

The Relatively New LSPR and zEC12 Performance Brief

SHARE San Francisco

EWCP

Gary King
IBM

February 6, 2013

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.

Topics

- What's "Relatively New" in the LSPR
 - ▶ and the theory and analysis behind it
- Performance drivers with zEC12
- zEC12 ITR Ratios
- Workload Variability

LSPR: Performance Showcase for z Processors

- IBM System z provides capacity comparisons among processors based on a variety of measured workloads which are published in the Large System Performance Reference (LSPR)
 - ▶ <https://www-304.ibm.com/servers/resourceLink/lib03060.nsf/pages/lspindex>
- Old and new processors are measured in the same environment with the same workloads at high utilizations
- Over time, workloads and environment are updated to stay current with customer profiles
 - ▶ old processors measured with new workloads/environment may have different average capacity ratios compared to when they were originally measured
- LSPR presents capacity ratios among processors
- Single number metrics MIPS, MSUs, and SRM Constants
 - ▶ based on the ratios for
 - the "average" workload
 - the "median" customer LPAR configuration

LSPR RNI-based Workload Categories

Validated and now zPCR default

- Historically, LSPR workload capacity curves (primitives and mixes) 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 starting with z10, the ability to gain insight into the interaction of workload and hardware exists.
- The LPSR for z196 introduced three new workload categories which replaced all prior primitives and mixes.
 - ▶ LOW, AVERAGE, HIGH Relative Nest Intensity
 - ▶ treated as a workload "hint" in zPCR
- Migration to z196 has validated this approach
 - ▶ detailed study of 16 customers and 75 LPARs
- RNI-based methodology for workload matching is now the default in zPCR

Fundamental Components of Workload Capacity Performance

Part 1

- Instruction Path Length for a transaction or job
 - ▶ Application dependent, of course
 - ▶ Can also be sensitive to Nway (due to MP effects such as locking, work queue searches, etc)
 - ▶ But generally doesn't change much on moves between processors of similar capacity and/or Nway
- 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, does not change very much

Fundamental Components of Workload Capacity Performance Part 2

- 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
 - ▶ **Relative Nest Intensity**
 - Activity beyond the private cache(s) is the most sensitive area
 - due to larger latencies involved
 - Reflects activity distribution and latency to chip-level caches, book-level caches and memory
 - Level 1 cache miss percentage also important
 - Data for calculation available from CPU MF (SMF 113) starting with z10

