Achieving CPU (& MLC) Savings Through Optimizing Processor Cache

Todd Havekost, IntelliMagic

Session 21045

Agenda

- Key Processor Cache Concepts and Metrics
- Optimize LPAR Topology
- Maximize Work Executing on Vertical High CPs
  - Optimize LPAR weights
  - Increase number of physical CPs

Key Processor Cache Concepts & Metrics

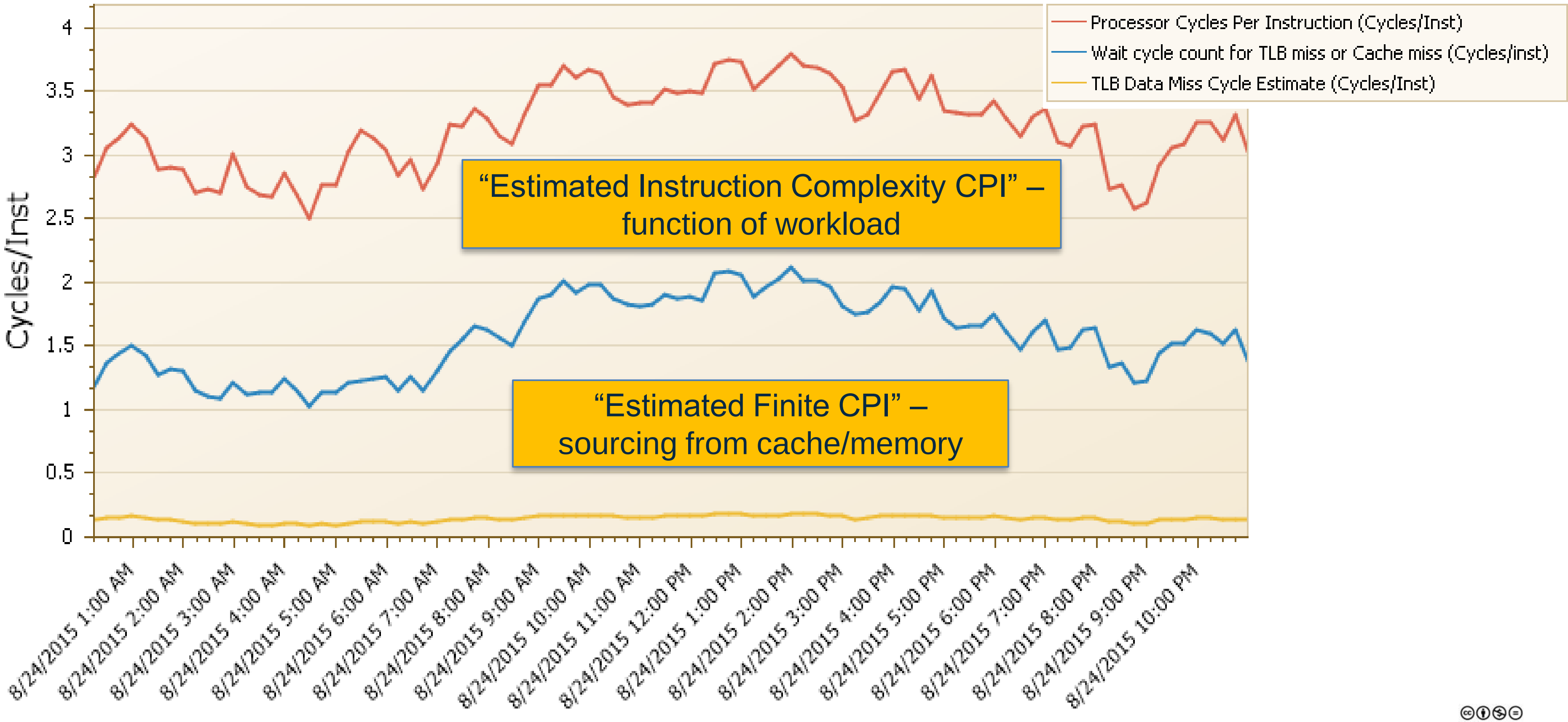
Cycles per Instruction (CPI)

- Number of processor cycles spent per completed instruction

- Processor cycles are spent
  - Productively – executing instructions present in L1 cache
  - Unproductively – waiting to stage data (L1 cache or TLB miss)

- Note: “Waiting” does not always mean waiting
  - Out Of Order (OOO) execution
  - Other pipeline enhancements

Cycles Per Instruction



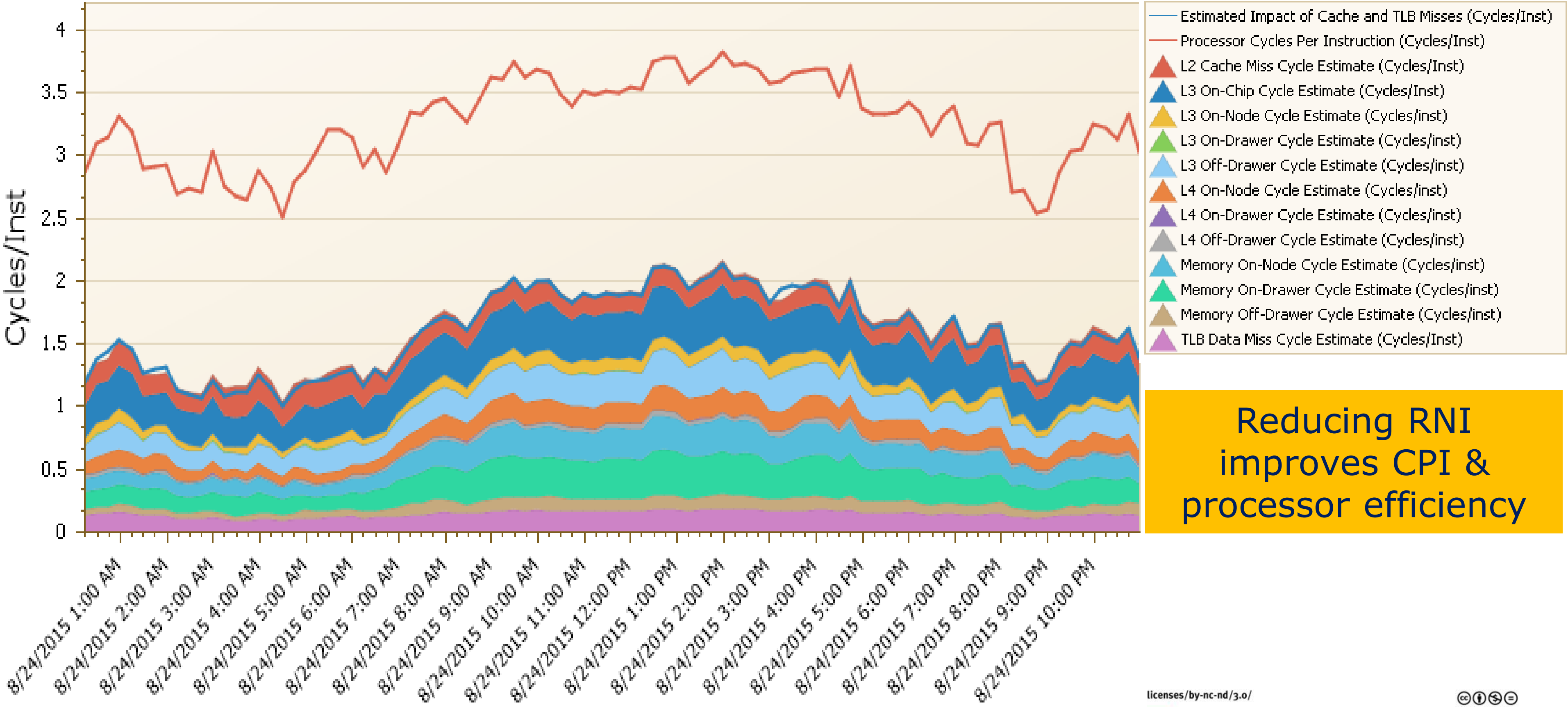
Relative Nest Intensity (RNI)

- How deep into the shared cache and memory hierarchy ("nest") the processor must go to retrieve data
- Access time increases significantly with each additional level (increasing processor wait time)

- Formulas are processor dependent: for z13, RNI = 2.3 * (0.4*L3P + 1.6*L4LP + 3.5*L4RP + 7.5*MEMP) / 100

- Reducing RNI improves processor efficiency

Estimated Impact Cache Misses

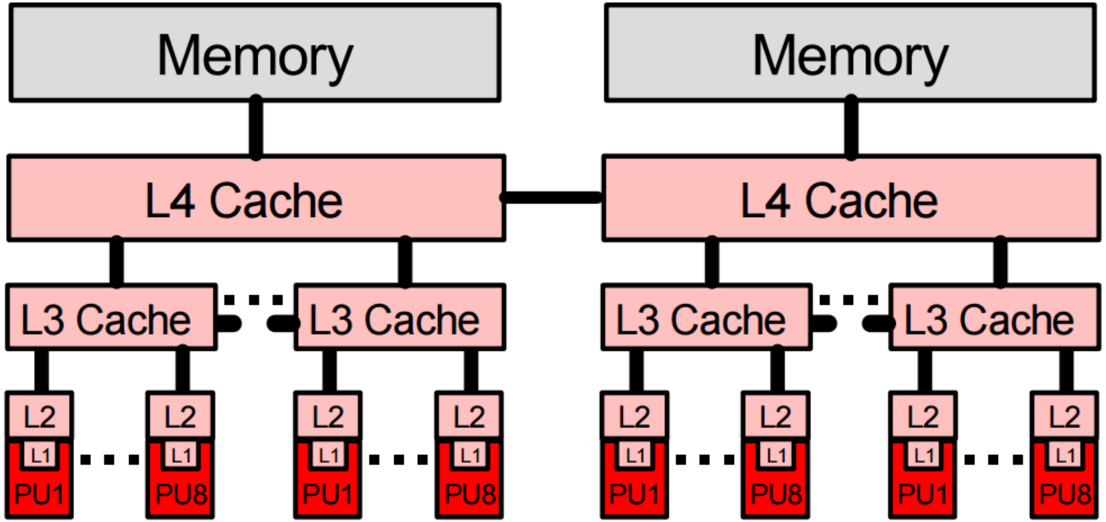


Reducing RNI improves CPI & processor efficiency

HiperDispatch

- Interfaces with PR/SM & z/OS Dispatchers to align work to logical processors (LPs) & align LPs to physical CPs

- Repeatedly dispatching the same work to the same or nearby CP is vital to optimizing processor cache hits

