



Advanced Technical Skills (ATS) North America

CPU MF – 2011 Update and WSC Experiences

SHARE - Session 9999

August 10, 2011

John Burg

jpburg@us.ibm.com

IBM
Washington Systems Center



Trademarks

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

AlphaBlox*	GDPS*	RACF*	Tivoli*
APPN*	HiperSockets	Redbooks*	Tivoli Storage Manager
CICS*	HyperSwap	Resource Link	TotalStorage*
CICS/VSE*	IBM*	RETAIN*	VSE/ESA
Cool Blue	IBM eServer	REXX	VTAM*
DB2*	IBM logo*	RMF	WebSphere*
DFSMS	IMS	S/390*	zEnterprise
DFSMSHsm	Language Environment*	Scalable Architecture for Financial Reporting	xSeries*
DFSMSrmm	Lotus*	Sysplex Timer*	z9*
DirMaint	Large System Performance Reference™ (LSPR™)	Systems Director Active Energy Manager	z10
DRDA*	Multiprise*	System/370	z10 BC
DS6000	MVS	System p*	z10 EC
DS8000	OMEGAMON*	System Storage	z/Architecture*
ECKD	Parallel Sysplex*	System x*	z/OS*
ESCON*	Performance Toolkit for VM	System z	z/VM*
FICON*	PowerPC*	System z9*	z/VSE
FlashCopy*	PR/SM	System z10	zSeries*
	Processor Resource/Systems Manager		

* Registered trademarks of IBM Corporation

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.

Techdocs provides the latest ATS technical collateral

www.ibm.com/support/techdocs

Techdocs - the Technical Sales Library

This site provides access to the Technical Sales Support organization's technical information databases. It gives you access to the most current installation, planning and technical support information available from IBM pre-sales support, and is constantly updated. You can browse or search these databases by date, document number, product, platform, keywords, etc.

New to Techdocs? Take a look at our [detailed introduction](#), which describes the document categories available (those listed on the navigation area on the left side of this page).

Rather than browse these categories, as a convenience you may enter a search of the full **Techdocs** database, or of any category you wish, here:

Search: Allow word variants
 "Fuzzy" search
 for:
 Hits: Order by:
 Include docs updated: [Help for Search](#)

Also available: our [Advanced search](#), where you can select documents based on various assigned document attributes.

New to Techdocs?
 Is this your first visit to **Techdocs** (the Technical Sales Library)?
 → [Learn more](#)

Returning to Techdocs?
 Looking for what's new in the **Techdocs Library**?
 → [Latest updates](#)

Need Technical Support?
 Looking for support resources or other documents and tools?
 → [Support & downloads](#)

Topics

- **CPU MF Introduction**
 - What is it and Why would you want to use?

- **Workload Characterization Update**
 - Step 1 completed

- **Key Performance Metrics for z10s and z196s**
 - CPI, Problem State, Cache / Memory Hierarchy
 - New metrics and formulas

- **2011 WSC Customer Experiences with SMF 113s**
 - New Uses
 - z196 Customer Initiated Power Savings Mode Detection
 - Utilization Effect in Capacity Planning
 - Crypto Function Positioning and Usage

 - What's New
 - COUNTER Data Loss - APAR OA36816
 - z/VM Support – APAR VM64961


- **Summary**

- **Back Up**
 - CPU MF Enablement
 - Old WSC Experiences

CPU Measurement Facility Introduction

What is the CPU Measurement Facility

- **New hardware instrumentation facility “CPU Measurement Facility” (CPU MF)**
 - Available on System z10 GA2 (EC and BC), z196, and z114
 - Supported by a new z/OS component, Hardware Instrumentation Services (HIS)

 - **Potential Uses – for this new “cool” virtualization technology**
 - COUNTERS
 - Supplement Current Performance Metrics – why performance may have changed
 - Workload characterization
 - SAMPLING
 - ISV product improvement
 - Application Tuning
- New LSPR Workloads**
- 

What is the z10 CPU Measurement Facility

- **IBM Research article**

- *“IBM System z10 performance improvements with software & hardware synergy”*
- <http://www.research.ibm.com/journal/rd/531/jackson.pdf>
- Contact IBM team for copy of the article

New March 2011:

- **Feb 2011 *Hot Topics* - A z/OS Newsletter - GA22-7501**

- *“A whole lot of benefits from HIS data” article page 24*
 - *COUNTERS and an update on SAMPLING - HIS report tool and STG Lab Services*

- **Redpaper *Setting Up and Using System z CPU Measurement Facility with z/OS***

- <http://www.redbooks.ibm.com/redpieces/pdfs/redp4727.pdf>

CPU MF COUNTERS – What and Why

■ What is CPU MF?

- A new z10 and later facility that provides cache and memory hierarchy COUNTERS
- Also capable of time-in-Csect type SAMPLES
- Data gathering controlled through z/OS HIS (HW Instrumentation Services)
 - Collected on an LPAR basis
 - Written to SMF 113 records
 - Minimal overhead

■ How can the COUNTERS be used today?

- To supplement current performance data from SMF, RMF, DB2, CICS, etc.
- To help understand why performance may have changed

■ How can the COUNTERS be used for future processor planning?

First Step in workload
Characterization

- They provide the basis for the new LSPR workload categories
- zPCR automatically processes CPU MF data to provide a workload "hint" based on RNI
 - zPCR still defaults to old IO-based methodology for workload selection
 - RNI "hint" is a "work in progress"

Recommend Capturing CPU MF SMF 113 Records on z10s, z196s, and z114s

Workload Characterization Update

New LSPR Workload Categories

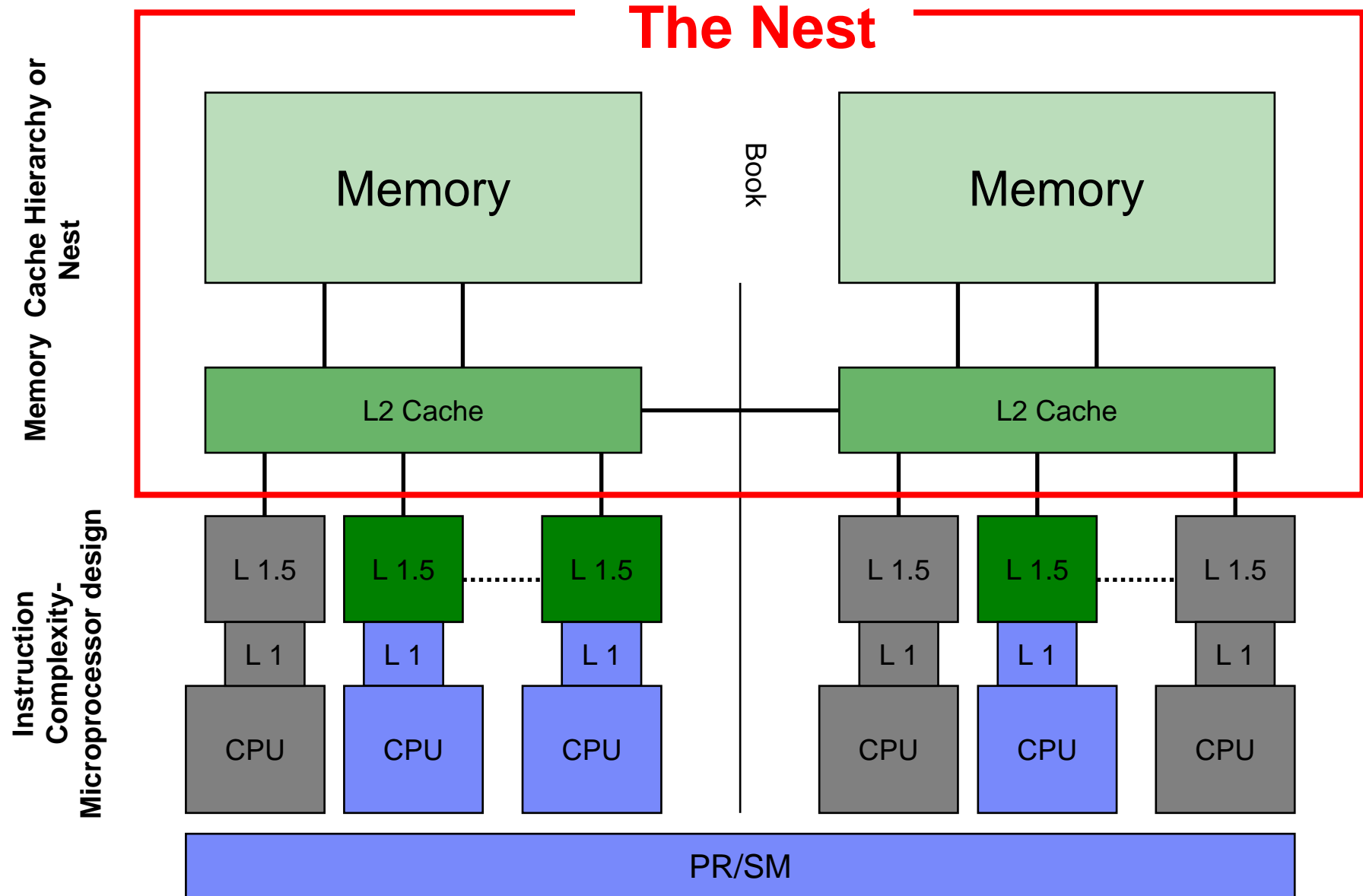
- Historically, **LSPR workload capacity curves** (primitives and mixes) have had application names or been **identified by a “software” captured characteristic**
 - For example, CICS, IMS, OLTP-T, CB-L, LoIO-mix, TI-mix, etc
- However, **capacity performance is more closely associated with how a workload is using and interacting with a processor “hardware” design**
- With the availability of CPU MF (**SMF 113**) data on z10, the ability to gain **insight into the interaction of workload and hardware has arrived**
- The knowledge gained is still evolving, but the **first step in the process is to produce LSPR workload capacity curves based** on the underlying hardware sensitivities
- Thus, the **LSPR for z196 will introduce three new workload categories** which replace all prior primitives and mixes
 - **Based on** new hardware defined metric called **Relative Nest Intensity**
 - Low, Average, High (Relative Nest Intensity)
- To simplify the transition, an easy and automatic translation of old names to new categories will be supplied in zPCR
 - E.G if you have been using LoIO-mix in your studies, you’ll simply use the new “Average”

Fundamental Components of Workload Capacity Performance

- **Instruction Complexity (Micro processor design)**
 - Many design alternatives
 - Cycle time (GHz), instruction architecture, pipeline, superscalar, Out-Of-Order, branch prediction and more
 - Workload effect
 - May be different with each processor design
 - **But once established for a workload on a processor, doesn't change very much**

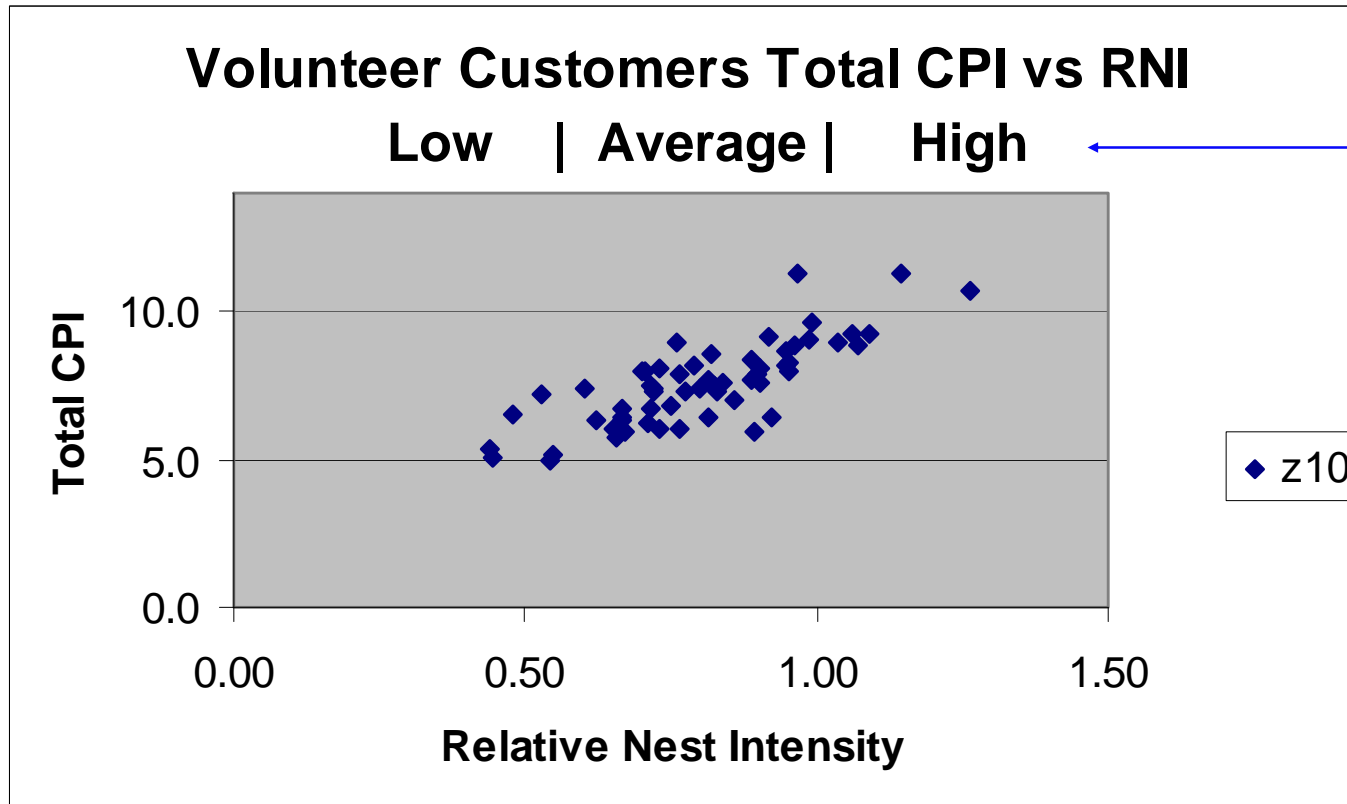
- **Memory Hierarchy or “Nest”**
 - Many design alternatives
 - Cache (levels, size, private, shared, latency, MESI protocol), controller, data buses
 - Workload effect
 - Quite variable
 - **Sensitive to many factors: locality of reference, dispatch rate, IO rate, competition with other applications and/or LPARs, and more**
 - Net effect of these factors represented in “Relative Nest Intensity”
 - **Relative Nest Intensity (RNI)**
 - **Activity beyond private-on-chip cache(s) is the most sensitive area**
 - **Reflects distribution and latency of sourcing from shared caches and memory**
 - Level 1 cache miss per 100 instructions (L1MP) also important
 - Data for calculation available from CPU MF (SMF 113) starting with z10