z196 versus z10 hardware comparison

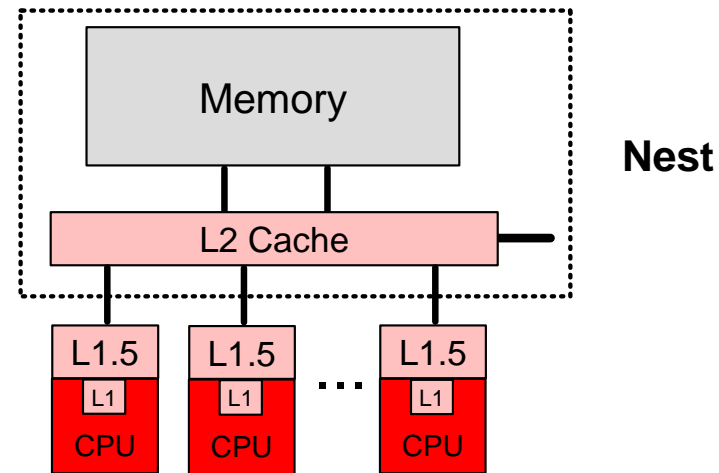
■ z10 EC

▶ CPU

- 4.4 GHz

▶ Caches

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



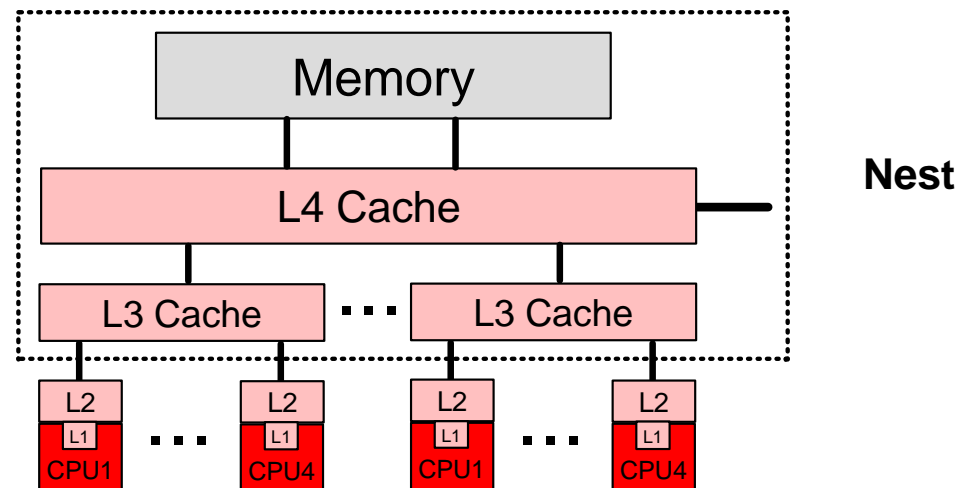
■ z196

▶ CPU

- 5.2 GHz
- Out-Of-Order execution

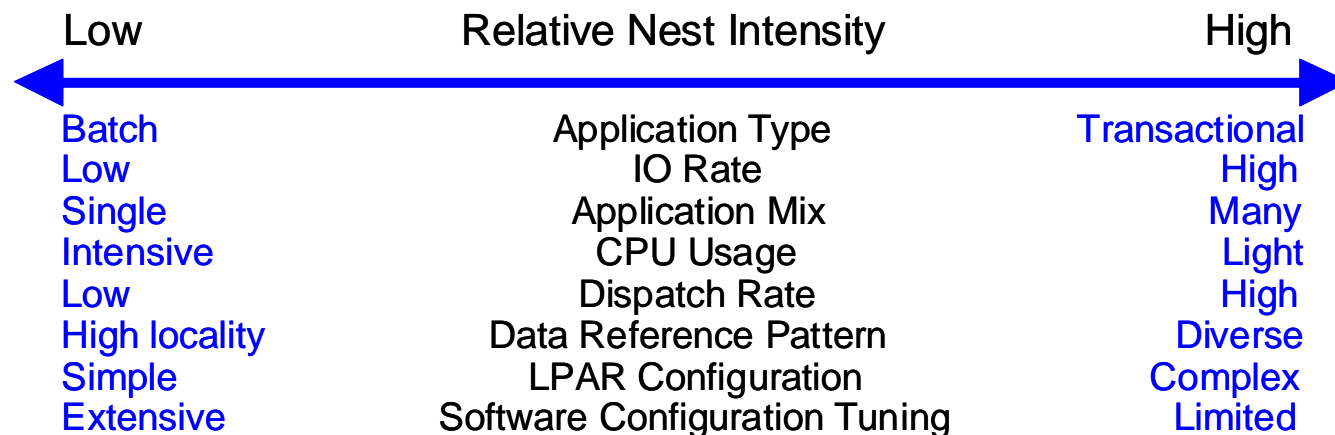
▶ Caches

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



The Most Influential Factor Underlying Workload Capacity Curves is Relative Nest Intensity (RNI)

- Many factors influence a workload's capacity curve
- However, what they are actually affecting is the workload's RNI
- It is the net effect of the interaction of all these factors that determines the capacity curve
- The chart below indicates the trend of the effect of each factor but is not absolute
 - ▶ for example, some batch will have high RNI while some transactional workloads will have low
 - ▶ for example, some low IO rate workloads will have high RNI, while some high IO rates will have low



LSPR Workload Categories

- Categories developed to match the profile of data gathered on customer systems
 - ▶ over 100 data points (LPARs) used in the profiling
- Various combinations of prior workload primitives are measured on which the new workload categories are based
 - ▶ Applications include CICS, DB2, IMS, OSAM, VSAM, WebSphere, COBOL, utilities
- **LOW** (relative nest intensity)
 - ▶ Workload curve representing light use of the memory hierarchy
 - ▶ Similar to past high Nway scaling workload primitives
- **AVERAGE** (relative nest intensity)
 - ▶ Workload curve expected to represent the majority of customer workloads
 - ▶ Similar to the past LoIo-mix curve
- **HIGH** (relative nest intensity)
 - ▶ Workload curve representing heavy use of the memory hierarchy
 - ▶ Similar to the past DI-mix curve
- zPCR extends these published categories
 - ▶ Low-Avg
 - 50% LOW and 50% AVERAGE
 - ▶ Avg-High
 - 50% AVERAGE and 50% HIGH

CPU MF

- What is CPU MF?
 - ▶ A z10 GA2 and later facility that provides 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?
 - ▶ They provide the basis for the LSPR workload categories
 - ▶ zPCR can automatically process CPU MF data to provide a workload match based on RNI