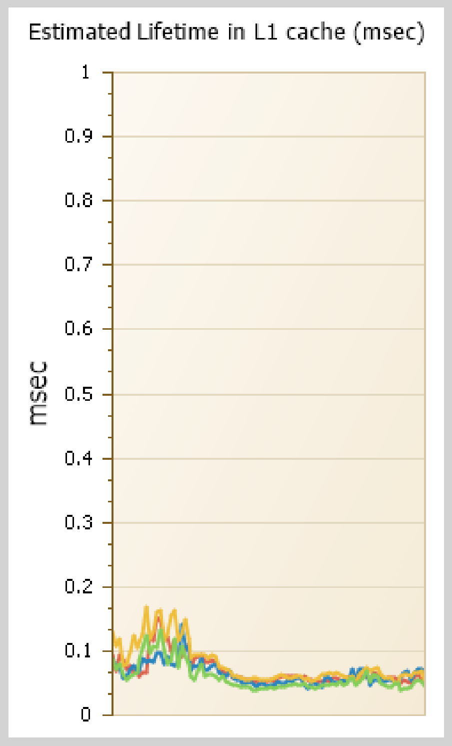


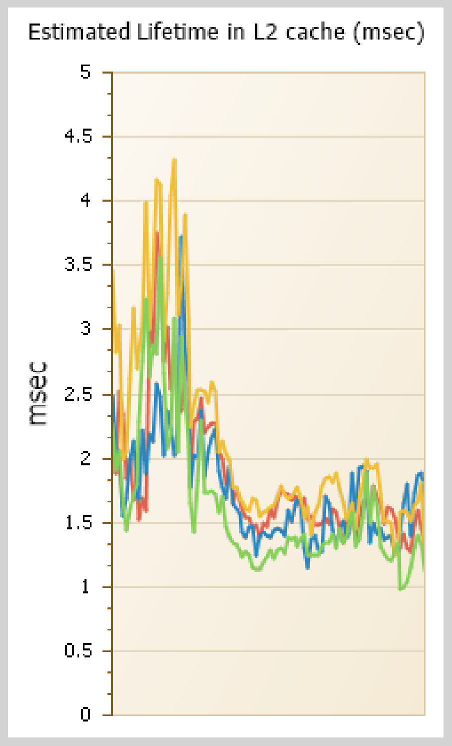
(© IBM)

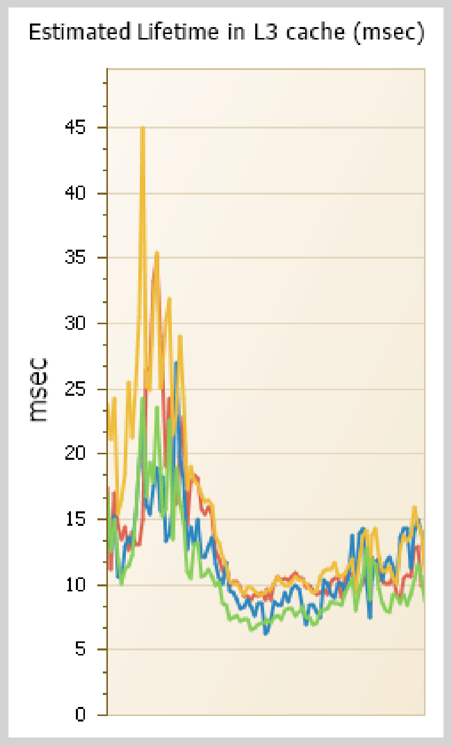
Vertical CP Assignments

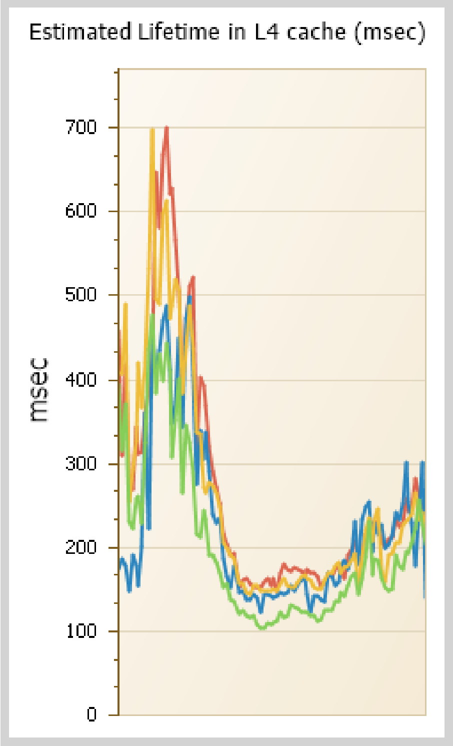
- Based on LPAR weights and the number of physical CPs PR/SM assigns logical CPs as
  - Vertical High (VH) – 1-1 relationship with physical CP
  - Vertical Medium (VM) – has at least 50% share of a CP
  - Vertical Low (VL) – has less than 50% share of a CP
- Work running on VHs has higher probability of cache hits
- Work running on VMs & VLs is subject to being dispatched on various CPs and contending with other LPARs

