z/OS Workload Management

Spring 2016 Update for IBM z13, z/OS V2.2, and z/OS V2.1

Horst Sinram, STSM, z/OS Workload and Capacity Management, sinram@de.ibm.com
BM Germany Research & Development 28 Apr 2016

Trademarks

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

Not all common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean that the product is not actively marketed or is not significant within its relevant market.

Those trademarks followed by ® are registered trademarks of IBM in the United States; all others are trademarks or common law marks of IBM in the United States.

For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml:

*, AS/400®, e business(logo)®, DBE, ESCO, eServer, FICON, IBM®, IBM (logo)®, iSeries®, MVS, OS/390®, pSeries®, RS/6000®, S/30, VM/ESA®, VSE/ESA, WebSphere®, xSeries®, z/OS®, zSeries®, z/VM®, System i, System i5, System p, System p5, System x, System z, System z9®, BladeCenter®

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
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.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* 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.

Notice Regarding Specialty Engines (e.g., zIIPs, zAAPs and IFLs):

Any information contained in this document regarding Specialty Engines ("SEs") and SE eligible workloads provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIIPs, zAAPs, and IFLs). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the "Authorized Use Table for IBM Machines" provided at
www.ibm.com/systems/support/machine_warranties/machine_code/aut.html ("AUT").

No other workload processing is authorized for execution on an SE.

IBM offers SEs at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

Agenda

New capping options

Reporting enhancements for CICS and IMS, and Mobile Workloads

IBM z13 Support

z/OS V2.2 enhancements

z/OS V2.1 highlights

Other service stream enhancements and recommendations

Statements regarding IBM future direction and intent are subject
to change or withdrawal, and represent goals and objectives only

Two new capping options

z/OS release Function	V2.2	V2.1
z13 GA2 LPAR Absolute group capping	**OA47752**	**OA47752**
Absolute MSU capping	**OA49201**	**OA49201**

- **LPAR absolute group capping**
  - Feature of PR/SM as of z13 GA2, and z13s

  - Like LPAR absolute capping but for a group of LPARs

  - Specified on the HMC as number of processors with 2 decimal places (like 3.75). All processor classes supported.
  - Recognized by WLM as possible limit to the LPAR capacity

- **WLM Absolute MSU capping**
  - Function of WLM provided by APAR OA49201

  - Similar to WLM defined capacity or group capacity but LPAR will always be capped
    - Independent of 4 hour rolling average consumption.
    - General purpose processor
  - Specified in IEAOPTxx. Limit is the LPAR defined capacity or group capacity specified on the HMC **in MSU**.

6

Using absolute MSU capping

IEAOPTxx ABSMSUCAPPING=	
<u>NO</u>	Defined capacity limits and group capacity limits should be enforced only while the long term four hour rolling average consumption exceeds the respective limit (existing and usually desired behavior).
YES	Defined capacity limits and group capacity limit should be enforced **permanently, independently of the long term four hour rolling average consumption**.

- AbsMSUcapping=Yes limits LPAR consumption to a certain MSU number at all times.
  - I.e., the system loses the flexibility of consuming above the defined capacity limit while the four hour rolling average is below the limit.

- Limit remains stable even when CEC configuration changes, e.g. through On/Off CoD or CBU activations or deactivations.
7
  - Absolute MSU capping is an effective means to permanently limit the consumption of an LPAR to a specific MSU figure at all times
    - including times when the *four-hour rolling average* does not exceed the defined limit.

Using absolute MSU capping with group capacity

- When used with an LPAR capacity group:
  - Limit on behalf of the group entitlement will always be enforced
    - Regardless of the *four-hour rolling group average* consumption.

  - As with AbsMSUcapping=NO, an LPAR is allowed to take benefit of the unused group capacity
    - Unless the LPAR is also capped via other LPAR limits.

  - All members of a capacity group that use AbsMSUcapping=YES will permanently enforce the limit on behalf of the capacity group.

  - All members of a capacity group that do *not* use AbsMSUcapping=YES will be capped while the group *four-hour rolling group average* consumption is greater or equal to the group limit

Comparison of capping types

→ New capping types (spring 2016)

Type of capping	Scope	Specification unit	Proc types	Stability of limit under configuration changes	Suitable to isolate LPARs or LPAR groups	Control point
Initial (hard) capping	LPAR	LPAR share of CPC capacity	Any	–	+	SE/HMC
LPAR Absolute capping (zEC12 GA2 and later)	LPAR	Fractional #processors	Any	O	+	SE/HMC
→ LPAR Group Absolute Capping (z13 GA2 and later)	Group of LPARs	Fractional #processors	Any	O	+	SE/HMC
Defined capacity (DC, soft capping)	LPAR	MSU (4HRA)	CP	+	–	SE/HMC
LPAR group capacity (GC, soft capping)	Group of LPARs	MSU (4HRA)	CP	+	–	SE/HMC
→ Absolute MSU Capping	LPAR or Group	MSU	CP	+	+ (CP only)	SE/HMC + IEAOPT
Resource group capping	Groups of service classes in Sysplex or per LPAR	Unweighted CPU SU/sec, fraction of LPAR share, or fractional #CPs	CP*	+	N/A	WLM Policy
Logical configuration	LPAR	Integer #processors	Any	O	+ but coarse grain	HMC+OS

PR/SM controlled   WLM controlled, PR/SM enforced   WLM controlled

© Copyright IBM Corp. 2016

Which capping techniques may be combined?

Type of capping➔	Initial (hard capping)	LPAR Absolute capping	LPAR Absolute group capping	Defined capacity [1]	LPAR group capacity [1]	Resource group capping
Initial (hard capping)		+	+	−	−	+
LPAR Absolute capping	+		+	+	+	+
LPAR Group Absolute capping	+	+		+	+	+
Defined capacity [1]	−	+	+		+	+
LPAR group capacity [1]	−	+	+	+		+
Resource group capping	+	+	+	+	+	

10   **(1)** Includes ABSMSUCAPPING=NO and ABSMSUCAPPING=YES

Agenda

New capping options

Reporting enhancements for CICS and IMS, and Mobile Workloads

IBM z13 Support

z/OS V2.2 enhancements

z/OS V2.1 highlights

Other service stream enhancements and recommendations

Statements regarding IBM future direction and intent are subject
to change or withdrawal, and represent goals and objectives only

Reporting enhancements for CICS and IMS, and Mobile Workloads

- Mobile Workload Pricing (MWP) is an IBM Software Pricing option, announced in May 2014.

- For eligible software it can reduce the cost of transactions that originate on mobile devices.
    - MWP can mitigate the impact of mobile workloads on sub-capacity license charges, specifically in the cases where higher mobile transaction volumes may cause a spike in machine utilization.

- *Reporting enhancements for CICS and IMS, and Mobile Workloads* introduce WLM enhancements that can simplify the identifying and reporting of the mobile-sourced transactions and their processor consumption.
    - WLM introduces a new transaction level attribute in the WLM classification rules that allows for the identification of mobile transactions and the reporting of their processor consumption.

- With exploiting levels of CICS and IMS, processor consumption data are made available on the transaction service and report classes including such transactions.

Enablement of Reporting Enhancements for Mobile Workloads

z/OS release Function	z/OS V2.1	z/OS V2.2	Other
WLM: Reporting Enhancements for Mobile Workloads	OA47042	OA47042	
RMF Reporting Enhancements for Mobile Workloads	OA48466	OA48466	
z/OSMF Reporting Enhancements for Mobile Workloads	PI47638	PI47638	
CICS TS Reporting Enhancements for Mobile Workloads			CICS 5.3
IMS TM Reporting Enhancements for Mobile Workloads			IMS 14 PI46933 (available) PI51948 (1H2016*)
SCRT and Billing System Support			SCRT 23.13.0 (1H2016*)

* Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.

WLM Support for Mobile Pricing

Mobile Workload Pricing can provide substantial savings for customers.

How are Mobile Workloads identified?

- **NEW** Via a transaction level **Reporting Attribute**
- Processor consumption data aggregated by WLM
- Reporting integrated into standard performance monitors (RMF) and low volume SMF records
- Applicable to wide range of workloads, including enclave work and CICS/IMS work

WLM Support for Mobile Pricing – Solution

- In their WLM classification rules, installations can classify transactions as **"mobile"**

- The assigned mobile attribute is independent from the assigned service and report class
  - Eliminates the need for using new dedicated classes for mobile workload reporting

- The mobile attribute is transparent to subsystems

- WLM tracks and reports the total and the mobile CPU consumption for all service and report classes
  - Eligible exploiters of the WLM Execution Delay Monitoring Services like CICS or IMS can provide CPU times for all their transactions
  - As soon as they do, total and mobile CPU consumption data is also available for CICS and IMS transaction service and report classes that previously did not report any CPU consumption data
  - Subsystems using independent enclaves can participate transparently; only the classification rules need to be updated.

- WLM also aggregates and reports the system-wide mobile consumption data

- Besides mobile there are two more categories (A and B) that are currently unused but may be used in the future

WLM Service Definition Changes:
New Reporting Attribute for Classification Rules

Classfication rules panel:

```
 Subsystem-Type   Xref   Notes   Options   Help
 ----------------------------------------------------------------------
                    Modify Rules for the Subsystem Type        Row 1 to 3 of 3
 Command ===> _____        Scroll ===> CSR

 Subsystem Type . : CICS          Fold qualifier names?    Y   (Y or N)
 Description  . . . CICS rules

 Action codes:    A=After        C=Copy        M=Move      I=Insert rule
                  B=Before       D=Delete row  R=Repeat    IS=Insert Sub-rule
                                                              <=== More
            ------Qualifier-------      Storage    Reporting  Manage Region
 Action     Type      Name      Start   Critical   Attribute  Using Goals Of

 _____  1 TC         BANKING   ____       NO       MOBILE     N/A
 _____  1 TC         HR        ____       NO       NONE       N/A
 _____  1 TN         ACCT      ____       NO       NONE       N/A
 ************************************ BOTTOM OF DATA *********************************
```

```
 File   Utilities   Notes   Options   Help
 ----------------------------------------------------------------------
 Functionality LEVEL030        Definition Menu       WLM Appl LEVEL030
 Command ===> _____

 Definition data set  . . : 'WLM.HBB7780.FCT.CRTME.SD4'

 Definition name  . . . . . SVDEF1    (Required)
 Description  . . . . . . . Service Definition 1

 Select one of the
 following options. . . . . ___   1.   Policies
                                  2.   Workloads
                                  3.   Resource Groups
                                  4.   Service Classes
                                  5.   Classification Groups
                                  6.   Classification Rules
                                  7.   Report Classes
                                  8.   Service Coefficients/Options
                                  9.   Application Environments
                                  10.  Scheduling Environments
                                  11.  Guest Platform Management Provider
```

The WLM Administrative Application level changes from 29 to 30

As soon as transactions are flagged as mobile, the functionality level of the service definition changes to 30

IBM

Reporting Attribute Mobile – An Example

- The Reporting Attribute is independent from the assigned service and report class
- Example:
  - Suppose part of the BANKING transactions flows in from mobile devices
  - Suppose mobile BANKING transactions flow in via TCP/IP Service TCP001
  - To differentiate those from „normal" BANKING transactions, insert a sub-rule
  - Specify the MOBILE reporting attribute for the sub-rule
  - ➤ WLM tracks and reports the total and the mobile CPU consumption for the service and report class of the BANKING transactions

```
                        Modify Rules for the Subsystem Type         Row 1 to 4 of 4
Command ===>                                                        Scroll ===> CSR

Subsystem Type . : CICS          Fold qualifier names?   Y   (Y or N)
Description  . . . CICS rules

Action codes:      A=After       C=Copy        M=Move       I=Insert rule
                   B=Before      D=Delete row  R=Repeat     IS=Insert Sub-rule
                                                               <=== More
            -------Qualifier--------        Storage    Reporting  Manage Region
Action      Type       Name       Start     Critical   Attribute  Using Goals Of

_____   1  TC         BANKING    ____        NO        NONE        N/A
_____   2    CT          TCP001  ____        NO        MOBILE      N/A
_____   1  TC         HR         ____        NO        NONE        N/A
_____   1  TN         ACCT       ____        NO        NONE        N/A
```

IBM.

Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.

New Qualifier Types for CICS and IMS

- The mobile classification can be based on any currently supported work qualifier, plus
  - Two new qualifiers for CICS:
    - Connection Type (CT): The name of the TCP/IP Service that received the request for this transaction
    - Transaction Class (TC): The name of the transaction class to which this transaction, or transid belongs
  - Two new qualifiers for IMS
    - Connection Type (CT): The port number of the TCP/IP Service that received the request for this transaction
    - Client Transaction Name (CTN): The name of the Transaction Pipe (TPIPE)
  - Expected to be used frequently to identify mobile transactions

Mobile Workload Pricing Reporting – RMF Sample Workload Activity Report

```
REPORT BY: POLICY=BASEPOL    WORKLOAD=CICS_WLD    SERVICE CLASS=CICSLOW    RESOURCE GROUP=*NONE
                                                  CRITICAL      =CPU
                                                  DESCRIPTION  =Low priority for CICS workloads


    -TRANSACTIONS-  TRANS-TIME HHH.MM.SS.TTT
AVG         0.00  ACTUAL                0
MPL         0.00  EXECUTION             2
ENDED        626  QUEUED                0
END/S       0.70  R/S AFFIN             0
#SWAPS         0  INELIGIBLE            0
EXCTD      15836  CONVERSION            0
AVG ENC     0.00  STD DEV               0
REM ENC     0.00
MS ENC      0.00


TRANSACTION APPL %:  TOTAL:  CP  15.30  AAP/IIP ON CP    0.00  AAP/IIP  29.50
                     MOBILE: CP  13.40  AAP/IIP ON CP    0.00  AAP/IIP  20.46


        RESPONSE TIME EX    PERF
SYSTEM       ACTUAL%    VEL% INDX


*ALL          100     N/A  0.5
CB8B          100     N/A  0.5
```

SMF Record Type 72 Changes: Service and Report Class Level

- SMF Record Type 72 – Workload Activity, Storage Data, and Serialization Delay
  Subtype 3 – Workload Activity
  Service /Report Class Period Data Section
  (RMF SMF record level x'77')

Offsets	Name	Length	Format	Description
624 270	R723TSUCP	8	Floating	Total Service units consumed by transactions, executed on general purpose processors.
632 278	R723TSUSP	8	Floating	Total Service units consumed by transactions, executed on specialty processors.
640 280	R723TSUOCP	8	Floating	Total Service units consumed by transactions, eligible to run on specialty processors, but executed on general purpose processors.
648 288	R723MSUCP	8	Floating	Service units consumed by transactions, classified with reporting attribute MOBILE, executed on general purpose processors CP.
656 290	R723MSUSP	8	Floating	Service units consumed by transactions, classified with reporting attribute MOBILE, executed on specialty processors.
664 298	R723MSUOCP	8	Floating	Service units consumed by transactions, classified with reporting attribute MOBILE, eligible to run on specialty processors, but executed on general purpose processors.

IBM®

SMF Record Type 72 Changes:
Service and Report Class Level (cont.)

Offsets	Name	Length	Format	Description
672 2A0	R723ASUCP	8	Floating	Service units consumed by transactions, classified with reporting attribute CATEGORYA, executed on general purpose processors.
680 2A8	R723ASUSP	8	Floating	Service units consumed by transactions, classified with reporting attribute CATEGORYA, executed on specialty processors.
688 2B0	R723ASUOCP	8	Floating	Service units consumed by transactions, classified with reporting attribute CATEGORYA, eligible to run on specialty processors, but executed on general purpose processors.
696 2B8	R723BSUCP	8	Floating	Service units consumed by transactions, classified with reporting attribute CATEGORYB, executed on general purpose processors.
704 2C0	R723BSUSP	8	Floating	Service units consumed by transactions, classified with reporting attribute CATEGORYB, executed on specialty processors.
712 2C8	R723BSUOCP	8	Floating	Service units consumed by transactions, classified with reporting attribute CATEGORYB, eligible to run on specialty processors, but executed on general purpose processors.

SMF Record Type 70 Changes: System Level

- SMF Record Type 70 – RMF Processor Activity
  Subtype 1 – CPU, PR/SM, and ICF Activity
  CPU Control Section
  (RMF SMF record level x'77')

Offsets	Name	Length	Format	Description
268 10C	SMF70LACM	4	Binary	Long-term average of CPU service (millions of service units) consumed by transactions classified with reporting attribute MOBILE
272 110	SMF70LACA	4	Binary	Long-term average of CPU service (millions of service units) consumed by transactions classified with reporting attribute CATEGORYA.
276 114	SMF70LACB	4	Binary	Long-term average of CPU service (millions of service units) consumed by transactions classified with reporting attribute CATEGORYB.

XML Format WLM service definitions recommended

- For many z/OS releases WLM has supported service definitions in **XML format**
  - z/OSMF WLM task
  - ISPF Administrative Application: "Save as XML"…

- XML format avoids particular problems with the ISPF tables format, namely coexistence behavior, when a new functionality level needs to be introduced, and the number of table columns needs to be extended.
  - For example, OA47042 introduces such a change.

- **Recommendation**:
  Use the XML-format for your WLM service definition data sets.

IBM z Systems
The innovation continues

Agenda

IBM z13 Support

> **z13 base support**
>
> **zIIP SMT support**
>
> **HiperDispatch and capping enhancements**

z/OS V2.2 enhancements

z/OS V2.1 highlights

Other service stream enhancements and recommendations

Agenda

New capping options

Reporting enhancements for CICS and IMS, and Mobile

IBM z13 Support

z/OS V2.2 enhancements

z/OS V2.1 highlights

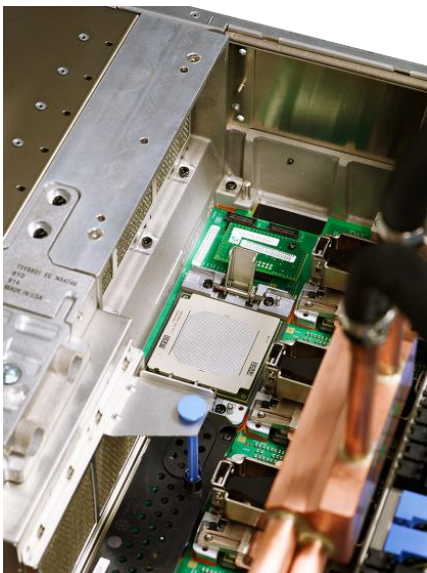Other service stream enhancements and recommendations

WLM/SRM support overview for IBM z13

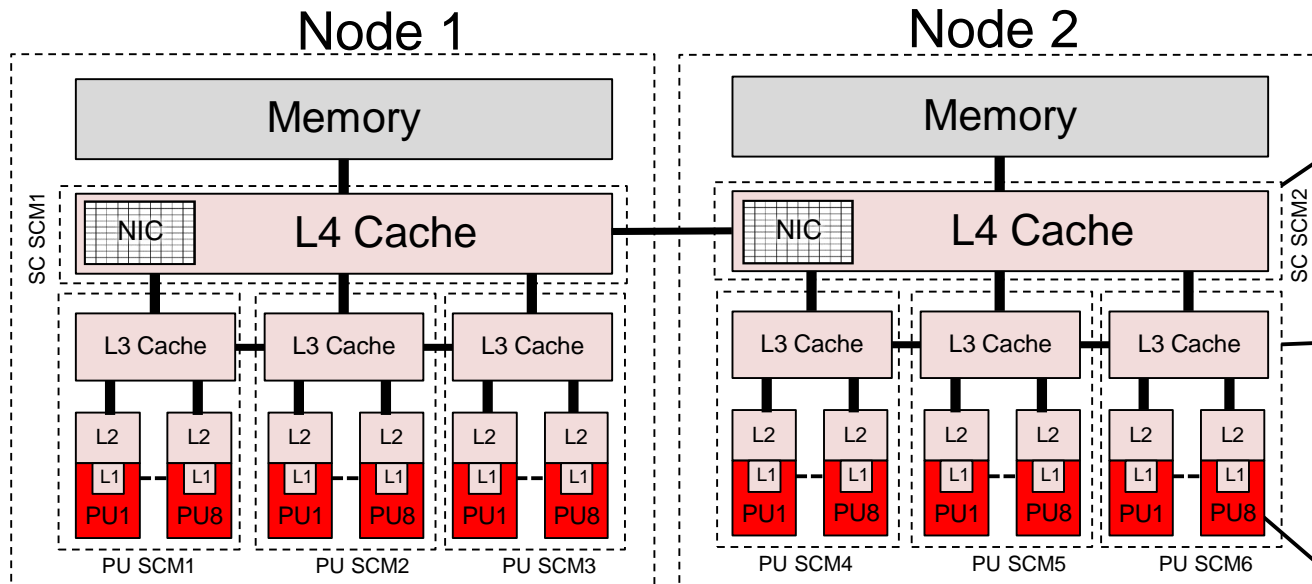| z/OS release ╱ Function	V2.2	V2.1	V1.13
z13 GA2 LPAR Absolute group capping	OA47752	OA47752	
z13 Support (base)	+	OA43622 OA47021	OA43622
z13 HiperDispatch Optimizations	OA47968 (Included in GA code)	OA47968	OA47968
zIIP SMT Support	+	OA43622	
Hiper-Dispatch z13 & zEC12 — Unpark while capped / Unused capacity refinement / Prime cycle elimination	+	OA43622	
SRM storage management changes in support of RSM for z13	OA48858	OA44504 OA46396 OA48858	OA44504 OA46396 OA48858

Base z13 support

- New limits for z13
  - 85 LPARs
  - Up to 141 processors per CPC
  - Up to 141-way on z/OS V2.1 (non-SMT mode)
  - Up to 128-way on z/OS V2.1 (SMT mode), or z/OS <V2.1
    - Maximum active threads in SMT mode is 213 with zIIP:CP ratio of 2:1

- New Cache topology
  - Chip, node, drawer
  - No longer using "books"
  - z/OS HiperDispatch uses new topology information to place work topologically close – to maximize cache efficiency

z13 CPC Drawer Cache Hierarchy Detail

Single CPC Drawer View (N30 Model) – 2 Nodes

Node 1

Memory

SC SCM1

NIC	L4 Cache

L3 Cache	L3 Cache	L3 Cache

L2	L2	L2	L2	L2	L2
L1	L1	L1	L1	L1	L1
PU1	PU8	PU1	PU8	PU1	PU8

PU SCM1 PU SCM2 PU SCM3

Node 2

Memory

SC SCM2

NIC	L4 Cache

L3 Cache	L3 Cache	L3 Cache

L2	L2	L2	L2	L2	L2
L1	L1	L1	L1	L1	L1
PU1	PU8	PU1	PU8	PU1	PU8

PU SCM4 PU SCM5 PU SCM6

2 SC SCM Chips
(1 per node)

L4 Cache

6 PU SCM Chips
(3 per node)

L3 Cache

* Up to 8 PU cores per chip

Single PU core

* Not all PU's active

Node 1 - Caches
- L1 private 96k i, 128k d
- L2 private 2 MB i + 2 MB d
- L3 shared 64 MB / chip
- L4 shared 480 MB / node
  - plus 224 MB NIC