z10 Memory / Cache Hierarchy



CPU MF

z10 Customer Workload Characterization Summary



3) Created new LSPR workloads

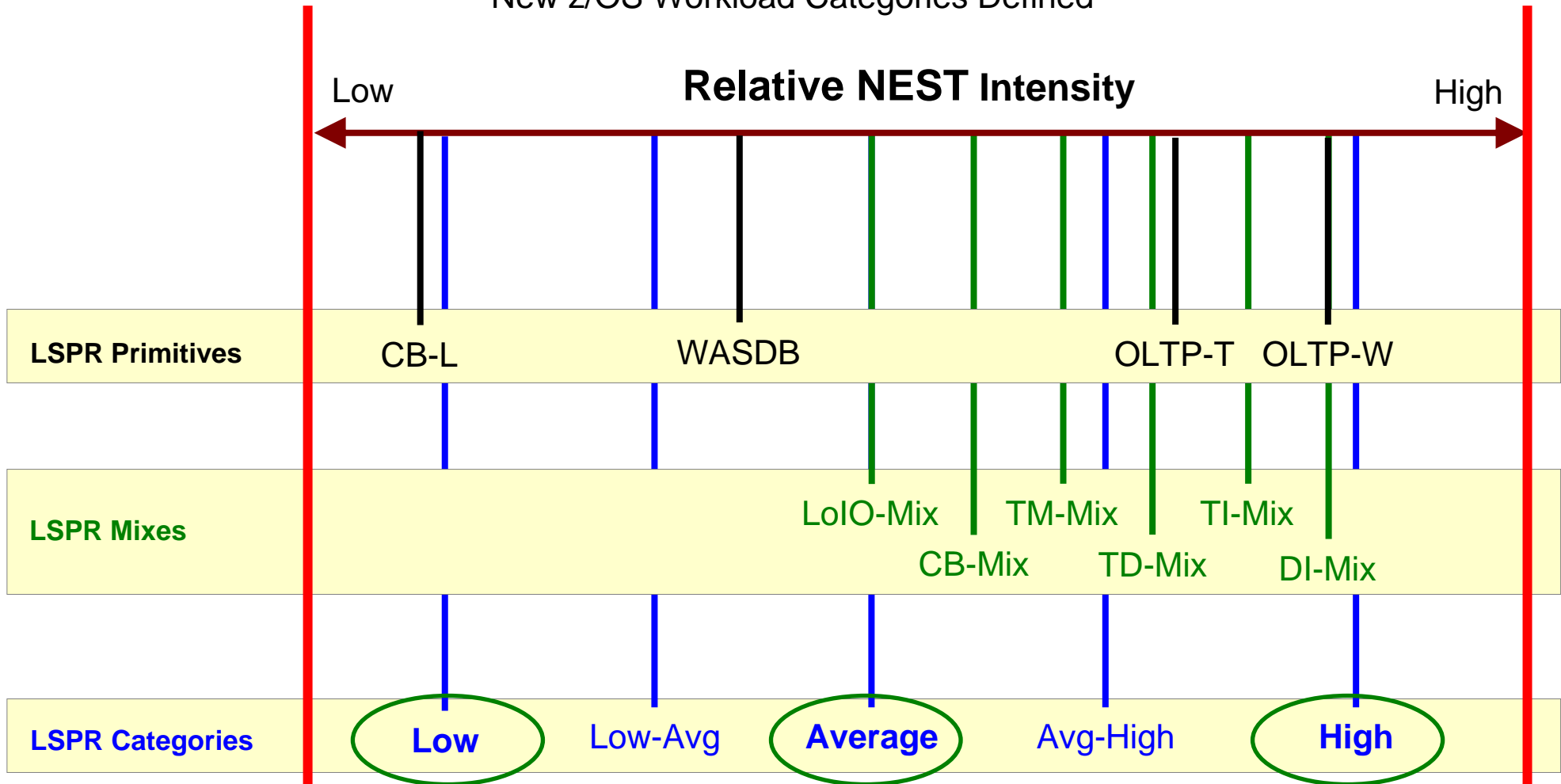
1) Customer CPI measurements

2) Created new RNI metric

zPCR Workload Characterization for z/OS

“Scope of Work” Definition Change

New z/OS Workload Categories Defined



Use zPCR’s Workload Selection Assistant to choose appropriate workload category

Automated with EDF input into zPCR

Note: Workload selection is automated in zCP3000

Workload Characterization Future Vision – Step 1 is Complete

- **Future vision to help identify workload characteristics and to provide better input for capacity planning and performance**
 - Step 1 – Created Workload Categories from SMF 113s - **completed**
 - **Over 150 z10 Customer/Partitions participated. Thank You!**
 - Measured LSPR with these new Categories
 - Step 2 – Refine Workload Selection Process
 - **As you move to z196 from z10, looking for “Before” and “ After” volunteers**
 - **Have received over 70 z196 Customer/Partitions thru Aug 1, but we need more!**

Still Looking for “Volunteers” – (3 days, 24 hours/day, SMF 70s, 72s, 113s per LPAR) “Before z10” and “After z196”

If interested send note to jpburg@us.ibm.com, No deliverable will be returned

Benefit: Opportunity to ensure your data is used to influence analysis

Key Performance Metrics for z10s and z196s

z196 versus z10 hardware comparison

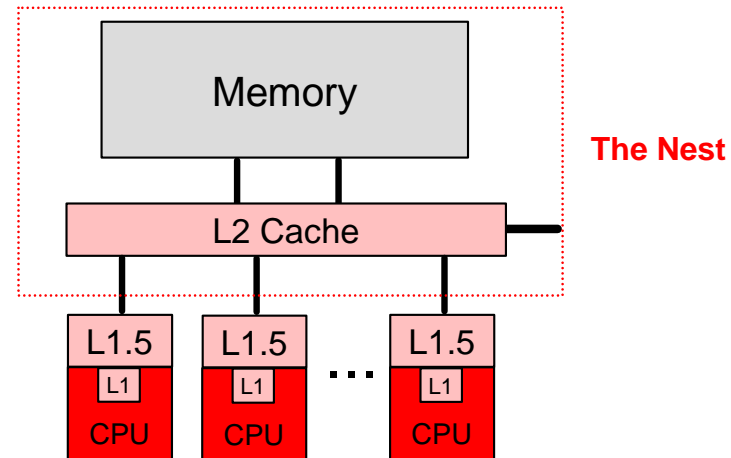
■ z10 EC

▶ CPU

- 4.4 GHz

▶ Caches

- L1 private 64k i, 128k d
- L1.5 private 3 MB
- L2 shared 48 MB / book
- book interconnect: star



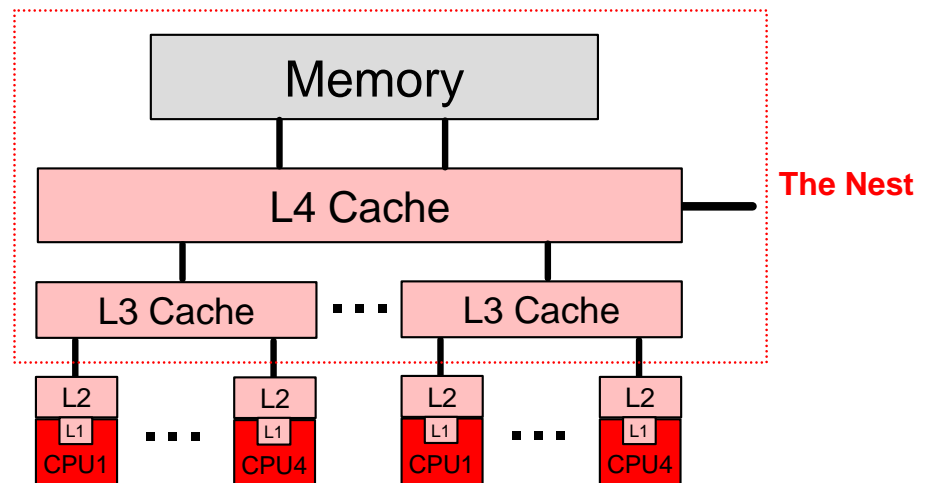
■ z196

▶ CPU

- 5.2 GHz
- Out-Of-Order execution

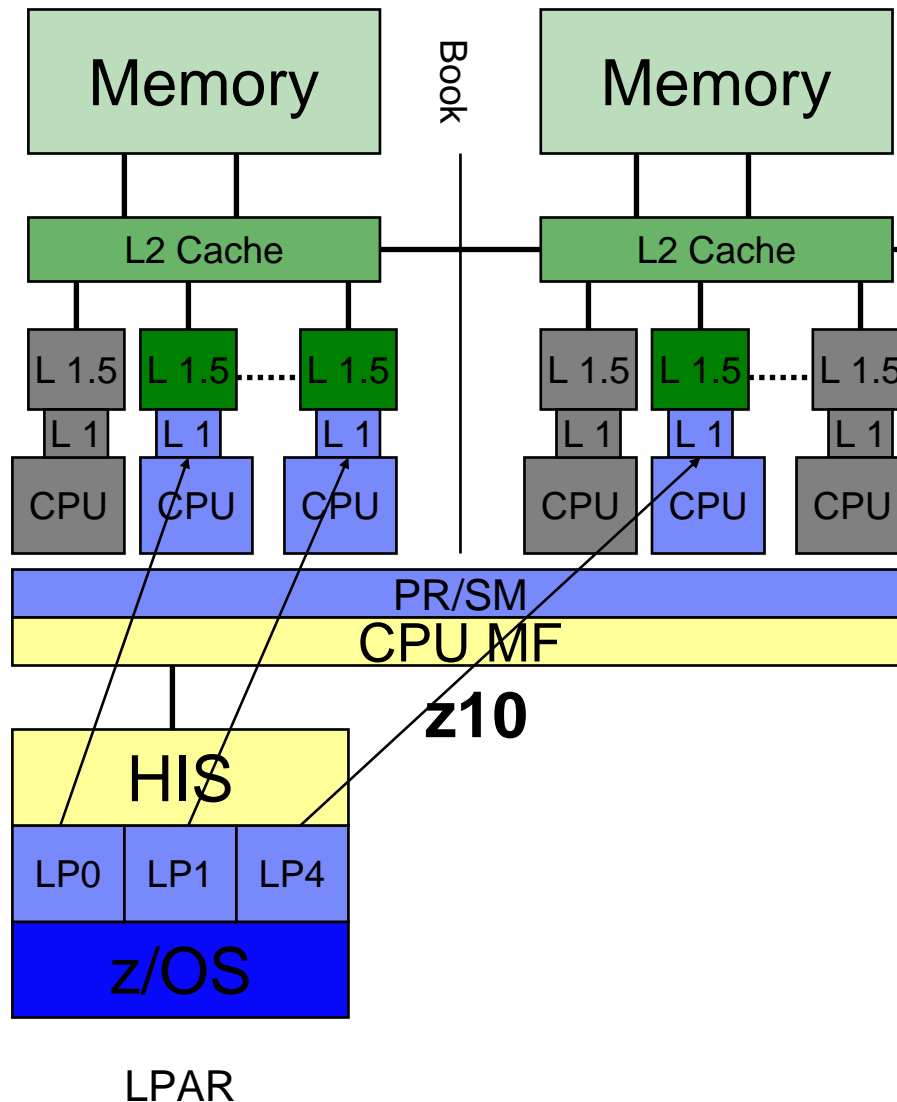
▶ Caches

- L1 private 64k i, 128k d
- L2 private 1.5 MB
- L3 shared 24 MB / chip
- L4 shared 192 MB / book
- book interconnect: star



z10 CPU MF and HIS provide a z/OS logical view of resource usage and cache / memory hierarchy sourcing

LPAR / Logical CP view:



Memory Accesses

Cache

- L 2 / (L4 z196) Accesses (local and remote)
- L3 Accesses on z196
- L1.5 / (L2 z196) Accesses
- L1 Sourced from Hierarchy

Instructions and Cycles

Crypto function

Current CPU MF Key Performance Metrics:

CPI	PRBSTATE	L1MP	L15P	L2LP	L2RP	MEMP	LPARCPU
-----	----------	------	------	------	------	------	---------

CPI – Cycles per Instruction

PRBSTATE - % Problem State

L1MP – Level 1 Miss Per 100 instructions


L15P – % sourced from L1.5 cache

L2LP – % sourced from Level 2 Local cache (on same book)

L2RP – % sourced from Level 2 Remote cache (on different book)