Cache Data Lifetime

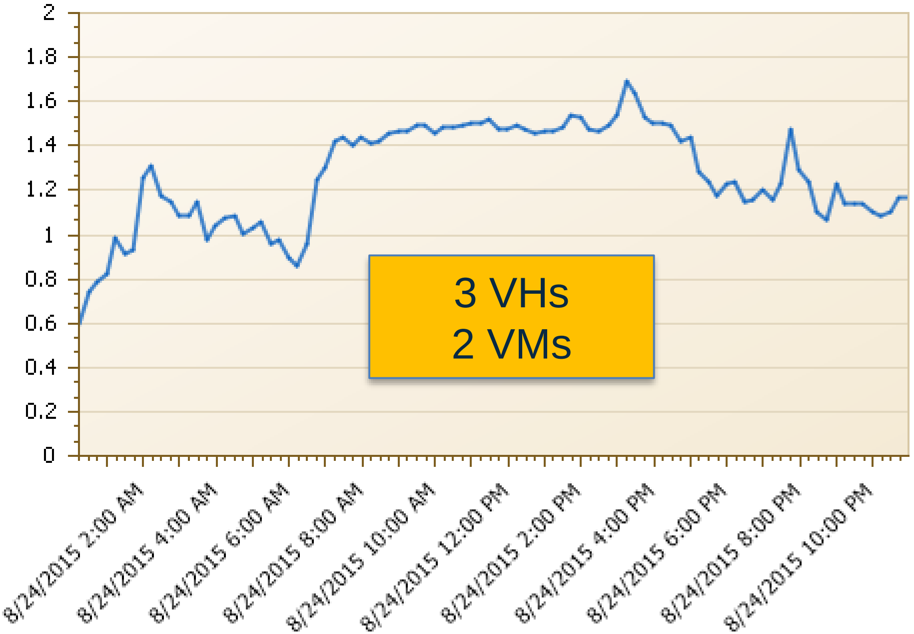


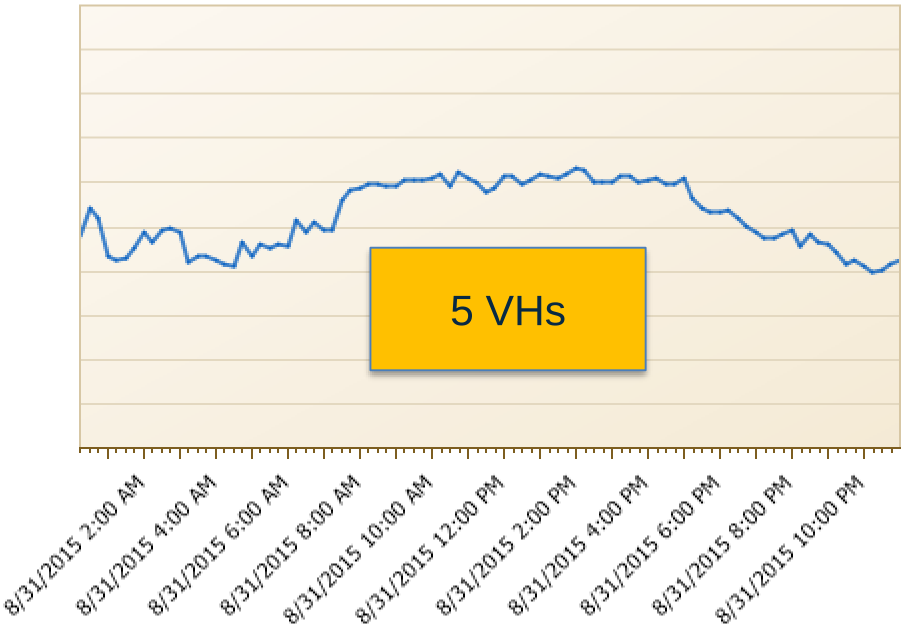




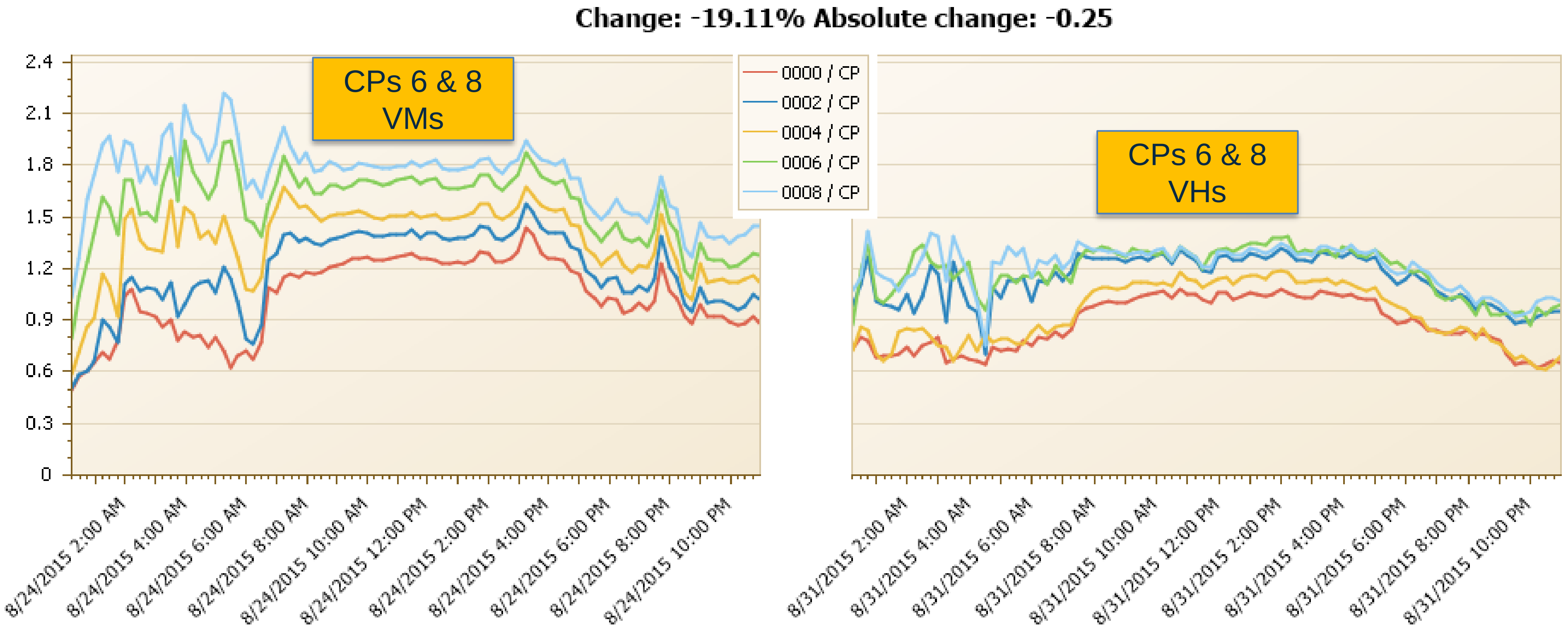


RNI Impact of More Work on VHs

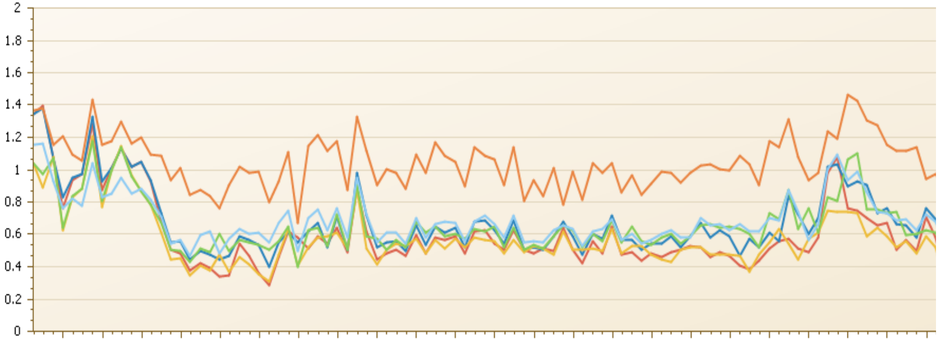


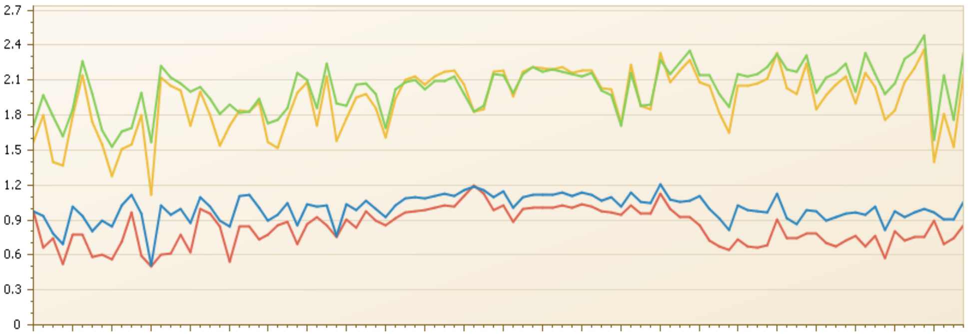


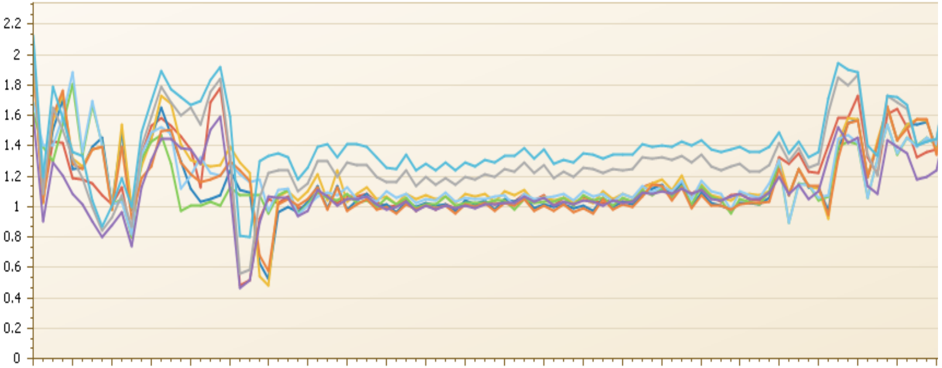
RNI Impact by Logical CP

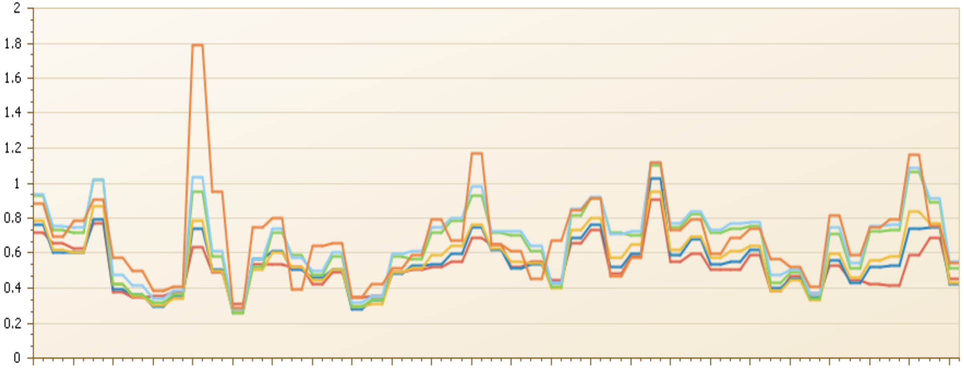


RNI Impact: VHs vs. VMs – 4 Sites

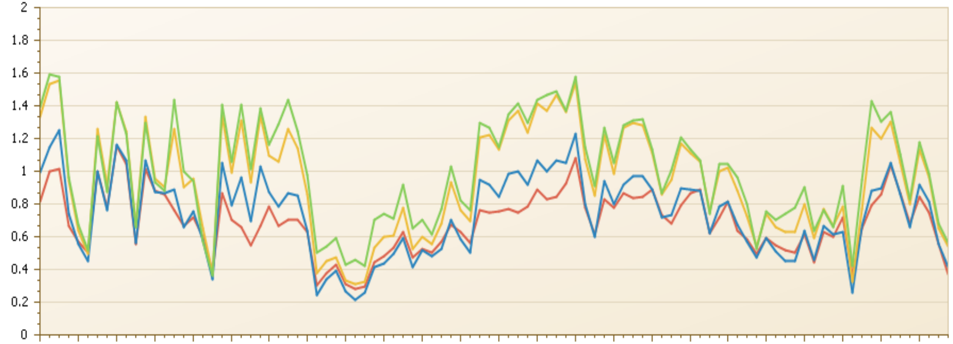


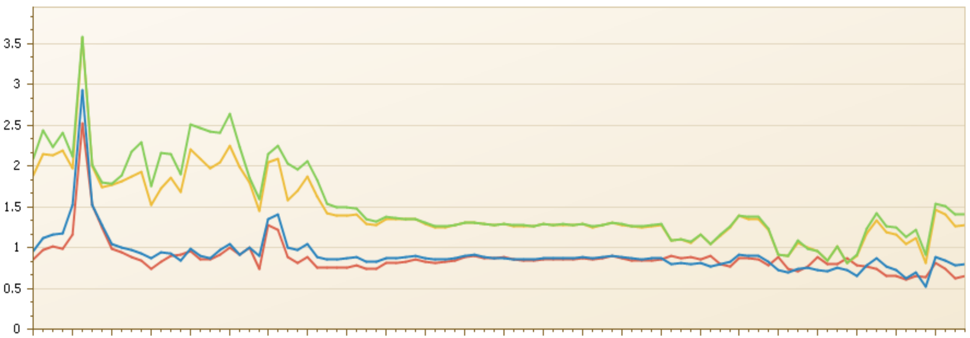


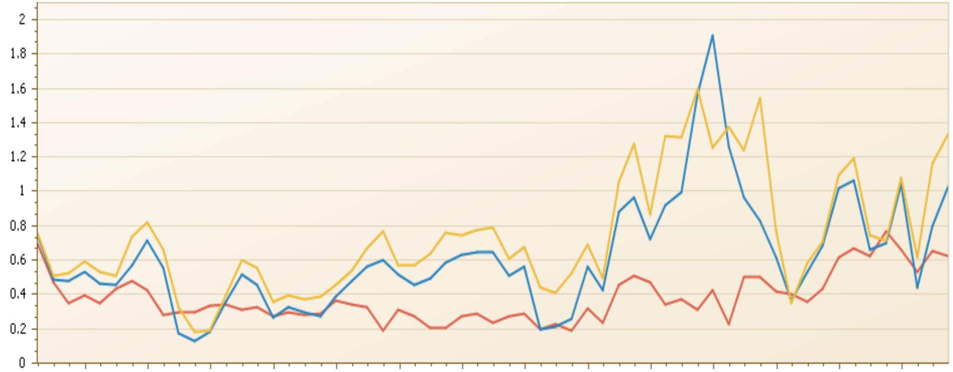


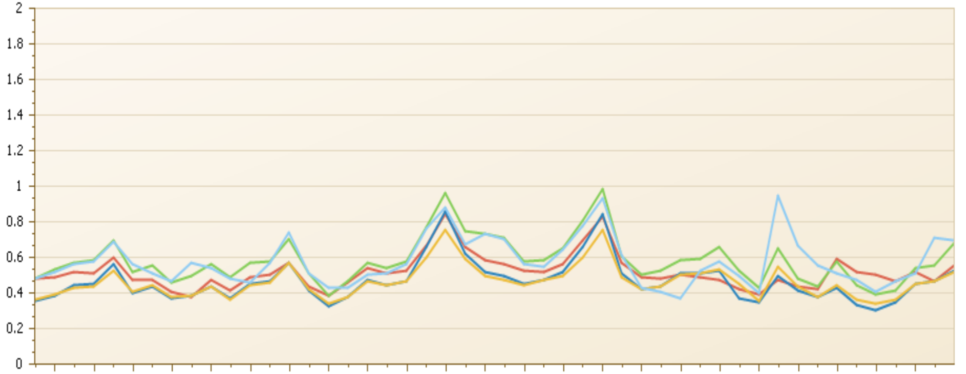


RNI Impact: VHs vs. VMs – 4 More Sites









Optimizing Processor Cache – Recap

- CPU consumption decreases when we reduce unproductive cycles waiting for data to be staged into L1 cache
  - Represent significant component of overall CPU
  - RNI metric correlates to unproductive waiting cycles
  - Reducing RNI reduces CPU (and thus MLC software expense)

- Ways to reduce RNI
  - Optimize LPAR topology
  - Maximize work executing on VHs

Optimize LPAR Topology

LPAR Topology – Concepts

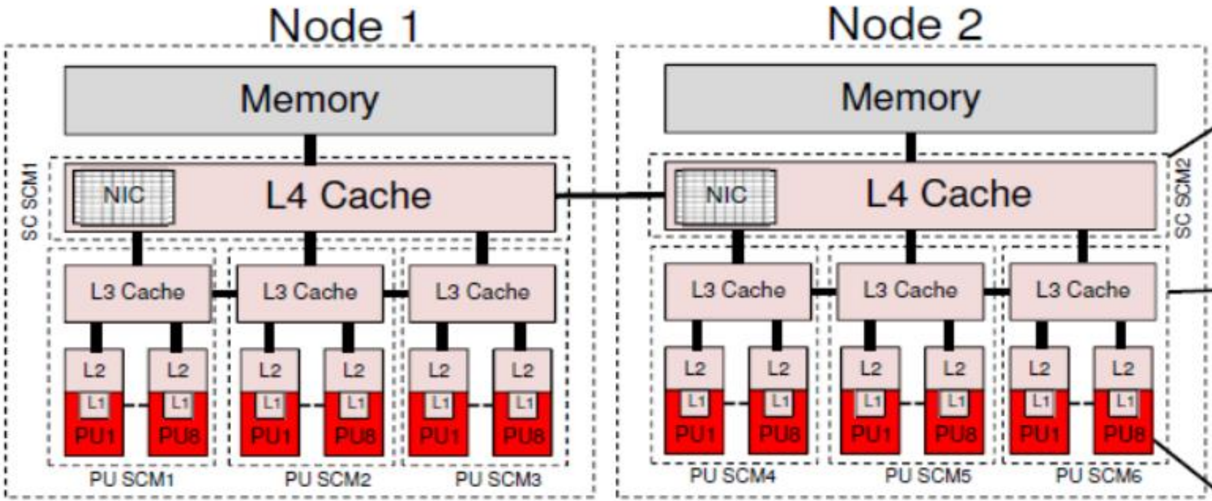
- PR/SM dynamically assigns LPAR CPs and memory to hardware chips, nodes and drawers, seeking to optimize processor cache efficiency

