The Relatively New LSPR and The IBM z13 Performance Brief

SHARE Seattle 16814

EWCP

Gary King
IBM

March 3, 2015

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*
* Registered trademarks of IBM Corporation	Processor Resource/Systems Manager		

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 z13
- z13 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/lsprindex
- 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
  - originally treated as a workload "hint" in zPCR
- Migrations to z196 and zEC12 have validated this approach
  - detailed study of 16 customers and 75 LPARs for each of the migration scenarios of z10 to z196 and z196 to zEC12
- RNI-based methodology for workload matching is now the default in zPCR

Fundamental Components of Workload Capacity Performance Part 1

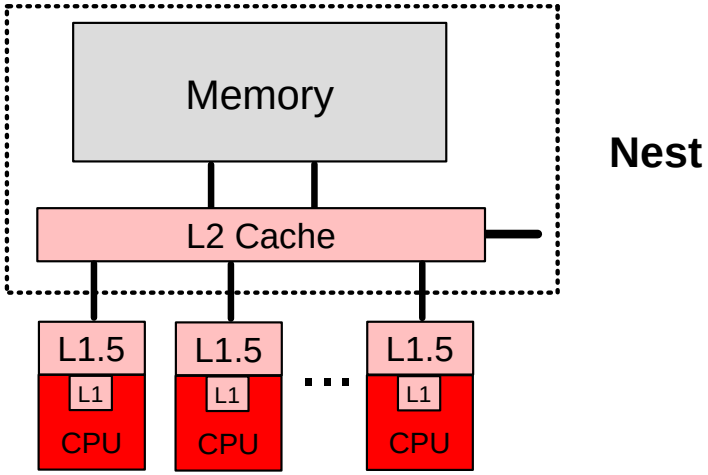
- Instruction Path Length for a transaction or job
  - Application dependent, of course
  - Can also be sensitive to Nway (due to MP effects such as locking, work queue searches, etc)
  - But generally doesn't change much on moves between processors of similar capacity and/or Nway
- Instruction Complexity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

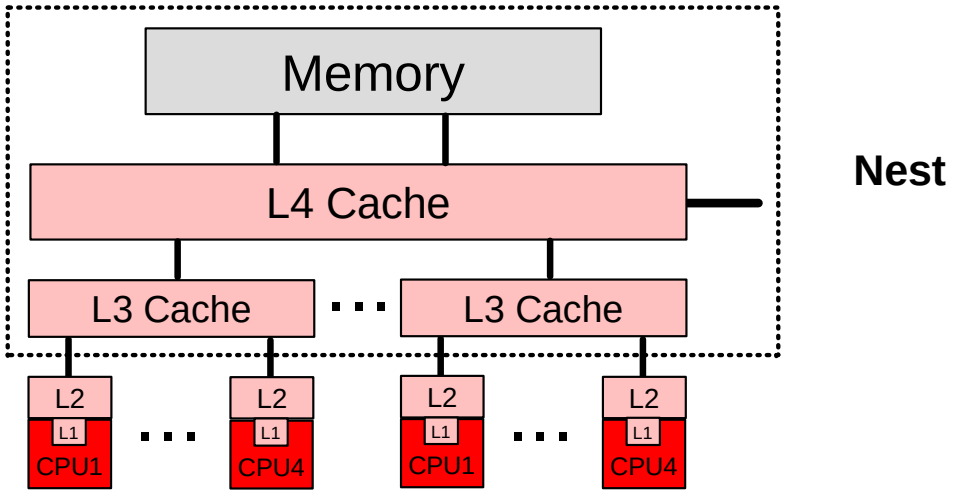
- Memory Hierarchy or "nest"
  - Many design alternatives
    - cache (levels, size, private, shared, latency, MESI protocol), controller, data buses
  - Workload effect
    - Quite variable
    - Sensitive to many factors: locality of reference, dispatch rate, IO rate, competition with other applications 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 cacluation 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

