The All New LSPR
and
z196 Performance Brief

SHARE Anaheim

EWCP

Gary King
IBM

March 2, 2011

Trademarks

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

AlphaBlox*
APPN*
CICS*
CICS/VSE*
Cool Blue
DB2*
DFSMS
DFSMShsm
DFSMSrmm
DirMaint
DRDA*
DS6000
DS8000
ECKD
ESCON*
FICON*
FlashCopy*

GDPS*
HiperSockets
HyperSwap
IBM*
IBM eServer
IBM logo*
IMS
Language Environment*
Lotus*
Large System Performance Reference™ (LSPR™)
Multiprise*
MVS
OMEGAMON*
Parallel Sysplex*
Performance Toolkit for VM
PowerPC*
PR/SM
Processor Resource/Systems Manager

RACF*
Redbooks*
Resource Link
RETAIN*
REXX
RMF
S/390*
Scalable Architecture for Financial Reporting
Sysplex Timer*
Systems Director Active Energy Manager
System/370
System p*
System Storage
System x*
System z
System z9*
System z10

Tivoli*
Tivoli Storage Manager
TotalStorage*
VSE/ESA
VTAM*
WebSphere*
zEnterprise
xSeries*
z9*
z10
z10 BC
z10 EC
z/Architecture*
z/OS*
z/VM*
z/VSE
zSeries*

* 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

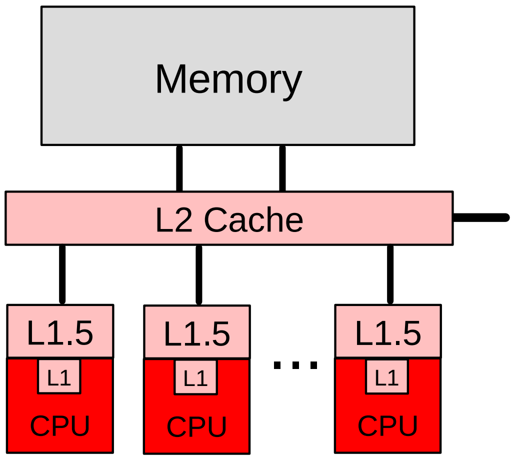
- Performance drivers with z196
- What's New in the LSPR
- z196 ITR Ratios
- Workload Variability
- Subcapacity Offerings

Performance Drivers with z196

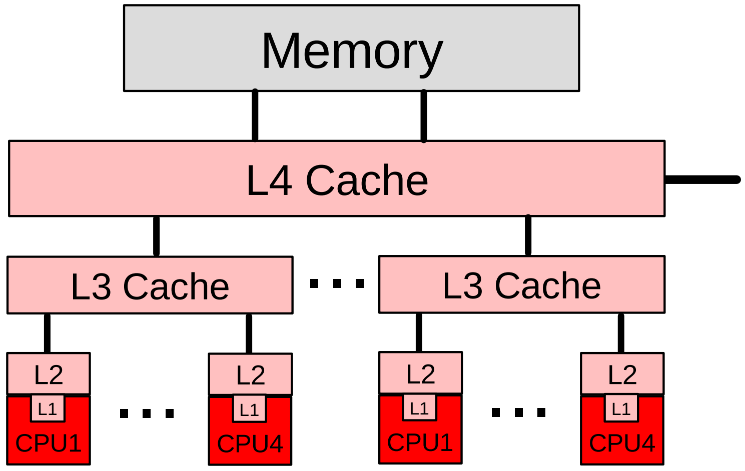
- Hardware
  - memory hierarchy
    - focus on keeping data "closer" to the processor unit
    - new chip-level shared cache
    - much larger book-level shared cache
  - processor
    - Out-Of-Order execution
    - new instructions to allow for ...
      - reduced processor quiesce effects
      - reduced cache misses
      - reduced pipeline disruption
  - up to 80 configurable processor units
  - 4 different uni speeds
- HiperDispatch
  - exploits new cache topology
  - reduced cross-book "help"
  - better locality for multi-task address spaces

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



Performance Drivers with z196

- Hardware
  - memory hierarchy
    - focus on keeping data "closer" to the processor unit
    - new chip-level shared cache
    - much larger book-level shared cache
  - processor
    - Out-Of-Order execution
    - over 100 new instructions to allow for ...
      - reduced processor quiesce effects
      - improved compiled code efficiency
  - up to 80 configurable processor units
  - 4 different uni speeds
- HiperDispatch
  - exploits new cache topology
  - reduced cross-book "help"
  - better locality for multi-task address spaces

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

What's new in the LSPR for z196

- Workload updates
  - upleveled software - z/OS 1.11, subsystems, compilers
  - three new hardware-characteristic-based workload categories
    - replace current workload primitives and mixes
      - there is a suggested translation from old to new names
- HiperDispatch continues to be turned on for all measurements
- LSPR Tables
  - 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
    - one z/OS image equal in size to Nway of model (z/OS V1R11 to 80way)
    - dropped from the LSPR website
    - continues to be input to zPCR

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 LPSR for z196 will introduce three new workload categories which replace all prior primitives and mixes.
- To simplify the transition, an easy and automatic translation of old names to new categories will be supplied with zPCR.
  - For example, if you have been using LoIO-mix in your studies, you will simply use the new "average" workload in the future

