z/OS Workload Manager

WLM – Update for z/OS Release 8 and 9
The Latest and Greatest

Horst Sinram

WLM Development

IBM Corporation, Boeblingen, Germany

mailto:sinram@de.ibm.com

09/28/2007

Trademarks

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

APPN*	HiperSockets	OS/390*	VM/ESA*
CICS*	HyperSwap	Parallel Sysplex*	VSE/ESA
DB2*	IBM*	PR/SM	VTAM*
DB2 Connect	IBM eServer	Processor Resource/Systems Manager	WebSphere*
DirMaint	IBM e(logo)server*	RACF*	z/Architecture
e-business logo*	IBM logo*	Resource Link	z/OS*
ECKD	IMS	RMF	z/VM*
Enterprise Storage Server*	Language Environment*	S/390*	z/VSE
ESCON*	MQSeries*	Sysplex Timer*	zSeries*
FICON*	Multiprise*	System z	z9
GDPS*	NetView*	System z9	
Geographically Dispersed Parallel Sysplex	On demand business logo	TotalStorage*	
		Virtualization Engine	

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries

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

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.

Red Hat, the Red Hat "Shadow Man" logo, and all Red Hat-based trademarks and logos are trademarks or registered trademarks of Red Hat, Inc., in the United States and other countries.

SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.

* All other products may be trademarks or registered trademarks of their respective companies.

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.

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.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

ENABLING BUSINESS.
A THROUGH Z.

Agenda

- New function overview

- Enhancements for Blocked Workloads
- Promotion of Canceled Jobs
- Starting Minimum Number of Servers Without Delay

- zAAP/zIIP Enhancements
- RSM/SRM Support for Large Real storage
- Group Capacity Limit

- Support for DB2 WLM Assisted Buffer Pool Management
- JES2 Enhancements for WLM-managed batch initiators
- New Resource Group Types
- Routing Enhancements

- WLM Tools
  - User Friendly Interface for WLM Administrative Application
  - Monitoring Tool for Application Environments
- Summary

Overview of New Functions (1) – zIIP and zAAP

Function \ Release	z/OS R9	z/OS R8	z/OS R7 w/ JBB772S	z/OS R7	z/OS R6 w/ JBB77S9
Routing Support for zAAP and zIIP	☺				
IIPHONORPRIORITY	☺	OA20045	OA21993		
zIIP Reporting Support		☺	☺ +OA16005		☺ +OA16005
zAAP Management Support		☺			
zAAP Processing Enhancements		☺	☺	OA14131 + OA13953	☺
zAAP Reporting Support			☺		

Overview of New Functions (2) – Storage and Routing

Function \ Release	z/OS R9	z/OS R8	z/OS R7	z/OS R6
z/OS R8 Coexistence	n/a	n/a	OA13837	OA13837
RSM/SRM Support for >128GB real storage	☺	☺		
Support for DB2 WLM Assisted Buffer Pool Mgmt	OA18461	OA18461		
Capacity based routing for WAS	☺	OA16486	OA16486	OA16486
Routing Support for zAAP and zIIP	☺			
Enhancements in LPAR free capacity computations	☺	☺	☺	OA10006
Routing (IWMSRSRS) supports >96 (up to 300) servers	☺	OA18531	OA18531	OA18531

Overview of New Functions (3) - Other

Enhancement	z/OS R9	z/OS R8	z/OS R7	z/OS R6
Process Entitlement Type 2 and 3 Resource Groups	☺		z/OS R8 Coexistence Support (OA13837) on previous chart	
Enhancements for blocked workloads	☺	**OA17735 (*)** (R8 + JBB772S)		
Promote Cancelled Jobs	☺			
Start min number of servers without delay	☺			
EWLM Monitoring Support 1	☺			**OA07196**
WLM-managed Initiator Start Considers System Affinities	☺		**OA10814**	

*: The function is not active with the IEAOPT default settings provided via this APAR

Agenda

- New function overview

- Enhancements for Blocked Workloads
- Promotion of Canceled Jobs
- Starting Minimum Number of Servers Without Delay

- zAAP/zIIP Enhancements
- RSM/SRM Support for Large Real storage
- Group Capacity Limit

- Support for DB2 WLM Assisted Buffer Pool Management
- JES2 Enhancements for WLM-managed batch initiators
- New Resource Group Types
- Routing Enhancements

- WLM Tools
  – User Friendly Interface for WLM Administrative Application
  – Monitoring Tool for Application Environments
- Summary

Support for "Blocked Workloads"

- **Blocked workloads:**
  - Lower priority work may not get dispatched for an elongated time
  - May hold a resource that more important work is waiting for

- **WLM allows some throughput for blocked workloads**
  - By dispatching low important workload from time to time, these "blocked workloads" are no longer blocked
  - Helps to resolve resource contention for workloads that have no resource management implemented

IEAOPT BLWLTRPCT and BLWLINTHD

BLWLTRPCT	Percentage of the CPU capacity of the LPAR to be used for promotion ▪Specified in units of 0.1% ▪Default on z/OS R9 is 5 (=0.5%) –Disabled on previous releases (OA17735) ▪Maximum is 200 (=20%)
BLWLINTHD	Specifies threshold time interval for which a blocked address space or enclave must wait before being considered for promotion. ▪Minimum is 5 seconds. Maximum is 65535 seconds. ▪Default is 60 seconds.

Blocked Workload Support: Implementation

Time Slices

BLWRTRPCT
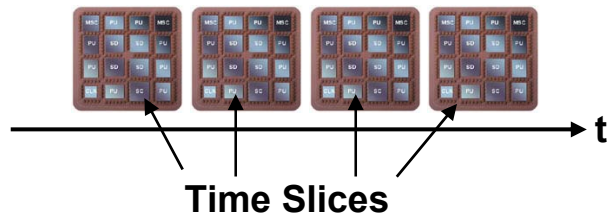
- ▪ **Usually work uses the CPU for a predefined amount of time**
  - – Named a time slice
    - • After that period the work gets re-dispatched
  - – z/OS uses a major and a minor time slice

- ▪ **Approach**
  - – Convert the capacity of the LPAR

$$LPAR\ WCap = \frac{LPAR\ Weight}{Total\ Weight + Phantom\ Weight} \bullet Shared\ Regular\ CPs$$

$$LPAR\ Capacity = MIN(LPAR\ WCap;\ Logical\ Regular\ CPs\ of\ LPAR)$$

  - – Define a percentage value of how many time slices can be used for blocked workloads (as trickles)
  - – Parameter: BLWLTRPCT
    - • Values: 0.. 200
    - • 200 = 20%
    - • Default: 5 = 0.5% of the LPAR capacity

Enhancements for Blocked Workloads - Example
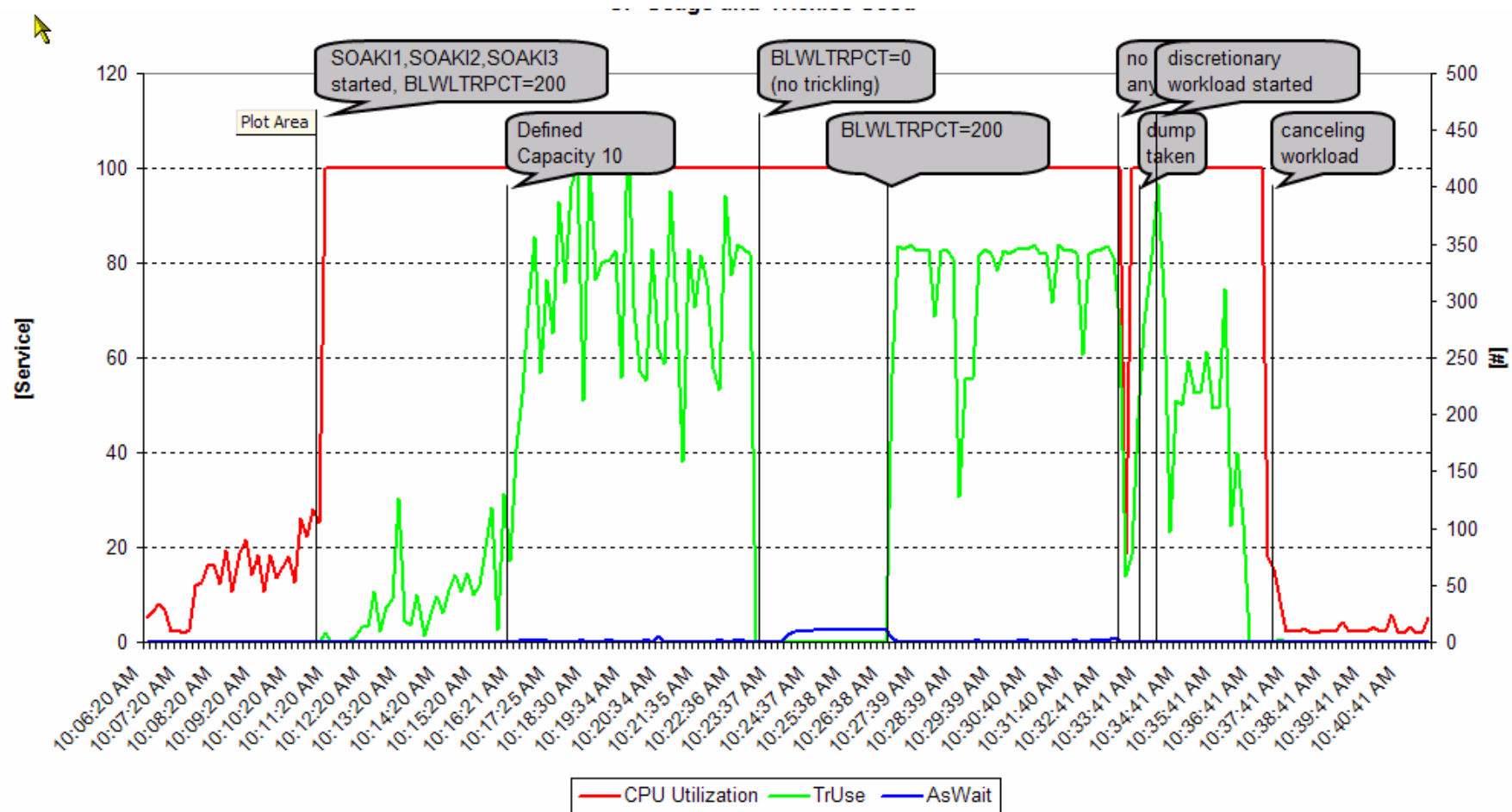
Enhancements for Blocked Workloads
RMF Support

- ## RMF provides enhancements:

  - Extensions of RMF Postprocessor CPU Activity and WLMGL reports with information about blocked worklads and the temporary promotion of their dispatching priority

  - SMF record 70-1 (CPU activity) and SMF 72-3 (Workload activity)

```
                        C P U   A C T I V I T Y
                              …



BLOCKED WORKLOAD ANALYSIS


OPT PARAMETERS: BLWLTRPCT (%)   0.5   PROMOTE RATE:   DEFINED    50000   WAITERS FOR PROMOTE:  AVG    0.001
                BLWLINTHD       60                    USED (%)      95                         PEAK      15
```

Starting Minimum Number of Servers Without Delay Overview

- Problem: **If a WLM managed server performs heavy processing before connecting to WLM, the startup of the minimum amount of servers can take a long time.**

- Solution:

  - When requested by the managed server, WLM starts the minimum amount of servers without waiting on the connect status of the previously started servers.

Starting Minimum Number of Servers Without Delay Invocation

- **The support is invoked by the managed server via:**
  - New parameter on the Application Environment Limit Service macro

IWM4SLI —— AE_SERVERMIN = **ServerMin** —— START_MINIMUM = SERIAL

START_MINIMUM = PARALLEL ... Ne
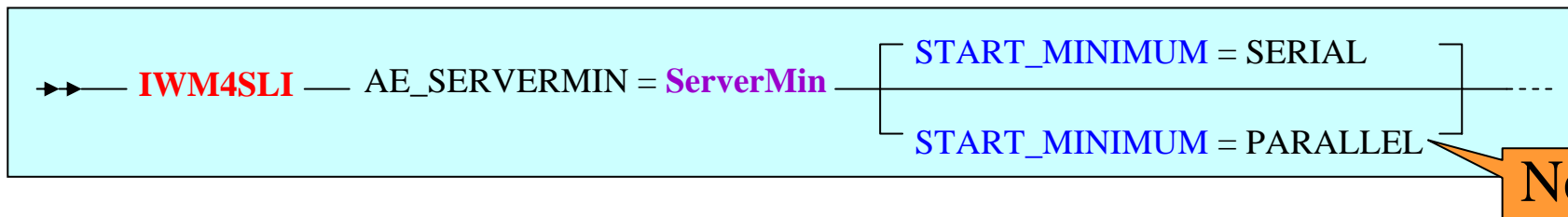
- **Called by:**
  - WLM started server address spaces.
    The first server address space decides in which way the remaining servers will be started.