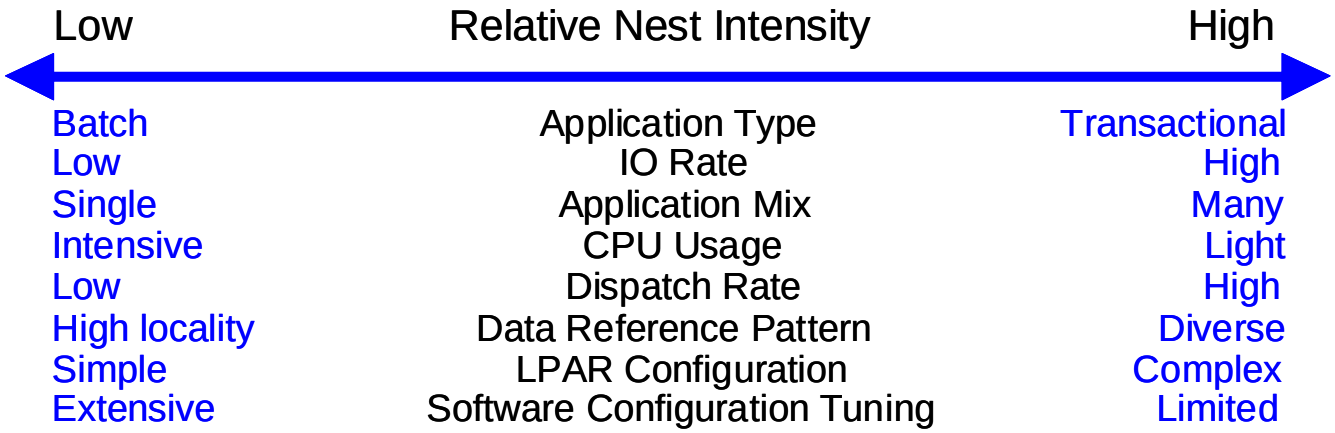


- z196
  - CPU
    - 5.2 GHz
    - Out-Of-Order execution
  - Caches
    - L1 private 64k i, 128k d
    - L2 private 1.5 MB
    - L3 shared 24 MB / chip
    - L4 shared 192 MB / book
    - book interconnect: star



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

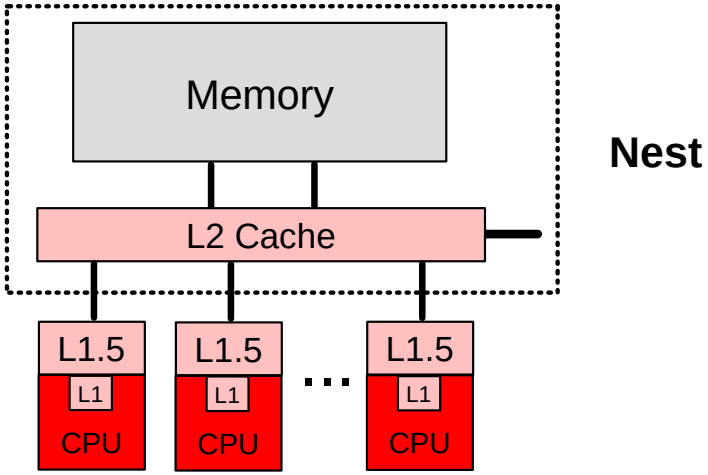
- Categories developed to match the profile of data gathered on customer systems
  - over 100 data points (LPARs) used in the profiling
- Various combinations of prior workload primitives are measured on which the new workload categories are based
  - Applications include CICS, DB2, IMS, OSAM, VSAM, 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

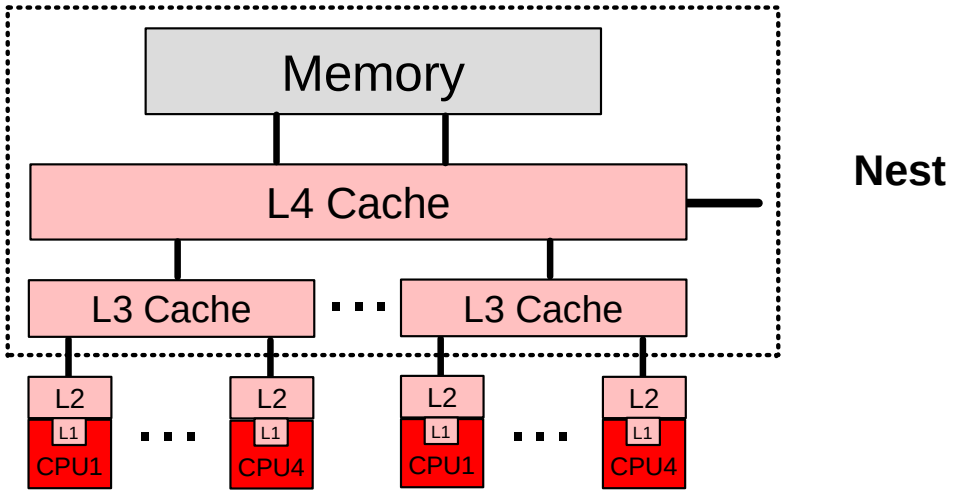
- What is CPU MF?
  - A z10 GA2 and later facility that provides memory hierarchy COUNTERS
  - Also capable of time-in-Csect type SAMPLES
  - Data gathering controlled through z/OS HIS (HW Instrumentation Services)
    - Collected on an LPAR basis
    - Written to SMF 113 records
    - Minimal overhead
- 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 processes CPU MF data to provide a workload match based on RNI
- Reference John Burg's CPU MF presentation at SHARE
  - March 3, 10:00-11:00

