Session V62

z/VM Resource Manager (VMRM)

Christine Casey

Senior Software Engineer

z/VM Development, Endicott, NY

IBM System z Technical Conference

April 2007 – Munich, Germany

DISCLAIMER

The information contained in this document has not been submitted to any formal IBM test and is distributed on an "AS IS" basis without any warranty either express or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

In this document, any references made to an IBM licensed program are not intended to state or imply that only IBM's licensed program may be used; any functionally equivalent program may be used instead.

Any performance data contained in this document was determined in a controlled environment and, therefore, the results which may be obtained in other operating environments may vary significantly. Users of this document should verify the applicable data for their specific environments.

It is possible that this material may contain reference to, or information about, IBM products (machines and programs), programming, or services that are not announced in your country. Such references or information must not be construed to mean that IBM intends to announce such IBM products, programming or services in your country.

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

CICS*
DB2*
DB2 Connect
DB2 Universal Database
DFSMS/MVS*
DFSMS/VM*
e business(logo)*
Enterprise Storage Server*
ESCON*
FICON
GDDM*
HiperSockets
IBM*
IBM(logo)*
*

Language Environment*
MQSeries*
Multiprise*
MVS
NetRexx
OpenEdition*
OpenExtensions
OS/390*
Parallel Sysplex*
PR/SM
QMF
RACF*
RAMAC*
RISC

S/370
S/390*
S/390 Parallel Enterprise Server
VisualAge*
VisualGen*
VM/ESA*
VTAM*
VSE/ESA
WebSphere*
z/Architecture
z/OS*
zSeries*
z/VM*

* Registered Trademanks of IBM Corporation

The following are trademarks or registered trademarks of other companies.
Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation.
Tivoli is a trademark of Tivoli Systems Inc.
Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
UNIX is a registered trademark of The Open Group in the United States and other countries.
Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.

Notes:
Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

IBM considers a product "Year 2000 ready" if the product, when used in accordance with its associated documentation, is capable of correctly processing, providing and/or receiving date data within and between the 20th and 21st centuries, provided that all products (for example, hardware, software and firmware) used with the product properly exchange accurate date data with it. Any statements concerning the Year 2000 readiness of any IBM products contained in this presentation are Year 2000 Readiness Disclosures, subject to the Year 2000 Information and Readiness Disclosure Act of 1998.
All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the 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.

Agenda

- VMRM objectives and overview
- Configuration file statements
- Systems Management VMRM-related APIs
- Workload selection and adjustments
- I/O priority queuing
- Monitor data and Performance toolkit
- Cooperative Memory Management
- Summary

VMRM Objectives

- Dynamically tune a system
- Manage workloads to CPU and DASD I/O velocity goals
- Allow I/O priority queuing to be exploited on behalf of VM-based workloads
- Provide an infrastructure for more extensive workload and resource management for future releases of z/VM

Overview

- Shipped as part of CMS component of VM
  - Executables on MAINT's 193 disk
- The Service Virtual Machine: VMRMSVM
  - PROFILE EXEC begins operation of the server by calling the IRMSERV EXEC
    - May also be invoked from the command line
  - IRMSERV reads the customer-supplied definition file
    - Default is VMRM CONFIG A
- Uses VM monitor data
  - Obtains 1-minute interval measurements of virtual machine resource consumption

Overview (cont.)

- Based on definition of workloads, goals, and priorities in the configuration file, the SVM…
  - Computes the achievement levels of interest (actuals) for each workload
  - Selects one workload to adjust:
    - For each goal type of CPU or DASD
    - Based on the customer-supplied importance value
  - Adjusts virtual machine tuning parameters to achieve defined goals
    - Using CP Commands Set Share and Set IOPriority
    - Issued for “eligible” guests in the workload

Overview (cont.)

- VMRM Cooperative Memory Management (VMRM-CMM)
  - A collaboration between VM and Linux to optimize memory management
  - Linux guests to be notified are identified in the VMRM configuration file, treated with equal priority
  - VMRM tracks system memory utilization/demand and computes target "resident footprint" for each guest
  - VMRM sends SMSG to guests to adjust footprint
  - Guest device driver receives messages
    - Uses existing guest logic to return the least valuable pages