MEMP - % sourced from Memory

LPARCPU - APPL% (GCPs, zAAPs, zIIPs) captured and uncaptured

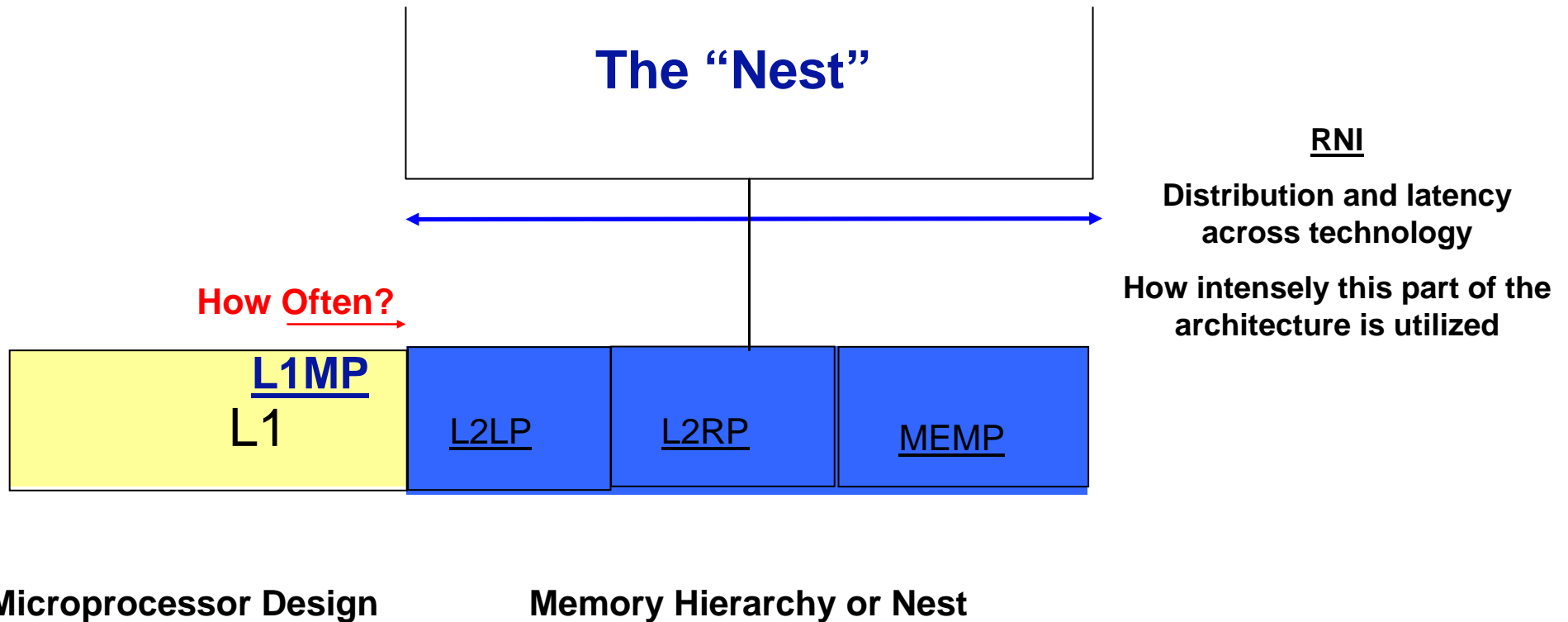
 Workload Characterization
L1 Sourcing from cache/memory hierarchy

Introducing the **new** Relative Nest Intensity (RNI) metric

■ Relative Nest Intensity

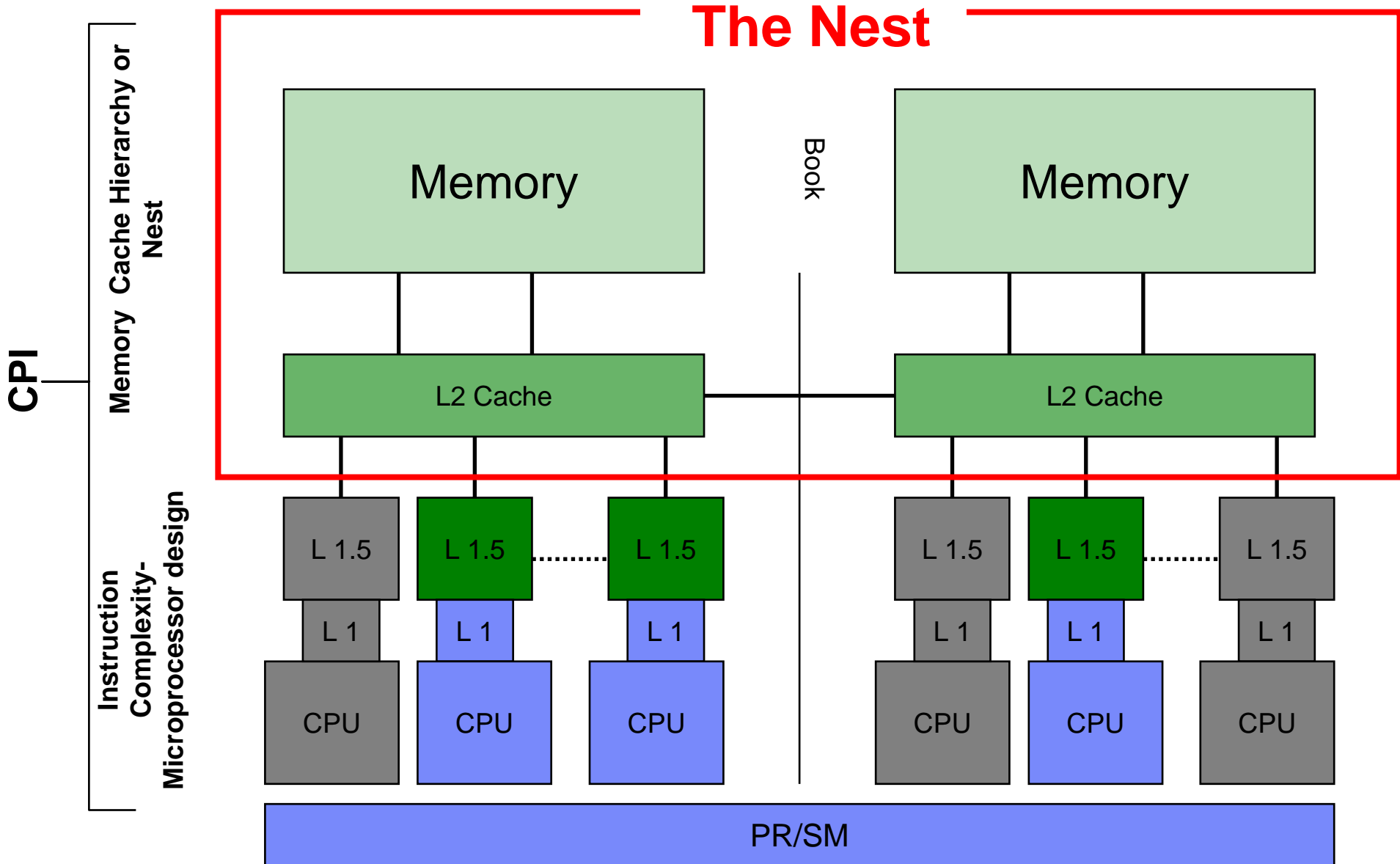
- Reflects the distribution and latency of sourcing from shared caches and memory
- For z10 Technology the Relative Nest Intensity = $(L2LP * 1 + L2RP * 2.4 + MEMP * 7.5) / 100$

Relative Nest Intensity

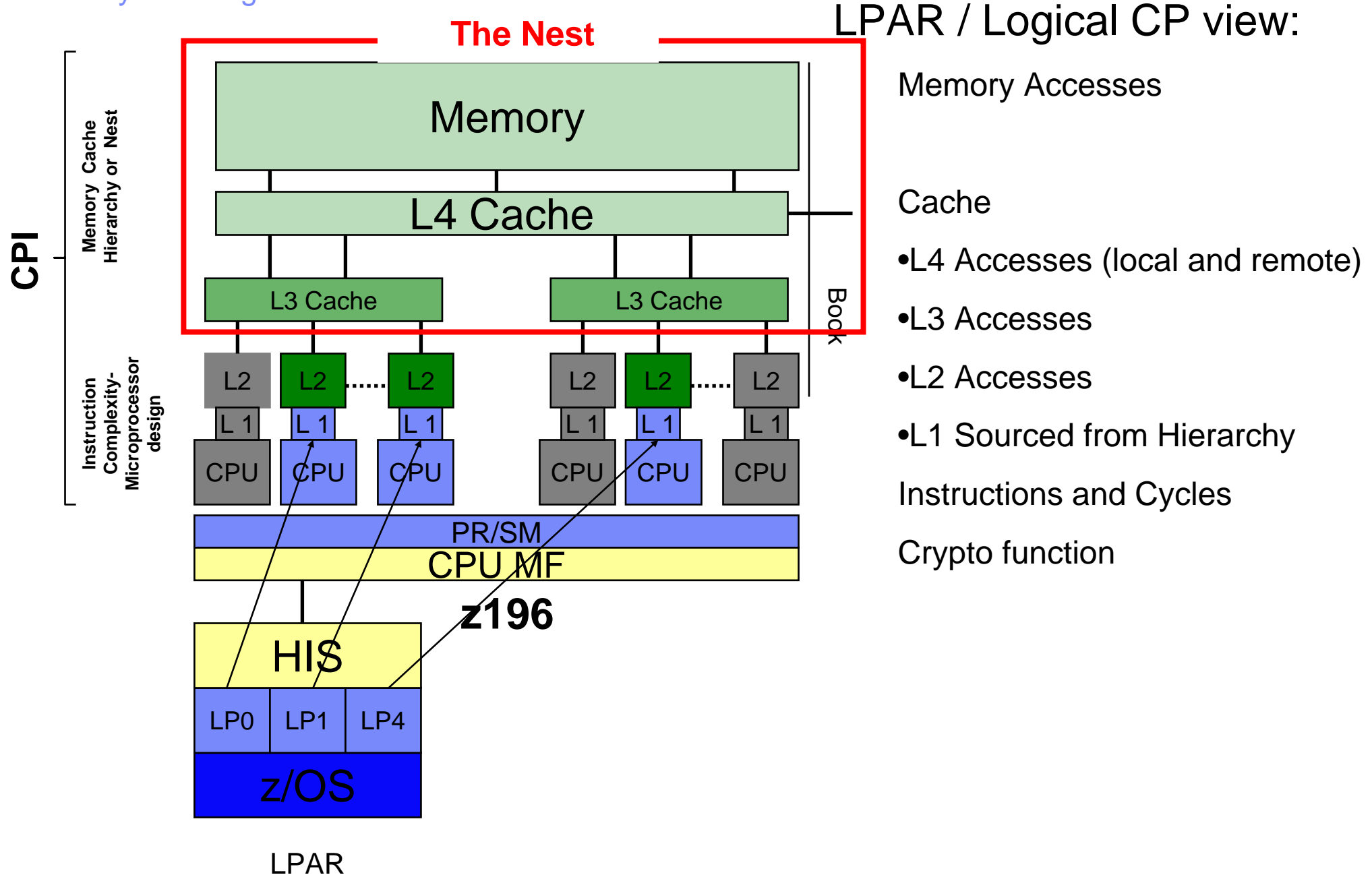


Note these Formulas may change in the future

New Estimated metrics: Instruction Complexity CPI, Finite CPI, and SCPL1M



z196 CPU MF and HIS provide a z/OS logical view of resource usage and cache / memory cache hierarchy sourcing



Updated z10 CPU MF Workload Characterization Summary



Customer	SYSID	MON	DAY	CPI	PRBSTATE	Est Instr Cmplx	Est Finite CPI	Est SCPL1M	L1MP	L15P	L2LP	L2RP	MEMP	Rel Nest Intensity	LPARCPU	Eff GHz
All Volunteers		Minimum		3.1	1.1	2.1	0.9	59.6	1.3	48.6	5.6	0.0	2.2	0.4	14.4	
All Volunteers		Average		7.2	31.2	3.2	3.9	101.4	3.9	68.9	21.2	1.6	8.3	0.9	376.3	
All Volunteers		Maximum		12.0	67.1	5.6	8.6	194.9	6.9	82.8	32.9	6.9	20.2	1.8	1442.3	4.40

New z10 columns are

1. Est Instr Cmplx CPI
2. Est Finite CPI
3. Est SCPL1M
4. Rel Nest Intensity
5. Eff GHz

CPI – Cycles per Instruction

Prb State - % Problem State

Est Instr Cmplx CPI – Estimated Instruction Complexity CPI (infinite L1)

Est Finite CPI – Estimated CPI from Finite cache/memory

Est SCPL1M – Estimated Sourcing Cycles per Level 1 Miss

L1MP – Level 1 Miss Per 100 instructions

L15P – % sourced from Level 2 cache

L2LP – % sourced from Level 2 Local cache (on same book)

L2RP – % sourced from Level 2 Remote cache (on different book)

MEMP - % sourced from Memory

Rel Nest Intensity – Reflects distribution and latency of sourcing from shared caches and memory

LPARCPU - APPL% (GCPs, zAAPs, zIIPs) captured and uncaptured

Eff GHz – Effective gigahertz for GCPs, cycles per nanosecond

Workload Characterization
L1 Sourcing from cache/memory hierarchy

WSC z196 Sample CPU MF from July – 5 Minute Synched Intervals

z196 **New z196 Metrics**

SYSID	Mon	Day	SH	Hour	CPI	Prb State	Est Instr Cmplx CPI	Est Finite CPI	Est SCPL1M	L1MP	L2P	L3P	L4LP	L4RP	MEMP	Rel Nest Intensity	LPARCPU	Eff GHz
SYSD	JUL	22	N	17.25	3.65	2.3	2.70	0.95	26	3.7	77.8	20.5	0.9	0.2	0.7	0.24	0.8	5.2
SYSD	JUL	22	N	17.33	3.68	2.3	2.73	0.95	26	3.6	77.4	20.8	0.9	0.2	0.7	0.24	0.8	5.2
SYSD	JUL	22	N	17.42	3.67	2.3	2.72	0.95	26	3.7	78.0	20.3	0.9	0.2	0.7	0.24	0.8	5.2
SYSD	JUL	22	N	17.50	3.64	2.3	2.71	0.93	26	3.6	77.8	20.5	0.9	0.2	0.7	0.24	0.8	5.2
SYSD	JUL	22	N	17.58	3.66	2.3	2.72	0.94	26	3.6	77.9	20.4	0.8	0.2	0.7	0.24	0.8	5.2
SYSD	JUL	22	N	17.67	3.65	2.3	2.72	0.94	26	3.6	77.0	21.1	0.9	0.2	0.7	0.24	0.8	5.2
SYSD	JUL	22	N	17.75	3.66	2.3	2.72	0.94	26	3.6	77.4	20.8	0.9	0.2	0.7	0.24	0.8	5.2
SYSD	JUL	22	N	17.83	3.64	2.3	2.70	0.94	26	3.6	77.1	21.0	0.9	0.2	0.7	0.24	0.8	5.2
SYSD	JUL	22	N	17.92	2.78	49.2	2.06	0.72	34	2.1	76.7	18.3	1.8	1.4	1.9	0.42	1.5	5.2
SYSD	JUL	22	N	18.00	3.65	3.2	2.71	0.94	26	3.6	77.0	21.1	1.0	0.2	0.7	0.25	0.8	5.2
SYSD	JUL	22	N	18.08	5.00	0.8	3.46	1.53	27	5.7	86.1	11.9	0.3	0.1	1.7	0.28	9.7	5.2
SYSD	JUL	22	N	18.17	3.72	3.2	2.76	0.96	27	3.6	76.8	21.0	1.1	0.2	0.8	0.26	0.9	5.2
SYSD	JUL	22	N	18.25	3.82	3.7	2.76	1.06	28	3.7	77.4	19.8	1.2	0.6	1.1	0.30	0.9	5.2