Node 2 - Caches
- L1 private 96k i, 128k d
- L2 private 2 MB i + 2 MB d
- L3 shared 64 MB / chip
- L4 shared 480 MB / node
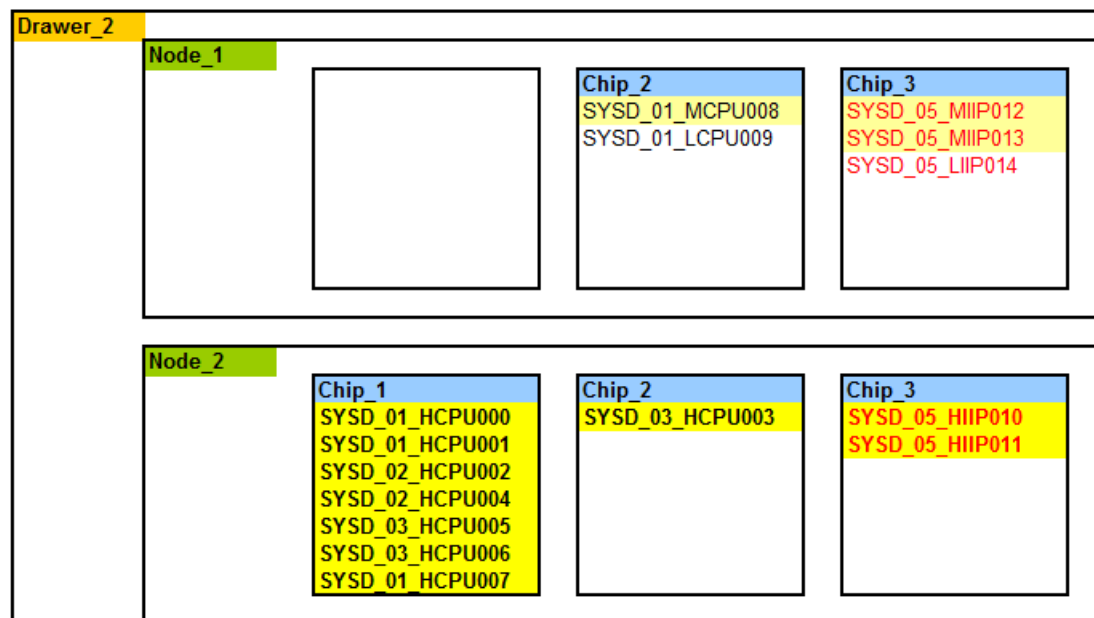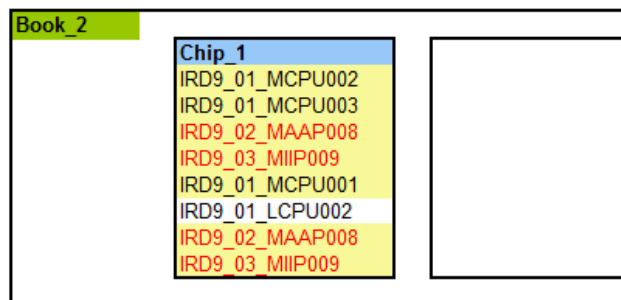 - plus 224 MB NIC

WLM Topology Report Tool (As-is)

- New *as-is* tool available for download from the WLM homepage
  - http://www.ibm.com/systems/z/os/zos/features/wlm/WLM_Further_Info_Tools.html#Topology

- Visualizes mapping of HiperDispatch affinity nodes to physical structure

- Supports IBM zEC10 and later

- To use:
  1. Download from above location
  2. Run installer
  3. Collect SMF99.14 records
  4. Upload Host code to a z/OS system

Sample output (z13):

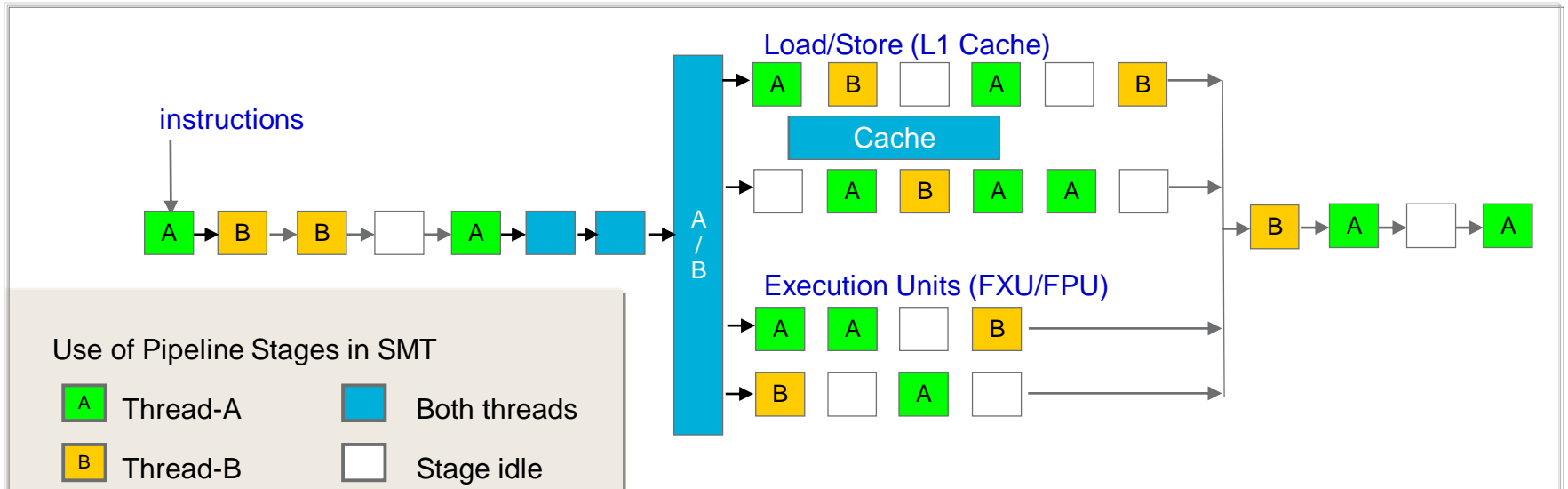**Topology for 01-30-2015-14:08:32 ,System: SYSD**

Drawer_2

Node_1

Chip_2	Chip_3
SYSD_01_MCPU008	SYSD_05_MIIP012
SYSD_01_LCPU009	SYSD_05_MIIP013
	SYSD_05_LIIP014

Node_2

Chip_1	Chip_2	Chip_3
SYSD_01_HCPU000	SYSD_03_HCPU003	SYSD_05_HIIP010
SYSD_01_HCPU001		SYSD_05_HIIP011
SYSD_02_HCPU002		
SYSD_02_HCPU004		
SYSD_03_HCPU005		
SYSD_03_HCPU006		
SYSD_01_HCPU007		

Sample output (zEC12): **Topology for 07-21-2014-13:44:27 ,Syste**

Book_2

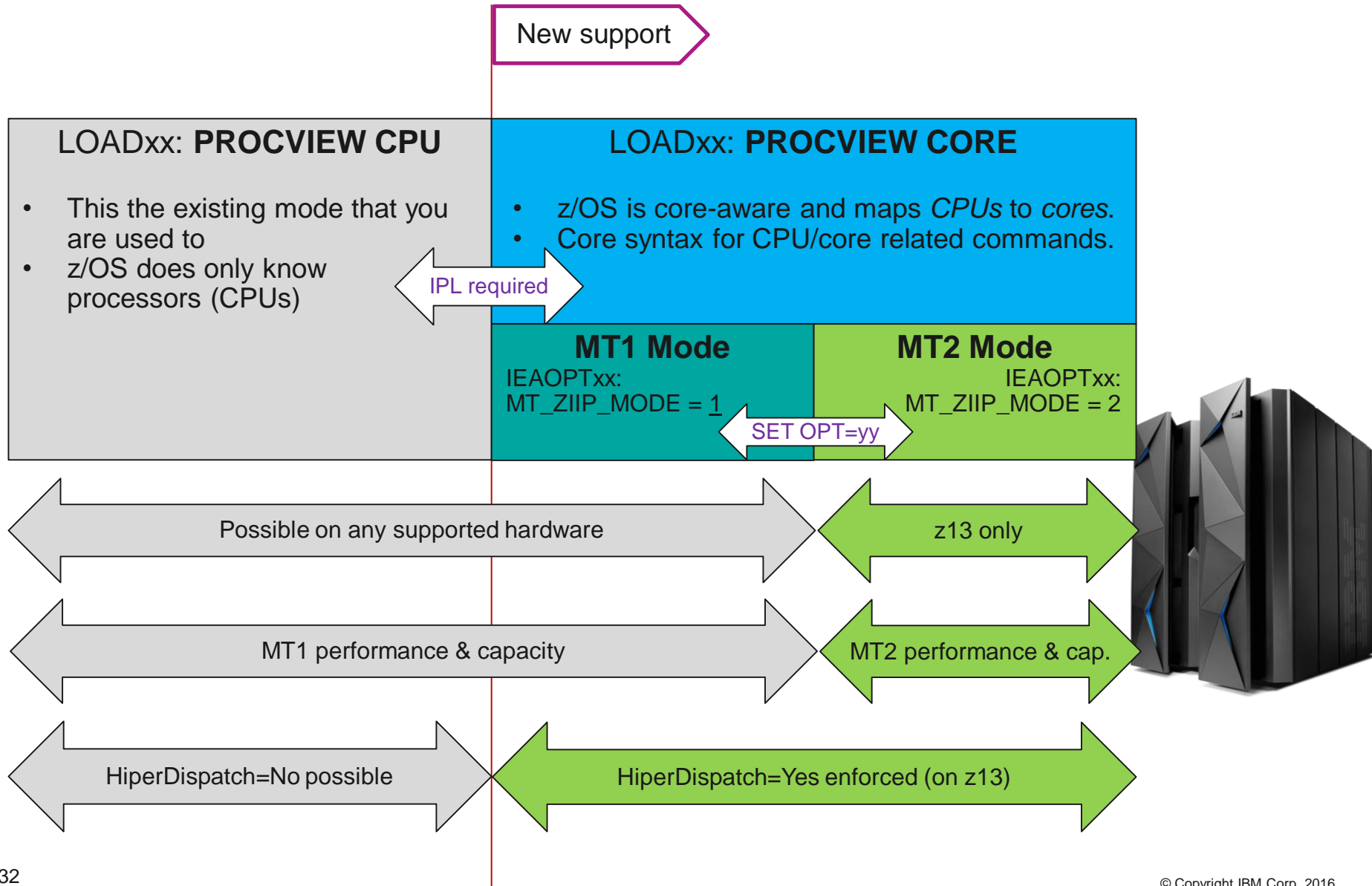| Chip_1 |
|---|
| IRD9_01_MCPU002 |
| IRD9_01_MCPU003 |
| IRD9_02_MAAP008 |
| IRD9_03_MIIP009 |
| IRD9_01_MCPU001 |
| IRD9_01_LCPU002 |
| IRD9_02_MAAP008 |
| IRD9_03_MIIP009 |

30

2016

Motivation for Simultaneous Multi Threading

- "Simultaneous multithreading (SMT) permits multiple independent threads of execution to better utilize the resources provided by modern processor architectures."*
- With z13, SMT allows up to two instructions streams per core to run simultaneously to get better overall throughput
- SMT is designed to make better use of processor hardware units

- On z/OS, SMT is available for zIIP processing:
    - Two concurrent threads are available per core
    - Capacity (throughput) usually increases
    - Performance may be superior using single threading

What is new with multithreading support?