Fundamental Components of Workload Capacity Performance Part 1

- Instruction Path Length for a transaction or job
  - Application dependent, of course
  - Generally invariant across processor designs
  - But can be sensitive to Nway (due to MP effects such as locking, work queue searches, etc)

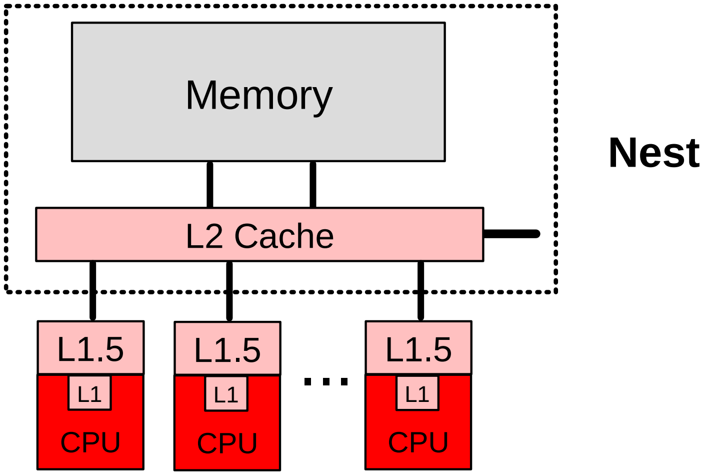
- Instruction Complexity (Micro processor design)
  - Many design alternatives
    - Cycle time (GHz), instruction architecture, pipeline, superscalar, Out-Of-Order, branch prediction and more
  - Workload effect
    - May be different with each processor design
    - But once established for a workload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

Fundamental Components of Workload Capacity Performance Part 2

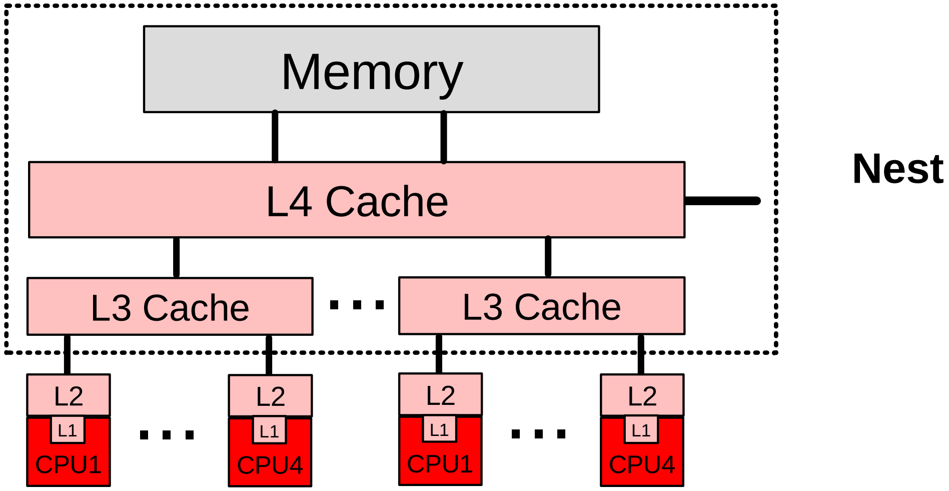
- Memory Hierarchy or "nest"
  - Many design alternatives
    - cache (levels, size, private, shared, latency, MESI protocol), controller, data buses
  - Workload effect
    - Quite variable
    - Sensitive to many factors: locality of reference, dispatch rate, IO rate, competition with other applications and/or LPARs, and more
  - Relative Nest Intensity
    - Activity beyond private-on-chip cache(s) is the most sensitive area
    - Reflects activity distribution and latency to shared caches and memory
    - Level 1 cache miss activity is 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

Low	Relative Nest Intensity	High
Batch	Application Type	Transactional
Low	IO Rate	High
Single	Application Mix	Many
Intensive	CPU Usage	Light
High Locality	Data Reference Pattern	Diverse
Simple	LPAR Configuration	Complex
Extensive	Software Configuration Tuning	Limited

New 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 to reflect the new workload categories
  - 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 new z10 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

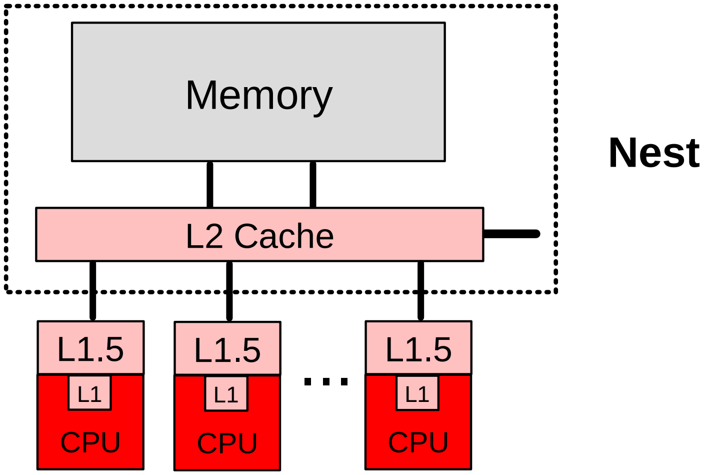
- How can the COUNTERS be used today?
  - To supplement current performance data from SMF, RMF, DB2, CICS, etc.
  - To help understand **why** performance may have changed

