channels that are not related to a specific control unit. With the help of the Workload Manager, channels can float between control units to best service the work according to their goals and their importance.

E

EMIF. ESCON multiple image facility

enclave. An enclave is a group of associated dispatchable units. More specifically, an enclave is a group of SRB routines that are to be managed and reported on as an entity.

EPDM. Enterprise Performance Data Manager/MVS, new name of the product is Performance Manager for MVS

execution velocity. A measure of how fast work should run when ready, without being delayed for processor or storage access.

ES. Expanded storage

ESCON multiple image facility (EMIF). A facility that allows channels to be shared among PR/SM logical partitions in an ESCON environment.

ETR. External throughput rate, see also ITR

exception reporting. So that you don't have to watch your monitor all the time to find out if there is a performance problem, you can define performance criteria. RMF will then send a report when the criteria are not being met (an exception has occurred).

expanded storage (ES). (1) An extension of processor storage. (2) Optional high-speed storage that transfers 4KB pages to and from central storage.

G

generalized trace facility (GTF). An optional OS/VS service program that records significant system events, such as supervisor calls and start I/O operations, for the purpose of problem determination.

GO mode. In this mode, the screen is updated with the interval you specified in your session options. The terminal cannot be used for anything else when it is in GO mode. See also *mode*.

goal mode. The implicit mode of an MVS system that has active service policies and performance goals defined by the workload manager. Contrast with *compatibility mode*.

GTF. generalized trace facility

H

high-speed buffer (HSB). A cache or a set of logically partitioned blocks that provides significantly faster access to instructions and data than provided by central storage.

HS. hiperspace

HSB. High-speed buffer

I

IFA. Integrated Facility for Applications, see zSeries Application Assist Processor (zAAP).

Integrated Facility for Applications (IFA). see zSeries Application Assist Processor (zAAP).

IMS. Information Management System

Information Management System (IMS). A database/data communication (DB/DC) system that can manage complex databases and networks. Synonymous with IMS/VS.

installation performance specification (IPS). In MVS, a set of installation-supplied control information used by the system workload manager. An IPS includes performance group definitions, performance objectives, and coefficients used to establish the service rate. See also service rate.

Intelligent Resource Director (IRD). The Intelligent Resource Director (IRD) is a functional enhancement exclusive to IBM's zSeries family that can continuously and automatically reallocate resources throughout the

system in response to user requirements and is based on a customer's business priorities.

ITR. Internal throughput rate, see also *ETR*.

L

Latency. The time waiting for a particular record on a track to rotate to the read/write head. This usually averages one half the rotation time.

LCU. Logical control unit

License Manager. The IBM License Manager is the base for a new software pricing model. It allows vendors to enable their products for licensed software management by customers and is the basic tool IBM will use to implement the Workload License Charges pricing model on z900 servers.

local shared resources. An option for sharing I/O buffers, I/O-related control blocks, and channel programs among VSAM data sets in a resource pool that serves one partition or address space.

logically partitioned (LPAR) mode. A central processor mode that is available on the Configuration frame when using the PR/SM feature. It allows an operator to allocate processor unit hardware resources among logical partitions. Contrast with *basic mode*.

logical partition (LP). A subset of the processor hardware that is defined to support an operating system. See also *logically partitioned (LPAR) mode*.

LP. Logical partition.

LPAR. Logically partitioned (mode).

LPAR cluster. An LPAR cluster is the subset of the systems that are running as LPARs on the same CEC. Based on business goals, WLM can direct PR/SM to enable or disable CP capacity for an LPAR, without human intervention.

LSPR methodology. Recommended methodology for assessing processor power.

LSR. Local shared resources

LUE. Low utilization effect

M

migration rate. The rate (pages/second) of pages being moved from expanded storage through central storage to auxiliary storage.

mintime. The smallest unit of sampling in Monitor III. Specifies a time interval during which the system is sampled. The data gatherer combines all samples gathered into a set of samples. The set of samples can be summarized and reported by the reporter.

mode. Monitor III can run in various modes: The GO mode displays a chosen report, and updates it according to the interval you have chosen, the STOP mode, which is the default mode. The GRAPHIC mode presents the reports in graphic format using the GDDM®* product. You can also use TABULAR mode by setting GRAPHIC OFF.

