SHARE in Orlando 2015

Workload Management (WLM) Update for z13, z/OS 2.1 and 2.2

Andreas Henicke (andreas.henicke@de.ibm.com)
IBM Corporation

Wednesday, August 12, 2015
Session 17637

#SHAREorg

SHARE is an independent volunteer-run information technology association that provides **education**, professional **networking** and industry **influence**.

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

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

IBM z Systems

The innovation continues

8

Agenda

IBM z13 Support

- **z13 base support**
- **zIIP SMT support**
- **HiperDispatch and capping enhancements**
- **SAN Fabric I/O priority**

z/OS V2.2 enhancements

z/OS V2.1 highlights

Other service stream enhancements and recommendations

IBM z13 Service

Do not use APAR numbers from this presentation for planning z13 service installation. Refer to the official fix categories:

- IBM.Device.Server.z13-2964.RequiredService
- IBM.Device.Server.z13-2964.Exploitation
- IBM.Device.Server.z13-2964.RecommendedService
- IBM.Device.Server.z13-2964.ParallelSysplexInfiniBandCoupling
- IBM.Device.Server.z13-2964.ServerTimeProtocol
- IBM.Device.Server.z13-2964.UnifiedResourceManager
- IBM.Device.Server.z13-2964.zHighPerformanceFICON
- IBM.Function.zEDC
- IBM.Device.Server.zBX-2458
- IBM.DB2.AnalyticsAccelerator.V2R1

10

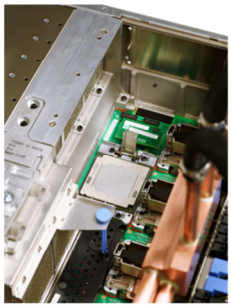
WLM/SRM support overview for IBM z13

Function \ z/OS release	V2.2	V2.1	V1.13
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	+	OA44504 OA46396	OA44504
SAN Fabric I/O Priority (Availability planned for 25 September 2015)	+	OA44431 OA44529	OA44431 OA44529

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

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

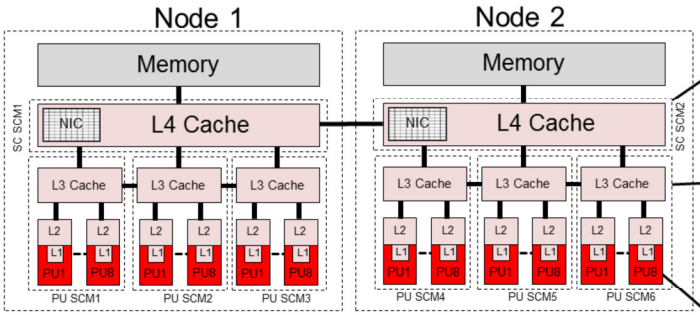


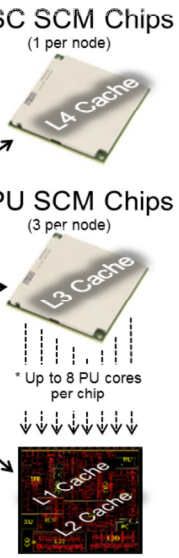
- New Cache topology
  - Chip, node, drawer
  - No longer using “books”
  - z/OS HiperDispatch uses new topology information to place work topologically close – to maximize cache efficiency

12

z13 CPC Drawer Cache Hierarchy Detail

Single CPC Drawer View (N30 Model) – 2 Nodes

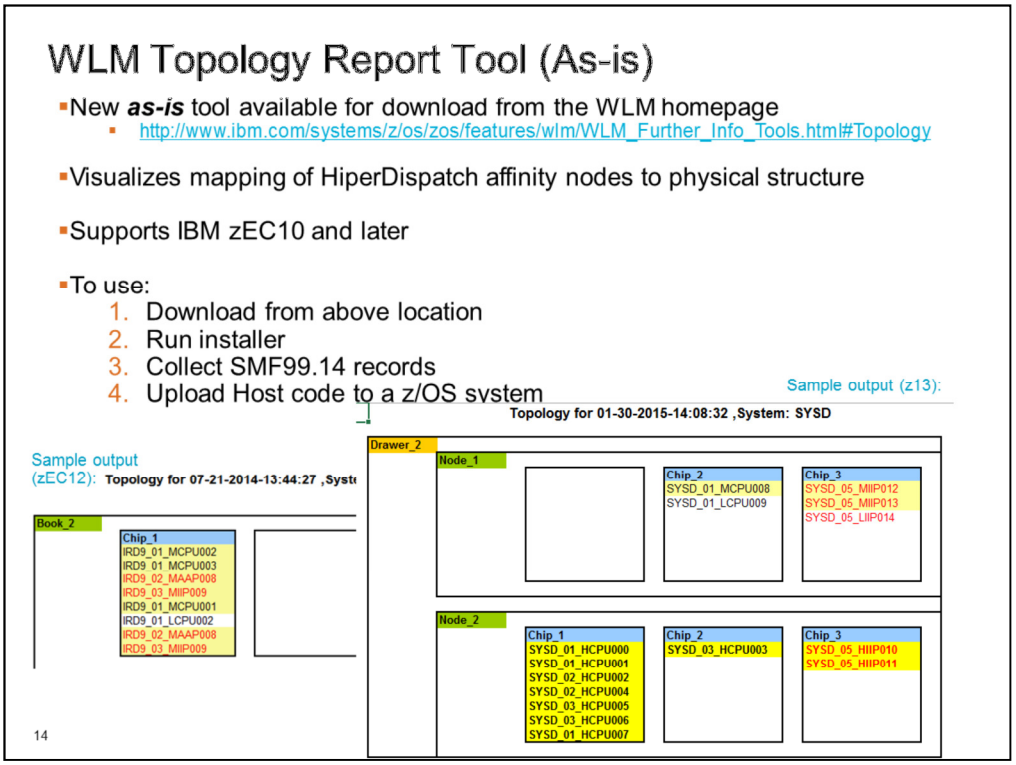




Node 1 - Caches
- L1 private 96k i, 128k d
- L2 private 2 MB i + 2 MB d
- L3 shared 64 MB / chip
- L4 shared 480 MB / node
  - plus 224 MB NIC

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

13



WLM Topology Report

The topology report displays the logical processor topology for systems running in Hiperdispatch mode. The Excel report on your workstation uses an input file (comma separated value) which must be first created on a z/OS system from SMF 99 subtype 14 records. The tool supports all System z environments from z10 to z13 for partitions running in Hiperdispatch mode. It displays the association of logical processors to books, chips, drawers, and nodes, the polarization of the processors (high, medium, low), the processor type (regular CP, zIIP, or zAAP), and the association to WLM nodes. The tool can be used to understand the processor placement and how it changes when topology changes occur.

In order to run the tool it is required to install the exe file from this webpage and afterwards two z/OS datasets on your local z/OS system. The install file creates two entries: "TopoReport.lnk" and "Topo Report Help.lnk" in the Windows program folder "IBM RMF Performance Management". Please select the "Topo Report Help" link and follow the instructions in topic "Processing SMF 99 data" to install and execute the z/OS datasets and programs. The other topics in the help file describe the usage of the Excel spreadsheet to display the information on your workstation.