Agenda

- New function overview

- Enhancements for Blocked Workloads
- Promotion of Canceled Jobs
- Starting Minimum Number of Servers Without Delay

- zAAP/zIIP Enhancements
- RSM/SRM Support for Large Real storage
- Group Capacity Limit

- Support for DB2 WLM Assisted Buffer Pool Management
- JES2 Enhancements for WLM-managed batch initiators
- New Resource Group Types
- Routing Enhancements

- WLM Tools
  – User Friendly Interface for WLM Administrative Application
  – Monitoring Tool for Application Environments
- Summary

Initial Handling of zAAP-Related IEAOPTxx Parameters

IFACROSSOVER	IFAHONORPRIORITY	Meaning
<u>YES</u>	<u>YES</u>	**CP's help zAAP's honoring priorities until discretionary work is reached.** **CP's always help zAAP discretionary work after all CP work.** – With soft-capping in effect: See next row
YES	NO	**CPs always help zAAP's after all CP work**
NO	YES	
NO	NO	**CPs never help zAAPs, except when no zAAPs are operational**

Changed

Release 8: IFAHONORPRIORITY Enhancements

- **Applies to**

  - z/OS Release 8, and

  - z/OS Release 6 and 7 with the JBB77S9, or JBB772S FMIDs

- **IEAOPT Parameters:**

  - **IFACROSSOVER** now obsolete

  - **IFAHONORPRIORITY**:

    - YES: General purpose processors dispatch regular work <u>and</u> Java work in priority order, <u>if</u> zAAP needs help.
    - No: General purpose processors dispatch regular work only
      - Unless no zAAP is online, or in case of contention

  - **PROJECTCPU**:

    - YES: Reports zAAP-on-CP usage; helps sizing

WLM Treatment of the zAAP Resource

- **A zSeries Application Assist Processor is a new resource type that WLM is aware of**
  - Contributes using and delay samples
  - Contributes service times
  - zAAP utilization is reported by RMF (SPE OA05731)

- **Up to z/OS 1.7 zAAPs are managed by WLM as extension of CPs**
  - Java work executing on zAAPs inherits the dispatch priority from its execution on regular CPs
  - Execution is accounted for in execution velocity and goal achievement (PI)

- **zAAP management support**
  - Beginning with z/OS 1.8 work on zAAP is managed independently

- **zAAP service not included in defined capacity computations and resource group management**

- **zAAPs are not varied by IRD Vary CPU Management**

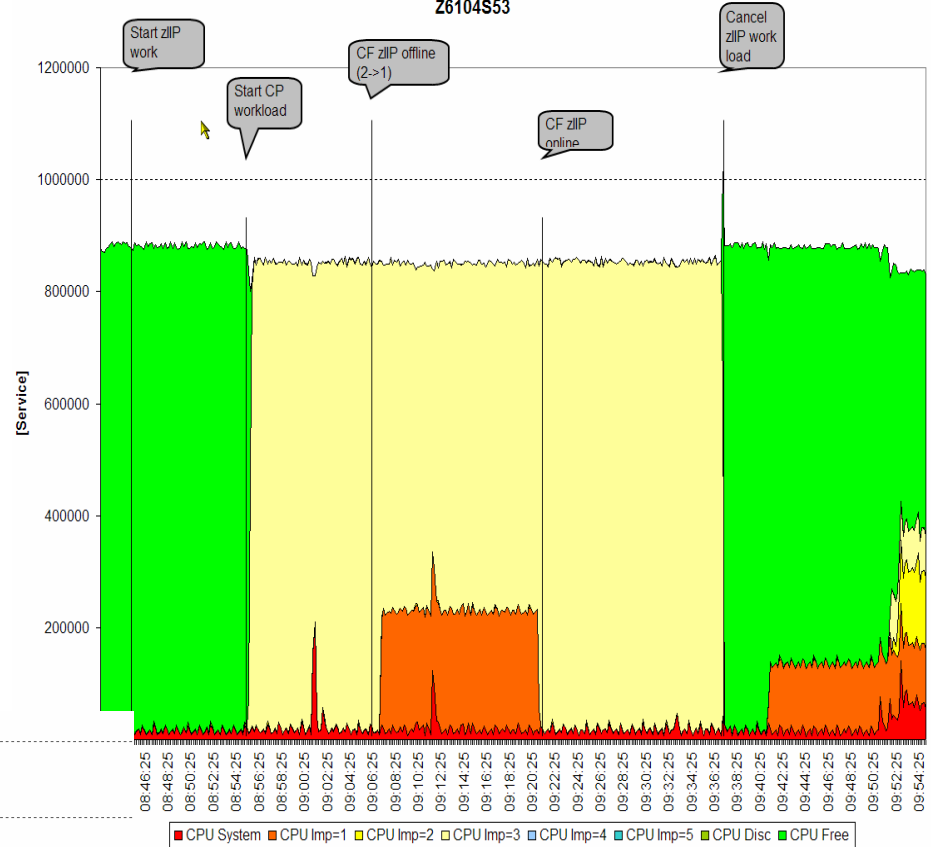- **On system z9 specialty engines, such as zAAPs, are in separate pools.**
  - Previously, on z890 and z990: zAAPs inherited their weight from CP pool but were part of the ICF/IFL pool

IBM System z9 Integrated Information Processor (zIIP)

- **New specialty engine for the System z9 EC and BC mainframes designed to help:**
  - Customers integrate data across the enterprise
  - Improve resource optimization
- **z/OS manages and directs work between the general purpose processor and the zIIP**
  - Number of zIIPs per z9-109 not to exceed number of standard processors
  - No changes anticipated to DB2 for z/OS V8 applications
- **DB2 for z/OS V8 first IBM exploiter of the zIIP with**
  - System z9 EC/BC
  - z/OS 1.6 or later (zIIP FMIDs for R6, R7, integrated into R8 and later)
  - DB2 for z/OS V8
- **More information available on the zIIP web site**
  - http://www.ibm.com/systems/z/ziip/about.html

zIIP Eligible Work and Workflow

- **z/OS dispatches work in either**
  - TCB (Task Control Block) mode or
  - SRB Supervisor Request Block) mode.

- **Pre-emptible enclaves are used to do the work on behalf of the originating TCB or SRB address space.**

- **Only a portion of the enclave SRB work is eligible to be redirected to the zIIP**
  - Client SRB work, non-preemptible SRB work, or TCB work is never eligible

	Task (TCB)	Non-preempt. SRB ▪Local ▪Global	Pre-empt. SRB	Client SRB	Enclave SRB
Pre-emptible	YES	No	Yes		
Type of Service	CPU	SRB	CPU		
Created via	ATTACH	IEAMSCHED, SCHEDULE			
zIIP Eligibility	Never				Controlled by enclave creator

Handling of zIIP Work

- **In many respects WLM/SRM support of zIIPs is equivalent to the zAAP monitoring support with the zAAP processing enhancements**
  - But: support only available via distinct FMIDs pre z/OS R8

- **zIIP work managed as an extension of CP work**

- **zIIP work will flow over to general purpose CPs**
  - Like zAAP work with the zAAP processing enhancements
    - "needs help"
  - Previously: No external controls.
    New IIPHonorPriority allows zAAP-like control.

- **zIIP service not included in defined capacity computations and resource group management**

- **zIIPs are not varied by IRD Vary CPU Management**

- **On System z9 specialty engines, such as zIIPs, are in separate processor pools.**

IEAOPT IIPHONORPRIORITY

IIPHONORPRIORITY	
YES	Standard processors run both zIIP processor eligible and non-zIIP processor eligible work in priority order when the zIIP processors indicate the need for help.
NO	Standard processors will not examine zIIP processor eligible work regardless of the demand for zIIP processors unless no zIIP processor is online, or contention on a suspend lock requires a zIIP eligible unit of work to be dispatched.

zIIP Reporting via RMF

- WLM Services provide data on zIIP usage comparable to data on zAAP

- **RMF Monitor I**
  - CPU and Workload Activity reports
- **RMF Monitor III**
  - CPC, SYSINFO and ENCLAVE reports

- **zIIP Usage projection available without real zIIPs in the configuration**
  - Via IEAOPT parameter PROJECTCPU=YES

- **Detailed information available via** http://www.ibm.com/systems/z/ziip/
  - See "Introduction to System z9 Integrated Information Processor (zIIP)" at http://publibz.boulder.ibm.com/zoslib/pdf/zosr6r7ziip.pdf and the Program Directory http://publibz.boulder.ibm.com/epubs/pdf/i1007730.pdf

Example:
Handling of zIIP Work
with IIPHonorpriority=YES

Service consumed in service class with zIIP work by processor type:
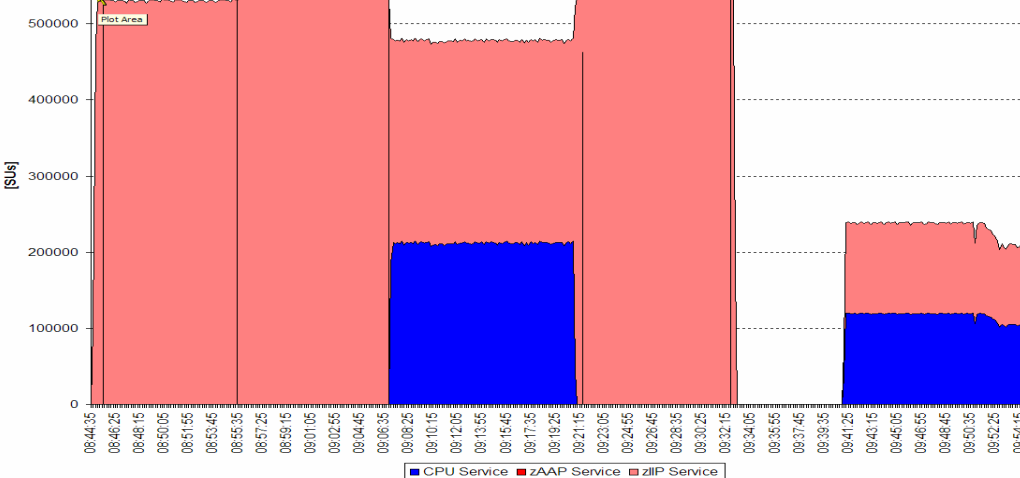
Agenda

- New function overview

- Enhancements for Blocked Workloads
- Promotion of Canceled Jobs
- Starting Minimum Number of Servers Without Delay

- zAAP/zIIP Enhancements
- RSM/SRM Support for Large Real storage
- Group Capacity Limit

- Support for DB2 WLM Assisted Buffer Pool Management
- JES2 Enhancements for WLM-managed batch initiators
- New Resource Group Types
- Routing Enhancements

- WLM Tools
  - User Friendly Interface for WLM Administrative Application
  - Monitoring Tool for Application Environments
- Summary

RSM/SRM Support for Large Real storage

- **Support up to 4 Terabytes of real storage**

- **Page Replacement and UIC**
  - The CPU cost of the current function continues to grow as we move to systems with more and more real storage while its effectiveness is diminishing. This algorithm was changed to perform more efficiently in large real systems.