New support

LOADxx: PROCVIEW CPU

- This the existing mode that you are used to
- z/OS does only know processors (CPUs)

IPL required

LOADxx: PROCVIEW CORE

- z/OS is core-aware and maps *CPUs* to *cores*.
- Core syntax for CPU/core related commands.

MT1 Mode
IEAOPTxx:
MT_ZIIP_MODE = 1

SET OPT=yy

MT2 Mode
IEAOPTxx:
MT_ZIIP_MODE = 2

Possible on any supported hardware

z13 only

MT1 performance & capacity

MT2 performance & cap.

HiperDispatch=No possible

HiperDispatch=Yes enforced (on z13)

32

New terminology for SMT…

- z/OS logical processor (CPU) ➔ Thread
  - A thread implements (most of) the System z processor architecture
  - z/OS dispatches work units on threads
  - In MT mode two threads are mapped to a logical core

- Processor core ➔ Core
  - PR/SM dispatches logical core on a physical core
    - Thread density 1 (TD1) - when only a single thread runs on a core
    - Thread density 2 (TD2) - when both threads run on a core

- MT1 Equivalent Time (MT1ET)
  - z/OS CPU times are normalized to the time it would have taken to run same work in MT-1 mode on a CP
    - ASCB, ASSB, …, SMF30, SMF32, SMF7x, …
  - You will usually not see the term MT1ET because it is implied

- Several new metrics to describe how efficiently core resources could be utilized…

33

…and several new metrics for SMT…

- New metrics:
  - WLM/RMF: Capacity Factor (CF), Maximum Capacity Factor (mCF)
  - RMF: Average Thread Density, Core busy time, Productivity (PROD)

- How are the new metrics derived?

  - Hardware provides metrics (counters) describing the efficiency of processor (cache use/misses, number cycles when one or two threads were active…)

  - LPAR level counters are made available to the OS

  - MVS HIS component and supervisor collect LPAR level counters. HIS provides HISMT API to compute average metrics between "previous" HISMT invocation and "now" (current HISMT invocation)

    - HIS address space may be active but is not required to be active

  - System components (WLM/SRM, monitors such as RMF) retrieve metrics for management and reporting

z/OS MT Capacity Factors - used by WLM/SRM

- **Capacity Factor (CF)**
  - How much work core <u>actually completes</u> for a given workload mix at current utilization - relative to single thread
  - Therefore, MT1 Capacity Factor is 1.0 (100%)
  - MT2 Capacity Factor is workload dependent
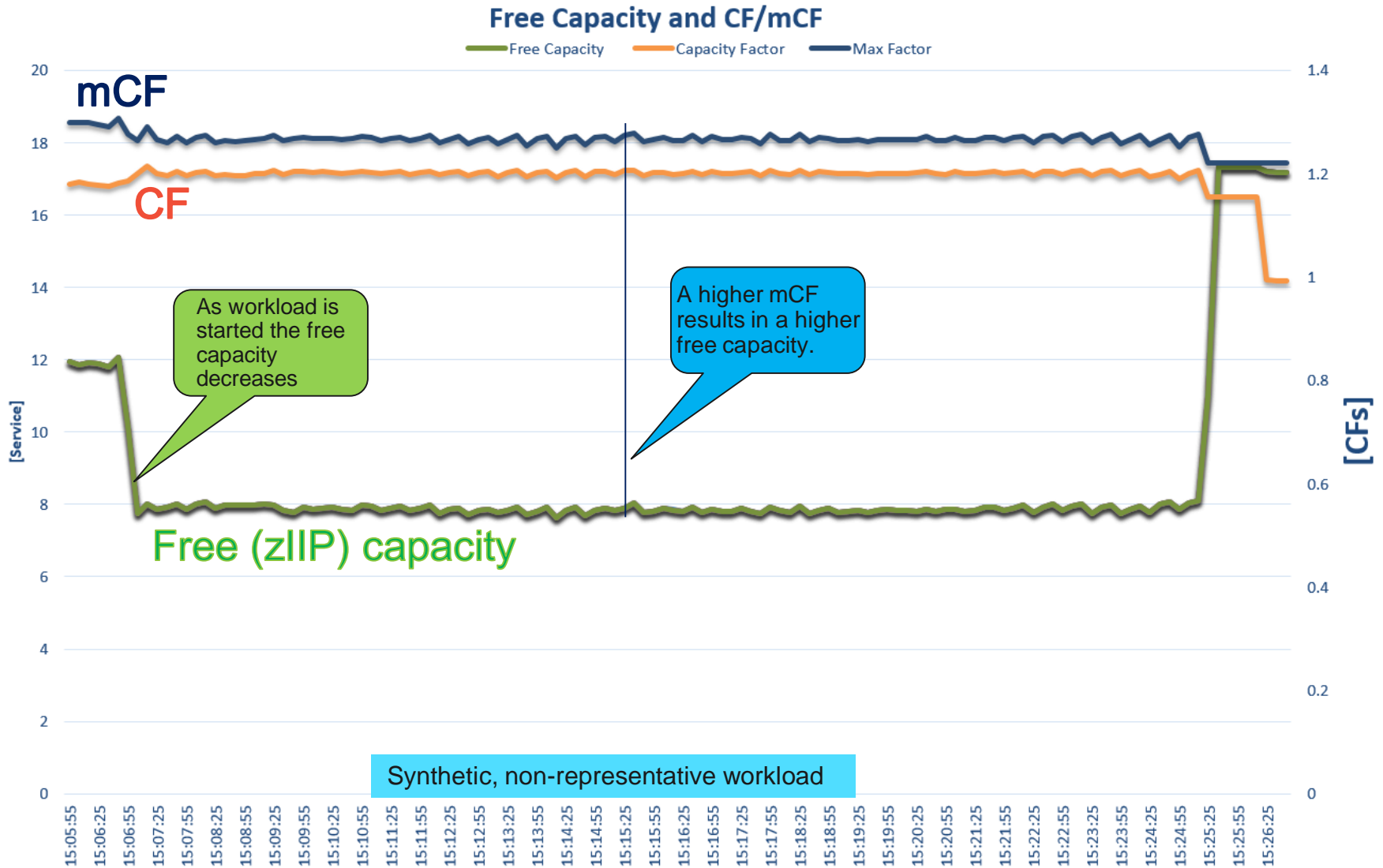 - Describes the actual, current efficiency of MT2

- **Maximum Capacity Factor (mCF)**
  - How much work a core <u>can complete</u> for a given workload mix <u>at most</u> relative to MT-1 mode
  - Used to estimate MT2 efficiency if the system was fully utilized
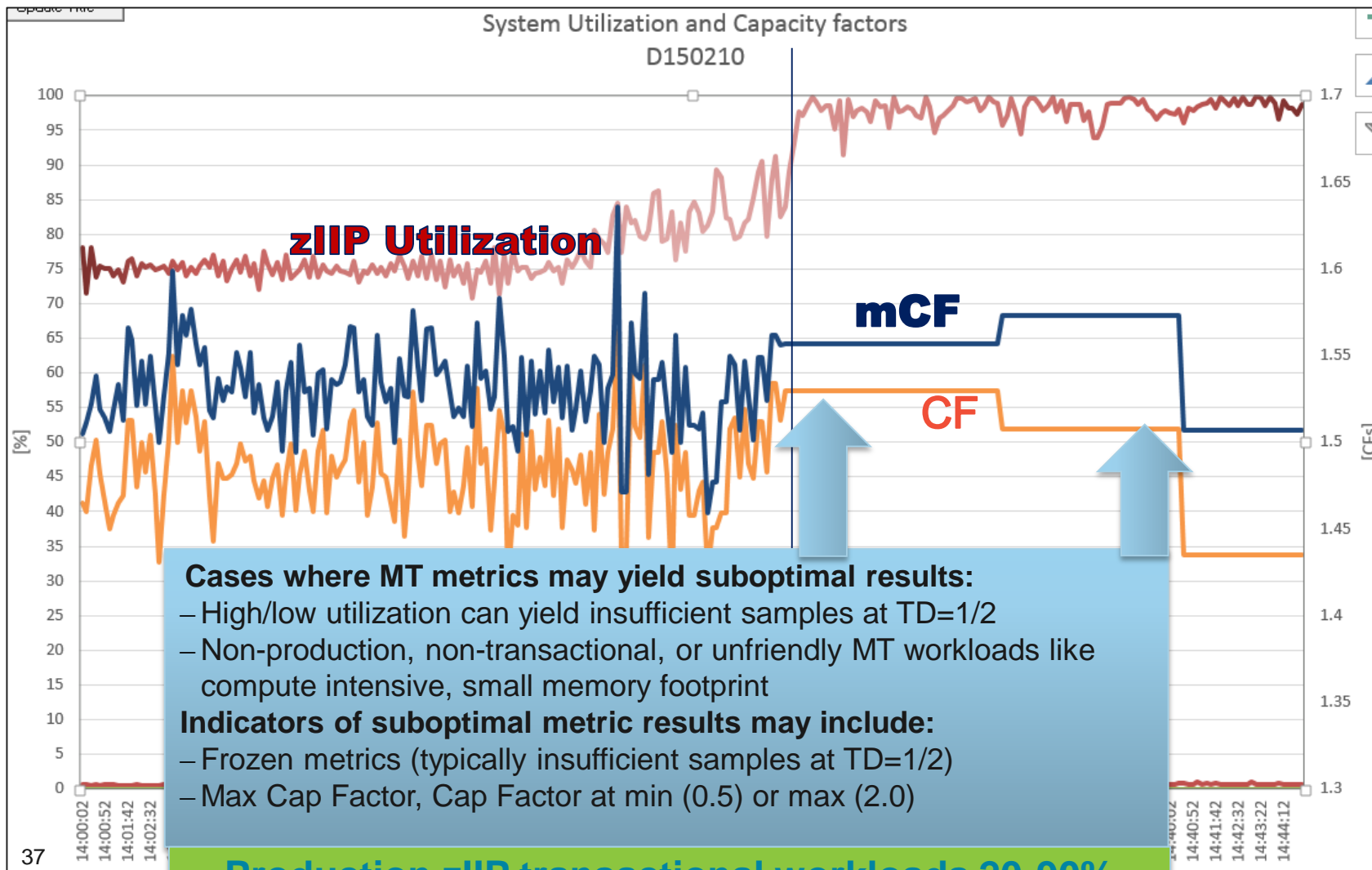   - E.g., to derive WLM view of total system capacity or free capacity

- Value range of CF and mCF is [0.5 … 2.0]
  - Expect CF in a range of 1.0 -1 .4 (100%-140%) for typical workloads
  - Untypical ("pathological") workloads may see untypical/pathological CF/mCFs, such as <1

Sample Capacity and maximum Capacity Factor
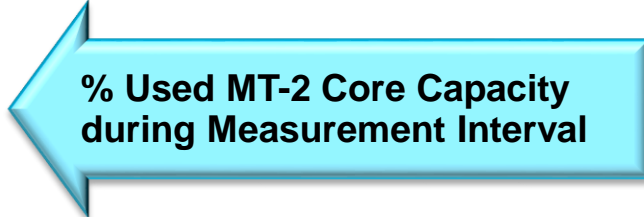

36

Atypical Capacity and maximum Capacity Factors

System Utilization and Capacity factors
D150210

zIIP Utilization

mCF

CF

Cases where MT metrics may yield suboptimal results:
– High/low utilization can yield insufficient samples at TD=1/2
– Non-production, non-transactional, or unfriendly MT workloads like compute intensive, small memory footprint

Indicators of suboptimal metric results may include:
– Frozen metrics (typically insufficient samples at TD=1/2)
– Max Cap Factor, Cap Factor at min (0.5) or max (2.0)

Production zIIP transactional workloads 20-90% busy generally receive optimal metric results. z/OS V2.2 includes optimizations.