Requirements: A z10 or newer System z environment with partitions running in Hiperdispatch mode

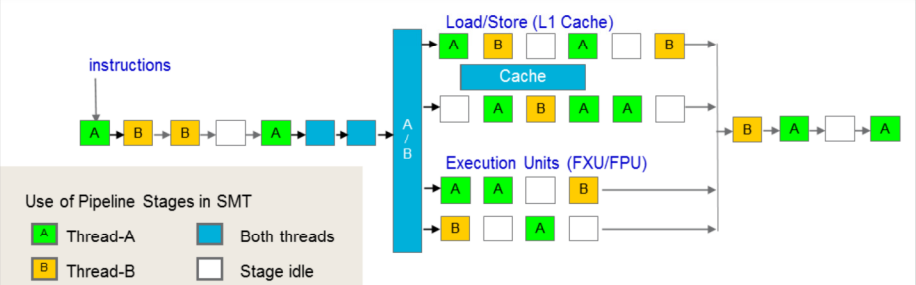
Collecting SMF 99 subtype 14 records

Excel Version 2013. The spreadsheet should also work on Excel 2007 and 2010

Motivation for Simultaneous Multi Threading

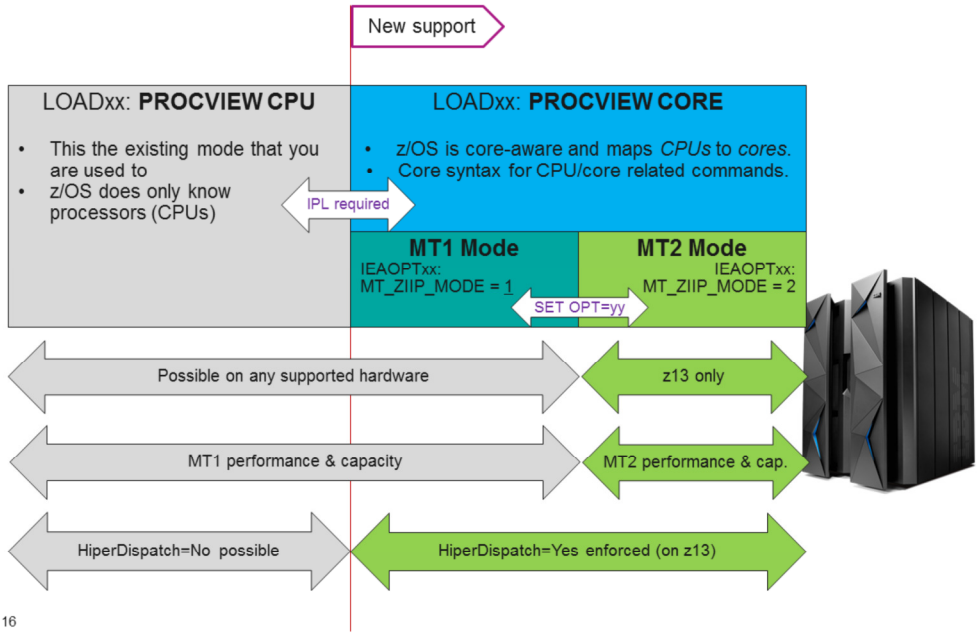
- "Simultaneous multithreading (SMT) permits multiple independent threads of execution to better utilize the resources provided by modern processor architectures."*
- With z13, SMT allows up to two instructions streams per core to run simultaneously to get better overall throughput
- SMT is designed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



15

What is new with multithreading support?



16

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…

17

…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

18

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

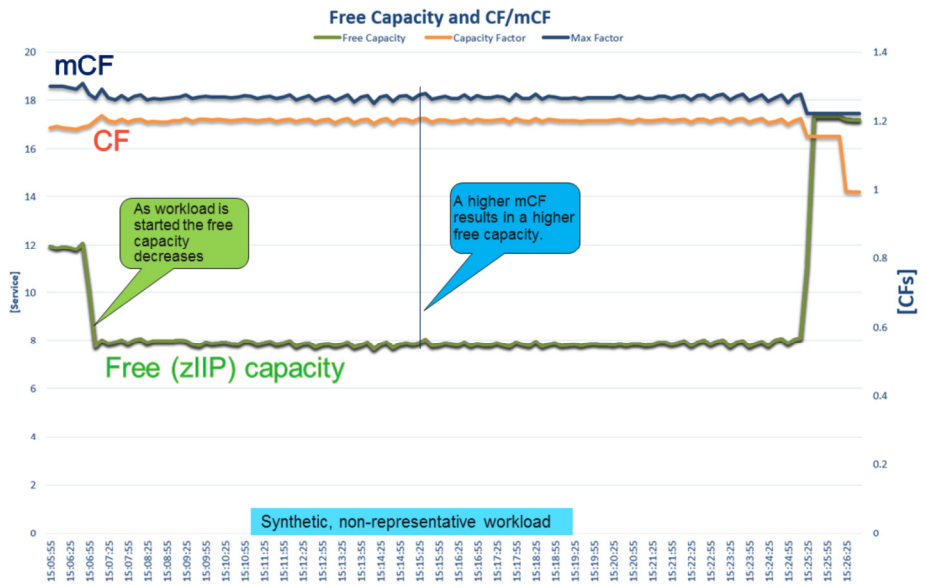
- **Capacity Factor (CF)**
  - How much work core actually completes for a given workload mix at current utilization - relative to single thread
  - Therefore, MT1 Capacity Factor is 1.0 (100%)
  - MT2 Capacity Factor is workload dependent
  - Describes the actual, current efficiency of MT2

- **Maximum Capacity Factor (mCF)**
  - How much work a core can complete for a given workload mix at most 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

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

Sample Capacity and maximum Capacity Factor



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

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

- z/OS SMT introduces several new metrics to describe how efficiently the core resources could be utilized and how efficiently they are actually utilized.

Transitioning into MT mode (Enablement)

- LOADxx PROCVIEW CORE enables use of SMT mode
  - IPL required to switch between PROCVIEW CPU and CORE
  - Causes syntax and semantic to change for core-aware commands.
    - -LOADxx CORE,CPU_OK allows using CPU as a synonym of CORE
  - HiperDispatch=YES enforced on SMT capable hardware

```
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
```

23

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

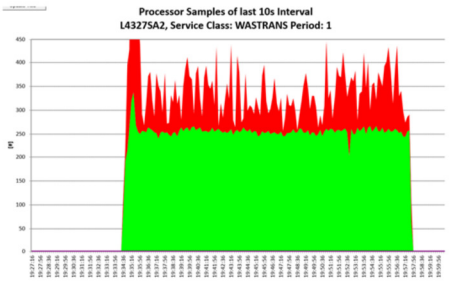
Transitioning into MT mode (Activation)