MPL. Multiprogramming level

MTW. Mean time to wait

N

NVS. Nonvolatile storage

nonvolatile storage (NVS). Additional random access electronic storage with a backup battery power source, available with a 3990 Model 3/6 Storage Control, used to retain data during a power failure. Nonvolatile storage, accessible from all storage directors, stores data during DASD fast write and dual copy operations.

P

partitioned data set (PDS). A data set in direct access storage that is divided into partitions, called members, each of which can contain a program, part of a program, or data. Synonymous with program library.

PDS. partitioned data set

peak-to-average ratio. The ratio between the highest CPU utilization and the average utilization (peak hour busy/prime shift average busy).

PER. Program event recording

performance management. (1) The activity which monitors and allocates data processing resources to applications according to goals defined in a service level agreement or other objectives. (2) The discipline that encompasses collection of performance data and tuning of resources.

performance group. Group of work with the same performance objectives managed by the SRM.

PG. Performance group

PGN. Performance group number

PR/SM. Processor Resource/Systems Manager™.

Processor Resource/Systems Manager (PR/SM). The feature that allows the processor to run several operating systems environments simultaneously and provides logical partitioning capability. See also *LPAR*.

program event recording (PER). A hardware feature used to assist in debugging programs by detecting and recording program events.

R

range. The time interval you choose for your report.

RMA. The 3990 RPS miss avoidance feature.

RHIT. Read hit

S

sample. Once in every cycle, the number of jobs waiting for a resource, and what job is using the resource at that moment, are gathered for all resources of a system by Monitor III. These numbers constitute one sample.

SCP. System control program

seek. The DASD arm movement to a cylinder. A seek can range from the minimum to the maximum seek time of a device. In addition, some I/O operations involve multiple imbedded seeks where the total seek time can be more than the maximum device seek time.

service level agreement (SLA). A written agreement of the information systems (I/S) service to be provided to the users of a computing installation.

Service Level Reporter (SLR). An IBM licensed program that provides the user with a coordinated set of tools and techniques and consistent information to help manage the data processing installation. For example, SLR extracts information from SMF, IMS, and CICS logs, formats selected information into tabular or graphic reports, and gives assistance in maintaining database tables.

service rate. In the system resources manager, a measure of the rate at which system resources (services) are provided to individual jobs. It is used by the installation to specify performance objectives, and used by the workload manager to track the progress of individual jobs. Service is a linear combination of processing unit, I/O, and main storage measures that can be adjusted by the installation.

shared page groups. An address space can decide to share its storage with other address spaces using a function of RSM. As soon as other address spaces use these storage areas, they can no longer be tied to only one address space. These storage areas then reside as *shared page groups* in the system. The pages of shared page groups can reside in central, expanded, or auxiliary storage.

SLA. service level agreement

SLIP. serviceability level indication processing

SLR. Service Level Reporter

SMF. System management facility

speed. See *workflow*

SRB. Service request block

SRM. System resource manager

SSCH. Start subchannel

sysplex. A complex consisting of a number of coupled MVS systems.

system control program (SCP). Programming that is fundamental to the operation of the system. SCPs include MVS, VM, and VSE operating systems and any other programming that is used to operate and maintain the system. Synonymous with *operating system*.

T

TCB. Task control block

threshold. The exception criteria defined on the report options screen.

throughput. A measure of the amount of work performed by a computer system over a period of time, for example, number of jobs per day.

THRU. These are DASD writes to devices behind 3990s that are not enabled for DFW.

TP. Teleprocessing

TPNS. Teleprocessing network simulator

TSO. Time Sharing Option, see *Time Sharing Option/Extensions*

Time Sharing Option Extensions (TSO/E). In MVS, a time-sharing system accessed from a terminal that allows user access to MVS system services and interactive facilities.

U

UIC. Unreferenced interval count

uncaptured time. CPU time not allocated to a specific address space.

using. Jobs getting service from hardware resources (PROC or DEV) are *using* these resources.

V

velocity. A measure of how fast work should run when ready, without being delayed for processor or storage access. See also *execution velocity*.