- How can the COUNTERS be used for future processor planning?
  - They provde the basis for the new LSPR workload categories
  - zPCR automically 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"

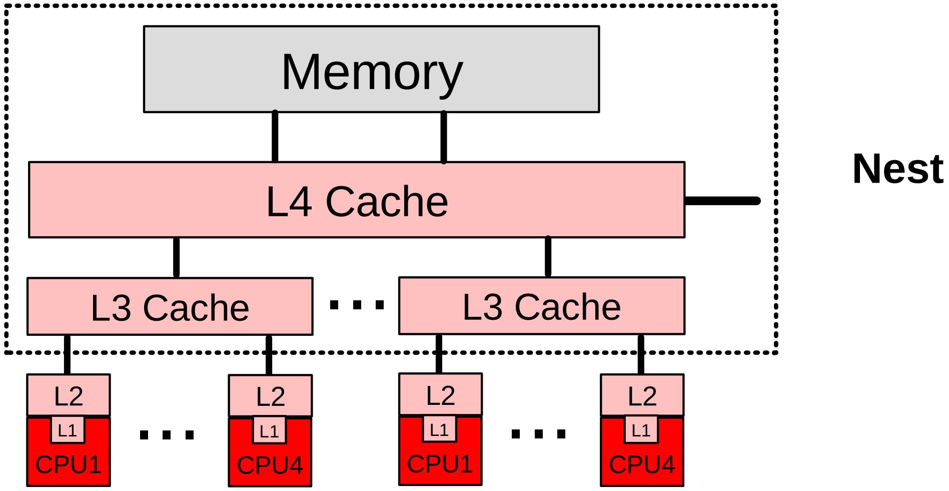
- Brand new RedPaper draft on CPU MF now available
  - http://www.redbooks.ibm.com/redpieces/abstracts/redp4727.html?Open

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

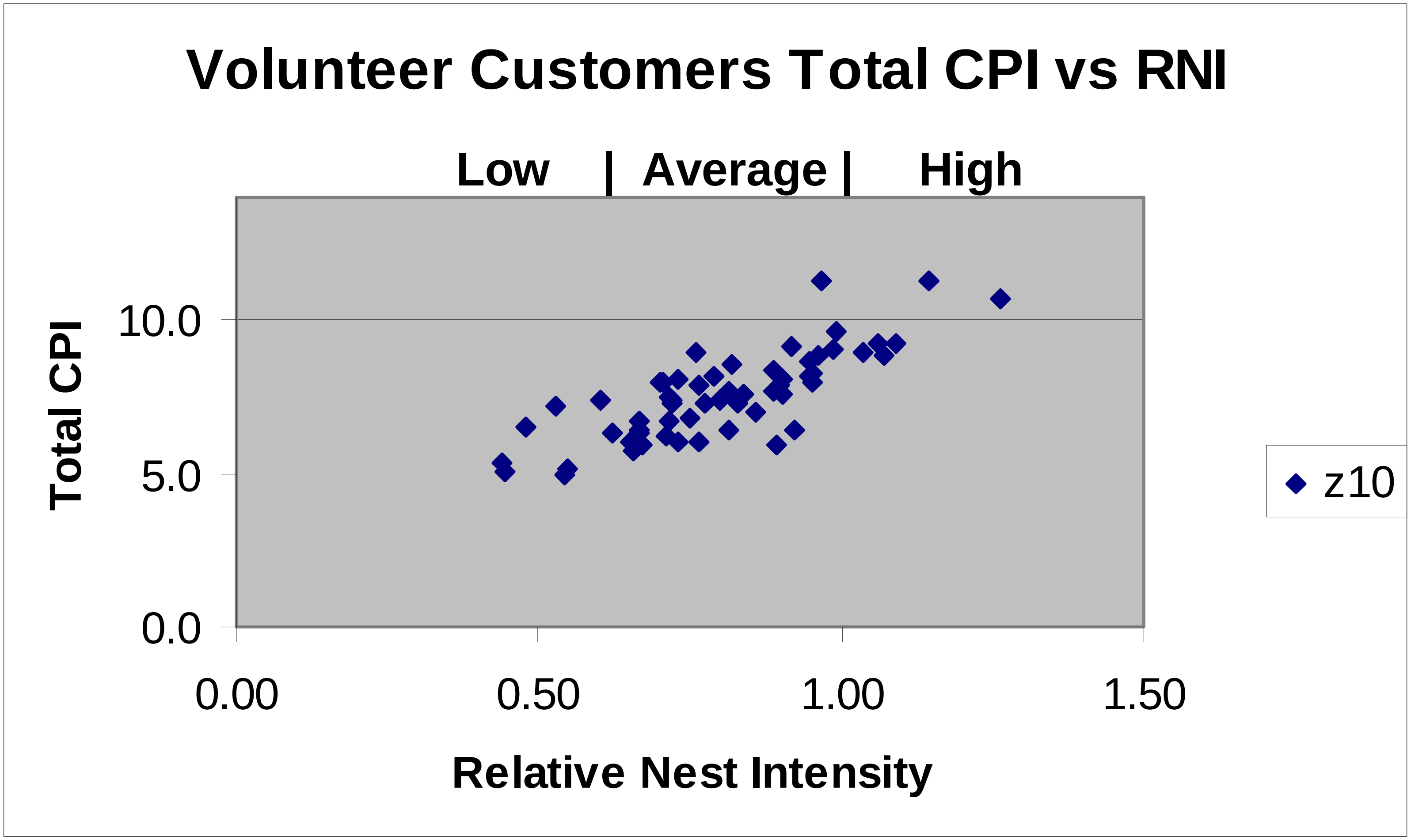
z10 RNI = (1.0xL2LP+2.4xL2RP+7.5xMEMP)/100