- LPAR topology can have a significant impact on performance, because remote accesses can take 100s of cycles

- This data is provided by SMF 99.14 records

z13 CPC Drawer Cache Hierarchy Detail

Single CPC Drawer View (N30 Model) – 2 Nodes



Node 1 - Caches

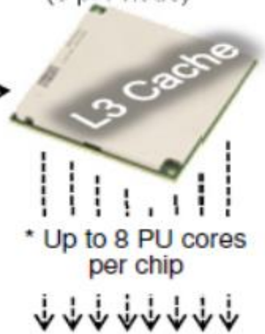
- L1 private 96k i, 128k d
- L2 private 2 MB i + 2 MB d
- L3 shared 64 MB / chip
- L4 shared 480 MB / node
  - plus 224 MB NIC

Node 2 - Caches

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

2 SC SCM Chips
(1 per node)

6 PU SCM Chips
(3 per node)



* Up to 8 PU cores per chip

Single PU core

* Not all PU's active

LPAR Topology – Sample Report

Processor Complex and LPAR information

For System 'SYS2'

Processors, LPARs and CECs with Hardware data: For System 'SYS2' by Processor Complex serial and Processor ID

Processor Complex serial	System	Processor ID	Processor A...	Processor Speed (Cycles/...	Processor Ty...	Relative Nest Inte...	Estimated TLB1 CP...
IBM-CEC1	SYS2	0000	z13	5000.00	CP	0.871	4.280
IBM-CEC1	SYS2	0002	z13	5000.00	CP	1.064	4.353
IBM-CEC1	SYS2	0004	z13	5000.00	CP	1.381	4.521
IBM-CEC1	SYS2	0006	z13	5000.00	CP	[illegible]	4.324

Logical Processors assigned to LPAR: For System 'SYS2' by Processor ID and Logical Processor/Core ID [illegible]AR

System	Processor ID	Logical Proc...	Processo...	Polarization	Core Capacity	Chip Id	Node/B...	Drawer ...	Logical...
SYS2	0000	0000	CP	Vertical High	2000000	1	1	2	0
SYS2	0002	0001	CP	Vertical High	2000000	1	1	2	0
SYS2	0004	0002	CP	Vertical Medium		2	1	2	0
SYS2	0006	0003	CP	Vertical Low		2	1	2	0

WLM Nodes for LPAR: For System 'SYS2' by WLM Node

System	WLM Node	Proces...	Vert...	Vert...	Vert...	WLM Node Flags	WLM ...	Chip...	Nod...	Draw...
SYS2	0001	CP	2	1	7	CPUs/cores on this node are boundary crossing		0	1	2
SYS2	0002	zIIP	0	1	3	CPUs/cores on this node are boundary crossing		0	1	2

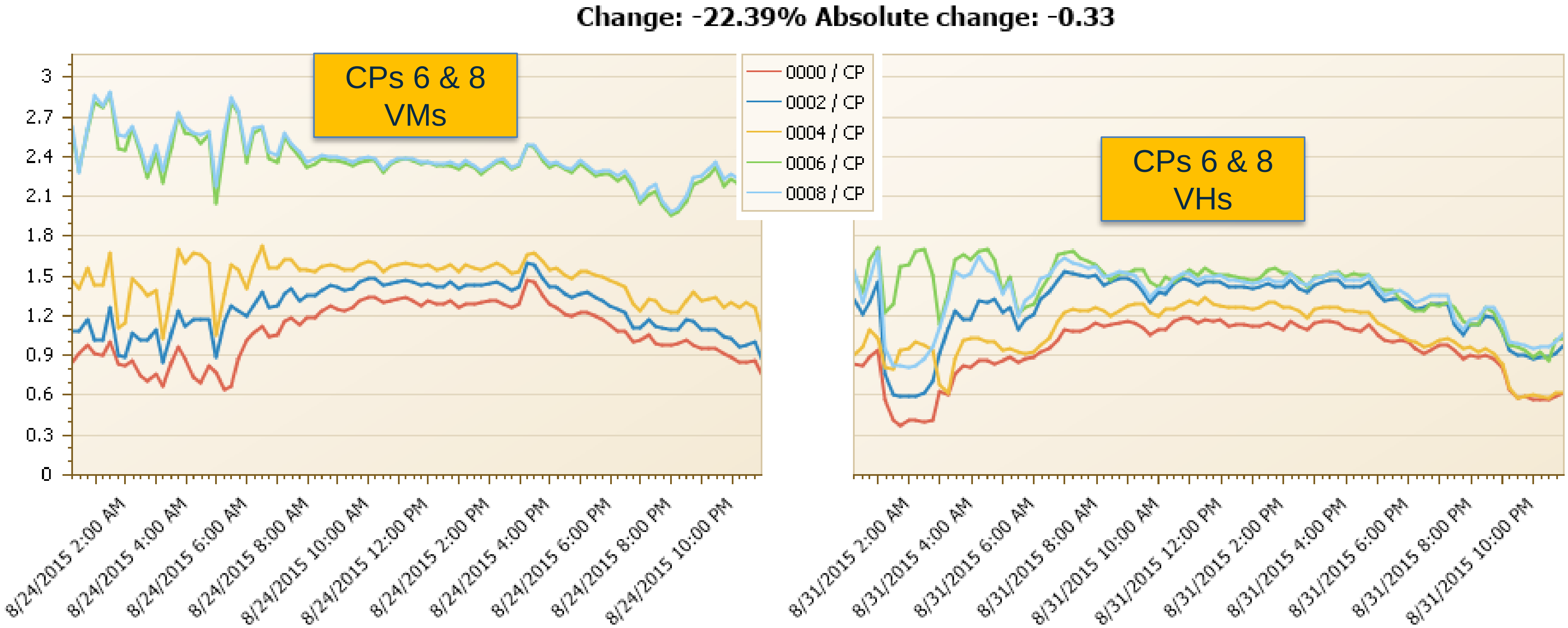
LPAR Topology – Scenario 1

	Chip 2
Drawer 2 Node 1	SY22 VM00
	SY22 VM01
	SY22 VL02
	SY18 VH00
	SY18 VH01
	SY18 VH02

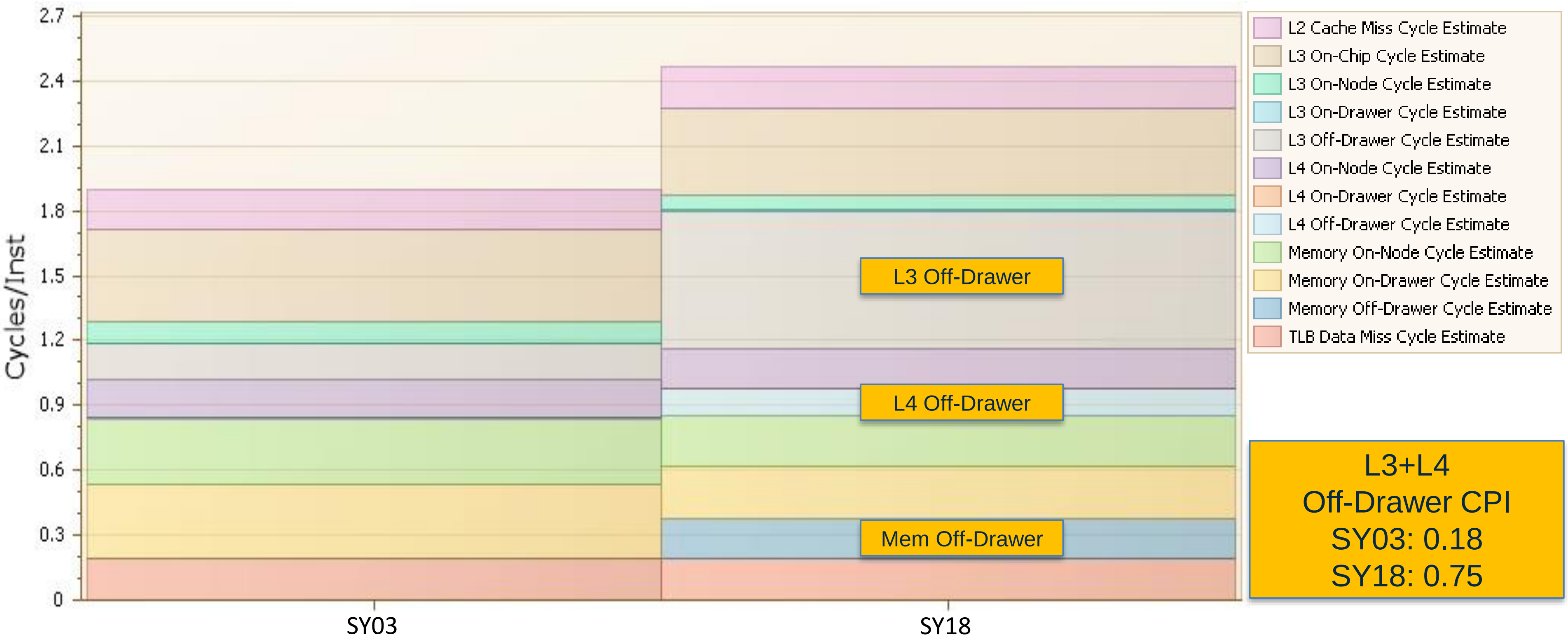
	Chip 2	Chip 3
Drawer 3 Node 1	SY18 VM03	**SY03 VH00**
	SY18 VM04	**SY03 VH01**
	SY16 VM00	**SY03 VH02**
	SY16 VM01	SY03 VM03
	SY16 VL02	SY03 VM04
	SY16 VL03	SY16 VL04
	SY20 VM00	
	SY20 VL01	