- **Physical Swap Processing**
  - Physical Swaps to Auxiliary storage were eliminated

- **Pageable Storage Shortages**
  - Used to be resolved by physical swaps
  - They are now resolved by frame exchanges

Page Replacement – pre z/OS R8

LSW / OUT

AS 3

IN

AS 1

AS 2

Page Datasets
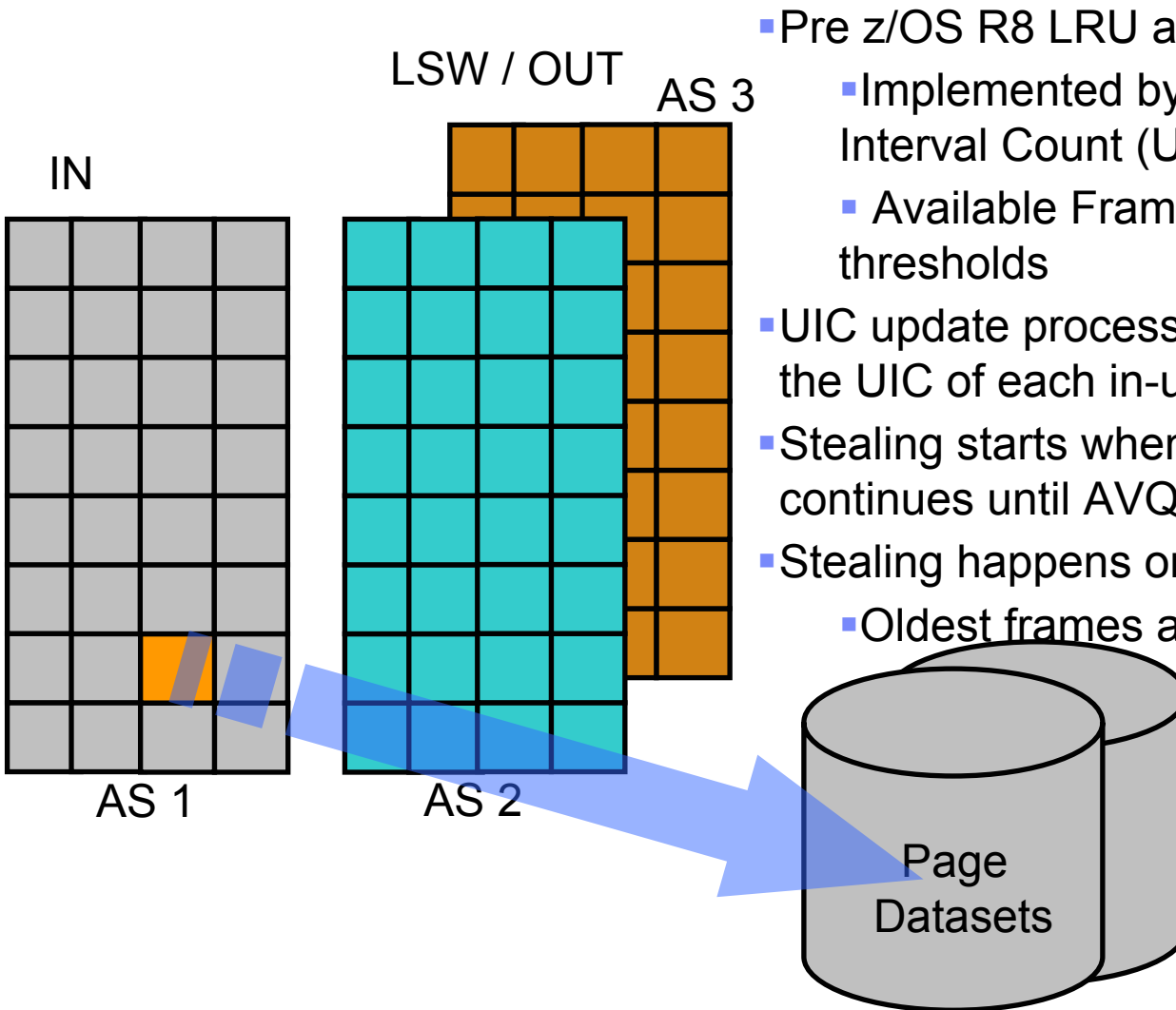
- Pre z/OS R8 LRU algorithms are AS oriented
  - Implemented by keeping an Unreferenced Interval Count (UIC)
  - Available Frame Queue (AVQ) Low/Okay thresholds
- UIC update process runs periodically to update the UIC of each in-use frame
- Stealing starts when we go below AVQ Low and continues until AVQ Okay is reached
- Stealing happens on an address space basis
  - Oldest frames are stolen first

Page Replacement – z/OS 1.8
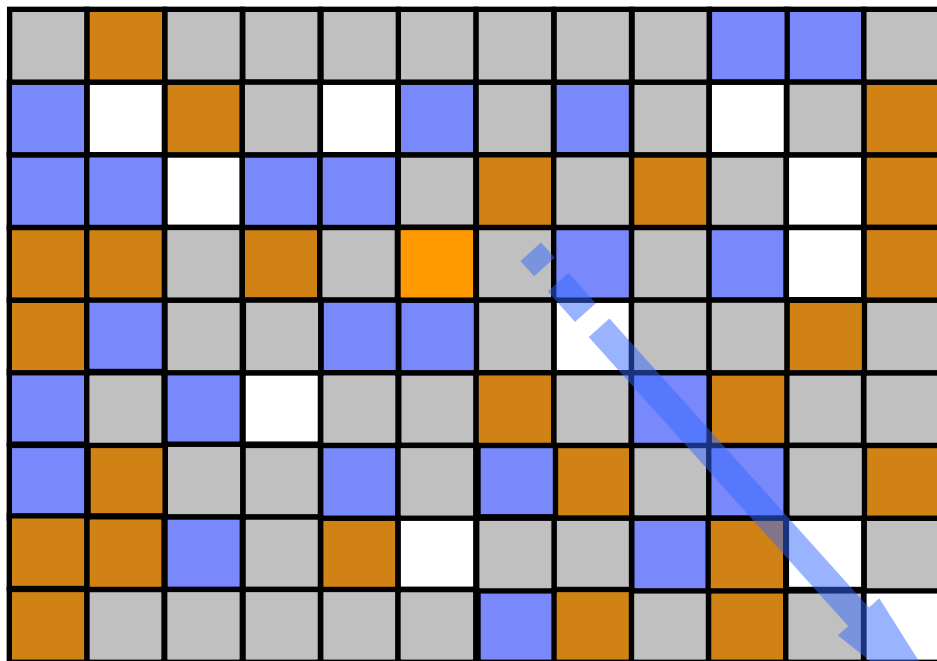
- Why Change?
  - Disruption of the UIC update process is intolerable as the amount of real storage allocated to address spaces increases
- The new page replacement algorithm
  - Runs when the Available Frame Queue needs to be replenished
  - Stealing happens on a global basis
- Storage oriented

Page Datasets
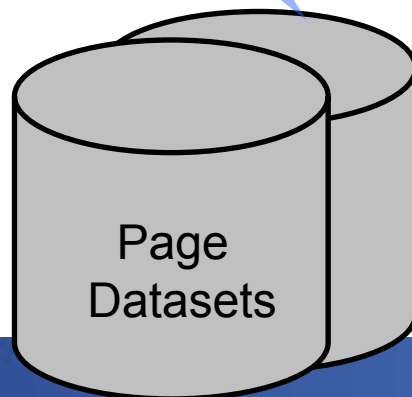
Physical Swap – z/OS 1.8 ...

- z/OS 1.8 will no longer physically swap address spaces to AUX
  - As the amount of fixed storage an address space owns increases, swapping an address space out to AUX is no longer viable.
- Analyze your pre-z/OS 1.8 system and pay special attention to the amount of physical swaps that occur on your system.
- Use the RMF CPU Activity Report to find out how many physical swaps occur in your particular system configuration (OUT Ready and OUT Wait).
  - Next for each address space on the OUT Ready or OUT Wait queue you need to find out the number of fixed frames each address space owns.

```
                        C P U   A C T I V I T Y   R E P O R T

SYSTEM ADDRESS SPACE ANALYSIS              SAMPLES =  3,600

NUMBER OF ADDRESS SPACES

        ------------------- QUEUE  TYPES -------------------     --------- ADDRESS SPACE TYPES ---------
        IN                  OUT       OUT     LOGICAL   LOGICAL   BATCH     STC     TSO     ASCH     OMVS
        READY       IN      READY     WAIT    OUT RDY   OUT WAIT

MIN        1        54        0         2        0         83        5     119      10       0        3
MAX       10        85        0         0        0        118       19     154      12       0       23
AVG      2.4      56.4      0.0       1.2      0.0      105.4      6.1    141.7    10.5     0.0      3.5
```

Agenda

- New function overview

- Enhancements for Blocked Workloads
- Promotion of Canceled Jobs
- Starting Minimum Number of Servers Without Delay

- zAAP/zIIP Enhancements
- RSM/SRM Support for Large Real storage
- Group Capacity Limit

- Support for DB2 WLM Assisted Buffer Pool Management
- JES2 Enhancements for WLM-managed batch initiators
- New Resource Group Types
- Routing Enhancements

- WLM Tools
  - User Friendly Interface for WLM Administrative Application
  - Monitoring Tool for Application Environments
- Summary

Group Capacity Limit Definitions

- **LPAR Definitions**
  - Defines a capacity group by entering the group name and group limit value for the partitions which should belong to the same group
    - Group Limit and Group Name new input values on the HMC/SE
    - Group Limit definition is independent from defined capacity definition
  - Exclusive to z9 EC/BC

- **Capacity Group**
  - Consists of multiple LPARs on the same CPC
  - LPARs must run z/OS 1.8
    - If they don't, the group limit may not be enforced correctly
  - It is possible to define multiple groups on a CPC
  - A partition can only belong to one group
  - A capacity group is independent of a sysplex and an LPAR cluster
  - WLM does only manage partitions with shared CPs
    - Dedicated partitions and partitions with wait completion equal to YES are ignored
      - > if they have been defined to a group they will be excluded from the group and only the partitions with wait completion equals NO and shared processors are managed towards the capacity limit

WLM Management of Group Capacity Limit

- **Each partition (z/OS system) manages itself independently from all other partitions**

- **Group capacity is based on defined capacity**
  - Therefore a 4 hour rolling average of the group MSU consumption is used as base for managing the partitions of the group
  - As usual, only general purpose processor service is considered

- **Each partition sees the consumption of all other partitions on the CPC**
  - If the partition belongs to a group it identifies the other partitions of the same group
  - Calculates its defined share of the capacity group
    - Based on the partition weight
      - This share is the target for the partition if all partitions of the group want to use as much CPU resources as possible
  - If one or more LPARs do not use their share, this donated capacity will be distributed over the LPARs which need additional capacity

- **Even when a partition receives capacity from another partition, it should not violates its defined capacity limit (if one exists)**

Defining LPAR Groups (1)
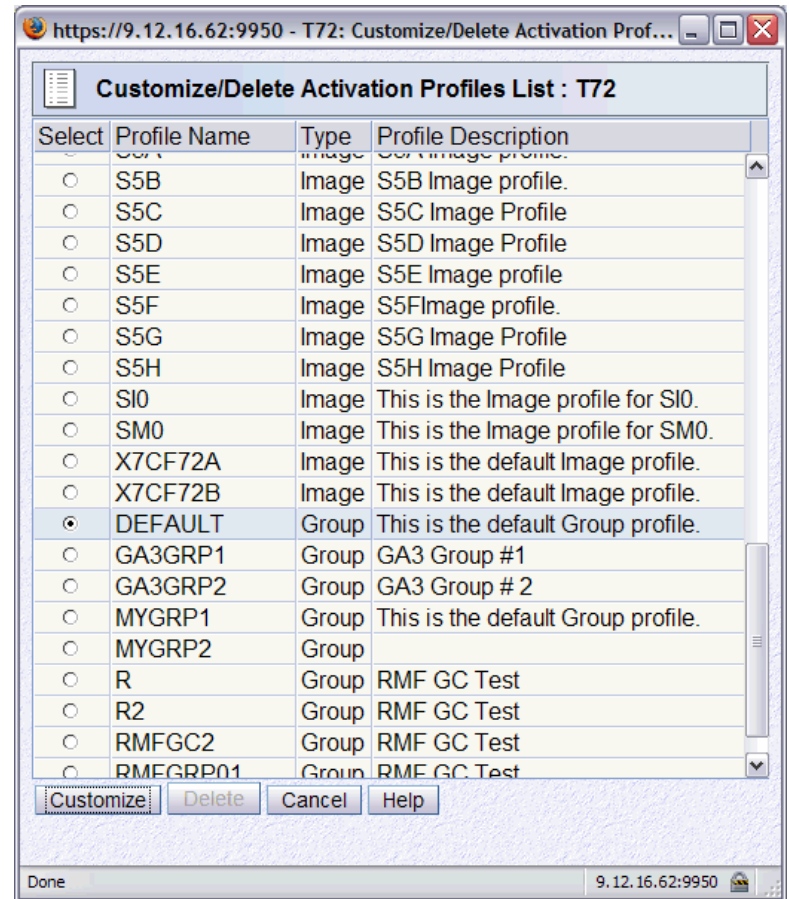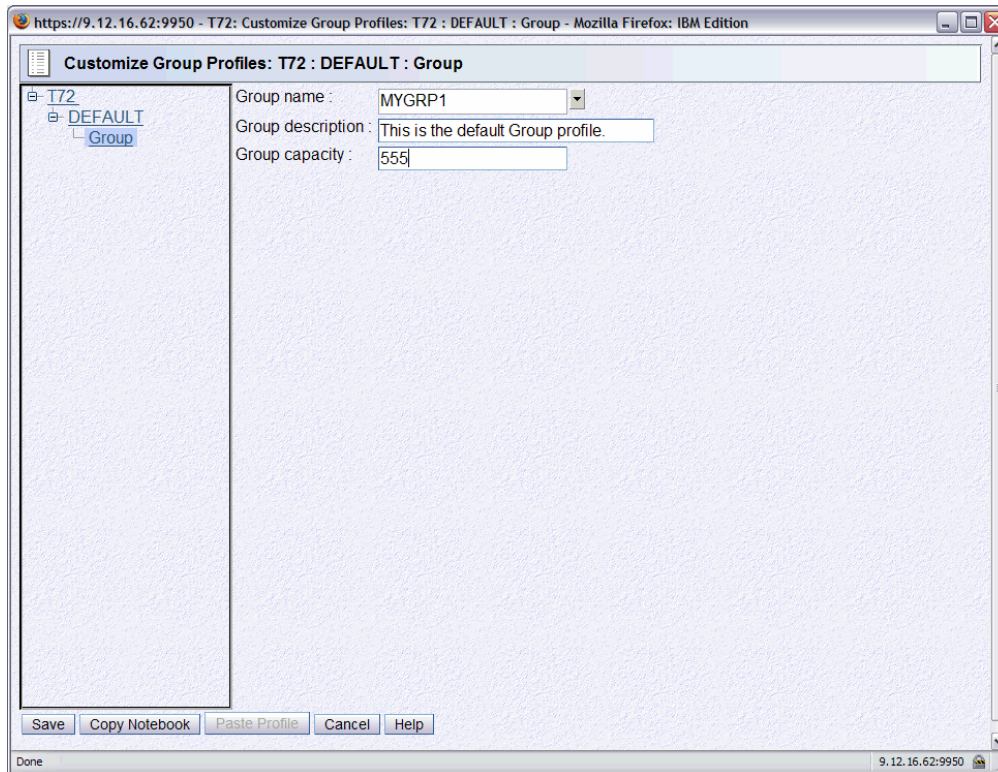
Defining LPAR Groups (2)