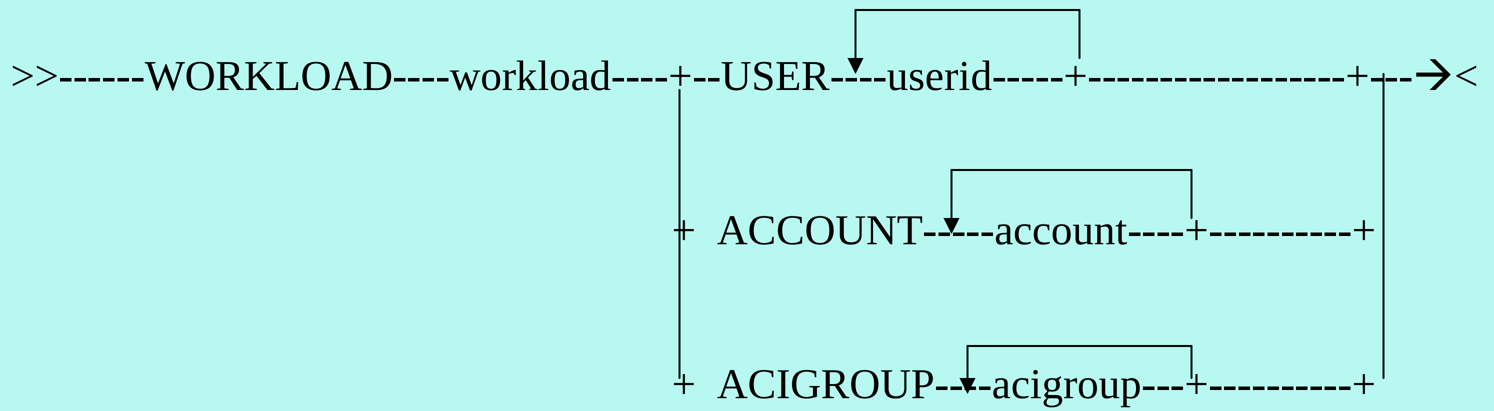
VMRM Configuration File Statements

- **WORKLOAD** - describes a workload by userid, account id, or acigroup
- **GOAL** - describes a DASD or CPU velocity goal
- **MANAGE** - associates a workload with a goal and assigns an importance value
- **ADMIN** - identifies a user to receive VMRM messages and/or a new config file name
- **NOTIFY** - identifies Linux user(s) to be notified when system memory is constrained

WORKLOAD Statement

- A workload is comprised of one or more virtual machines identified by user ID, account ID, or ACI group name

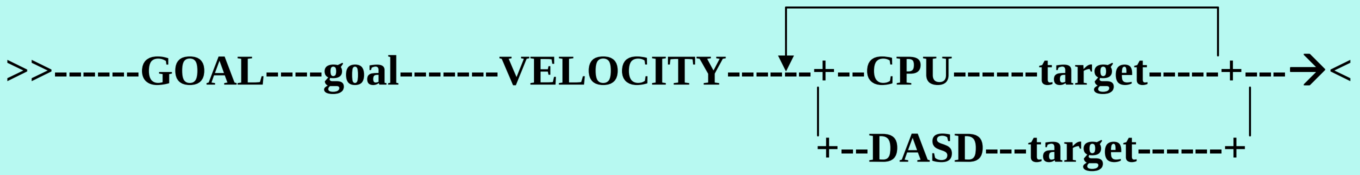
  ex: WORKLOAD work1 USER Linux* chrisC Alan



```
>>------WORKLOAD----workload----+--USER----userid-----+------------------+---→<
                                |                                         |
                                + ACCOUNT-----account----+----------+
                                |
                                + ACIGROUP----acigroup---+----------+
```

GOAL Statement

- A GOAL statement specifies velocity goals for:
  - **CPU**: percentage of time the user should receive CPU resources when it is ready
  - **DASD**: percentage of time that the user’s DASD I/O requests are not outprioritized



```
>>------GOAL----goal-------VELOCITY------+--CPU------target-----+---→<
                                          +--DASD---target------+
```

MANAGE Statement

- Associates a workload with a goal
- Assigns an importance value to the relationship
  - – value range 1-10 (10 is most important)
- Only one manage statement is allowed for each workload

```
>>-----MANAGE----workload------GOAL----goal----IMPORTANCE----value----→<
```

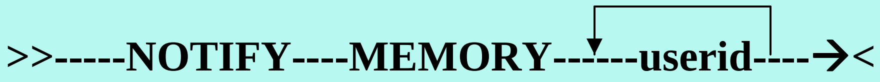
ADMIN Statement

- **MSGUSER** specifies a user ID on the same system where messages can be sent by VMRM
  - Also logged to “VMRM LOG1 A” logfile
- **NEWCFG** specifies a new configuration file on an SFS directory
  - Allows dynamic restart of the server with a new configuration

```
>>-----ADMIN----MSGUSER----userid----NEWCFG----fn----ft----dirid----→<
```

