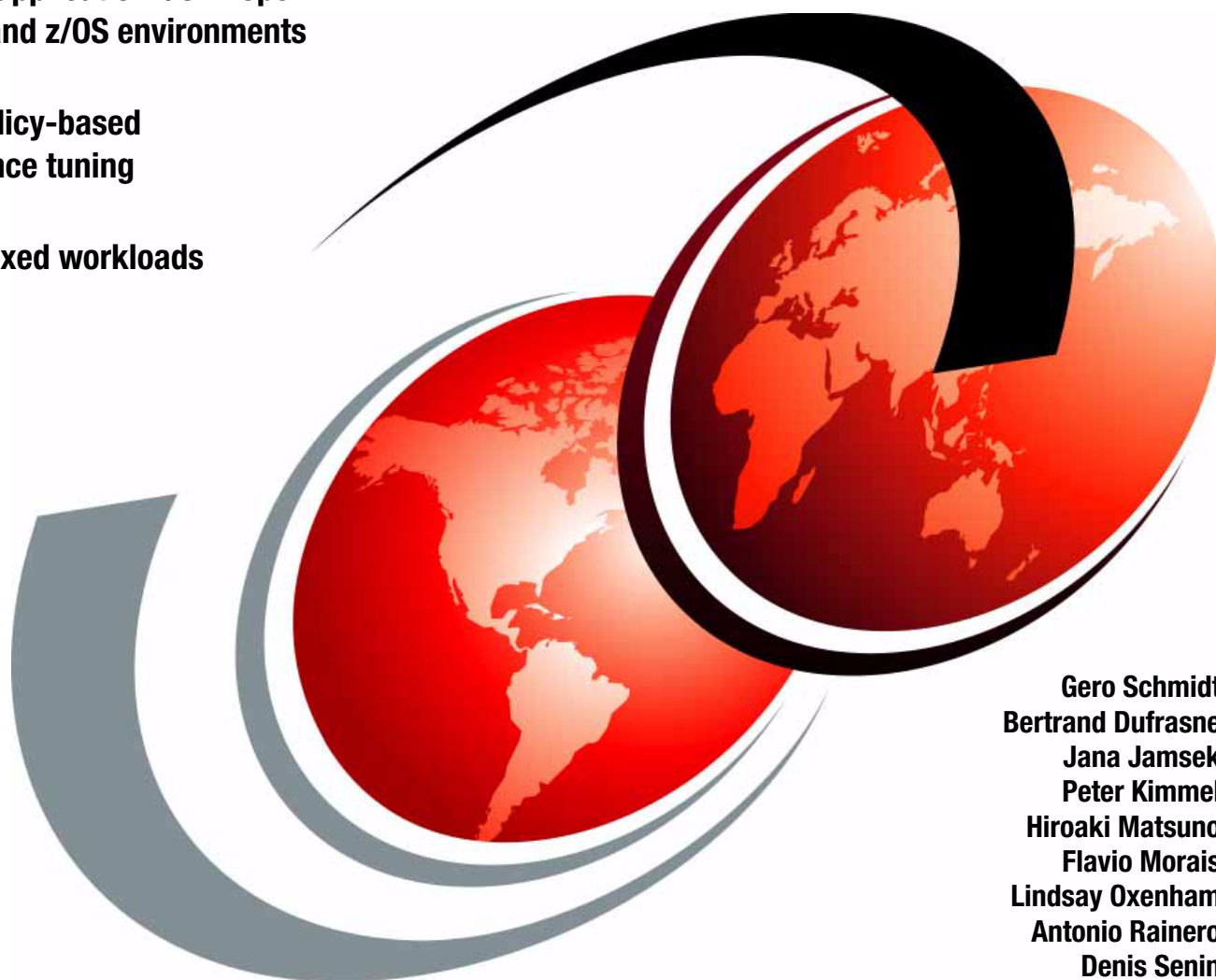


DS8000 I/O Priority Manager

Prioritize application I/O in open systems and z/OS environments

Enable policy-based performance tuning

Handle mixed workloads



Gero Schmidt
Bertrand Dufrasne
Jana Jamsek
Peter Kimmel
Hiroaki Matsuno
Flavio Morais
Lindsay Oxenham
Antonio Rainero
Denis Senin



International Technical Support Organization

DS8000 I/O Priority Manager

January 2012

Preparation for use: Before using this information and the product it supports, read the information in “Notices” on page v.

Second Edition (January 2012)

This edition applies to the IBM System Storage DS8700 with DS8000 Licensed Machine Code (LMC) level 6.6.2x.xxx and the IBM System Storage DS8800 with DS8000 Licensed Machine Code (LMC) level 7.6.2x.xxx.

This document was created or updated on March 21, 2012.

© Copyright International Business Machines Corporation 2012. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

| | |
|--|-----|
| Notices | v |
| Trademarks | vi |
| Preface | vii |
| The team who wrote this paper | vii |
| Now you can become a published author, too! | x |
| Comments welcome | x |
| Stay connected to IBM Redbooks | x |
| Chapter 1. I/O Priority Manager overview | 1 |
| 1.1 Introduction | 2 |
| 1.2 What the DS8000 I/O Priority Manager provides | 3 |
| 1.3 Business motivation for I/O Priority Manager | 3 |
| Chapter 2. I/O Priority Manager concept, design, and implementation | 5 |
| 2.1 Quality of Service with DS8000 | 6 |
| 2.2 I/O Priority Manager for open systems | 7 |
| 2.2.1 Performance policies for open systems | 7 |
| 2.2.2 Performance groups for open systems | 8 |
| 2.3 I/O Priority Manager for System z | 9 |
| 2.3.1 Performance policies and groups for System z | 9 |
| 2.3.2 I/O Priority Manager without z/OS software support | 11 |
| 2.3.3 I/O Priority Manager with z/OS software support | 11 |
| 2.3.4 Software input and effect on I/O priority for System z | 12 |
| 2.3.5 Service Class information passed by WLM | 14 |
| 2.4 I/O Priority Manager modes of operation | 18 |
| 2.5 I/O Priority Manager with Easy Tier | 19 |
| Chapter 3. Using I/O Priority Manager | 21 |
| 3.1 Licensing | 22 |
| 3.2 Activating I/O Priority Manager | 22 |
| 3.3 z/OS WLM settings to activate I/O Priority Manager | 23 |
| 3.4 Setting I/O Priority Manager mode using DS CLI | 24 |
| 3.5 Setting I/O Priority Manager mode using DS GUI | 25 |
| 3.6 Assigning I/O performance groups to FB volumes using DS CLI | 26 |
| 3.7 Assigning I/O performance groups to FB volumes using DS GUI | 28 |
| 3.8 Assigning I/O performance groups to CKD volumes using DS CLI | 30 |
| 3.9 Assigning I/O performance groups to CKD volumes using DS GUI | 31 |
| Chapter 4. I/O Priority Manager usage scenarios | 33 |
| 4.1 FB scenario: Architecture environment | 34 |
| 4.1.1 FB scenario: Test description | 35 |
| 4.1.2 FB scenario: test results | 37 |
| 4.2 CKD scenario: Architecture environment | 43 |
| 4.2.1 CKD scenario: test description | 43 |
| 4.2.2 CKD scenario: test results | 45 |
| Chapter 5. Monitoring and reporting | 59 |
| 5.1 Monitoring and reporting with the DS CLI | 60 |

| | | |
|-------|--|----|
| 5.1.1 | Displaying I/O performance groups | 60 |
| 5.1.2 | Performance reports | 61 |
| 5.2 | Monitoring and reporting with the DS GUI | 64 |
| | Related publications | 67 |
| | IBM Redbooks | 67 |
| | Other publications | 67 |
| | Online resources | 67 |
| | How to get IBM Redbooks publications | 68 |
| | Help from IBM | 68 |
| | Index | 69 |

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.

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 websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.


COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Trademarks

IBM, the IBM logo, and [ibm.com](http://www.ibm.com) are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

| | | |
|----------------------------|---|--------------------------------|
| AIX® | IBM® | Resource Measurement Facility™ |
| DB2® | MVS™ | RMF™ |
| DS4000® | Power Systems™ | System i® |
| DS6000™ | PowerHA® | System Storage® |
| DS8000® | PowerVM® | System z® |
| Easy Tier® | Power® | XIV® |
| Enterprise Storage Server® | Redbooks® | z/OS® |
| FICON® | Redpaper™ | |
| i5/OS® | Redbooks (logo)  ® | |

The following terms are trademarks of other companies:

Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Preface

This IBM® Redpaper™ publication describes the concepts and functions of the IBM System Storage® DS8000® I/O Priority Manager. The DS8000 I/O Priority Manager enables more effective storage consolidation and performance management combined with the ability to align quality of service (QoS) levels to separate workloads in the system.

With DS8000 I/O Priority Manager, the system can prioritize access to system resources to achieve the volume's desired QoS based on defined performance goals (high, medium, or low) of any volume. I/O Priority Manager constantly monitors and balances system resources to help applications meet their performance targets automatically, without operator intervention. Starting with DS8000 Licensed Machine Code (LMC) level R6.2, the DS8000 I/O Priority Manager feature supports open systems and IBM System z®.

DS8000 I/O Priority Manager, together with IBM z/OS® Workload Manager (WLM), provides more effective storage consolidation and performance management for System z systems. Now tightly integrated with Workload Manager for z/OS, DS8000 I/O Priority Manager improves disk I/O performance for important workloads. It also drives I/O prioritization to the disk system by allowing WLM to give priority to the system's resources automatically when higher priority workloads are not meeting their performance goals. Integration with zWLM is exclusive to DS8000 and System z systems.

The paper is aimed at those who want to get an understanding of the DS8000 I/O Priority Manager concept and its underlying design. It provides guidance and practical illustrations for users who want to exploit the capabilities of the DS8000 I/O Priority Manager.

The team who wrote this paper

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, San Jose Center.

Gero Schmidt is an IT Specialist in the IBM ATS technical sales support organization in Germany. He joined IBM in 2001, working at the European Storage Competence Center (ESCC) in Mainz. There he provided technical support for a broad range of IBM storage systems (ESS, DS4000®, DS5000, IBM DS6000™, DS8000, SVC, and IBM XIV®) in open systems environments. He primarily focused on IBM enterprise disk storage solutions, storage system performance, and IBM Power® Systems™ with AIX® including IBM PowerVM® and IBM PowerHA®. He also participated in the product rollout and major release beta test programs of the IBM System Storage DS6000/DS8000 series. He has been a speaker at several international IBM technical conferences and has co-authored several IBM Redbooks® publications. He holds a degree in Physics (Dipl.-Phys.) from the Technical University of Braunschweig, Germany.

Bertrand Dufrasne is an IBM Certified Consulting IT Specialist and Project Leader for System Storage disk products at the International Technical Support Organization, San Jose Center. He has worked at IBM in various IT areas, authored many IBM Redbooks publications, and developed and taught technical workshops. Before joining the ITSO, he worked for IBM Global Services as an Application Architect. He holds a master's degree in Electrical Engineering.

Jana Jamsek is an IT Specialist for IBM Slovenia. She works in Storage Advanced Technical Support for Europe as a specialist for IBM Storage Systems and the IBM i (i5/OS®) operating system. Jana has eight years of experience working with the IBM System i® platform and its predecessor models and eight years of experience working with storage systems. She has a master's degree in computer science and a degree in mathematics from the University of Ljubljana in Slovenia.

Peter Kimmel is an IT Specialist and ATS team lead of the Enterprise Disk Solutions team at the European Storage Competence Center (ESCC) in Mainz, Germany. He joined IBM Storage in 1999 and has since worked with all the various IBM Enterprise Storage Server® (ESS) and System Storage DS8000 generations with a focus on architecture and performance. He has been involved in the Early Shipment Programs (ESPs) of these early installs and has co-authored several DS8000 IBM Redbooks publications. Peter holds a MSc degree in physics from the University of Kaiserslautern.

Hiroaki Matsuno is an IT Specialist in IBM Japan. He has three years of experience in IBM storage system solutions working in the IBM ATS System Storage organization in Japan. His areas of expertise include DS8000 Copy Services, SAN, and Real-time Compression Appliance in open systems environments. He holds a master's of engineering degree from the University of Tokyo, Japan.

Flavio Morais is a GTS Storage Specialist in Brazil and has six years of experience in the SAN/storage field. He holds a degree in computer engineering from Instituto de Ensino Superior de Brasilia. His areas of expertise include DS8000 Planning, Copy Services, TPC and Performance Troubleshooting. He has worked extensively with performance problems in open systems.

Lindsay Oxenham is a Mainframe Storage Specialist working in Melbourne, Australia. He has over thirty years of experience in the mainframe environment working as an application programmer in performance and tuning areas. He joined IBM in 1998 and has been working in the storage area since 2005. He has a bachelor's degree in applied science (computing), and has presented papers at SAS user meetings and at Australia's Computer Management Group (CMG) conferences.

Antonio Rainero is a Certified IT Specialist working for Integrated Technology Services organization in IBM Italy. He joined IBM in 1998 and has more than ten years of experience in the delivery of storage services both for z/OS and open systems customers. His areas of expertise include storage subsystems implementation, performance analysis, storage area networks, storage virtualization, disaster recovery, and high availability solutions. Antonio holds a degree in computer science from University of Udine, Italy.

Denis Senin is an IT Specialist in IBM Russia. He has ten years of experience in the IT industry and has worked at IBM for six years. Denis holds a master's degree in design and engineering of computer systems from the Moscow State Institute of Radiotechnics, Electronics, and Automatics and has a background in systems design and development. His current areas of expertise include open systems high-performing and disaster recovery storage solutions.



The team: Hiroaki, Antonio, Flavio, Peter, Bertrand, Denis, Jana, Gero, Lindsay

Thanks to the authors of the previous edition of this paper, published in August 2011:

- ▶ Stéphane Couteau
- ▶ Blake Elliott
- ▶ Martin Jer
- ▶ Brenton Keller
- ▶ Stephen Manthorpe
- ▶ Richard Murke
- ▶ Massimo Rosichini

Thanks also to the following people for their contributions to this project:

- ▶ Werner Bauer
- ▶ David Chambliss
- ▶ Ingo Dimmer
- ▶ Wolfgang Heitz
- ▶ Frank Krüger
- ▶ Michael Lopez
- ▶ Richard Ripberger
- ▶ Günter Schmitt
- ▶ Brian Sherman
- ▶ William Sherman
- ▶ Horst Sinram
- ▶ Paulus Usong
- ▶ Stefan Wirag

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at: ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

- ▶ Use the online **Contact us** review Redbooks form found at:

ibm.com/redbooks

- ▶ Send your comments in an email to:

redbooks@us.ibm.com

- ▶ Mail your comments to:

IBM Corporation, International Technical Support Organization
Dept. HYTD Mail Station P099
2455 South Road
Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:

<http://www.facebook.com/IBMRedbooks>

- ▶ Follow us on Twitter:

<http://twitter.com/ibmredbooks>

- ▶ Look for us on LinkedIn:

<http://www.linkedin.com/groups?home=&gid=2130806>

- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>

- ▶ Stay current on recent Redbooks publications with RSS Feeds:

<http://www.redbooks.ibm.com/rss.html>



I/O Priority Manager overview

In this chapter, we introduce DS8000 I/O Priority Manager, describe its business motivation, and discuss what it provides.

1.1 Introduction

One of the most critical considerations in today's business world is data storage. Storage drives business decisions, business protection, and Quality of Service (QoS) for clients. Many data center managers struggle to provide optimal performance for all users with the limited resources at their disposal, and storage systems are called upon to provide high performance and availability to a diverse set of clients.

As cloud computing becomes a way for many companies to efficiently host multiple operating systems (OS), middleware, and applications in consolidated environments, data center managers face many issues. One long-standing issue is workload contention, where one application imposes a heavy workload on system resources and the QoS offered to others is significantly degraded. Also, back-end I/O contention on storage devices can be a serious problem for enterprise data centers, and achieving the desired QoS in these complex and dynamic environments becomes harder as the center grows.

DS8000 I/O Priority Manager is a new licensed function feature introduced for IBM System Storage DS8700 and DS8800 storage systems with DS8000 Licensed Machine Code (LMC) R6.1 or higher. It enables more effective storage consolidation and performance management and the ability to align QoS levels to separate workloads in the system which are competing for the same shared and possibly constrained storage resources.

DS8000 I/O Priority Manager constantly monitors system resources to help applications meet their performance targets automatically, without operator intervention. The DS8000 storage hardware resources that are monitored by the I/O Priority Manager for possible contention are the RAID ranks and device adapters.

Many I/O-based QoS implementations require the user to define fixed thresholds for competing workloads like upper limits for the I/O rate or data bandwidth. The DS8000 I/O Priority Manager provides a dynamic throttle that lowers the prioritized workload only when actual resource constraint conditions are occurring such that the targeted QoS for a higher prioritized workload might otherwise be missed. In workload contention situations, DS8000 I/O Priority Manager delays I/Os with lower priority performance policies to help I/Os with higher performance policies meet their QoS targets.

I/O Priority Manager is easy to use and requires only minimal supervision from the storage system administrator.

1.2 What the DS8000 I/O Priority Manager provides

The DS8000 I/O Priority Manager provides more effective storage consolidation and performance management, combined with the ability to align QoS levels to separate workloads in the system. I/O Priority Manager prioritizes access to system resources to achieve the desired QoS based on defined performance goals (high, medium, or low) for either the volume or single I/O request. The Priority Manager constantly monitors and balances system resources to help applications meet their performance targets automatically, without operator intervention.

For open systems, I/O Priority Manager accomplishes this goal once activated and correctly set up, at which time the administrator assigns volumes to the appropriate performance groups. The administrator defines performance goals of high priority (performance groups 1 - 5), medium priority (performance groups 6 - 10), low priority (performance groups 11 - 15), and 'no monitor no manage' (performance group 0), as shown in Table 2-1 on page 8.

For System z, the administrator defines the performance goals of high priority (performance groups 19 - 21), medium priority (performance groups 22 - 25), low priority (performance groups 26 - 31), and 'no monitor no manage' (performance group 0), as shown in Table 2-2 on page 10. I/O Priority Manager with z/OS Workload Manager support adds more granularity enabling the I/O prioritization at single I/O operation level.

1.3 Business motivation for I/O Priority Manager

It is increasingly common to use one storage system to serve many categories of workloads with separate characteristics and requirements. The widespread use of virtualization and the advent of cloud computing that facilitates consolidating applications into a shared storage infrastructure has become common practice. However, this consolidation creates the likelihood that business critical applications suffer performance degradation because of resource contention with less important applications. Workloads are forced to compete for resources such as disk storage capacity, bandwidth, device adapters, and ranks.

Clients are requiring that their desired QoS is preserved for selected applications, even when resource contention occurs. Achieving the target QoS is challenging given the dynamic and diverse characteristics of storage workloads, the separate types of storage system components, and real-time requirements.

A cost-effective solution to this problem is to incorporate QoS-aware scheduling into the storage system, which can dynamically ensure that appropriate performance is delivered to important applications. Setting I/O priorities helps maintain high performance of important I/Os when the system is constrained by existing resources. I/O Priority Manager protects the performance of important I/Os by throttling (slowing down) less important I/Os when I/O load on system components, such as device adapters and ranks, reach their physical capacity.

I/O Priority Manager, together with z/OS Workload Manager (WLM), enables more effective storage consolidation and performance management. This function, tightly integrated with WLM, improves disk I/O performance for important workloads. The function drives I/O prioritization to the disk system by allowing WLM to give priority to the system's resources automatically when higher priority workloads are not meeting their performance goals.

I/O Priority Manager and IBM Easy Tier® (described in IBM Redpaper *IBM System Storage DS8000 Easy Tier*, REDP-4667) together provide independent benefits to improve application performance.



I/O Priority Manager concept, design, and implementation

The DS8000 I/O Priority Manager enables more effective storage consolidation and performance management, combined with the ability to align quality of service (QoS) levels to separate workloads in the system.

For open systems, the DS8000 I/O Priority Manager system can prioritize access to system resources to achieve the volume's desired QoS based on defined performance goals (high, medium, or low) of any volume. I/O Priority Manager constantly monitors and balances system resources to help applications meet their performance targets automatically, without operator intervention.

For System z, I/O Priority Manager can use the z/OS Workload Manager (zWLM) to detect when a higher-priority application can be prioritized over lower-priority application that is competing for the same system resources. I/O Priority Manager delays the lower-priority I/O data to assist the more critical I/O data in meeting their performance targets. Without WLM support, I/O Priority Manager for System z can be used in the same way as it is for open systems.

2.1 Quality of Service with DS8000

The performance requirements for a set of I/O operations performed for a given application can generally be expressed in term of a QoS specification.

There are several approaches to measuring a QoS. With regard to I/O Priority Manager in the DS8000, QoS is a metric that uses the ratio of *optimal response time* to *average response time* for I/Os to a rank. QoS is defined as follows:

$$\text{Quality of Service (QoS)} = \frac{\text{optimal response time}}{\text{average response time}}$$

A high priority QoS target of 70% means that the average response time is targeted not to exceed the optimal response time by more than 43%, with the optimal response time based on 35% rank utilization.

The optimal response time for a rank is estimated by the system based on the RAID type and the type (size and speed) of the disks in the rank and the workload on the rank.

QoS is inversely proportional to the average response time. A lower response time generally means a higher QoS. However, because QoS also takes into account the type of disks in the rank and the type of workload being performed on the rank, similar response times can produce separate QoSs. For example, a response time of 8 ms on a read miss yields a high QoS for nearline disks, but it yields a lower QoS for enterprise disks and a very low QoS for solid-state drives (SSDs). If the workload had a high proportion of cache hits, the expected response time would be in the sub-millisecond range, so a response time of 2 - 3 ms would yield a low QoS.

QoS normally ranges from 0 to 100, but it can sometimes exceed 100, depending on the situation. A volume with a QoS of 100 means that I/Os are currently experiencing near optimal response times. A volume with a QoS of 50 means the I/Os are currently experiencing response times that are approximately twice as long as their optimal response time.

In the DS8000, each volume is assigned a *performance policy*. More precisely, each volume is assigned a performance group, and each performance group has a QoS target. This QoS target is used to determine whether or not a volume is experiencing appropriate response times. If the volume has a QoS that is less than its target QoS, the volume is not experiencing the types of response times that are expected for a volume in that performance group. If the volume's QoS is above its QoS target, the volume is experiencing response times that are within an acceptable range for a volume in that performance group. For System z with z/OS WLM support, the granularity of the assignment to the performance group is at single I/O operation level.

2.2 I/O Priority Manager for open systems

In the following section, we introduce the main I/O Priority Manager definitions that apply to open systems.

2.2.1 Performance policies for open systems

A performance policy sets the priority of a volume relative to other volumes.

The DS8000 has four defined performance policies: *default*, *high priority*, *medium priority*, and *low priority*. All volumes fall into one of these four performance policies.

- ▶ Default performance policy

The default performance policy does not have a QoS target associated with it, and I/Os to volumes which are assigned to the default performance policy are never delayed by I/O Priority Manager. Volumes on existing DS8000 storage systems that are upgraded to R6.1 are assigned the default performance policy.

- ▶ High priority performance policy

The high priority performance policy has a QoS target of 70. This means that I/Os from volumes associated with the high performance policy attempt to stay under approximately 1.5 times the optimal response of the rank. I/Os in the high performance policy are never delayed.