- Reference John Burg's CPU MF presentation at SHARE
 - ▶ This afternoon, February 6, 1:30-2:30

z196 versus z10 hardware comparison

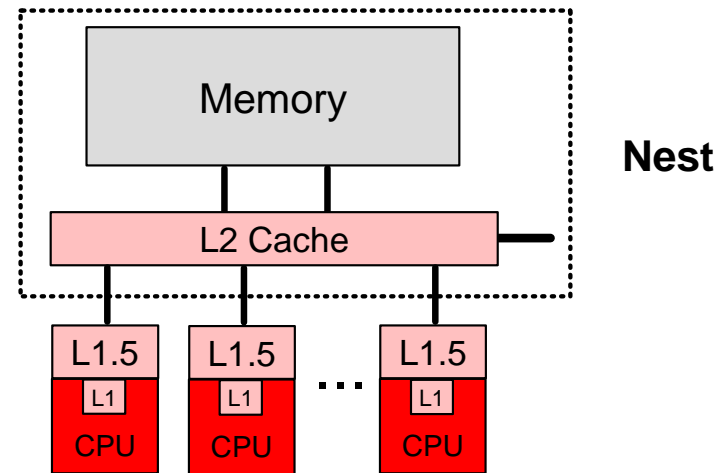
■ z10 EC

▶ CPU

- 4.4 GHz

▶ Caches

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



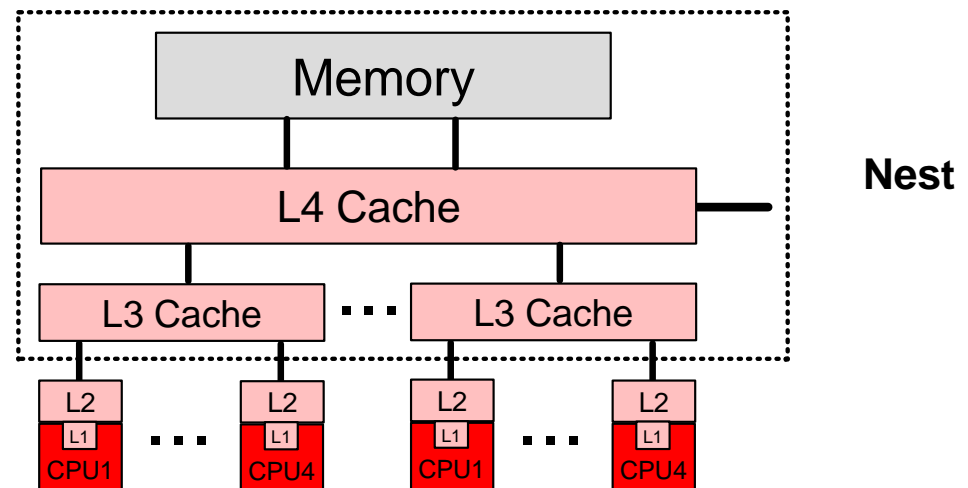
■ z196

▶ 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



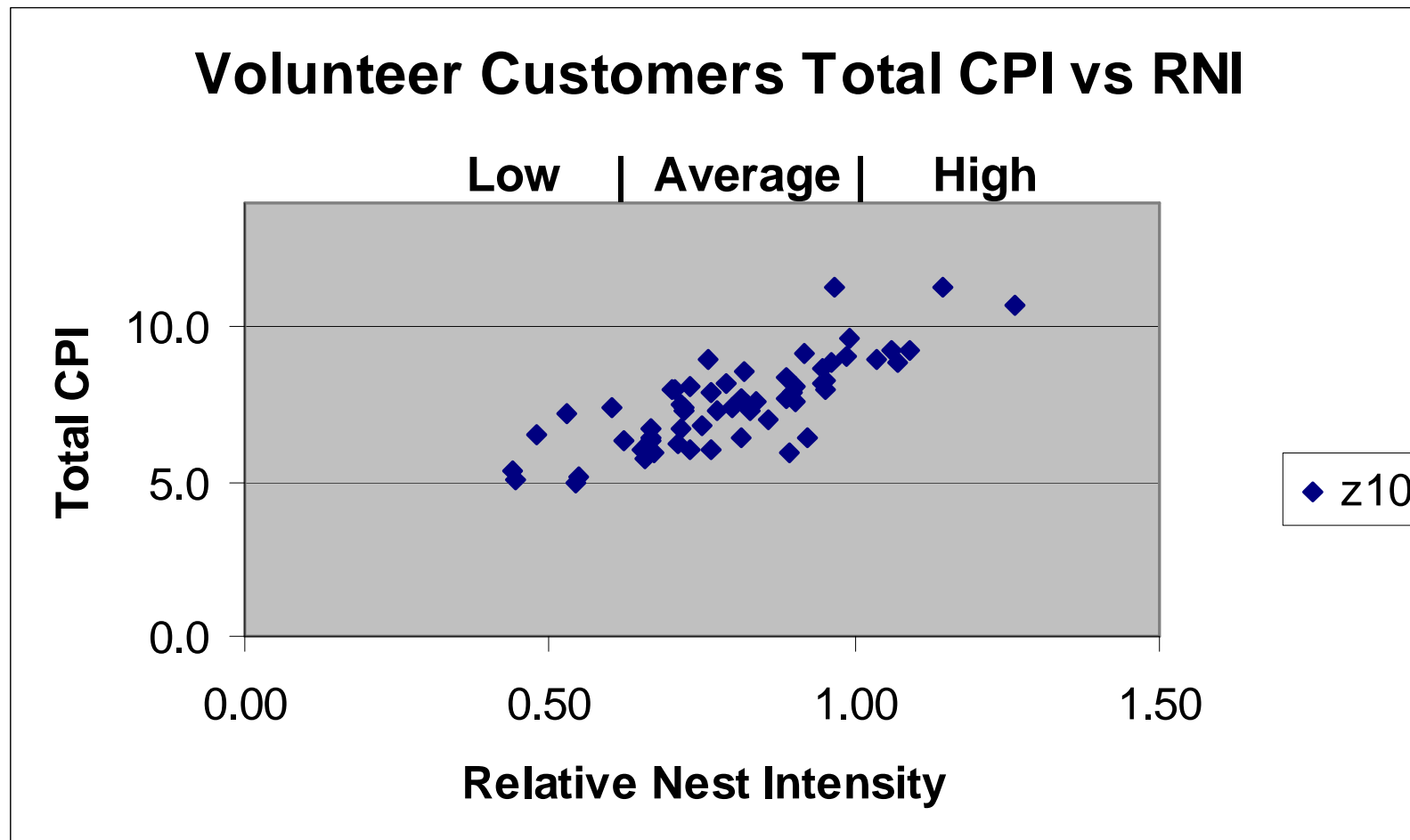
z10 CPU MF Memory Hierarchy Counters and Workload Characterization Stats

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		72	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

- 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

CPU MF

z10 Customer Workload Characterization Summary



RNI-based LSPR Workload Decision Table

L1MP	RNI	LSPR Workload Match
<3	≥ 0.75	AVERAGE
	< 0.75	LOW
3 to 6	>1.0	HIGH
	0.6 to 1.0	AVERAGE
	< 0.6	LOW
>6	≥ 0.75	HIGH
	< 0.75	AVERAGE

Notes: Applies to z10, z196 and zEC12 CPU MF data
Table may change based on feedback

What's new in the LSPR for zEC12

- Workload updates
 - ▶ upleveled software - z/OS 1.13, subsystems, compilers
 - ▶ minor tweaks to hardware-characteristic-based workload categories
 - based on CPU MF data from customers' z10s and z196s
 - updated ratios for LOW, HIGH; no change for AVERAGE