RNI-based Workload "Hint" Decision Table

L1MP	RNI	Workload Hint
<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 z10 CPU MF data
table may change based on feedback

CPU MF
z10 Customer Workload Characterization Summary



Using the z/OS V1R11 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 V1R11 MI Table
Comparing V1R11 to V1R9 using z10 EC

z10 EC	z/OS V1R9 MI	z/OS V1R11 MI	z/OS V1R9 MI	z/OS V1R11 MI
	old SW old Wklds	new SW new Wklds	old SW old Wklds	new SW new Wklds
	ITR	ITR	PCI	PCI
701	1.00	1.00	923	902
708	6.41	6.52	5921	5879
716	11.22	11.56	10360	10429
732	19.51	20.33	18008	18388
764	33.38	35.28	30811	31826

Notes: ITR ratios are normalized to 2097-701 (actual table normalized 2094-701 = 1.00)
V1R11 ITR ratios higher than V1R9 due to upleveled SW and new workloads
V1R11 PCIs slightly "re-centered" to minimize change from V1R9 across the models

LSPR website z/OS V1R11 Tables
z196 versus z10 EC

Multi Image Table

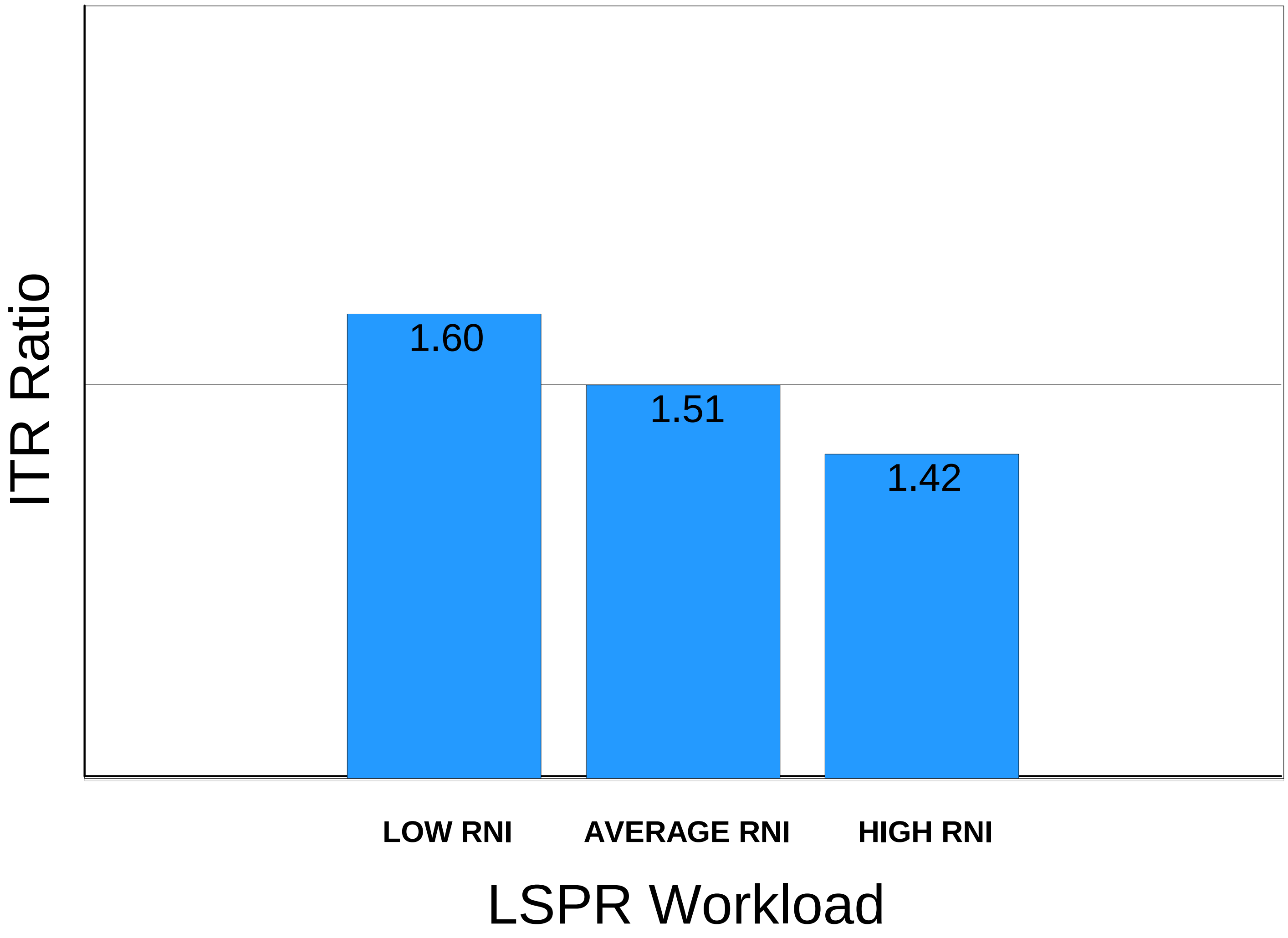
	z/OS V1R11 AVERAGE	z/OS V1R11 AVERAGE	z/OS V1R11 AVERAGE	z/OS V1R11 AVERAGE
	z10 EC ITR	z196 ITR	z196:z10 EC ratio	z196 PCI
701	1.61	2.15	1.33	1202
708	10.50	14.42	1.37	8072
716	18.63	25.67	1.38	14371
732	32.76	45.09	1.38	25241
764	56.85	80.30	1.41	44953
z196 780 vs z10 764	56.85	93.40	1.64	52286