CPI – Cycles per Instruction

Prb State - % Problem State

Est Instr Cmplx CPI – Estimated Instruction Complexity CPI (infinite L1)

Est Finite CPI – Estimated CPI from Finite cache/memory

Est SCPL1M – Estimated Sourcing Cycles per Level 1 Miss

L1MP – Level 1 Miss Per 100 instructions

L2P – % sourced from Level 2 cache

L3P – % sourced from Level 3 on same Chip cache

L4LP – % sourced from Level 4 Local cache (on same book)

L4RP – % sourced from Level 4 Remote cache (on different book)

MEMP - % sourced from Memory

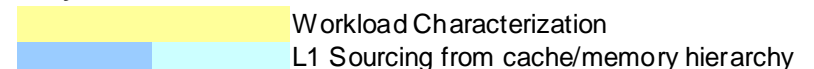
Rel Nest Intensity – Reflects distribution and latency of sourcing from shared caches and memory

LPARCPU - APPL% (GCPs, zAAPs, zIIPs) captured and uncaptured

Eff GHz – Effective gigahertz for GCPs, cycles per nanosecond

CPU MF provides measurement of the z196 Level 3 shared cache

These numbers come from a synthetic Benchmark and do not represent a production workload



Formulas – z10

Workload Characterization
L1 Sourcing from cache/memory hierarchy

Metric	Calculation – <i>note all fields are deltas between intervals</i>
CPI	$B0 / B1$
PRBSTATE	$(P33 / B1) * 100$
L1MP	$((B2+B4) / B1) * 100$
L15P	$((E128+E129) / (B2+B4)) * 100$
L2LP	$((E130+E131) / (B2+B4)) * 100$
L2RP	$((E132+E133) / (B2+B4)) * 100$
MEMP	$((E134+E135) + (B2+B4-E128-E129-E130-E131-E132-E133-E134-E135)) / (B2+B4)) * 100$
LPARCPU	$(((1/CPSP/1,000,000) * B0) / \text{Interval in Seconds}) * 100$

CPI – Cycles per Instruction

PRBSTATE - % Problem State

L1MP – Level 1 Miss Per 100 instructions

L15P – % sourced from L1.5 cache

L2LP – % sourced from Level 2 Local cache (on same book)

L2RP – % sourced from Level 2 Remote cache (on different book)

MEMP - % sourced from Memory

LPARCPU - APPL% (GCPs, zAAPs, zIIPs) captured and uncaptured

B* - Basic Counter Set - Counter Number

P* - Problem-State Counter Set - Counter Number

See “The Set-Program-Parameter and CPU-Measurement Facilities”
SA23-2260-0 for full description

E* - Extended Counters - Counter Number

See “IBM The CPU-Measurement Facility Extended Counters Definition
for z10” SA23-2261-0/1 for full description

CPSP - SMF113_2_CPSP “CPU Speed”

Formulas – z10 Additional

Metric	Calculation – note all fields are <i>deltas</i> between intervals
Est Instr Cmplx CPI	CPI – Estimated Finite CPI
Est Finite CPI	$((B3+B5) / B1) * .84$
Est SCPL1M	$((B3+B5) / (B2+B4)) * .84$
Rel Nest Intensity	$(1.0*L2LP + 2.4*L2RP + 7.5*MEMP) / 100$
Eff GHz	CPSP / 1000

Note these Formulas may change in the future

Est Instr Cmplx CPI – Estimated Instruction Complexity CPI (infinite L1)

Est Finite CPI – Estimated CPI from Finite cache/memory

Est SCPL1M – Estimated Sourcing Cycles per Level 1 Miss

Rel Nest Intensity – Reflects distribution and latency of sourcing from shared caches and memory


Eff GHz – Effective gigahertz for GCPs, cycles per nanosecond

B* - Basic Counter Set - Counter Number

P* - Problem-State Counter Set - Counter Number

See “The Set-Program-Parameter and CPU-Measurement Facilities”
SA23-2260-0 for full description

CPSP - SMF113_2_CPSP “CPU Speed”


 Workload Characterization
 L1 Sourcing from cache/memory hierarchy

Formulas – z196

Workload Characterization
L1 Sourcing from cache/memory hierarchy

Metric	Calculation – <i>note all fields are deltas between intervals</i>
CPI	$B0 / B1$
PRBSTATE	$(P33 / B1) * 100$
L1MP	$((B2+B4) / B1) * 100$
L2P	$((E128+E129) / (B2+B4)) * 100$
L3P	$((E150+E153) / (B2+B4)) * 100$
L4LP	$((E135+E136+E152+E155) / (B2+B4)) * 100$ L3 off chip in local L4
L4RP	$((E138+E139+E134+E143) / (B2+B4)) * 100$ L3 off book in remote L4
MEMP	$((E141+E142) + (B2+B4-E128-E129-E150-E153-E135-E136-E152-E155-E138-E139-E134-E143-E141-E142)) / (B2+B4)) * 100$
LPARCPU	$(((1/CPSP/1,000,000) * B0) / \text{Interval in Seconds}) * 100$

CPI – Cycles per Instruction

Prb State - % Problem State

L1MP – Level 1 Miss Per 100 instructions

L2P – % sourced from Level 2 cache

L3P – % sourced from Level 3 on same Chip cache

L4LP – % sourced from Level 4 Local cache (on same book)

L4RP – % sourced from Level 4 Remote cache (on different book)

MEMP - % sourced from Memory

LPARCPU - APPL% (GCPs, zAAPs, zIIPs) captured and uncaptured

B* - Basic Counter Set - Counter Number

P* - Problem-State Counter Set - Counter Number

See “The Set-Program-Parameter and CPU-Measurement Facilities” SA23-2260-0 for full description

E* - Extended Counters - Counter Number

See “The CPU-Measurement Facility Extended Counters Definition for z10 and z196” SA23-2261-01 for full description

CPSP - SMF113_2_CPSP “CPU Speed”

Formulas – z196 Additional

Metric	Calculation – <i>note all fields are deltas between intervals</i>
Est Instr Cmplx CPI	CPI – Estimated Finite CPI
Est Finite CPI	$((B3+B5) / B1) * (.57 + (0.1*RNI))$ updated *
Est SCPL1M	$((B3+B5) / (B2+B4)) * (.57 + (0.1*RNI))$ updated *
Rel Nest Intensity	$1.6*(0.4*L3P + 1.0*L4LP + 2.4*L4RP + 7.5*MEMP) / 100$
Eff GHz	CPSP / 1000

Note these Formulas may change in the future

* Updated March 2011

Est Instr Cmplx CPI – Estimated Instruction Complexity CPI (infinite L1)

Est Finite CPI – Estimated CPI from Finite cache/memory

Est SCPL1M – Estimated Sourcing Cycles per Level 1 Miss

Rel Nest Intensity – Reflects distribution and latency of sourcing from shared caches and memory

Eff GHz – Effective gigahertz for GCPs, cycles per nanosecond

Workload Characterization

L1 Sourcing from cache/memory hierarchy

B* - Basic Counter Set - Counter Number

P* - Problem-State Counter Set - Counter Number

See “The Set-Program-Parameter and CPU-Measurement Facilities”
SA23-2260-0 for full description

CPSP - SMF113_2_CPSP “CPU Speed”

2011 WSC Experiences

New Uses

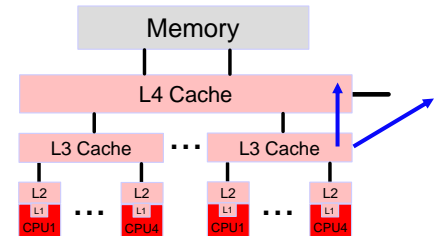
What's New?

CPU MF – WSC Experiences since March 2011

- **Customers continue to successfully running CPU MF COUNTERS in Production Today**

- **z196**

- HD=YES is even more important on z196, ensure HD=YES, 0-11% for 1 Book z196
 - See “Planning Considerations for HiperDispatch Mode **Version 2**” **WP101229**
<http://www.ibm.com/support/techdocs>
- CPU MF provides measurement of z196 L3 cache
- L3 off chip and off book sourced from respective L4s - see slide 27
 - Typically evenly distributed when HD=YES



- **New Uses**

- Measurement of z196 Customer Initiated Power Savings Mode - APAR OA29530
- Measurement of “Utilization Effect”
- Crypto Function Positioning and CPACF Usage

- **What’s New**

- Support for COUNTER Data Loss – APAR OA36816
- New z/VM CPU MF COUNTERS support – APAR VM64961

New Uses

z196 Customer Initiated Power Savings Mode

z196 - Power Save Mode - Customer Initiated

- Reduce the energy consumption of your system
- Can be done on a scheduled basis
- A zCPC can be placed in power saving mode only once per day
- In z/OS when a Power Save event occurs:
 - SMF interval is ended and new one started
 - MSU and SU/SEC values are changed
 - SMF records record change (30, 70, 72, 89, [113.2](#), new 90.34)
 - Requires CPU times to be normalized, service units would be correct
- Support Provided by APAR OA29530

WSC Test for z196 Power Save Mode with APAR OA29530

17:20:19.07 HIS032I STATE CHANGE DETECTED, ACTION=SAVE.

17:20:19.33 IWM063I WLM POLICY WAS REFRESHED DUE TO A PROCESSOR SPEED CHANGE

17:20:19.33 IWM064I THE SYSTEM IS RUNNING WITH REDUCED CAPACITY BECAUSE OF A MANUAL CONTROL SETTING

...

17:41:34.18 HIS032I STATE CHANGE DETECTED, ACTION=SAVE.

17:41:34.46 IWM063I WLM POLICY WAS REFRESHED DUE TO A PROCESSOR SPEED CHANGE

17:41:34.46 IWM064I THE SYSTEM IS RUNNING AT NOMINAL CAPACITY.

WSC z196 Power Save SMF 113 Test Results – Jan 9 2011

SMF 113 - Subtype 2

Time Stamp	CPU #	CpuProcClass	CPSP
11 JAN 09 17:20:00.00	0	1	5206
11 JAN 09 17:20:00.00	1	1	5206
11 JAN 09 17:20:00.00	2	6	5206
11 JAN 09 17:20:00.00	3	6	5206
11 JAN 09 17:20:19.07	0	1	5206
11 JAN 09 17:20:19.07	1	1	5206
11 JAN 09 17:20:19.07	2	6	5206
11 JAN 09 17:20:19.07	3	6	5206
11 JAN 09 17:20:19.17	0	1	4452
11 JAN 09 17:20:19.17	1	1	4452
11 JAN 09 17:20:19.17	2	6	4452
11 JAN 09 17:20:19.17	3	6	4452
11 JAN 09 17:25:00.00	0	1	4452
11 JAN 09 17:25:00.00	1	1	4452
11 JAN 09 17:25:00.00	2	6	4452
11 JAN 09 17:25:00.00	3	6	4452
11 JAN 09 17:41:34.18	0	1	4452
11 JAN 09 17:41:34.18	1	1	4452
11 JAN 09 17:41:34.18	2	6	4452
11 JAN 09 17:41:34.18	3	6	4452
11 JAN 09 17:41:34.35	0	1	5206
11 JAN 09 17:41:34.35	1	1	5206
11 JAN 09 17:41:34.35	2	6	5206
11 JAN 09 17:41:34.35	3	6	5206

RMF CPU Activity Report - Nominal (NONE) and POWERSAVE

C P U A C T I V I T Y

```

z/OS V1R12                SYSTEM ID SYSD                DATE 01/09/2011
                           RPT VERSION V1R12 RMF                TIME 17.15.00
CPU      2817  CPC CAPACITY  6053                SEQUENCE CODE 000000000000C7675
MODEL    778   CHANGE REASON=NONE                HIPERDISPATCH=YES
H/W MODEL M80
    
```

C P U A C T I V I T Y

```

z/OS V1R12                SYSTEM ID SYSD                DATE 01/09/2011
                           RPT VERSION V1R12 RMF                TIME 17.20.00
CPU      2817  CPC CAPACITY  5024                SEQUENCE CODE 000000000000C7675
MODEL    778   CHANGE REASON=POWERSAVE            HIPERDISPATCH=YES
H/W MODEL M80
    
```

Comparisons of MSUs and Effective GHz Ratio in Power Save Mode

	MSUs	Effective GHz
Nominal (NONE)	6053	5.206
POWERSAVE	5024	4.452
POWERSAVE / NONE	83.0%	85.5%

New Uses Utilization Effect

CPU MF COUNTERS can be used to assess Utilization Effect

Common Capacity Planning knowledge

- Capacity Planning can be affected by machine utilization
- Growth in CPU time per transaction occurs as utilization grows
 - Transactions being processed on a system share the physical hardware resources (CPs, caches, memory buses) and software resources (z/OS, subsystems)
 - At 50% Utilization 2x fixed shared resources and ½ software work units to manage Vs 100% Utilization
 - See “Running IBM System z at High CPU Utilization” by Gary King **WP101208**
<http://www.ibm.com/support/techdocs>

New Insights and Guidelines

- Experience has shown a 3-5% change in CPU per 10% machine utilization
- With CPU MF, can analyze partition CPI Vs Machine utilization (SMF 70s) to empirically build a unique Utilization Effect with a regression best fit formula
 - Or use following guidelines: % change in CPU per 10% machine utilization
 - RNI Hint “Low” 3%
 - RNI Hint “Avg” 4%
 - RNI Hint “High” 5%