- ▶ Medium priority performance policy

The medium priority performance policy has a QoS target of 40. This means I/Os from volumes with the medium performance policy attempt to stay under 2.5 times the optimal response time of the rank.

- ▶ Low performance policy

Volumes with a low performance policy have no QoS target and have no goal for response times.

Users associate logical volumes with the performance policy that characterizes the expected type of workload for that volume. This policy assignment allows the I/O Priority Manager to make decisions about the relative priority of I/O operations.

In workload contention situations, I/O Priority Manager delays I/Os with lower performance policies to help I/Os with higher performance policies meet their QoS targets.

2.2.2 Performance groups for open systems

A performance group associates the I/O operations of a logical volume with a performance policy. Performance groups are used for reporting statistics on the logical volumes that are associated with them.

For open systems, there are 16 performance groups: five performance groups each for the high, medium, and low performance policies, and one performance group for the default performance policy, as shown in Table 2-1.

A performance group can have multiple volumes assigned to it. A volume can reside in exactly one performance group.

Table 2-1 Performance group to performance policy mapping for open systems

| Performance group | Performance policy | Priority (as seen in GUI/CLI) | QoS target | Name |
|-------------------|--------------------|-------------------------------|------------|-----------------------------|
| 0 | 1 | 0 | 0 | Default |
| 1 | 2 | 1 | 70 | Fixed Block High Priority |
| 2 | 2 | 1 | 70 | Fixed Block High Priority |
| 3 | 2 | 1 | 70 | Fixed Block High Priority |
| 4 | 2 | 1 | 70 | Fixed Block High Priority |
| 5 | 2 | 1 | 70 | Fixed Block High Priority |
| 6 | 3 | 5 | 40 | Fixed Block Medium Priority |
| 7 | 3 | 5 | 40 | Fixed Block Medium Priority |
| 8 | 3 | 5 | 40 | Fixed Block Medium Priority |
| 9 | 3 | 5 | 40 | Fixed Block Medium Priority |
| 10 | 3 | 5 | 40 | Fixed Block Medium Priority |
| 11 | 4 | 15 | 0 | Fixed Block Low Priority |
| 12 | 4 | 15 | 0 | Fixed Block Low Priority |
| 13 | 4 | 15 | 0 | Fixed Block Low Priority |
| 14 | 4 | 15 | 0 | Fixed Block Low Priority |
| 15 | 4 | 15 | 0 | Fixed Block Low Priority |

Performance groups are assigned to a volume at the time of the volume creation. If you migrate from an earlier DS8000 microcode level to the release 6.1 code level, the fixed block (FB) volumes are assigned to the default performance group PG0. You can then use either the DS GUI or the DS CLI **chfbvo1** command to assign a separate performance group to the volume. Performance group 0 is the default performance group for new volumes that do not specify a performance group during creation.

Users need to assign a performance group to a volume that selects the appropriate performance policy for the expected workloads on that volume:

- ▶ Volumes that require higher performance relative to other volumes need to be assigned to performance groups PG1 through PG5 so that they are associated with the high priority performance policy.
- ▶ Volumes that do not require as much performance relative to other volumes, need to be assigned to performance groups PG6 through PG10 to be associated with the medium priority performance policy.
- ▶ Volumes that have no performance requirements need to be assigned to performance groups PG11 through PG15 to be associated with the low priority performance policy.

There is no difference between PG1, PG2, PG3, PG4, and PG5 (the same is true for PG6 through PG10 and PG11 through PG15). All volumes in these performance groups have the same performance policy. Multiple performance groups are defined for each policy so that more than one group of volumes can have the same performance policy. This allows independent monitoring and reporting for each group even though they share the same performance policy.

As an example, say a user has an IBM DB2® application running under IBM AIX, a mail server running under Windows, and a backup server. The user then assigns the DB2 storage to PG1, the mail server storage to PG3, and the backup server storage to PG11. In this example, DB2 and the mail server share the same high priority policy, and the backup server is assigned to the low priority performance policy. The performance groups allow the user to get individual statistics on other volumes or groups of volumes with the same performance policy. In this example, the user can get individual statistics for the DB2 storage, the mail server storage, and the backup storage.

You can monitor and report on each performance group using the `lspgrprpt` command.

2.3 I/O Priority Manager for System z

The following section introduces the main I/O Priority Manager definitions that apply to the System z environment. With System z, two operation modes are available for I/O Priority Manager: *without software support* or *with software support*.

z/OS and I/O Priority Manager: Currently only z/OS operating system uses the I/O Priority Manager with software support.

2.3.1 Performance policies and groups for System z

For System z, there are 14 performance groups:

- ▶ Three performance groups for high-performance policies (19 - 21)
- ▶ Four performance groups for medium-performance policies (22 - 25)
- ▶ Six performance groups for low-performance policies (26 - 31),
- ▶ One performance group (0) for the default performance policy. Performance groups 16 - 18 are not used but are mapped to performance policy 0.

See Table 2-2 for information about performance groups, their correspondence to performance policies, and QoS targets for System z.

Table 2-2 Performance group to performance policy mapping for System z

| Performance group | Performance policy | Priority (as seen in GUI/CLI) | QoS target | Name |
|-------------------|--------------------|-------------------------------|------------|-----------------------|
| 0 | 1 | 0 | 0 | Default |
| 16 | 16 | 0 | 0 | No Management |
| 17 | 17 | 0 | 0 | No Management |
| 18 | 18 | 0 | 0 | No Management |
| 19 | 19 | 1 | 80 | CKD High Priority 1 |
| 20 | 20 | 2 | 80 | CKD High Priority 2 |
| 21 | 21 | 3 | 70 | CKD High Priority 3 |
| 22 | 22 | 4 | 45 | CKD Medium Priority 1 |
| 23 | 23 | 4 | 5 | CKD Medium Priority 2 |
| 24 | 24 | 5 | 45 | CKD Medium Priority 3 |
| 25 | 25 | 6 | 5 | CKD Medium Priority 4 |
| 26 | 26 | 7 | 5 | CKD Low Priority 1 |
| 27 | 27 | 8 | 5 | CKD Low Priority 2 |
| 28 | 28 | 9 | 5 | CKD Low Priority 3 |
| 29 | 29 | 10 | 5 | CKD Low Priority 4 |
| 30 | 30 | 11 | 5 | CKD Low Priority 5 |
| 31 | 31 | 12 | 5 | CKD Low Priority 6 |

The performance policy number is equal to the performance group number. For example, performance group 19 is equivalent to performance policy 19. The performance group number used for count-key-data (CKD) volumes is separate from the performance numbers used for fixed-block (FB) volumes.

Performance priority 0 has been made common between FB and CKD, but none of the other policies are overlapping. Policy 0 is not managed, which means that there is no delay added in the I/O in the absence of software support and there is no Quality of Service (QoS) target for that particular volume. Thus, the volume does not impact any other I/O if it is running slow.

Unlike the FB priorities, in which all the high priorities are the same, all medium priorities are the same, and all low priorities are the same, for CKD, the priorities are ordered with priority 1 having a higher priority than priority 2, and so forth. You have a wide range of granular priorities available, ranging from high to low for CKD volumes. The IOPM ensures that the performance goals are met for priority 1 more aggressively than for priority 2 or any other lower priority.

Delays to I/Os are added only on a rank with contention. Delays are added if there is I/O with a higher priority that is not meeting the QoS target. That is, an I/O with a lower priority is delayed up to a maximum allowed time before delaying I/O with higher priority. The maximum delay that can be added to an I/O is 200 ms.

When I/O of the lowest priority is delayed at its maximum, the I/O with the second lowest priority gets delayed if certain higher priority I/O still does not meet its QoS. If I/O to the highest performance policy is still not meeting its QoS target, then I/O to all lower levels are already delayed to their maximum value.

2.3.2 I/O Priority Manager without z/OS software support

When z/OS WLM support to I/O Priority Manager is not active on the DS8000 and on ranks in saturation, volume I/O is managed according to the performance policy of the performance group the volume is assigned to on the DS8000. Without activated WLM software support on z/OS, I/O Priority Manager on DS8000 works the same way for CKD volumes as it does for FB volumes except that you have slightly different performance group and policy definitions for CKD volumes as for FB volumes.

2.3.3 I/O Priority Manager with z/OS software support

In an I/O Priority Manager environment, the z/OS user provides input to WLM to indicate the priority of the applications or workloads and setups, input to tell WLM how to manage the I/O for the applications, and the schedule for those applications. For a given workload, when z/OS issues the I/O to the storage subsystem, there are two parameters in the prefix command of the CKD chain:

- ▶ *Importance* value based on what is specified in the WLM service class definition of the specific workload
- ▶ *Achievement* value based on what the WLM has determined as the historic performance trend in that particular service class versus the performance goal defined for that service class.

Together, the importance and achievement values determine the performance policy for the I/O. If the I/O is going to a rank that is in contention, the I/O execution is managed by the DS8000 according to the I/Os performance policy.

2.3.4 Software input and effect on I/O priority for System z

When any kind of work enters the z/OS system, such as transactions or batch jobs, this work is classified by the zWLM by assigning the new work a *service class*. For each service class, a *performance goal* is assigned, and for each performance goal a *business importance* is assigned that consists of a number from 1 to 5.

You can assign the following kinds of performance goals to service classes:

- ▶ Average response time
- ▶ Response time with percentile
- ▶ Velocity
- ▶ Discretionary

According to the definitions and their goal achievement levels of the service classes, WLM passes to I/O Priority Manager two performance metrics: the *I/O importance* (or simply *importance*) and the *I/O goal achievement* (or simply *achievement*).

An I/O importance of 0 means no I/O importance is specified. There are six levels that you can specify, 1 being the highest and 6 being the lowest, and anything in between ranked from high to low. This value is assigned by z/OS based on the information the user supplied to the WLM about application behavior and the association of this I/O with the highest priority process, medium priority process, or low priority process. Note that the I/O importance does not match exactly the business importance specified in the WLM service class definition. See Table 2-3 on page 14 for the correct matching.

For the I/O achievement, 0 means no I/O achievement is recorded. There are seven levels of achievement, where 1 means that the workload is significantly overachieving its QoS target and 7 means that the workload is significantly underachieving its target.

The combination of the two input metrics from the software is used to map to a performance policy. Thus, if both the I/O importance and the I/O achievement are 0, the volume performance policy is used to manage the I/O. However, if one or both of the metrics are non-zero, the DS8000 assigns a priority to that particular I/O.

This range gets mapped to the performance groups 19 to 31 in order of highest to lowest priority. If the metrics show high importance and significant underachievement, the DS8000 assigns this I/O the highest priority (priority 1). On the other hand, if the metrics show that the lowest important I/O is significantly overachieving, this I/O is given the lowest priority (priority 6).

With the performance policies assigned to each I/O operation, I/O Priority Manager determines which I/O requests are more important than others and which I/O requests need to be processed faster to fulfill the performance goals for the corresponding workload in z/OS. In case of resource contention (that is, rank saturation) the I/O Priority Manager throttles I/O requests with a lower performance priority to help I/O requests with higher priority. Figure 2-1 summarizes the I/O request handling through the various components from the z/OS to the DS8000.

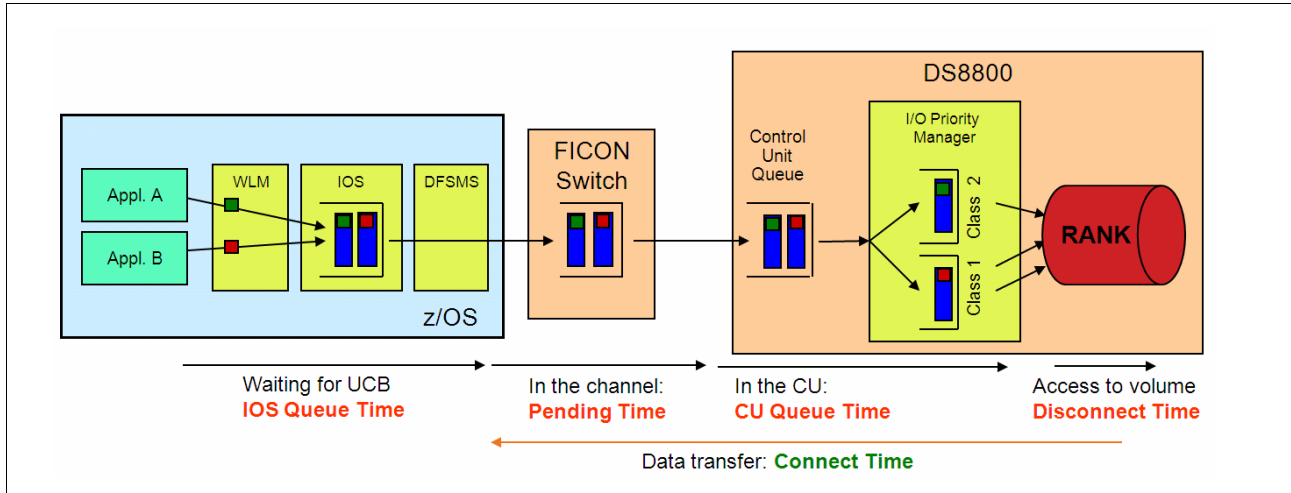


Figure 2-1 I/O request handling

This figure shows two applications, Appl.A and Appl.B, initiating I/O requests directed to the same DS8000 rank. The WLM adds to the channel program of both I/O requests the importance and goal achievement values according to the applications service classes. These requests eventually arrive at the DS8000, where the I/O Priority Manager component assigns a performance policy and then a performance group to the two I/O requests. At this point, if the rank is overloaded, the I/O Priority Manager begins throttling the I/O with the lower priority.

Usually when the WLM support is enabled, the volume assignment to a performance group is not taken in account by I/O Priority Manager anymore. I/O Priority manager assign a performance group to every single I/O operation according to the importance and achievement values and then sets the proper performance policy to the I/O operation. However, when either the importance or achievement value is set to 0 for a given I/O request, the I/O Priority Manager uses the performance policy of the performance group where the target I/O request volume resides. For this reason, even with the WLM support enabled, the volume assignment to a performance group must be planned carefully.

Performance groups are assigned to a volume at the time of the volume creation. If you migrated from an earlier DS8000 microcode level to the release 6.2 code level, the CKD volumes are assigned to the default performance group PG0. You can then use either the DS GUI or the DS CLI `chckdvo1` command to assign a separate performance group to the CKD volume. Performance group 0 is the default performance group for new volumes that do not specify a performance group during creation.

Avoiding I/O delay: To avoid any delay to an I/O request with either importance or achievement set to 0, assign all the volumes to performance group PG16.

You can monitor and report on each performance group using the `lspfrgrprt` command.

2.3.5 Service Class information passed by WLM

WLM derives performance policy assignment for service class periods, as summarized in Table 2-3. For CKD management, the performance policy number is always equal to the performance group number. For this reason, in Table 2-4 on page 15 through Table 2-7 on page 16, performance policy values also denote the performance group assignments. Depending on the specific type of goal, separate values for Importance and Achievement, and thus separate performance policies, apply.

Table 2-3 Service class information passed to WLM

| Goal type | Importance | Achievement | Goal |
|-----------------------------|------------|-------------|----------|
| System service class (SYS*) | 0 | None (0) | None (0) |
| Execution velocity | 1-5 | None (0) | 1-99 |
| Response time | 1-5 | 1-7 | None (0) |
| Discretionary | 6 | 1 | None (0) |

Table 2-3 shows each type of performance goal:

- ▶ For the *System service class (SYS*)* goal, Importance and Achievement values are 0. In this case, the performance policy is set to 0 (no management).
- ▶ For the *Execution velocity* goal, the Importance values can be 1 to 5, according to the WLM service class importance, but the Achievement value is not used. In this case, there is a performance policy *static* assignment according to the importance and the execution velocity goal. This means that for a given service period with an execution velocity goal, all the I/O requests are assigned to the same performance policy.

Table 2-4 shows the mapping between the input values (Importance and Execution Velocity Goal) and the resulting performance policy.

Table 2-4 Mapping of importance and execution velocity goal to performance policies (or groups)

| | | Execution Velocity Goal | | | | |
|------------|---|-------------------------|-------|-------|-------|-------|
| | | 1-15 | 16-40 | 41-59 | 60-85 | 86-99 |
| Importance | 1 | 28 | 26 | 22 | 21 | 20 |
| | 2 | 28 | 26 | 22 | 21 | 20 |
| | 3 | 29 | 28 | 25 | 22 | 21 |
| | 4 | 29 | 28 | 25 | 22 | 21 |
| | 5 | 30 | 29 | 27 | 24 | 22 |
| | 6 | 30 | 29 | 27 | 24 | 22 |

- For the *Response time* goal, again the Importance values can be 1 to 5, according to the WLM service class period importance. The Achievement value is derived from the performance index as shown in Table 2-5. In this case, the performance policy assignment is *dynamic* because during a service period a performance index might change.

Table 2-5 . Mapping of service class performance index and goal achievement value

| Performance index | Goal achievement value |
|---------------------|---------------------------------|
| $PI \leq 0.5$ | 1 (significantly overachieve) |
| $0.5 < PI \leq 0.7$ | 2 (overachieve) |
| $0.7 < PI \leq 0.9$ | 3 (slightly over achieve) |
| $0.9 < PI < 1.4$ | 4 (achieve) |
| $1.4 \leq PI < 2.5$ | 5 (slightly under achieve) |
| $2.5 \leq PI < 4.5$ | 6 (under achieve) |
| $PI \geq 4.5$ | 7 (significantly under achieve) |

- For the *Discretionary* goal, the Importance value is 6 and the Achievement value is 1. In this case, the performance policy assignment is static and can be derived from Table 2-6. This table reports the matching between the input values (Importance and Achievement) and the performance policy

Table 2-6 Mapping of importance and goal achievement values to performance policies (or groups)

| | | Achievement | | | | | | | |
|------------|---|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Importance | 0 | LV-PG ¹ | LV-PG ¹ | LV-PG ¹ | LV-PG ¹ | LV-PG ¹ | LV-PG ¹ | LV-PG ¹ | LV-PG ¹ |
| | 1 | LV-PG ¹ | 27 | 26 | 25 | 23 | 21 | 20 | 19 |
| | 2 | LV-PG ¹ | 27 | 26 | 25 | 23 | 21 | 20 | 19 |
| | 3 | LV-PG ¹ | 29 | 27 | 26 | 25 | 22 | 21 | 20 |
| | 4 | LV-PG ¹ | 29 | 27 | 26 | 25 | 22 | 21 | 20 |
| | 5 | LV-PG ¹ | 31 | 30 | 28 | 26 | 24 | 22 | 21 |
| | 6 | LV-PG ¹ | 31 | 30 | 28 | 26 | 24 | 22 | 21 |

Table notes:

1. If the program specifies a zero value for either a service class importance or service class achievement, the I/O Priority Manager uses the specified logical volume performance group (LV-PG) object for this I/O operation.

There is currently no mapping to performance groups 23 and 25

Table 2-7 shows the CKD performance policy attributes.

Table 2-7 CKD performance policy attributes

| Performance policy | Priority | Quality of service target | Maximum delay factor |
|--------------------|------------------|---------------------------|--|
| 0 | N.A ⁴ | N.A ⁴ | N.A ¹ |
| 19 | 1 | 80% | 1.0 ¹ (for example, QoS=100.0%) ² |
| 20 | 2 | 80% | 1.0 ¹ (for example, QoS=100.0%) ² |
| 21 | 3 | 70% | 1.0 ¹ (for example, QoS=100.0%) ² |
| 22 | 4 | 45 | 1.4 (for example, QoS=71.4%) ² |
| 23 | 4 | 5 | 1.4 (for example, QoS=71.4%) ² |

| Performance policy | Priority | Quality of service target | Maximum delay factor |
|--------------------|----------|---------------------------|--|
| 24 | 5 | 45 | 2.0 (for example, QoS=50.0%) ² |
| 25 | 6 | 5 | 2.7 (for example, QoS=37.0%) ² |
| 26 | 7 | 5 | 3.8 (for example, QoS=26.3%) ² |
| 27 | 8 | 5 | 5.3 (for example, QoS=18.8%) ² |
| 28 | 9 | 5 | 7.5 (for example, QoS=13.3%) ² |
| 29 | 10 | 5 | 10.5 (for example, QoS=9.5%) ² |
| 30 | 11 | 5 | 14.7 (for example, QoS=6.8%) ² |
| 31 | 12 | N/A ³ | 20 (for example, QoS=5%) ² |

Table notes:

1. I/Os in this performance policy are not delayed. As such, the maximum delay factor is not applicable. Equivalently, one can say the maximum delay factor is 1.0
2. The QoS index indicated is applicable for cases where the service time is equal to the optimal response time.
3. There is no quality of service target for this performance policy because there are no lower priority performance policies to impact to achieve a quality of service target.
4. There is no priority or quality of service target for this performance policy because by definition, the I/O operations in the default performance group do not impact other I/O operations, no matter what quality of service they receive. However, because they are not delayed, they do in fact tend to reap the benefits of any delays that are introduced to other performance groups to achieve other quality of service targets.

For further information about z/OS WLM, see *z/OS MVS Planning: Workload Management*, SA22-7602, *System Programmer's Guide to: Workload Manager*, SG24-6472, and *ABCs of z/OS System Programming Volume 12*, SG24-7621.

2.4 I/O Priority Manager modes of operation

I/O Priority Manager can operate in the following modes:

- ▶ **Disabled:** I/O Priority Manager does not monitor any resources and does not alter any I/O response times.
- ▶ **Monitor:** I/O Priority Manager monitors resources (ranks) and updates statistics that are available in performance data. This performance data can be accessed from the DS CLI or the GUI. No I/O response times are altered.
- ▶ **Manage:** I/O Priority Manager monitors resources (ranks) and updates statistics that are available in performance data. This performance data can be accessed from the DS CLI or the GUI. I/O response times are altered on volumes that are in performance groups 6-15 if resource contention occurs.

In both monitor and manage modes, it is possible to have I/O Priority Manager send Simple Network Management Protocol (SNMP) traps to alert the user when certain resources have detected a saturation event. Saturation of a rank occurs when the workload on a rank causes it to operate at or above a certain performance threshold defined by the system. A rank enters a saturation state if it is saturated for five consecutive one-minute samples (five of five samples within a five minute period have detected a saturation condition).

A rank exits a saturation state if it is not saturated for three out of five consecutive one-minute samples. An SNMP trap is sent to the user when a rank enters a saturation state and also every eight hours for a rank remains in saturation. The SNMP report contains the storage facility image (SFI) and rank where the saturation event occurred.

Number of SNMP traps: I/O Priority Manager sends out a maximum of eight SNMP traps per SFI server in a 24-hour period (for CKD and FB).

In manage mode, the I/O Priority Manager feature attempts to optimize the overall I/O performance of the SFI by using user-supplied information to drive policy-based decisions on prioritizing I/O operations.

A given Redundant Array of Independent Disks (RAID) array, or rank, has a certain capability to process I/O operations. When the workload on the rank is well below the maximum capacity of the rank, average I/O response times are typically in the nominal response range. When the workload on the rank approaches the maximum capacity of the rank, I/Os get queued and average response times can be significantly higher than nominal response times. I/O Priority Manager can actively manage the queuing of I/O operations on the rank to give priority to more critical I/O operations. By delaying less critical I/Os, more critical I/Os receive better response times and get more throughput and better access to available processing power.