Workload Variability with z196

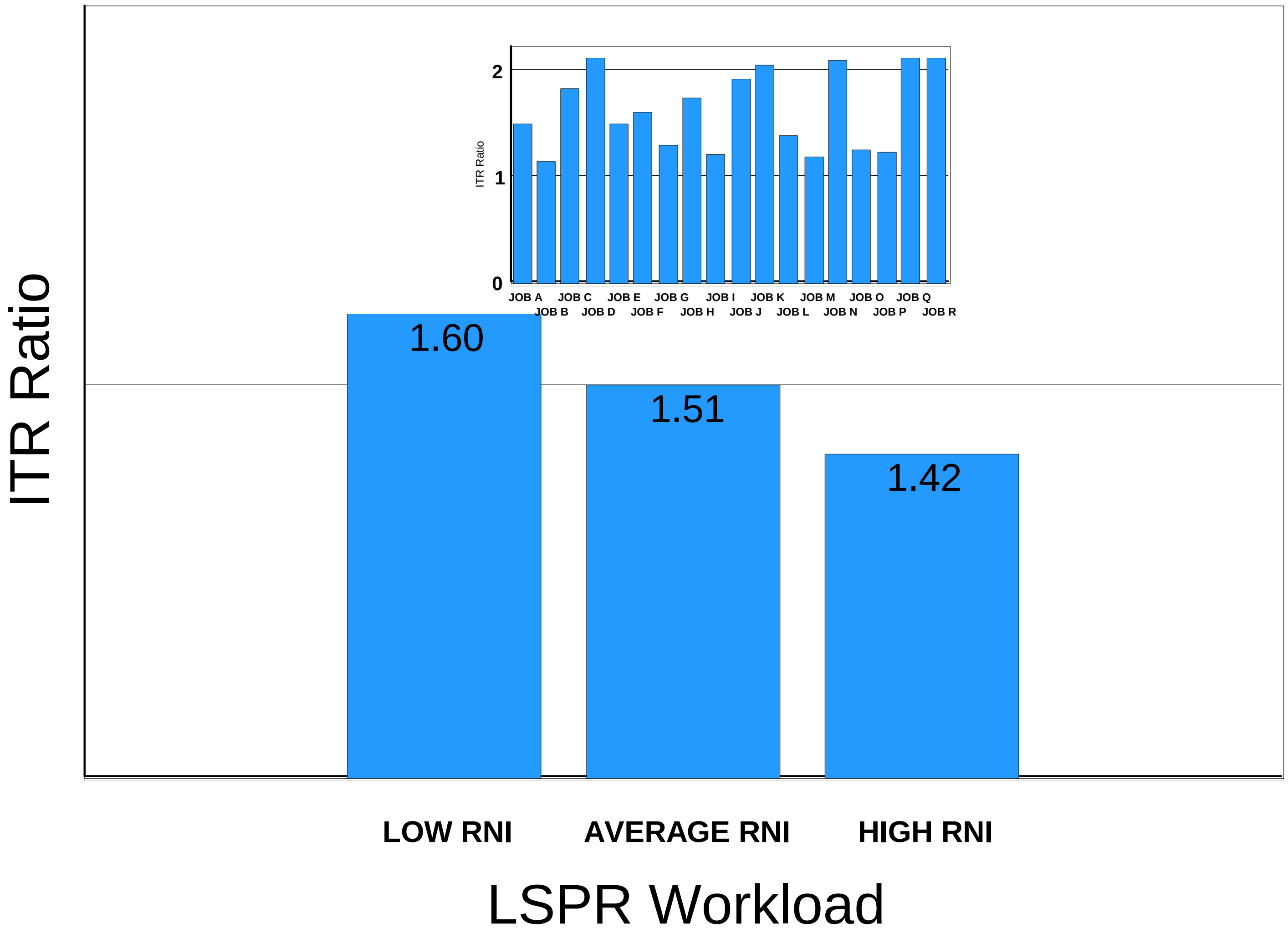
- Workloads moving onto a z196 will see less variability around "average" performance than z10 but more than past upgrades
  - variability generally related to fast clock speed and physics
    - increased memory hierarchy latencies relative to clock speed
  - z196 memory hierarchy enhancements dampens this effect
  - above and below average workloads typically the reverse of moves to z10
  - workload characteristics determinant, not application type