Note these guidelines may change in the future

Partition CPI Vs Total Machine GCP Busy Methodology

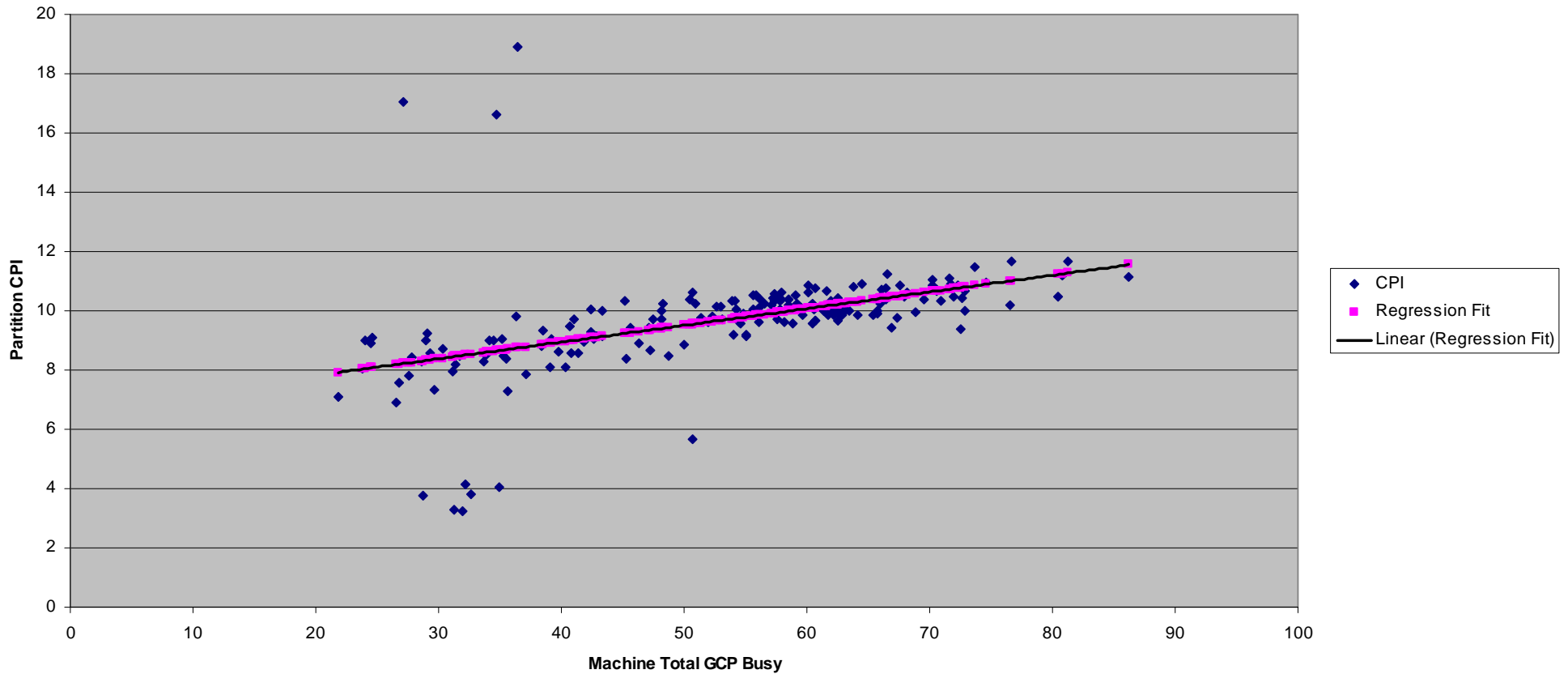
- **Regression Best Fit Methodology**
 - SMF 113 data provides CPI (Cycles Per Instruction) for each Partition
 - Proxy for CPU / trans encompassing all work with assumption that within each LPAR
 - Consistent mix of transactions and jobs
 - Consistent instructions per transaction and job
 - SMF 70 data provides overall machine utilization (total GCP utilization)
 - Plot Partition CPI versus total machine GCP utilization for each interval, 8AM-5PM
 - Select representative business cycle days
 - Regression fit to plotted points
 - Compare growth in best-fit CPI from 80% to 90% box utilization

Partition CPI Vs Total Machine GCP Busy

Regression Best Fit

$$Y = .05636X + 6.6576$$

CPI Vs Machine Busy Regression Fit



Regression Analysis and RNI "Hint" Guideline

Empirical Determined Regression Best Fit Analysis

- Formula is unique to this customer analysis
- $Y = .05636X + 6.6902$

		SYSA Machine Utilization %			
		70	80	90	← Estimate CPI at 70%, 80% and 90% machine Utilization
X					
Slope		5.636E-02	5.636E-02	5.636E-02	
Y Intercept		6.6902	6.6902	6.6902	
Estimated CPI	Y	10.64	11.20	11.76	← CPI and
	Increase		1.05	1.05	← % Difference

Best Fit results in 5% change in CPU per 10% machine utilization

e.g. so difference between 70% and 90% machine utilization is 10% more CPU

Results per 10% change in machine utilization

- Regression Best Fit: Partition average growth in best-fit CPI is 5.0%
- RNI Guideline: Average of 5 days is 4.2%
 - $(4+4+5+4+4) / 5 = 4.2$

SYSID	Mon	Day	SH	Hour	CPI	Prb State	Est Instr Cmplx CPI	Est Finite CPI	Est SCPL1M	L1MP	L15P / L2P	L3P	L2LP / L4LP	L2RP / L4RP	MEMP	Rel Nest Intensity	LPARCPU	Eff GHz	Machine Type	LSPR Wkld Hint
SYSA	MAY	12	P	TOTAL	8.35	0.2	3.35	5.00	108	4.6	69.7	0.0	18.0	3.8	8.5	0.91	135.1	4.4	Z10	AVG
SYSA	MAY	13	P	TOTAL	8.24	0.3	3.36	4.88	104	4.7	70.6	0.0	17.5	3.5	8.5	0.89	104.7	4.4	Z10	AVG
SYSA	MAY	16	P	TOTAL	6.93	0.7	3.11	3.82	117	3.3	71.0	0.0	15.0	2.7	11.3	1.06	54.1	4.4	Z10	HIGH
SYSA	MAY	17	P	TOTAL	7.89	0.3	3.27	4.62	111	4.2	70.4	0.0	16.8	3.4	9.3	0.95	95.7	4.4	Z10	AVG
SYSA	MAY	18	P	TOTAL	7.79	0.2	3.24	4.55	107	4.3	70.0	0.0	17.8	3.8	8.4	0.90	147.2	4.4	Z10	AVG

New Uses

Crypto Function Positioning and Usage by CPACF

Crypto Functions and CPU MF Measurements

- **There are 4 Crypto Functions:**

- Data Confidentiality
 - Symmetric
 - Asymmetric
- Data Integrity (Modification Detection / Modification Authentication)
- **Key Management**
- Financial PIN Functions

- **These Functions can be implemented in System z as follows:**

Software Routines

Data Confidentiality (Symmetric)
Data Confidentiality (Asymmetric)
Data Integrity (Modification Detection)
Data Integrity (Message Authentication)

CPACF - Crypto CoProcessor

Data Confidentiality (Symmetric DES/AES)
Data Integrity (Modification Detection)

Crypto Express Card

Data Confidentiality (Symmetric DES/AES)
Data Confidentiality (Asymmetric)
Data Integrity (Message Authentication)
Key Management
Financial PIN Functions

Crypto Exploiters and CPU MF Measurements

■ Crypto Exploiters by System z implementation include:

Software Routines

System SSL Handshake Phase 1 (2nd Choice)
System SSL Record Phase 2 (2nd Choice)

CPACF - Crypto CoProcessor

System SSL Record Phase 2 (1st Choice)

Data Encryption Tool for IMS and DB2

IBM Encryption Facility (CLRTDES or CLEAES option to encrypt with clear key)
IBM Encryption Facility (use Passphrase to generate data key)

DB2 BIFs

Crypto Express Card

System SSL Handshake Phase 1 (1st Choice)

Data Encryption Tool for IMS and DB2

IBM Encryption Facility (ENCTDES option to encrypt with secure key)
IBM Encryption Facility (use RSA option to protect the data key)

Tape/DASD Encryption (with ICSF as your keystore)



CPU MF COUNTERS can count of the type of functions

and CPU time for the Crypto Functions that execute on the CPACF Crypto CoProcessors

Sample Report – Crypto COUNTERS provide measurement of CPACF Crypto Co-Processor Usage

```

*** z10 Summary - CRYPTO Counters Information ***
***          TOTAL for all CPUS          ***

PRNG Function Count           0/Sec
PRNG Cycle Count              0/Sec
PRNG Blocked Function Count   0/Sec
PRNG Blocked cycle Count      0/Sec
SHA Function Count            0.73/Sec
SHA Cycle Count               592.47/Sec
SHA Blocked Function Count    0/Sec
SHA Blocked cycle Count       0/Sec
DEA Function Count            6277.39/Sec
DEA cycle count               332273396.24/Sec
DEA Blocked Function Count    0/Sec
DEA Blocked cycle Count       0/Sec
AES Function Count            0/Sec
AES Cycle Count               0/Sec
AES Blocked Function Count    0/Sec
AES Blocked cycle Count       0/Sec

***          CRYPTO BUSY SUMMARY          ***

PRNG Crypto Busy:  0.00% - for the 3 CPUS
SHA  Crypto Busy:  0.00% - for the 3 CPUS
DEA  Crypto Busy:  2.55% - for the 3 CPUS
AES  Crypto Busy:  0.00% - for the 3 CPUS
-----
Total Crypto Busy:  2.55% - for the 3 CPUS

```

This information may be useful in determining:

- A count of How Many CPACF encryption functions were executed
- How much CPU Time (cycles) were used

The encryption facility executed both SHA functions and TDES functions for this specific test.

Ran DASD dumps sequentially over 20 minute duration

With option: ENCRYPT(CLRTDES) -

These numbers come from a synthetic Benchmark and do not represent a production workload

It is important to remember that other Crypto functions may be executing in software and/or on Crypto Express Cards (if installed & implemented). This is not measured by the CPU MF Crypto COUNTERS

CPU MF Crypto COUNTERS can help assess how many of the Crypto Functions are occurring on the CPACF Co-Processors

See “A Synopsis of System z Crypto Hardware” by Greg Boyd **WP100810**
<http://www.ibm.com/support/techdocs>

What's New?

DATALOSS Parameter

OA36816 - DATALOSS Parameter Expanded to COUNTERS

- **F HIS “DATALOSS” parameter expanded to COUNTERS and Sampling for hardware data loss**
 - Was already valid for SAMPLING buffer overflow
 - Without APAR, HIS and SMF 113 recording stops until HIS restarted and modified to collect COUNTERS again
 - HIS025I DATA COLLECTION IS PREMATURELY ENDING.
MACHINE REPORTED COUNTER DATA LOSS

- **With APAR can now specify IGNORE for any COUNTER Data Loss**
 - Options are to IGNORE or STOP: DATALOSS= IGNORE|STOP
 - If STOP, HIS and recording of SMF 113s stops (until HIS started and modified to enable COUNTERS)

- **Recommendation is use IGNORE (default) as SMF 113s will reflect condition**
 - Can be abbreviated DL=I
 - New message HIS034I gets issued when DATALOSS=IGNORE and condition is detected
 - If COUNTER data loss occurs with IGNORE specified, HIS flags the SMF 113 records for that interval
 - SMF113_2_CF bit 4 set “hardware has lost counter data during the current interval”
 - If no COUNTER data loss in subsequent intervals, SMF 113_2_CF bit 4 is not set
 - **Advantage is no operational support needed to continue collecting SMF 113s for COUNTER data loss**

- **Planned availability is August 12, 2011**

What's New?

z/VM Supports CPU MF COUNTERS

New z/VM capability supports CPU MF COUNTERS

- **z/VM now supports CPU MF COUNTERS**
 - Provided by APAR VM64961
 - Planned availability is August 19, 2011
 - For z/VM 6.1 and z/VM 5.4 on z10s and z196s
- **Future vision is to help identify z/VM workload characteristics and to provide better input for capacity planning and performance**

CPU MF for z/VM Partitions

- The CPU Measurement Counter Facility provides for z/VM partitions:
 - Sets of counters for each logical processor that count events such as cycle, instruction, and cache directory-write counts
 - Same COUNTER information as z/OS partitions
 - Their accumulation is a relatively low-overhead activity and is performed automatically by the machine when the counters are authorized, enabled, and activated.
 - Authorization is controlled by a logical partition's Security settings in its activation profile. Enablement and activation are controlled by z/VM.