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

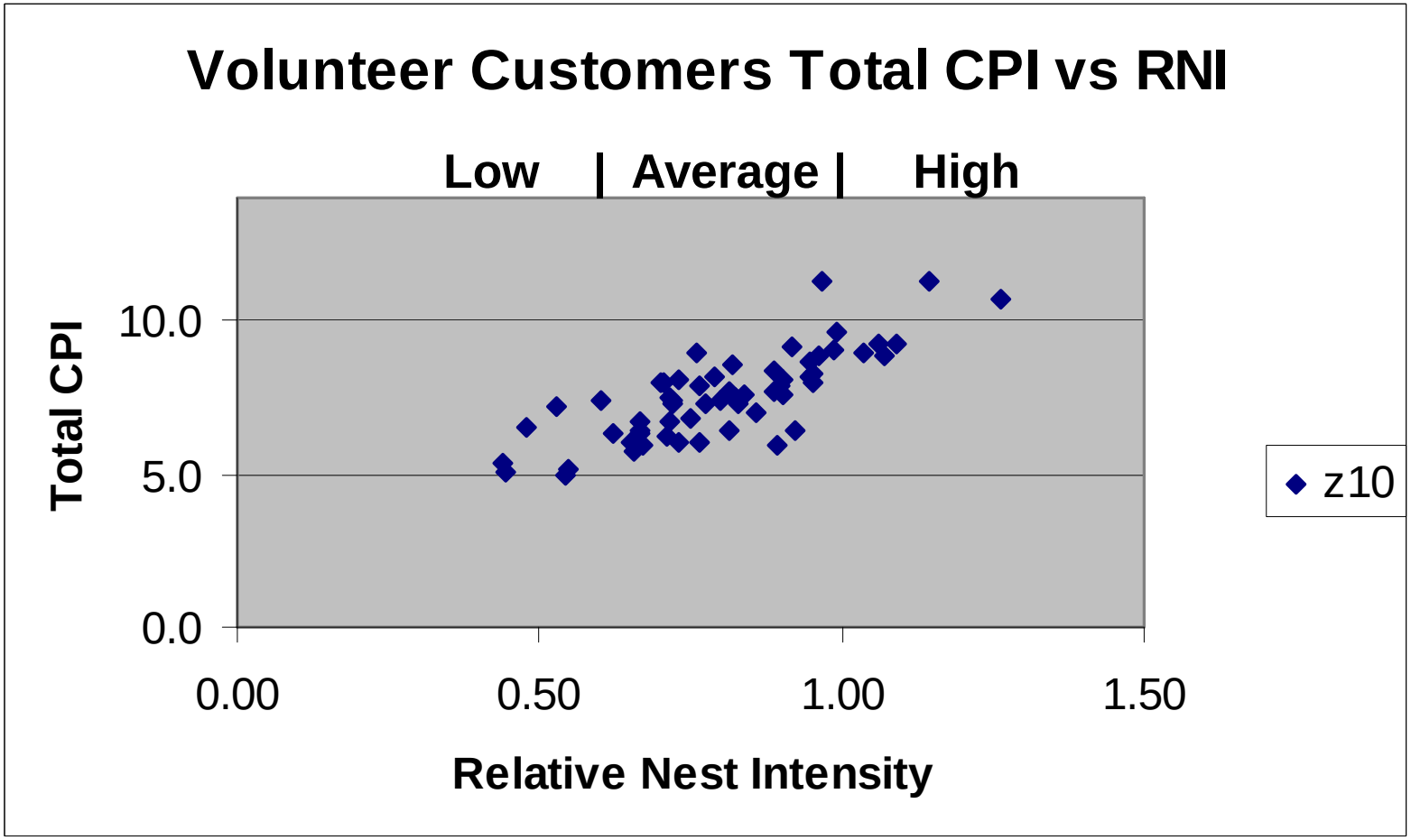


z10 CPU MF Memory Hierarchy Counters and Workload Characterization Stats

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		**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

- 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 < 0.75	AVERAGE LOW
3 to 6	>1.0 0.6 to 1.0 < 0.6	HIGH AVERAGE LOW
>6	>=0.75 < 0.75	HIGH AVERAGE

Notes: Applies to all processors z10 and later
Table may change based on feedback