37

Additional z/OS MT metrics reported by RMF

- Core Busy Time
  - Time any thread on the core is executing instructions when core is dispatched to physical core

- Average Thread Density
  - Average number of executing threads during Core Busy Time (Range: 1.0 - 2.0)

- Productivity
  - Core Busy Time Utilization (percentage of used capacity) for a given workload mix
  - Productivity represents capacity in use (CF) relative to capacity total (mCF) during Core Busy Time.

% Used MT-2 Core Capacity during Core Busy Time

- Core Utilization
  - Capacity in use relative to capacity total over some time interval
  - Calculated as Core Busy Time x Productivity

% Used MT-2 Core Capacity during Measurement Interval

38

Transitioning into MT mode (Enablement)

- LOADxx PROCVIEW CORE <u>enables</u> use of SMT mode

  - IPL required to switch between PROCVIEW CPU and CORE
  - While an LPAR is in PROCVIEW CORE mode with zIIP in MT1 mode, the HMC may recognize an exception (can be ignored)

⛓ ZOSLP2	Z13	🔲 ❌ Exception ESA/390	SYS1	z/OS	V2R2	SYS1		LPAR Image

  - Causes syntax and semantic to change for <u>core-aware commands</u>.
    - -LOADxx …CORE,CPU_OK allows using CPU as a synonym of CORE

- ☐   ure

```
CORE STATUS: HD=Y    MT=2   MT_MODE: CP=1   zIIP=1
ID      ST      ID RANGE      VP   ISCM   CPU THREAD STATUS
0000    +       0000-0001   M    FC00    +N
0001    -       0002-0003
0002    -       0004-0005
0003    +I      0006-0007   M    0200    +N
0004    -I      0008-0009
0005    -       000A-000B
```

z/OS Commands requiring CORE keyword

- Config Core(x),Online   Configs core online for MT Mode

- Config Core(x),Offline   Configs all threads on core offline

- Config Member=xx   Configs cores according to CONFIGxx

- Config Online or Config Offline   Lists eligible cores to config

- Reply to IEE522D accepts CORE(x) to configure

- Display Matrix=Core   Displays core status (new message)

- Display Matrix=Config(xx)   CONFIGxx vs system differences

40

Transitioning into MT mode (Activation)
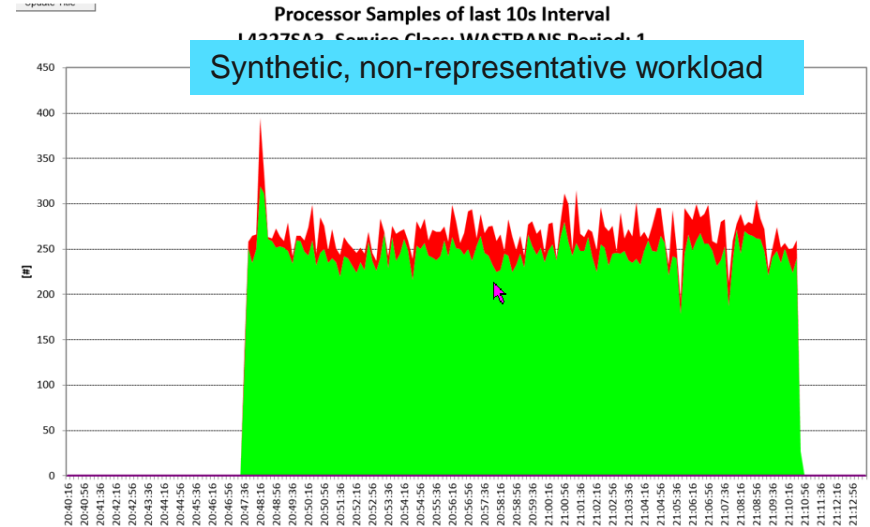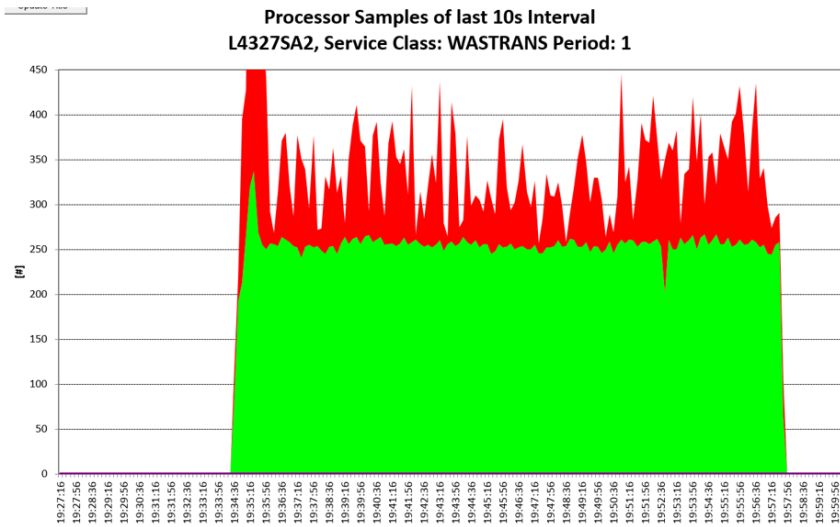
MT-2 mode Activation and Deactivation

- IEAOPTxx new parameter

  - MT_ZIIP_MODE=**1** specifies MT-1 mode for zIIPs
  - MT_ZIIP_MODE=2 specifies MT-2 mode for zIIPs

- Switch dynamically between MT-1 and MT-2 mode via SET OPT=xx

- Performance-wise, MT-1 mode and PROCVIEW CPU are equivalent

- Some WLM considerations…Details later

```
IWM066I MT MODE CHANGED FOR PROCESSOR CLASS zIIP. THE MT MODE WAS
CHANGED FROM 1 TO 2.
IWM063I WLM POLICY WAS REFRESHED DUE TO A PROCESSOR SPEED CHANGE OR MT
MODE CHANGE
```

Processor samples may change when going to MT-2 mode

Processor Samples of last 10s Interval
L4327SA2, Service Class: WASTRANS Period: 1

Synthetic, non-representative workload

- In MT-2 mode we see less processor delays resulting in a higher execution velocity

42

© Copyright IBM Corp. 2016

Free capacity and service consumption MT-1 vs. MT-2

zIIP Service Consumption by Importance Level
S50411I1 J0929678 R75

Synthetic, non-representative workload

43

Sample execution velocity MT-1 vs. MT-2

Last zIIP Delay and Last zIIP Using
S50411I1 J0929678 R75,

Synthetic, non-representative workload

MT1 performance & capacity

MT2 performance & cap.

Workload level
1x   3x   5x

Workload level
1x   3x   5x

Workload level
1x   3x   5x

Last zIIPD — Last zIIPU

44

WLM/SRM message changes (OA43622)

- **IRA866I HIPERDISPATCH=YES FORCED DUE TO PROCVIEW=CORE**
  - HIPERDISPATCH=YES is enforced because PROCVIEW=CORE was specified in the load parameter member (LOADxx) on HW capable of supporting MT.

- **IWM066I MT MODE CHANGED FOR PROCESSOR CLASS zIIP. MT MODE CHANGED FROM nn TO mm.**
  - The System successfully changed the MT Mode for the respective processor class. ProcessorClass specifies the processor for which the MT Mode was changed . nn specifies the previous MT Mode, mm specifies the new effective MT Mode

- **IWM067I SETTING MT MODE FAILED FOR PROCESSOR CLASS zIIP DUE TO THE FOLLOWING: reason, problem.**
  - The System could not change the MT Mode. "problem" can be one of the following
    - SPECIFIED VALUE IS NOT SUPPORTED BY Z/OS
    - SPECIFIED VALUE IS NOT SUPPORTED BY HARDWARE
    - HIPERDISPATCH FUNCTION IS NOT ACTIVE
    - WAITCOMPLETION=YES IS SET
    - CONFIGURATION OF PROCESSORS FAILED
    - FUNCTIONAL PROBLEM

  SoD: IBM plans to offer only event-driven dispatching (Wait Completion = No) and not to offer time-driven dispatching (Time Slicing or Wait Completion = Yes) on the high end z System server following z13. Event-driven dispatching, the default for many years, better manages processor resource to adjust for fluctuations in demand among partitions.

Control block changes (IRARMCTZ)

```
OFFSET     OFFSET
DECIMAL      HEX   TYPE            LENGTH   NAME (DIM)        DESCRIPTION
========  ========  ==================  ======  ==============  ==============================
  1264    (4F0) CHARACTER           12     RMCTZ_MT_AREA    MT section
  1264    (4F0) BIT(8)               1     RMCTZ_MT_FLAGS   MT Flags
               1... ....                   RMCTZ_PROCVIEW   1:=core
               .1.. ....                   RMCTZ_MT         1:=Multiple threads per core

  1268    (4F4) UNSIGNED             4     RMCTZ_MT_STAT    Current status
  1270    (4F6) UNSIGNED             1     RMCTZ_MT_ZIIP    …for zIIPs

  1272    (4F8) UNSIGNED             4     RMCTZ_MT_OPT OPT Requested status
  1274    (4FA) UNSIGNED             1     RMCTZ_MT_OPT_ZIIP…for zIIPs
```

z13 – SMT: Postprocessor CPU Activity Report

- PP CPU activity report provides new metrics when SMT is active
  - MT Productivity and Utilization of each logical core
  - MT Multi-Threading Analysis section displays MT Mode, MT Capacity Factors and average Thread Density

- Contains core and thread level metrics, e.g.
  - LPAR Busy: PR/SM dispatching logical core to physical
  - MVS Busy: Unparked logical CPU not waiting
  - Parked: Logical CPU parked

% Used MT-2 Core Capacity during Core Busy Time

% Used MT-2 Core Capacity during Measurement Interval

```
                                        C P U   A C T I V I T Y
z/OS V2R1              SYSTEM ID CB8B              DATE 02/2015        INTERVAL 15.00.004
                      RPT VERSION V2R1 RMF         TIME 11.00.00       CYCLE 1.000 SECONDS
---CPU---      --------------- TIME % -----------   --- MT % ----  LOG PROC    --I/O INTERRUPTS-
NUM  TYPE    ONLINE   LPAR BUSY  MVS BUSY  PARKED    PROD    UTIL   SHARE %      RATE   % VIA TPI
 0    CP     100.00    68.07      67.94     0.00     100.00  68.07  100.0 HIGH   370.1   13.90
 1    CP     100.00    46.78      46.78     0.00     100.00  46.78   52.9 MED      5.29  16.93
...
TOTAL/AVERAGE           8.66      54.17              100.00   8.66  152.9        375.3   13.95
 A    IIP    100.00    48.15      41.70     0.00      85.84  41.33  100.0 HIGH
                                  35.66     0.00
 B    IIP    100.00    38.50      32.81     0.00      85.94  33.09  100.0 HIGH
                                  26.47     0.00
...
TOTAL/AVERAGE          29.48      23.23               86.47  25.39  386.7
------------ MULTI-THREADING ANALYSIS ---------------
CPU TYPE     MODE     MAX CF          CF        AVG TD
   CP          1      1.000        1.000        1.000
   IIP         2      1.485        1.279        1.576
```

z13 – SMT: Monitor III CPC Report

MT Mode and Productivity for zIIP processors

```
                              RMF V2R1    CPC Capacity                    Line 1
of 50

Samples: 60         System: CB88  Date: 02/02/15  Time: 11.00.00  Range: 60
Sec

Partition:  CB88          2964 Model 731
CPC Capacity:     3935    Weight % of Max: 50.1    4h Avg:  138    Group:
N/A
Image Capacity: 1777      WLM Capping %:    0.0    4h Max:  177    Limit:
N/A
MT Mode IIP:       2    Prod % IIP:       80.9


Partition  --- MSU ---  Cap   Proc    Logical Util %    - Physical Util % -
            Def    Act  Def   Num     Effect    Total   LPAR    Effect   Total

*CP                            390                       0.8     43.7     44.5
CB8B         0    192   NO    10.0     15.0     15.1     0.0      4.8      4.9
CB8D         0    134   NO    15.0      7.0      7.0     0.0      3.4      3.4
CB8E         0    330   NO    14.0     18.4     18.6     0.1      8.3      8.4
CB88         0    182   NO    14.0     10.2     10.3     0.0      4.6      4.6
C05          0    140   NO    14.0      7.9      7.9     0.0      3.5      3.6
C06          0    150   NO    14.0      8.4      8.4     0.0      3.8      3.8
LP1          0    507   NO     4.0      100      100     0.0     12.9     12.9
```