For long term gathering recommend Basic and Extended Counters, by setting only the Basic and Extended Counters Security settings

For z/VM Control Program, Problem State is 0 and CPACF is not utilized

- Enhancements in z/VM 6.1 and z/VM 5.4 provide the ability to collect CPU Measurement Facility counter data in the monitor stream.
- Support for the CPU-Measurement Sampling Facility and virtualization of the CPU-Measurement Facility interfaces for guest use are not provided.

z/VM CPU MF operation support

- New support includes:
 - New command syntax for the MONITOR SAMPLE command, along with a new message.
 - A new command response for the QUERY MONITOR command.
 - One new monitor record.
 - Domain 5 (Processor), Record 13 (CPU-Measurement Facility Counters)
 - A modification to an existing monitor record.
 - Domain 1 (Monitor), Record 14 (Domain Detail)
 - A new message for CP MONITOR SAMPLE ENABLE PROCESSOR and CP MONITOR SAMPLE ENABLE ALL



Sample z/VM CPU MF COUNTER Report

SYSID	Mon	Day	SH	Hour	Pool	CPI	Prb State	Est Instr Cmplx CPI	Est Finite CPI	Est SCPL1M	L1MP	L15P / L2P	L3P	L2LP / L4LP	L2RP / L4RP	MEMP	Rel Nest Intensity	LPARCPU	Eff GHz	Machine Type	LSPR Wkld Hint
Shift Summary																					
GDLBOFVM	OCT	26	N	TOTA	T	2.43	0.0	2.42	0.01	112	0.0	76.0	0.0	21.9	0.1	2.1	0.37	41.6	4.4	2097	LOW
GDLRCT1	OCT	27	P	TOTA	T	2.35	0.0	2.32	0.04	27	0.1	95.1	2.3	1.4	1.2	0.1	0.09	12.2	5.2	2817	LOW
Hourly																					
GDLBOFVM	OCT	26	N	19.50	T	4.53	0.0	3.01	1.52	121	1.3	44.2	0.0	52.1	0.0	3.8	0.80	0.1	4.4	2097	AVG
GDLBOFVM	OCT	26	N	19.60	T	4.55	0.0	2.99	1.56	124	1.3	44.5	0.0	51.4	0.0	4.1	0.82	0.4	4.4	2097	AVG
GDLBOFVM	OCT	26	N	19.70	T	2.52	0.0	2.47	0.04	101	0.0	69.8	0.0	26.1	0.0	4.1	0.57	16.8	4.4	2097	LOW
GDLBOFVM	OCT	26	N	19.80	T	2.45	0.0	2.44	0.01	124	0.0	78.4	0.0	19.7	0.1	1.9	0.34	202.3	4.4	2097	LOW
GDLBOFVM	OCT	26	N	19.90	T	2.43	0.0	2.42	0.01	109	0.0	78.7	0.0	19.5	0.1	1.7	0.33	226.8	4.4	2097	LOW
GDLBOFVM	OCT	26	N	20.00	T	2.43	0.0	2.42	0.01	106	0.0	79.0	0.0	19.4	0.1	1.5	0.31	242.3	4.4	2097	LOW
GDLBOFVM	OCT	26	N	20.10	T	2.40	0.0	2.40	0.01	113	0.0	77.9	0.0	20.3	0.1	1.7	0.33	143.0	4.4	2097	LOW
GDLRCT1	OCT	27	P	12.80	T	2.36	0.0	2.32	0.04	26	0.1	95.6	2.1	1.1	1.1	0.1	0.08	34.8	5.2	2817	LOW
GDLRCT1	OCT	27	P	12.90	T	2.35	0.0	2.31	0.04	28	0.1	94.8	2.4	1.5	1.2	0.1	0.09	87.3	5.2	2817	LOW
Pool Summary																					
GDLBOFVM	OCT	26	N	TOTA	1	2.43	0.0	2.42	0.01	112	0.0	76.0	0.0	21.9	0.1	2.1	0.37	41.6	4.4	2097	LOW
GDLRCT1	OCT	27	P	TOTA	1	2.29	0.0	2.27	0.02	33	0.1	89.9	5.0	2.8	2.1	0.2	0.18	11.8	5.2	2817	LOW
GDLRCT1	OCT	27	P	TOTA	3	12.38	0.0	5.74	6.65	46	14.4	88.2	1.0	3.8	6.8	0.2	0.36	0.0	5.2	2817	AVG
GDLRCT1	OCT	27	P	TOTA	4	10.09	0.0	7.60	2.50	24	10.5	97.9	0.9	0.6	0.6	0.0	0.04	0.4	5.2	2817	AVG
GDLRCT1	OCT	27	P	TOTA	5	12.96	0.0	5.46	7.49	51	14.7	88.3	0.6	1.3	9.3	0.5	0.44	0.0	5.2	2817	AVG
GDLRCT1	OCT	27	P	TOTA	6	12.39	0.0	5.63	6.76	47	14.5	87.9	0.8	4.7	6.1	0.5	0.37	0.0	5.2	2817	AVG

These numbers come from a synthetic Benchmark and do not represent a production workload

Workload Characterization Future Vision – z/VM Starting

- **Future vision to help identify workload characteristics and to provide better input for capacity planning and performance**
 - Step 1 – Need samples from Customer Production z/VM partitions to develop Workload Categories - **starting now August 2011**
 - **Hope to develop z/VM workload “Hint” like done for z/OS**
 - **As you move to z196 from z10, looking for z/VM “Before” and “ After” volunteers**

Looking for “Volunteers” – (1 hour MONWRITE with D5 R13 enabled)

If interested send note to jpburg@us.ibm.com and rflewis@us.ibm.com for instructions. No deliverable will be returned

Benefit: Opportunity to ensure your data is used to influence analysis

Summary

z10 and z196 CPU MF COUNTERS Summary

- **Traditional metrics continue to provide the best view of Performance**
 - CPU MF can help explain *why* a change occurred
- **First Step completed in Workload Characterization for Capacity Sizing**
 - Relative Nest Intensity calculation today gives a hint to zPCR
- **z196 Volunteers are still needed for our Workload Characterization study for refinement**
 - Feedback from Volunteers is this is very easy to enable, with a minimal time investment\
- **CPU MF has a very low overhead to run and is easy to implement**
 - Less than 1/100 of a second for HIS address space in 15 minute interval
 - Customers continue to successfully running CPU MF COUNTERS in Production Today
- **Recommend enabling CPU MF COUNTERS on z10s, z196s and z114s today!**
 - To supplement current performance metrics (e.g. from SMF, RMF, DB2, CICS), turn on and leave on
 - APAR OA30486 required for z196s and recommended for z10s
 - For long term gathering recommend CTR=(B,E) to get Extended Counters
- **CPU MF Overview and WSC Experiences Techdoc TC000066**
 - <http://www.ibm.com/support/techdocs/atmastr.nsf/WebIndex/TC000066>
 - CPU MF presentation and a detailed write up for enabling CPU MF
- **z/VM Supports CPU MF – Volunteers Needed**



**First Step in workload
Characterization**

New LSPR Workloads

Thank You
for attending!

Acknowledgements

- **Many people contributed to this presentation including:**

Riaz Ahmad

Greg Boyd

Jane Bartik

Harv Emery

Gary King

Frank Kyne

Marty Moroch

Steve Olenik

Bob Rogers

Bill Schray

Brian Smith

Bob St John

Elpida Tzortzatos

Kathy Walsh

Richard Lewis

Gretchen Frye

Brian Wade

Stephen Jones

Romney White

Les Geer

Kevin Adams

CPU MF COUNTERS Enablement

CPU MF Requirements to utilize z10 and z196

- **z196 or System z10 machine**
 - z10 must be at GA2 Driver 76D – Bundle #20 or higher

- **z10 z/OS LPAR being measured must be at z/OS 1.8 or higher with APARs:**
 - OA25755, OA25750, and OA25773 – also **OA30486 for z/OS 1.10 and higher for new functionality**
 - OA27623 also recommended to add “CPU Speed” to SMF 113s and HIS COUNTERS output
 - Not currently supported for z/OS running as a z/VM guest – **z/VM native prototype support in process**


- **z196 z/OS LPARs being measured at z/OS 1.9 or higher require APAR OA30486**
 - z/OS 1.8 requires OA33052

- **Recommended New SAMPLING APARS**
 - z/OS Mapping
 - z/OS 1.9 APAR OA32113
 - z/OS 1.10 APAR OA32113 and APAR OA34485
 - z/OS 1.11 APAR OA30429 and OA34485
 - z/OS 1.12 APAR OA34485
 - CICS Mapping
 - APAR PM08568 (for CTS 3.2) or APAR PM08573 (for CTS 4.1).

Steps to utilize z10 and z196 CPU MF

Operationally CPU MF works the same on z196

Steps to utilize CPU MF

- Configure the z10 or z196 to collect CPU MF Data
 - Update LPAR Security Tabs (See appendix)
- Configure HIS on z/OS to collect CPU MF Data
 - Set up HIS Proc 
 - Set up OMVS Directory
 - Collect SMF 113s via SMFPRMxx
- Collect CPU MF Data
 - Start HIS – Modify with Begin/End – for COUNTERS or SAMPLING
 - “F HIS,B,TT='Text',PATH='/his/',CTRONLY,CTR=ALL” or
 - “F HIS,B,TT='Text',PATH='/his/',CTRONLY,CTR=(B,E)”
- Analyze the CPU MF Data – SMF 113s

```
//HIS PROC
//HIS EXEC PGM=HISINIT,REGION=0K,TIME=NOLIMIT
//SYSPRINT DD SYSOUT=*
```

CPU MF has a very low overhead to run, is easy to implement, and is a very small SMF record

For long term gathering recommend CTR=(B,E) to get Extended Counters!

OA30486 - New support for Sync Interval, PU Type and STATECHANGE

- **APAR OA30486 with z/OS 1.12 GA – will be rolled down to z/OS V1R11 and z/OS V1R10**
 - **Applicable for z10s and z196s for new functionality**
 - **A. New CPU MF capability to SYNC the SMF 113 Interval with other SMF records**
 - SMFINTVAL=SYNC
 - Synchronize records with the SMF global recording interval
 - ..or choose Interval time 1-60
 - **Recommendation is “SYNC”:**

Recommend SMFINTVAL=SYNC or SI=SYNC

 - “F HIS,B,TT='Text',PATH='/his/',**CTRONLY,CTR=(B,E),SMFINTVAL=SYNC**”
 - **B. Identification of PU Type (GCP, zIIP or zAAP) in SMF 113 record**
 - SMF113_2_CpuProcClass '0' - GCP / '2' - zAAP / '4' - zIIP
 - **C. STATECHANGE – (see next slide)**
 - Both SMFINTERVAL and STATECHANGE can be abbreviated, e,g, SI=SYNC, SC=SAVE
 - “F HIS,B,TT='Text',PATH='/his/',**CTRONLY,CTR=(B,E),SI=SYNC,SC=SAVE**”
- **In SMF 113s the z196 processor is identified by**
 - SMF113_2_CTRVN2 = **'2'** for z196, '1' for z10

z196 Extended Counters have changed, use CTRVN2 to determine if z10 or z196

HIS STATECHANGE

- **HIS detects and handles significant hardware events (state change)**

- Replacement Capacity (Customer Initiated Upgrade)
- On/Off Capacity on demand
- z196 Power Save Mode

Verify with SMF 113s that “CPU Speed” or “Effective GHz” changed as expected

- **How HIS reacts depends on the STATECHANGE parameter specified**

- STATECHANGE=STOP
 - Stop the collection run when the event was detected
- STATECHANGE=IGNORE
 - Continue the collection run as if the event never happened
- STATECHANGE=SAVE (Default)
 - Record the previous state of the system (Save all data)
 - Write and close the .CNT file
 - Close all .SMP files (1 per CPU)
 - Cut SMF Type 113 Records (1 per CPU)
 - Continue the collection run with the new state
 - Create new .SMP files (1 per CPU)
 - Cut SMF Type 113 Records (1 per CPU)

Recommend STATECHANGE=SAVE, (the default) so don't need to specify

- **STATECHANGE information not directly reported in the SMF 113**

- You will see additional record(s) and an increase/decrease in CPIDs or “CPU Speed”

New HIS APAR OA30486 support for z196 – WSC Example

15:33:13.24 JPBURG 00000200 F HIS,B,TT='Z196 w/ TEST',CTRONLY,CTR=ALL,SMFINTVAL=SYNC

15:33:14.22 STC01226 00000000 HIS011I HIS DATA COLLECTION STARTED

Time Stamp	CPU #	CpuProcClass	CTNVN1	CTNVN2	CPSP
10 JUL 22 15:33:14.22	0	0	1	2	5206
10 JUL 22 15:33:14.22	1	0	1	2	5206
10 JUL 22 15:33:14.22	4	4	1	2	5206
10 JUL 22 15:33:14.22	5	4	1	2	5206
10 JUL 22 15:35:00.00	0	0	1	2	5206
10 JUL 22 15:35:00.00	1	'0' GCP → 0	1	2	5206
10 JUL 22 15:35:00.00	4	'4'-zIIP → 4	1	2	5206
10 JUL 22 15:35:00.00	5	4	1	2	5206

'5.2' GHz

z196

'2' z196

SMF 113 Synched with SMF Global Recording Interval - 5 Minutes

New HIS APAR OA30486 support for z196 – WSC Example

12:19:11.82 JPBURG 00000200 F HIS,B,TT='Z196 TEST', CTRONLY,CTR=ALL,SMFINTVAL=1

12:19:11.82 STC32434 00000000 HIS011I HIS DATA COLLECTION STARTED

Time Stamp	CPU #
10 JUL 16 12:19:11.82	0
10 JUL 16 12:19:11.82	1
10 JUL 16 12:19:11.82	2
10 JUL 16 12:19:11.82	3
10 JUL 16 12:20:11.82	0
10 JUL 16 12:20:11.82	1
10 JUL 16 12:20:11.82	2
10 JUL 16 12:20:11.82	3