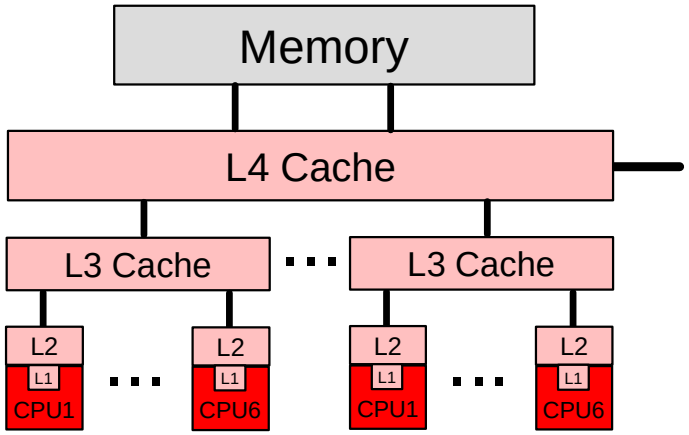
Performance Drivers with z13

- Hardware
  - memory subsystem
    - continued focus on keeping data "closer" to the processor unit
      - larger L1, L2, L3, L4 caches
      - improved IPC (Instructions Per Cycle)
    - 3x configurable memory
  - processor
    - 2x instruction pipe width, re-optimized pipe depth for power/performance
      - improved IPC
    - SMT for zIIPs and IFLs
      - includes metering for capacity, utilization and adjusted chargeback (zIIPs)
    - SIMD unit for analytics
    - up to 8 processor units per chip
  - up to 141 configurable processor units
  - 4 different uni speeds
- HiperDispatch
  - exploits new chip configuration
  - required for SMT on zIIPs
- PR/SM
  - 85 customer partitions (up from 60)
  - memory affinity
    - keep LPAR's CPs and memory local to drawer as much as possible

z13 versus zEC12 hardware comparison

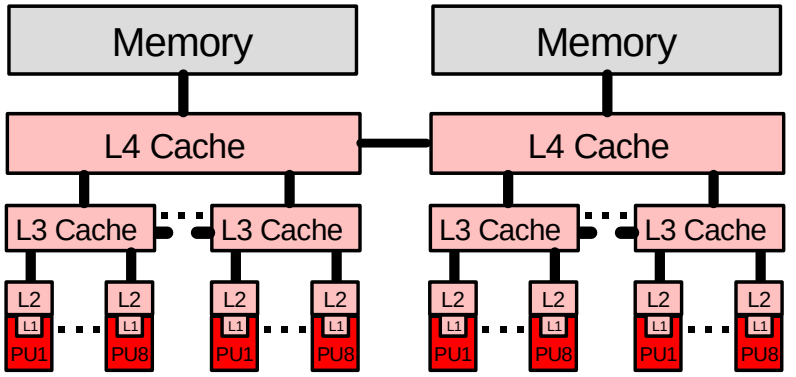
- zEC12
  - CPU
    - 5.5 GHz
    - Enhanced Out-Of-Order
  - Caches
    - L1 private 64k i, 96k d
    - L2 private 1 MB i + 1 MB d
    - L3 shared 48 MB / chip
    - L4 shared 384 MB / book

Single Book View



- z13
  - CPU
    - 5.0 GHz
    - Major pipeline enhancements
  - 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

Single Drawer View



z13 Capacity Performance Highlights

- Full speed capacity models ... capacity ratio to zEC12
  - average 1.10x at equal Nway
  - average 1.40x max capacity (141w z13 versus 101w zEC12)
- Subcapacity models
  - Uniprocessor capacity ratio to full speed z13
    - 0.15x (target 250 MIPS)
    - 0.44x
    - 0.63x
  - up to 30 CPs (general purpose processors) for each subcap model
- SMT capacity option
  - IFL's and zIIPs can optionally choose to run 2 HW threads per processor engine or "core"
    - opt-in or opt-out at the LPAR level
    - added HW threads appear as additional logical processors to z/VM and z/OS
  - may see wide range in capacity improvement per core over single thread: +10% to +40%
- Variability amongst workloads
  - workloads moving to z13 can expect to see more variability than last migration
    - performance driven by improved IPC in core and nest
      - workloads will not react the same to the improvements in these areas
      - micro benchmarks are particularly susceptible to this effect

SMT Overview

- SMT allows for the enablement of a second hardware thread per processor engine or "core"
  - Appears as another logical processor to z/VM and z/OS
  - LPARs may opt-in or opt-out to SMT on IFLs or zIIPs
- Capacity gain per core will vary
  - Dependent on the overlap and interference between the two threads
    - overlap
      - many core resources are replicated so each thread can make progress
      - while one thread waits for a cache miss, the other thread can continue to run
    - interference
      - some serialization points within the core
      - threads share the same caches, thus cache misses can increase
  - Benchmarks observe +10% to +40% capacity increase versus single HW thread per core
    - no clear predictor of where a workload will fall
- With SMT, individual tasks (SW threads) run slower but dispatcher delays reduce
  - For example, a 1.3x capacity gain is spread over 2 HW threads which means each thread runs at 1.3/2 = .65x a single thread or about the speed of a z196 core
  - But with twice as many HW threads (logical processors) to dispatch to, dispatching delays (CPU queuing) can be reduced
- Metering available through RMF and z/VM Performance Reports
  - Thread density, utilization, capacity factors