SMT mode enabled:
Processor data at logical core granularity

SMT mode disabled:
Processor data at logical processor granularity

Transitioning into MT2 mode: WLM considerations (1)

- **Less overflow from zIIP to CPs** may occur because
  - zIIP capacity increases, and
  - number of zIIP CPUs double

- CPU time and CPU service **variability may increase**, because
  - Threads which are running on a core at the same time influence each other
  - Threads may be dispatched at TD1 or TD2
    - Unlike other OS, z/OS attempts to dispatch threads densely

- Sysplex workload routing: routing recommendation may change because
  - zIIP capacity will be adjusted with the mCF to reflect MT2 capacity
  - mCF may change as workload or workload mix changes

Transitioning into MT2 mode: WLM Considerations (2)

- **Goals should be verified** for zIIP-intensive work, because

    - The number of zIIP CPUs double and the achieved velocity may change
        - "Chatty" (frequent dispatches) workloads may profit because there is a chance of more timely dispatching
        - More capacity is available
        - Any single thread will effectively run at a reduced speed and the achieved velocity will be lower.
        Affects processor speed bound work, such as single threaded Java batch

- MT-2 APPL% numbers can continue to be used to understand relative core utilization in a given interval, at times of comparable maxCFs.
However, the maxCF needs to be considered when comparing APPL% across different workloads or times with different maxCF values.
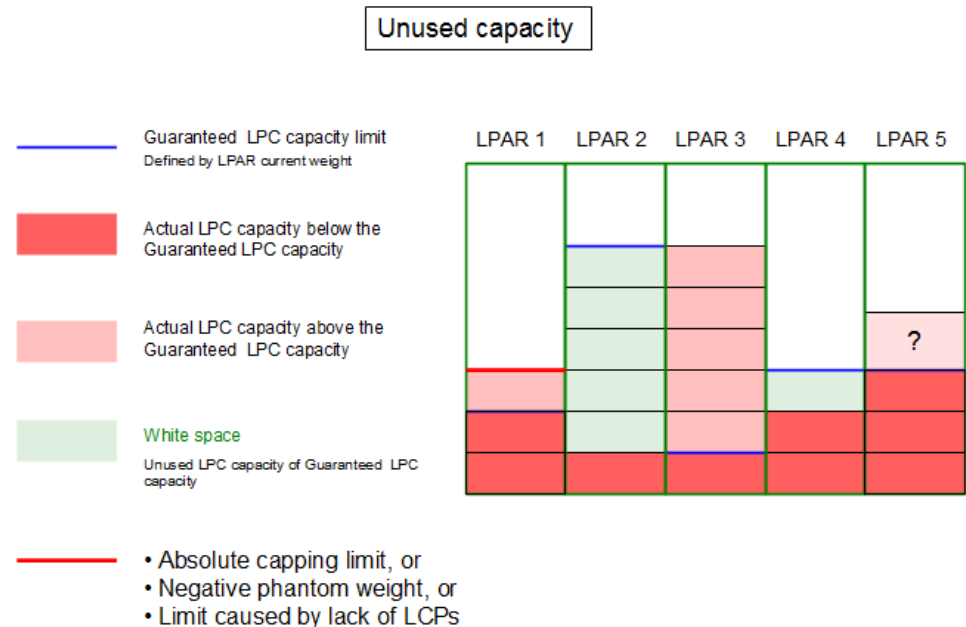
HiperDispatch "Unpark while capped"

- Previously, HiperDispatch

  - Parked all Vertical Low (VL) processors when a system capped via positive phantom weight
    - VLs are used for discretionary capacity and not required to absorb the LPAR weight
    - However, it was seen that, for some workloads, the reduced number of logical processors made it difficult to fully utilize the cap target capacity.
  - Unparked all VL processors when a system was capped by negative phantom weight, or some cases of PR/SM absolute capping

- Now, HiperDispatch can unpark VL processors if the processors can be used efficiently.

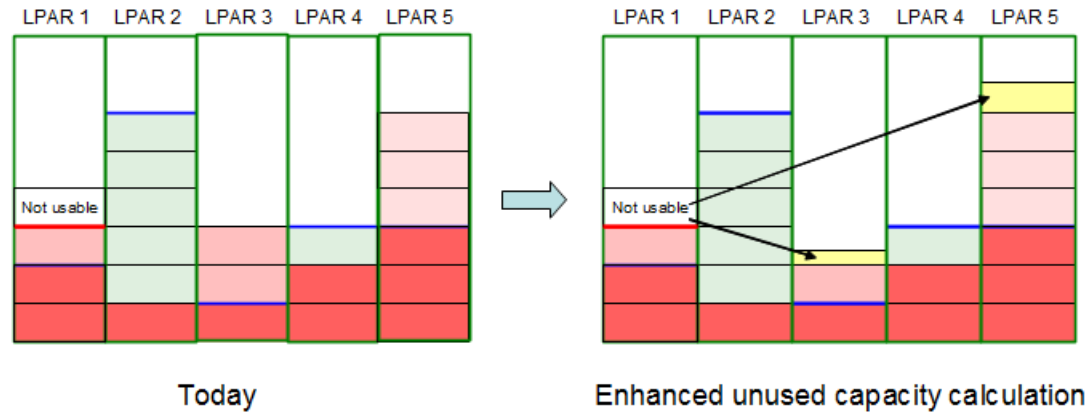# HiperDispatch refinement of "unused capacity" use

▪ HiperDispatch decisions consider the CPC-wide 'unused capacity share' situation

▪ The 'unused capacity share' calculation was enhanced to also include the LPAR configuration values
  – absolute capping value
  – negative phantom weight
  – number of logical processors
  – effective defined capacity
    and group capacity limit
  of possible 'unused capacity'
  receivers

CPC with 5 LPARs. LPAR1 has an absolute capping limit, which is indicated with the red line. LPAR2, and LPAR4 are unused capacity donors, while LPAR1 / 3 / 5 are unused capacity receivers.

Unused capacity

Guaranteed LPC capacity limit
Defined by LPAR current weight

Actual LPC capacity below the
Guaranteed LPC capacity

Actual LPC capacity above the
Guaranteed LPC capacity

White space
Unused LPC capacity of Guaranteed LPC
capacity

LPAR 1   LPAR 2   LPAR 3   LPAR 4   LPAR 5

?

• Absolute capping limit, or
• Negative phantom weight, or
• Limit caused by lack of LCPs

HiperDispatch refinement of "unused capacity" use

Enhanced unused capacity calculation

Today

Enhanced unused capacity calculation

- Figure on the left shows today's unused capacity calculation, which does not consider LPAR capping limits.

- Unused capacity calculation is only based on the receiver's weight share.

- Figure on the right shows an example of enhanced unused capacity calculation. It considers the capping limits of the receivers.

- Because LPAR1 is not able to use its total unused capacity share its 'not usable' unused capacity share portion increases the unused capacity share of LPAR5.

OA47968: HiperDispatch Optimizations for z13

- Vertical Low (VL) processors are used to absorb discretionary ("above the weight) processor capacity. VLs may float between different physical processors – consuming free physical capacity not used by other logical processors

- With OA47968 HiperDispatch takes benefit of the fact that lower VL numbers are likely to be topologically "closer" to the LPAR's VH and VM processors

- Visible effect is that the park time in the RMF CPU activity report should be increasing from the low to the high processor numbers
    - Due to weight changes numbers can still decrease

- On z13, even in the presence of free CPC capacity, unparking can be more restrictive, based on effective capacity used on the VM and VL processors.

Sample CPU Activity Report…
showing high VL numbers unparked

0---CPU---		---------------- TIME % ----------------				LOG PROC		--I/O INTERRUPTS--	
NUM	TYPE	ONLINE	LPAR BUSY	MVS BUSY	PARKED	SHARE %		RATE	% VIA TPI
0	CP	100.00	73.07	73.01	0.00	100.0	HIGH	331.0	47.48
...									
D	CP	100.00	62.53	62.49	0.00	100.0	HIGH	12768	14.71
E	CP	100.00	50.63	53.18	0.00	50.0	MED	134.8	60.51
F	CP	100.00	5.03	41.30	85.77	0.0	LOW	0.00	0.00
10	CP	100.00	5.14	38.64	84.88	0.0	LOW	0.00	0.00
11	CP	100.00	4.10	42.47	88.22	0.0	LOW	0.00	0.00
12	CP	100.00	0.00	-----	100.00	0.0	LOW	0.00	0.00
13	CP	100.00	0.00	-----	100.00	0.0	LOW	0.00	0.00
14	CP	100.00	0.00	-----	100.00	0.0	LOW	0.00	0.00
15	CP	100.00	0.00	-----	100.00	0.0	LOW	0.00	0.00
16	CP	100.00	0.00	-----	100.00	0.0	LOW	0.00	0.00
17	CP	100.00	0.00	-----	100.00	0.0	LOW	0.00	0.00
18	CP	100.00	8.81	46.39	76.66	0.0	LOW	0.00	0.00
19	CP	100.00	0.00	-----	100.00	0.0	LOW	0.00	0.00
1A	CP	100.00	0.00	-----	100.00	0.0	LOW	0.00	0.00
1B	CP	100.00	0.00	-----	100.00	0.0	LOW	0.00	0.00
1C	CP	100.00	0.00	-----	100.00	0.0	LOW	0.00	0.00
TOTAL/AVERAGE			33.37	62.25		1450		35779	15.49