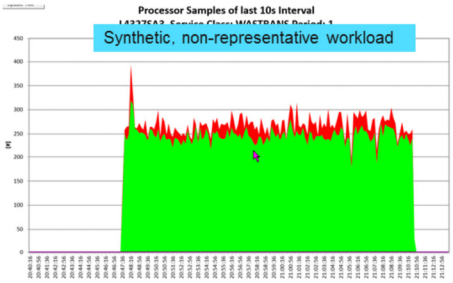
MT-2 mode Activation and Deactivation

- 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
```

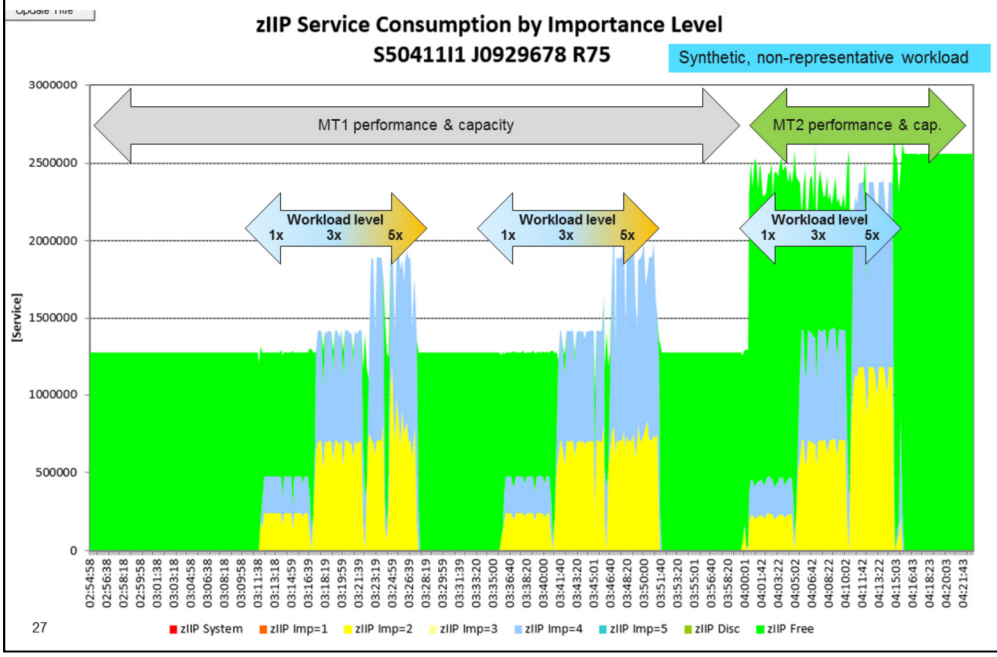
Processor samples may change when going to MT-2 mode





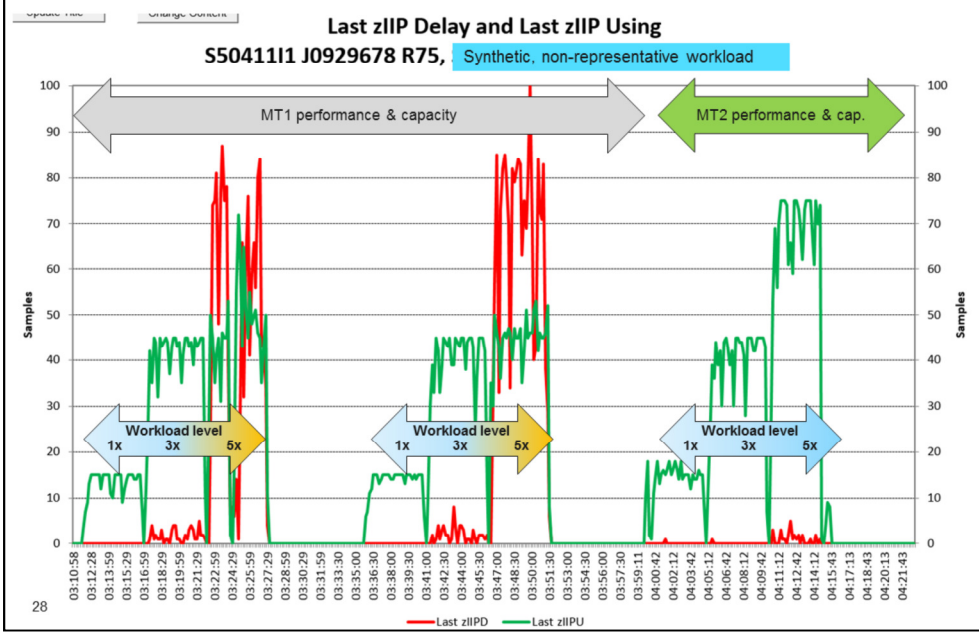
- In MT-2 mode we see less processor delays resulting in a higher execution velocity

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



27

Sample execution velocity MT-1 vs. MT-2



28

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 DECIMAL	OFFSET 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

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

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
z/OS V2R1                SYSTEM ID CB8B                DA          2/2015                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        1.289        1.976
```

The CPU Activity section reports on logical core and logical processor activity. For each processor, the report provides a set of calculations that are provided at a particular granularity that depends on whether multithreading is disabled (LOADxx PROCVIEW CPU parameter is in effect) or enabled (LOADxx PROCVIEW CORE parameter is in effect).

If multithreading is disabled for a processor type, all calculations are at logical processor granularity.

If multithreading is enabled for a processor type, some calculations are provided at logical core granularity and some are provided at logical processor (thread) granularity. The CPU Activity section displays exactly one report line per thread showing all calculations at logical processor granularity. Those calculations that are provided at core granularity are only shown in the same report line that shows the core id in the CPU NUM field and which is representing the first thread of a core.

The following calculations are on a per logical processor basis when multithreading is disabled and on a per logical core basis when multithreading is enabled

- Percentage of the interval time the processor was online
- LPAR view of the processor utilization (LPAR Busy time percentage)
- Percentage of a physical processor the logical processor is entitled to use
- Multithreading core productivity (only reported when multithreading is enabled)
- Multithreading core utilization (only reported when multithreading is enabled)

The following calculations are on a per logical processor basis regardless whether multithreading is enabled or disabled:

- MVS view of the processor utilization (MVS Busy time percentage)
- Percentage of the online time the processor was parked (in HiperDispatch mode only)
- I/O interrupts rate (general purpose processors only)

Percentage of I/O interrupts handled by the I/O supervisor without re-enabling (general purpose processors only)

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

RMF Monitor III CPC report displays performance data for all partitions belonging to the CPC

If multithreading is enabled the processor data is reported at logical core granularity, otherwise processor data is reported at logical processor granularity

The report header is enhanced with the information about MT Mode and Productivity for the zIIP processors.

Additional SMT metrics are available as hidden report header fields:

- Multi-Threading Maximum Capacity Factor for IIP
- Multi-Threading Capacity Factor for IIP
- Average Thread Density for IIP

These hidden report header fields can be displayed, if the CPC report is invoked in the RMF Data Portal for z/OS web browser frontend.