https://9.12.16.62:9950 - T72: Primary Support Element Workplace (Version 2.9.2) - Mozilla Firefox: IBM Edition

Views

CPC Operational Customization

https://9.12.16.62:9950 - T72: Change LPAR Group Controls - Mozilla Firefox: IBM Edit...

Change LPAR Group Controls

Edit ▾

Input/Output configuration data set (IOCDS): a3

Group Name	Member Partitions	Group Capacity Value
DEFAULT		0
MYGRP1		555
MYGRP2		777
RMFGRP01		122
RMFGC2		255
RMFTST7		649
R		199
R2		200
GA3GRP1		1200
GA3GRP2		1400

Save to Profiles | Change Running System | Save and Change | Reset | Cancel | Help

Done   9.12.16.62:9950

Change LPAR Security

Enable I/O Priority Queuing

Storage Information

Change LPAR I/O Priority Queuing

System Activity Profiles

Change CP/SAP Allocation

Enable/Disable Dynamic Channel Subsystem

Change LPAR Group Controls

View LPAR Cryptographic Controls

Export/Import Profile Data

9.12.16.62:9950

Defining LPAR Groups (3)

Group Capacity Limit: Example

- **Partitions A, B, and C are part of GROUP1**
  - Partitions D and E do not belong to any group
- **The limit for the group is 200 MSUs**
  - Assuming the CPC capacity is >200

- **The guaranteed minimum for each partition is based on the weights.**

LP	Group Name	Group Limit [MSU]	Weight	Target consumption based on weight [MSU]	LP Defined Capacity [MSU]
A	GROUP1		70	93	none
B	GROUP1	200	50	67	80
C	GROUP1		30	40	30
D	none	n/a	100	n/a	120
E	none	n/a	50	n/a	n/a

The target consumptions for each partition within the group are:

- Partition A: 93
  - Can use up to 200 MSU based on 4 hour rolling average. If all three partitions want to use as much as possible, partition A will get 200 * 70/(70+50+30) ≈93 MSU (if possible at all)
- Partition B: 67
  - Can use up to 80 MSU because an individual softcap is defined. If all three partitions want to use as much as possible, partition B will get 67 MSU (if possible at all)
- Partition C: 40
  - Can use up to 30 MSU because the defined capacity is smaller than the target based on partition weight

Group Capacity Limit Considerations

- **A 4 hour bonus exists after IPL for the group**

  – Meaning the group wide capping will start when the 4 hour rolling group average will reach the group capacity limit - similar to standard defined capacity.

- **When new members join a group…**

  – If a new system is IPLed and joins a group it does not have the history of the MSU consumption of the complete group

  – Therefore it can take up to 4 hours until all systems in the group have the same view

    • During that time period the group limit cannot be guaranteed
    • If that happens the other partitions will be reduced to their capacity based on their weight in the group

- **A partition can be dynamically removed from a group and/or added to another group**

  – The changed partition has no knowledge about unused capacity of the new group and it does not keep a history from previous activity

  – All systems must learn again about the new situation - therefore the group limit cannot be guaranteed

- **Works together with IRD Weight Management and Vary CPU Management**

  – IRD Weight Management may change the weight of partitions in a capacity group and thus changes the target share of the partition in the capacity group

Example 1: Three Partitions with Different Load Pattern –
Test Setup

- **Scenario setup**

    – 3 partitions with different weights and **no individual defined capacity limits**

    – Group Limit: 50

- **Test scenario**

Partition	Limit	Weight	Target MSU
IRD3	n/a	52	~8.5
IRD4	n/a	102	~16.7
IRD5	n/a	152	~24.8
Group	50	306	50

1. Initially only IRD3 and IRD4 demand CPU

    - IRD5 just IPLs w/o any load

2. Then all partitions have high CPU demands

3. IRD5 demand drops to minimum (just system)

- **System load is not constant - the test tries to include load changes on the three systems**

Example 1: Three Partitions with Different Load Pattern Test Execution
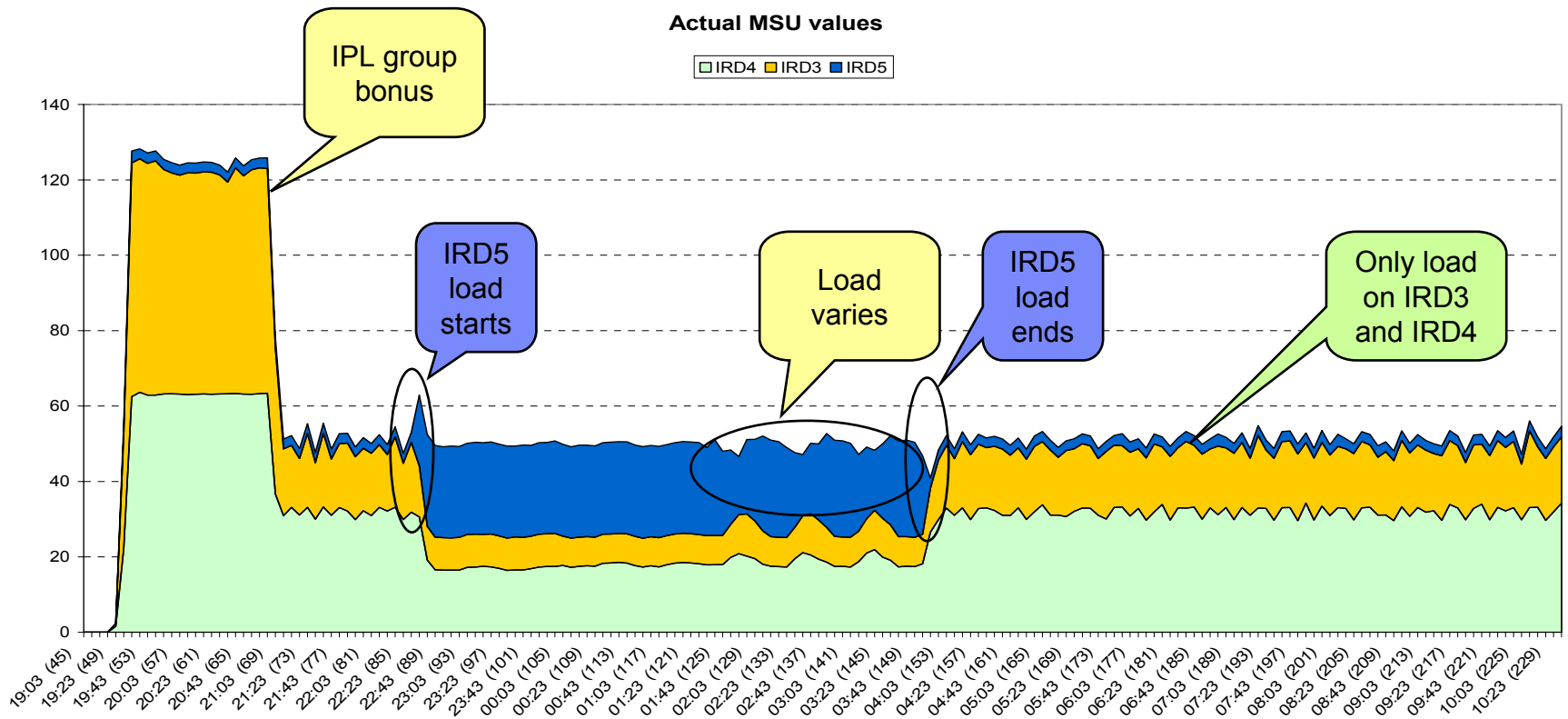
Example 3: Partitions <u>with</u> Defined Capacity Limits

Partition		Limit	Weight	Target MSU	Based on Limits
IRD3	🟩	n/a	51	~8.5	~40
IRD4	🟧	5	101	~16.7	5
IRD5	🟧	5	151	~24.8	5
Group		50	303	50	

IPL bonus for IRD4 and IRD5 based on their defined capacities is expired. The partitions IRD4 and IRD5 are capped now to 5 MSU

IPL bonus for the group is expired. Partition IRD3 is now capped, too, but can use about 40 MSU because IRD4 and IRD5 use only 5 MSU each
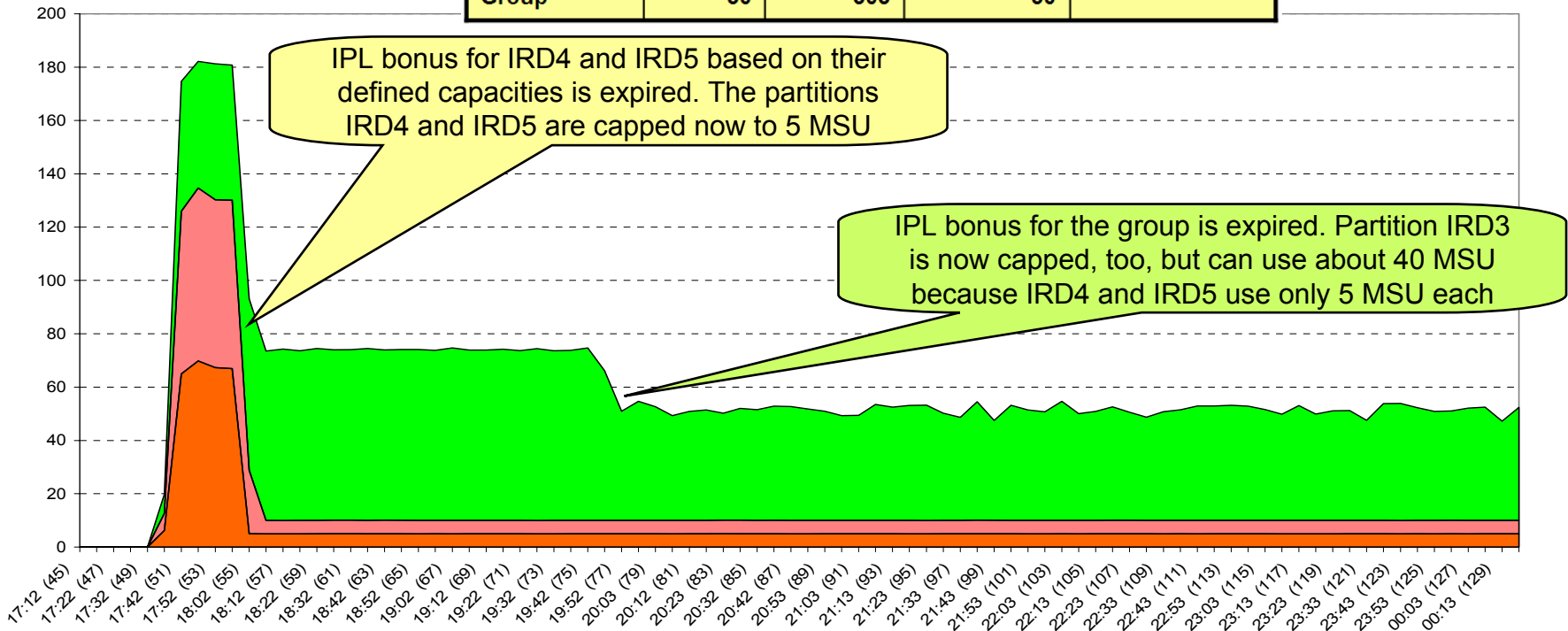
RMF Support of Group Capacity Limits:

- **RMF extends its existing CPU Activity reporting with information about capacity groups**
  - The Monitor I / Postprocessor CPU activity report provides a new report section for capacity groups with information about
    - Defined MSU limits
    - The share each partition in the group can take
    - The guaranteed share of the partitions
    - The overall MSU consumption within the group
  - SMF record type 70 subtype 1 is extended
  - Monitor III CPC report extended
  - An asterisk in the reports indicates that the partition is a member of the capacity group for less than four hours with regard to the RMF interval start.

```
                              P A R T I T I O N   D A T A   R E P O R T


       z/OS V1R8              SYSTEM ID SYS1           DATE 30/04/2006           INTERVAL 15.00.999
                              RPT VERSION V1R8 RMF     TIME 13.30.00             CYCLE 1.000 SECONDS


MVS PARTITION NAME                      A     NUMBER OF PHYSICAL PROCESSORS  16    GROUP NAME      GROUP1
IMAGE CAPACITY                        120                               CP    8       LIMIT        200 *
NUMBER OF CONFIGURED PARTITIONS         6                               IFA   2
WAIT COMPLETION                        NO                               IFL   5
DISPATCH INTERVAL                 DYNAMIC                               ICF   1
```

Agenda

- New function overview

- Enhancements for Blocked Workloads
- Promotion of Canceled Jobs
- Starting Minimum Number of Servers Without Delay

- zAAP/zIIP Enhancements
- RSM/SRM Support for Large Real storage
- Group Capacity Limit

- Support for DB2 WLM Assisted Buffer Pool Management
- JES2 Enhancements for WLM-managed batch initiators
- New Resource Group Types
- Routing Enhancements

- WLM Tools
  - User Friendly Interface for WLM Administrative Application
  - Monitoring Tool for Application Environments
- Summary

Support for DB2 WLM Assisted Buffer Pool Management

- **Allows Data Base Managers to benefit from WLM/SRM goal-driven resource management capabilities**

- **Allows for dynamic adjustment of Buffer Pool sizes considering the requirements of data requesters and system / sysplex workload.**

- **Implementation:**
  - Introduce a new delay-type being reported to WLM
  - Report these new delays via dedicated performance blocks (PBs)
  - Optimum Buffer Pool sizes are determined by means of the Data requesters' requirements.

- **Exploitation**
  - DB2 Version 9.1 for z/OS
    - ALTER BUFFERPOOL… AUTOSIZE(YES|NO)

DB2 WLM Assisted Buffer Pool Management Overall Flow
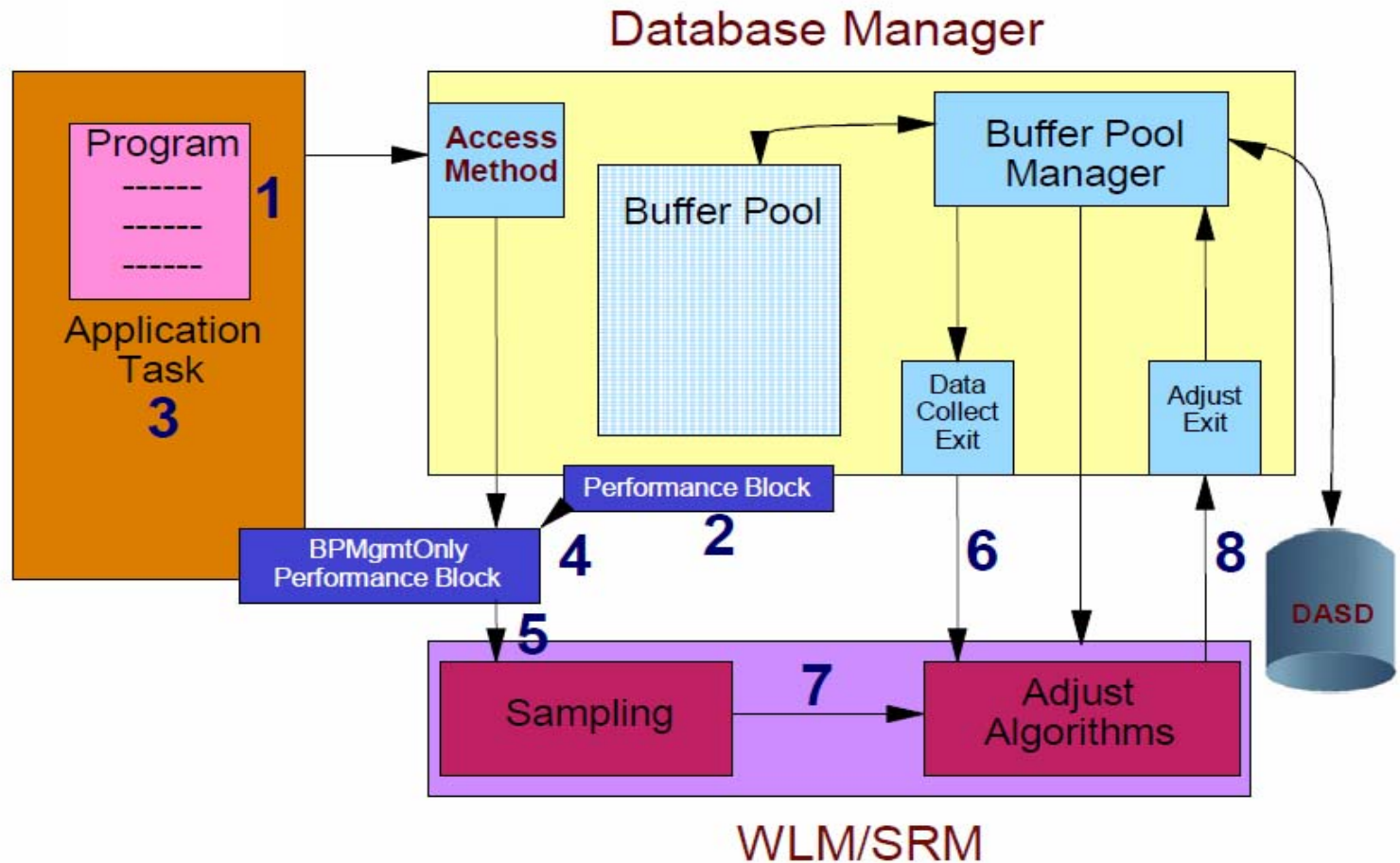
Buffer Pool Size Adjustment - Example
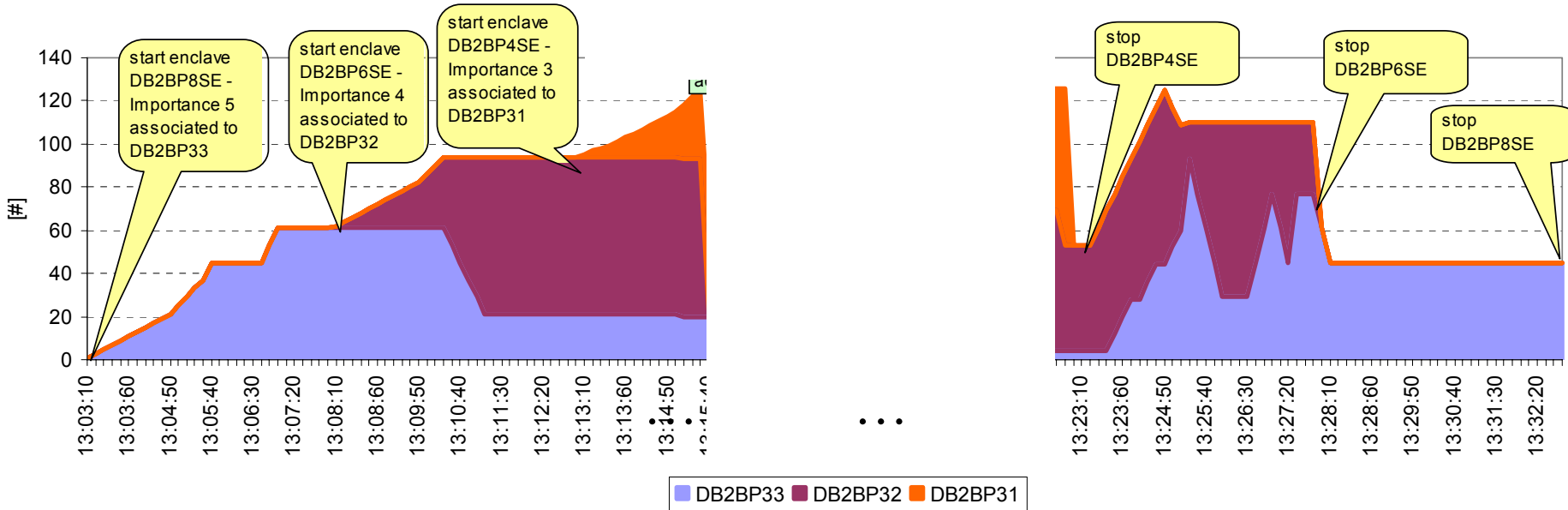
- **Remarks:**
  - 3 buffer pools compete for limited amount of storage
  - Each buffer pool serves different service classes which are assigned different importance levels
  - Buffer pools serving service classes with higher importance may steal storage from buffer pools serving enclaves of lower importance
  - Classifications of the buffer pool managers themselves do not influence buffer pool adjustment

Support for DB2 WLM Assisted Buffer Pool Management Remarks

- **For more information refer to DB2 V9.1 documentation**

  – E.g. **DB2 V 9 for z/OS Technical Overview**
    http://www.redbooks.ibm.com/abstracts/sg247330.html

  – Need to consider impact on real storage requirements with PGFIX(YES)

JES2 Improved Batch Initiator Balancing

- **Problem prior to z/OS R8:**

  - In a WLM managed initiator environment JES2 will preferably start jobs on the *submitting* system

- **Enhancement in R8 attempts to use *approximately* the same percentage of active WLM-managed initiators in each service class on each system.**

  - May help balance batch workloads across a MAS configuration within a sysplex

  - In the scenario on the next chart we submit one job per minute (for a total of 50 min) on SYS1 (top row)

    - With R7, the blue lines (Active servers = initiators) differ on the two systems. E.g. SYS1 peaks at 52, SYS2 at 45.

    - With R8, the lines are very similar and peak at 50.