NOTIFY Statement

- Notifies specified Linux users when there is memory constraint in the system
  - Collaborates with Linux guest via SMSG
- Supports Cooperative Memory Management in 5.2.0 (VMRM-CMM) with APAR VM64085



```
>>-----NOTIFY----MEMORY-------userid----→<
```

Sample VMRM Configuration File

```
*   This is a valid comment line  *
/*  So is this                    */
;   and this
ADMIN    MSGUSER  Chris,
         NEWCFG   Mycfg config VMSYS:VMRMSVM.
WORKLOAD work1    USER abcde,
                  a123 456
WORKLOAD work2    USER fghij*
WORKLOAD workabcd USER qrst
WORKLOAD work3    ACCOUNT 1234 5678
WORKLOAD work4    ACIGROUP  ABC
GOAL   goal1,     /* continuation allowed */
             VELOCITY CPU  10
GOAL   goal2 VELOCITY DASD 50
GOAL   goal3 VELOCITY CPU  80  DASD 20
MANAGE work1 GOAL goal1,
                        IMPORTANCE 10
MANAGE work2 GOAL goal1 IMPORTANCE  5
MANAGE work3 GOAL goal2 IMPORTANCE  2
MANAGE work4 GOAL goal3 IMPORTANCE 10
MANAGE workabcd   GOAL  goal2 IMPORTANCE  7
NOTIFY MEMORY  Linux1 Linux5 LinUserX
```

Configuration File APIs

- Systems Management APIs for VMRM
  - VMRM_Configuration_Update
    - Updates a VMRM configuration file remotely from a client using the NEWCFG support
  - VMRM_Configuration_Query
    - Query a VMRM configuration file remotely from a client
  - VMRM_Measurement_Query
    - Query workload measurements from a client - - returns workload goal and actual data

Verifying a Configuration File

- SYNCHECK option on server invocation

  IRMSERV TEST CONFIG A (syncheck

  - Syntax checks a configuration file without starting the server
  - Allows Class G users to check a configuration file before it is put into use by the server
  - VMRM_Configuration_Update API always performs syncheck before updates go into production

VMRM Log File

- **VMRM LOG1 A** file is used to log:
  - Messages sent to MSGUSER
  - VMRM events, monitor fields, commands issued
  - Measurement data
  - Debug messages

- VMRM **LOG1** A will be copied to VMRM **LOG2** A when it reaches 10,000 records.
  - VMRM LOG1 will then be erased and rewritten

Sample VMRM Log File

```
2007-03-28   17:02:02   ServExe MSG

MSG        IRMSER0022I  VM Resource Manager Initialization Started
PCfg       VMRM CONFIG A1  03/28/07  17:01:55
MSG        IRMSER0008W  The ADMIN message user ID is not logged on …
InitEnv    Monitor sample started – recording is pending
InitEnv    HCPMNR6224I  Sample recording is pending because there …
InitEnv    MONITOR EVENT INACTIVE   BLOCK     4    PARTITION     0
InitEnv    MONITOR DCSS NAME  -  NO DCSS NAME DEFINED
InitEnv    CONFIGURATION SIZE          68 LIMIT          1  MINUTES
InitEnv    CONFIGURATION AREA IS FREE
InitEnv    USERS CONNECTED TO *MONITOR – NO USERS CONNECTED
InitEnv    ….
InitEnv    ….   more data from Q Monitor command ….
MSG        IRMSER0023I  VM Resource Manager Initialization complete.
           Proceeding to connect to Monitor.
Exit       STARMON completed.  RC=0
ExitSVM    Monitor sample stopped
MSG        IRMSER0012I  VM Resource Manager shutdown in progress
```

Some Terminology

- Absolute vs. Relative
  - Used to prioritize real CPU consumption
  - **Absolute** specifies a user is to receive a target minimum of nnn% of the scheduled system resources
  - Amount of resources available to relative share users = total resources available less the amount allocated to absolute share users
  - **Relative** portion that a user receives is nnnn/sum of all relative share users
  - VM Resource Manager will **not** adjust Absolute users

- Limithard vs. Limitsoft
  - **Limithard** specifies the user’s share of CPU resource is limited (can’t receive more than maximum share of CPU resource)
  - **Limitsoft** specifies the user’s share of CPU resource is limited, **but** the limit can be exceeded if the capacity is available

Workload Selection Criteria

- Workloads are selected first based on importance value
- If a workload was selected in the last interval either for improvement or degradation, it is skipped and an attempt is made to select another
- If there are workloads of equal importance, the workload farthest from its goal is selected
- Eligible users within a workload will have their SHARE or IOPRIORITY adjusted appropriately based on how far they are from the workload goal