- HiperDispatch continues to be turned on for all measurements
 - ▶ particularly valuable on smaller z196 and zEC12 configurations due to sensitivity to L3 chip-level cache
- LSPR will continue to publish only the multi-image table
 - ▶ multi-image (MI) table
 - median LPAR configuration for each model based on customer profile
 - including effect of average number of ICFs and IFLs
 - most representative for vast majority of customers
 - basis for single-number metrics MIPS, MSUs, SRM constants
 - ▶ single-image (SI) table
 - no longer published starting with z/OS 1.11 LSPR
 - avoid confusion
 - continues to be used in zPCR

Median LPAR Configuration Profiles for the Multi-image Table

- Total number of z/OS images
 - ▶ 5 images at low-end models to 9 images at high-end
- Number of major images (>20% weight each)
 - ▶ 2 images across full range of models
- Size of images
 - ▶ low- to mid-range models have at least one image close to Nway of model
 - ▶ high-end models generally have largest image well below Nway of model
 - these models tend to be used for consolidation
- Logical to physical CP ratio
 - ▶ low-end near 5-1
 - ▶ most of the range 2-1
 - ▶ high-end near 1.3-1
- Book configuration
 - ▶ 1 "extra" book beyond what is needed to contain CPs
- ICFs/IFLs
 - ▶ 3 ICFs/IFLs