- Examples of z9 to z10 and z10 to z196 on next several slides

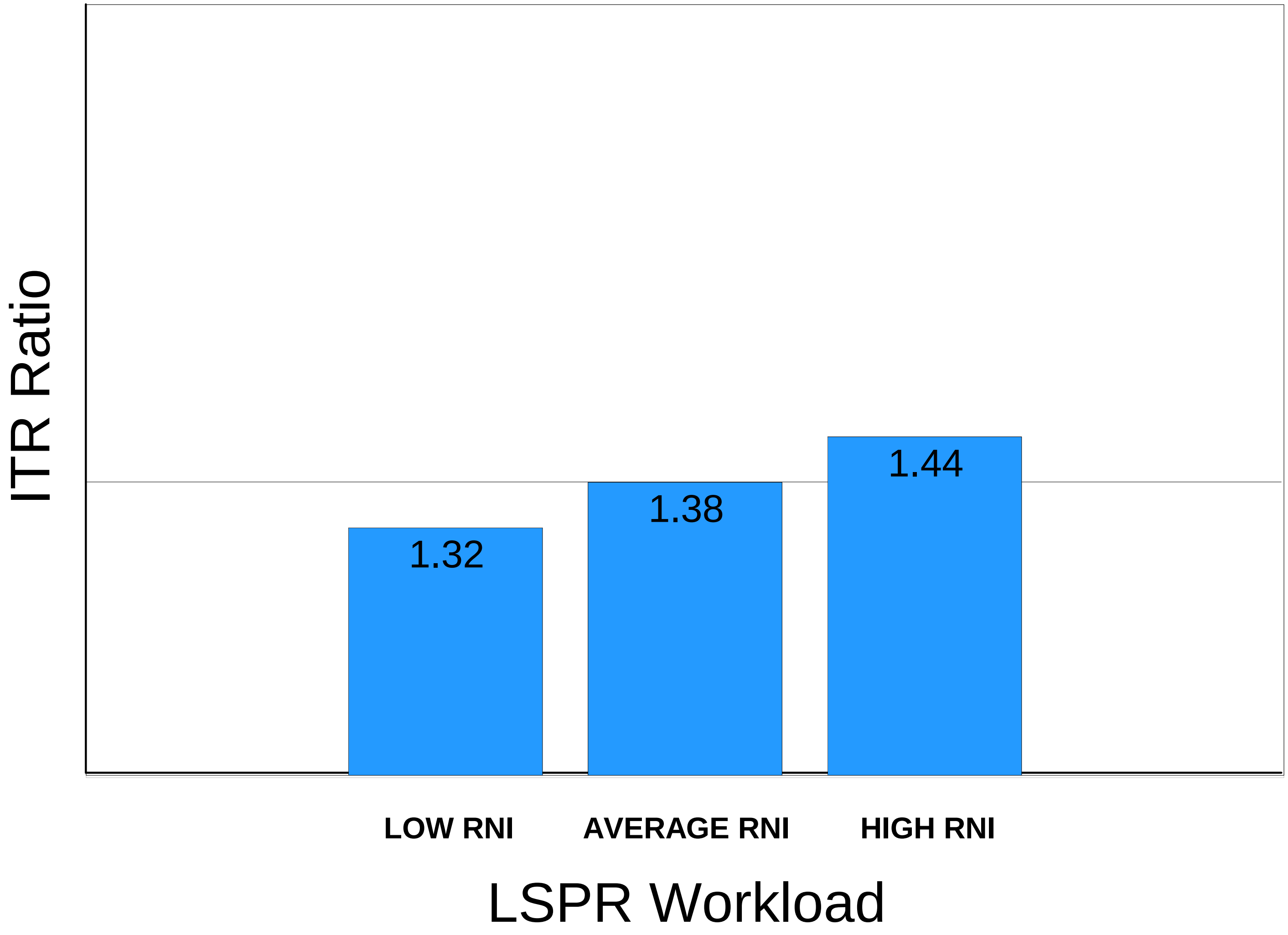
LSPR Workload ITR Ratios
z10 vs z9 @ SI 10way
Workload Variability