JES2 Improved Batch Initiator Balancing - Comparison

z/OS R7

z/OS R8
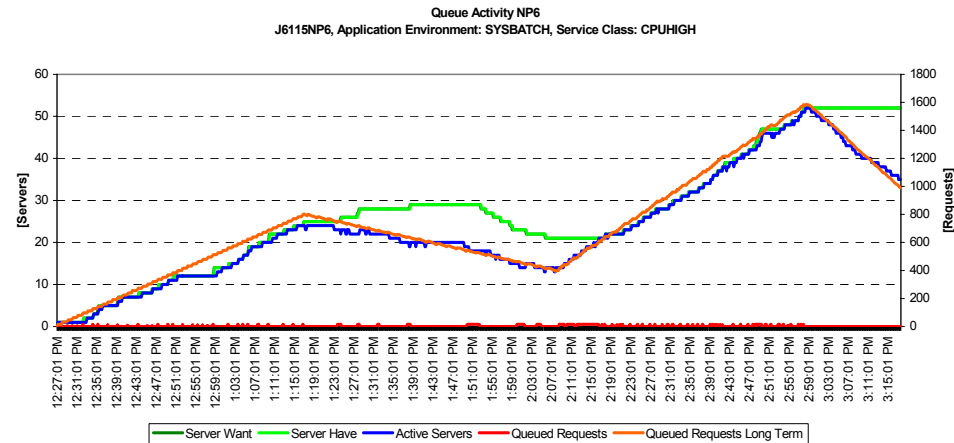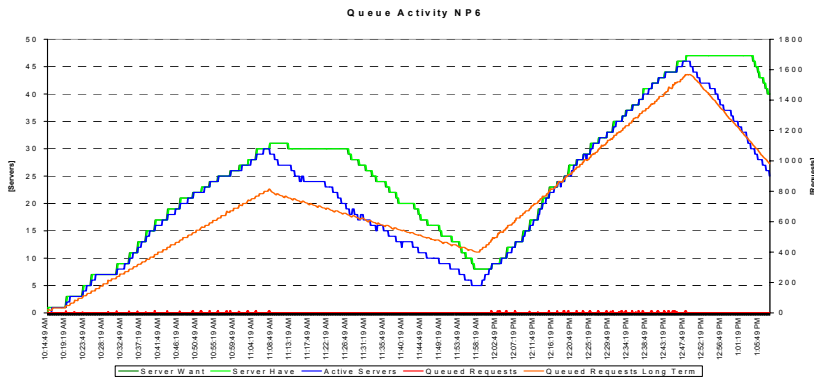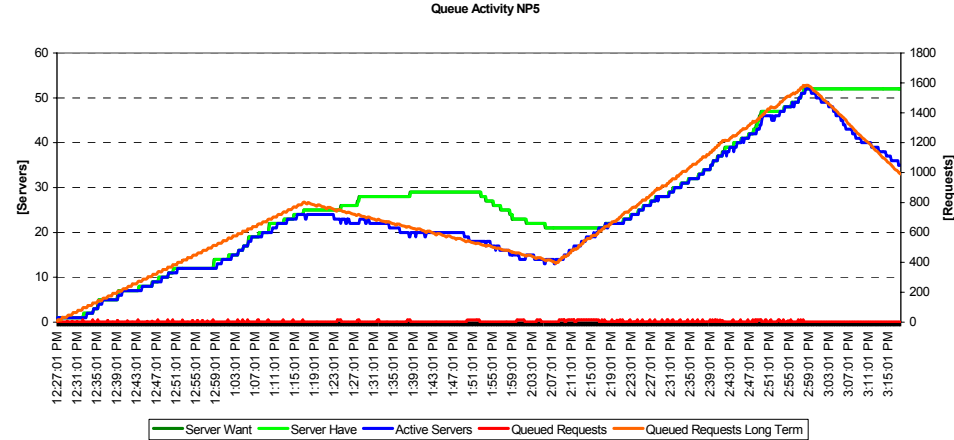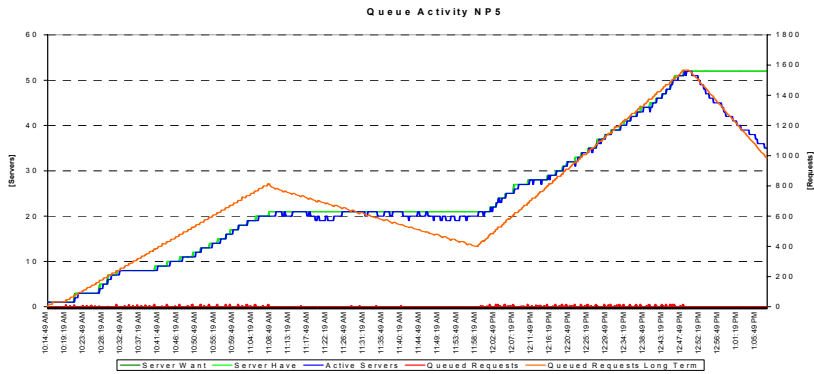
Agenda

- New function overview

- Enhancements for Blocked Workloads
- Promotion of Canceled Jobs
- Starting Minimum Number of Servers Without Delay

- zAAP/zIIP Enhancements
- RSM/SRM Support for Large Real storage
- Group Capacity Limit

- Support for DB2 WLM Assisted Buffer Pool Management
- JES2 Enhancements for WLM-managed batch initiators
- New Resource Group Types
- Routing Enhancements

- WLM Tools
  – User Friendly Interface for WLM Administrative Application
  – Monitoring Tool for Application Environments
- Summary

Existing Resource Group Concept

⇒Type 1 Resource Group

- **Sysplex-wide defined in unweighted service units per second**
- **Sysplex-wide managed**
- **General Considerations**
  - Multiple service classes may be assigned to a resource group
    - With different utilizations on different systems
  - Systems may have different capacities
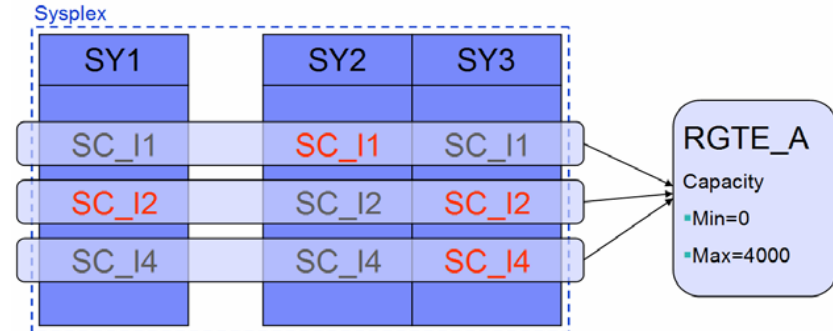
- **Results:**
  - Not easy to understand how much is consumed on which system
  - Consumption depends highly on the capacity of the systems!
  - Resource Group definitions need to be revisited every time
    - Systems are upgraded
    - Workload utilization changes
- **Resource Groups are sometimes hard to understand**

Sysplex

SY1	SY2	SY3
SC_I1	SC_I1	SC_I1
SC_I2	SC_I2	SC_I2
SC_I4	SC_I4	SC_I4

RGTE_A
Capacity
- Min=0
- Max=4000

New Types of Resource Groups (Type 2 and 3)

- **Sysplex-wide defined,
  but definition applies to each system**

- **Managed by each system**

- **General Considerations**
  - Multiple service classes can be assigned to a resource group but this has no sysplex-wide effect anymore
  - Definition is based on one of two possible units:
    - **LPAR capacity: based on system weight (+…)** ⇒Type 2 Resource Group
    - **LCP capacity** ⇒Type 3 Resource Group
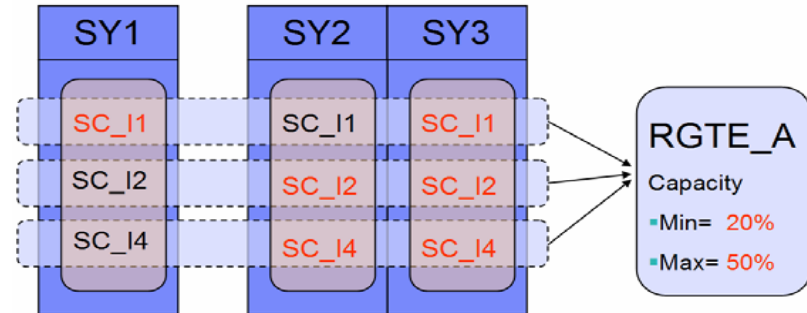
- **Results**
  - New resource groups are managed by system thus they must be evaluated on a per system base
  - Resource groups "grow" automatically when systems are upgraded

- **Easier to understand**

SY1 SY2 SY3

SC_I1 SC_I1 SC_I1
SC_I2 SC_I2 SC_I2
SC_I4 SC_I4 SC_I4

RGTE_A

Capacity
▪Min= 20%
▪Max= 50%

Type 2 Resource Group Specification

- **Capacity is specified as a percentage of the effective LPAR share**
  - Range 0..99
  - Effective LPAR share is the minimum of
    - LPAR share:
      CPC shared pool capacity · (LPAR weight / sum of weights)
    - Logical processor capacity
    - Defined capacity limit (only when soft capping is in effect)
  - Use WLM Planning: Appendix B - CPU Capacity Table
    to locate the CPC capacity that would be equivalent to your shared pool configuration

Numbers should be meaningful for configuration, e.g. sum of all minimum LPAR shares $\leq$ 99

$$RG_{[SUs]} = \begin{cases} Softcap_{[SUs]} \bullet RG_{Limit[\%]} & \text{if } softcap < ShareCapacity \text{ or } LCPCapacity \\ LCPCapacity_{[SUs]} \bullet RG_{Limit[\%]} & \text{if } LCPCapacity < ShareCapacity \text{ or } softcap \\ ShareCapacity_{[SUs]} \bullet RG_{Limit[\%]} & \text{if } ShareCapacity < LCPCapacity \text{ or } softca \end{cases}$$

$$ShareCapacity_{[SUs]} = CECCapacity_{[SUs]} \bullet LPARShare$$

$$CECCapacity_{[SUs]} = \text{Capacity based on shared physical processors for the CEC}$$

$$LPARShare = \frac{Weight(Current\ Partition)}{\sum_i^{all\ active\ partitions} Weight(i)}$$

$$LCPCapacity_{[SUs]} = \text{Capacity based on shared processors available to the LPAR}$$

Type 3 Resource Group Specification

- **Capacity is expressed as an equivalent number of general purpose processors**

  - Scaled by 100, i.e. 100 equivalent to 1 CP
    Range: 0…999999

  - Use *WLM Planning: Appendix B - CPU Capacity Table*
    to locate the model that would be equivalent to your virtual processor configuration.

Numbers should be "meaningful" for configuration.

```
 Resource-Group  Xref  Notes  Options  Help
 ------------------------------------------------------------
                           Modify a Resource Group
 Command ===> _____

 Enter or change the following information:

 Resource Group Name  . . . . : DIVB
 Description  . . . . . . . . . Division B service

 Define Capacity:
  3  1.  In Service Units (Sysplex Scope)
     2.  As Percentage of the LPAR capacity (System Scope)
     3.  As a Number of CPs times 100 (System Scope)
 Minimum Capacity . . . . . . . 10
 Maximum Capacity . . . . . . . 20
```

Comparison of Resource Group Types

Definition	Scope	Advantages	Possible Disadvantages
Type 1 RG: Service Units	Sysplex	▪ Allows balancing of resources across members in a sysplex	▪ Requires adjustments for Migration ▪ Difficult to monitor ▪ Not applicable to constrained work on a single system in a sysplex environment
Type 2 RG: % of LPAR	System	▪ Allows to control work on single members in a sysplex ☑ <u>Stable for migrations</u>	▪ Need to know the LPAR capacity
Type 3 RG: % of LCPs	System	▪ Allows to control work on single members in a sysplex ☑ <u>Easy and straight forward definition</u>	▪ May requires adjustments for migration

Resource Groups: Administrative Application

```
Resource-Group  Xref  Notes  Options  Help
------------------------------------------------------------
                     Modify a Resource Group
Command ===> _____

Enter or change the following information:

Resource Group Name  . . . . : DIVB
Description  . . . . . . . . . Division B service

Define Capacity:
2   1.  In Service Units (Sysplex Scope)
    2.  As Percentage of the LPAR capacity (System Scope)
    3.  As a Number of CPs times 100 (System Scope)
Minimum Capacity . . . . . . . 10
Maximum Capacity . . . . . . . 20
```

- **Resource group definition panel has been extended and does now allow for definitions of the three variations for resource groups**

  1. In service units with a sysplex-wide scope (existing)

  2. As a percentage of the LPAR capacity (new concept)
     - Based on LPAR weight
     - System-wide managed

  3. As a percentage of the LCP capacity
     - System-wide managed

Resource Groups: Example

**Resource Group Overview
System: IRD1, Resource Group: ELPMAX , Type 2: 60% of LPAR**

CEC Capacity	~	111758 SU/s	(2084-306)
LPAR Capacity	~	8697 SU/s	Based on Weight (7.7%)
ELPMAX Capacity	~	5218 SU/s	60% of the LPAR capacity