RNI Impact by Logical CP – Scenario 1

SHARE
EDUCATE • NETWORK • INFLUENCE

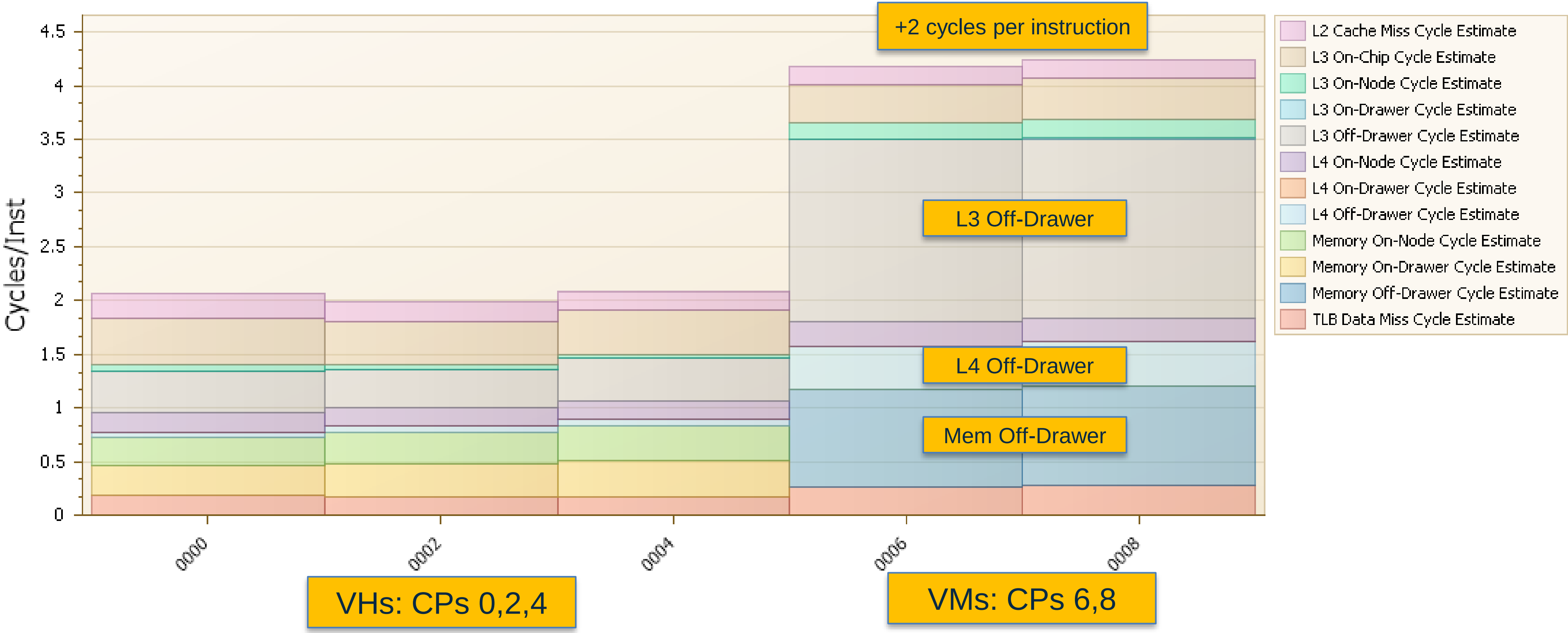


Estimated Impact Cache Misses – S1

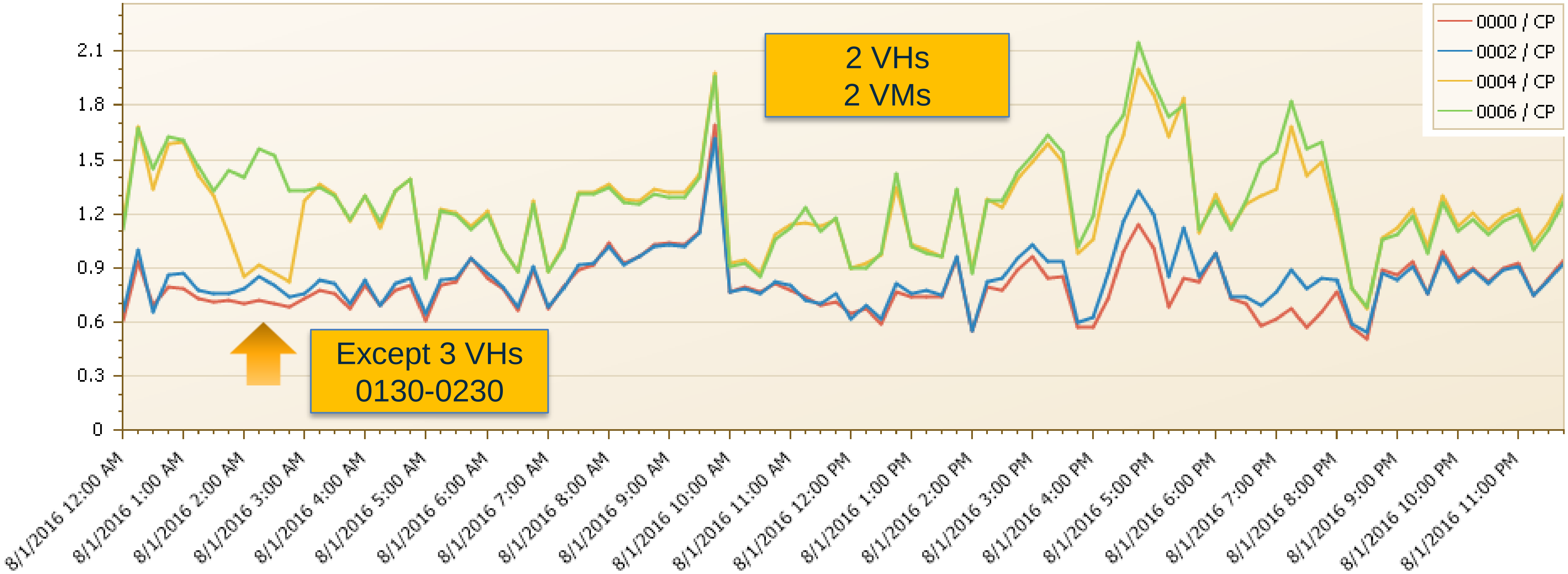


L3+L4
Off-Drawer CPI
SY03: 0.18
SY18: 0.75

Impact Cache Misses by Logical CP



LPAR Topology – Scenario 2

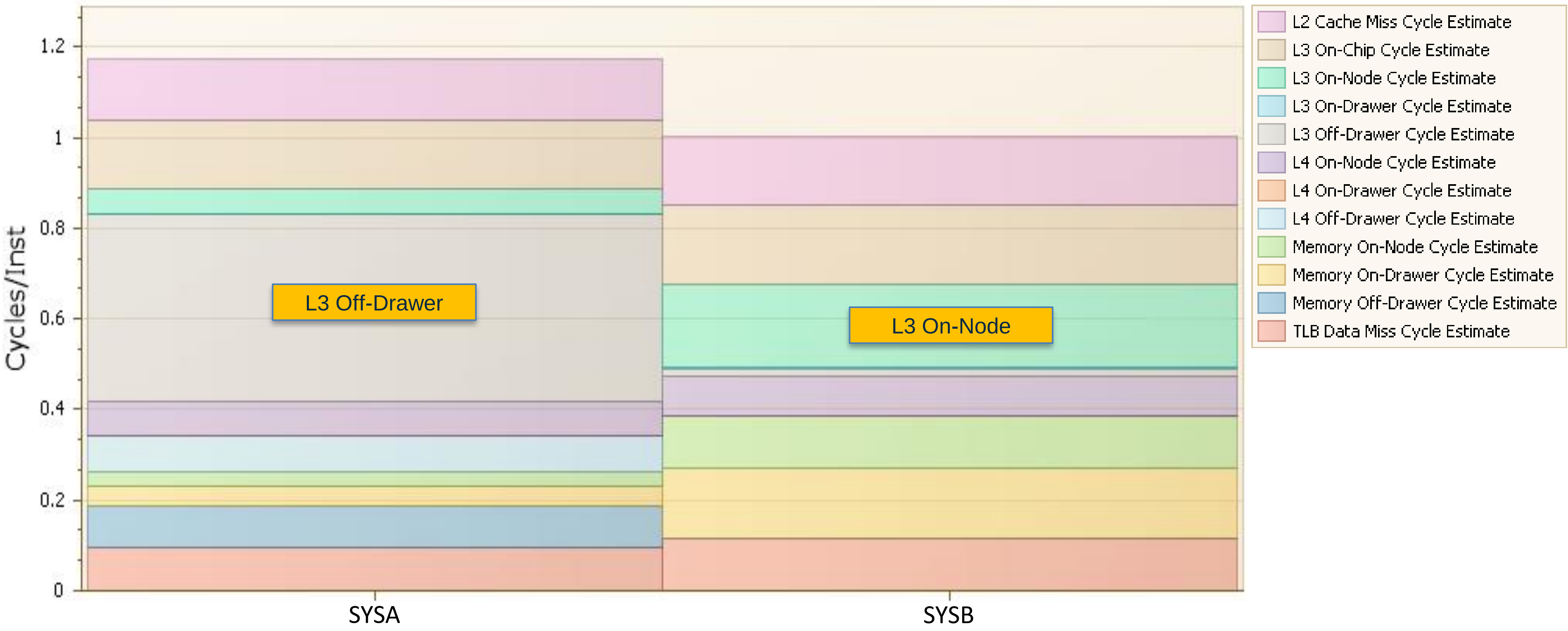


LPAR Topology – Scenario 2

	Chip 1	Chip 2	Chip 3
Drawer 3 Node 1	SYSA VM02	SYSA VM03	**SYSB VH00**
	SYSB VM04	SYSB VL05	**SYSB VH01**
		SYSB VL06	**SYSB VH02**
			SYSB VH03

		Chip 2	
Drawer 4 Node 1		**SYSA VH00**	
		SYSA VH01	

Estimated Impact Cache Misses – S2



LPAR Topology – Scenario 3

- Initially PR/SM allocated VHs for both primary LPARs in same node
- LPAR memory increase forced PR/SM to distribute VHs across drawers

Drawer 2 Node 1	Chip 1	Chip 2
Before: 1 node 480 MB L4L cache	SY07 VH00	SY17 VH00
	SY07 VH01	SY17 VH01
	SY07 VH02	SY17 VH02
	SY07 VH03	SY17 VH03
	SY07 VH04	SY17 VH04

Drawer 2 Node 1	Chip 2
	SY17 VH00
	SY17 VH01
	SY17 VH02
	SY17 VH03
	SY17 VH04

After: 2 nodes 960 MB L4L cache

Drawer 3 Node 1	Chip 2
	SY07 VH00
	SY07 VH01
	SY07 VH02
	SY07 VH03
	SY07 VH04

Impact of Topology Change – S3

- Improved % L1 misses sourced from L4 local cache (L4LP)

  2.3*(0.4*L3P + 1.6*L4LP + 3.5*L4RP + 7.5*MEMP) / 100

- 11.5% reduction in RNI ➡ 6% reduction in CPU

	L4LP	L4RP	MEMP	RNI
Before	4.38%	0.91%	4.85%	1.48
After	5.84%	0.59%	3.82%	1.31

Maximize Work Executing on VHs

Vertical CP Assignments

- PR/SM assigns logical CPs as VH / VM / VL based on LPAR weights and the number of physical CPs
- Methodology for vertical CP assignments for each LPAR
  - % Share = LPAR Weight / Sum of LPAR Weights
  - CP Share = % Share * Physical CPs
  - Assign that CP Share as VHs, VMs, and VLs
    - 1 or 2 VMs with >= 50% share of a CP
    - Remaining integer CP Share as VHs