What's new in the LSPR for z13

- Workload updates
  - upleveled software - z/OS 2.1, subsystems, compilers
  - minor tweaks to three hardware-characteristic-based workload categories
    - based on CPU MF data from customers' z196 to zEC12 migrations
- HiperDispatch continues to be turned on for all measurements
  - important even on smaller Nway processors starting with z196 and above due to sensitivity to L3 chip-level cache
- LSPR will publish only single HW thread capacity in 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
- zPCR allows any configuration to be modelled
  - customized LPAR configurations and workloads (as always)
  - SMT capacity effect will be included via a user controlled "dial"
    - set dial to reflect the estimated capacity increse of 2 threads over 1 thread
    - pre-install guideance in setting dial to be provided based on internal testing and eventual field experience (Defaults to 20% for IFLs, 25% for zIIPs)
    - post-install guideance in setting dial from metering data available in RMF and z/VM Performance Reports

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 V2R1 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)

LSPR website z/OS V2R1 Tables
z13 versus zEC12

Multi Image Table

	z/OS V2R1 AVERAGE	z/OS V2R1 AVERAGE	z/OS V2R1 AVERAGE	z/OS V2R1 AVERAGE
	zEC12 ITR	z13 ITR	z13:zEC12 ratio	z13 PCI
701	2.70	3.03	1.12	1695
708	17.98	19.99	1.11	11188
716	31.98	35.13	1.10	19665
732	55.85	61.55	1.10	34456
764	98.70	108.44	1.10	60706
7A1	140.10	154.99	1.11	86761
z13 7E1 vs zEC12 7A1	140.10	199.28	1.42	111556

z13 includes 3 subcapacity offerings

Subcapacity Offerings vs Full Speed

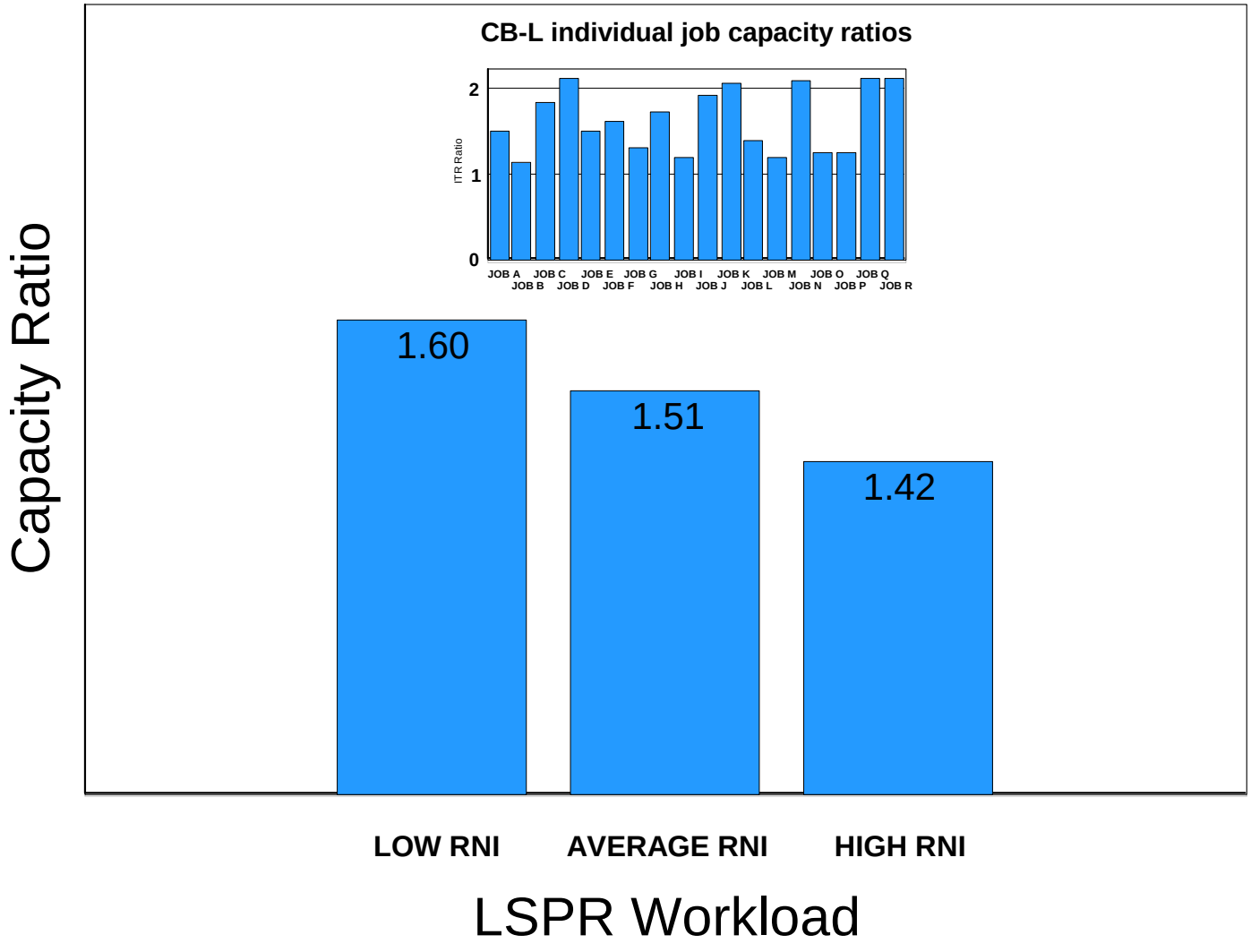
z13	z/OS V2R1 MI AVG ITRR	Ratio to 701	PCI	Max #CPs
701	3.03	1.00	1695	141
601	1.91	.63	1068	30
501	1.33	.44	746	30
401	.45	.15	250	30