Using the LSPR z/OS V1R13 Tables

- For the most accurate capacity sizing ...
 - ▶ use zPCR customized LPAR configuration planning function
 - should always be used for final configuration planning for any upgrade
- LSPR tables may be used for high level capacity comparisons
 - ▶ Multi-image table represents average LPAR configuration and is the basis for all single-number metrics
- Tables at the LSPR website and those in zPCR will have slight differences
 - ▶ Precision
 - LSPR rounded to two digits to right of decimal point
 - zPCR carries maximum significant digits internally (displayed result is rounded to show 5 significant digits for the largest processor)
 - ▶ Reference (base) processor
 - LSPR fixed at *2094-701*
 - zPCR chosen by *you* (the user)

Performance Drivers with zEC12

■ Hardware

▶ memory subsystem

- continued focus on keeping data "closer" to the processor unit
- larger L2 cache
- double size of both chip-level and book-level shared caches

▶ processor

– enhanced Out-Of-Order execution

- better branch prediction
 - ◆ added BTB2 (Branch Target Buffer)
- add another execution unit pair
- improved grouping of instructions (pipeline)

– up to 6 processor units per chip

- ▶ up to 101 configurable processor units
- ▶ 4 different uni speeds

■ HiperDispatch

- ▶ exploits new chip configuration

zEC12 versus z196 hardware comparison

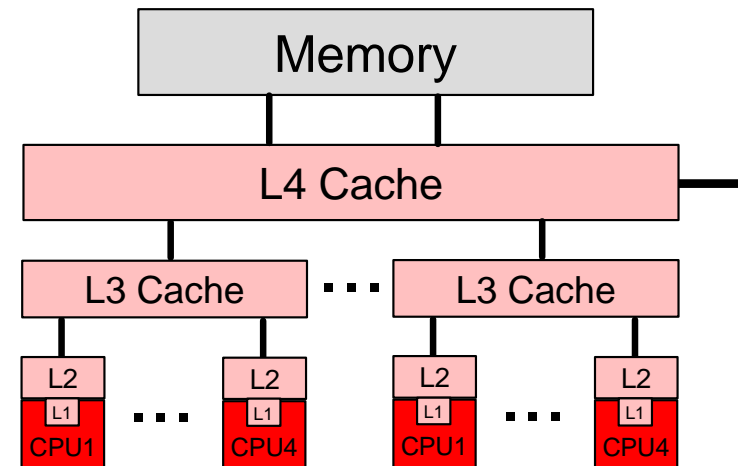
■ z196

▶ CPU

- 5.2 GHz
- Out-Of-Order execution
- Up to 4 PUs per chip

▶ Caches

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



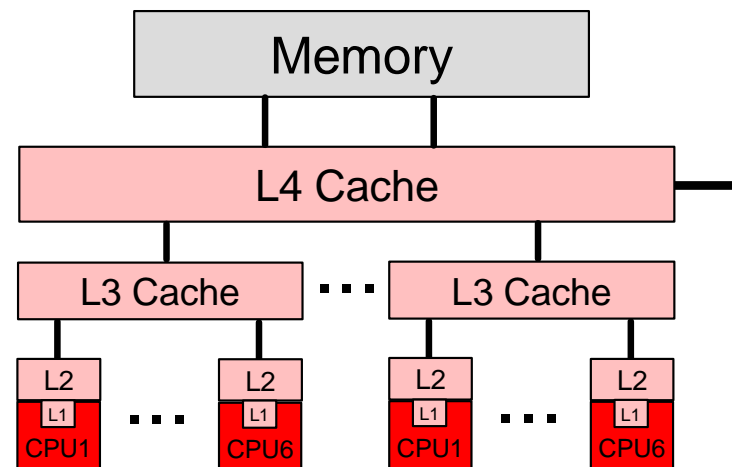
■ zEC12

▶ CPU

- 5.5 GHz
- Enhanced Out-Of-Order
- Up to 6 PUs per chip

▶ Caches

- L1 private 64k i, 96k d
- L2 private 1 MB i + 1 MB d
- L3 shared 48 MB / chip
- L4 shared 384 MB / book