Users specify a performance group for each logical volume. With z/OS WLM-enabled management, a performance group is assigned to each I/O request no matter where the target I/O request volume is assigned. Each performance group is associated with specific performance policies. Given the assigned performance policies, I/O Priority Manager is able to make decisions about the relative priority of I/O operations and insert a delay in certain stage or destage operations to regulate the workload on the rank.

I/O operations are *only delayed in the event that a rank is experiencing a deviation from its normal I/O operation performance*. The choice of which I/O operations to impact and the extent to which they are impacted is based on the expected workload behavior associated with the performance policy.

Important: I/O operations are impacted only when there is already a situation where the set of I/O operations in progress on a given RAID array is already causing a deviation from normal I/O operation performance. Additionally, the I/O operations that are considered for such impact are limited to those involving the RAID arrays that are experiencing a performance deviation.

Throttling takes place when a higher performance policy volume, or I/O request, is not meeting its QoS target, when there is rank saturation, or when there is device adapter (DA) saturation. I/O operations with a specific priority are impacted only to improve the QoS of a higher priority I/Os. The throttling impact to an I/O operation depends proportionately to its priority: the lower the priority, the more aggressive the throttling. I/Os with priority 1 and 0 (default priority) are never delayed. Throttling is done on a rank by rank basis, so it is possible for certain ranks to be impacting I/Os while other ranks are not.

If a rank or DA is close to saturation, I/O Priority Manager delays the low performance policy I/Os up to the maximum delay factor to try and meet the QoS targets of high performance policy I/Os. If delaying low performance policy I/Os is not enough, I/O Priority Manager adds delay to the medium performance policy I/Os, up to its maximum delay factor, in an attempt to help high priority I/Os meet their QoS target. I/Os from the medium and low priority groups are not delayed indefinitely. The maximum added delay to an I/O is 200 ms.

Important: Because there is a maximum delay that can be added to the I/Os, there is no guarantee that a performance group meets its QoS target.

2.5 I/O Priority Manager with Easy Tier

I/O Priority Manager and Easy Tier, if both enabled, provide independent benefits:

- ▶ I/O Priority Manager attempts to make sure that the most important I/O operations get serviced when a given rank is overloaded by the workload on the system by delaying less important I/Os. It does not move any extents.
- ▶ Easy Tier (automatic mode) attempts to locate allocated extents on the storage tier that are most appropriate for the frequency of host access. This is done to try and maximize the throughput of the most active extents. Easy Tier also relocates extents between ranks within a storage tier in an attempt to distribute the workload evenly across available ranks to avoid rank overloading.

Together, these functions can help the various applications running on DS8000 systems meet their respective service levels in a simple and cost effective manner. The DS8000 can help address storage consolidation requirements, which in turn helps to manage increasing amounts of data with less effort and lower infrastructure costs.



Using I/O Priority Manager

This chapter shows how to enable and activate the DS8000 I/O Priority Manager (IOPM), how to select a mode of operation, and how to assign volumes used by specific applications into separate I/O priority levels and priority groups.

3.1 Licensing

Support for DS8000 I/O Priority Manager is an optional feature for DS8700 model 941 and DS8800 model 951. It is available with the DS8000 microcode level 6.6.1.xx (DS8700) or 7.6.1.xx (DS8800) for FB support. The DS8000 microcode level 6.6.20.xx (DS8700) or 7.6.20.xx (DS8800) includes CKD support. It requires I/O Priority Manager licensed feature indicator DS8000 Function Authorization (239x-LFA), I/O Priority Manager feature number 784x. However, monitoring is available without purchasing a license code.

For System z, z/OS release V1.11 or greater is required. Ficon or High Performance IBM FICON® for System z (zHPF) is also required. The following APARs are needed:-

- ▶ For z/OS V1.11, OA32298, OA34063, OA34662
- ▶ For z/OS V1.12, OA32298, OA34063, OA34662
- ▶ For z/OS V1.13, OA32298
- ▶ IBM RMF™ APAR OA35306

3.2 Activating I/O Priority Manager

Like many other DS8000 licensed functions, activating the license keys can be done after the IBM service representative has completed the storage complex installation. Based on your 239x licensed function order, you need to obtain the necessary key from the IBM Disk Storage Feature Activation (DSFA) website at the following address:

<http://www.ibm.com/storage/dsfa>

Before connecting to the IBM DSFA website to obtain your feature activation codes, ensure that you have the following items:

- ▶ The IBM License Function Authorization documents. If you are activating codes for a new storage unit, these documents are included in the shipment of the storage unit. If you are activating codes for an existing storage unit, IBM sends the documents to you in an envelope.
- ▶ A USB memory device can be used for downloading your activation codes if you cannot access the DS Storage Manager from the system that you are using to access the DSFA website. Instead of downloading the activation codes in softcopy format, you can also print the activation codes and manually enter them using the DS Storage Manager GUI. However, this is slow and can be error-prone because the activation keys are 32-character long strings.

Refer to *IBM System Storage DS8000: Architecture and Implementation*, SG24-8886, for details about how to obtain the necessary information for the DSFA website and how to apply the license keys.

You can use DS CLI `lskey IBM.2107-xxxxxx1` command to verify that you have correctly applied the license key for the I/O Priority Manager feature, as illustrated in Example 3-1.

Example 3-1 Checking DS8000 license code

```
dscli> lskey IBM.2107-1300281
Activation Key                               Authorization Level (TB) Scope
=====
I/O Priority Manager                          18.5                          All
Operating environment (OEL)                  18.5                          FB
```

Alternatively, you can use the DS GUI to display the licence key information by selecting **Storage Image** → **Add Activation Key**, as shown in Figure 3-1.

Activation keys information:

| Type | Activation Code |
|-----------------------------|---------------------------------------|
| Operating environment (OEL) | E3E4-EDD7-E8D8-CC8E-D8EC-F2FE- |
| Point in time copy (PTC) | 73C6-0D32-3E2D-C06F-D8EC-F2FE- |
| FlashCopy SE | 55DF-9392-595B-622E-D8EC-F2FE- |
| Metro Mirror | 7ECB-41E8-6387-B7C2-D8EC-F2FE- |
| Global Mirror | C341-29A0-E8C7-73CB-D8EC-F2FE- |
| Metro/Global Mirror (MGM) | DFBE-6B5F-9BCB-813A-D8EC-F2FE- |
| Thin Provisioning | 5BF9-7AC1-420C-B5D3-D8EC-F2FE- |
| Easy Tier | 18D5-1183-6887-2B60-D8EC-F2FE- |
| I/O Priority Manager | B556-2827-CE2E-482A-D8EC-F2FE- |

Figure 3-1 GUI display of license key

Tip: If you just want to monitor the performance groups, then you do not need the I/O Priority Manager licensed feature indicator. If you want to manage the performance groups, then you need the I/O Priority Manager licensed feature indicator.

3.3 z/OS WLM settings to activate I/O Priority Manager

In the Workload Manager (WLM) Service Definition, the service option “I/O priority management” has to be set to Yes. Refer to Figure 3-2 for the setting in “Option 8. Service Coefficients/Server Definition Options” of the WLM definition Interactive System Productivity Facility (ISPF) panel.

| Coefficients/Options | Notes | Options | Help |
|---|-------|---------------|------------------|
| ----- Service Coefficient/Service Definition Options | | | |
| Command ==> | | | |
| Enter or change the Service Coefficients: | | | |
| CPU | | <u>1.0</u> | (0.0-99.9) |
| IOC | | <u>0.1</u> | (0.0-99.9) |
| MSO | | <u>0.1000</u> | (0.0000-99.9999) |
| SRB | | <u>1.0</u> | (0.0-99.9) |
| Enter or change the service definition options: | | | |
| I/O priority management | | <u>YES</u> | (Yes or No) |
| Dynamic alias tuning management | | <u>YES</u> | (Yes or No) |

Figure 3-2 WLM Setting for IOPM

The SYS1.PARMLIB parmlib member IEAOPTxx must contain the following new statement:

```
STORAGESERVERMGT=YES
```

This settings can be activated dynamically using the **SET OPT=xx IBM MVS™** command.

3.4 Setting I/O Priority Manager mode using DS CLI

A new parameter, **-iopmode**, is available with the **chsi** command to let you enable and control the Priority Manager operation mode.

You must specify one of the following values with the parameter:

| | |
|--------------------|--|
| disable | Disables I/O Priority Manager. |
| monitor | Monitors resources associated with performance groups that specify I/O management, but does not manage them. |
| monitorsnmp | Same as monitor with added SNMP trap support. When a rank is saturated, SNMP alerts are sent out to the monitor server. |
| manage | Manages resources associated with performance groups that specify I/O management. |
| managesnmp | Same as manage with added SNMP trap support. When a rank is saturated, SNMP alerts are sent out to the monitor server. |

License requirement: The **manage** and **managesnmp** modes require LIC feature code.

Example 3-2 illustrates the use of DS CLI **chsi** command to change I/O Priority Manager to manage mode.

Example 3-2 Changing I/O Priority Manager to managed mode

```
dsccli> chsi -iopmode manage IBM.2107-1300281
CMUC00042I chsi: Storage image IBM.2107-1300281 successfully modified.
```

You can run **shows** to display current I/O Priority Manager mode, as illustrated in Example 3-3.

Example 3-3 Displaying I/O Priority Manager mode

```
dsccli> shows IBM.2107-75TV181
numegsupported    0
ETAutoMode        tiered
ETMonitor         all
IOPMmode        Managed
```

3.5 Setting I/O Priority Manager mode using DS GUI

In the DS GUI, hover the mouse over the **Home** icon and click **System Status** as illustrated in Figure 3-3.

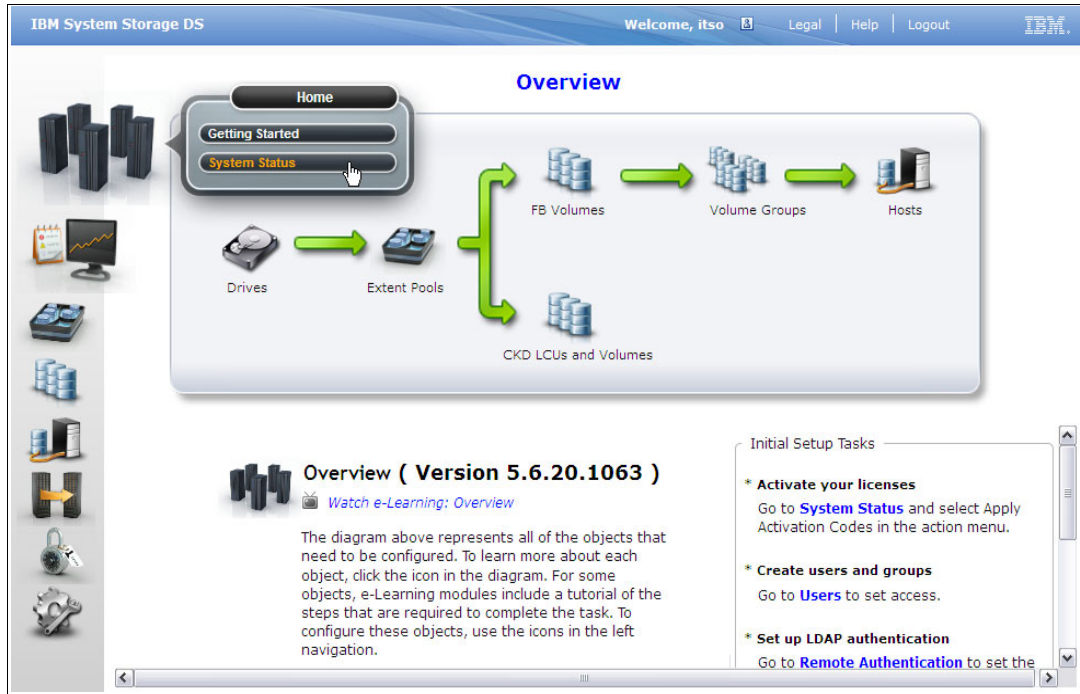


Figure 3-3 Selecting System Status

In the System Status panel, highlight the storage image, and then select **Action** → **Storage Image** → **Properties** from the drop-down menu. See Figure 3-4.

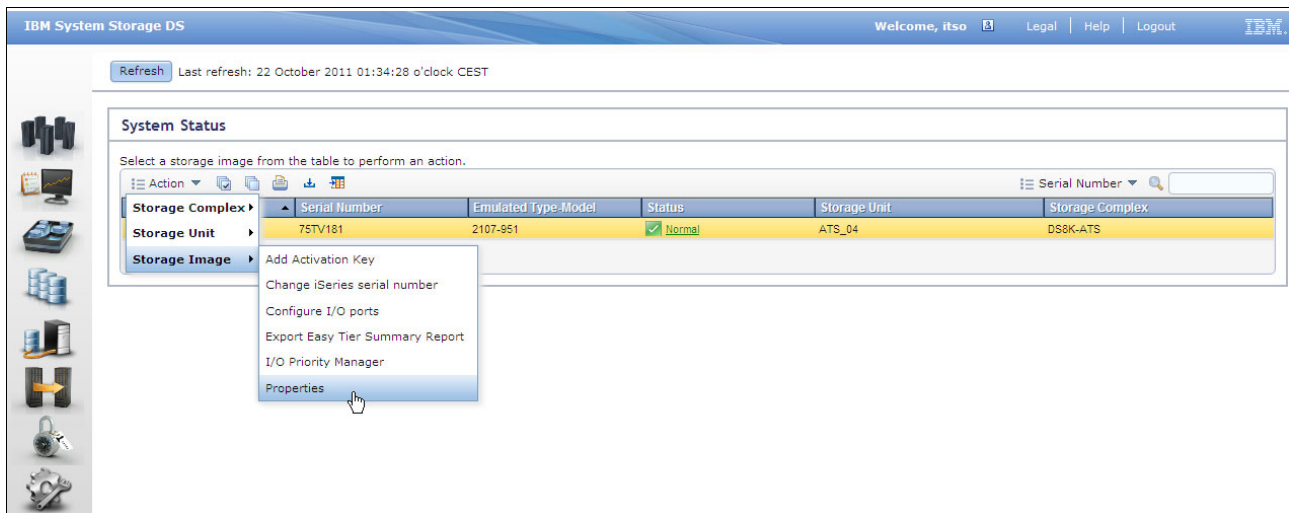


Figure 3-4 Selecting Properties

In the Storage Image Properties dialog, click the **Advanced** tab. In the **Advanced** tab window, you can select one of the three choices under I/O Priority Manager, as shown in Figure 3-5. If you select **Monitor** or **Manage** mode, you can also select the **Send SNMP Traps** check box if you want to send SNMP traps when a rank is saturated.

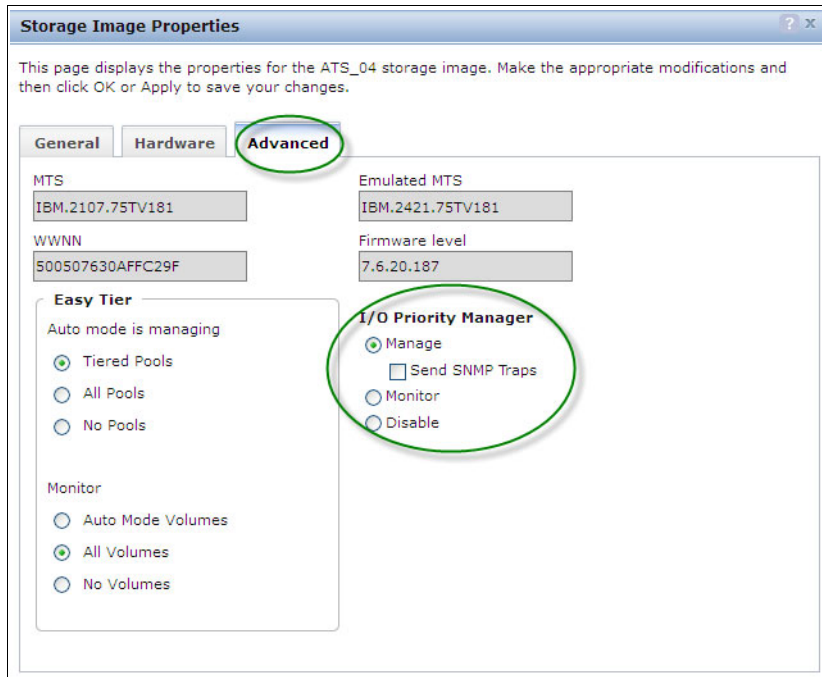


Figure 3-5 Selecting Advanced tab and I/O Priority Manager mode

3.6 Assigning I/O performance groups to FB volumes using DS CLI

The DS CLI `mkfbvol` command that is used to create a FB volume includes a new parameter `-perfmode` that lets you specify a performance group (PGx) for the volume. See Table 2-1 on page 8 for a list of the performance group numbers and their corresponding priority levels. If the `-perfmode` parameter is not specified, the new volume defaults to performance group 0 (PG0).

Example 3-4 shows that we did not include `-perfgrp PGx` in the command, and, as a result, the volume defaults to PG0, which is *not* a managed performance group.

Example 3-4 Creating a FB volume

```
dsccli> mkfbvol -extpool p2 -cap 5 -type ds -name ITS0-#h 0013
CMUC00025I mkfbvol: FB volume 0013 successfully created.
```

Example 3-5 shows the **showfbvol** command that is used to check the I/O performance group of the volume that was just created.

Example 3-5 Displaying I/O performance group information for FB volume using showfbvol

```
dsccli> showfbvol 0013
Name           ITS0-0013
ID             0013
...
perfgrp      PG0
migratingfrom -
resgrp         RGO
```

You can change the I/O performance group of a logical unit number (LUN) using the **chfbvol** command with the **-perfgrp** parameter to specify a new performance group. In Example 3-6, we change LUN number 0010 to performance group 3.

Example 3-6 Changing the I/O performance group of a LUN to PG3

```
dsccli> chfbvol -perfgrp pg3 0010
CMUC00026I chfbvol: FB volume 0010 successfully modified.
```

You can then use **showfbvol** command to verify the change.

3.7 Assigning I/O performance groups to FB volumes using DS GUI

A new Performance Group drop-down menu has been added to the **Fixed Block Volume Creation** window (**Add Volumes**), as shown in Figure 3-6, to allow selection of a performance group when creating a volume.

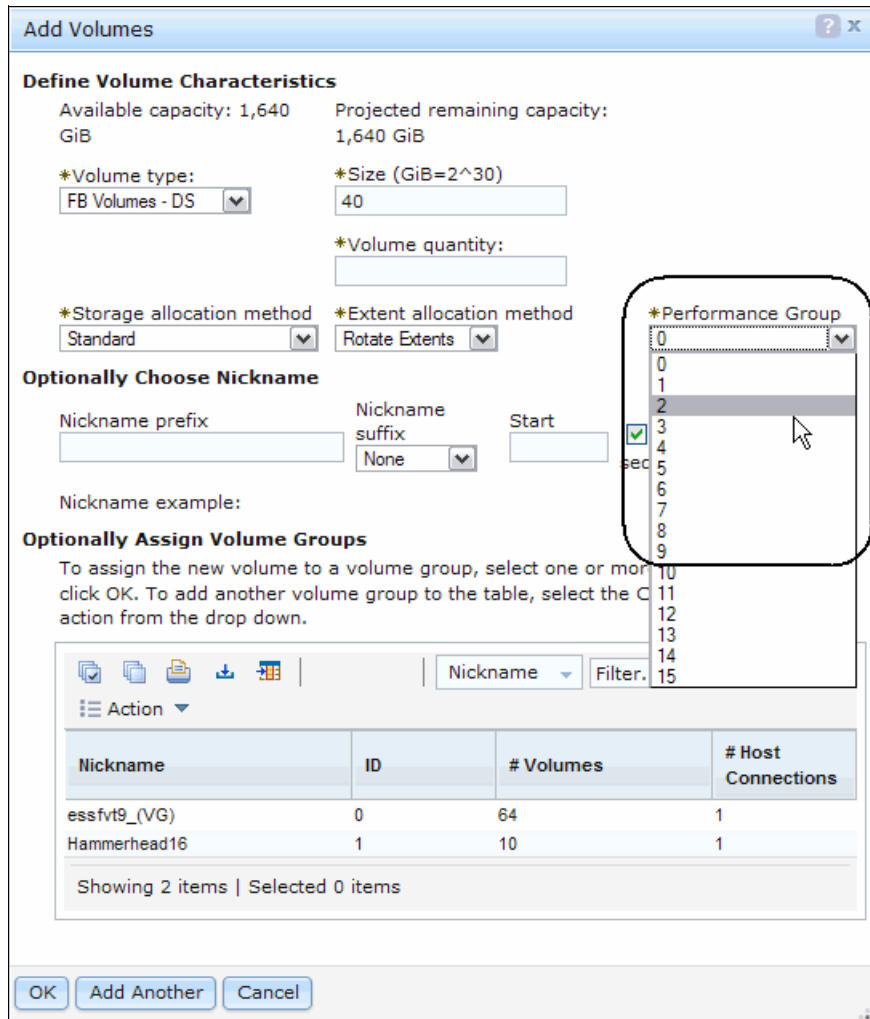


Figure 3-6 Creating a FB volume specifying performance group

To change the performance group previously assigned to a FB volume, select **Manage Volumes** from the **DS GUI Welcome** window. In the **Manage Volumes** window, select the FB volume that you want to change. Then, select **Action** → **Properties**, as shown in Figure 3-7.

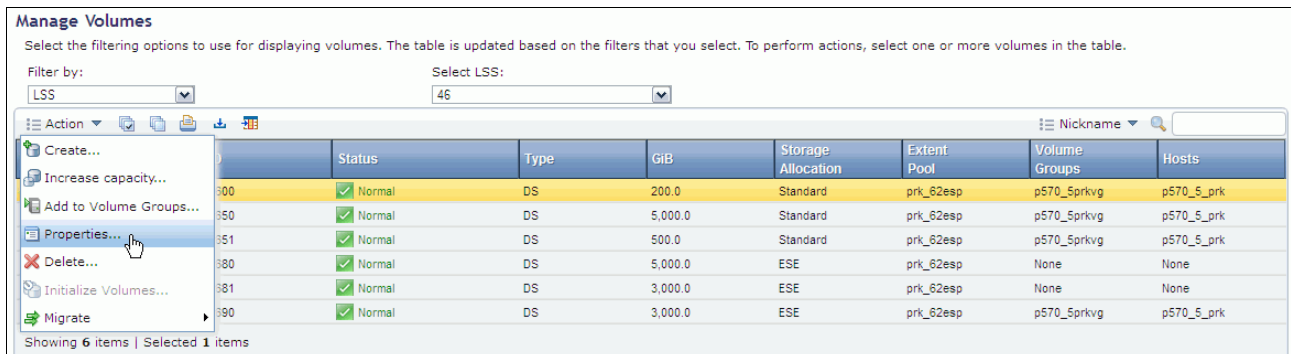


Figure 3-7 Selecting FB volume properties to change performance group of FB volume

In the **Single Volume Properties** panel, a new **Performance Group** drop-down menu allows you to reassign a performance group for the volume. See Figure 3-8.