Workload Adjustment Criteria

- Individual users within selected workload may be adjusted based on calculations from monitor data
  - User must have **Relative** Share and I/O Priority settings
  - User does not have **Limithard** specified for CPU Share
  - Sum of wait and run deltas is > current sample size of 5
  - Sum of I/O and Outprioritized deltas is > current sample size of 5
  - CPU actual = run delta / (run delta + wait delta) * 100
  - DASD actual = IO delta / (IO delta + outprior delta) * 100
- If above criteria is met and user is not within 5% of goal, then they can be adjusted

Adjustment Algorithms

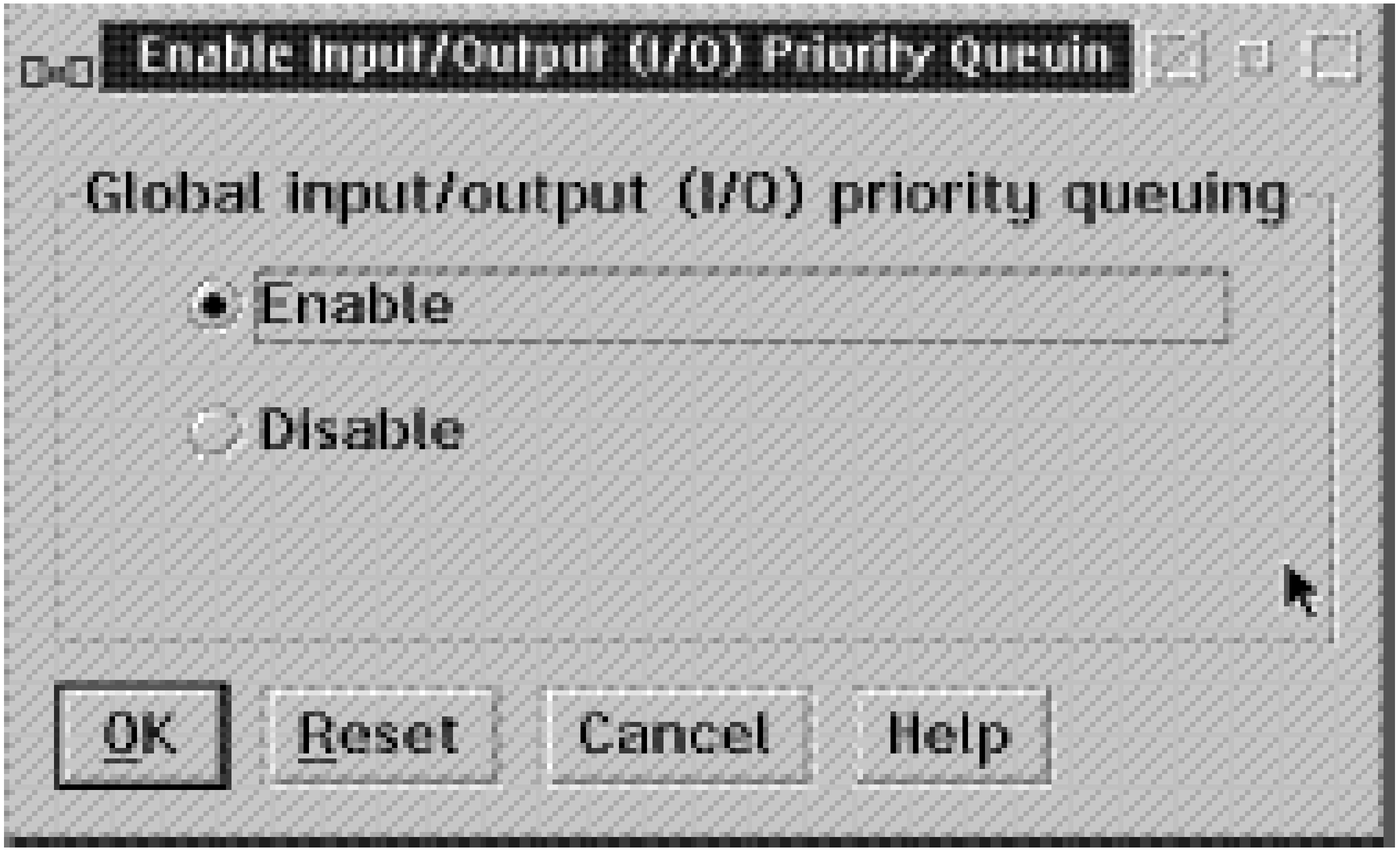
- Determine how much to adjust each user
  - relvalue = (CPU goal / actual) * User current share
  - relvalueLo =(DASD goal / actual) * User current IO Lo
  - relvalueHi = relvalueLo + (curr Hi – curr Lo)
- Set Share and/or Set IOPriority command is issued on behalf of the user