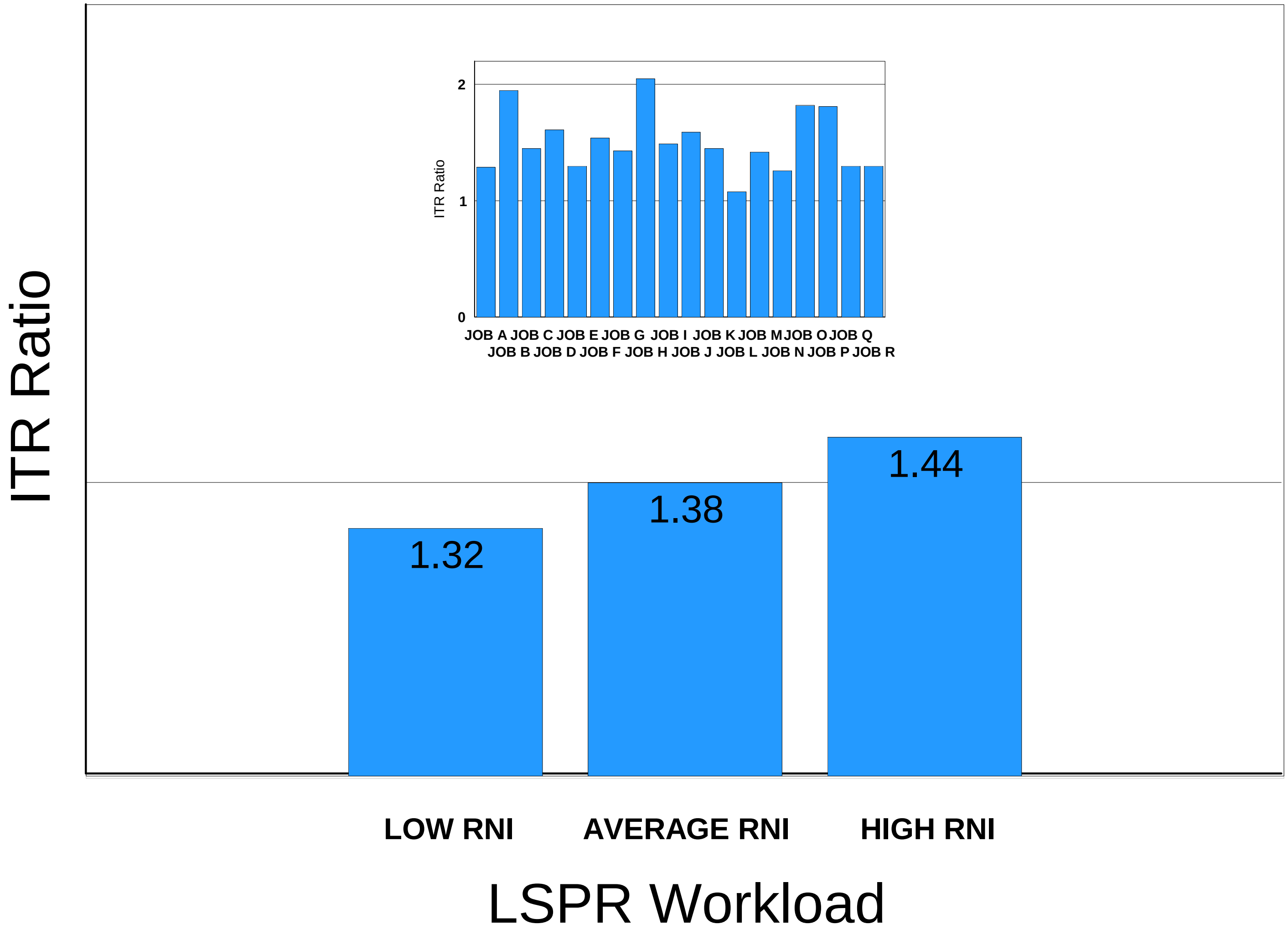
LSPR Workload ITR Ratios
z10 vs z9 @ SI 10way
Variability of Individual Jobs



LSPR Workload ITR Ratios
z196 vs z10 @ SI 10way
Workload Variability



LSPR Workload ITR Ratios
z196 vs z10 @ SI 10way
Variability of Individual Jobs



z196 includes 3 subcapacity offerings

Subcapacity Offerings vs Full Speed

z196	z/OS V1R11 MI AVG ITRR	Ratio to 701	Max #CPs
701	2.15	1.00	80
601	1.37	.64	15
501	1.05	.49	15
401	.43	.20	15

Notes: Uni speeds range from 20% to 64% of full speed uni
Each subcapacity offering has a maximum of 15 CPs

Backup

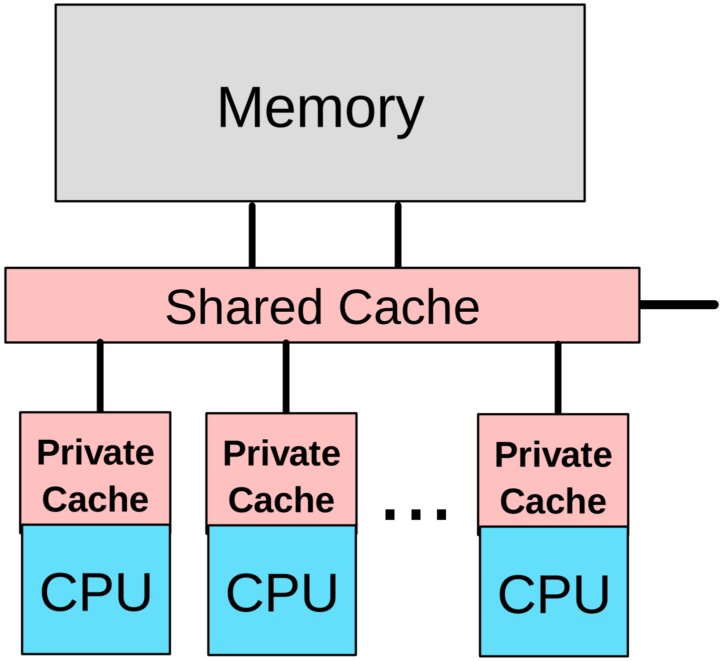
Average LPAR Configuration Profiles for the Multi-image Table