Legend: —— Minimum Service Rate —— Maximum Service Rate —— Actual Service Rate —— Capping Slices
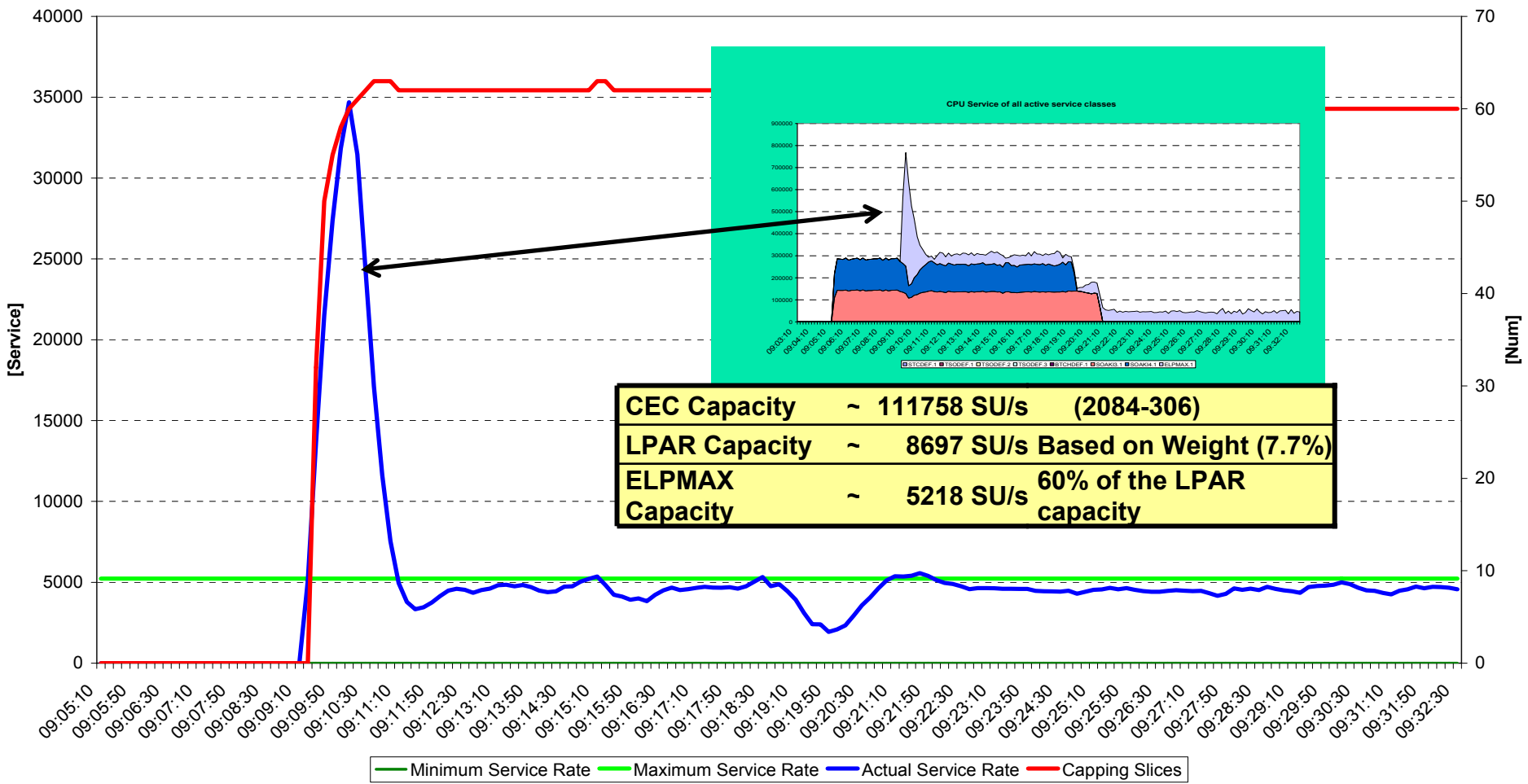
Resource Groups: Migration Considerations

- **New system service classes and new resource group concept introduce incompatible changes to the WLM service definition**

  – Introduce new WLM policy level to LEVEL019

  – APAR OA13837 is required on pre-R8 systems

    • listed in GA22-7499 z/OS V1R8.0 Migration: z/OS V1Rx coexistence and fallback PTFs

- **The compatibility APAR OA13837**

  – Ensures that type 2 and 3 resource groups are not managed on pre-1.8 systems

  – As long as policies are only installed / activated from back level systems the compatibility code is not required

    • Compatibility PTFs can be installed via rolling IPL
      – no Sysplex-wide IPL necessary

Resource Groups: WLM API Changes

- **Resource Groups**

  - IWMSVPOL Service

    - The values for SVPOLGMN and SVPOLGMX (IWMSVPOL) mapping can now be in percentages
      - of LPAR capacity (SVPOLGPV bit turned on)
      - of single processor capacity (SVPOLGPC bit turned on)
    - If both bits are off the values are in unweighted service units

  - For mixed-release sysplexes (z/OS 1.8 and older versions)

    - On pre-z/OS 1.8 releases the SVPOLMXS and SVPOLMNS bits indicate that NO limits have been specified.

Agenda

- New function overview

- Enhancements for Blocked Workloads
- Promotion of Canceled Jobs
- Starting Minimum Number of Servers Without Delay

- zAAP/zIIP Enhancements
- RSM/SRM Support for Large Real storage
- Group Capacity Limit

- Support for DB2 WLM Assisted Buffer Pool Management
- JES2 Enhancements for WLM-managed batch initiators
- New Resource Group Types
- Routing Enhancements

- WLM Tools
  - User Friendly Interface for WLM Administrative Application
  - Monitoring Tool for Application Environments
- Summary

Sysplex Routing: Shortcomings before z/OS 1.7

- **Recommendation purely capacity based**
  - New option on routing services will allow a router to receive weights which consider the goal achievement of the work
  - **z/OS 1.7   (IWMSRSRS FUNCTION=SPECIFIC)**
- **TCP/IP only registers the stack and doesn't know how the work is being processed**
  - Introduce new routing service e.g. for TCP/IP which allows to provide routing recommendations for TCP/IP backend applications
  - **z/OS 1.7 (IWM4SRSC)**

- **Abnormal conditions of the work receiver or consumer not factored into the routing recommendations**
  - In a first step the queue time of the work is accounted for DDF regions and its proportion to the execution time is factored into the routing recommendations
  - **z/OS 1.7 (incorporated in IWMSRSRS FUNCTION=SPECIFIC)**

  – Current activities to provide broader solutions
  - Include abnormal terminations which are recorded by CICS
  - New interface/option which allows a server to provide its state to WLM

- **Recommendations (weights) returned by routing services include not only general purpose capacity but also zAAP and zIIP capacity**

Routing Services: Summary

Service	Function	Remark	Description
IWMSRSRS	**Existing function SELECT**		**Returns capacity for the system relative to other systems in the sysplex (accounts for the number of registered server instances)**
IWMSRSRS	**New parameter SPECIFIC**	**(1)**	**Returns capacity for the registered server relative to all other registered servers of the same type in the sysplex.** **With R9, contains information on general purpose, zAAP and zIIP capacities. Considers:** ▪ **Goal Achievement (PI)** ▪ **Queue Time for Enclaves** ▪ Health Indicator
IWMSRSRG	**New Parameter HEALTH**		**Allows the server which registers to provide a health indicator from ok=100 to not ok=0. The factor is considered as part of the weight.** **IWMSRSRG HEALTH can be updated by the server at any time**
IWM4HLTH	**New Service**		**For address spaces which are not registered and which want to set a health status. This status is factored into IWM4SRSC return data**
IWM4SRSC	**New service**	**(1)**	**Returns capacity for another address space to which the request is provided by the registered server. Considers** ▪ **Goal Achievement (PI)** ▪ abnormal termination rate expressed as the number of abnormal terminations (as passed to WLM by the IWMRPT interface) per 1000 total terminations.
	New parameter ABNORM_COUNT		

(1)   Goal achievement is derived from the service class the working is running in. Can be an enclave service class, a transaction manager service class or the service class to which the address space is classified too

Sysplex Routing: Enhancements for IWMSRSRS Service

- **New Function code „SPECIFIC"**

- **Weight calculation: Product of four factors**
  - System Utilization Factor:
    - Same as the resulting system weight for old „SELECT" function (as described in the earlier foils)
  - PI Factor
    - This gives an indication of how good this server, respective the work that is related to this server, is achieving its goals as defined in the active WLM policy.
  - Queue Time Ratio:
    - If the server owns independent enclaves, the ratio of queue time to elapsed time of those enclaves
  - Health Indicator
    - A registered server can use the IWMSRSRG HEALTH= parameter to inform WLM about additional states which are unknown to WLM but should be factored into the returned weight
  - Remark
    - If multiple servers are registered on the same system, the weight is divided by the number of the servers.

Sysplex Routing: New IWM4SRSC Service

- **Characteristics**
  - Provides routing recommendation for a given server
  - System scope
  - No registration of servers necessary
  - The intention of this service is to obtain routing recommendations for work which is passed through the registered server (e.g. TCP/IP) to a server which consumes/processes the work
  - Input: STOKEN that identifies the address space of a server
  - Output: Recommendation of how good this server is suitable to receive work → the weight
    - A value between 1 and 64

- **Weight calculation: Product of four factors**
  - The PI Factor (Performance Indicator Factor)
    - This gives an indication of how good this server, respective the work that is related to this server, is achieving its goals as defined in the active WLM policy.
  - The Importance factor
    - This is a measurement of how much CPU Capacity is displaceable by work of the server's importance, respective the work that is related to this server
  - The health factor
    - This is an indicator an address space can set via the new IWM4HLTH service
  - Abnormal termination rate

Effect of Performance Factor on WLM Routing Recommendations

SYS	Avail Cap	Original Server weight	PI	WLM weight
SYS1	110	18	1.3	14
SYS2	100	16	0.8	16
SYS3	95	15	1.0	15
SYS4	95	15	2.0	8
SUM		64		53

- **Example assuming a 4-way Sysplex**

- **If the server specific PI is >1 the weight is divided by the PI.**

- **Consequently the sum of all weight is no longer normalized to 64.**

- **WLM server weight influences the distribution of work across the sysplex by the subsystem.**

z/OS R9 IWMSRSRS: Enhancements for zAAPs and zIIPs

- **Base calculation for FUNCTION=SPECIFIC and FUNCTION=SELECT is the same**

- **In previous releases**
    - Only one weight (SYSR_WEIGHT) is returned
    - It is based only on regular CP capacity
    - A system was NOT returned if it had less than 5% of displaceable capacity at the selected importance level

- **With z/OS 1.9**
    - SYSR_WEIGHT now is the combined weight of all available processor resources
    - The individual weights have been added:
        - SYSR_CPU_WEIGHT (corresponds to SYSR_WEIGHT of previous releases)
        - SYSR_ZAAP_WEIGHT
        - SYSR_ZIIP_WEIGHT
    - A system is NOT returned if it has NO displaceable capacity for regular CPs at the selected importance level
        - As a result potentially more systems may be returned
        - If a system has no displaceable for an assist processor it is returned and missing capacity is reflected in the weight

z/OS R9 IWMSRSRS…

Additional Enhancements

– Restriction that no more than 96 servers per system could be returned is relieved

- Now up to 300 servers per system will be returned
- Also valid for IWMSRSRS FUNCTION=QUERY
- This support is available for z/OS V1R6 and above via APAR OA18531

– The C interface to IWMSRSRS IWMDNSRV has also been extended to return the new weights

IWMWSYSQ
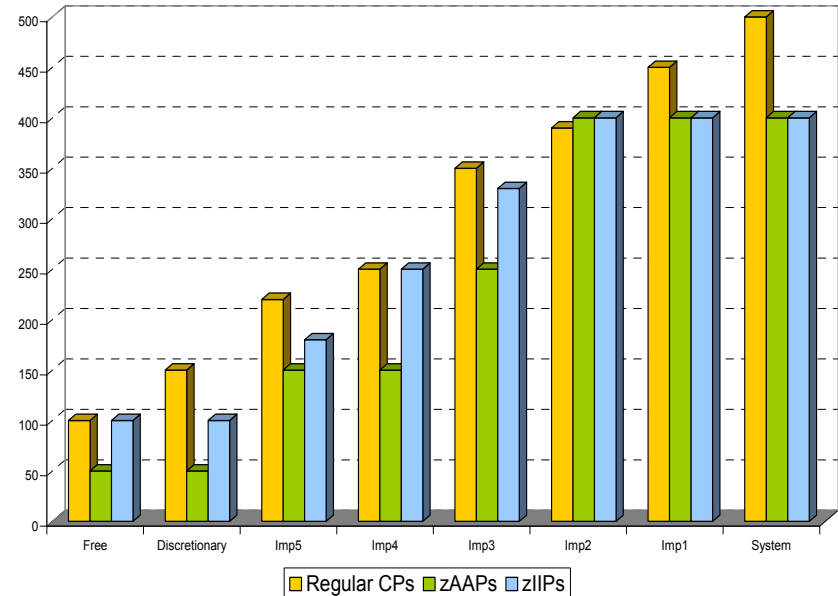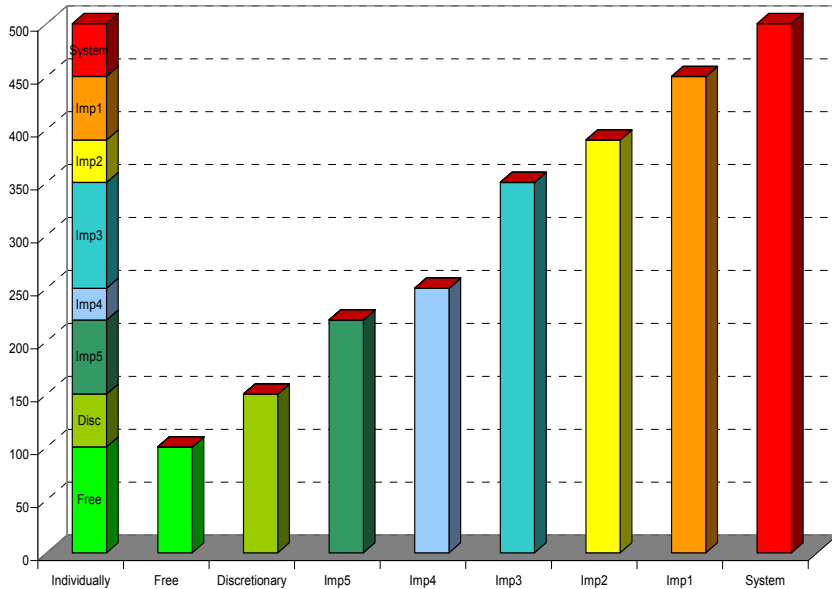
- **Returns capacity information of systems in a sysplex**
  - Each bucket contains the accumulated capacity of all lower importance levels

z/OS R9 IWMWSYSQ Enhancements

- **Enhancement for z/OS V1R9 is available via new parameter EXTENDED_DATA**

  - With EXTENDED_DATA=YES additional information is returned in the output area
    - The system level (0) which contains the total system capacity was added
    - Data now returned for all processor types
    - In addition:
      - Uniprocessor speed of a single processor
      - zAAP and zIIP normalization factors (deviation from regular processor speed if applicable)

  - EXTENDED_DATA=NO returns the output area as before (pre V1R9 layout)

Capacity Based Routing for WebSphere

- In the past the routing algorithm was round-robin.