SMF 113 Written every 1 Minute

The Display HIS command (D HIS) provides useful status information

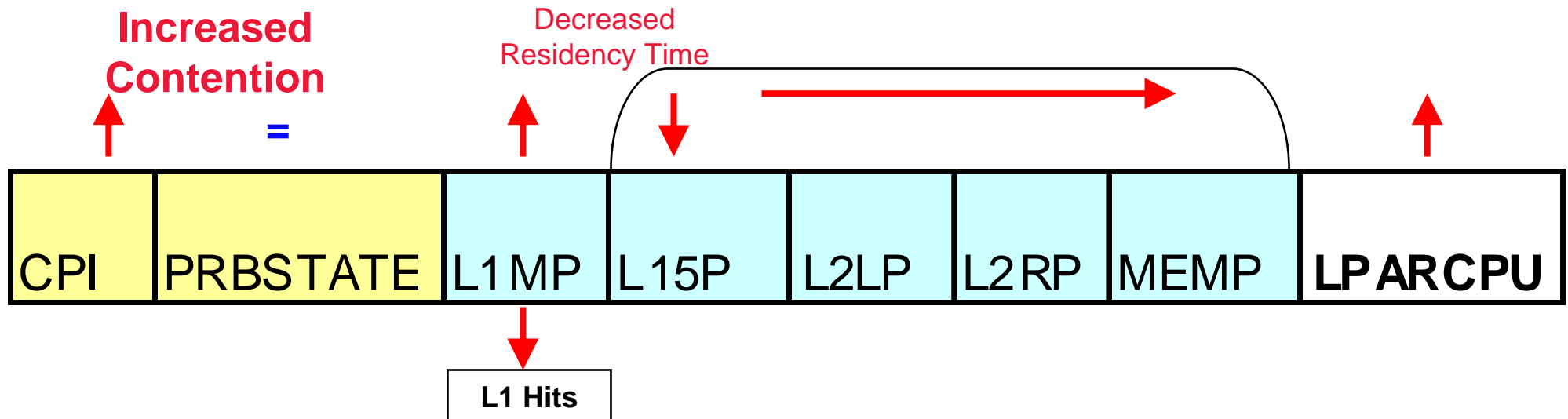
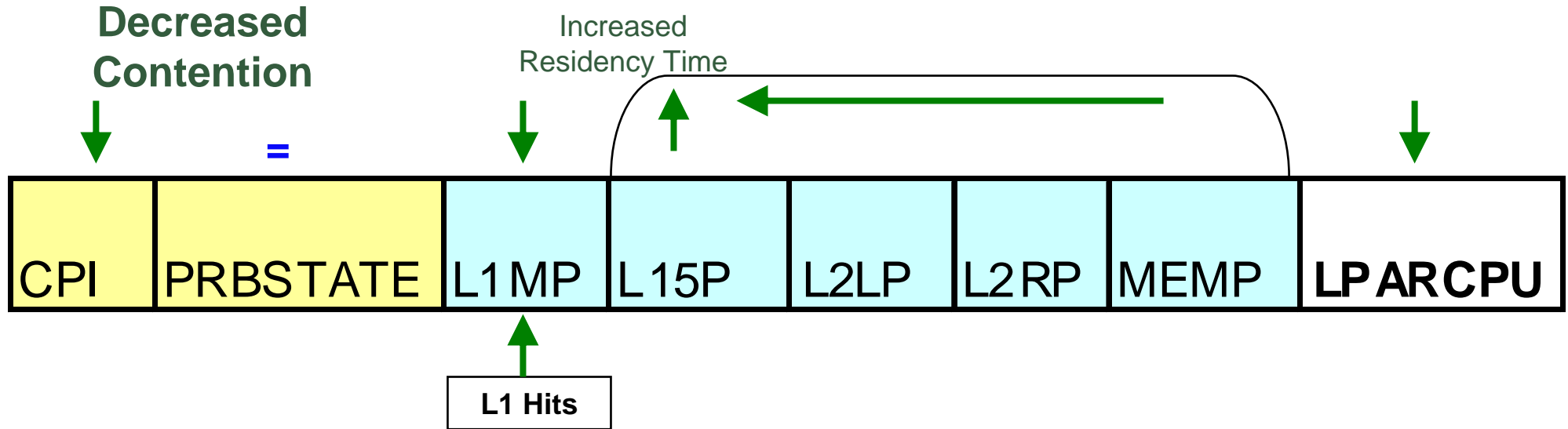
Information includes:

- Modify command used to enable CPU MF / HIS
- COUNTER sets enabled
- SMF interval

```
HIS015I 11.32.39 DISPLAY HIS 056
HIS      0025 ACTIVE
COMMAND: MODIFY HIS,B,TT='CMU MF COUNTERS ENABLED',CTRONLY,CTR=ALL,SI=
        SYNC
START TIME: 2010/09/23 09:52:22
END TIME:   ----/--/--  --:--:--
COMPLETION STATUS: -----
FILE PREFIX: SYSHIS20100923.095222.
COUNTER VERSION NUMBER 1: 1   COUNTER VERSION NUMBER 2: 2
COMMAND PARAMETER VALUES USED:
  TITLE=  CMU MF COUNTERS ENABLED
  PATH=   .
  COUNTER SET= BASIC, PROBLEM-STATE, CRYPTO-ACTIVITY, EXTENDED
  DURATION= NOLIMIT
  CTRONLY
  STATECHANGE= SAVE
  SMFINTVAL= SYNC
```

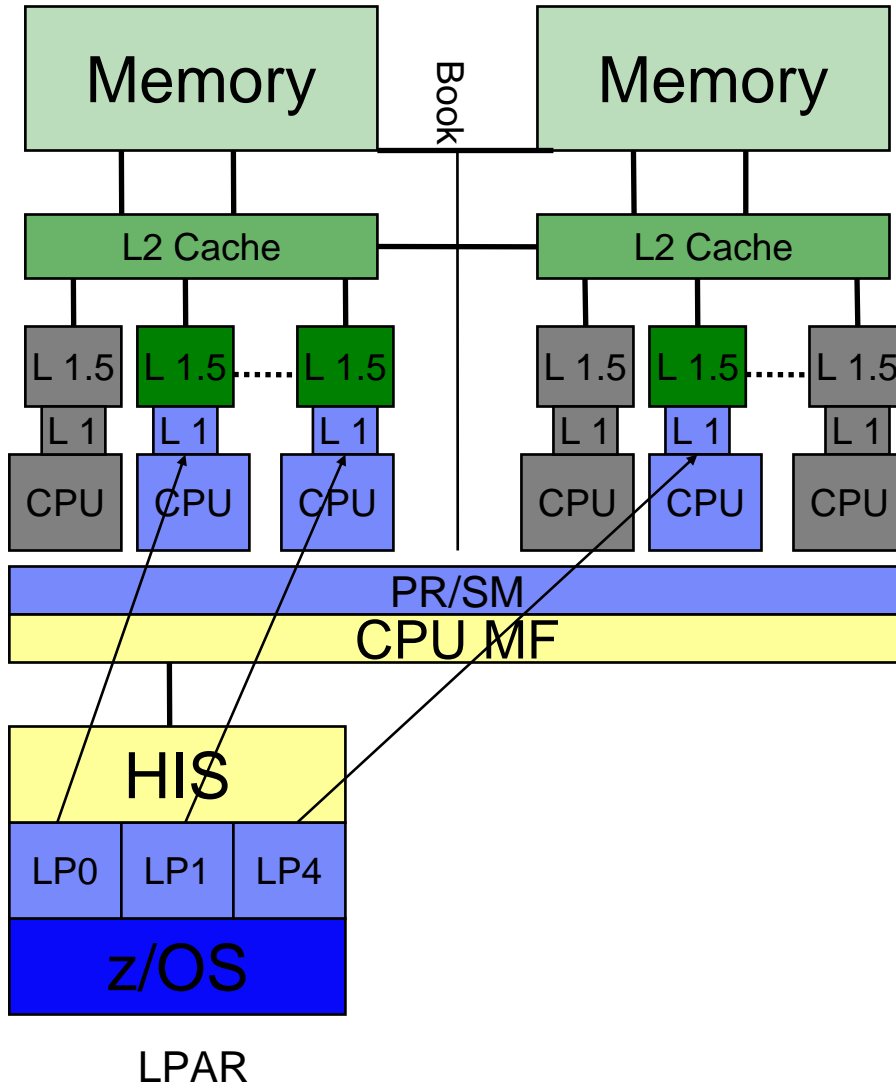

Old Experiences

CPU MF can help provide cache/memory resource change insights

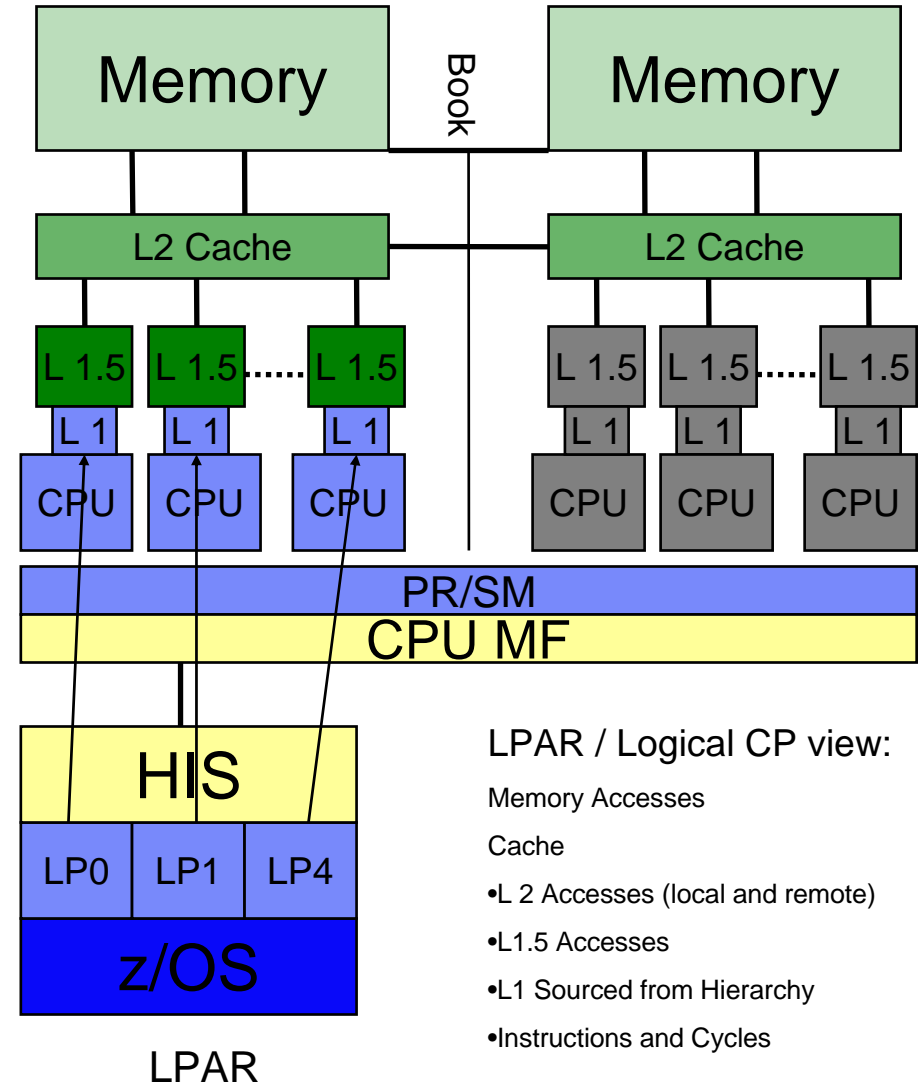


z10 HiperDispatch attempts to align Logical CPs with PUs in the same Book

HD=NO



HD=YES



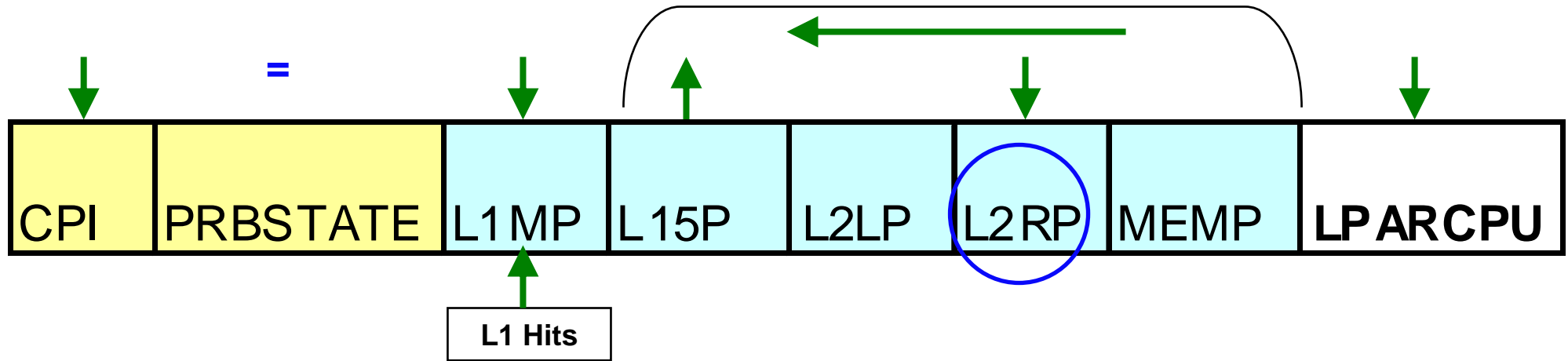
LPAR / Logical CP view:

Memory Accesses

Cache

- L 2 Accesses (local and remote)
- L1.5 Accesses
- L1 Sourced from Hierarchy
- Instructions and Cycles

From CPU MF, z10 HiperDispatch=YES May Decrease the L2 Remote %



CPI – Cycles per Instruction

PRBSTATE - % Problem State

L1MP – Level 1 Miss Per 100 instructions

L15P – % sourced from L1.5 cache

L2LP – % sourced from Level 2 Local cache (on same book)

L2RP – % sourced from Level 2 Remote cache (on different book)

MEMP - % sourced from Memory

LPARCPU - APPL% (GCPs, zAAPs, zIIPs) captured and uncaptured

Potential Workload Characterization
 z10 L1 sourcing from cache/memory hierarchy

HiperDispatch=Yes Customer Improvement on z10 721

	Day	Hour	CPI	Prb State	Est Instr Cmplx CPI	Est Finite CPI	Est SCPL1M	L1MP	L15P	L2LP	L2RP	MEMP	Rel Nest Intensity	LPARCPU	HD ?
	12	11.0	8.2	53.7	3.79	4.45	115	3.9	63.7	23.1	6.7	6.6	0.89	1745.1	No
	11	11.0	7.5	52.8	3.74	3.71	97	3.8	70.3	19.8	3.7	6.3	0.76	1632.3	Yes
HD=Yes % Improvement			1.10	1.02	1.01	1.20	1.19	1.01	0.91	1.17	1.80	1.05	1.17	1.07	

CPI – Cycles per Instruction

Prb State - % Problem State

Est Instr Cmplx CPI – Estimated Instruction Complexity CPI (infinite L1)

Est Finite CPI – Estimated CPI from Finite cache/memory

Est SCPL1M – Estimated Sourcing Cycles per Level 1 Miss

L1MP – Level 1 Miss per 100 instructions

L15P – % sourced from Level 2 cache

L2LP – % sourced from Level 2 Local cache (on same book)

L2RP – % sourced from Level 2 Remote cache (on different book)

MEMP - % sourced from Memory

Rel Nest Intensity – Reflects distribution and latency of sourcing from shared caches and memory

LPARCPU - APPL% (GCPs, zAAPs, zIIPs) captured and uncaptured

HiperDispatch=YES resulted in a ~ +10% improvement as measured by CPU MF.