- 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

Processor Design Basics

- Processor Design
  - CPU (core)
    - cycle time
    - pipeline, execution order
    - branch prediction
    - hardware vs. millicode
  - memory subsystem (nest)
    - high speed buffers (caches)
      - on chip, on book
      - private, shared
    - buses
      - number, bandwidth
    - latency
      - distance
      - speed of light

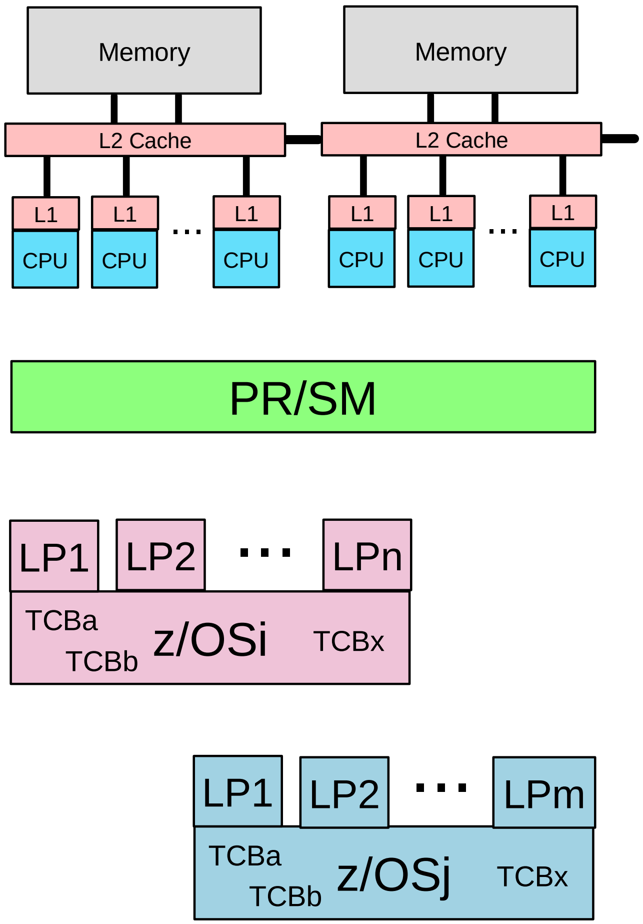
Logical View of Single Book



Hipervisor and OS Basics

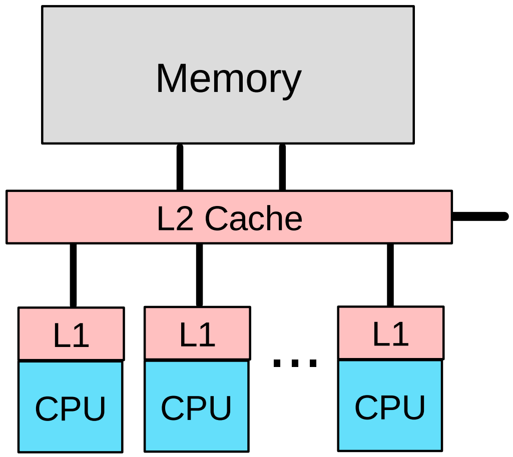
- Hipervisor (PR/SM)
  - virtualization layer at OS level
  - distributes physical resources
    - memory
    - processors
      - logicals dispatched on physicals
      - dedicated
      - shared
      - affinities
- OS
  - virtualization layer at addrspc level
  - distributes logical resources
    - memory
    - processors
      - tasks dispatched on logcials
- Enhanced cooperation
  - HiperDispatch with z10 EC
    - z/OS + PR/SM

Logical View of 2 books

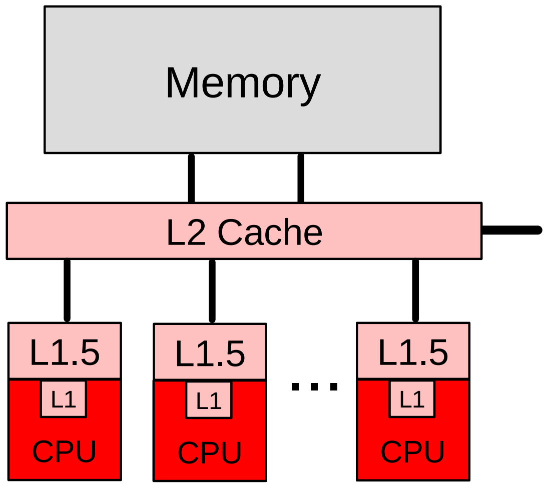


z10 versus z9 hardware comparison

- z9 EC
  - CPU
    - 1.7 GHz
    - superscalar
  - Caches
    - L1 private 256k i, 256k d
    - L2 shared 40 MB / book
    - book interconnect: ring



- z10 EC
  - CPU
    - 4.4 GHz
    - redesigned pipeline
    - superscalar
  - Caches
    - L1 private 64k i, 128k d
    - L1.5 private 3 MB
    - L2 shared 48 MB / book
    - book interconnect: star


Memory
L2 Cache
L1.5
L1.5
L1.5
L1
L1
L1
CPU
CPU
CPU
...