I/O Priority Queuing

- Enables prioritization of virtual machine I/O
  - If I/O Priority queuing is available and enabled:
    - Queuing low/high range is obtained from the hardware
    - Guest queuing values are mapped to fall within that range
    - CP I/O uses highest value available
  - If not available nor enabled, CP simulates range of 0-255
  - Range may be changed/set by the CP SET IOPRIORITY command or IOPRIORITY directory statement
  - For I/O priority-aware guests, the priority associated with the guest I/O requests will be enforced, otherwise CP assigns a value.

Enabling I/O Priority Queuing on zSeries Processors

From the Hardware Management Console (HMC)
use the “Enable I/O Priority Queuing” task



Enable Input/Output (I/O) Priority Queuin

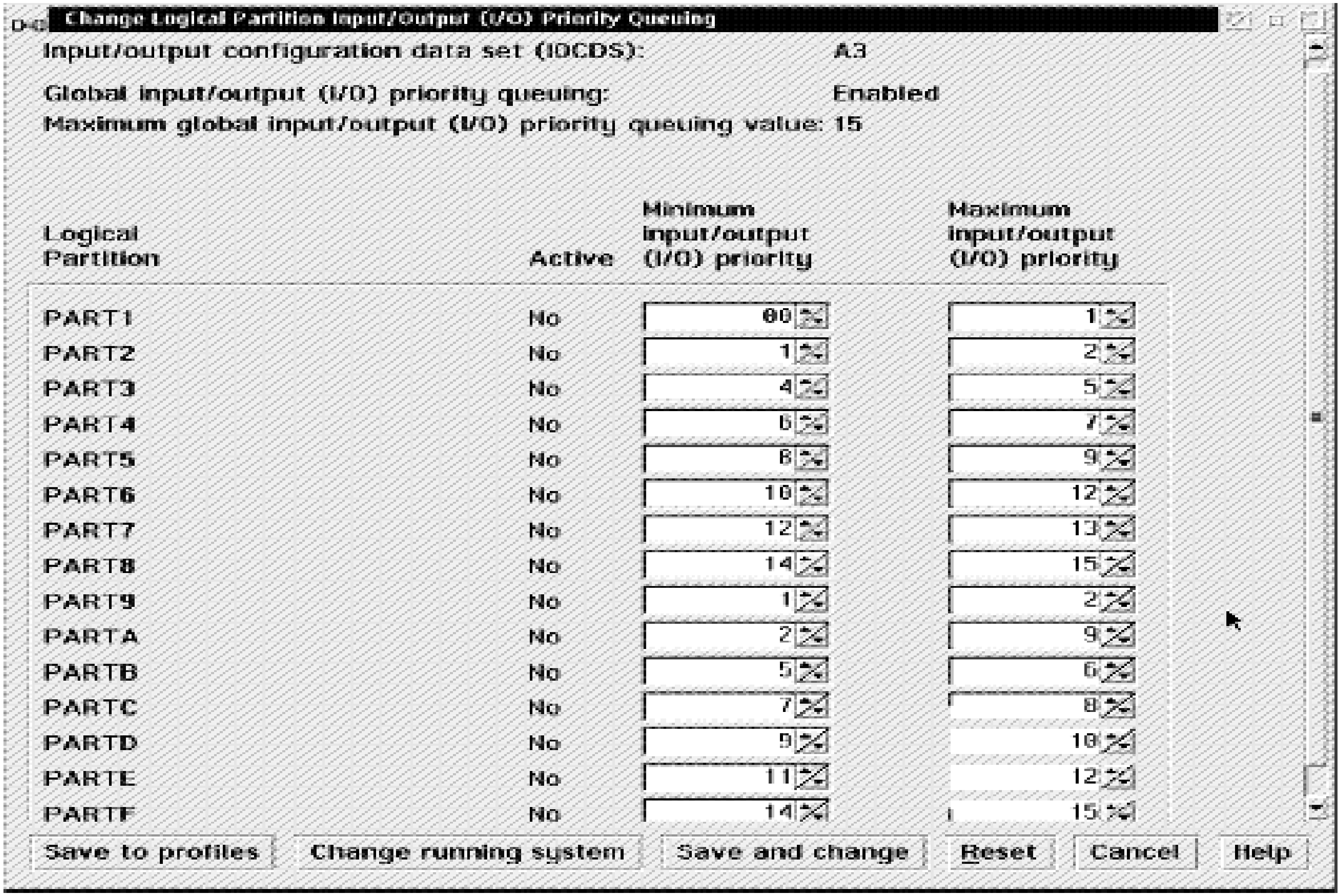
Global input/output (I/O) priority queuing