LSPR website z/OS V1R13 Tables

zEC12 versus z196

Multi Image Table

	z/OS V1R13 AVERAGE	z/OS V1R13 AVERAGE	z/OS V1R13 AVERAGE	z/OS V1R13 AVERAGE
	z196 ITR	zEC12 ITR	zEC12:z196 ratio	zEC12 PCI
701	2.15	2.70	1.26	1514
708	14.42	17.98	1.25	10063
716	25.67	31.98	1.25	17904
732	45.09	55.85	1.24	31266
764	80.30	98.70	1.23	55251
780	93.40	117.71	1.26	65895
zEC12 7A1 vs z196 780	93.40	140.10	1.50	78426

zEC12 includes 3 subcapacity offerings

Subcapacity Offerings vs Full Speed

zEC12	z/OS V1R13 MI AVG ITRR	Ratio to 701	PCI	Max #CPs
701	2.70	1.00	1514	101
601	1.69	.63	947	20
501	1.13	.42	631	20
401	.43	.16	240	20

Notes: Uni speeds range from 16% to 63% of full speed uni
Each subcapacity offering has a maximum of 20 CPs

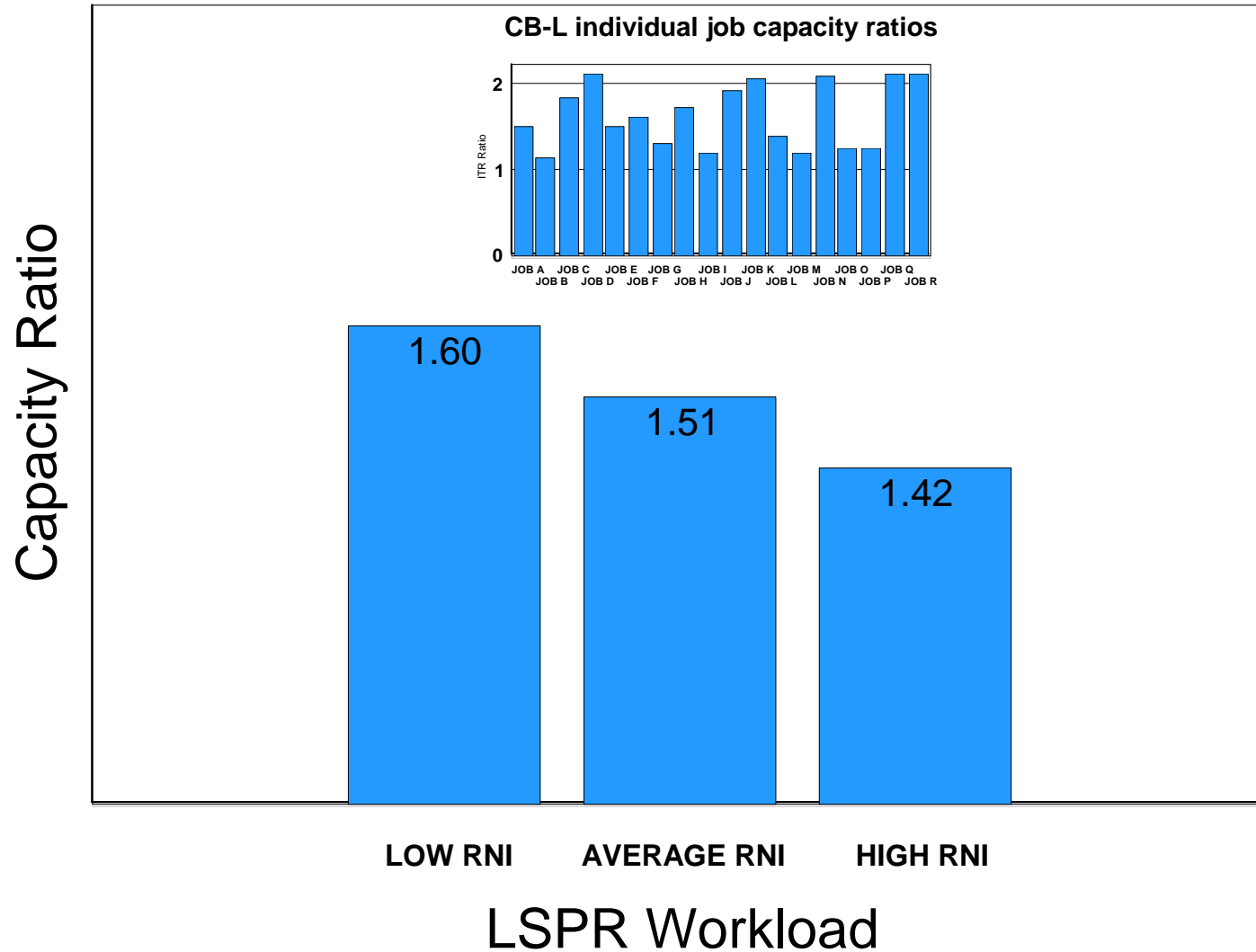
Workload Variability with zEC12

- Performance variability is generally related to fast clock speed and physics
 - ▶ increasing memory hierarchy latencies relative to micro-processor speed
 - ▶ increasing sensitivity to frequency of "missing" each level of processor cache
 - ▶ workload characteristics are determining factor, not application type
- zEC12 has fairly balanced improvements to both micro-processor and memory subsystem
 - ▶ workloads moving from z196 to zEC12 expected to have less variation than last few migrations
 - ▶ moves from z10 and earlier machines expected to have similar variation to moves to z196
- Examples of z9 to z10 and z10 to z196 and z196 to zEC12 on next several slides