Transitioning into MT2 mode: WLM considerations (1)

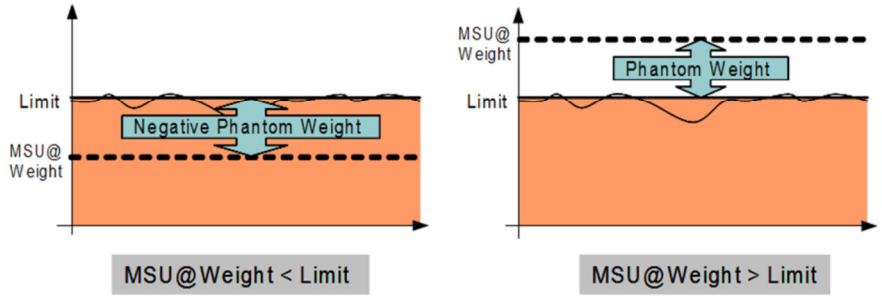
- **Less overflow from zIIP to CPs** may occur because
  - zIIP capacity increases, and
  - number of zIIP CPUs double
- CPU time and CPU service **variability may increase**, because
  - Threads which are running on a core at the same time influence each other
  - Threads may be dispatched at TD1 or 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.

34

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.

35

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 <u>if</u> the processors can be used efficiently.

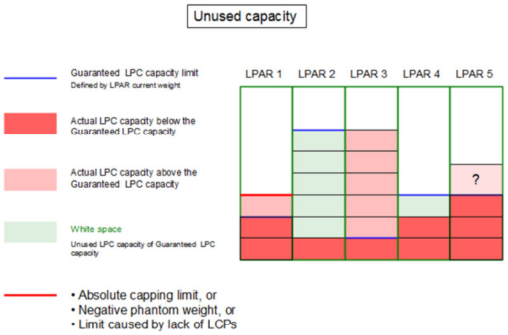
33

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

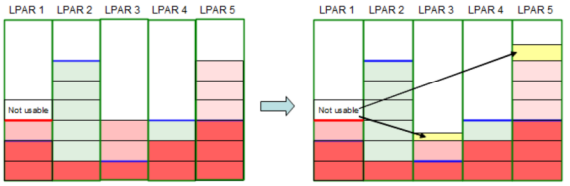
  of possible ‘unused capacity’ receivers

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.



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

(Planned availability 8/2015)

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

Prior to OA47968, VL processors will be unparked from the low to the high numbers,; and *also* parked from the low to the high numbers.

With this APAR, VLs will be parked from the high to the low numbers. On z13, efficiency can be improved because the lower numbers logical processors may share cache structures with the VH and VM processors.
Also, on z13 only, the unpark can occur a bit more restrictive. Unparking will stop earlier when the VLs can no longer be efficiently used.

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

This chart shows a “dangling” high CPU number being unparked. This will be mostly eliminated by OA47968.

Fabric I/O Priority

- z/OS V2.2 planned to support additional I/O priority capabilities
  - Like other I/O priorities already set by IOS and WLM
    - Control unit, Channel subsystem, Tape, or DS8000 I/O Priority Manager importance
  - Used today by channel subsystem and IBM System Storage DS8000 series for both read and write operations
  - Intended to provide end-to-end prioritization according to WLM policy for write operations
- Planned to be extended to provide additional prioritization data for the FICON fabric so that the highest priority write operations can be done first when the fabric becomes congested
- Will require:
  - z13 processor
  - z/OS V2.2; or, z/OS V1.13 or z/OS V2.1 with PTFs for APARs OA44529 and OA44431
- Availability planned for 25 September 2015; to be enabled by IOS
- See also Enhancing Value to Existing and Future Workloads with IBM z13

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

Storage Area Network (SAN) Fabric I/O Priority

This new function on the IBM z13 provides the ability for z/OS to specify an I/O priority for the SAN fabric to utilize. This capability allows z/OS to extend the z/OS Work Load Manager (WLM) to manage the SAN fabric, completing the management

of the entire end-to-end flow of an I/O operation. WLM will assign an I/O priority consistent with the client-specified goals for the workloads within the supported range of I/O priorities in the SAN fabric. SAN fabric I/O priority is especially useful in circumstances that can lead to SAN fabric contention such as workload spikes and

hardware failures to provide additional resilience and allow z/OS WLM to deliver the highest I/O priority to the most important work first.

SAN Fabric Priority on IBM DS8870: IBM will be the first platform to exploit this industry feature with a fully integrated workload management solution provided by z/OS and supported by DS8870. Intelligent access to data and greater efficiencies are reached with SAN Fabric I/O Priority enabled by DS8870. The DS8870 will also propagate the fabric priority for write operations to the resulting Metro Mirror traffic to provide a consistent prioritization with FICON when sharing the same SAN

infrastructure and Inter Switch Links (ISLs).

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

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

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
- 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

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

Today, the demand initiator interface is used with "$S JOB"

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 planned to be 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.

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

WLM enhances service IWM4HLTH (setting server health indicator)

WLM differentiates between health values of the address space reported by itself or reported by another space

The algorithm for determining the health indicator for an address space is changed. The value is no longer the last value being reported but the minimum of the values reported by the different callers.

An additional function of IWM4HLTH refers to RAS considerations regarding a server's health state. The RESET function restarts setting of a composite health value by specifying an initial value and discarding the values reported by other callers before.

Callers of service can specify a reason for cause of change

Callers can identify themselves by a subsystem type and subsystem name. WLM uses these parameters to recognize different callers of the service.

Users of the service need to check their programs for sufficient program authorization

WLM provides a new query service (IWM4QHLT) to obtain reported health indicators for diagnostic purposes

Callers of service can obtain health values for particular address spaces or

for all spaces for which a health value has been set

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.

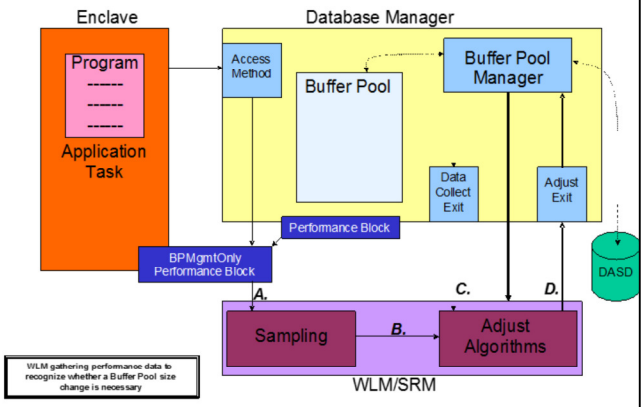
XRC Write Pacing

- z/OS Global Mirror (XRC) designed to work with…
  - z/OS WLM; and,
  - DS8000 with the z/OS Global Mirror feature

- …to throttle low-priority writes when they would cause significant delays that might affect response time
- Will be designed to allow you to specify that write delays be imposed for different classes of work based on WLM definitions
- Exploits WLM support for the DS8000 I/O Priority Manager
- Intended to:
  - Make it unnecessary to adjust write pacing settings and monitor data set residency
  - Improve system responsiveness to more important work
- Requires a DS8870 with an MCL
- Available now for z/OS V1.13 and z/OS V2.1 with the PTFs for APARs OA41906, OA44004, and OA43453