Agenda

IBM z13 Support

z/OS V2.2 planned enhancements
- **Support for JES concurrent job execution**
- **API to retrieve IEAOPT keywords and values**
- **Health based routing enhancements**
- Global Mirror (XRC) exploitation of I/O Priority Manager support
- **WLM-managed DB2 bufferpools enhancements**
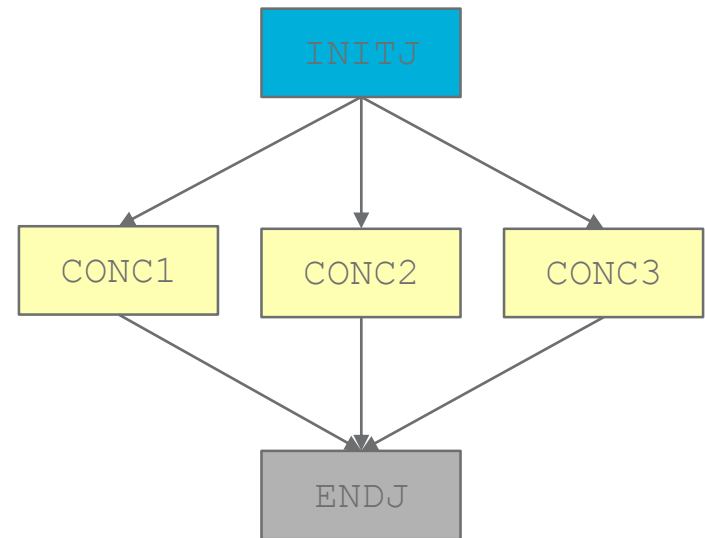- **SRM enhancements for large real storage**

z/OS V2.1 highlights

Other service stream enhancements and recommendations

Dependent Job Control for JES2

- JES2 in z/OS V2.2 provides a new job scheduling scheme similar to "JES3' Dependent Job Control" which in turn allows for a set of concurrent jobs to be run

```
//JGROUP   JOBGROUP
//INITJ    GJOB
//CONC1    GJOB
//         AFTER NAME=INITJ
//CONC2    GJOB
//CONC3    GJOB
//         CONCURRENT NAME=(CONC1,CONC2)
//ENDJ     GJOB
//         AFTER NAME=CONC1,WHEN=(RC=0)
//JGROUP   ENDGROUP
//INITJ    JOB TIME=NOLIMIT,REGION=0K,MSGCLASS=H,CLASS=J
//         SCHEDULE JOBGROUP=JGROUP
//STEP1    EXEC  PGM=...
//CONC1    JOB TIME=NOLIMIT,REGION=0K,MSGCLASS=H,CLASS=J
//         SCHEDULE JOBGROUP=JGROUP
//STEP1    EXEC  PGM=...
...
```

Dependent Job Control for JES2

- JES2 in z/OS V2.2 provides a new job scheduling scheme similar to "JES3' Dependent Job Control" which in turn allows for a *set of concurrent jobs* to be run

- For a set of concurrent jobs, WLM extends the *demand batch initiator* interface with JES2:

  - WLM returns the most eligible system for starting the demand batch initiators, or indicates that all candidate systems are too constrained

  - If a system is eligible, then
    - WLM reuses drained initiators, or
    - starts demand batch initiators.
    Both select the concurrent jobs specified by JES2

  - When the jobs are finished, both the reused and the newly started initiators go to the drained state

* Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.

IWM4HLTH: Extensions for health based routing

- WLM Sysplex routing services provide advice for routing work within a Sysplex
  - Enable distributed client/server environments to balance work among multiple servers based, on capacity, performance, **server health**
  - Utilized e.g. by Sysplex distributor (SERVERWLM), DB2 DDF

- The IWM4HLTH service allows to modify the health value when the health status of the server changes for the worse or better

- Before z/OS V2.2 the server health value solely is based on self-assessment with only the last value reported is being kept by WLM

- With z/OS V2.2 the IWM4HLTH service is extended to work with *multiple* components providing their views of the health of a server address space.
  The new IWM4QLTH service allows to query the health.

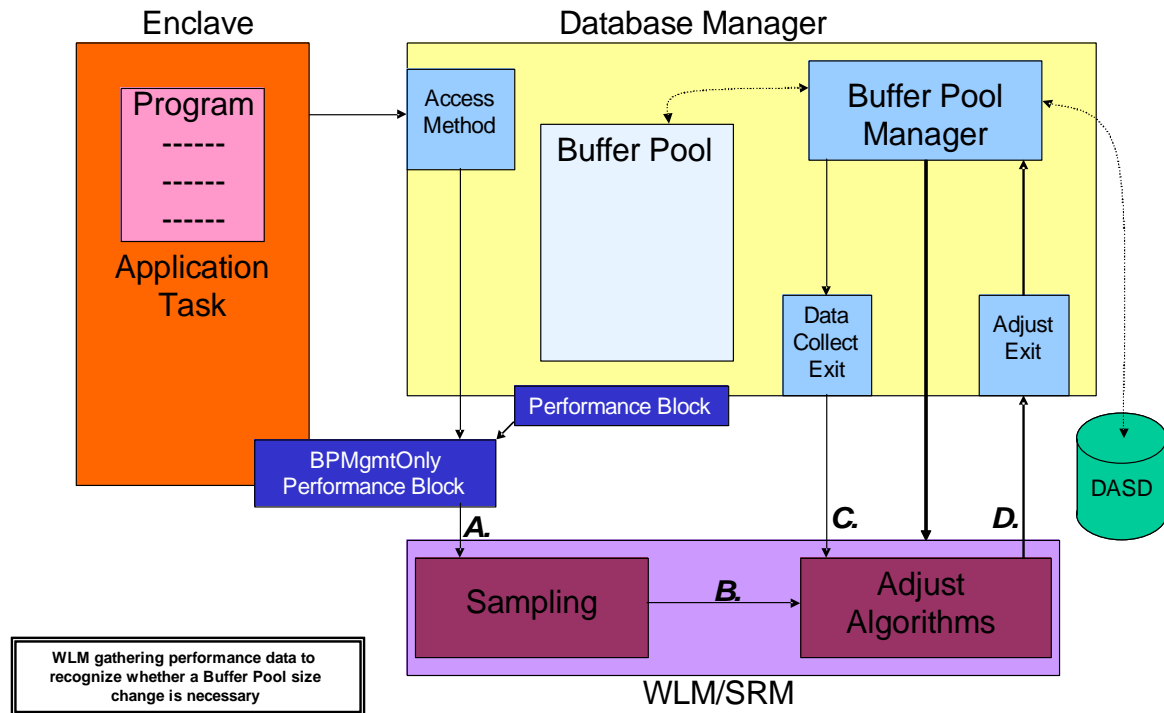# IWM4HLTH: Extensions for health based routing

```
                                    +--,FUNCTION=SET----+
                                    |                   |
>>-+------+---IWM4HLTH---STOKEN=stoken--+------------------+---,HEALTH=health--------->
   |      |                         |                   |
   +-name-+                         +--,FUNCTION=RESET--+


  +--,SUBSYS=NO_SUBSYS--+    +--,SUBSYSNM=NO_SUBSYSNM--+   +--,HEALTHRSN=NO_RSN-----+
  |                     |    |                         |   |                        |
>-+---------------------+----+-------------------------+---+------------------------+-->
  |                     |    |                         |   |                        |
  +--,SUBSYS=subsys-----+    +--,SUBSYSNM=subsysnm-----+   +--,HEALTHRSN=healthrsn--+


>-+--------------------+---+--------------------+-------------------------------------->
  |                    |   |                    |
  +--,RETCODE=retcode--+   +--,RSNCODE=rsncode--+


  +--,PLISTVER=IMPLIED_VERSION--+    +--,MF=S----------------------------------+
  |                             |    |                                         |
>-+-----------------------------+----+-----------------------------------------+---------><
  |                             |    |                                         |
  +--,PLISTVER=MAX--------------+    |                          +-,0D---+      |
  |                             |    +--,MF=(L-,list addr-+------+-)------+
  +--,PLISTVER=0----------------+    |                    +-,attr-+       |
  |                             |    |                                    |
  +--,PLISTVER=1----------------+    |                          +-,COMPLETE-+  |
                                     +--,MF=(E-,list addr-+----------+-)--+
```

* Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.

Runtime Diagnostics of Health Values

- z/OS Runtime Diagnostic display & analyze all server health values <100 through F HZR,ANALYZE command:

```
-----------------------------------------------------------------
EVENT 17: HIGH - SERVERHEALTH - SYSTEM: SYS1     2016/04/19 - 08:00:30
JOB NAME: DB1XDIST  ASID: 01CC  CURRENT HEALTH VALUE: 75
CURRENT LOWEST HEALTH VALUES:
               SUBSYSTEM    HEALTH                    REPORTED
 SUBSYSTEM    NAME         SETTING          REASON   DATE AND TIME
 DB1TDIST                    75                       2016/04/19 06:01:04
  ERROR: ADDRESS SPACE SERVER CURRENT HEALTH VALUE LESS THAN 100.
  ERROR: THIS VALUE MAY IMPACT YOUR SYSTEM OR SYSPLEX TRANSACTION
  ERROR: PROCESSING.
 ACTION: USE YOUR SOFTWARE MONITORS TO INVESTIGATE THE ASID AND TO
 ACTION: DETERMINE THE IMPACT OF THE HEALTH OF THE ADDRESS SPACE TO
 ACTION: OVERALL TRANSACTION PROCESSING.
-----------------------------------------------------------------
```

WLM-managed DB2 Bufferpools: Overall flow

- DB2 registers bufferpool with WLM

- WLM will recommend to grow the size of the bufferpool when the Performance Index of a Service Class Period is impacted and bufferpool delays are a significant contributor

- WLM will recommend to shrink the size of the bufferpool due to donation to a suffering Service Class period, or due to regular housekeeping cycles

- DB2 de-registers bufferpool from WLM management

- DB2 10: ALTER BUFFERPOOL [VPSIZE(s)] AUTOSIZE(YES)
  - ➔ MIN size = 0.75 x VPSIZE
  - ➔ MAX size = 1.25 x VPSIZE

- DB2 11: MIN/MAX can be specified

- Initial USED size between MIN size and MAX size

- Management range between MIN and MAX sizes

WLM-Managed DB2 Bufferpool: Changes in z/OS V2.2 plus V2.1

A bufferpool can be increased when

- Performance index impacted and buffer pool delays are a significant contributor

A bufferpool may shrink…

- Due to donation to a suffering service class period
  - May suffer storage related delays

- Due to regular housekeeping cycles
  - Consider one BP reduction candidate per 10 sec interval
    - BP idle - had no references
    - No delays, i.e. 100% hit ratio
    - Least important period showing buffer pool delays
  Any bufferpool may shrink no more than once per 5 min

- When WLM recommends to increase the size of a bufferpool, DB2 accepts the recommended size as new current VPSIZE. DB2 does not necessarily use the entire recommended size ➔ Used size of the bufferpool will be less or equal the current VPSIZE

Historic observations

Hit Ratio

min Buffer Pool Size max

Bufferpool
BP: BP11   ,0100002000000002,AS: DBX2DBM1,SC: $SRMS059

Legend: CurSize, TarSize, MinSize, MaxSize, UseSize

SRM Enhancements for large real storage

Function	z/OS release	V2.2	V2.1	V1.13
MCCFXPTR Limit		+	OA44668 OA44207(RSM)	
New AUTO keyword for RCCFXET and RCCFXTT		+		

Service Stream Enhancement:
OA44668: SRM – New Function

1. On LPARs with large real storage, lock contention may be seen in SRM and RSM when SRM calls RSM to determine frame counts.

2. The MCCFXTPR keyword in the IEAOPTxx specifies the percentage of online storage that may be page fixed before a **pageable storage shortage** is detected and message IRA400E is issued.
   - Before OA44668, MCCFXTPR default of 80% requires that 20% (100 minus MCCFXTPR) of storage remain pageable, regardless of the amount of online storage. On systems with large amounts of central storage, the MCCFXTPR default of 80% can result in a pageable storage shortage being detected when there is still plenty of pageable storage.
   - With OA44688 at most 64GB of pageable online storage will be required before a pageable storage shortage is recognized.