LSPR Single Image Capacity Ratios

10way: z10 EC versus z9 EC

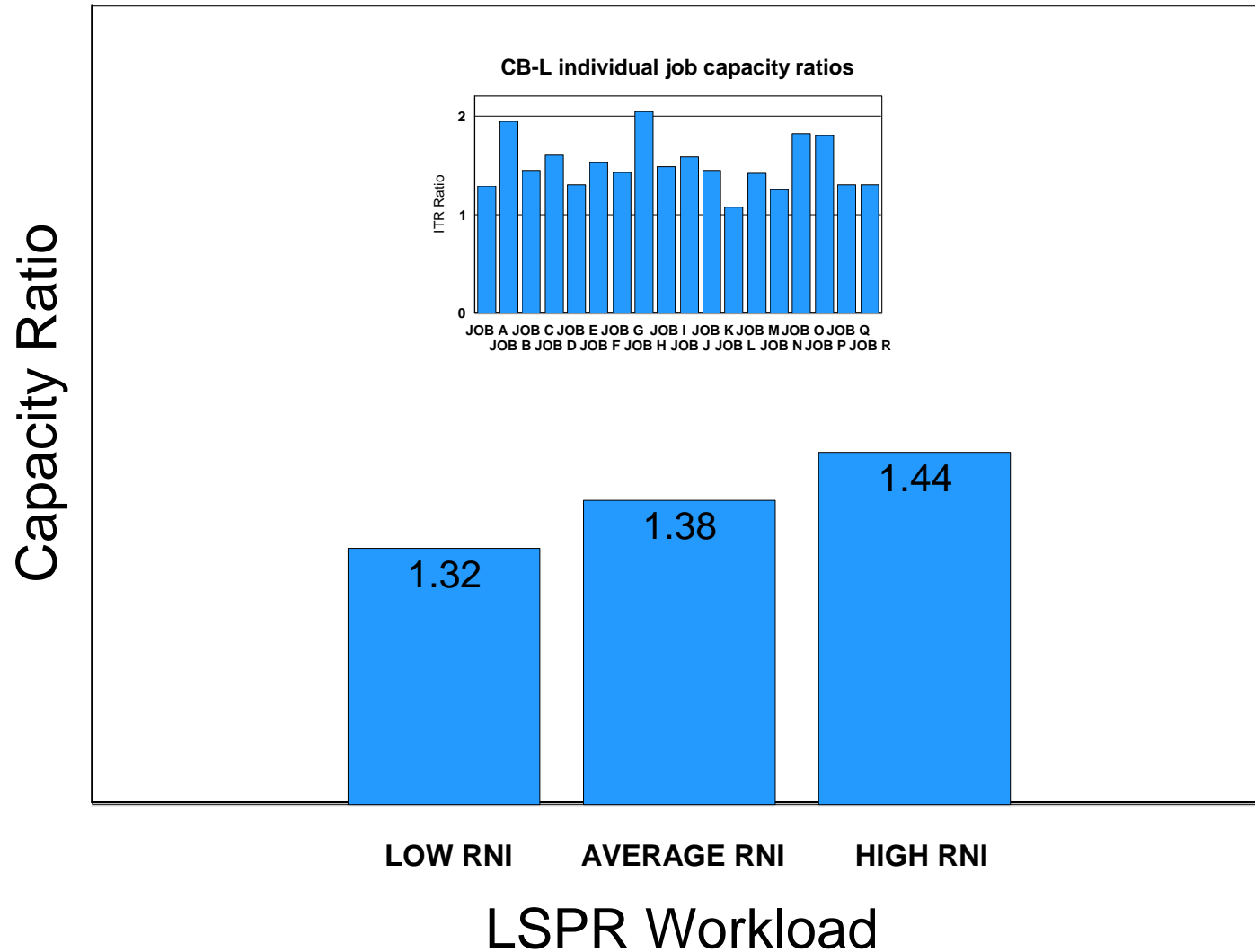
Example of Workload Variability



LSPR Single Image Capacity Ratios

10way: z196 versus z10 EC

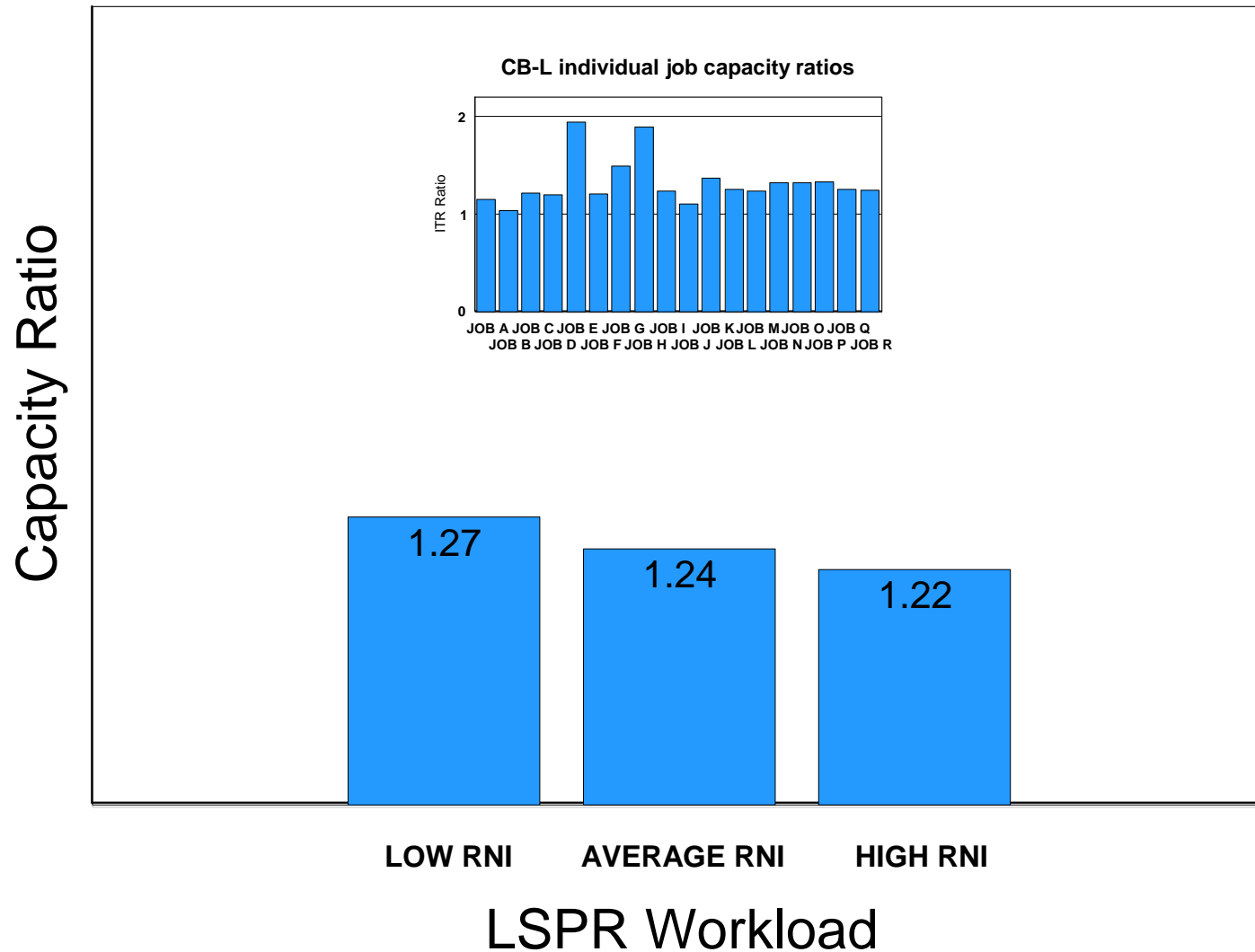
Example of Workload Variability



LSPR Single Image Capacity Ratios

16way: zEC12 versus z196

Example of Workload Variability



Summary

- "Relatively New" RNI-based LSPR
 - ▶ Validated and now default in zPCR

- zEC12
 - ▶ approximately 25% faster engines vs z196
 - ▶ max config provides approximately 50% more capacity vs z196
 - ▶ variability somewhat less than past few generations