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

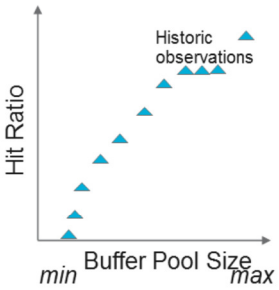


- ALTER BUFFERPOOL [VPSIZE(s)] AUTOSIZE(YES)
  - ➔ MIN size = 0.75 x VPSIZE
  - ➔ MAX size = 1.25 x VPSIZE
- Initial USED size between MIN size and MAX size
- Management range between MIN and MAX sizes

This sequence runs asynchronously from all (prior or following) WLM service invocations in the Database Manager

- WLM collects Performance Block states
- The collected states are reported to the Adjust Algorithms
- Periodically poll current Buffer Pool size
- If Buffer Pool delays are compelling, calculate and instruct the Buffer Pool Manager to adjust the Buffer Pool size

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…

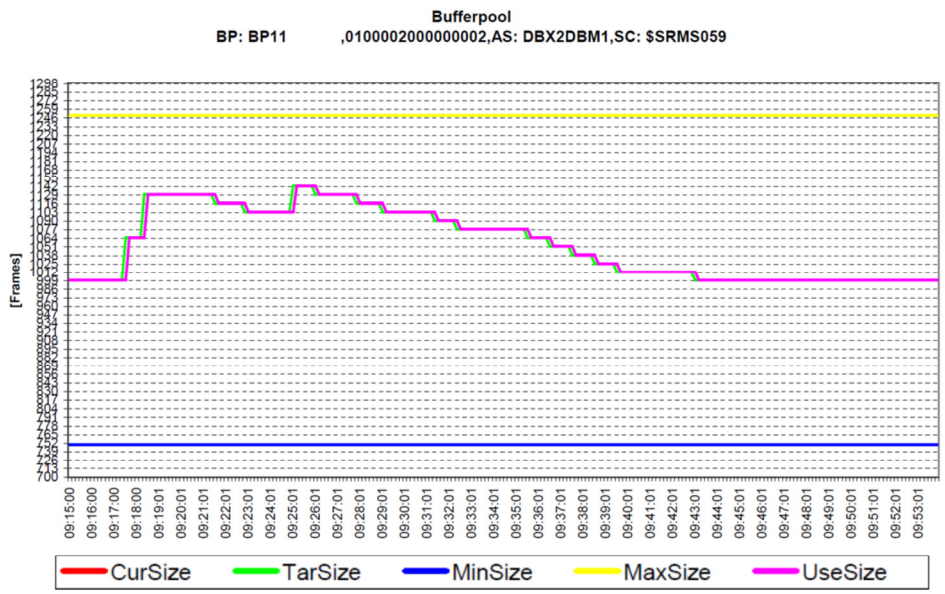
- Due to donation to a suffering service class period
  - May suffer storage related delays
- Due to regular housekeeping cycles
  - Consider one BP reduction candidate per 10 sec interval
    - BP idle - had no references
    - No delays, i.e. 100% hit ratio
    - Least important period showing buffer 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 <u>not</u> necessarily use the entire recommended size ➔ Used size of the bufferpool will be less or equal the current VPSIZE

53

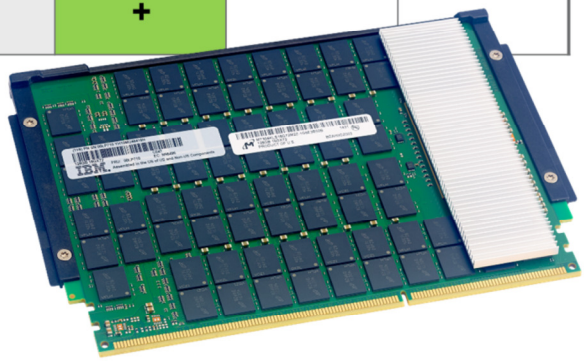
Row 1+2 40 sec


Bufferpool
BP: BP11 ,0100002000000002,AS: DBX2DBM1,SC: $SRMS059
[Frames]
CurSize
TarSize
MinSize
MaxSize
UseSize

54

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 RCCEXTT and RCCFXTT	+		



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

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.

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

This APAR addresses two areas:

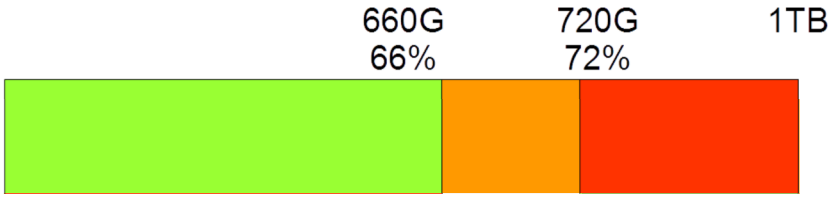
1. Lock contention seen when SRM call RSM IARXNCNT will be reduced
2. If 100% - MCCFXTPR * the total amount of online frames is greater than 64GB, the MCCFXTPR keyword will no longer be used in determining the threshold at which a shortage of pageable storage exists. Instead, on larger systems with more than 320GB of storage, a pageable storage shortage will be detected when less than 64GB of online storage is pageable. When calculating the number of frames that can be page fixed before a pageable storage shortage is detected, SRM now uses the maximum of MCCFXTPR * total online storage and total online storage minus 64GB.

New AUTO keyword for RCCEXTT and RCCFXTT

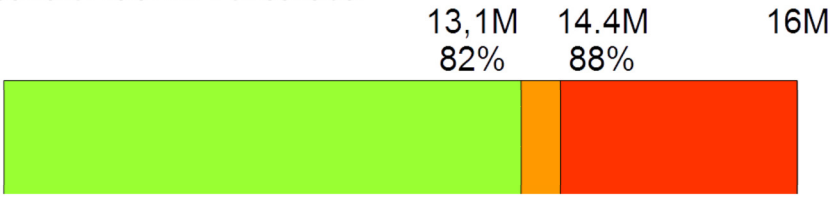
- The IEAOPTxx *RCCFXTT* keyword specifies low and high threshold of fixed real storage:
  - SRM uses these thresholds to determine if the **system MPL** needs to be increased/decreased. The default is 66% and 72%.
  - On small systems such percentages are not a problem.
  - On a 1TB LPAR these percentages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

Current RCCFXTT thresholds:



Current RCCEXTT thresholds:



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

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

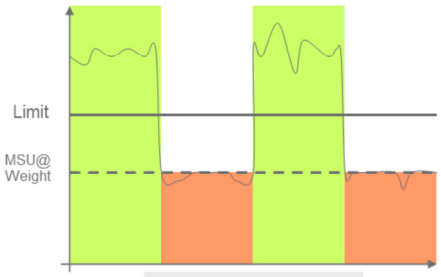
z/OS release / Function	V2.1	V1.13	V1.12
Smoother capping	+		
Group capacity to use initial weight	+	***OA41125***	***OA41125***
Absolute capping	+	***OA41125***	***OA41125***

63

Capping algorithms for defined capacity prior to zEC12 GA2

Pattern capping

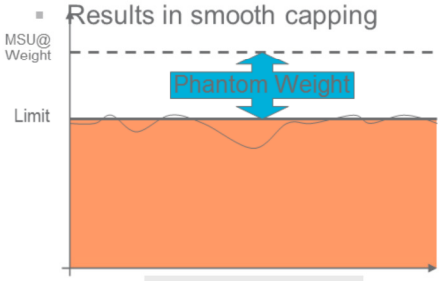
- Must be used when *MSU@LPARweight* < definedLimit
- Periods with LPAR capped at weight and running uncapped
- Can result in "pulsing" potentially impacting online workloads