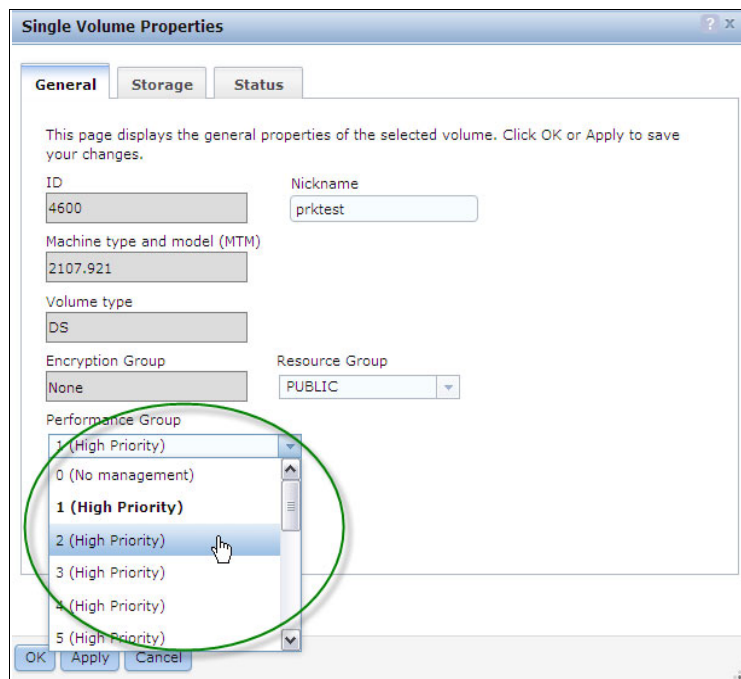


Figure 3-8 Changing FB volume performance group

3.8 Assigning I/O performance groups to CKD volumes using DS CLI

The DS CLI **mkckdvol** command that is used to create a CKD volume includes a new parameter “**-perfgrp performance_group_ID**” that lets you specify a performance group (PGx) for the volume. See Table 2-2 on page 10 for a list of the performance group numbers and their corresponding priority levels. If the **-perfgrp** parameter is not specified, the new volume defaults to performance group 0, PG0.

Example 3-7 shows how to create a Mod1 CKD volume with a performance group 22 (PG22 with medium priority).

Example 3-7 Creating a CKD volume

```
dsccli>mkckdvol -extpool P0 -cap 1113 -name ckd_p0_7000 -perfgrp PG22 7000
CMUC00021I mkckdvol: CKD volume 7000 successfully created.
```

Example 3-8 shows how to use the **showckdvol** command to check the I/O performance group of the volume that was just created.

Example 3-8 Checking I/O performance group of volume

```
dsccli> showckdvol 7000
Name          ckd_p0_7000
ID            7000
...
cap (cyl)     1113
...
perfgrp     PG22
migratingfrom -
resgrp        RG0
```

You can change the I/O performance group of a volume using the **chckdvol** command with the new **-perfgrp** parameter to specify a new performance group. In Example 3-9, we change the volume number 7000 to performance group 19 (high priority).

Example 3-9 Changing the I/O performance group of a CKD volume to PG19

```
dsccli> chckdvol -perfgrp pg19 7000
CMUC00022I chckdvol: CKD Volume 7000 successfully modified.
```

You can then use **showckdvol** command to verify the change.

3.9 Assigning I/O performance groups to CKD volumes using DS GUI

A new **Performance Group** drop-down menu has been added to the CKD **Create Volumes** window, as shown in Figure 3-9, to allow selection of a performance group when creating a volume.

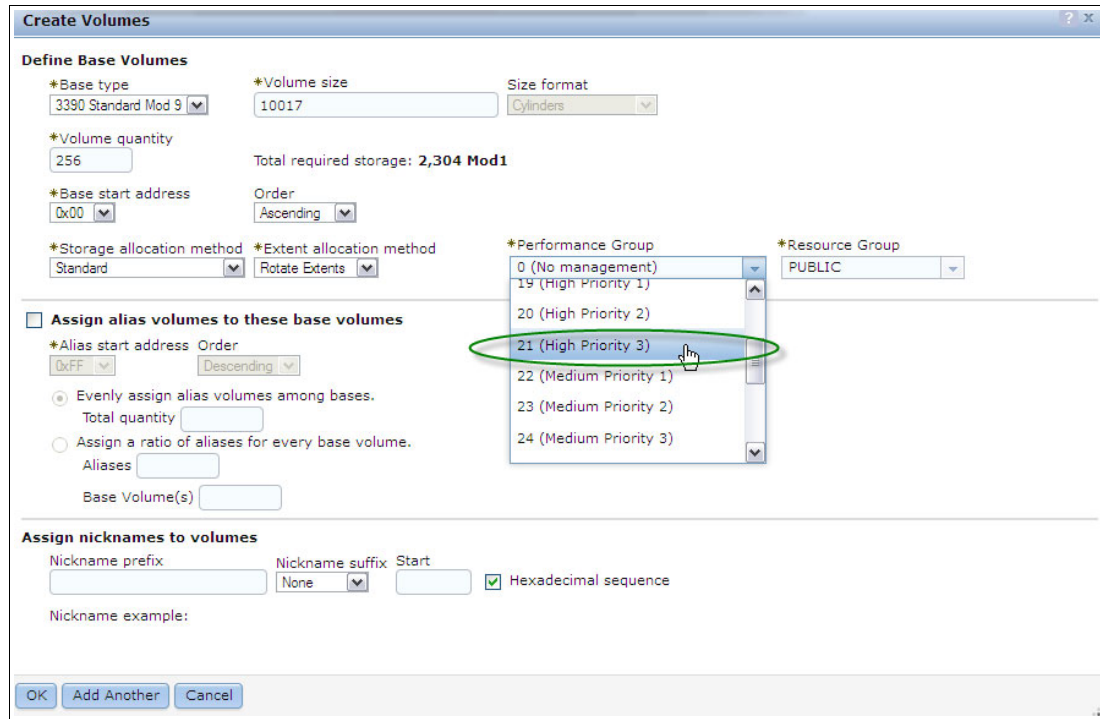


Figure 3-9 Creating a CKD volume with performance group

To change the performance group previously assigned to a CKD volume, select **CKD LCUs and Volumes** from the **DS GUI Welcome** window. In the **Manage CKD Volumes** window, click **Manage existing LCUs and volumes**, and select the CKD volume that you want to change. Then select **Action** → **Properties**, as shown in Figure 3-10.

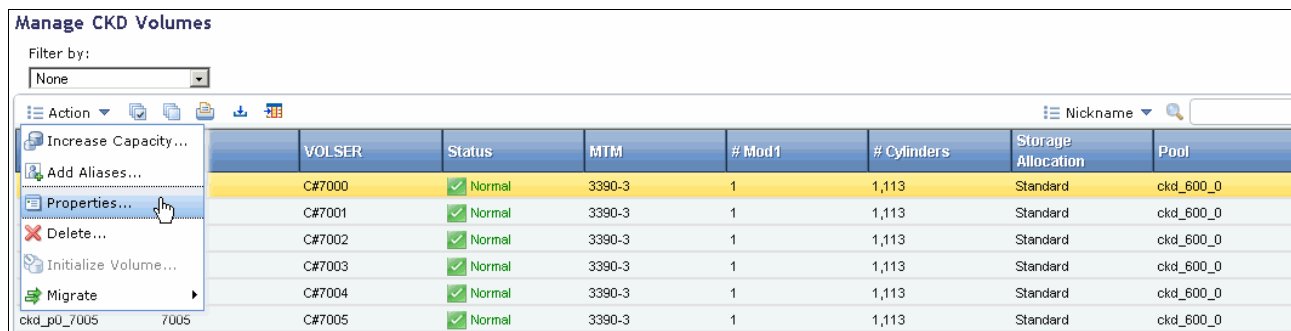


Figure 3-10 Selecting CKD volume properties to change performance group of a CKD volume

In the **Single Volume Properties** panel, a new **Performance Group** drop-down menu allows you to reassign a performance group for the volume. See Figure 3-11.

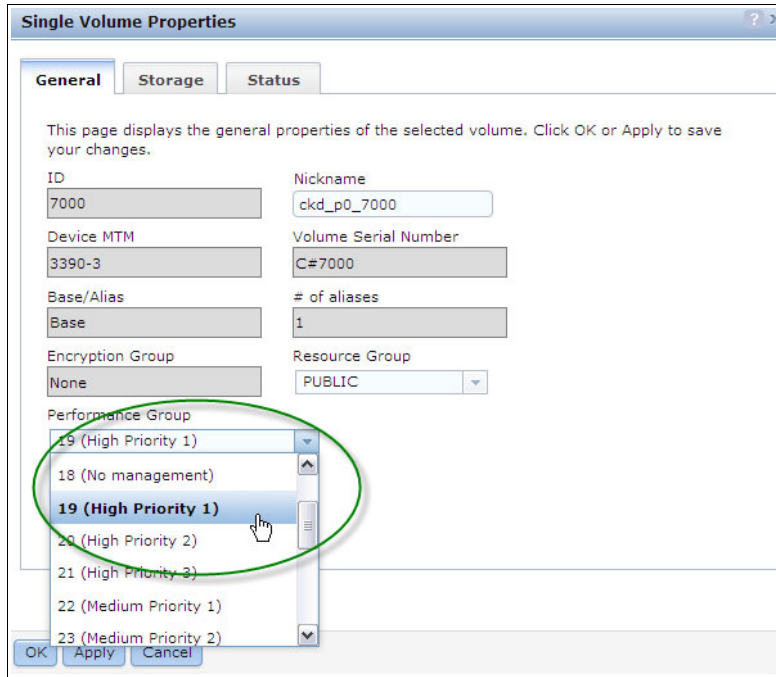


Figure 3-11 Changing CKD volume performance group



I/O Priority Manager usage scenarios

This chapter demonstrates the effect of I/O Priority Manager (IOPM) policy usage on I/O activity level and response times.

We consider two separate scenarios. In the first scenario, we test the I/O Priority Manager in an open systems environment where we use three groups of volumes connected to a Linux server, with a separate workload for each group. We present the statistics from the server side and show the effect of volume priority changes on each workload performance level.

In the second scenario, we test the I/O Priority Manager in a z/OS environment with Workload Manager (WLM) support.

4.1 FB scenario: Architecture environment

For the first scenario, we test the I/O Priority Manager in the environment depicted in Figure 4-1:

- ▶ One SUSE Linux Enterprise Server (SLES) v11.1, 64-bit server with Fibre Channel (FC) attachments
- ▶ One DS8800 with R6.1 code (LIC Version 7.6.10.507), 28 ranks (4 x 300 GB SSD, 6 x 146 GB / 15K SAS, 22 x 146 GB / 15 K SAS)
- ▶ One extent pool (P4) composed of three RAID 5 ranks with 146 GB / 15K SAS disks (R19, R20, and R21)
- ▶ Nine volumes (ID 4000 to 4008) of 100 GB each, defined on extent pool P4 with rotate extents using extent allocation method (EAM) and mapped to the Linux server
- ▶ FC SAN configuration to allow, for each volume, access over four separate paths from the server to the DS8800, managed by Linux native multipath (device mapper).

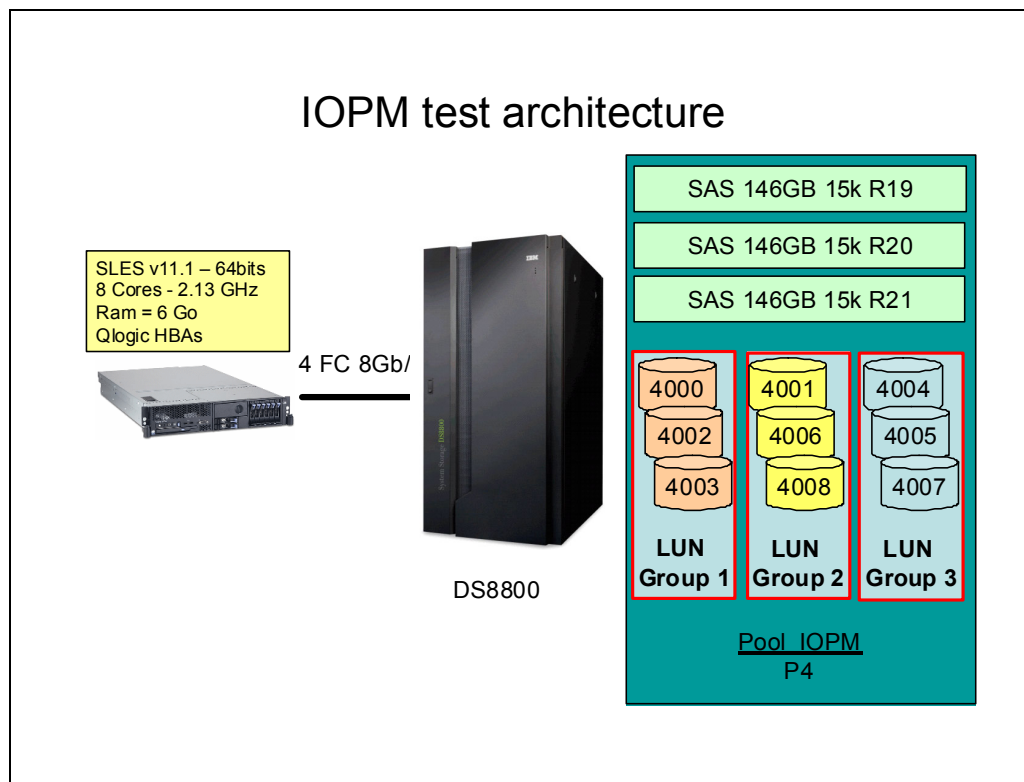


Figure 4-1 IOPM policy usage test architecture

We have groups of three LUNs each, and each LUN group is set to a separate performance group to allow separate statistics at the DS8800 level.

Let us assume, for this example, that the following associations are established:

- ▶ LUNs in the first group are associated to a business critical production application.
- ▶ LUNs in the second group are associated to a development environment.
- ▶ LUNs in the third group are associated to a test application.

Using an I/O simulator, a workload comparable to online transaction processing (OLTP) (70% read, 50% hit ratio, 8K blocks) is run on each LUN group, with a load factor that we can vary separately during the test.

Linux raw devices: The I/O simulation tool is applied on Linux “raw devices” to avoid any server cache effect. Raw device is a UNIX concept that allows write I/Os to bypass the server’s cache and write or read directly to/from the LUNs. It was often used for large databases in older UNIX releases.

4.1.1 FB scenario: Test description

As depicted in Table 4-1, scenario 1, test 1 starts with all volumes (PG1, PG2, and PG3) using the same performance policy (High Priority Performance Policy) and with an identical load factor (LF). This means that we apply the same level of load on all groups of LUNs, and because all LUNs have the same performance policy, I/O Priority Manager does not try to optimize the I/Os on the separate LUNs in the extent pool.

Performance policies: Performance policies are described in Chapter 4, “I/O Priority Manager usage scenarios” on page 33.

Table 4-1 IOPM policy usage test overview

| Test number | LUN group 1 | LUN group 2 | LUN group 3 |
|---------------------|---------------|--------------|----------------|
| 1 | PG 1 LF* 8 | PG 2 LF 8 | PG 3 LF 8 |
| 2 | PG 1 LF 8 | PG 2 LF 8 | PG 3 LF 24 |
| 3 | PG 1 LF 8 | PG 6 LF 8 | PG 11 LF 24 |
| 4 | PG 1 LF 8 | PG 6 LF 8 | PG 11 / |
| 5 | PG 1 LF 24 | PG 6 LF 8 | PG 11 / |
| 6 | PG 1 LF 24 | PG 2 LF 8 | PG 3 / |
| * LF -= Load factor | | | |

For scenario 1, test 2, the load factor is tripled on the third LUN group, which in practice means that someone has started to overload the test environment. Because the priority groups (PGs) have not been changed, IOPM still does not try to optimize the I/Os, and both production and development environments can suffer from the test environment overload and increase in their response times.

At the beginning of scenario 1, test 3, we modify the performance policy for the LUN groups to preserve the higher priority (PG1) for LUN group 1, give a medium priority (PG6) to LUN group 2, and a low priority (PG11) to LUN group 3. The load factor is the same as in test 2. In this situation, if there is saturation on the ranks of the LUNs, IOPM adds delay to the low PG LUNs and, if necessary, on the medium PG LUNs to reduce the saturation of the rank and thus give the high PG LUNs a better performance level.

In scenario 1, test 4, we stop the load on the third LUN group, which means that the run on the test environment has reached an end. No change is made at the priority group level, which means that, like in test 3, IOPM only applies a delay to medium priority LUNs if the ranks reach saturation. However, this might not be the case here due to the global load decrease. Even in such a case, it might still take time to return to a situation where IOPM does not influence the workload at all.

In scenario 1, test 5, without changing the PGs, we increase the load factor from 8 to 24 on the first LUN group to simulate a workload peak on the main production application. If the ranks are saturated because of the new workload, IOPM adds a delay to the medium priority LUNs on the same ranks to help the high priority LUNs to reach a better performance level.

Finally, in scenario 1, test 6, we change the PGs on the LUN groups to return to the startup situation, with all LUNs in high priority groups. This last step allows a comparison with the previous situation and thus shows how IOPM can help to guarantee the highest performance level to the most critical applications.

Frequency of IOPS monitoring: When activated, IOPM checks I/O operations per second (IOPS) thresholds every 20 seconds, which means that it can react quickly to any workload change if needed.

4.1.2 FB scenario: test results

In this section, we examine the statistics gathered during our tests.

Linux server statistics

Figure 4-2 shows input/output operations per second (IOPS) statistics at the Linux server during the test series. The picture has been divided into six parts, corresponding to the six tests depicted in Figure 4-2. Test numbers are shown at the bottom of the picture.

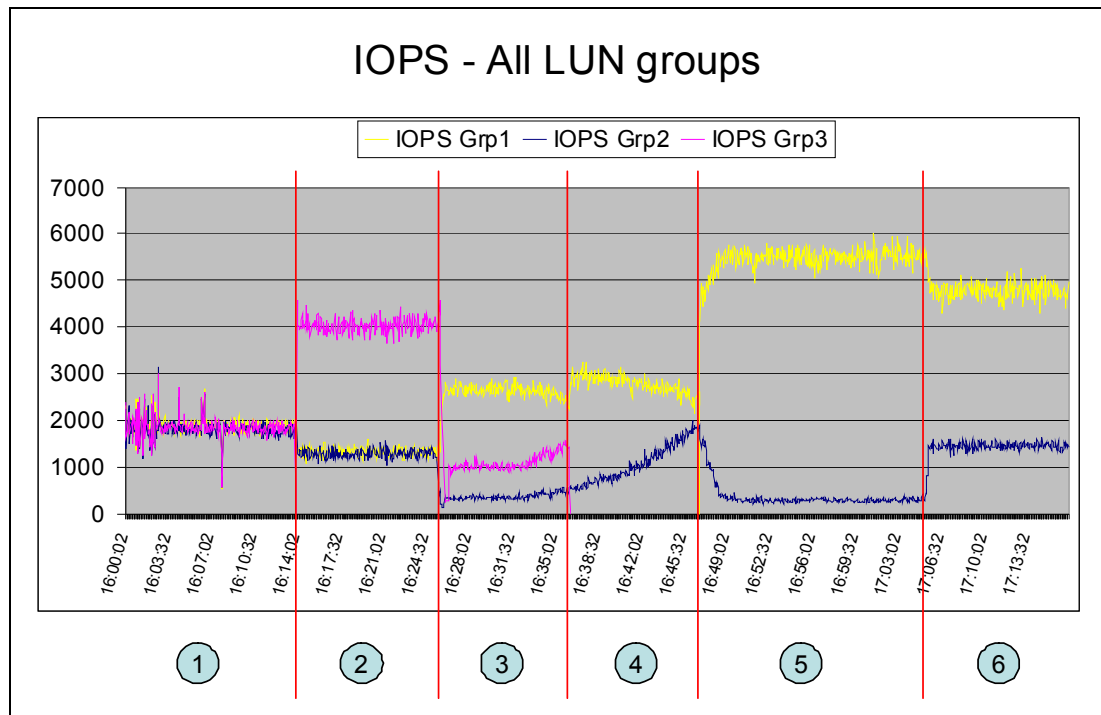


Figure 4-2 IOPS statistics on the Linux server

For test no. 1, all LUN groups roughly maintain the same IOPS level. When test 2 starts, IOPS decrease for LUN groups 1 and 2 due to the increased workload on LUN group 3. This typically means that the ranks (for those LUNs) reach a saturation level, but as IOPM only has to manage high priority LUNs at that moment, it does not influence (delay) the I/Os.

Soon after test 3 has started with the changed PGs, IOPM reacts to saturation of the ranks and adds delay to both medium priority and low priority LUN I/Os. The effect is almost immediate, and the IOPS level increases on the LUNs in PG1 (even better than at the beginning of the entire test), while it decreases on both PG6 and PG11 LUNs. Because the load is three times higher on low priority LUNs, they show more IOPS than the medium priority LUNs, but with a ratio a bit less than 3:1. This change is due to the fact that the delay added by IOPM is a bit smaller on the medium priority group compared to the lower priority group.

During test 3, we can also see that IOPM constantly readjusts the delay if it detects that the rank saturation is not reached after the first change. This is why, as time passes, we see the IOPS decrease slowly on PG1 LUNs, while it slowly increases on both PG6 and PG11 LUNs.

We can observe the same kind of behavior during test 4, when workload has been stopped on the third LUN group. After an initial limited increase, the IOPS level on LUN group 1 slowly decreases as it increases for LUN group 2, due to the reduction of the delay applied on the medium priority LUNs. In this case, it seems that the ranks are not saturated anymore, but it takes time for IOPM to return to a situation where the I/Os are not delayed to favor the high priority LUNs.

During test 5, as workload is tripled on the first LUN group, saturation occurs on the ranks, and IOPM helps the high priority LUNs by applying a delay on medium priority LUNs to try to minimize the saturation level. In this case, we see that the delay is maintained throughout the test, which means that IOPM had detected that the rank saturation was quite high.

When the LUNs are returned to their original PGs in the test 6, IOPM stops helping the I/Os as all of them are now in high priority groups. The result is that the IOPS level on the first LUN group decreases and is now three times higher than that of the second LUN group, which has increased because delay has disappeared. In this case, the IOPS level is directly representative of the workload level applied on the LUNs.

Figure 4-3 shows the bandwidth statistics (MBps) for the Linux server during the test series. The curves are similar to the one for IOPS because the block size is constant (8K) during the test scenario.