● Enable

○ Disable

Setting Hardware I/O Priority Queuing Ranges

Use the “Change LPAR I/O Priority Queuing” task to set minimum and maximum I/O priority queuing values


Change Logical Partition Input/Output (I/O) Priority Queuing

Input/output configuration data set (IOCDS): A3

Global input/output (I/O) priority queuing: Enabled

Maximum global input/output (I/O) priority queuing value: 15

Logical Partition	Active	Minimum input/output (I/O) priority	Maximum input/output (I/O) priority
PART1	No	00	1
PART2	No	1	2
PART3	No	4	5
PART4	No	6	7
PART5	No	8	9
PART6	No	10	12
PART7	No	12	13
PART8	No	14	15
PART9	No	1	2
PARTA	No	2	9
PARTB	No	5	6
PARTC	No	7	8
PARTD	No	9	10
PARTE	No	11	12
PARTF	No	14	15

IOPRIORITY
Directory Statement

- Specifies the I/O priority range to be set when the user logs on
- If hardware priority queuing is available and enabled…
  - Absolute priority ranges outside the range available to CP are clipped to fall within that range
  - Relative ranges are mapped to fall within the range available to CP
- If IOPRIORITY is not specified in the directory, low and high are set to a relative value of 0

Set and Query Commands

- CP Set IOPRIORITY (class A privilege)
  - Set IOPRIORITY {userid | *} {Absolute | Relative} low {low value | high value}
  - Absolute must fit in range available to CP (or it will be clipped)
  - Relative maps proportionally to the available range
- CP Query IOPRIORITY (class A or E)
  - Query IOPRIORITY {userid | * | system}
    - **userid** requests the range of a given user
    - ***** requests the range of the user issuing the command
    - **system** requests the priority range available to CP

Query IOPRIORITY Responses

- userid   REQUESTED RANGE nnn mmm ABSOLUTE

  EFFECTIVE RANGE   xxx yyy

- userid   REQUESTED RANGE nnn mmm RELATIVE

  EFFECTIVE RANGE   xxx yyy

where:

- **requested range** indicates low and high ranges requested
- **effective range** is the low and high range that CP will allow for this user

Example of Absolute I/O Priority Queuing Ranges

- If the I/O priority queuing range available to CP is 50-75…
  - Virtual machine requests for ranges from 0-49 will be assigned an absolute value of 50
  - Virtual machine requests for ranges 50-75 will be accepted without change
  - Virtual machine requests for ranges 75-255 will be assigned an absolute value of 75