MSU@Weight < Limit

Phantom weight capping

- Is used when *MSU@LPARweight* ≥ definedLimit
- Internally PR/SM uses an additional weight to limit LPAR consumption below weight
  - Phantom weight must be non-negative pre-zEC12 GA2
- Results in smooth capping



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
  - However, IRD weight management gets suspended when capping is in effect
  - Because entitlement of an LPAR within a capacity group is currently derived from the current weight the LPAR might get stuck at a low weight
    - Consequently, a low group capacity entitlement can result

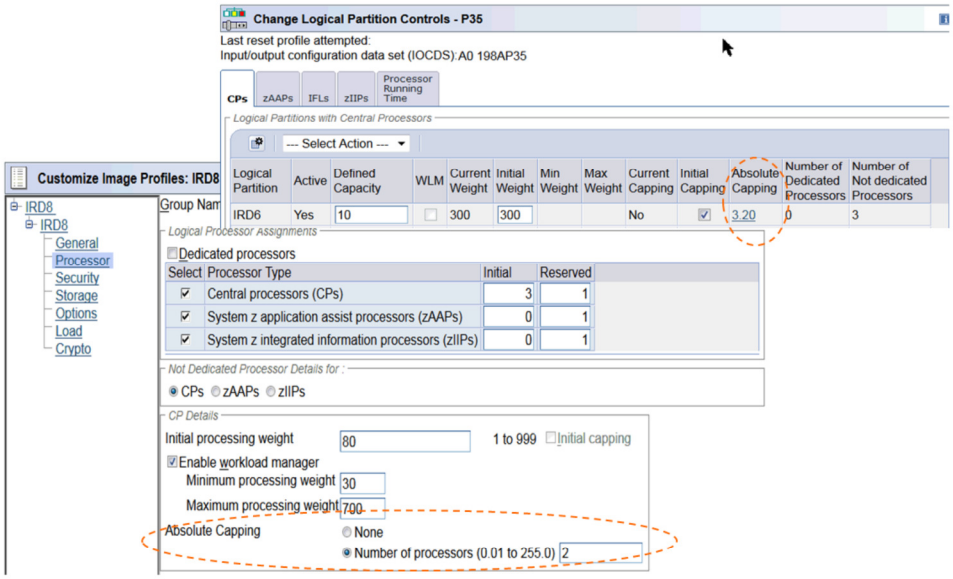
- On zEC12 GA2 the **initial** LPAR weight will be used for group capacity
  - Only if **all** systems in a capacity group run
    - z/OS V2.1, or
    - z/OS V1.12, V1.13 with OA41125 applied.
  - Results in more predictive and better controllable group capacity entitlement

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

zEC12 GA2 Absolute Capping Limit - Examples



New Classification Qualifiers and Groups: Overview

- With z/OS V2R1, WLM/SRM introduces
  - New types of classification groups, and
  - Some new and modified types of work qualifiers for use in classification rules in the WLM service definition
- Can be used to improve the structure of your WLM service definition when masking or wild-carding are not sufficient to simplify classification rules.
- New and modified qualifier types allow better classification of new DB2 and DDF workload
  ➔ For use with DB2 11
- More notepad information about a service definition allowed

New Classification Qualifiers and Groups

- z/OS V2.1 extends classification groups to all non-numeric work qualifier types.
- For long qualifier types, a start position for group members, and nesting is allowed.
- **New Groups:**
  - Accounting Information Group
  - Client Accounting Information Group
  - Client IP Address Group
  - Client Transaction Name Group
  - Client Userid Group
  - Client Workstation Name Group
  - Collection Name Group
  - Correlation Information Group
  - Procedure Name Group
  - Process Name Group
  - Scheduling Environment Group
  - Subsystem Collection Group
  - Subsystem Parameter Group
  - Sysplex Name Group

New Classification Qualifiers and Groups

- Subsystems (DB2) require longer and additional work qualifiers:
  - Work qualifier type “Package Name”: 128 characters (instead of 8)
  - Work qualifier type “Procedure Name”: 128 characters (instead of 18)

  - New work qualifier types:
    - Client Accounting Information (max. 512 characters)
    - Client IP Address (max. 39 characters)
    - Client Transaction Name (max. 255 characters)
    - Client User ID (max. 128 characters)
    - Client Workstation Name (max. 255 characters)

- The maximum number of “Notepad” lines the has been increased from 500 to 1000 lines

- Note: New and modified work qualifier types are only supported by the new 64-bit classify IWM4CLSY (planned to be used by DB2 V11).

These enhancements were implemented per requests from DB2.

WLM ISPF application enhancements

- Option 5 Classification Groups: Groups can be defined for all non-numeric work qualifier types.
  - Except: Priority (numeric), zEnterprise Service Class

```
 File  Utilities  Notes  Options  Help
--------------------------------------------------------------------------------
Functionality LEVEL029          Definition Menu          WLM Appl LEVEL029
Command ===>

Definition data set  . . : none

Definition name  . . . . . coeffs     (Required)
Description  . . . . . . . Service coefficients

Select one of the
following options. . . . . 5     1.   Policies
                          Classification Group Menu

 Select one of the following options.
 __   1.  Accounting Information Groups    14. Plan Name Groups
      2.  Client Accounting Info Groups    15. Procedure Name Groups
      3.  Client IP Address Groups         16. Process Name Groups
      4.  Client Transaction Name Groups   17. Scheduling Environment Groups
      5.  Client Userid Groups             18. Subsystem Collection Groups
      6.  Client Workstation Name Groups   19. Subsystem Instance Groups
      7.  Collection Name Groups           20. Subsystem Parameter Groups
      8.  Connection Type Groups           21. Sysplex Name Groups
      9.  Correlation Information Groups   22. System Name Groups
      10. LU Name Groups                   23. Transaction Class Groups
      11. Net ID Groups                    24. Transaction Name Groups
      12. Package Name Groups              25. Userid Groups
      13. Perform Groups
  F1=Help        F2=Split       F5=KeysHelp    F9=Swap       F12=Cancel
```

WLM ISPF application samples

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

Enter or change the following information:

Qualifier type . . . . . . . : Accounti
Group name . . . . . . . . . : SLOWACCT
Description  . . . . . . . . .
Fold qualifier names?  . . . . Y  (Y or

Qualifier Name   Start  Des
020175
030275
040375
```

Use to group work when there is no naming convention that allows for masking or wild-carding

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

Enter or change the following information:

Qualifier type . . . . . . . : Accounting Information
Group name . . . . . . . . . : FASTDEPT
Description  . . . . . . . . .
Fold qualifier names?  . . . . Y  (Y or N

Qualifier Name   Start  Description
PURCHASE         8
SALES            8
SHIPPING         8
ITDEP*           11
HRDEP*           11
```

Use a start position for each group member to indicate how far to index into the character string for a match. The start position may differ across group members.