Vertical CP Assignments – Examples

- Example 1
- LP01 VM: 80% share
  LP02 VMs: 60% share

8 CPs	Wgt	% Shr	CP Shr	Log CP	VHs	VMs	VLs
LP01	300	60%	4.8	6	4	1	1
LP02	200	40%	3.2	5	2	2	1

- Example 2
- LP11 VMs: 65% share
  LP12 VMs: 70% share
  LP13 VM: 30% share

8 CPs	Wgt	% Shr	CP Shr	VHs	VMs	VLs
LP11	550	55%	3.3	2	2	1
LP12	400	40%	2.4	1	2	1
LP13	50	5%	0.3	0	1	1

Vertical CPs – z13 Exception

- z13 prior to mid-2016
- LP02 VM: 80% share

9 CPs	% Shr	CP Shr	VHs	VMs	VLs
LP01	80%	7.2	6	2	1
LP02	20%	1.8	1	1	1

- Benefit: PR/SM more likely to configure 2 VMs on same chip

- 0 VHs as of z13 MCL Bundle 24 (6/2016)
- LP02 VMs: 90% share

9 CPs	% Shr	CP Shr	VHs	VMs	VLs
LP01	80%	7.2	6	2	1
LP02	20%	1.8	0	2	1

Maximize Work on VHs – LPAR Weights

- Increase weights for high CPU LPARs
- Tailor weights to maximize assignment of VHs
- Customize weights by shift to reflect changes in workload
- Configure fewer, larger LPARs

Maximize Work on VHs – Example 1

- Small adjustments to LPAR weights may increase work executing on VHs (and thus reduce RNI)

12 CPs	% Shr	CP Shr	VHs	VMs	VLs	RNI
Before	70%	8.4	7	2	3	
After	71%	8.52	8	1	3	-2%

Maximize Work on VHs – Example 2

- "Ordinary" LPAR weight configurations can result in 0 VHs (post Bundle 24)

6 CPs	% Shr	CP Shr	VHs	VMs	VLs
LP01	30%	1.8	0	2	2
LP02	30%	1.8	0	2	2
LP03	20%	1.2	0	2	1
LP04	20%	1.2	0	2	1

6 CPs	% Shr	CP Shr	VHs	VMs	VLs
LP01	34%	2.04	1	2	1
LP02	34%	2.04	1	2	1
LP03	16%	0.96	0	1	2
LP04	16%	0.96	0	1	2

- With minor weight changes 33% of workload could execute on VHs

Maximize Work on VHs – Example 3

- LP04 will share 50% of workload with LP01 in future but is currently idle

13 CPs	% Shr	CP Shr	VHs	VMs	VLs
LP01	40%	5.2	4	2	4
LP02	10%	1.3	0	2	1
LP03	10%	1.3	0	2	1
LP04	40%	5.2	4	2	4

13 CPs	% Shr	CP Shr	VHs	VMs	VLs
LP01	75%	9.75	9	1	3
LP02	10%	1.3	0	2	1
LP03	10%	1.3	0	2	1
LP04	5%	0.65	0	1	1

- Reduced overall RNI for the primary Production workload by over 10%

Maximize Work on VHs – # of Physical CPs

- Utilize sub-capacity processor models
- Activate On/Off Capacity on Demand (CoD) during monthly peak intervals
- Install or deploy additional hardware

Sub-capacity Hardware Models

SHARE
EDUCATE • NETWORK • INFLUENCE

- If single engine speeds or other considerations do not require full capacity models, sub-capacity models can be selected to add CPs without incurring hardware expense
- More VHs, reduced RNI, more realized capacity
- From z13-710, sub-capacity models could add 7-15 CPs

Total	zEC12-711	z13-710	z13-617	z13-525
MSUs	1593	1632	1610	1603

Activate On/Off Capacity on Demand

SHARE EDUCATE • NETWORK • INFLUENCE

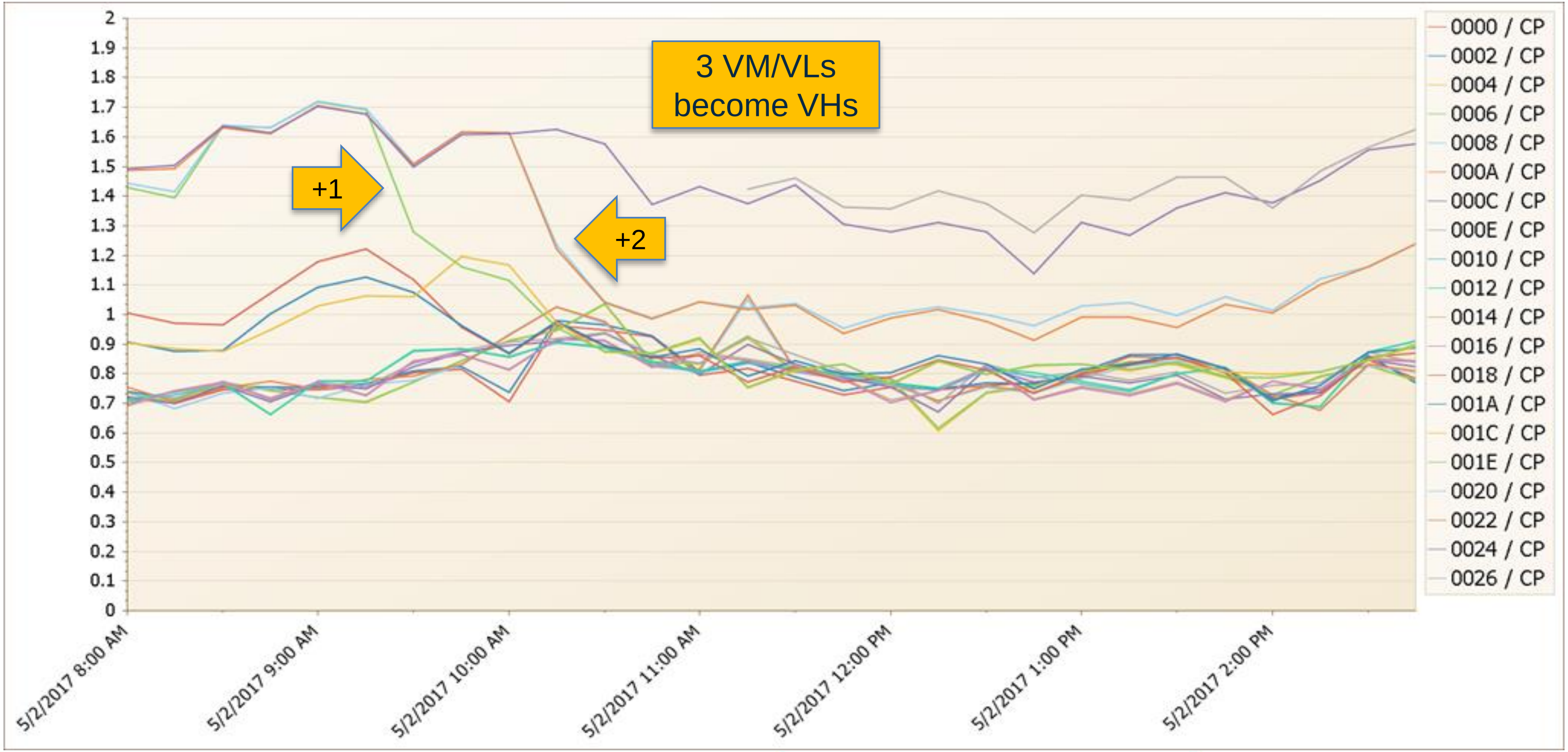
- If monthly peak intervals are predictable, On/Off Capacity on Demand (CoD) can be activated during those peak intervals with minimal incremental hardware expense

- Processor cache impact
  - More VHs and more work executing on VHs
  - Reduced RNI
  - Reduced CPU consumption
  - Reduced MLC expense

More VHs with On/Off CoD

CEC / Sysid	CEC Model		CEC VHs		Sys VHs	
	Std	CoD	Std	CoD	Std	CoD
C4 / SYS1	z13-728	z13-732	20	24	15	18
C3 / SYS2	z13-724	z13-728	18	21	15	17

Reduced RNI with On/Off CoD



Reduced MSUs & MLC with On/Off CoD

Sysid	CEC Model		MSUs / CP	Add VHs	% CP Util	VH RNI Reduc	MSUs Reduc
	Std	CoD					
SYS1	z13-728	z13-732	105	3	95%	96%	144
SYS2	z13-724	z13-728	102	2	95%	82%	79

Sysid	CEC Model		MSUs Reduc	$ / MSU	$ Savings	
	Std	CoD			Monthly	Annually
SYS1	z13-728	z13-732	144	$181	$ 25,999	$ 311,986
SYS2	z13-724	z13-728	79	$181	$ 14,382	$ 172,583

Perspectives on Hardware Capacity

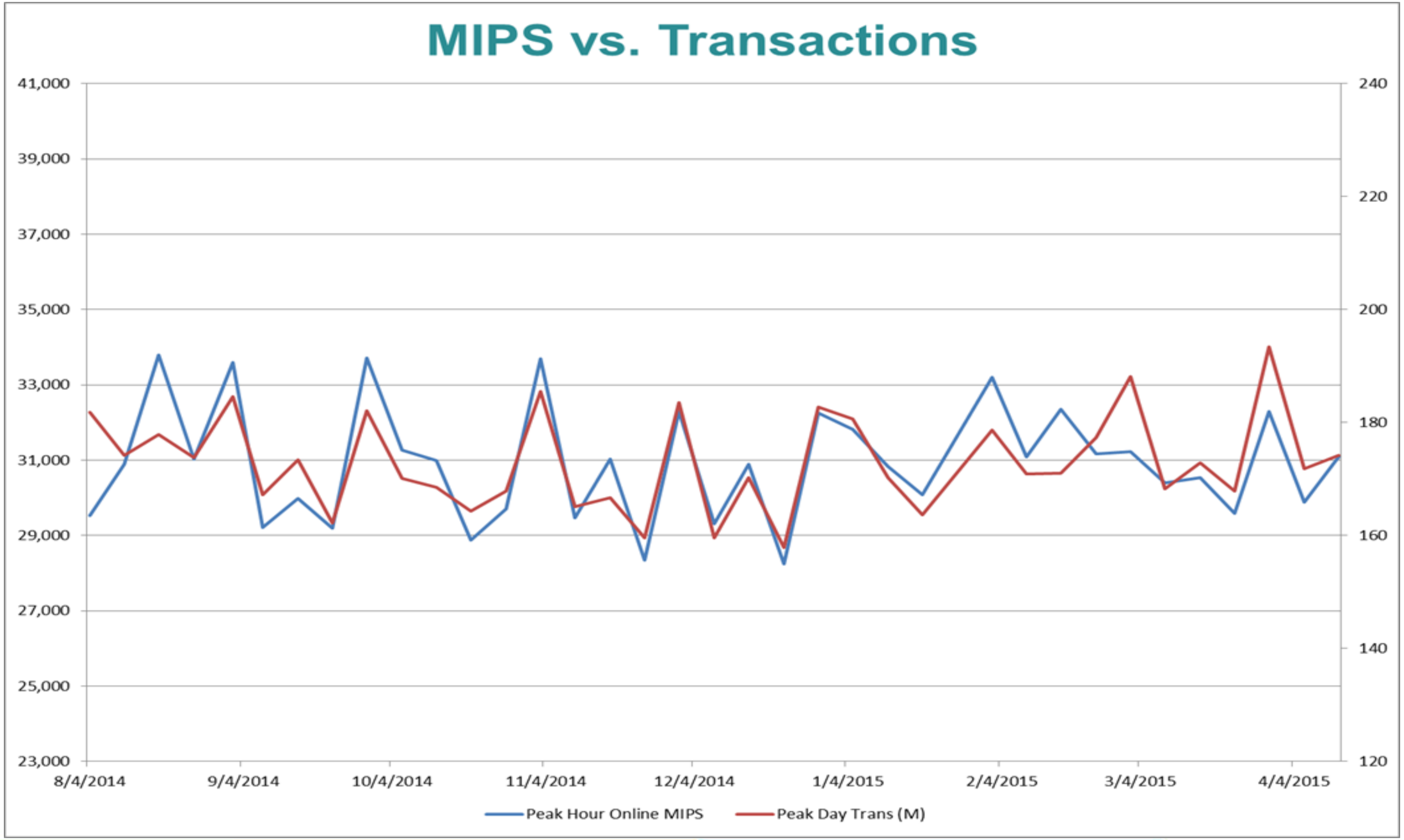
SHARE
EDUCATE • NETWORK • INFLUENCE

- Reexamine traditional perspective that running mainframe at high utilization is most cost-effective strategy
- Reexamine traditional "just in time" approach to deploying previously purchased capacity
- In today's mainframe budgets software typically represents a much larger expense than hardware
  - <u>Recurring</u> savings from substantially reducing MLC through processor cache efficiencies may justify <u>one-time</u> expense for acquisition of additional hardware capacity