Notes: Uni speeds range from 15% to 63% of full speed uni
Each subcapacity offering has a maximum of 30 CPs

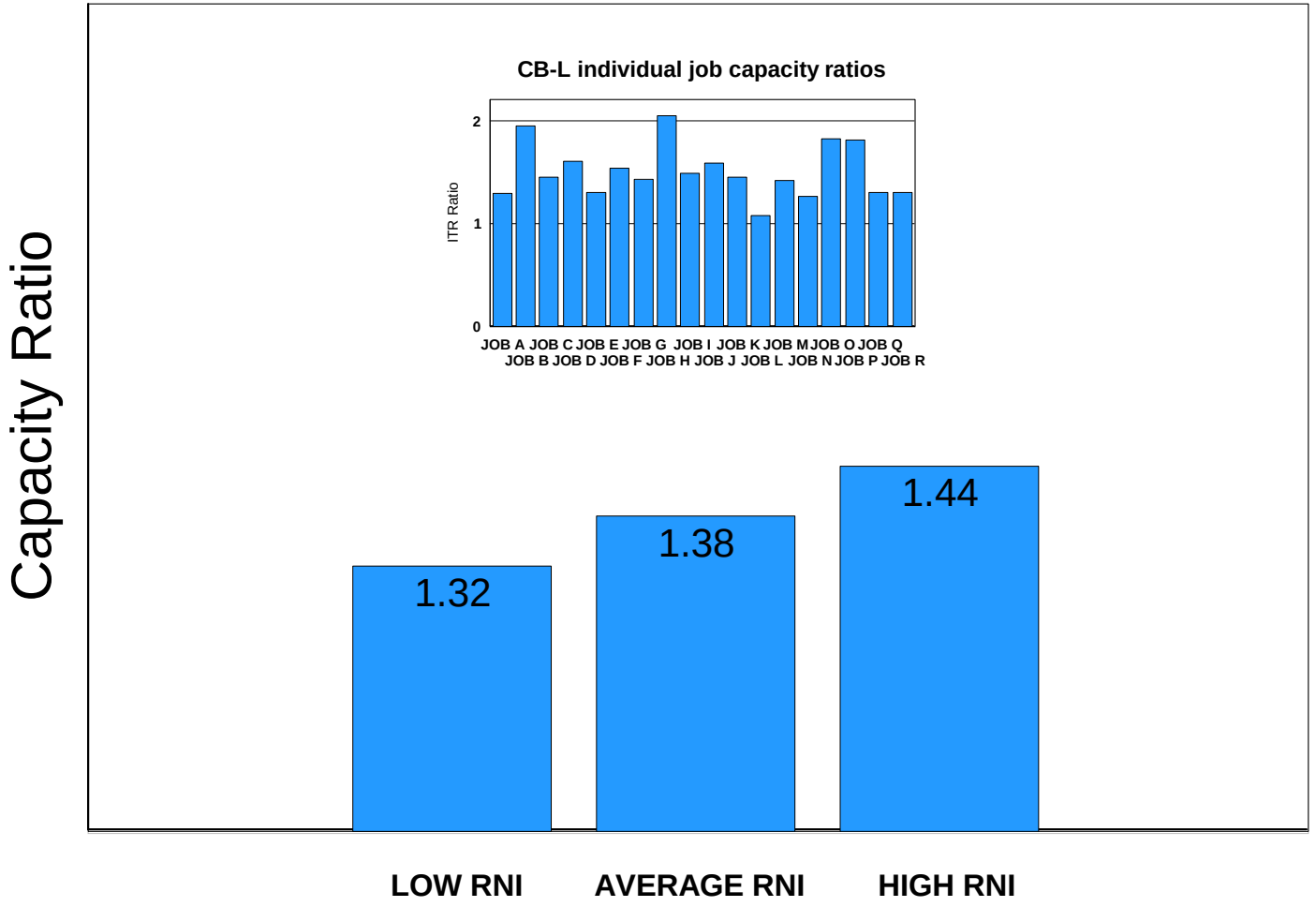
Workload Variability with z13

- 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
- z13 performance comes from improved IPC (instructions per cycle) in both the micro-processor and the memory subsystem (clock speed is 10% slower but tasks run on average 10% or more faster)