73

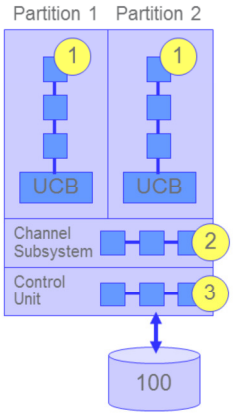
Coexistence and migration considerations for new classification qualifiers and groups

- If you plan to use more than 500 lines of notepad information, re-allocate the WLM couple data set on the z/OS V2R1 system before installing the service definition
  - By using z/OS V2.1 to allocate the WLM couple data set, the space allocated is sufficient for the increased notepad size
  - Else you may receive error message "WLM couple data set is too small to hold the service definition. (IWMAM047)"

z/OS release / Function	V2.1	V1.13 – V1.10
Groups of SPM rules & new classification qualifiers	+	**Toleration OA36842**

I/O Priority Groups

- **Rationale**
  - I/O Priority is used to control DASD I/O queuing.
  - WLM dynamically adjusts the I/O priority based on goal attainment and whether the device can contribute to achieve the goal.
  - Every 10 minutes, WLM determines which service classes use which devices and builds so called device sets.
  - Typically, different workloads use distinct device sets and WLM changes I/O priorities between service classes using the same device set.
  - If a workload starts to use a device outside from its previously used device sets and experiences significant I/O delay, it may take minutes until WLM refreshes the device sets and adapts the I/O priority of the corresponding service class.



- **Solution:**
  - Important service classes which are sensitive to I/O delay can now be assigned to I/O priority group HIGH which ensures that they get always higher I/O priorities than the service classes assigned to group NORMAL.

I/O priority group support is similar to what „CPU critical“ is for CPU management.

z/OS V2R1 allows to define I/O Priority Groups

z/OS V1R12 and z/OS V1R13 can expoit with **OA37824**

I/O Priority Groups Specification in WLM ISPF Application

z/OS release / Function	V2.1	V1.13	V1.12
I/O Priority Groups	+	Toleration OA37824	Toleration OA37824

I/O Priority Group is specified in the service class definition:

```
                    Create a Service Class

 Command ===> ____________________________________________________________
 Service Class Name . . . . . . ________   (Required)
 Description  . . . . . . . . . ________
 Workload Name  . . . . . . . . ________   (name or ?)
 Base Resource Group  . . . . . ________   (name or ?)
 Cpu Critical . . . . . . . . . NO_        (YES or NO)
 I/O Priority Group . . . . . . HIGH       (NORMAL or HIGH)
```

This indicates the I/O priority group of the service class. HIGH ensures that work in this service class always has higher I/O priority than work in service classes assigned to I/O priority group NORMAL. See the discussion of "Long-term I/O Protection" in the "Defining Special Protection Options for Critical Work" chapter of z/OS MVS Planning: Workload Management.

NORMAL and HIGH are the only valid groups. The default is NORMAL.

Group HIGH is only allowed if dynamic I/O priority management is enabled, that is, workload management dynamically manages your I/O priorities based on service class goals and importance. To turn on I/O priority management, specify YES on the Service Coefficients/Options panel.

I/O Priority Groups – Validation

But I/O Priority Group HIGH is only honored by WLM if both "I/O priority management" and "I/O priority groups" are enabled for the service definition:

```
Service Coefficient/Service Definition Options

I/O priority management  . . . . . . . . YES  (Yes or No)
Enable I/O priority groups . . . . . . . YES  (Yes or No)
Dynamic alias tuning management  . . . . NO_  (Yes or No)
```

The "Validate definition" option can be used to check whether service classes assigned to I/O priority group HIGH although I/O priority management is not enabled

```
Service Definition Validation Results

IWMAM918W  Service class(es) assigned to I/O priority group HIGH but
_________  I/O priority management or I/O priority groups are not
_________  enabled. The I/O priority group will not be honored.
```

80

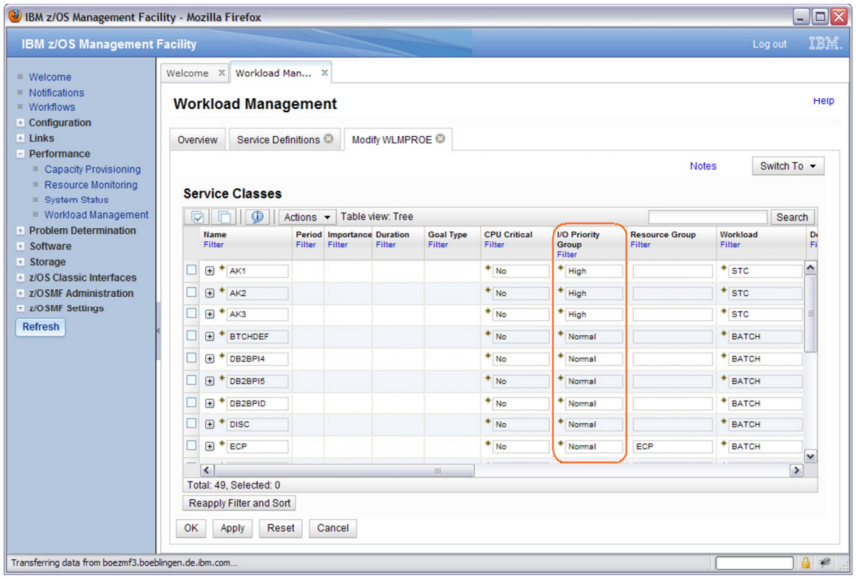
This indicates the I/O priority group of the service class. HIGH ensures that work in this service class always has higher I/O priority than work in service classes assigned to I/O priority group NORMAL. See the discussion of "Long-term I/O Protection" in the "Defining Special Protection Options for Critical Work" chapter of z/OS MVS Planning: Workload Management.

NORMAL and HIGH are the only valid groups. The default is NORMAL.

Group HIGH is only allowed if dynamic I/O priority management is enabled, that is, workload management dynamically manages your I/O priorities based on service class goals and importance. To turn on I/O priority management, specify YES on the Service Coefficients/Options panel.

I/O Priority Groups – Specification in z/OSMF

z/OSMF Workload Management task provides new option, too.



81

This indicates the I/O priority group of the service class. HIGH ensures that work in this service class always has higher I/O priority than work in service classes assigned to I/O priority group NORMAL. See the discussion of "Long-term I/O Protection" in the "Defining Special Protection Options for Critical Work" chapter of z/OS MVS Planning: Workload Management.

NORMAL and HIGH are the only valid groups. The default is NORMAL.

Group HIGH is only allowed if dynamic I/O priority management is enabled, that is, workload management dynamically manages your I/O priorities based on service class goals and importance. To turn on I/O priority management, specify YES on the Service Coefficients/Options panel.