Example of Relative I/O Priority Queuing Ranges

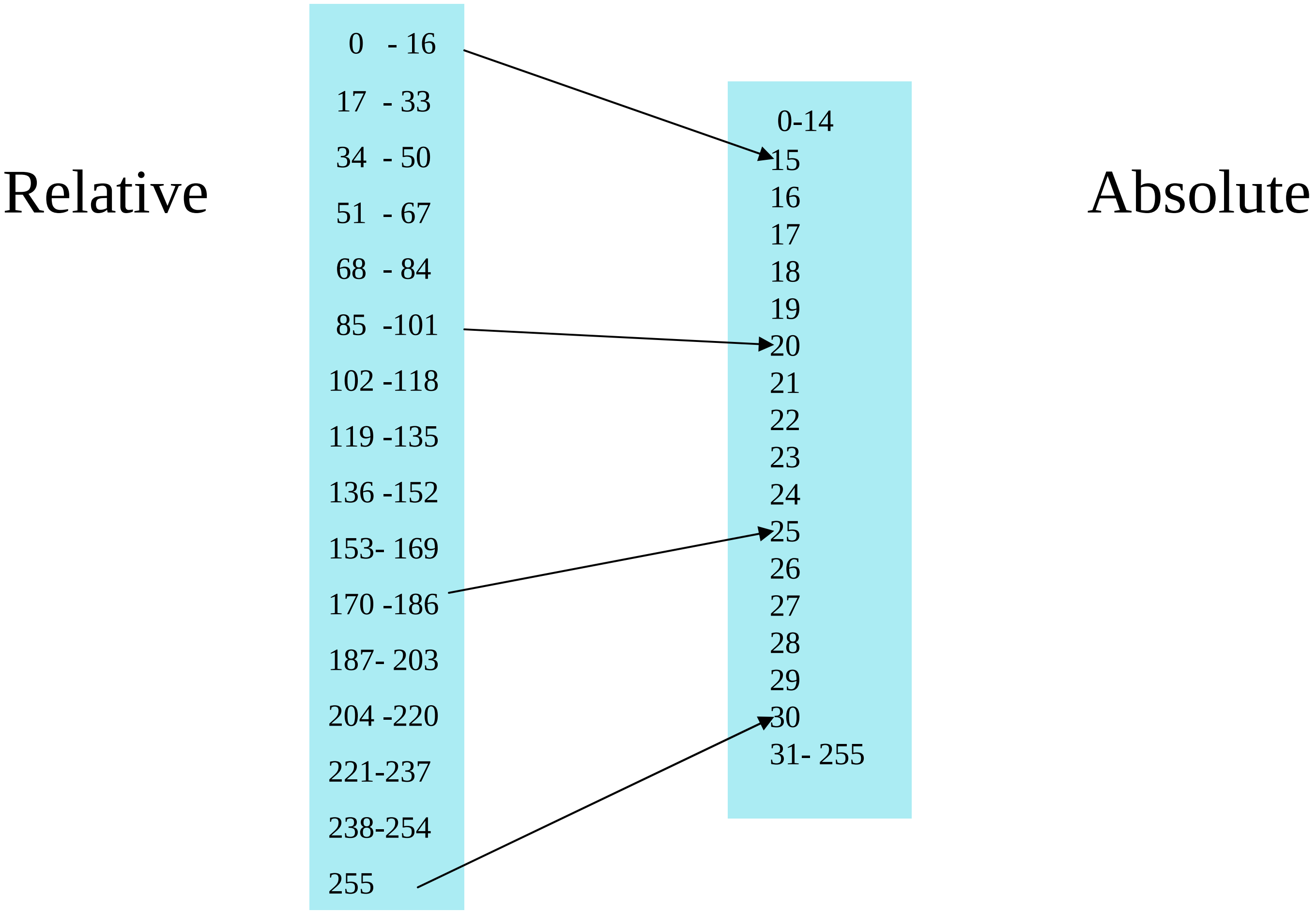
- The effective value is calculated from the requested value and the range available to CP

$$\text{Eff\_val} = \text{Trunc}\left(\frac{\text{Rel\_val} + (\text{CPhi} - \text{CPlo})}{255}\right) + \text{CPlo}$$

Eff_val is the effective I/O priority
Rel_val is the relative I/O priority
CPhi   is the highest I/O priority value available to CP
CPlo   is the lowest I/O priority value available to CP

Example of Relative I/O Priority Queuing Ranges

- If the range of I/O priority values available to CP is 15-30 then relative priorities map to absolute as follows:



Relative

0 - 16
17 - 33
34 - 50
51 - 67
68 - 84
85 -101
102 -118
119 -135
136 -152
153- 169
170 -186
187- 203
204 -220
221-237
238-254
255