800G   1TB
80%

6
4
G

New AUTO keyword for RCCFXETand RCCFXTT

- The IEAOPTxx *RCCFXTT* keyword specifies low and high threshold of fixed real storage:
  - SRM uses these thresholds to determine if the **system MPL** needs to be increased/decreased. The default is 66% and 72%.
  - On small systems such percentages are not a problem.
  - On a 1TB LPAR these percentages imply that WLM will stop increasing the MPL. when 660G of storage is fixed

- Similarly, *RCCFXET* specifies the low and high thresholds of fixed real storage below 16M. SRM uses these thresholds to determine if the **system MPL** needs to be increased/decreased. The default is 82% and 88%.
  - This OPT keyword is also enhanced, mainly for consistency with the RCCFXTT keyword. The default is still: RCCFXET=(82, 88)

- Both keywords were enhanced to accept a value of **AUTO**
  - AUTO allows SRM to compute thresholds based on available storage.
  - Allows to higher utilize available storage in large systems without risking system shortages
  - AUTO needs to be specified in IEAOPTxx (not default)

New AUTO keyword for RCCFXET and RCCFXTT

Current RCCFXTT thresholds:

660G 66%   720G 72%   1TB

Current RCCFXET thresholds:

13,1M 82%   14.4M 88%   16M

Agenda

IBM z13 Support

z/OS V2.2 enhancements

z/OS V2.1 highlights

      zEC12 GA2 Support

      New Classification Qualifiers and Groups

      I/O Priority Groups

Other service stream enhancements and recommendations
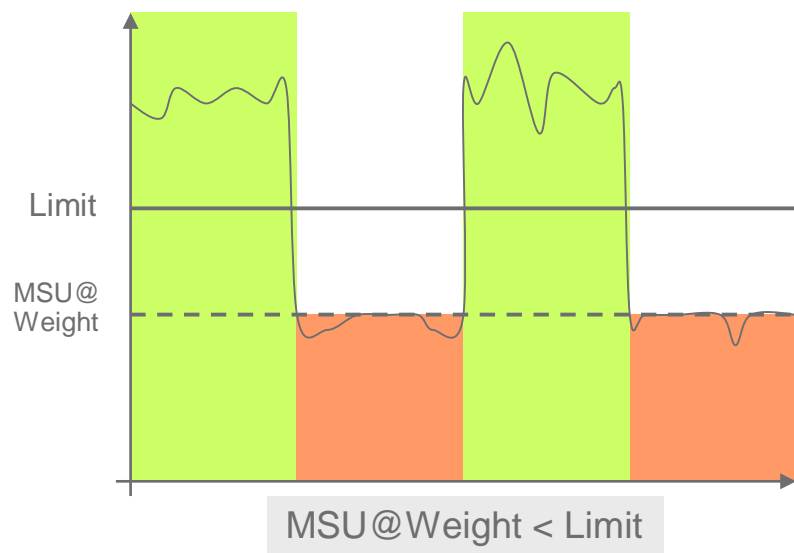
IBM zEnterprise EC12 GA2 Support Overview

- zEnterprise BC12 and EC12 (zEC12) GA2 (firmware driver 15F) offer new functions for hard and soft capping:
  - Smoother capping with WLM managed softcapping
  - When IRD weight management is active the group capacity of an LPAR may be derived by the initial weight
  - New "Absolute Capping Limit" LPAR control

z/OS release Function	V2.1	V1.13	V1.12
Smoother capping	**+**		
Group capacity to use initial weight	**+**	*OA41125*	*OA41125*
Absolute capping	**+**	*OA41125*	*OA41125*

Capping algorithms for defined capacity prior to zEC12 GA2

Pattern capping

- Must be used when *MSU @LPARweight* < definedLimit

- Periods with LPAR capped at weight and running uncapped

- Can result in "pulsing" potentially impacting online workloads

Phantom weight capping

- Is used when *MSU @LPARweight* ≥ definedLimit

- PR/SM uses an additional "phantom weight" to limit LPAR consumption below weight
  - Phantom weight must be non-negative pre-zEC12 GA2

- Results in smooth capping

MSU@Weight < Limit

MSU@Weight > Limit

zEC12 GA2 Negative Phantom Weight

- zEC12 GA2 allows using a *negative* phantom weight
  for soft capping

- Therefore, when *MSU @LPARweight* < definedLimit
  WLM can now use a negative phantom weight instead of pattern capping
  - I.e., phantom weight capping becomes the only mechanism

- z/OS V2.1 will exploit this feature
  - Eliminates pulsing effects caused by cap patterns

With IRD, zEC12 GA2 can use initial weight for group capping

- It is possible to combine Intelligence Resource Director weight management with capacity groups
  - IRD changes the –current- weight in order to shift capacity within an LPAR cluster

- On zEC12 GA2 the **initial** LPAR weight will be used for group capacity
  - Only if **all** systems in a capacity group run
    - z/OS V2.1, or
    - z/OS V1.12, V1.13 with OA41125 applied.
  - Results in more predictive and better controllable group capacity entitlement

zEC12 GA2 Absolute Capping Limit

- zEC12 GA2 allows to define an "absolute capping limit"
  - Primarily intended for non z/OS images
  - Expressed in terms of 1/100ths of a processor
  - Therefore, it is insensitive to LPAR (de)activations and less sensitive to capacity changes
  - Can be specified independently from the LPAR weight
  - Can be specified per processor type in image profile and partition controls panel

- Unlike initial capping it may be used *concurrently* with defined capacity and/or group capacity management

  - The minimum of all specified limits will be used
  - WLM/SRM recognizes new cap, e.g. for routing decisions.
  - RCTIMGWU = MIN(absolute cap, defined capacity, group cap) when all capping types are in effect
    - RMF provides RCTIMGWU in SMF70WLA
    - In addition, SMF70HW_Cap_Limit value in hundredths of CPUs

Agenda

IBM z13 Support

z/OS V2.2 enhancements

z/OS V2.1 highlights

Other service stream enhancements and recommendations

Recent changes for DB2 stored procedures and IDAA environments

- WLM OA43538 (z/OS V1.12, z/OS V1.13, V2.1): "Unbound Servers"
  Server address spaces, such as for DB2 Application Environments were not started due to incorrect assessment of available capacity
    - Symptom could be DB2 stored procedure timeouts with SQLCODE -471
  - Could occur even when minimum number of servers were requested via MNSPAS=n parameter

- WLM OA45658 - DB2 Stored Procedure Timeouts due to capped dependent enclave (triggered by Discretionary Goal Management)

- WLM OA45716 - When the CEC is less than 90% busy, this algorithm ignores the capping status of the system and therefore tends to overestimate the available CPU capacity of capped systems.

- WLM OA49442 - WLM may not start additional server address spaces for a service class with combined general purpose processor and specialty processor demand even though there are available resources

- OA50291 option to eliminate ICH7000I messages for WLM Application Environment Servers

OA50291 – New option to eliminate ICH7000I messages for WLM Application Environment Servers

- With Security=User, use of WLM application environments can result in a huge number of ICH70001I xxxxxx LAST ACCESS AT 10:38:32 ON THURSDAY, NOVEMBER 15, 2015 messages
  - Fill up logs
  - RACF data base updates

- WLM APAR OA50291 allows for requesting only one daily RACF message

- To enable daily logon statistics it is required to
  - Update security (RACF) definitions
  - Specify new IEAOPTxx keyword: SuppSAFinfoMsg=YES

IEAOPTxx SuppSAFinfoMsg =	
	Subject to change. Refer to the APAR documentation for implementation.
NO	WLM performs a standard verification of the application environment server address spaces through the security product.
YES	WLM performs a verification of the application environment server address spaces through the security product and passes an APPL class profile with a name equal to the procedure name associated with the application environment. SAF informational messages, such as RACF message ICH70001I will be suppressed.

OA50291 – Option to eliminate ICH7000I messages - Sample security definitions-

- The following security definitions allow to reduce the impact of recording logon statistics by recording for only the first daily logon by each user, rather than for every logon by each user:
  - Define a RACF APPL profile
    RDEFINE APPL *applname* UACC(NONE) APPLDATA('RACF-INITSTATS(DAILY)')
    where *applname* is the name of the JCL procedure name used for the WLM application environment
    - Ensure that no such APPL already exists
    - Generic profiles may be used
  - If UACC(NONE) was specified on the APPL, permit users READ access to the APPL profile
    - PERMIT *applname* CLASS(APPL) ID(*userid*) ACCESS(READ)
  - Activate or refresh the APPL profile:
    - if not yet active, issue
      SETROPTS CLASSACT(APPL) RACLIST(APPL)
    - SETROPTS RACLIST(APPL) REFRESH

- For more details refer to Reducing application logon statistics (https://ibm.biz/Bd4S3s)

Service Stream Enhancements for more aggressive Blocked Workload support (OA44526)

- Problem addressed:
  - The current minimum value that can be specified for the Blocked Workload interval threshold BLWLINTHD is 5 sec.
    DB2 could profit from earlier or more frequent trickling.

- More aggressive specifications will be enabled by OA44526

  - New lower limit is 1 sec

  - BLWLINTHD default and BLWLTRPCT remain unchanged
    - Consider lowering BLWLTRPCT with very small BLWLINTHD values if amount of trickle cycles that may be handed out is a concern.

XML Format WLM service definitions recommended

- For several releases WLM has supported service definitions in **XML format**
  - z/OSMF WLM task
  - ISPF Administrative Application: "Save as XML"…

- XML format avoids particular problems with the ISPF tables format, namely coexistence behavior, when a new functionality level needs to be introduced, and the number of table columns needs to be extended.

- **Recommendation**:
  Consider using the XML-format for your WLM service definition data sets.

z/OS Workload Management
- More Information -

- z/OS WLM Homepage:

  http://www.ibm.com/systems/z/os/zos/features/wlm/
  – Inside WLM: https://ibm.biz/BdF4L4

- z/OS MVS documentation
  – z/OS MVS Planning: Workload Management:
    http://publibz.boulder.ibm.com/epubs/pdf/iea3w101.pdf
  – z/OS MVS Programming: Workload Management Services:
    http://publibz.boulder.ibm.com/epubs/pdf/iea3w201.pdf

- *IBM Redbooks publication*s:
  – System Programmer's Guide to: Workload Manager:
    http://publib-b.boulder.ibm.com/abstracts/sg246472.html?Open
  – ABCs of z/OS System Programming Volume 12
    http://publib-b.boulder.ibm.com/abstracts/sg247621.html?Open

WLM Topology Report Tool (As-is)

- New *as-is* tool available for download from the WLM homepage
  - https://ibm.biz/BdE74v

- Visualizes mapping of HiperDispatch affinity nodes to physical structure

- Supports IBM zEC10 and later

- To use:
  1. Download from above location
  2. Run installer
  3. Upload Host code to a z/OS system
  4. Collect SMF99.14 records

Sample output (zEC12):

Topology for 07-21-2014-13:44:27 ,System: IRD9

Book_2

Chip_1		Chip_3	Chip_4	Chip_5	Chip_6
IRD9_01_MCPU002		IRD9_01_LCPU004	IRD9_01_LCPU007	IRD9_01_LCPU006	IRD9_01_HCPU000
IRD9_01_MCPU003		IRD9_01_LCPU005	IRD9_01_LCPU010	IRD9_01_LCPU007	IRD9_01_HCPU001
IRD9_02_MAAP008		IRD9_01_LCPU006	IRD9_01_LCPU011	IRD9_01_LCPU010	IRD9_01_HCPU000
IRD9_03_MIIP009			IRD9_01_LCPU003		IRD9_01_LCPU011
IRD9_01_MCPU001			IRD9_01_LCPU004		
IRD9_01_LCPU002			IRD9_01_LCPU005		
IRD9_02_MAAP008					
IRD9_03_MIIP009					

Background: Capping algorithm with negative phantom weight (zEC12 GA2 and later)

The phantom weight instructs PR/SM at what capacity an LPAR needs to be capped.

- A positive phantom weight also lowers the priority of a partition,

- A negative phantom weight caps the partition at a higher defined capacity without changing the priority of the partition.