  - magnitude of improvement in IPC will vary by workload
  - workloads moving into a z13 will likely see more variation than last migration
- Examples of workload variation for moves to new technology starting with the z9 appear on the next few slides

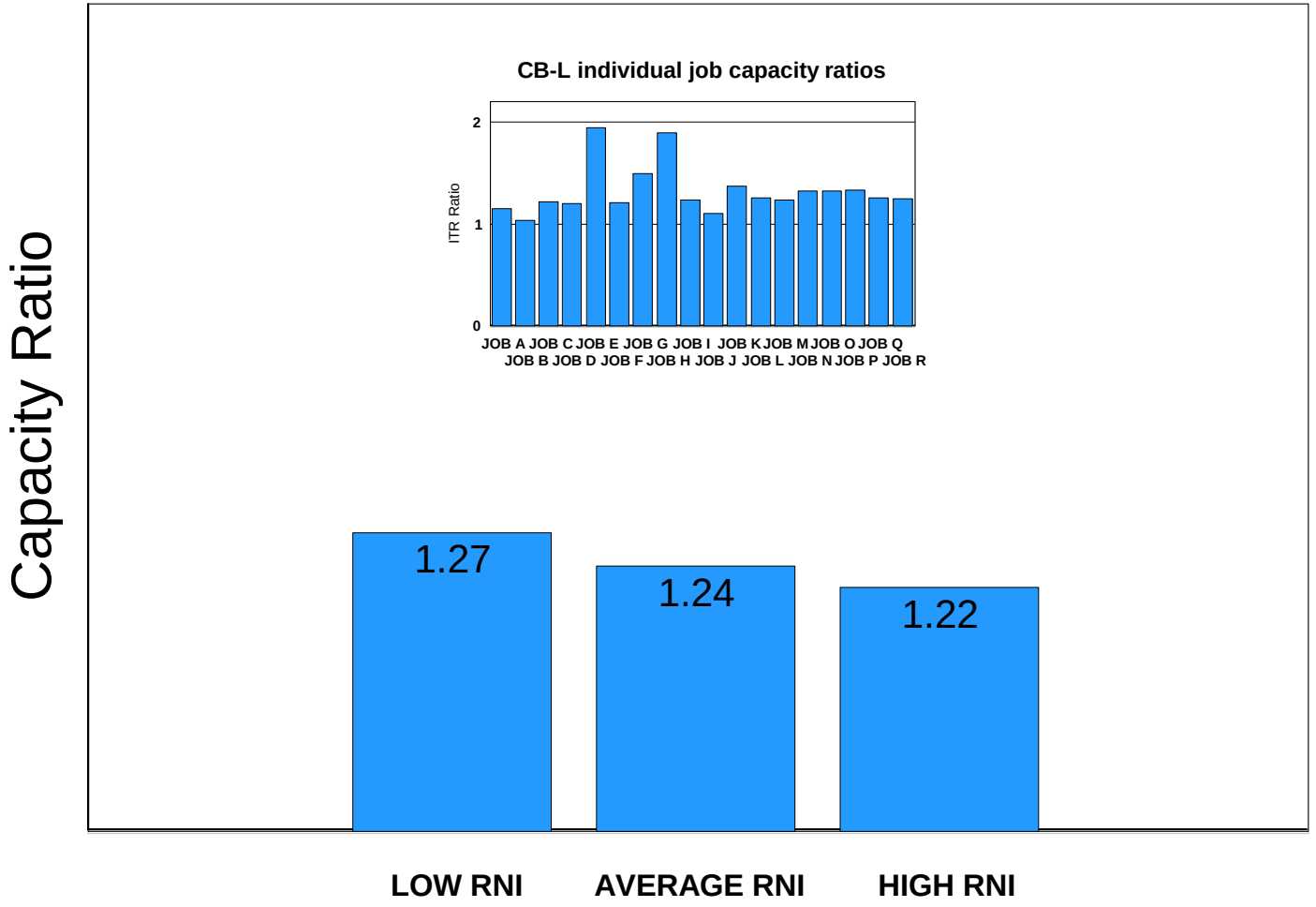
LSPR Single Image Capacity Ratios
10way: z10 EC versus z9 EC
Example of Workload Variability