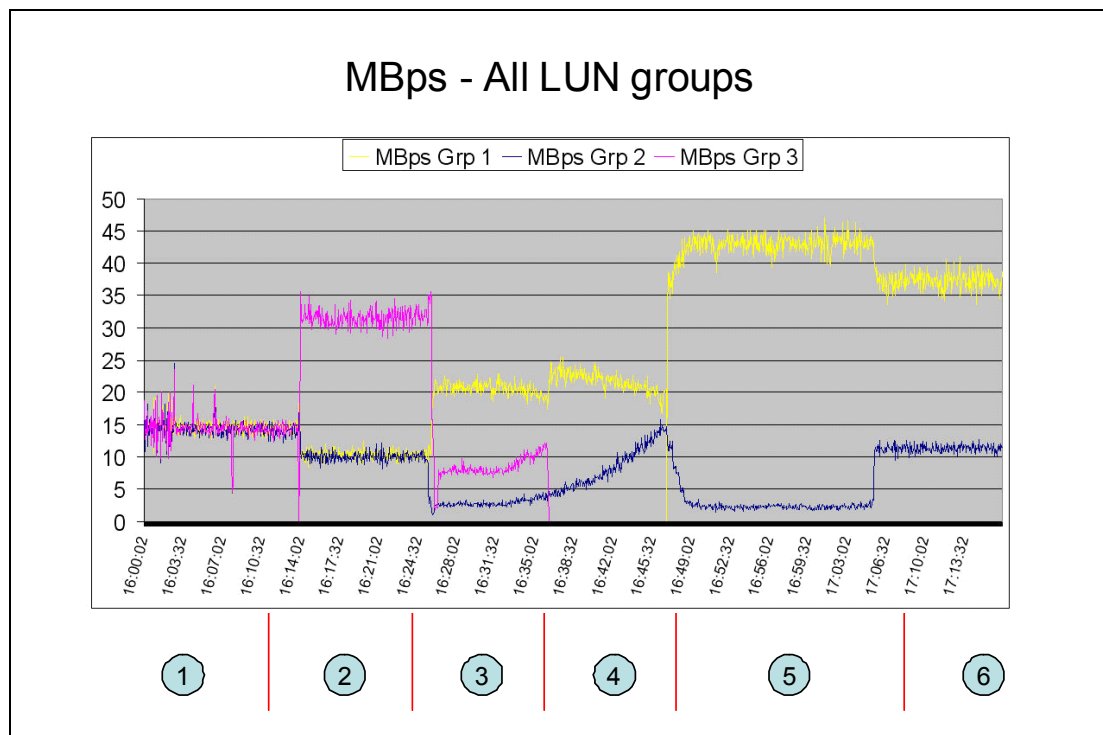


Figure 4-3 Bandwidth statistics on the Linux server

Figure 4-4 shows the service time statistics, in milliseconds, on the Linux server throughout the test.

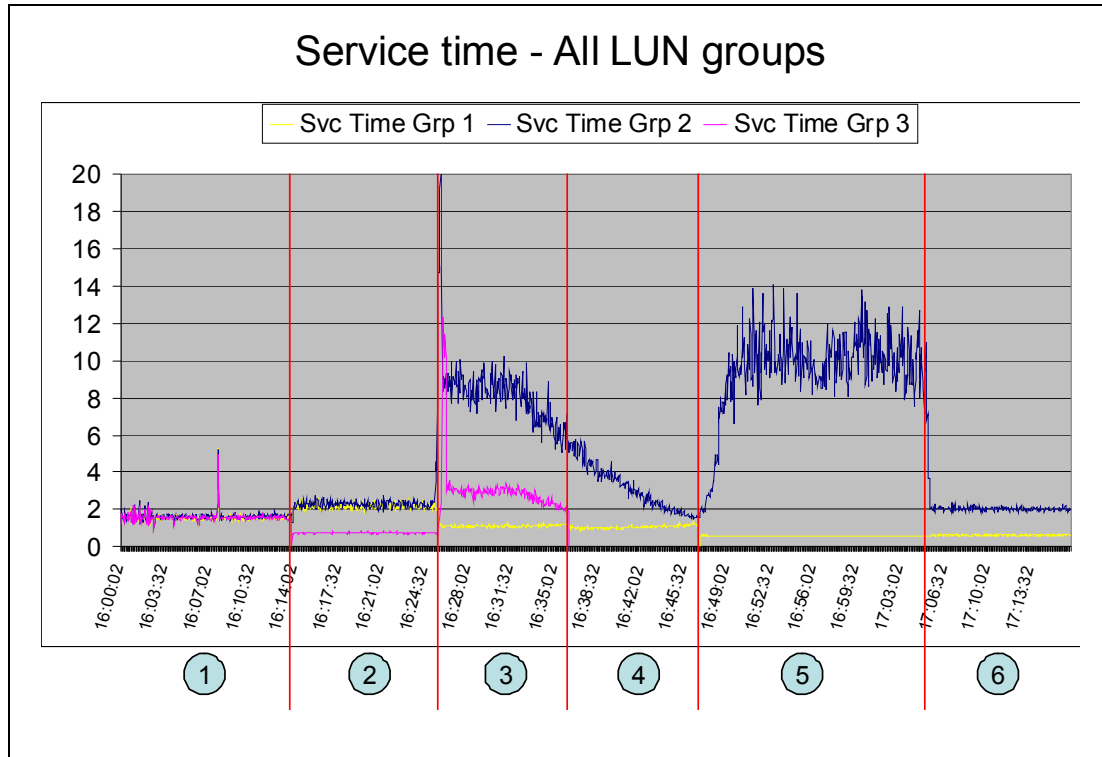


Figure 4-4 Service time statistics on the Linux server

It is interesting to see that, during tests 3 and 4, even if the delay applied on medium priority and low priority LUN groups decreases significantly (and thus the service time on the server side as well), it does not affect the service time of the high priority group LUNs. This is because IOPM smoothly readjusts the delay to avoid any sudden peak that can affect the performance level of the high priority LUNs.

We can also see that the added delay is maintained at a higher level during test 5 because the workload remains high on the ranks, as opposed to test 3, where the added delay on PG11 LUNs decreases the global workload.

DS8000 IOPM statistics

DS8000 IOPM statistics are collected using the `lspperfgrprpt` command, which is explained in 5.1, “Monitoring and reporting with the DS CLI” on page 60. This command allows us to retrieve various metrics at the priority group and rank levels.

Figure 4-5 shows the IOPS statistics for the three LUN groups on the DS8000 during the entire test series.

These statistics are similar to the IOPS server statistics shown in Figure 4-2 on page 37 because IOPM takes all I/Os in consideration, including the ones that cached at the DS8000.

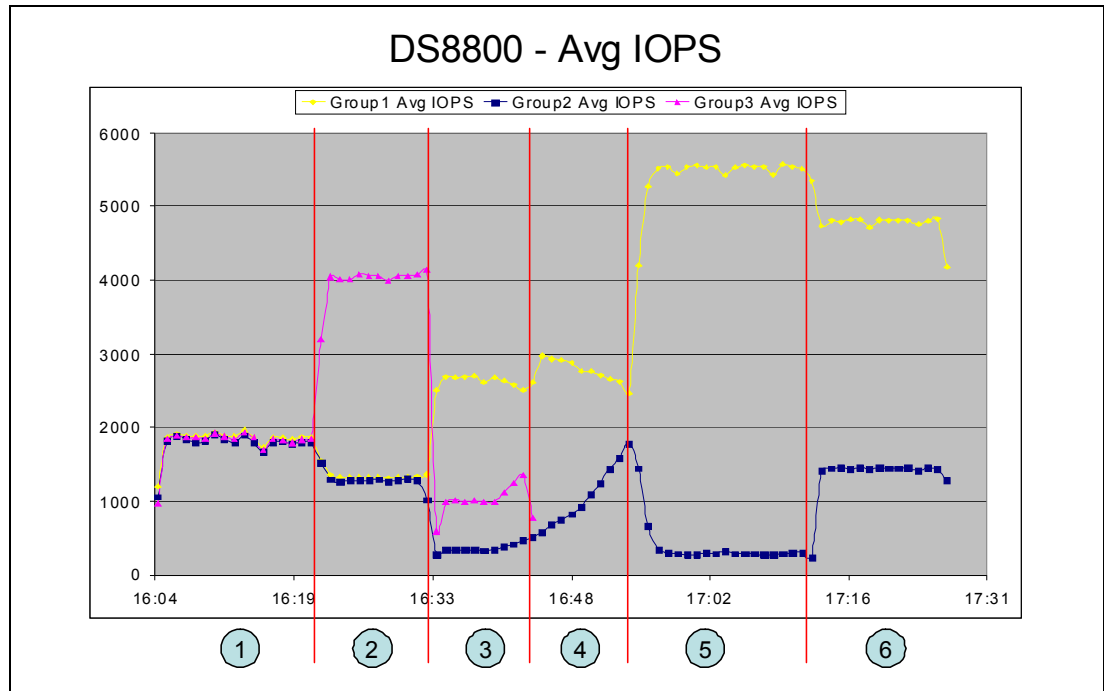


Figure 4-5 Rank IOPS statistics on the DS8800

Figure 4-6 shows the average response time statistics for the three LUN groups during the test series. In this case, these numbers represent the back-end disk service time without considering the cache service time. This data explains the differences in the service times from the server side as shown in Figure 4-4 on page 39, where the I/Os returned from the DS8000 cache tend to decrease the average service time.

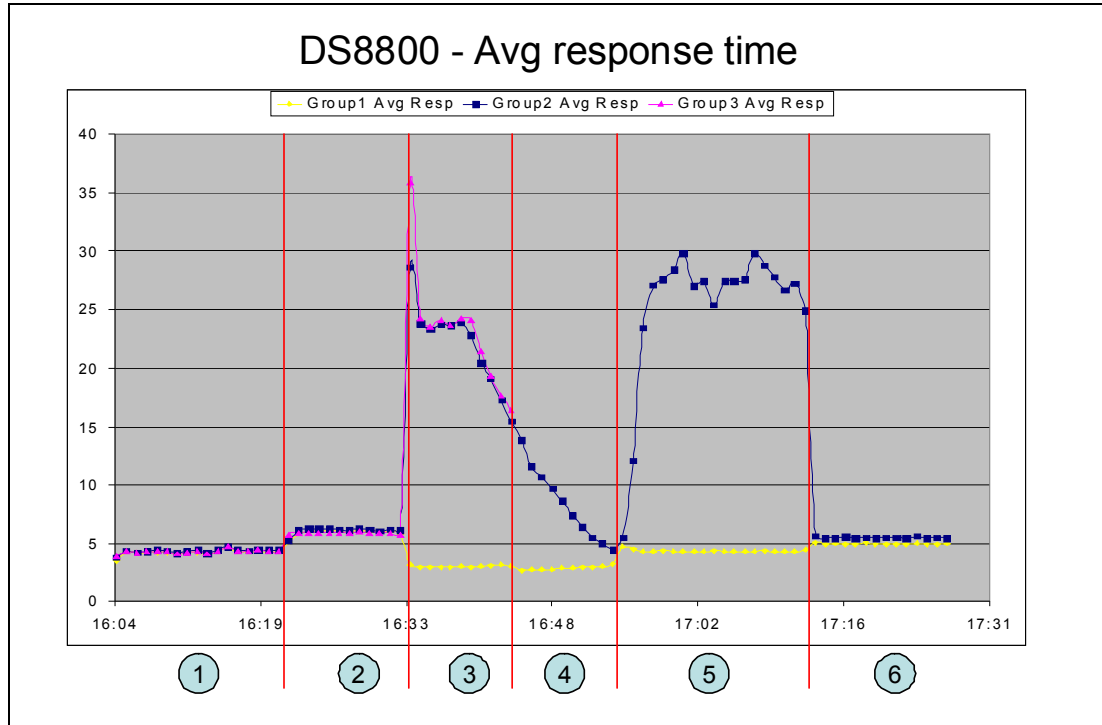


Figure 4-6 Back-end response time statistics on the DS8800

The most noticeable information about this graph is the effect of IOPM delay, either on the medium priority or low priority LUNs, and also how this delay helps the I/Os on the high priority LUNs during tests 3, 4, and 5, particularly compared to the tests 1, 2, and 6. In these latter tests, the back-end response time is the same for the three LUN groups, even if the load varies independently for each group.

Figure 4-7 shows the average Quality of Service (avQoS) statistics for the three LUN groups throughout the test series. This value is known as the “Quality of Service Index,” a percentage computed from the response times for writes, read hits, and read misses. A higher value for this index reflects better performance, and any value above 70% can be regarded as excellent performance. If a PG is being throttled, then the avQoS reflects the effect of throttling.

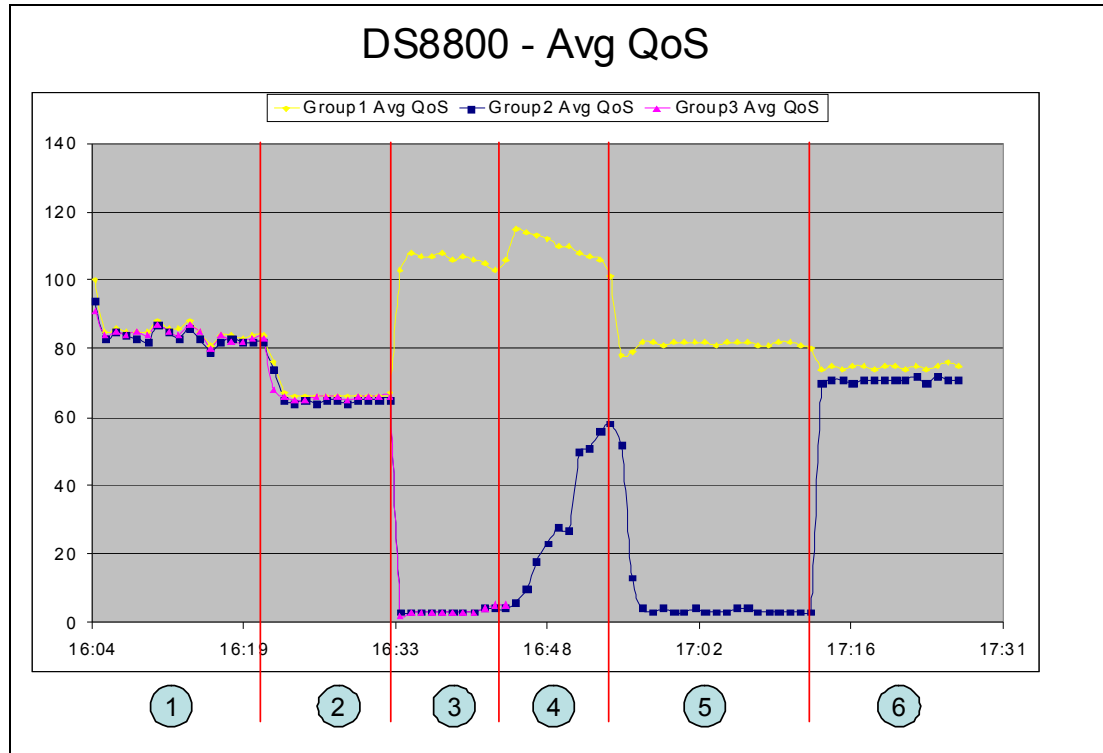


Figure 4-7 Average QoS statistics on the DS8800

The avQoS value is compared to the tgtQoS (QoS target) value, which, for a given PG, is a fixed policy parameter. When avQoS < tgtQoS, an attempt is made to identify other PGs whose I/O can be throttled to help this PG. This throttling only applies to lower priority PGs, and in this case, IOPM tries to reach the QoS values listed in Table 4-2.

Table 4-2 QoS target values

| Performance policy | Name | QoS target | Maximum delay factor |
|--------------------|----------------------|------------|----------------------|
| 0 | No Manage/No Monitor | N/A | N/A |
| 1 | High priority | 70% | N/A |
| 2 | Medium priority | 40% | 5 (QoS = 20%) |
| 3 | Low priority | N/A | 20 (QoS = 5%) |

IOPM algorithms always try to privilege the highest priority performance policy, which means that IOPM does not try to reach the 40% QoS target for the medium policy LUNs if it decreases the avQoS of the high priority LUNs below their QoS target.

In our example, we see that when the workload of the third LUN group increases, IOPM throttles the I/Os on both medium priority and low priority group LUNs to increase avQoS on PG1 LUNs, which is under its 70% objective. In that case, the throttling has a big impact on the avQoS of the medium priority and low priority LUNs, and the IOPM tries to readjust this value during test 3 and especially test 4 by assigning them a much lighter workload.

During test 5, the I/Os are delayed on the medium priority LUN group to keep the avQoS on the high priority LUNs close to their target QoS.

4.2 CKD scenario: Architecture environment

The CKD scenario tests are performed in the following environment:

- ▶ One IBM System z z900 (2064-116) with a LPAR running z/OS v1.12
- ▶ One DS8800 with R6.2 code (LIC Version 7.6.20.185), 24 ranks (2 x 300 GB SSD, 12 x 146 GB / 15K SAS, 6 x 600 GB / 10K SAS, 3 x 3 TB / 7.2 SAS)
- ▶ One extent pool (P0) composed of one RAID 5 rank with 600 GB / 10K SAS disks (R5)
- ▶ Two LCU (SSID 7201 and 7401) each with one 16 volume 3390 model 1 (ID 7220 to 723F and 7400 to 741F). All the volumes have been defined on extent pool P0
- ▶ Four FICON chpid to connect both LCUs.

4.2.1 CKD scenario: test description

In the CKD scenario, we perform three tests. In the first test, we run two batch jobs (WORK1 and WORK2), performing the same workload with the same WLM service class. In the second and third tests, we run the same two jobs as in the first test, but using separate combinations of WLM service classes. The main settings for the WLM service classes used for the tests are summarized in Table 4-3.

Table 4-3 WLM service classes

| Service class name | Importance | Performance goal |
|--------------------|------------|----------------------------------|
| ONLHI | 1 | 90% complete within 00:00:00.300 |
| BATHI | 1 | Execution velocity of 70 |
| BATLO | 3 | Execution velocity of 30 |

Note that the service class ONLHI is not typical for a batch workload because it has a response time goal.

Table 4-4 shows the service class assignment for each job in the tests.

Table 4-4 Job service class assignment

| Test number | Batch jobs | Service class |
|-------------|------------|---------------|
| 1 | WORK1 | BATLO |
| | WORK2 | BATLO |
| 2 | WORK1 | BATHI |
| | WORK2 | BATLO |
| 3 | WORK1 | ONLHI |
| | WORK2 | BATLO |

All the batch jobs run the same PL1 program, creating a workload with the following characteristics:

- ▶ A total of 20,000 random accesses to a Virtual Storage Access Method (VSAM) relative record data set (RRDS). The two batch jobs use two separate target data sets with the same allocation characteristics:
 - WORK1 used the data set TEAM12.RRDS1 allocated in 32 volumes (7220-723F) and loaded with 6 million records.
 - WORK2 used the data set TEAM12.RRDS2 allocated in 32 volumes (7400-741F) and loaded with 6 million records.
- ▶ Read/Write ratio is 5 (one update every five reads)
- ▶ The I/O blocksize is 4 KB
- ▶ The average hit cache measured by DS8000 is 30%

Before starting the tests, all the volumes are assigned to performance group PG16. No other workload that can interfere with the test is running.

4.2.2 CKD scenario: test results

In this section we analyze the results of the tests performed.

Test 1

This test shows that the I/O operations performed by two jobs with the same service class are assigned to the same performance group. In this scenario, the expected result is that IOPM does not throttle any I/O operations coming from the two jobs.

```
JES2 JOB LOG -- SYSTEM CEBC -- NODE MCECEBC

15.56.15 JOB07101 ---- WEDNESDAY, 09 NOV 2011 ----
15.56.15 JOB07101 IRR010I USERID TEAM12 IS ASSIGNED TO THIS JOB.
15.56.35 JOB07101 ICH70001I TEAM12 LAST ACCESS AT 15.56.15 ON WEDNESDAY, NOVEMBER 9, 2011
15.56.35 JOB07101 $HASP373 TEAM12X STARTED - INIT 1 - CLASS A - SYS CEBC
15.56.35 JOB07101 IEF403I TEAM12X - STARTED - TIME=15.56.35
16.05.00 JOB07101 -
16.05.00 JOB07101 -STEPNAME PROCSTEP RC EXCP CONN TCB SRB CLOCK SERV WORKLOAD PAGE SWAP VIO SWAPS
16.05.00 JOB07101 -WORK1 00 40196 133K .05 .00 8.4 69065 BAT_WKL 0 0 0 0
16.05.00 JOB07101 IEF404I TEAM12X - ENDED - TIME=16.05.00
16.05.00 JOB07101 -TEAM12X ENDED. NAME=TEAM12 TOTAL TCB CPU TIME= .05 TOTAL ELAPSED TIME= 8.4
16.05.00 JOB07101 $HASP395 TEAM12X ENDED
```

Figure 4-8 Test 1: WORK1 job sysout

```
JES2 JOB LOG -- SYSTEM CEBC -- NODE MCECEBC

15.56.15 JOB07102 ---- WEDNESDAY, 09 NOV 2011 ----
15.56.15 JOB07102 IRR010I USERID TEAM12 IS ASSIGNED TO THIS JOB.
15.56.35 JOB07102 ICH70001I TEAM12 LAST ACCESS AT 15.56.15 ON WEDNESDAY, NOVEMBER 9, 2011
15.56.35 JOB07102 $HASP373 TEAM12Y STARTED - INIT 2 - CLASS A - SYS CEBC
15.56.35 JOB07102 IEF403I TEAM12Y - STARTED - TIME=15.56.35
16.05.38 JOB07102 -
16.05.38 JOB07102 -STEPNAME PROCSTEP RC EXCP CONN TCB SRB CLOCK SERV WORKLOAD PAGE SWAP VIO SWAPS
16.05.38 JOB07102 -WORK2 00 40196 165K .05 .00 9.0 68708 BAT_WKL 0 0 0 0
16.05.38 JOB07102 IEF404I TEAM12Y - ENDED - TIME=16.05.38
16.05.38 JOB07102 -TEAM12Y ENDED. NAME=TEAM12 TOTAL TCB CPU TIME= .05 TOTAL ELAPSED TIME= 9.0
16.05.38 JOB07102 $HASP395 TEAM12Y ENDED
```

Figure 4-9 Test 1: WORK2 job sysout

As reported in the job sysouts shown in Figure 4-8 on page 45 and Figure 4-9 on page 45, the jobs start at the same time and complete in nearly the same elapsed time. Also, note that the EXCP field reports the same values for both jobs, stating the workload created is exactly the same for the two jobs. Figure 4-10 shows that all the I/O operations are performed in the performance group PG28, as expected, because both jobs are using the same service class BATLO with a execution velocity goal. Also note that the performance group assignment is based on Table 2-5 on page 15.

```

dsccli> lsfprgrprpt -interval 1m -start 1m
time          grp          resrc          avIO  avMB  avresp pri  avQ  tgtQ %hlpT %dlyT %impt
-----
2011-11-09/15:58:00 IBM.2107-75TV181/PG0 IBM.2107-75TV181 10009 151.570 0.136 0 112 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG1 IBM.2107-75TV181 0 0.000 0.000 1 0 70 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG2 IBM.2107-75TV181 0 0.000 0.000 1 0 70 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG3 IBM.2107-75TV181 0 0.000 0.000 1 0 70 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG4 IBM.2107-75TV181 0 0.000 0.000 1 0 70 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG5 IBM.2107-75TV181 0 0.000 0.000 1 0 70 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG6 IBM.2107-75TV181 0 0.000 0.000 5 0 40 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG7 IBM.2107-75TV181 0 0.000 0.000 5 0 40 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG8 IBM.2107-75TV181 0 0.000 0.000 5 0 40 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG9 IBM.2107-75TV181 0 0.000 0.000 5 0 40 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG10 IBM.2107-75TV181 0 0.000 0.000 5 0 40 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG11 IBM.2107-75TV181 0 0.000 0.000 15 0 0 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG12 IBM.2107-75TV181 0 0.000 0.000 15 0 0 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG13 IBM.2107-75TV181 0 0.000 0.000 15 0 0 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG14 IBM.2107-75TV181 0 0.000 0.000 15 0 0 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG15 IBM.2107-75TV181 0 0.000 0.000 15 0 0 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG16 IBM.2107-75TV181 0 0.000 0.000 0 0 0 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG17 IBM.2107-75TV181 0 0.000 0.000 0 0 0 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG18 IBM.2107-75TV181 0 0.000 0.000 0 0 0 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG19 IBM.2107-75TV181 0 0.000 0.000 1 0 80 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG20 IBM.2107-75TV181 0 0.000 0.000 2 0 80 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG21 IBM.2107-75TV181 0 0.000 0.000 3 0 70 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG22 IBM.2107-75TV181 0 0.000 0.000 4 0 45 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG23 IBM.2107-75TV181 0 0.000 0.000 4 0 5 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG24 IBM.2107-75TV181 0 0.000 0.000 5 0 45 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG25 IBM.2107-75TV181 0 0.000 0.000 6 0 5 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG26 IBM.2107-75TV181 0 0.000 0.000 7 0 5 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG27 IBM.2107-75TV181 0 0.000 0.000 8 0 5 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG28 IBM.2107-75TV181 141 8.086 13.762 9 54 5 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG29 IBM.2107-75TV181 0 0.000 0.000 10 0 5 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG30 IBM.2107-75TV181 0 0.000 0.000 11 0 5 0 0 0 0
2011-11-09/15:58:00 IBM.2107-75TV181/PG31 IBM.2107-75TV181 0 0.000 0.000 12 0 5 0 0 0 0