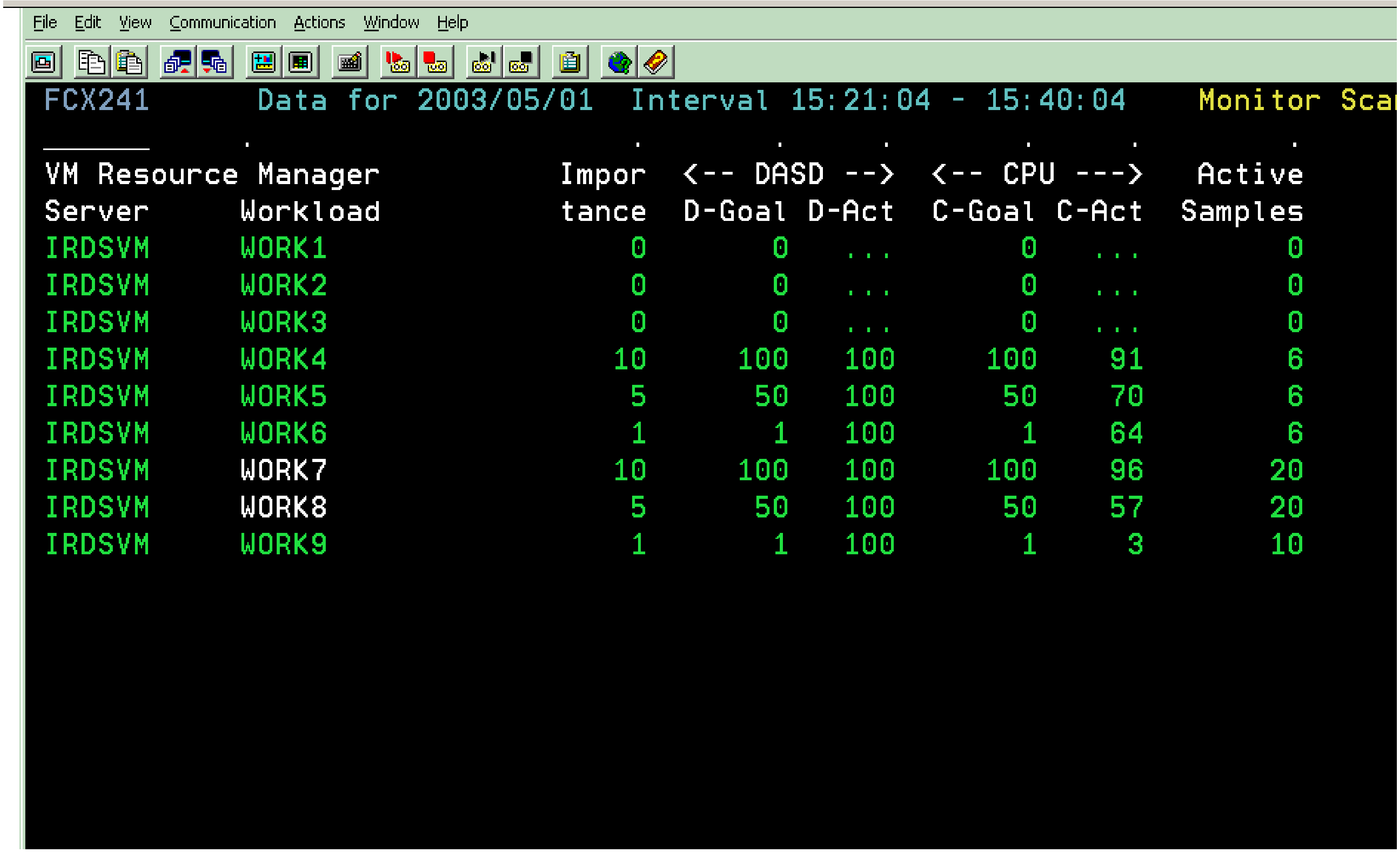
Absolute

0-14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31- 255

Monitor Data

- VMRM Application Monitor Data (APPLDATA) is provided
- Shows workloads, goals, and actual workload achievements
- Performance Toolkit for VM is enhanced to interpret this data
  - detects when a new configuration file is put into production and refreshes data accordingly
- Documented in the z/VM Performance publication - Appendix G

Performance Toolkit screen with VMRM data



```
FCX241       Data for 2003/05/01  Interval 15:21:04 - 15:40:04    Monitor Sca
______            .                           .       .      .       .      .        .
VM Resource Manager                  Impor  <-- DASD -->  <-- CPU --->   Active
Server      Workload                 tance  D-Goal D-Act  C-Goal C-Act  Samples
IRDSVM      WORK1                        0       0   ...       0   ...        0
IRDSVM      WORK2                        0       0   ...       0   ...        0
IRDSVM      WORK3                        0       0   ...       0   ...        0
IRDSVM      WORK4                       10     100   100     100    91        6
IRDSVM      WORK5                        5      50   100      50    70        6
IRDSVM      WORK6                        1       1   100       1    64        6
IRDSVM      WORK7                       10     100   100     100    96       20
IRDSVM      WORK8                        5      50   100      50    57       20
IRDSVM      WORK9                        1       1   100       1     3       10
```

Cooperative Memory Management

- VMRM-CMM support in base z/VM 5.3.0, and enablement APAR VM64085 for z/VM 5.2.0
- Notifies Linux guests specified in the VMRM Notify list when there is memory constraint
  - Communicates via SMSG
  - SHRINK message to suggest how much memory to release
    - Based on calculations from various CP Monitor data fields
- Linux guest that is “CMM-aware” will release pages via Diagnose x’10’
- Subsequent SHRINK messages may also indicate how much memory to reclaim
  - If SHRINK value is less than previous value

Summary

- Use VMRM to dynamically tune your system
- Manage guests in workloads according to CPU and DASD velocity goals
- Enables Cooperative Memory Management between VM and enabled Linux guests
- Designed to easily add more management constructs in the future

Questions ?

Contact Info: caseyct@us.ibm.com

Documentation: z/VM Performance, SC24-6109-00

Webpage: http://www.vm.ibm.com/sysman/vmrm/

Linux updates for CMM support:

http://www-128.ibm.com/developerworks/linux/linux390/linux-2.6.5-s390-34-april2004.html