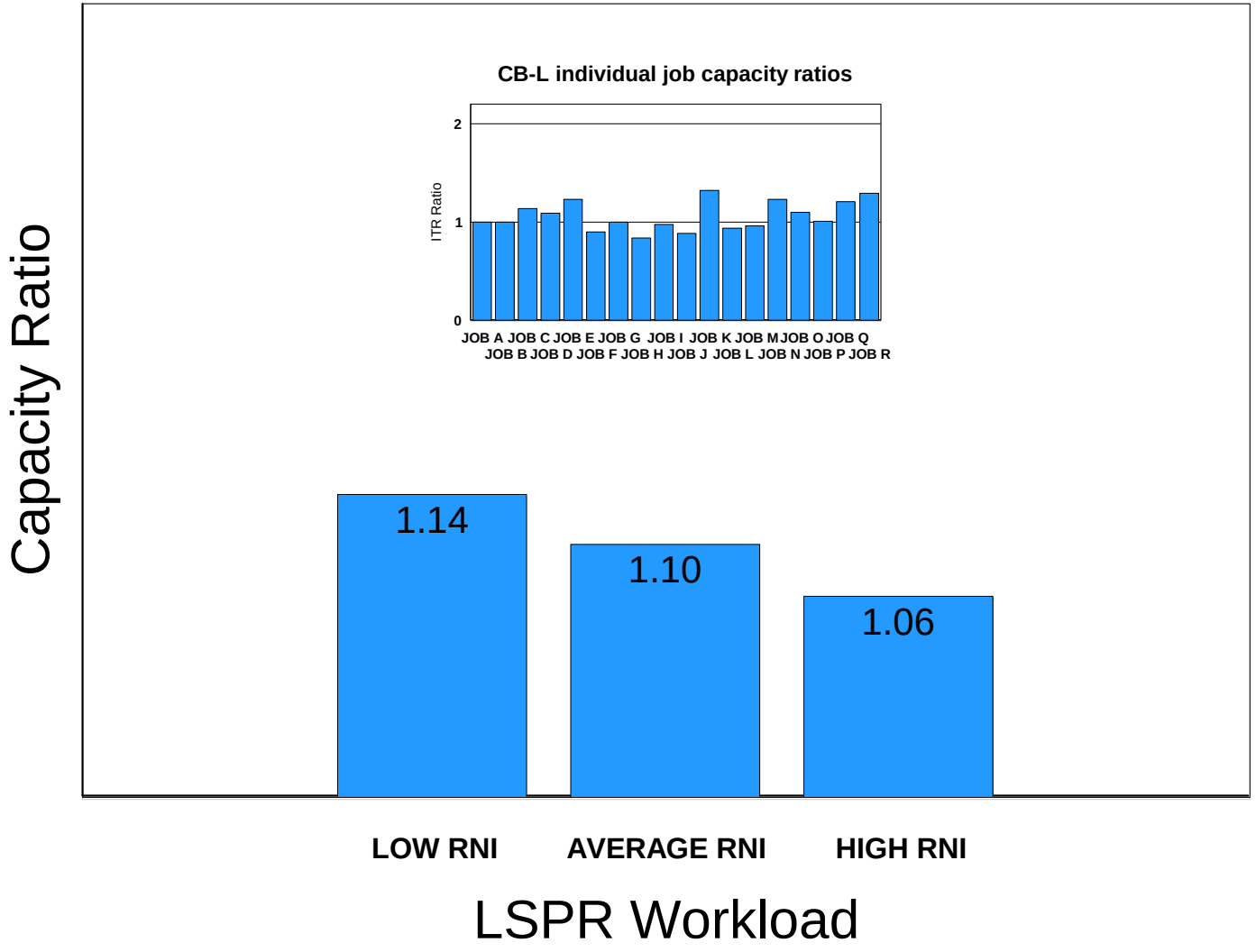
LSPR Single Image Capacity Ratios
10way: z196 versus z10 EC
Example of Workload Variability



LSPR Single Image Capacity Ratios
16way: zEC12 versus z196
Example of Workload Variability



LSPR Single Image Capacity Ratios
16way: z13 versus zEC12
Example of Workload Variability



Summary

- "Relatively New" RNI-based LSPR
  - Validated and now default in zPCR
- z13 traditional performance
  - approximately 10% more capacity per engine than zEC12
  - max config provides approximately 40% more capacity vs zEC12
- z13 new performance opportunities
  - 3x memory
  - SMT for IFLs and zIIPs add another 10% to 40% per engine capacity
  - SIMD for analytics
- Workload variability will be higher than past few generations