I/O Priority Groups – RMF: Workload Activity Report

- Postprocessor Workload Activity (WLMGL) report shows I/O priority group
- If service class is assigned to I/O priority group HIGH, an indication is displayed in the SERVICE CLASS(ES) and SERVICE CLASS PERIODS sections.

```
-------------------------------------------------------------------------------- SERVICE CLASS(ES)

REPORT BY: POLICY=WLMPOL     WORKLOAD=ONLINE      SERVICE CLASS=ONLTOP     RESOURCE GROUP=*NONE
                                                CRITICAL     =CPU+STORAGE                          I/O PRIORITY GROUP=HIGH
                                                DESCRIPTION  =Batch Workload

-TRANSACTIONS-  TRANS-TIME HHH.MM.SS.TTT  --DASD I/O--  ---SERVICE---   SERVICE TIME  ---APPL %---  --PROMOTED--  ----STORAGE----
AVG       0.74  ACTUAL                0  SSCHRT   0.0  IOC         0   CPU    6.429  CP      0.66  BLK    0.000  AVG     7663.01
MPL       0.74  EXECUTION             0  RESP     0.0  CPU    287332   SRB    0.000  AAPCP   0.00  ENQ    0.000  TOTAL   5698.61
ENDED        0  QUEUED                0  CONN     0.0  MSO    537297   RCT    0.002  IIPCP   0.00  CRM    0.000  SHARED     0.00
```

This indicates the I/O priority group of the service class. HIGH ensures that work in this service class always has higher I/O priority than work in service classes assigned to I/O priority group NORMAL. See the discussion of "Long-term I/O Protection" in the "Defining Special Protection Options for Critical Work" chapter of z/OS MVS Planning: Workload Management.

NORMAL and HIGH are the only valid groups. The default is NORMAL.

Group HIGH is only allowed if dynamic I/O priority management is enabled, that is, workload management dynamically manages your I/O priorities based on service class goals and importance. To turn on I/O priority management, specify YES on the Service Coefficients/Options panel.

Use of I/O Priority Ranges

	I/O Priority Management=YES	
Priority	I/O PriorityGroups NOT enabled	I/O PriorityGroup enabled
FF	SYSTEM	SYSTEM
FE	SYSSTC	SYSSTC
FD	Dynamically managed	Priority Group = HIGH
FC		
FB		
FA		
F9		
F8		Priority Group = NORMAL
F7		
F6		
F5		
F4		
F3		
F2	Discretionary	Discretionary

I/O Priority Groups require some migration and coexistence considerations

- Toleration **APAR OA37824** required on z/OS V1R12 and z/OS V1R13 systems because dynamic I/O priority management is a sysplex-wide function
- Recommend to turn on I/O priorities only if all systems sharing disk systems run on z/OS V2R1 or on z/OS V1R12 / R13 with OA37824
- When the Enable I/O Priority Groups option is turned on in one sysplex, turn it also on in other sysplexes even if they do not exploit I/O priority group HIGH.
  - Ensures that all systems sharing a disk system work with an identical range of I/O priorities
- Assigning service classes to I/O priority group HIGH is only possible with the z/OS V2R1 WLM ISPF Application or z/OSMF V2R1
- If a service class is assigned to I/O priority group HIGH, the functionality level of the service definition is increased to **LEVEL029**
  - A service definition at functionality level 29 cannot be extracted, displayed, modified, installed, or activated by an WLM Application prior z/OS V2R1
- RMF support is only available with z/OS V2R1

Agenda

IBM z13 Support

z/OS V2.2 enhancements

z/OS V2.1 highlights

Other service stream enhancements and recommendations

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.

In addition, the structures of the following HD SMF 99 subtype records were published:

Subtype 12 record - HD interval data

Subtype 14 record - HD topology data

Service Stream Enhancements for reduced WLM Address Space utilization (OA48161 for z/OS V2.1)
Minimizing Sampling Overhead for Performance Block (PBs)

- In some environments a –relatively- higher percentage of WLM address space utilization may be observed

- Typically these are mostly idle test systems hosting many CICS or DB2 subsystems
  - In such cases the WLM task responsible for sampling *Performance Block*s (PBs) allocated by theses subsystems may incur the highest CPU cost

- Recommendations:

  - On z/OS releases up to z/OS V1.13 install PTF for OA38280
    - Eliminates an IVSK instruction that is more costly on IBM z196 and later hardware
  - **On z/OS V2.1 install PTF for OA48161**
    - **Eliminates an MVCSK instruction that was introduced with z/OS V2.1**

  - Reduce the number of PB control blocks allocated by the subsystems:
    - CICS: Max task (MXT)
    - DB2 MSTR+DBM1: 1000 + MAXDBAT
      + (n · 500) as needed

    Reducing the effective max task and MAXDBAT numbers can help to reduce the WLM address space consumption

91

IVSK = Insert Virtual Storage Key

Recent changes for DB2 stored procedures and IDAA environments

- DB2 PM90151
  - In the case where a stored procedure spawns a thread and the spawned thread calls another stored procedure, the inner stored procedure can exceed the STORTIME zparm.
  - With this APAR change, DB2 will use the DEPENDENT(YES) attribute when inserting the WLM request to schedule the stored procedure
  - Provided there are system resources available, WLM will give increased priority to this request. This should help prevent the sqlcode471 rc00E79002.

- 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.

XML Format WLM service definitions recommended

- For several releases WLM has supported to store a 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

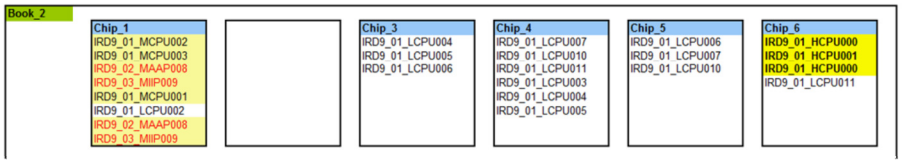
- z/OS MVS documentation
  - z/OS MVS Planning: Workload Management: http://publibz.boulder.ibm.com/epubs/pdf/iea3w101.pdf
  - z/OS MVS Programming: Workload Management Services: http://publibz.boulder.ibm.com/epubs/pdf/iea3w201.pdf
- *IBM Redbooks publications*:
  - System Programmer's Guide to: Workload 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



96

WLM Topology Report

The topology report displays the logical processor topology for systems running in Hiperdispatch mode. The Excel report on your workstation uses an input file (comma separated value) which must be first created on a z/OS system from SMF 99 subtype 14 records. The tool supports all System z environments from z10 to z13 for partitions running in Hiperdispatch mode. It displays the association of logical processors to books, chips, drawers, and nodes, the polarization of the processors (high, medium, low), the processor type (regular CP, zIIP, or zAAP), and the association to WLM nodes. The tool can be used to understand the processor placement and how it changes when topology changes occur.

In order to run the tool it is required to install the exe file from this webpage and afterwards two z/OS datasets on your local z/OS system. The install file creates two entries: "TopoReport.lnk" and "Topo Report Help.lnk" in the Windows program folder "IBM RMF Performance Management". Please select the "Topo Report Help" link and follow the instructions in topic "Processing SMF 99 data" to install and execute the z/OS datasets and programs. The other topics in the help file describe the usage of the Excel spreadsheet to display the information on your workstation.

Requirements: A z10 or newer System z environment with partitions running in Hiperdispatch mode

Collecting SMF 99 subtype 14 records

Excel Version 2013. The spreadsheet should also work on Excel 2007 and 2010