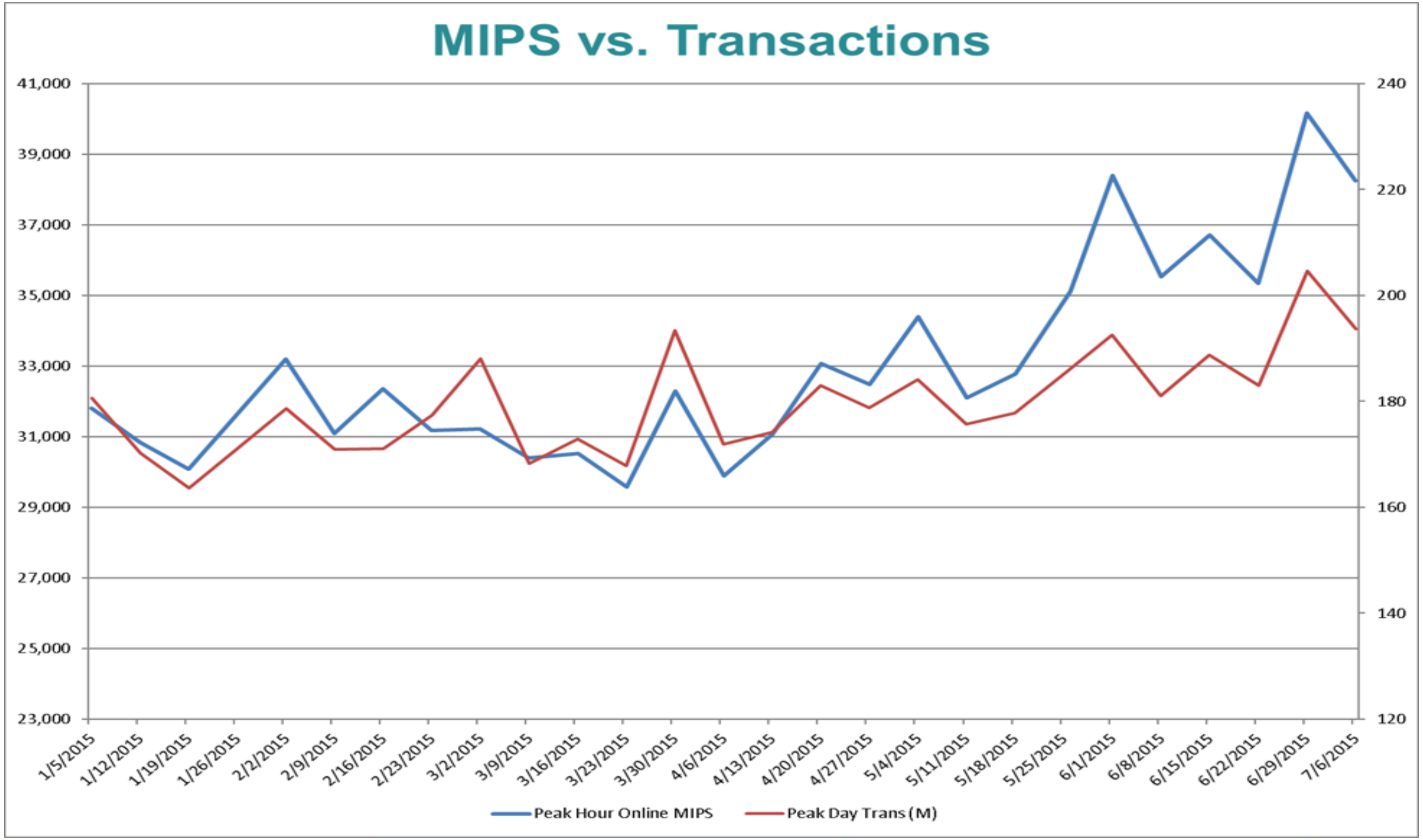
ISVs as Barrier to Deploying Capacity

- ISV licenses appear to be a primary barrier
  - To deploying owned surplus capacity
  - To considering the acquisition of surplus capacity
- Strongly encourage effort to proactively convert all ISV licenses from capacity-based to usage-based
- It can be done … and will position you to have options for substantial financial savings going forward!