```

Figure 4-10 Test 1: Performance group assignment

Figure 4-11 shows the performance group statistics during the jobs run for PG28. The values reported in the %hlpT and %dlyT fields show that IOPM did not take any action to help or delay the I/O operations in this performance group.

```
dsccli> lsfprgrprpt -interval 1m -start 20m pg28
```

| time | grp | resrc | avIO | avMB | avresp | pri | avQ | tgtQ | %hlpT | %dlyT | %impt |
|---------------------|-----------------------|------------------|------|-------|--------|-----|-----|------|-------|-------|-------|
| 2011-11-09/15:51:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 0 | 0.000 | 0.000 | 9 | 0 | 5 | 0 | 0 | 0 |
| 2011-11-09/15:52:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 0 | 0.000 | 0.000 | 9 | 0 | 5 | 0 | 0 | 0 |
| 2011-11-09/15:53:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 0 | 0.000 | 0.000 | 9 | 0 | 5 | 0 | 0 | 0 |
| 2011-11-09/15:54:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 0 | 0.000 | 0.000 | 9 | 0 | 5 | 0 | 0 | 0 |
| 2011-11-09/15:55:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 0 | 0.000 | 0.000 | 9 | 0 | 5 | 0 | 0 | 0 |
| 2011-11-09/15:56:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 0 | 0.000 | 0.000 | 9 | 0 | 5 | 0 | 0 | 0 |
| 2011-11-09/15:57:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 69 | 4.002 | 14.548 | 9 | 55 | 5 | 0 | 0 | 0 |
| 2011-11-09/15:58:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 141 | 8.086 | 13.762 | 9 | 54 | 5 | 0 | 0 | 0 |
| 2011-11-09/15:59:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 143 | 8.201 | 13.539 | 9 | 54 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:00:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 163 | 9.387 | 11.790 | 9 | 58 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:01:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 124 | 7.130 | 15.663 | 9 | 50 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:02:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 163 | 9.356 | 11.805 | 9 | 58 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:03:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 170 | 9.785 | 11.285 | 9 | 60 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:04:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 160 | 9.227 | 11.996 | 9 | 58 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:05:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 164 | 9.406 | 11.101 | 9 | 60 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:06:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 32 | 1.865 | 15.450 | 9 | 63 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:07:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 0 | 0.000 | 0.000 | 9 | 0 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:08:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 0 | 0.000 | 0.000 | 9 | 0 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:09:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 0 | 0.000 | 0.000 | 9 | 0 | 5 | 0 | 0 | 0 |
| 2011-11-09/16:10:00 | IBM.2107-75TV181/PG28 | IBM.2107-75TV181 | 0 | 0.000 | 0.000 | 9 | 0 | 5 | 0 | 0 | 0 |

Figure 4-11 Test 1: Performance statistics for PG28

Figure 4-12 shows a snapshot of the response time picked from the IBM Resource Measurement Facility™ (RMF) Monitor II.

```
RMF - DEV Device Activity
```

Line 31 of 1068
Scroll ==> PAGE

Command ==>

CPU= 3/ 2 UIC= 65K PR= 0 System= CEBC Total

| 16:04:14 | I=44% | DEV | ACTV | RESP | IOSQ | -DELAY- | PEND | DISC | CONN | %D | %D | | | |
|----------|--------|--------|------|------|-------|---------|------|------|------|------|------|------|----|----|
| STG | GRP | VOLSER | NUM | PAV | LCU | RATE | TIME | TIME | CMR | DB | TIME | TIME | UT | RV |
| SG2 | DE740D | 740D | 2 | 02AA | 0.400 | 13.1 | .000 | .05 | .00 | .142 | 10.4 | 2.54 | 0 | 0 |
| SG2 | DE7409 | 7409 | 2 | 02AA | 0.504 | 13.1 | .000 | .04 | .00 | .131 | 7.06 | 5.88 | 0 | 0 |
| SG2 | DE7225 | 7225 | 2 | 02A8 | 0.445 | 13.0 | .000 | .04 | .00 | .155 | 9.49 | 3.35 | 0 | 0 |
| SG2 | DE7233 | 7233 | 2 | 02A8 | 0.341 | 12.8 | .000 | .05 | .00 | .144 | 11.2 | 1.52 | 0 | 0 |
| SG2 | DE7404 | 7404 | 2 | 02AA | 0.445 | 12.7 | .000 | .03 | .00 | .134 | 8.10 | 4.48 | 0 | 0 |
| SG2 | DE7412 | 7412 | 2 | 02AA | 0.252 | 12.7 | .000 | .04 | .00 | .131 | 12.4 | .105 | 0 | 0 |
| SG2 | DE7228 | 7228 | 2 | 02A8 | 0.445 | 12.6 | .000 | .03 | .00 | .142 | 8.03 | 4.46 | 0 | 0 |
| SG2 | DE7227 | 7227 | 2 | 02A8 | 0.400 | 12.4 | .000 | .03 | .00 | .151 | 9.72 | 2.52 | 0 | 0 |
| SG2 | DE7402 | 7402 | 2 | 02AA | 0.371 | 12.4 | .000 | .04 | .00 | .133 | 8.22 | 4.03 | 0 | 0 |
| SG2 | DE740F | 740F | 2 | 02AA | 0.534 | 12.1 | .000 | .04 | .00 | .140 | 9.12 | 2.83 | 0 | 0 |
| SG2 | DE741C | 741C | 2 | 02AA | 0.385 | 11.9 | .000 | .02 | .00 | .135 | 9.16 | 2.62 | 0 | 0 |
| SG2 | DE723F | 723F | 2 | 02A8 | 0.504 | 11.7 | .000 | .03 | .00 | .152 | 10.5 | 1.03 | 0 | 0 |
| SG2 | DE7414 | 7414 | 2 | 02AA | 0.460 | 11.4 | .000 | .04 | .00 | .130 | 10.1 | 1.16 | 0 | 0 |
| SG2 | DE722A | 722A | 2 | 02A8 | 0.474 | 11.1 | .000 | .03 | .00 | .138 | 8.78 | 2.14 | 0 | 0 |
| SG2 | DE7417 | 7417 | 2 | 02AA | 0.356 | 11.0 | .000 | .02 | .00 | .136 | 10.7 | .114 | 0 | 0 |

Figure 4-12 Test 1: RMF monitor II snapshot

As you can see, the response time is similar for volume used by the WORK1 (device numbers 72XX) and WORK2 (device numbers 74XX) job.

Test 2

In this test, the jobs WORK1 and WORK2 run with separate service classes, as shown in Table 4-4 on page 44. Because this case creates a rank overloading situation, the IOPM might start throttling the I/O operations coming from the job with lower priority service class. As a result, we expect to have a longer run time for the jobs with the lower priority service class (that is, WORK2).

Figure 4-13 and Figure 4-14 show that the two jobs spend separate amounts of time to complete. The job WORK1 completes in 8 minutes and 27 seconds, and the job WORK2 completes in 14 minutes and 55 seconds.

```
JES2 JOB LOG -- SYSTEM CEBC -- NODE MCECEBC

14.36.36 JOB07088 ---- WEDNESDAY, 09 NOV 2011 ----
14.36.36 JOB07088 IRR010I USERID TEAM12 IS ASSIGNED TO THIS JOB.
14.36.36 JOB07088 ICH70001I TEAM12 LAST ACCESS AT 14.36.36 ON WEDNESDAY, NOVEMBER 9, 2011
14.36.36 JOB07088 $HASP373 TEAM12X STARTED - INIT 2 - CLASS A - SYS CEBC
14.36.36 JOB07088 IEF403I TEAM12X - STARTED - TIME=14.36.36
14.45.03 JOB07088 -
14.45.03 JOB07088 -STEPNAME PROCSTEP RC EXCP CONN TCB SRB CLOCK SERV WORKLOAD PAGE SWAP VIO SWAPS
14.45.03 JOB07088 -WORK1 00 40197 135K .05 .00 8.4 69107 BAT_WKL 0 0 0 0
14.45.03 JOB07088 IEF404I TEAM12X - ENDED - TIME=14.45.03
14.45.03 JOB07088 -TEAM12X ENDED. NAME=TEAM12 TOTAL TCB CPU TIME= .05 TOTAL ELAPSED TIME= 8.4
14.45.03 JOB07088 $HASP395 TEAM12X ENDED
```

Figure 4-13 Test 2: WORK1 job sysout

```
JES2 JOB LOG -- SYSTEM CEBC -- NODE MCECEBC

14.36.36 JOB07087 ---- WEDNESDAY, 09 NOV 2011 ----
14.36.36 JOB07087 IRR010I USERID TEAM12 IS ASSIGNED TO THIS JOB.
14.36.36 JOB07087 ICH70001I TEAM12 LAST ACCESS AT 14.36.36 ON WEDNESDAY, NOVEMBER 9, 2011
14.36.36 JOB07087 $HASP373 TEAM12Y STARTED - INIT 1 - CLASS A - SYS CEBC
14.36.36 JOB07087 IEF403I TEAM12Y - STARTED - TIME=14.36.36
14.53.31 JOB07087 -
14.53.31 JOB07087 -STEPNAME PROCSTEP RC EXCP CONN TCB SRB CLOCK SERV WORKLOAD PAGE SWAP VIO SWAPS
14.53.31 JOB07087 -WORK2 00 40196 509K .05 .00 14.9 72867 BAT_WKL 0 0 0 0
14.53.31 JOB07087 IEF404I TEAM12Y - ENDED - TIME=14.53.31
14.53.31 JOB07087 -TEAM12Y ENDED. NAME=TEAM12 TOTAL TCB CPU TIME= .05 TOTAL ELAPSED TIME= 14.9
14.53.31 JOB07087 $HASP395 TEAM12Y ENDED
```

Figure 4-14 Test 2: WORK2 job sysout

The performance group assignment for the I/O operation performed by the two jobs is shown in the Figure 4-15.

```

dsccli> lperfgrprpt -interval 1m -start 1m -l
time          grp          resrc          avIO avMB  avresp pri avQ tgtQ %hlpT %dlyT %impt
-----
2011-11-09/14:39:00 IBM.2107-75TV181/PG0 IBM.2107-75TV181 2 0.030 0.160 0 113 0 0 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG1 IBM.2107-75TV181 0 0.000 0.000 1 0 70 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG2 IBM.2107-75TV181 0 0.000 0.000 1 0 70 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG3 IBM.2107-75TV181 0 0.000 0.000 1 0 70 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG4 IBM.2107-75TV181 0 0.000 0.000 1 0 70 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG5 IBM.2107-75TV181 0 0.000 0.000 1 0 70 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG6 IBM.2107-75TV181 0 0.000 0.000 5 0 40 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG7 IBM.2107-75TV181 0 0.000 0.000 5 0 40 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG8 IBM.2107-75TV181 0 0.000 0.000 5 0 40 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG9 IBM.2107-75TV181 0 0.000 0.000 5 0 40 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG10 IBM.2107-75TV181 0 0.000 0.000 5 0 40 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG11 IBM.2107-75TV181 0 0.000 0.000 15 0 0 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG12 IBM.2107-75TV181 0 0.000 0.000 15 0 0 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG13 IBM.2107-75TV181 0 0.000 0.000 15 0 0 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG14 IBM.2107-75TV181 0 0.000 0.000 15 0 0 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG15 IBM.2107-75TV181 0 0.000 0.000 15 0 0 20 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG16 IBM.2107-75TV181 0 0.000 0.000 0 0 0 0 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG17 IBM.2107-75TV181 0 0.000 0.000 0 0 0 0 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG18 IBM.2107-75TV181 0 0.000 0.000 0 0 0 0 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG19 IBM.2107-75TV181 0 0.000 0.000 1 0 80 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG20 IBM.2107-75TV181 0 0.000 0.000 2 0 80 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG21 IBM.2107-75TV181 74 4.283 12.933 3 63 70 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG22 IBM.2107-75TV181 0 0.000 0.000 4 0 45 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG23 IBM.2107-75TV181 0 0.000 0.000 4 0 5 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG24 IBM.2107-75TV181 0 0.000 0.000 5 0 45 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG25 IBM.2107-75TV181 0 0.000 0.000 6 0 5 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG26 IBM.2107-75TV181 0 0.000 0.000 7 0 5 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG27 IBM.2107-75TV181 0 0.000 0.000 8 0 5 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG28 IBM.2107-75TV181 37 2.127 26.508 9 12 5 7 7 163
2011-11-09/14:39:00 IBM.2107-75TV181/PG29 IBM.2107-75TV181 0 0.000 0.000 10 0 5 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG30 IBM.2107-75TV181 0 0.000 0.000 11 0 5 14 0 0
2011-11-09/14:39:00 IBM.2107-75TV181/PG31 IBM.2107-75TV181 0 0.000 0.000 12 0 5 14 0 0

```

Figure 4-15 Test 2: Performance group assignment

Also, in this case, the performance group assignment is static because both the service classes used (BATLO and BATHI) are defined with an execution velocity goal. The I/O operation performed by WORK1 and WORK2 jobs are assigned respectively to PG21 and PG28.

The charts depicted in Figure 4-16 and Figure 4-17 on page 51 show the IOPS and the response time for the performance groups PG21 and PG28 during the test.

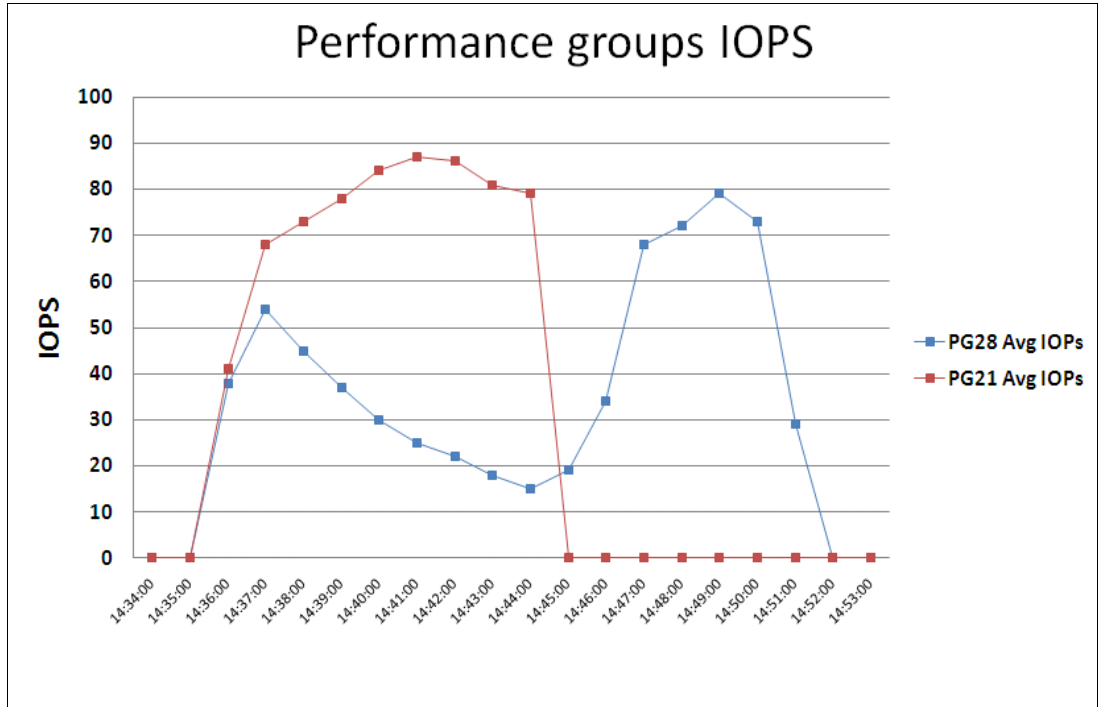


Figure 4-16 Test 2: Performance groups PG21 and PG28 IOPS

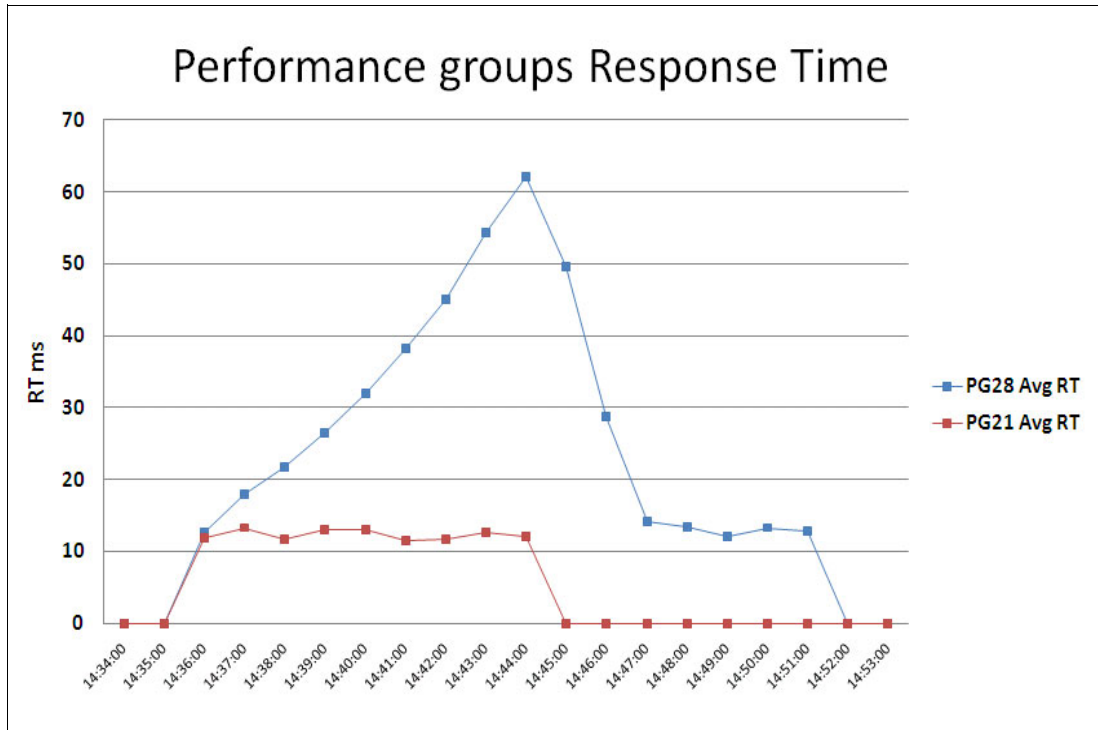


Figure 4-17 Test 2: Performance groups PG21 and PG28 response time

The graphs clearly show that right after the jobs start, IOPM starts throttling the I/Os on performance group PG28 to favor the I/Os in performance group PG21. As a result, the response time for the I/Os in PG28 increase to more than 60 ms. Right after the job WORK1 ends, the IOPM stops throttling the I/Os on PG28, and the response times and IOPS return to normal values.

Figure 4-18 and Figure 4-19 show two RMF Monitor II snapshots indicating that only volumes used by WORK2 job (device numbers 74XX) are affected by delay.

```

RMF - DEV Device Activity                               Line 1 of 1068
Command ==>                                           Scroll ==> PAGE

                                CPU= 2/ 3 UIC= 65K PR= 0      System= CEBC Total

14:44:13 I=88% DEV                                ACTV RESP IOSQ -DELAY- PEND DISC CONN %D %D
STG GRP VOLSER NUM PAV LCU RATE TIME TIME CMR DB TIME TIME TIME UT RV

SG2      DE7417 7417  2 02AA 0.318 64.6 .000 .00 36 16.2 9.94 38.5 1 0
SG2      DE7412 7412  2 02AA 0.378 64.5 .000 .00 32 12.0 11.9 40.6 1 0
SG2      DE7405 7405  2 02AA 0.265 63.1 .000 .00 33 12.9 13.2 37.0 1 0
SG2      DE7406 7406  2 02AA 0.378 62.8 .000 .00 36 16.0 13.5 33.4 1 0
SG2      DE7403 7403  2 02AA 0.318 62.3 .000 .00 33 12.8 11.0 38.5 1 0
SG2      DE7401 7401  2 02AA 0.356 61.4 .000 .00 34 14.1 11.8 35.5 1 0
SG2      DE7408 7408  2 02AA 0.291 61.2 .000 .00 35 14.9 16.1 30.1 1 0
SG2      DE7409 7409  2 02AA 0.303 61.1 .000 .00 32 12.4 12.7 36.0 1 0
SG2      DE741C 741C  2 02AA 0.333 61.1 .000 .00 30 10.4 11.4 39.3 1 0
SG2      DE740B 740B  2 02AA 0.340 59.9 .000 .00 34 14.4 12.4 33.1 1 0
SG2      DE741D 741D  2 02AA 0.424 59.2 .000 .00 31 11.4 13.8 34.0 1 0
SG2      DE7414 7414  2 02AA 0.401 58.4 .000 .00 35 15.1 12.3 30.9 1 0
SG2      DE7413 7413  2 02AA 0.310 58.3 .000 .00 33 13.4 12.4 32.5 1 0
SG2      DE7415 7415  2 02AA 0.348 57.9 .000 .00 34 13.8 12.0 32.2 1 0
SG2      DE7418 7418  2 02AA 0.416 57.9 .000 .00 33 12.9 12.0 33.0 1 0

```

Figure 4-18 Test 2: RMF monitor II snapshot with WORK2 volumes

```

RMF - DEV Device Activity                               Line 33 of 1068
Command ==>                                           Scroll ==> PAGE

                                CPU= 2/ 3 UIC= 65K PR= 0      System= CEBC Total

14:44:13 I=88% DEV                                ACTV RESP IOSQ -DELAY- PEND DISC CONN %D %D
STG GRP VOLSER NUM PAV LCU RATE TIME TIME CMR DB TIME TIME TIME UT RV

SG2      DE7233 7233  2 02A8 1.916 17.2 .000 .02 .00 .135 12.1 4.89 2 0
SG2      DE7221 7221  2 02A8 2.242 13.8 .000 .02 .00 .137 11.6 2.08 2 0
SG2      DE7231 7231  2 02A8 1.696 13.7 .000 .02 .00 .139 11.7 1.78 1 0
SG2      DE723D 723D  2 02A8 2.098 13.6 .000 .02 .00 .137 11.8 1.71 2 0
SG2      DE7224 7224  2 02A8 2.234 13.3 .000 .02 .00 .137 12.1 1.09 2 0
SG2      DE7222 7222  2 02A8 2.121 13.1 .000 .02 .00 .135 11.6 1.43 2 0
SG2      DE722E 722E  2 02A8 1.954 12.6 .000 .02 .00 .137 11.8 1.65 2 0
SG2      DE722A 722A  2 02A8 2.000 12.5 .000 .02 .00 .133 11.7 0.70 2 0
SG2      DE723A 723A  2 02A8 2.045 12.4 .000 .02 .00 .137 11.2 1.10 2 0
SG2      DE7234 7234  2 02A8 1.954 12.3 .000 .02 .00 .136 11.6 1.53 1 0
SG2      DE7228 7228  2 02A8 2.219 12.2 .000 .02 .00 .135 11.1 1.00 2 0
SG2      DE722F 722F  2 02A8 1.969 12.2 .000 .02 .00 .135 11.1 1.01 1 0
SG2      DE7227 7227  2 02A8 1.977 12.1 .000 .02 .00 .135 10.8 1.24 1 0
SG2      DE7235 7235  2 02A8 2.030 12.1 .000 .02 .00 .135 11.7 0.28 2 0
SG2      DE7230 7230  2 02A8 2.037 12.1 .000 .02 .00 .138 11.0 0.87 2 0

```

Figure 4-19 Test 2: RMF monitor II snapshot with WORK1 volumes

Test 3

Test 3 is similar to test 2, but it uses a higher-prioritized service class for WORK1 (see Table 4-4 on page 44). Test 3 shows that a more demanding service class introduces a more aggressive throttling of the I/Os.