- With z/OS R9 the default routing is based on available processor capacity of the systems.

  – May result in changed routing recommendations compared to current behavior.

- The new IEAOPT parameter WasRoutingLevel=1 allows going back to the old round-robin routing algorithm RoundRobin.

- Function can be enabled on z/OS Releases ≥6 with **OA16486**

IEAOPT WASROUTINGLEVEL

WASROUTINGLEVEL	
<u>0</u>	**Use the most advanced routing algorithm supported by all systems in the sysplex.** ■ If release level of the lowest system is z/OS R9 (or above) , or APAR OA16486 is installed, routing decisions are based on available /displaceable capacity of standard and assist processors. ▪If the release level of the lowest system is below z/OS R9 and runs without the APAR OA16486 installed, the routing algorithm round robin is used.
1	**WLM uses the routing algorithm round robin.** If this option is used, set it on all systems of the sysplex so that all are using the same algorithm. Otherwise, the WebSphere routing service on each system uses the specific algorithm setting for the system, which can lead to inconsistent results.

Advanced Routing For WebSphere

Find the importance level at which at least one system has 5% capacity, compare the capacities of the systems and calculate the system weights.

SYS 1		
CP	zAAP	zIIP
10	21	15

SYS 2		
CP	zAAP	zIIP
20	21	15

SYS 3		
CP	zAAP	zIIP
34	21	34

Calculate the workload distribution over the processor types.

CP	zAAP	zIIP
30	20	14

Calculate a combined system weight considering all processor types.

S1	S2	S3
15	19	30

Agenda

- New function overview

- Enhancements for Blocked Workloads
- Promotion of Canceled Jobs
- Starting Minimum Number of Servers Without Delay

- zAAP/zIIP Enhancements
- RSM/SRM Support for Large Real storage
- Group Capacity Limit

- Support for DB2 WLM Assisted Buffer Pool Management
- JES2 Enhancements for WLM-managed batch initiators
- New Resource Group Types
- Routing Enhancements

- WLM Tools
  - User Friendly Interface for WLM Administrative Application
  - Monitoring Tool for Application Environments
- Summary

User Friendly Interface for WLM Administrative Application
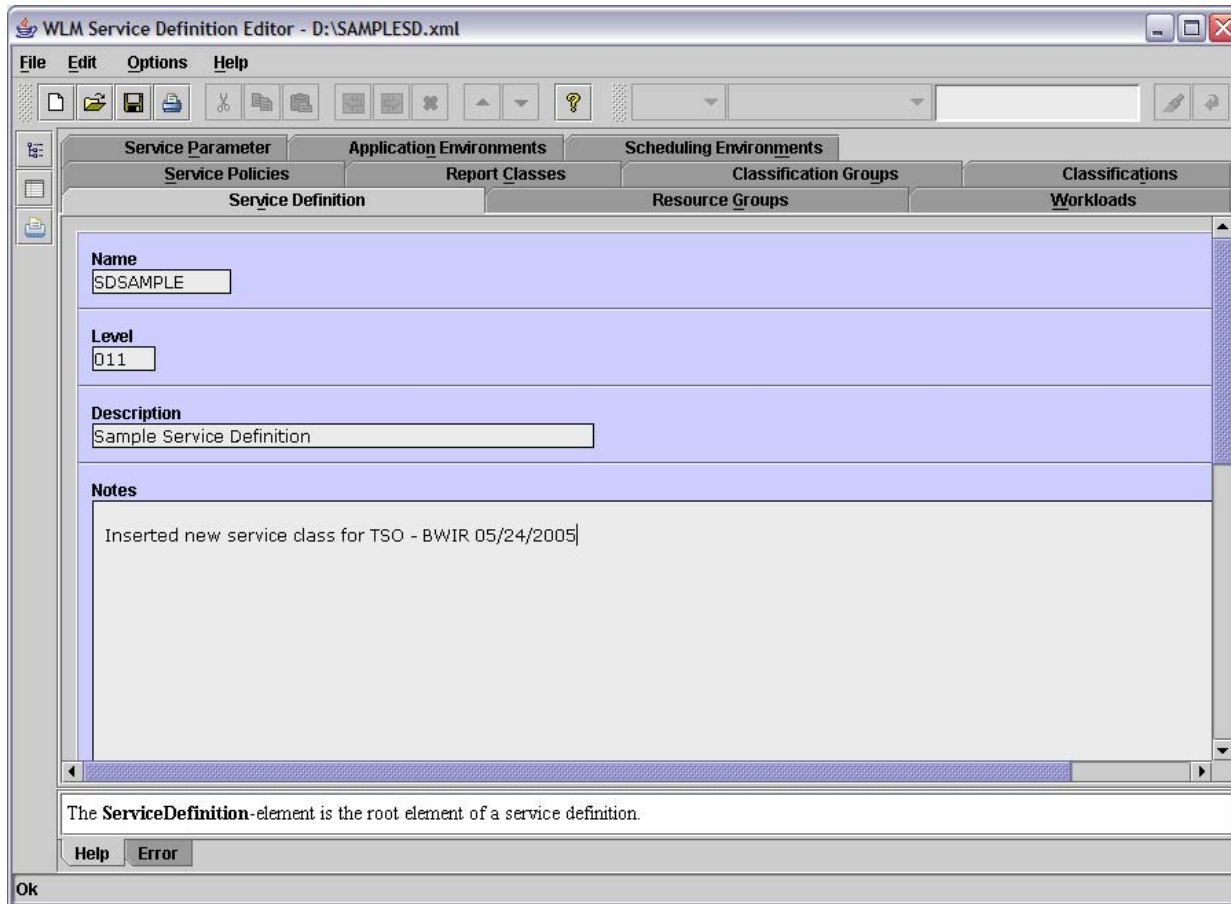
- **Java application**

- **Interface to WLM ISPF datasets**

- **Different views to show relations between policy elements**

- **Real-time error checking**

- **Real-time help**

- **No popup menus**

- **Tabular input processing**

User Friendly Interface for WLM Administrative Application
Example: Workload and Service Class Panel/View

- **Information is represented as tables**

- **Direct manipulation of tables**

- **Display of relations between policy elements**

- **Real-time error checking**

- **Context-sensitive help**

User Friendly Interface for WLM Administrative Application
Example: Easy generation of new views

- **Table columns can be moved as needed**

- **Table rows can be sorted as needed**

- **Rearranged tables can be edited**

- **Search function for arbitrary text within tables**

WLM Administrative Application
Example: Print View

- **Service Definition is converted to HTML-Document**

- **In order to print HTML-Document the configured WWW-Browser of the Operating System is launched**

WLM Administrative Application
Link to existing WLM Administrative Application

Workstation | z/OS

install
activate

read/write

read/write

WLM
ISPF
Tables

WLM
CDS
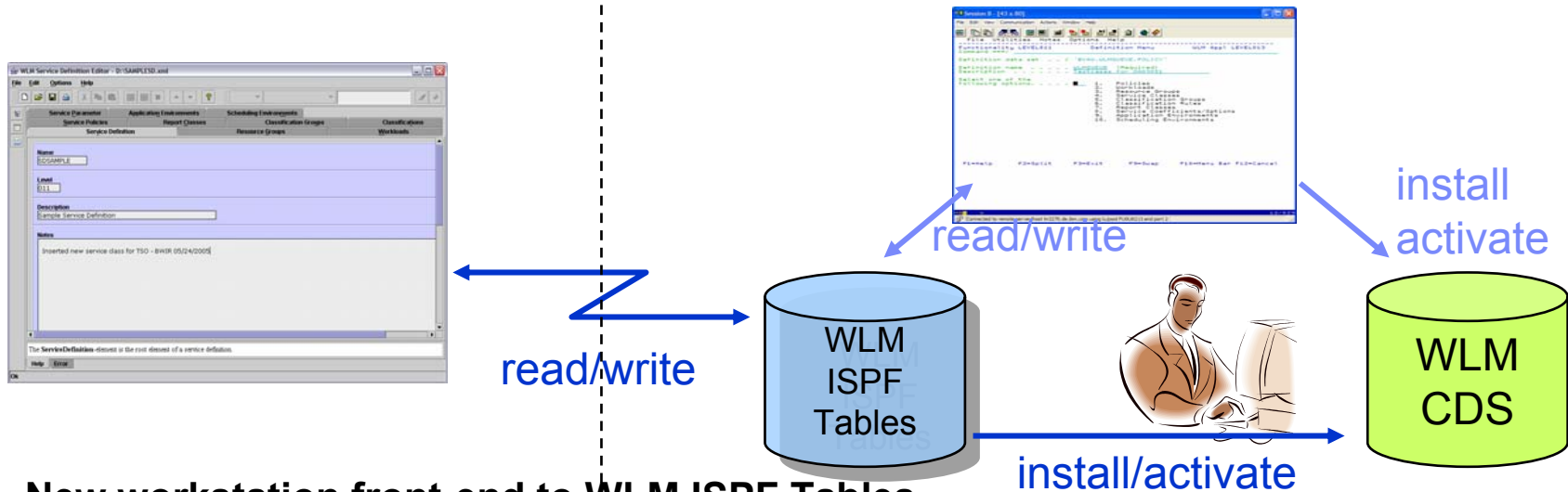
install/activate

- **New workstation front-end to WLM ISPF Tables**
  - FTP connection with automatic upload and download
  - FTP connection profiles
- **Policy install and activate via operator command**

Delivery:
  - As a tool ("as-is") that can be downloaded via the WLM homepage
    http://www.ibm.com/servers/eserver/zseries/zos/wlm/tools/

Tool for Application Environment Monitoring

- **Application Environments are difficult to monitor**

- **This little tool can help …AE knowledge required, though.**

- **REXX/ISPF based**

- **Delivery:**

  – As a tool ("as-is") that can be downloaded from the WLM homepage
  http://www.ibm.com/servers/eserver/zseries/zos/wlm/tools/

New ITSO Redbook

- **System Programmer's Guide to: Workload Manager (SG24-6472)**
  - Content:
    1. Introduction
    2. How WLM works
    3. WLM functions
    4. Implementation and workload classifications
    5. Batch considerations
    6. TSO, STC, and APPC workloads
    7. DB2 workload considerations
    8. WebSphere Application Server workload considerations
    9. UNIX System Services considerations
    10. Transactional workload considerations

  - Download via:
    http://www.redbooks.ibm.com/abstracts/sg246472.html?Open

Documentation

Documentation

z/OS MVS Planning: Workload Management (SA22-7602)

http://publibz.boulder.ibm.com/epubs/pdf/iea2w151.pdf

z/OS MVS Programming: Workload Manager Services (SA22-7619)

http://publibz.boulder.ibm.com/epubs/pdf/iea2w252.pdf

Redbook – System Programmer's Guide to: Workload Manager (SG24-6472)

http://www.redbooks.ibm.com/abstracts/sg246472.html?Open

Redbook – z/OS Intelligent Resource Director (SG24-5952)

Redbook – Effective zSeries Performance Monitoring Using RMF (SG24-6645)

Internet Links

WLM	http://www.ibm.com/s390/wlm
SRM	http://www.ibm.com/s390/srm
RMF	http://www.ibm.com/s390/rmf
WSC	http://www.ibm.com/support/techdocs
Software Pricing	http://www-1.ibm.com/servers/eserver/zseries/swprice/
IRD	http://www-1.ibm.com/servers/eserver/zseries/ird/
Redbooks	http://www.redbooks.ibm.com/