Deploy HW Case 1 – zEC12 Baseline

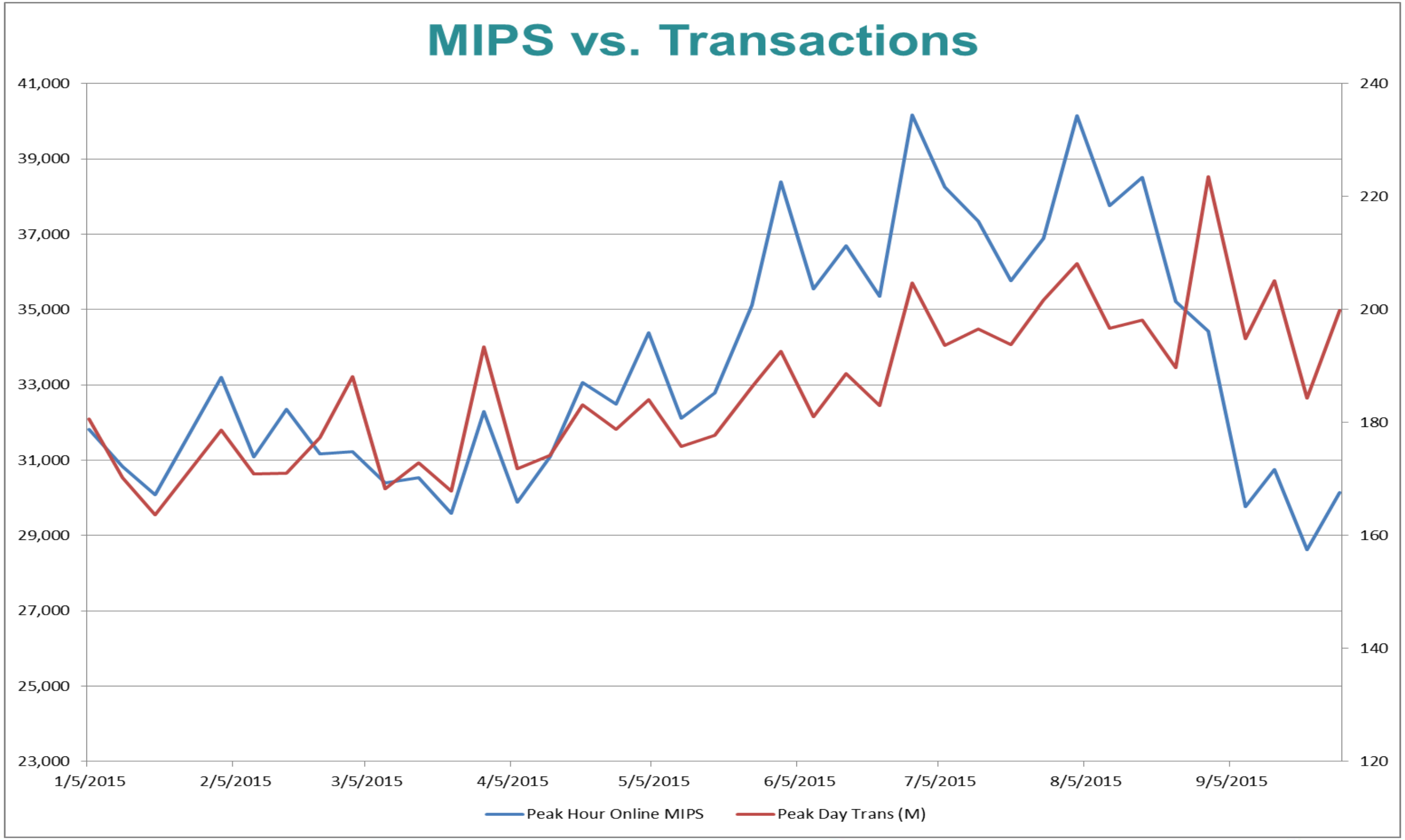


Deploy HW Case 1 – z13 CPU Lift



4K MIPS vs. zEC12

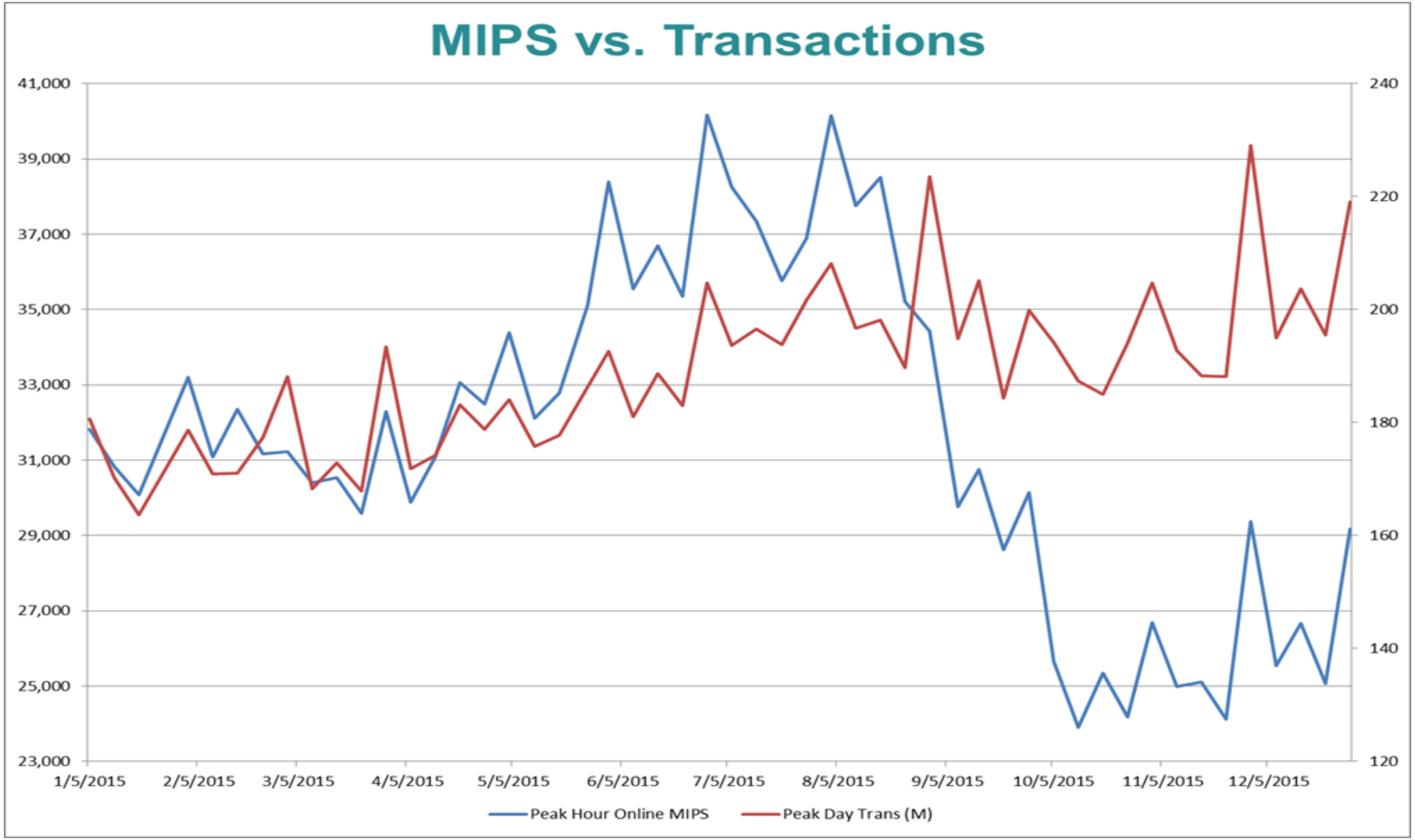
Deploy HW Case 1 – 716 Upgrades



-5K MIPS vs. zEC12

-9K MIPS vs. 711

Deploy HW Case 1 – 726 Upgrades



-9K MIPS vs. zEC12

-13K MIPS vs. 711

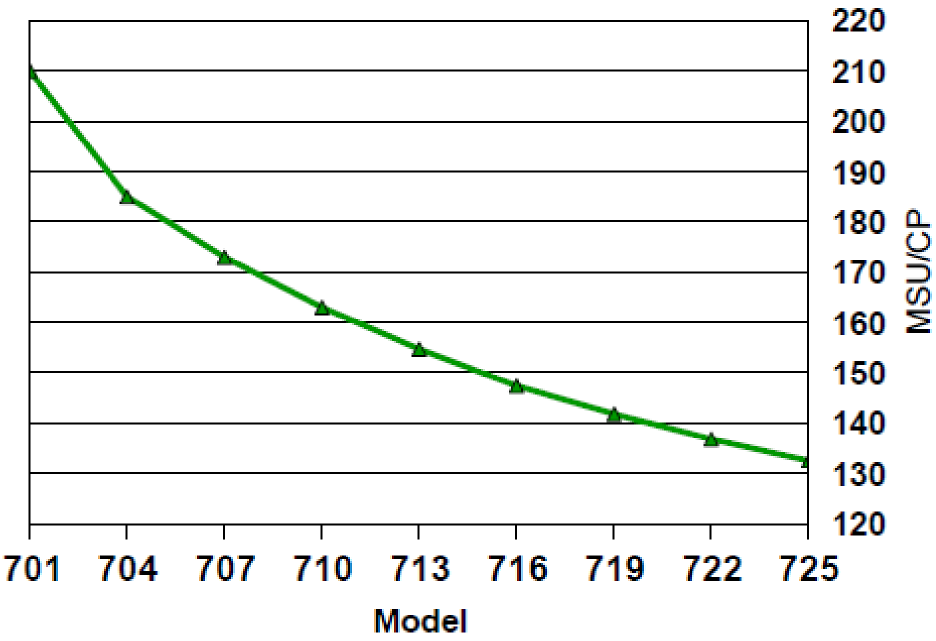
-4K MIPS vs. 716

Multiprocessing Effect

- Adding CPs increases overhead required to manage interactions between hardware & workloads

- Thus MSU/CP ratios for IBM processor ratings not linear

- Ratings based on LSPR workloads at 90% utilization



Savings from MSU/CP Ratings

- If workload remains same, CEC utilization decreases, and MP overhead will be minimal
- Lower MSU/CP rating translates directly into reduced MSUs for same workload

CEC	MSUs/CP	vs zEC12-711	vs z13-711
zEC12-711	144.8		-9.7%
z13-711	160.4	10.7%	
z13-716	147.4	1.8%	-8.1%
z13-726	131.3	-9.3%	-18.1%

Twofold Benefits of Deploying Capacity

- Consume less CPU due to operating efficiencies
  - Optimize LPAR Topology
  - Maximize work executing on VHs

- CPU that is consumed translates into fewer MSUs
  - Lower processor MSU/CP ratings

Twofold Benefits of Deploying Capacity

SHARE
EDUCATE • NETWORK • INFLUENCE

- "Acquiring more hardware capacity than needed to support the workload allows multiple benefits.
  - First it enables taking advantage of the lower MSU per engine rating of larger processor configurations, and
  - Second it enables benefits from the impacts of low utilization on processor capacity."

  - Kathy Walsh, IBM Whitepaper, rev. April 2017

Deploy HW Case 2 – z13 Upgrade

- Another High RNI workload
- Same two primary takeaways
  - Capacity shortfall migrating to the z13
  - Significant reduction in RNI by deploying additional hardware

Deploy HW Case 2 – Double Whammy

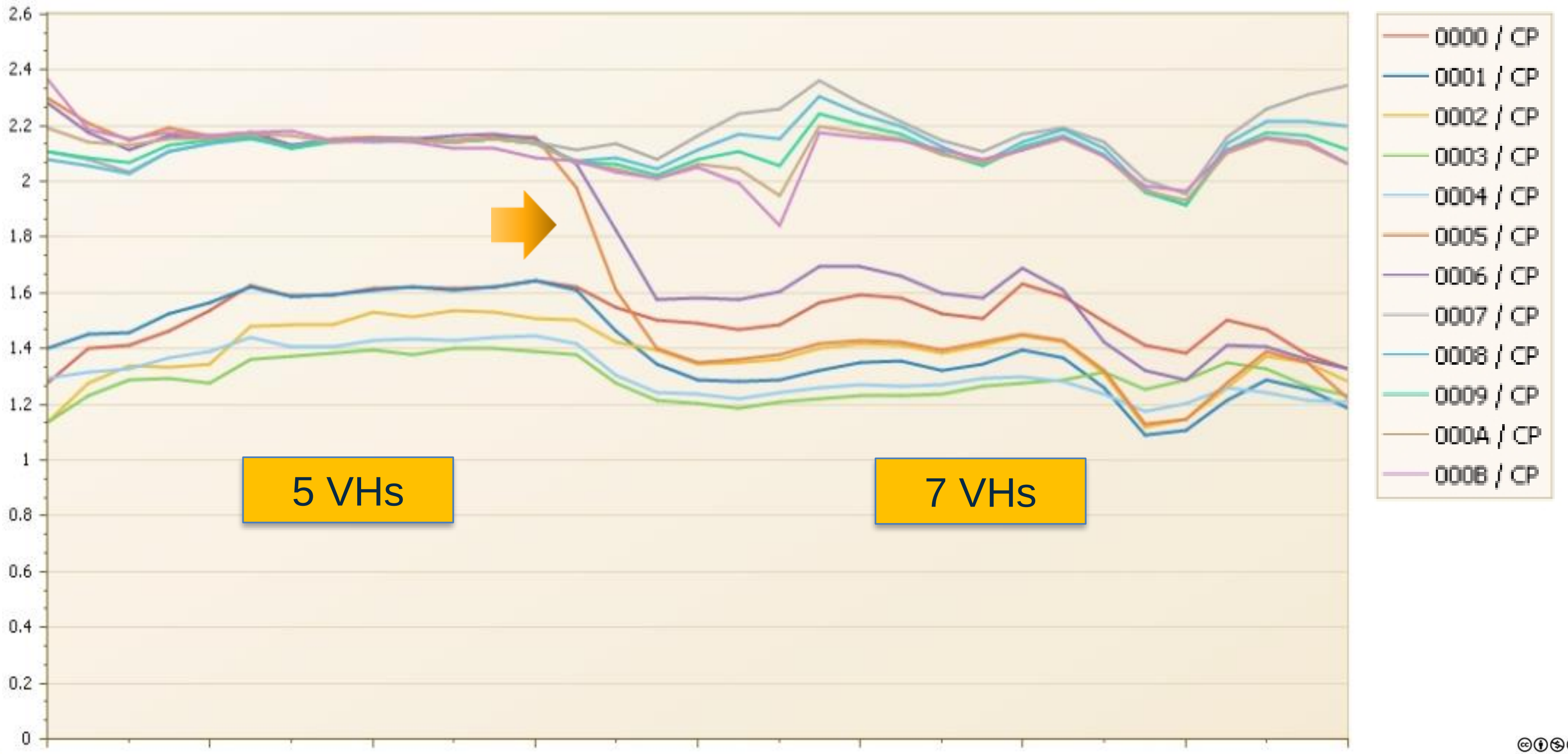
SHARE EDUCATE • NETWORK • INFLUENCE

- Lateral MIPS upgrades reduce number of physical CPs
- Compounds the negative impact of the upgrade
  - Underachieve the z13 10% capacity rating increase
  - Reduce the number of VHs further impacting RNIs

Ratings	MIPS	MSUs
zEC12-728	28023	3301
z13-725	28130	3313

SYS1	zEC12	z13	z13+Upgr
VHs	6	5	7
RNI	1.25	1.67	1.40

Deploy HW Case 2 – RNI Impact



Summary

- Key Processor Cache Concepts and Metrics
- Optimize LPAR Topology
- Maximize Work Executing on Vertical High CPs
  - Optimize LPAR weights
  - Increase number of physical CPs

Sources

- IBM TechDocs Library
  - TD106388, “Number of Logical CPs Defined for an LPAR” (6/14/2016)
  - TD106389, “z13 HiperDispatch – New MCL Bundle Changes Vertical CP Assignment for Certain LPAR Configurations” (6/30/2016)
  - WP102705, “Leveraging MSU Ratings for Better Value”, Kathy Walsh (rev. 4/24/2017)
- Processor cache tuning and deploy hardware use case
  - Frank Kyne, “A Holistic Approach to Capacity Planning”, Cheryl Watson’s Tuning Letter (CWTL) 2015 #4
  - Kyne, “Introduction to CPU Measurement Facility”, CWTL 2016 #4
  - IntelliMagic White Paper, “How to use Processor Cache Optimization to Reduce z Systems Costs”, www.intellimagic.com/mlc
  - Todd Havekost, “Achieving Significant Capacity Improvements on the IBM z13 Processor – User Experience”, SHARE 8/2016

Questions?

- Please fill out your session evaluations

- Thank you for attending – particularly on Friday morning!!