Again, the two jobs complete with separate elapsed times, as shown in Figure 4-20 and Figure 4-21, but in this case, the difference is much greater than in test 2. The job WORK1 completes in 8 minutes and 25 seconds, while the job WORK2 completes in 20 minutes and 2 seconds.

```
JES2 JOB LOG -- SYSTEM CEBC -- NODE MCECEBC

15.32.50 JOB07097 ---- WEDNESDAY, 09 NOV 2011 ----
15.32.50 JOB07097 IRR010I USERID TEAM12 IS ASSIGNED TO THIS JOB.
15.32.50 JOB07097 ICH70001I TEAM12 LAST ACCESS AT 15.32.04 ON WEDNESDAY, NOVEMBER 9, 2011
15.32.50 JOB07097 $HASP373 TEAM12X STARTED - INIT 1 - CLASS A - SYS CEBC
15.32.50 JOB07097 IEF403I TEAM12X - STARTED - TIME=15.32.04
15.41.15 JOB07097 -
15.41.15 JOB07097 -STEPNAME PROCSTEP RC EXCP CONN TCB SRB CLOCK SERV WORKLOAD PAGE SWAP VIO SWAPS
15.41.15 JOB07097 -WORK1 00 40198 140K .05 .00 8.4 71005 ONL_WKL 0 0 0 0
15.41.15 JOB07097 IEF404I TEAM12X - ENDED - TIME=15.41.15
15.41.15 JOB07097 -TEAM12X ENDED. NAME=TEAM12 TOTAL TCB CPU TIME= .05 TOTAL ELAPSED TIME= 8.4
15.41.15 JOB07097 $HASP395 TEAM12X ENDED
```

Figure 4-20 Test 3: WORK1 job sysout

```
JES2 JOB LOG -- SYSTEM CEBC -- NODE MCECEBC

15.32.50 JOB07098 ---- WEDNESDAY, 09 NOV 2011 ----
15.32.50 JOB07098 IRR010I USERID TEAM12 IS ASSIGNED TO THIS JOB.
15.32.50 JOB07098 ICH70001I TEAM12 LAST ACCESS AT 15.32.50 ON WEDNESDAY, NOVEMBER 9, 2011
15.32.50 JOB07098 $HASP373 TEAM12Y STARTED - INIT 2 - CLASS A - SYS CEBC
15.32.50 JOB07098 IEF403I TEAM12Y - STARTED - TIME=15.32.50
15.52.52 JOB07098 -
15.52.52 JOB07098 -STEPNAME PROCSTEP RC EXCP CONN TCB SRB CLOCK SERV WORKLOAD PAGE SWAP VIO SWAPS
15.52.52 JOB07098 -WORK2 00 40196 785K .05 .01 20.0 74521 BAT_WKL 0 0 0 0
15.52.52 JOB07098 IEF404I TEAM12Y - ENDED - TIME=15.52.52
15.52.52 JOB07098 -TEAM12Y ENDED. NAME=TEAM12 TOTAL TCB CPU TIME= .05 TOTAL ELAPSED TIME= 20.0
15.52.52 JOB07098 $HASP395 TEAM12Y ENDED
```

Figure 4-21 Test 3: WORK2 job sysout

The performance group assignment for the I/O operation performed by the two jobs is shown in Figure 4-22. Note that in this case the performance group assignment for the WORK1 I/Os is dynamic because the service class ONLHI is defined with a response time goal. However, because the response time goal can hardly be achieved by a batch job, the performance group selected for the I/Os becomes almost immediately PG19, the highest performance priority for CKD volumes.

```

dsccli> lsperfgrprpt -interval 1m -start 25m -stop 24m
time          grp          resrc          avIO avMB  avresp pri  avQ tgtQ %hlpT %dlyT %impt
-----
2011-11-09/15:33:00 IBM.2107-75TV181/PG0 IBM.2107-75TV181 3651 38.379 0.279 0 124 0 0 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG1 IBM.2107-75TV181 0 0.000 0.000 1 0 70 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG2 IBM.2107-75TV181 0 0.000 0.000 1 0 70 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG3 IBM.2107-75TV181 0 0.000 0.000 1 0 70 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG4 IBM.2107-75TV181 0 0.000 0.000 1 0 70 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG5 IBM.2107-75TV181 0 0.000 0.000 1 0 70 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG6 IBM.2107-75TV181 0 0.000 0.000 5 0 40 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG7 IBM.2107-75TV181 0 0.000 0.000 5 0 40 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG8 IBM.2107-75TV181 0 0.000 0.000 5 0 40 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG9 IBM.2107-75TV181 0 0.000 0.000 5 0 40 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG10 IBM.2107-75TV181 0 0.000 0.000 5 0 40 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG11 IBM.2107-75TV181 0 0.000 0.000 15 0 0 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG12 IBM.2107-75TV181 0 0.000 0.000 15 0 0 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG13 IBM.2107-75TV181 0 0.000 0.000 15 0 0 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG14 IBM.2107-75TV181 0 0.000 0.000 15 0 0 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG15 IBM.2107-75TV181 0 0.000 0.000 15 0 0 20 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG16 IBM.2107-75TV181 0 0.000 0.000 0 0 0 0 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG17 IBM.2107-75TV181 0 0.000 0.000 0 0 0 0 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG18 IBM.2107-75TV181 0 0.000 0.000 0 0 0 0 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG19 IBM.2107-75TV181 66 3.824 14.561 1 61 80 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG20 IBM.2107-75TV181 0 0.000 0.000 2 0 80 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG21 IBM.2107-75TV181 0 0.000 0.000 3 0 70 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG22 IBM.2107-75TV181 0 0.000 0.000 4 0 45 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG23 IBM.2107-75TV181 0 0.000 0.000 4 0 5 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG24 IBM.2107-75TV181 0 0.000 0.000 5 0 45 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG25 IBM.2107-75TV181 0 0.000 0.000 6 0 5 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG26 IBM.2107-75TV181 0 0.000 0.000 7 0 5 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG27 IBM.2107-75TV181 0 0.000 0.000 8 0 5 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG28 IBM.2107-75TV181 34 1.981 28.475 9 13 5 7 7 170
2011-11-09/15:33:00 IBM.2107-75TV181/PG29 IBM.2107-75TV181 0 0.000 0.000 10 0 5 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG30 IBM.2107-75TV181 0 0.000 0.000 11 0 5 14 0 0
2011-11-09/15:33:00 IBM.2107-75TV181/PG31 IBM.2107-75TV181 0 0.000 0.000 12 0 5 14 0 0

```

Figure 4-22 Test 3: Performance Group assignment

Figure 4-23 and Figure 4-24 show the IOPS and response times for the performance groups PG19 and PG28 during test 3.

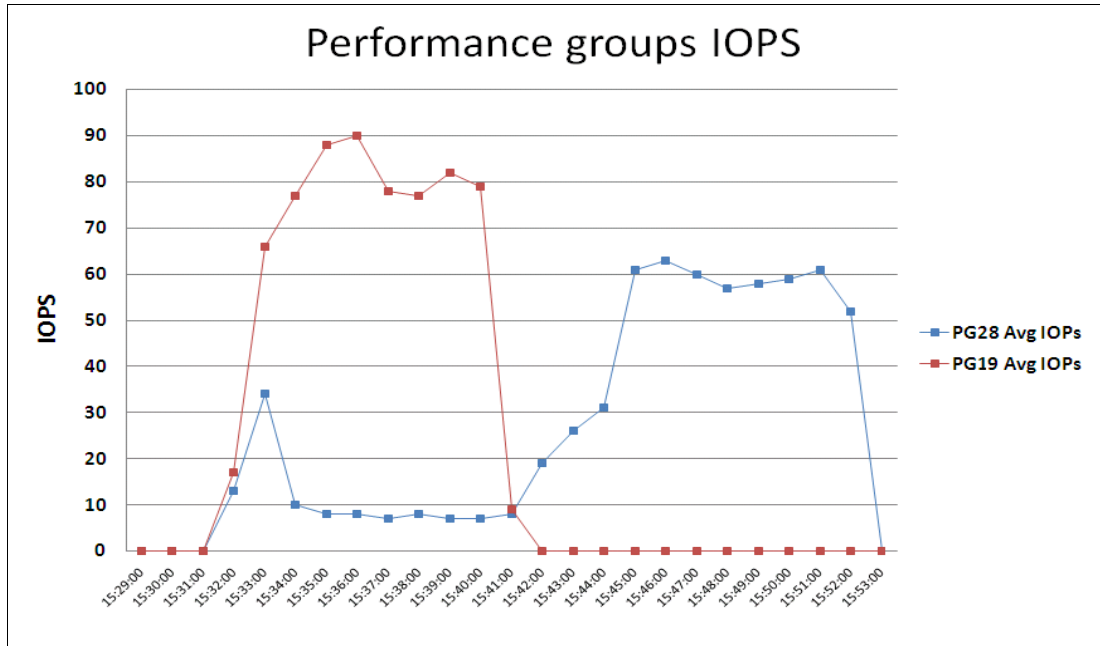


Figure 4-23 Test 3: Performance groups PG19 and PG28 IOPS

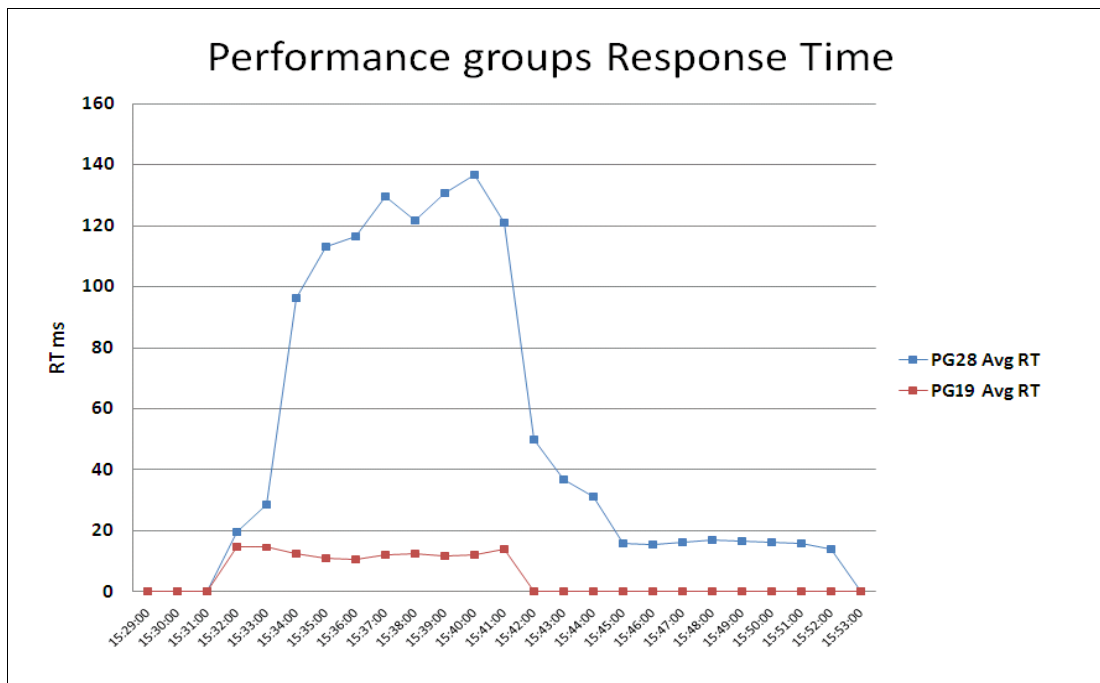


Figure 4-24 Test 2: Performance groups PG19 and PG28 response time

These graphs show an aggressive I/O throttling on performance group PG28 to favor the I/Os in performance group PG19. As result, the response time for the I/Os in PG28 increases to more than 130ms, and the job WORK2 almost stops (processing only 7-8 operations per second). As in test 2, right after the job WORK1 ends, the IOPM stops throttling the I/Os on PG28, and the response times and IOPS return to normal values.

RMF Monitor II snapshots are depicted in the Figure 4-25 and Figure 4-26 on page 57. Note that the response time reported by RMF matches to the one collected by DS8000 (for the interval 15:40).

| RMF - DEV Device Activity | | | | | | | | | | | | | Line 33 of 1068 | | | |
|---------------------------|-----|--------|------|-----|------|-------|------|------|-----|-----|------|------|---|----|----|--|
| Command ==> | | | | | | | | | | | | | Scroll ==> PAGE | | | |
| CPU= 2/ 1 UIC= 65K PR= 0 | | | | | | | | | | | | | System= CEBC Total | | | |
| 15:40:11 I=23% DEV | | | | | | | | | | | | | ACTV RESP IOSQ -DELAY- PEND DISC CONN %D %D | | | |
| STG | GRP | VOLSER | NUM | PAV | LCU | RATE | TIME | TIME | CMR | DB | TIME | TIME | TIME | UT | RV | |
| SG2 | | D£7233 | 7233 | 2 | 02A8 | 1.712 | 14.7 | .000 | .03 | .00 | .136 | 13.4 | 1.22 | 2 | 0 | |
| SG2 | | D£723D | 723D | 2 | 02A8 | 2.105 | 12.9 | .000 | .03 | .00 | .142 | 11.8 | 0.95 | 2 | 0 | |
| SG2 | | D£7231 | 7231 | 2 | 02A8 | 1.740 | 12.2 | .000 | .03 | .00 | .133 | 11.1 | 0.02 | 2 | 0 | |
| SG2 | | D£7223 | 7223 | 2 | 02A8 | 1.487 | 12.0 | .000 | .03 | .00 | .141 | 11.6 | 0.27 | 1 | 0 | |
| SG2 | | D£7232 | 7232 | 2 | 02A8 | 1.431 | 11.5 | .000 | .03 | .00 | .141 | 10.8 | 0.52 | 1 | 0 | |
| SG2 | | D£723C | 723C | 2 | 02A8 | 1.571 | 11.5 | .000 | .03 | .00 | .136 | 10.0 | 0.36 | 2 | 0 | |
| SG2 | | D£722D | 722D | 2 | 02A8 | 2.077 | 11.1 | .000 | .04 | .00 | .138 | 10.6 | 0.41 | 2 | 0 | |
| SG2 | | D£7237 | 7237 | 2 | 02A8 | 1.628 | 11.0 | .000 | .04 | .00 | .140 | 10.5 | 0.44 | 2 | 0 | |
| SG2 | | D£723B | 723B | 2 | 02A8 | 1.964 | 11.0 | .000 | .02 | .00 | .141 | 10.2 | 0.59 | 2 | 0 | |
| SG2 | | D£7227 | 7227 | 2 | 02A8 | 1.880 | 10.4 | .000 | .02 | .00 | .137 | 10.3 | 0.94 | 2 | 0 | |
| SG2 | | D£7234 | 7234 | 2 | 02A8 | 2.077 | 10.4 | .000 | .03 | .00 | .140 | 10.1 | 0.29 | 2 | 0 | |
| SG2 | | D£7235 | 7235 | 2 | 02A8 | 2.105 | 10.4 | .000 | .03 | .00 | .144 | 9.30 | 0.90 | 2 | 0 | |
| SG2 | | D£723E | 723E | 2 | 02A8 | 1.628 | 9.41 | .000 | .03 | .00 | .136 | 8.10 | 1.17 | 1 | 0 | |
| SG2 | | D£722E | 722E | 2 | 02A8 | 1.768 | 9.32 | .000 | .02 | .00 | .142 | 7.82 | 0.85 | 2 | 0 | |
| SG2 | | D£7229 | 7229 | 2 | 02A8 | 2.189 | 9.14 | .000 | .02 | .00 | .141 | 8.21 | 0.82 | 2 | 0 | |

Figure 4-25 Test 3: RMF monitor II snapshot with WORK1 volumes

| RMF - DEV Device Activity | | | | | | | | | | | Line 1 of 1068 | | | |
|---------------------------|-----|--------|------|-----|------|-------|------|------|---------|------|--------------------|------|------|-----|
| Command ==> | | | | | | | | | | | Scroll ==> PAGE | | | |
| CPU= 2/ 1 UIC= 65K PR= 0 | | | | | | | | | | | System= CEBC Total | | | |
| 15:40:11 I=23% DEV | | | | | | | | | | | | | | |
| STG | GRP | VOLSER | NUM | PAV | LCU | ACTV | RESP | IOSQ | -DELAY- | PEND | DISC | CONN | %D | %D |
| | | | | | | RATE | TIME | TIME | CMR | DB | TIME | TIME | UT | RV |
| SG2 | | D£7403 | 7403 | 2 | 02AA | 0.168 | 150 | .000 | .00 | 31 | 31.5 | 12.6 | 106 | 1 0 |
| SG2 | | D£7405 | 7405 | 2 | 02AA | 0.224 | 142 | .000 | .02 | 36 | 36.4 | 14.5 | 91.1 | 1 0 |
| SG2 | | D£740B | 740B | 2 | 02AA | 0.196 | 141 | .000 | .00 | 47 | 47.5 | 9.85 | 83.2 | 1 0 |
| SG2 | | D£7407 | 7407 | 2 | 02AA | 0.196 | 137 | .000 | .00 | 40 | 40.4 | 13.4 | 83.1 | 1 0 |
| SG2 | | D£7413 | 7413 | 2 | 02AA | 0.168 | 136 | .000 | .00 | 31 | 31.5 | 9.77 | 95.3 | 1 0 |
| SG2 | | D£7418 | 7418 | 2 | 02AA | 0.224 | 136 | .000 | .00 | 39 | 39.6 | 13.5 | 83.2 | 1 0 |
| SG2 | | D£7408 | 7408 | 2 | 02AA | 0.252 | 136 | .000 | .00 | 48 | 47.8 | 13.0 | 74.9 | 1 0 |
| SG2 | | D£7417 | 7417 | 2 | 02AA | 0.252 | 135 | .000 | .00 | 32 | 31.6 | 10.5 | 93.1 | 2 0 |
| SG2 | | D£741C | 741C | 2 | 02AA | 0.140 | 135 | .000 | .00 | 25 | 25.4 | 17.9 | 92.2 | 1 0 |
| SG2 | | D£7419 | 7419 | 2 | 02AA | 0.252 | 134 | .000 | .00 | 40 | 39.9 | 11.6 | 92.1 | 1 0 |
| SG2 | | D£740F | 740F | 2 | 02AA | 0.196 | 133 | .000 | .00 | 31 | 31.3 | 9.63 | 93.4 | 1 0 |
| SG2 | | D£740D | 740D | 2 | 02AA | 0.196 | 133 | .000 | .00 | 25 | 24.9 | 6.58 | 102 | 1 0 |
| SG2 | | D£741A | 741A | 2 | 02AA | 0.280 | 132 | .000 | .00 | 36 | 36.2 | 7.94 | 88.1 | 2 0 |
| SG2 | | D£7416 | 7416 | 2 | 02AA | 0.140 | 132 | .000 | .00 | 43 | 43.0 | 10.7 | 98.5 | 1 0 |
| SG2 | | D£7411 | 7411 | 2 | 02AA | 0.280 | 132 | .000 | .00 | 40 | 40.6 | 12.6 | 98.5 | 2 0 |

Figure 4-26 Test 3: RMF monitor II snapshot with WORK2 volumes



Monitoring and reporting

This chapter introduces the main monitoring features related to I/O Priority Manager that are provided through the DS CLI and DS GUI interfaces.

5.1 Monitoring and reporting with the DS CLI

Three new reporting commands have been added to the DS CLI for I/O Priority Manager: **lspfergrp**, **lspferescript**, and **lspfergrprpt**. The following section describes these commands.

5.1.1 Displaying I/O performance groups

Example 5-1 shows how the **lspfergrp** command is used to display the I/O performance groups and the policy of each group. PG0 is not monitored nor managed. If you have migrated from an earlier microcode level to the R 6.1 code level, all of the CKD and FB volumes are assigned to this default I/O performance group.

PG1 through PG5 are given the high performance policy (2) for FB volumes. These FB volumes are monitored and managed by the same performance policy, but reports are generated separately for each performance group. PG6 - PG10 are managed by the medium performance policy (3), and PG11 - PG15 are assigned the low performance policy (4) for FB volumes. For CKD volumes, PG19 - PG31 are given performance policies 19 to 31, respectively.

Example 5-1 Displaying I/O performance group and policy

```
dscli>lspfergrp
```

```
ID    pol
```

```
=====
```

```
PG0   1
PG1   2
PG2   2
PG3   2
PG4   2
PG5   2
PG6   3
PG7   3
PG8   3
PG9   3
PG10  3
PG11  4
PG12  4
PG13  4
PG14  4
PG15  4
PG16 16
PG17 17
PG18 18
PG19 19
PG20 20
PG21 21
PG22 22
PG23 23
PG24 24
PG25 25
PG26 26
PG27 27
PG28 28
PG29 29
PG30 30
PG31 31
```

5.1.2 Performance reports

I/O Priority Manager generates performance statistics every 60 seconds for device adapters (DAs), ranks, and performance groups. These performance statistics samples are kept for a specified period of time. I/O Priority Manager maintains statistics for the last immediate set of values, as follows:

- ▶ Sixty 1-minute intervals
- ▶ Sixty 1-minute intervals
- ▶ Sixty 15-minute intervals
- ▶ Sixty 1-hour intervals
- ▶ Sixty 4-hour intervals
- ▶ Sixty 1-day intervals

Rank reports

The `lspersfrescrpt` command displays performance statistics for individual ranks, as follows:

- ▶ The first three columns indicate the average number of IOPS (avIO), throughput (avMB), and average response time (avresp), each in milliseconds, for all I/Os on that rank.
- ▶ The %Hutl column shows the percentage of time the rank has had utilization high enough (above 33%) to warrant workload control.
- ▶ The %hlpT column indicates the average percentage of time that the IOPM has helped I/Os on this rank for all performance groups, that is, the percentage of time where lower priority I/Os have been delayed to help higher priority I/Os.
- ▶ The %dlyT column indicates the average percentage of time that I/Os have been delayed for all performance groups on this rank.