Additional measurements over multiple days from traditional CPU/Transaction metrics should be used to validate HD=No Vs. Yes results

- Partition has 21 logical processors
- 2 additional partitions on the CEC

z196 Customer Benchmark - CICS Threadsafes Vs QR

SYSID	Mon	Day	SH	Hour	CPI	Prb State	Est Instr Cmplx CPI	Est Finite CPI	Est SCPL1M	L1MP	L2P	L3P	L4LP	L4RP	MEMP	Rel Nest Intensity	LPARCPU	Eff GHz	CICS
PAR1	SEP	1	P	16	7.0	31.8	2.8	4.2	50	8.5	69.3	19.0	7.6	3.4	0.7	0.46	1459.4	5.2	QR
PAR2	SEP	1	P	16	4.7	23.4	3.0	1.7	24	7.0	89.2	5.6	4.5	0.1	0.7	0.20	1546.2	5.2	Threadsafes

■ Benchmark Description:

- Comprises of CICS transactions and some Batch...
 - All Batch is heavy Update and running on both LPARs
 - The CICS transactions are cloned pairs. One group is left to run in QR mode and the other is marked threadsafes in the CICS PPT definition. This test Focused all the Quasi-Reentrant transactions in one LPAR and all the Threadsafes transactions in the other LPAR. Transaction concurrency was establish in order to drive the LPARs to 90%+ utilization levels.

■ Threadsafes Vs QR Results

- CICS 110s
 - Increase of 52% of transactions
 - Decrease of 42% in CPU per Transaction
 - Decrease of average response time by 67% (3.0x)
- RMF 72s – CICS Storage Class
 - Ended Transactions up 2.4x
 - Response Time down 3.6x
- SMF 113s – LPAR
 - CPI down 1.48x from 7.0 to 4.7
 - L1MP down by 1.5 from 8.5 to 7.0
 - L2P up 19.9% from 69.3% to 89.2%

CICS Threadsafes is an option that may help you reduce CPU cost for applicable transactions by reducing switches between different TCB types

CPU MF example to supplement CICS and RMF performance metrics

As a secondary data source to understand why performance may have changed

These numbers come from a synthetic Benchmark and do not represent a production workload

***** New - This is an evolving use of CPU MF *****

CPU MF can help measure the impact of 1 MB Pages in your environment

Test	CPI	PRBSTATE	Est Instr Cmplx	Est Finite CPI	Est SCPL1M	L1MP	L15P	L2LP	L2RP	MEMP	Rel Nest Intensity	LPARCPU	GHz	TLB1 Miss CPU% of Total CPU	TLB1 Cycles per Miss	PTE% of all TLB1 Misses
DB2 V10 4K PageFix=YES	4.46	1.29	2.63	1.83	26	7.13	94.72	4.64	0.01	0.63	0.09	28.2	4.4	16.0	83	19.2
DB2 V10 1MB PageFix=YES	4.26 1.05	1.13	2.58	1.68	23	7.25 0.98	96.56 0.98	3.03 1.53	0.01	0.41	0.06	33.9	4.4	15.6 1.03	65 1.28	13.7 1.40

- **DB2 10 for z/OS Beta provides ability to specify 1 MB Pages for DB2 Buffer Pools**
 - **1 MB Pages can help reduce TLB Page Table Entry Misses**
 - **CPU MF can be used to help measure the 1 MB Page impact for your environment**
 - DB2 10 for z/OS Beta Customer ran DB2 Batch job that exercised 4k and 1MB pages (PageFix=Yes). LFArea=40M
 - The batch job executed 30M Selects, 20M Inserts, and 10M Fetchs
 - *CPU MF showed the following – but this is not necessarily representative of 1 MB Page results*
 - 40% reduction in Page Table Entry % (PTE) of all TLB1 Misses
 - 28% reduction TLB1 Cycles per Miss, 3% reduction TLB1 Miss CPU% of Total CPU
 - Lower CPI and Nest Intensity
 - DB2 Accounting report showed 1.4 % reduction in CPU time
- Warning: These numbers come from a synthetic Benchmark and do not represent a production workload**
- **As you implement 1 MB Page exploiters, use CPU MF to help measure the impact**
 - **Measure it in its intended Production LPAR**
 - **See white paper “IBM System z10 Support for large pages”**
 - <http://www.research.ibm.com/journal/abstracts/rd/531/tzortzatos.html>



DB2 10 for z/OS Beta Customer – RMF for 1 MB Page

PAGING ACTIVITY

z/OS V1R10 SYSTEM ID SYSA START 06/30/2010-12.45.00 INTERVAL 000.15.00
 CONVERTED TO z/OS V1R11 RMF END 06/30/2010-13.00.01 CYCLE 5.000 SECONDS
 DPT = IEAOPTXX MODE = ESAME CENTRAL STORAGE MOVEMENT RATES - IN PAGES PER SECOND

HIGH UIC (AVG) = 65535 (MAX) = 65535 (MIN) = 65535
 WRITTEN TO READ FROM
 CENTRAL STOR CENTRAL STOR
 --- CENTRAL STORAGE FRAME COUNTS ---
 MIN MAX AVG
 HIPERSPACE RT 0.00 0.00 1 1 1
 PAGES
 VIO RT 0.00 0.00 0 0 0
 PAGES

FRAME AND SLOT COUNTS

	CENTRAL STORAGE		
	MIN	MAX	AVG
(15 SAMPLES)			
AVAILABLE	158,574	161,884	159,692
SQA	10,497	10,595	10,529
LPA	5,734	5,735	5,734
CSA	39,739	39,921	39,850
LSQA	15,186	15,198	15,193
REGIONS+SWA	539,686	542,913	541,822
TOTAL FRAMES	786,432	786,432	786,432

	FIXED FRAMES		
	MIN	MAX	AVG
NUCLEUS	2,608	2,608	2,608
SQA	9,636	9,733	9,668
LPA	94	94	94
CSA	1,550	1,550	1,550
LSQA	14,324	14,334	14,331
REGIONS+SWA	49,347	49,392	49,359
BELOW 16 MEG	77	77	77
BETWEEN 16M-2G	13,456	13,498	13,467
TOTAL FRAMES	77,568	77,680	77,611

STORAGE REQUEST RATES	
GETMAIN REQ	0
FRAMES BACKED	0
FIX REQ < 2 GB	0
FRAMES < 2 GB	0
REF FAULTS 1ST	0
NON-1ST	0

	LOCAL PAGE DATA SET SLOT COUNTS		
	MIN	MAX	AVG
AVAILABLE SLOTS	2,854,758	2,854,758	2,854,758
VIO SLOTS	0	0	0
NON-VIO SLOTS	757	757	757
BAD SLOTS	0	0	0
TOTAL SLOTS	2,855,515	2,855,515	2,855,515

SHARED FRAMES AND SLOTS			
CENTRAL STORAGE	6,428	6,557	6,489
FIXED TOTAL	98	98	98
FIXED BELOW 16 M	0	0	0
AUXILIARY SLOTS	0	0	0
TOTAL	8,389	8,518	8,450

MEMORY OBJECTS AND FRAMES			
OBJECTS COMMON	3	3	3
SHARED	6	6	6
LARGE	40	40	40
FRAMES COMMON	3,791	3,811	3,801
COMMON FIXED	0	0	0
SHARED	7,504	7,645	7,590
1 MB	40	40	40



Formulas – Additional TLB

Metric – z10	Calculation – <i>note all fields are deltas between intervals</i>
TLB1 CPU Miss % of Total CPU	$((E145+E146) / B0) * 100$
TLB1 Cycles per TLB Miss	$(E145+E146) / (E138+E139)$
PTE % of all TLB1 Misses	$(E140 / (E138+E139)) * 100$

Metric – z196	Calculation – <i>note all fields are deltas between intervals</i>
TLB1 CPU Miss % of Total CPU	$((E130+E131) / B0) * 100$
TLB1 Cycles per TLB Miss	$(E130+E131) / (E144+E145)$
PTE % of all TLB1 Misses	$(E146 / (E144+E145)) * 100$

Note these Formulas may change in the future

TLB1 CPU Miss % of Total CPU - TLB CPU % of Total CPU

TLB1 Cycles per TLB Miss – Cycles per TLB Miss

PTE % of all TLB1 Misses – Page Table Entry % misses

B* - Basic Counter Set - Counter Number

See “The Set-Program-Parameter and CPU-Measurement Facilities” SA23-2260-0 for full description

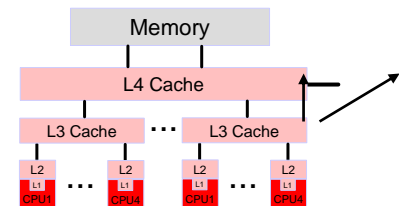
E* - Extended Counters - Counter Number

See “IBM The CPU-Measurement Facility Extended Counters Definition for z10” SA23-2261-0 for full description or “The CPU-Measurement Facility Extended Counters Definition for z10 and z196” SA23-2261-01 for full description

Appendix

CPU MF – Lessons Learned since August 2010

- **CPU MF Metrics continue to help understand why performance changed**
 - LPAR Configuration Changes including – HD=Yes/No
 - CICS QR Vs Threadsafe
 - 1 MB Vs 4k Pages
 - GHz measurement for State Changes including Power Savings Mode
- **Customers continue to successfully run CPU MF COUNTERS**
 - Over days/months without any reported performance impact, Turning on and leaving on
 - Volunteer Feedback: easy to enable, minimal time investment
 - For long term gathering on z196s recommend CTR=(B,E)
- **Update to z196 formulas for Estimated SCPL1M and Estimated Finite CPI – see slide 27**
- **z196**
 - Ensure HD=YES
 - L3 off chip and off book sourced from respective L4s - see formulas on slide 26



CPU MF – Lessons Learned since March 2010

- **CPU MF Performance Metrics continues to help understand why performance changed**
 - LPAR Configuration Changes including
 - HD= Yes/No
 - 1 MB Vs 4k Pages
 - GHz measurement for State Changes

- **Customers continue to successfully run CPU MF COUNTERS collecting SMF 113s**
 - Over days/months without any reported performance impact, Turning on and leaving on
 - Volunteer Feedback: easy to enable, minimal time investment

- **SMF 113 Logical CPU IDs are equal to the SMF 70 Logical CPU IDs**
 - Directly identifies GCPs, zIIPs or zAAPs in SMF 113s with [APAR OA30486](#) for z10s and z196

- **LPAR Management Time is NOT included in LPARCPU time (SMF 113 Cycles)**

- **Utilize the Counter Version Number fields to map to technology**
 - SMF113_2_CTRVN2 – Crypto or Extended counter sets = “2” for z196 “1” for z10

- **z/VM CPU MF – native prototype in process**

- **D HIS command provides useful status information**

CPU MF – Lessons Learned since August 2009

- **CPU MF Performance Metrics can be used to help understand why performance changed**
- **Customers are successfully running CPU MF COUNTERS collecting SMF 113s**
 - Over days and months without any reported performance impact
 - Feedback from Volunteers is this is very easy to enable, with a minimal time investment
- **SMF 113 Logical CPU IDs are equal to the SMF 70 Logical CPU IDs**
 - Can match up SMF 113s & SMF 70s to identify GCPs, zIIPs or zAAPs
 - Can see the unique Vertical Polarity Logical CPs cache/memory characteristics
 - E.G. Vertical Mediums may have higher L2 Remote activity
- **In multi-book z10 ECs there can be L2 Remote Activity even if ≤ 12 GCPs**
 - Because of I/O activity from SAPs as the data is initially stored in the Remote L2
- **Utilize the Counter Version Number fields to map to technology**
 - Number is increased for a change in meaning or number of counters
 - SMF113_2_CTRVN1 – Basic or Problem-State counter sets
 - SMF113_2_CTRVN2 – Crypto or Extended counter sets

CPU MF Update – Lessons Learned since March 2009

- L1 Miss per 100 instructions can be determined from CPU MF COUNTERS
- z10 EC must be at bundle #20 or higher for CPU MF COUNTERS
- IRD considerations
 - If CPU goes offline, only activity within internal is recorded in an Intermediate record, then
 - If no activity in follow on 15 minute interval(s), Intermediate record is not cut for the CPUID
 - No Final record when HIS is ended
 - When activity resumes, Intermediate record is written for CPUID
- New **APAR OA27623** to add “CPU Speed” to SMF 113 and to HIS COUNTERS output
 - *Processor speed for which the hardware event counters are recorded. Speed is in cycles / microsecond* - “4404” for z10 EC
 - SMF 113 new field: SMF113_2_CPSP - 4 byte binary
 - Simplifies conversion of Cycles into “Time”
- Customers are successfully running CPU MF COUNTERS (and collecting SMF 113s) over 24 hours
- Analyze the “major” LPARs on a z10 at the same time

Documentation

- *MVS Commands SA22-7627-19*
 - Setting up hardware event data collection 1-39
- *The Set-Program-Parameter and CPU-Measurement Facilities SA23-2260-0*
 - Full description of Basic, Problem-State and Crypto Counter Sets
- *IBM The CPU-Measurement Facility Extended Counters Definition for z10 SA23-2261-0*
- *IBM The CPU-Measurement Facility Extended Counters Definition for z10 and z196 SA23-2261-01*
- *WSC Short Stories and Tall Tales*
 - SHARE Summer 2009 Denver - Session 2136 – John Burg
- *CPU MF Overview and WSC Experiences Techdoc TC000041 available March 26 2010*
 - <http://www.ibm.com/support/techdocs/atmastr.nsf/WebIndex/TC000041>
 - SHARE Winter 2010 presentation and detailed write up for enabling CPU MF – John Burg
- *ITSO Red Book reference Planned for 4QT 2010*
 - *Exploiting System z LPAR Capacity Controls - SG24-7846. 2 Part Book:*
 - Part 1 - CPU MF
 - Part 2 - HiperDispatch, Group Capacity Controls, hard/soft capping
 - Draft available ~April 1
 - <http://www.redbooks.ibm.com/redbooks.nsf/home?ReadForm&page=drafts>