VLF. Virtual Lookaside Facility

VTOC. Volume table of contents

W

WLM. Workload Manager

workflow. (1) The workflow of an address space represents how a job uses system resources and the speed at which the jobs moves through the system in relation to the maximum average speed at which the job could move through the system. (2) The workflow of resources indicates how efficiently users are being served.

workload. A logical group of work to be tracked, managed, and reported as a unit. Also, a logical group of service classes.

WSM. Working Set Manager

Z

zAAP. see zSeries Application Assist Processor.

zSeries Application Assist Processor (zAAP). A specialized processing assist unit configured for running Java programming on selected zSeries machines.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Corporation
Mail Station P300
2455 South Road
Poughkeepsie New York 12601-5400
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

IBM World Trade Asia Corporation
Licensing
2-31 Roppongi 3-chome, Minato-ku
Tokyo 106, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS"

WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Programming Interface Information

This book is intended to help you work with the reports RMF provides to understand your system's performance and to plan for appropriate tuning activities. The book documents information that is Diagnosis, Modification or Tuning Information provided by RMF.

Warning: Do not use this Diagnosis, Modification or Tuning Information as a programming interface.

Trademarks

The following terms are trademarks of the IBM Corporation in the United States and/or other countries:

- CICS
- CICS/ESA
- DB2
- DFSMS
- DFSMS/MVS
- DFSORT
- Enterprise Storage Server
- ESCON
- eServer
- FICON
- GDDM
- Hiperbatch
- Hipersorting
- IBM
- IMS
- MVS
- NetView
- OS/390
- Parallel Sysplex
- PR/SM
- Processor Resource/Systems Manager
- RACF

- RAMAC
- RMF
- S/390
- Seascape
- S/390 Parallel Enterprise Server
- Versatile Storage Server
- VTAM
- z/OS
- zSeries

UNIX® is a registered trademark of The Open Group in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Other company, product, and service names may be trademarks or service marks of others.

Index

Special characters

A

B

C

D

E

G

H

S

T

U

V

W

Z

Readers' Comments — We'd Like to Hear from You

z/OS
Resource Measurement Facility
Performance Management Guide

Publication No. SC33-7992-03

Overall, how satisfied are you with the information in this book?

	Very Satisfied	Satisfied	Neutral	Dissatisfied	Very Dissatisfied
Overall satisfaction	☐	☐	☐	☐	☐

How satisfied are you that the information in this book is:

	Very Satisfied	Satisfied	Neutral	Dissatisfied	Very Dissatisfied
Accurate	☐	☐	☐	☐	☐
Complete	☐	☐	☐	☐	☐
Easy to find	☐	☐	☐	☐	☐
Easy to understand	☐	☐	☐	☐	☐
Well organized	☐	☐	☐	☐	☐
Applicable to your tasks	☐	☐	☐	☐	☐

Please tell us how we can improve this book:

Thank you for your responses. May we contact you? ☐ Yes ☐ No

When you send comments to IBM, you grant IBM a nonexclusive right to use or distribute your comments in any way it believes appropriate without incurring any obligation to you.

Name

Address

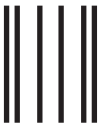
Company or Organization

Phone No.

Readers' Comments — We'd Like to Hear from You
SC33-7992-03

Cut or Fold
Along Line

Fold and Tape | **Please do not staple** | Fold and Tape



NO POSTAGE
NECESSARY
IF MAILED IN THE
UNITED STATES

BUSINESS REPLY MAIL

FIRST-CLASS MAIL PERMIT NO. 40 ARMONK, NEW YORK

POSTAGE WILL BE PAID BY ADDRESSEE

IBM Deutschland Entwicklung GmbH
eServer Performance Management Development
Schoenaicher Strasse 220
D-71032 Boeblingen
Federal Republic of Germany

Fold and Tape | **Please do not staple** | Fold and Tape

Cut or Fold
Along Line

Program Number: 5694–A01, 5655–G52

Printed in USA

SC33-7992-03

Spine information:

z/OS

z/OS V1R6.0 RMF Performance Management Guide

SC33-7992-03