For example, a 25 in either the %hlpT or %dlyT column can mean that four performance groups (25%) have been helped or delayed 100% of time, all performance groups have been helped or delayed 25% of the time, or a certain combination of ranks helping or delaying I/Os is producing this percentage.

- ▶ The %impt column indicates, on average, the delay time.

Example 5-2 shows a DS8000 performance report on rank 19.

Example 5-2 Rank performance resource report

```
dsccli> lspersfrescrpt R19
time          resrc avIO avMB  avresp %Hutl %hlpT %dlyT %impt
=====
2011-05-25/12:30:00 R19  1552 6.357 7.828 100  24  7  202
2011-05-25/12:35:00 R19  1611 6.601 8.55  100  4  6  213
2011-05-25/12:40:00 R19  1610 6.596 7.888  50  7  4  198
2011-05-25/12:45:00 R19  1595 6.535 6.401 100 12  7  167
2011-05-25/12:50:00 R19  1586 6.496 8.769  50  5  3  219
```

Looking at the first row in Example 5-2, you can see that for the first five-minute sample period, rank 19 averaged 1552 IOPS at 6.357 MBps, with an average response time of 7.828 milliseconds. For 100% of the time period, the rank has been used in a state with a utilization above 33%, which did warrant I/O Priority Manager workload control. Seven percent of the I/Os over the five-minute time period have been delayed. The average delay for each delayed I/O is 202% of the normal response time for rank 19.

Performance group reports

The `lspgrprpt` command is used to display statistics for individual performance groups.

Example 5-3 shows the output of the `lspgrprpt` command for performance group 6 broken up into two halves (for formatting purposes). The top half is the output from the “short” version of the `lspgrprpt` command, and the bottom half is the extra output that is displayed when the “long” version (with the `-l` switch) of the command is used.

Example 5-3 Performance group report

```

dsccli> lspgrprpt -l PG6
time                grp                resrc                avIO avMB  avresp pri avQ tgtQ %hlpT %dlyT %impt
=====
2011-05-25/13:45:00 IBM.2107-75LX521/PG6 IBM.2107-75LX521 2039 8.353 3.870 5 74 40 6 0 104
2011-05-25/13:50:00 IBM.2107-75LX521/PG6 IBM.2107-75LX521 2020 8.275 3.908 5 90 40 0 0 100
2011-05-25/13:55:00 IBM.2107-75LX521/PG6 IBM.2107-75LX521 1902 7.793 4.153 5 87 40 0 0 100
2011-05-25/14:00:00 IBM.2107-75LX521/PG6 IBM.2107-75LX521 1789 7.328 4.416 5 70 40 2 1 106
2011-05-25/14:05:00 IBM.2107-75LX521/PG6 IBM.2107-75LX521 1909 7.821 4.129 5 55 40 6 1 109

mnIO mxIO %idle mnMB mxMB mnresp mxresp %Hutl %VHutl %loQ %hiQ %hlp# %req %dly# %ceil
=====
0 2156 93 0.000 8.832 -0.001 4.392 30 12 0 99 0 0 5 500
1928 2090 86 7.902 8.562 3.774 4.097 27 14 0 100 0 0 0 500
1829 2002 86 7.495 8.204 3.942 4.320 27 14 0 100 0 0 0 500
0 1914 93 0.000 7.844 -0.001 4.992 47 12 0 98 0 0 14 500
0 2044 93 0.000 8.375 -0.001 4.668 37 11 0 98 0 0 18 500
1928 2095 86 7.901 8.585 3.765 4.094 26 14 0 100 0 0 0 500
0 2194 93 0.000 8.989 -0.001 4.400 23 10 0 99 0 0 5 500

```

The first three columns indicate what time the report was run and which resource the report is measuring. The `avIO`, `avMB`, and `avresp` columns indicate the average number of I/Os, throughput, and average response time of the performance group, respectively. Note that `avresp` indicates the average response time in tenths of a second for track I/O operations during this interval. The `pri` column indicates which performance policy the performance group is in. Table 5-1 and Table 5-2 show which priority corresponds to which performance policy for FB volumes and CKD volumes, respectively.

Table 5-1 Priority to performance policy mapping for FB volumes

| Priority (pri field) | Performance policy |
|----------------------|--------------------|
| 0 | Default |
| 1 | High |
| 5 | Medium |
| 15 | Low |

Table 5-2 Priority to performance policy mapping for CKD volumes

| Priority (pri field) | Performance policy |
|----------------------|--------------------|
| 0 | Default |
| 1 | High 1 |
| 2 | High 2 |
| 3 | High 3 |

| Priority (pri field) | Performance policy |
|----------------------|--------------------|
| 4 | Medium 1 |
| 5 | Medium 2 |
| 6 | Medium 3 |
| 7 | Low 1 |
| 8 | Low 2 |
| 9 | Low 3 |
| 10 | Low 4 |
| 11 | Low 5 |
| 12 | Low 6 |

The avQ and tgtQ columns in Example 5-3 on page 62 indicate the average quality of service (QoS) and target QoS for the performance group, respectively. For more information about QoS, see 2.1, “Quality of Service with DS8000” on page 6.

The %hlpT and %dlyT columns indicate the average percentage of time across all ranks in the entire system where I/Os have been delayed or helped. For example, for a system with 16 ranks, a value of 25 in either the %hlpT or %dlyT column can mean that four ranks (25%) have helped or delayed I/Os 100% of time, all ranks have been helping or delaying I/Os 25% of the time, or a certain combination of ranks that are helping or delaying I/Os is producing this percentage.

The %impt (percentage impact) column indicates the average percentage delay for I/Os in the performance group. A %impt value of 100 indicates no delay has been added.

The long performance group report is more detailed and includes more columns. Of particular interest are the following columns:

- ▶ Minimum and maximum I/O (mnIO and mxIO), minimum and maximum throughput (mnMB and mxMB), and minimum and maximum response time (mnresp and mxresp). A zero or negative number in any of these fields should be ignored.
- ▶ Percentage of ranks in the entire system that are not performing any I/O for the performance group in the report (%idle).
- ▶ High utilization (%Hutl), indicating what percentage of total ranks in the system have a utilization rate above 33%.
- ▶ Very high utilization (%VHutl), indicating what percentage of total ranks in the system have a utilization rate above 66%.
- ▶ Percentage of I/Os in the performance group that were delayed (%dly#)
- ▶ Ceiling (%ceil), showing the maximum value that the %impt field can be for a performance group. Note that certain %impt values might exceed this ceiling during periods of high throttling.

5.2 Monitoring and reporting with the DS GUI

Performance group monitoring can also be done from the DS GUI. To reach the I/O Priority Manager window, follow these steps:

1. From the DS GUI Welcome window, hover over the **Home** icon and click **System Status**.
2. In the System Status window, select the storage image that you want to monitor.
3. Select **Action** → **Storage Image** → **I/O Priority Manager**. The monitoring window opens as shown in Figure 5-1.

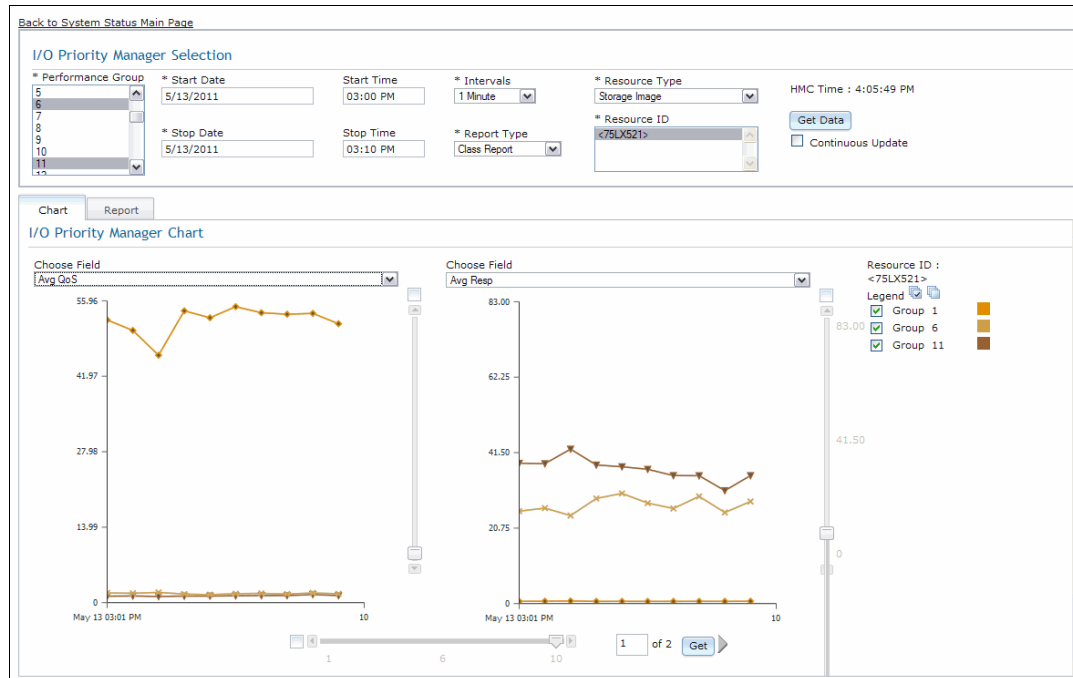


Figure 5-1 Report of I/O priority groups

The following fields are required to generate a report or graph:

- ▶ Performance Group: Selects which performance groups to run the report on. More than one performance group can be selected by using the Ctrl or Shift key in the selection process.
- ▶ Start Date and Stop Date: Specify the day(s) when the statistics are to be collected.
- ▶ Start/Stop Time: Specifies the time period of the report.
- ▶ Intervals: Specifies the sample rate of the report.

Report retention: Only the previous 60 reports are kept for each interval.

- ▶ Report Type: Indicates that the requested report is a resource report
- ▶ Resource Type: Specifies the type of resource to perform the report on. Values include storage image, rank, or device adapter (DA).
- ▶ Resource ID: Specifies the resource(s) to perform the report on. More than one resource can be selected by using the Ctrl or Shift key.

Click the **Get Data** button to generate a report. After the report has been generated, two graphs appear in the I/O Priority Manager Chart section of the window, as shown in Figure 5-1 on page 64, which shows a report for a performance group. Other types of charts can be displayed by selecting a new metric from the drop-down menus above the charts. Lines on the charts can be toggled on and off by selecting the appropriate check box in the legend on the right side of the window.

Separate resources have separate metrics to choose from and display. Figure 5-2 shows the output from a rank report. Note that the report type is Resource Report and that the resource type is Rank.

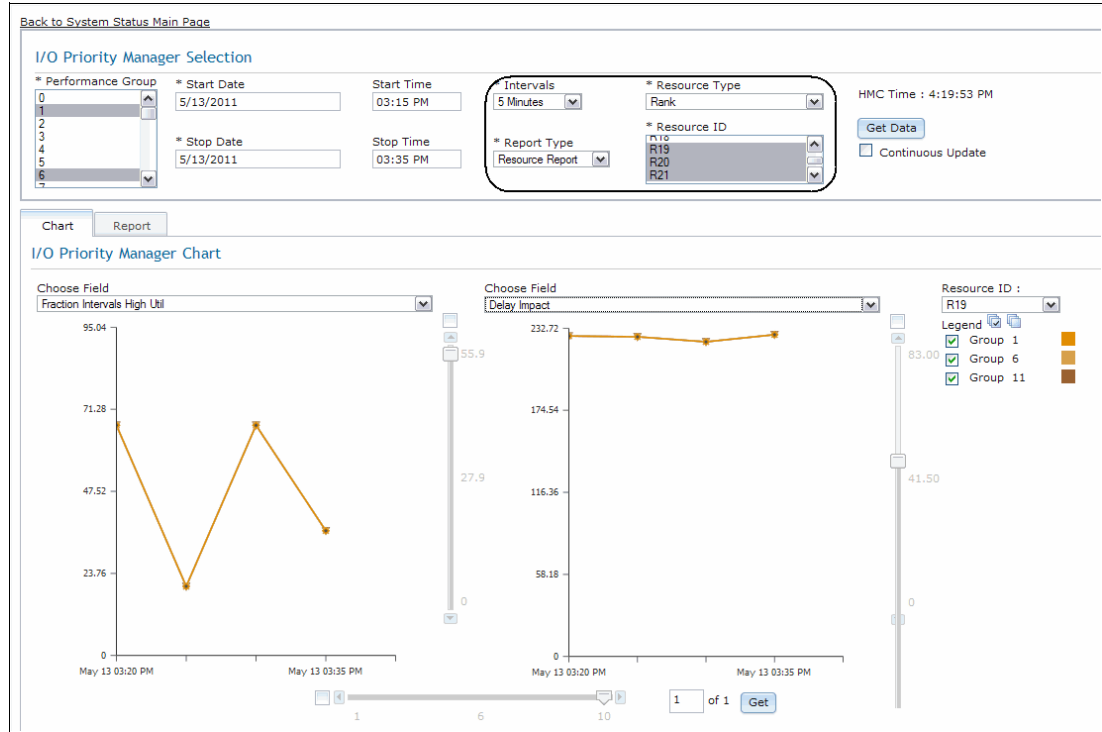


Figure 5-2 Report on a rank

To change the scale on any chart, first click the check box above the axis you want to scale. Then move the slider up or down to change the scale of the chart. Figure 5-3 shows an example of changing the Y-axis scale on the average response chart.

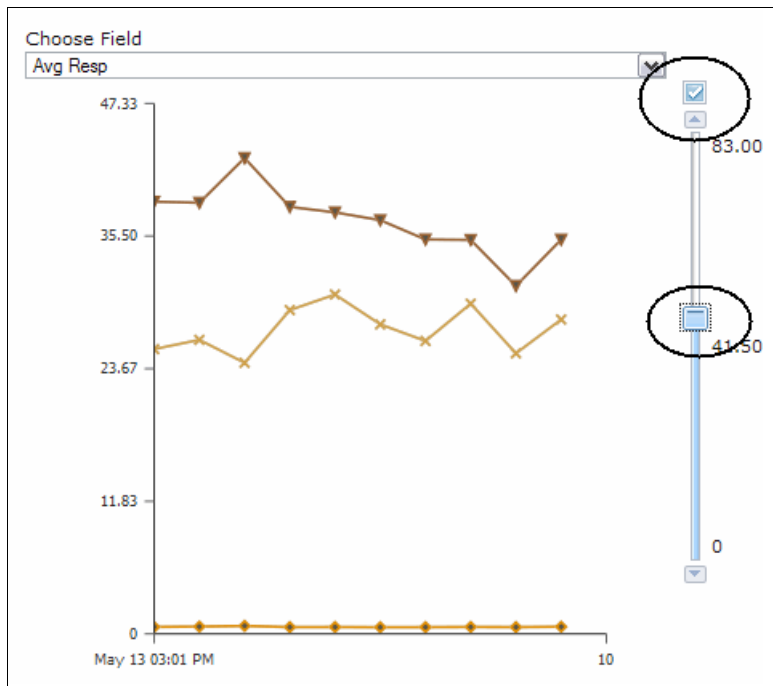


Figure 5-3 Change the scale of the display

To view the performance data in a table, click the **Report** tab as shown in Figure 5-4. This table can be downloaded as a comma separated file by clicking the Download Spreadsheet button. To change which fields are displayed, either right-click any column title or click the Choose Table Columns button.

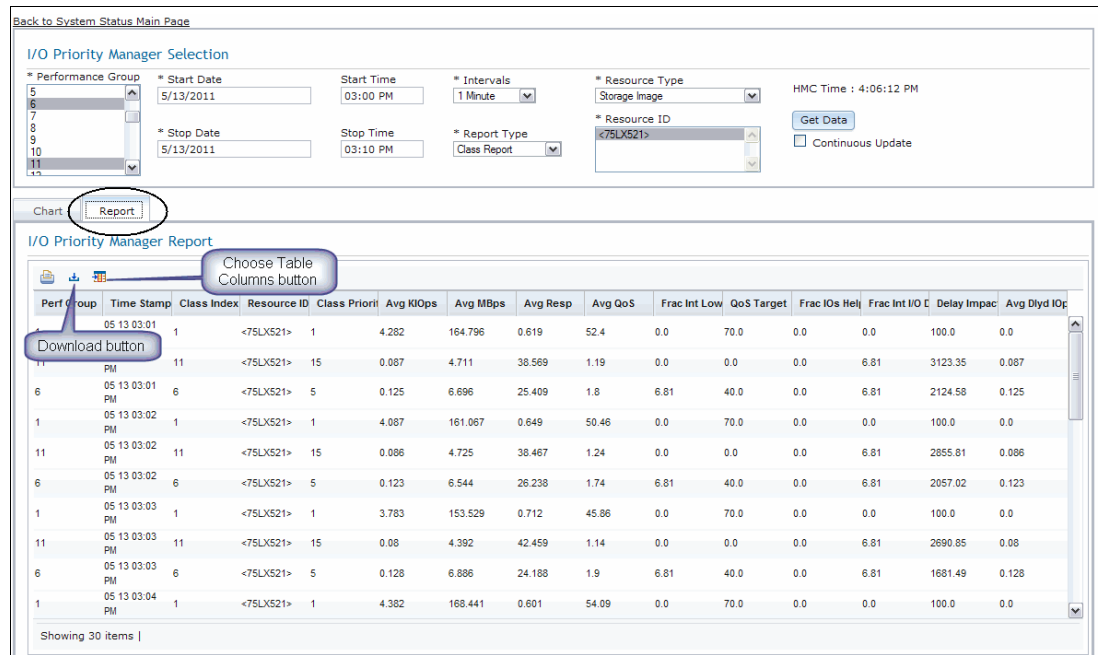


Figure 5-4 Tabular report on performance group

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that certain publications referenced in this list might be available in softcopy only.

- ▶ *ABCs of z/OS System Programming Volume 12, SG24-7621*
- ▶ *DS8000 Performance Monitoring and Tuning, SG24-8013*
- ▶ *IBM System Storage DS8000: Architecture and Implementation, SG24-8886*
- ▶ *IBM System Storage DS8000 Host Attachment and Interoperability, SG24-8887*
- ▶ *System Programmer's Guide to: Workload Manager, SG24-6472*

Other publications

These publications are also relevant as further information sources:

- ▶ *IBM i Shared Storage Performance using IBM System Storage DS8000 I/O Priority Manager, WP101935*
- ▶ *IBM System Storage DS8700 and DS8800 (M/T 242x) delivers DS8000 I/O Priority Manager and advanced features to enhance data protection for multi-tenant copy services, ZG11-0282 (IBM Europe, Middle East, and Africa Hardware Announcement, October 11, 2011 - DS8000 R6.2)*
- ▶ *IBM System Storage DS8700 and DS8800 (M/T 242x) delivers DS8000 I/O Priority Manager and advanced features to enhance data protection for multi-tenant copy services, ZG11-0130 (IBM Europe, Middle East, and Africa Hardware Announcement, May 9, 2011 - DS8000 R6.1)*
- ▶ *IBM System Storage DS8700 and DS8800 Introduction and Planning Guide, GC27-2297-07*
- ▶ *z/OS MVS Planning: Workload Management, SA22-7602*

Online resources

These websites are also relevant as further information sources:

- ▶ IBM Disk storage feature activation (DSFA)
<http://www.ibm.com/storage/dsfa>
- ▶ IBM Offering Information (searchable, such as for R6.2)
<http://www.ibm.com/common/ssi/index.wss>

- ▶ IBM System Storage DS8000 series
<http://www.ibm.com/systems/storage/disk/ds8000/index.html>
- ▶ IBM System Storage Interoperation Center (SSIC)
<http://www.ibm.com/systems/support/storage/config/ssic/index.jsp>
- ▶ IBM Techdocs - The Technical Sales Library
<http://www.ibm.com/support/techdocs/atmastr.nsf/Web/Techdocs>

How to get IBM Redbooks publications

You can search for, view, or download IBM Redbooks publications, Redpapers, Hints and Tips, draft publications, and additional materials at the following website. You can also order hardcopy IBM Redbooks publications or CD-ROMs here:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

A

achievement 11

B

business importance 12

C

chckdvol 14

chfbvol 8, 27

chsi 24

D

DA 19

default 7

delay 18

device adapter (DA) 19

DFSA 22

Discretionary 16

Disk Storage Feature Activation (DSFA) 22

E

Easy Tier 19

execution velocity 14

H

high priority 7

I

I/O Priority Manager

disabled 18

manage 18

modes of operation 18

monitor 18

importance value 11

L

licensed feature 22

licensed function 22

low priority 7

lskey 22

lsperfgrp 60

lsperfgrppt 9, 14, 62

lsperfrescrt 61

M

manage mode, IOPM 26

medium priority 7

mkckdvol 30

mkfbvol 26

modes of operation, IOPM 18

monitor mode, IOPM 26

P

performance goal 12

performance group 8, 14, 18

monitoring 64

performance policy 7, 10, 14

performance statistics 61

Q

QoS 2

metric 6

targets 10

Quality of Service (QoS) 2, 6

Index 42

R

RAID array 18

Redbooks website 68

Contact us x

response time 15

average 6, 18

nominal 18

optimal 6

S

saturation 18

Send SNMP Traps 26

service class 14

SFI 18

showckdvol 30

showfbvol 27

Simple Network Management Protocol 18

SNMP 18

storage facility image (SFI) 18

System Service Classes 14

System z 9

T

throttling 19

W

WLM 12



DS8000 I/O Priority Manager



Redpaper™

Prioritize application I/O in open systems and z/OS environments

Enable policy-based performance tuning

Handle mixed workloads

This IBM Redpaper publication describes the concepts and functions of the IBM System Storage DS8000 I/O Priority Manager. The DS8000 I/O Priority Manager enables more effective storage consolidation and performance management combined with the ability to align quality of service (QoS) levels to separate workloads in the system.

With DS8000 I/O Priority Manager, the system can prioritize access to system resources to achieve the volume's desired QoS based on defined performance goals (high, medium, or low) of any volume. I/O Priority Manager constantly monitors and balances system resources to help applications meet their performance targets automatically, without operator intervention. Starting with DS8000 Licensed Machine Code (LMC) level R6.2, the DS8000 I/O Priority Manager feature supports open systems and IBM System z.

DS8000 I/O Priority Manager, together with IBM z/OS Workload Manager (WLM), provides more effective storage consolidation and performance management for System z systems. Now tightly integrated with Workload Manager for z/OS, DS8000 I/O Priority Manager improves disk I/O performance for important workloads. It also drives I/O prioritization to the disk system by allowing WLM to give priority to the system's resources automatically when higher priority workloads are not meeting their performance goals. Integration with zWLM is exclusive to DS8000 and System z systems.

The paper is aimed at those who want to get an understanding of the DS8000 I/O Priority Manager concept and its underlying design. It provides guidance and practical illustrations for users who want to exploit the capabilities of the DS8000 I/O Priority Manager.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks