Achieving Significant Capacity Improvements on the IBM z13: User Experience

Todd Havekost, USAA

Session #18345

#SHAREorg

SHARE is an independent volunteer-run information technology association that provides **education**, professional **networking** and industry **influence**.

Agenda

- USAA Context and Configuration
- z13 Capacity Concepts and Initial Experiences
- Key Processor Hardware Metrics
- Actions to Improve Capacity and Impact on Metrics
- Financial and Other Considerations

USAA

- Financial services company facilitating the financial security of the military community
- IT at USAA plays a critical role in fulfilling this mission
  - 100% availability 24x7 so that service members can rapidly complete their business online wherever deployed
  - Cost-effective to support highly competitive products and services for our members

Configuration Overview

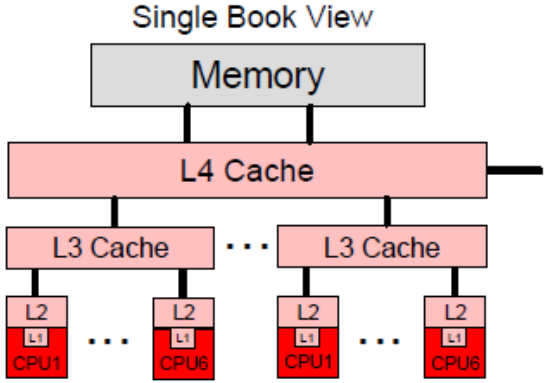
- Software – z/OS 2.1, IMS V13, CICS V5.1, DB2 V11, MQ V8
- Workload – 3 primary sysplexes across 4 CECs

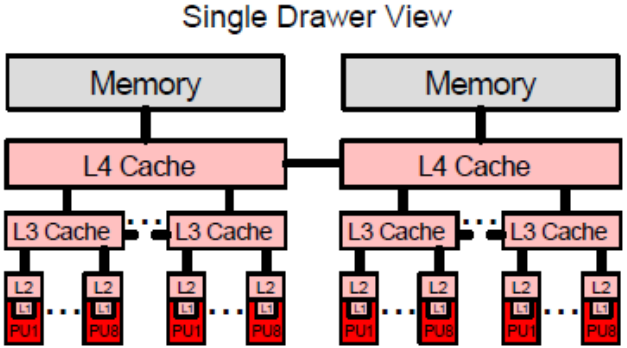
CEC A	CEC B	CEC F	CEC G
Prod Onln 1 Prod Onln 5 Prod Spcl 1	Prod Onln 2 Prod Onln 6	Prod Onln 3 Prod Onln 7 Prod Spcl 2	Prod Onln 4 Prod Onln 8
Bank Onln 1	Bank Onln 2	Bank Onln 3 Bank Spcl 1	Bank Onln 4 Bank Spcl 2
Dev Onln 1	Dev Onln 2 Dev Spcl 1	Dev Onln 3	Dev Onln 4

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

- 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





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

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

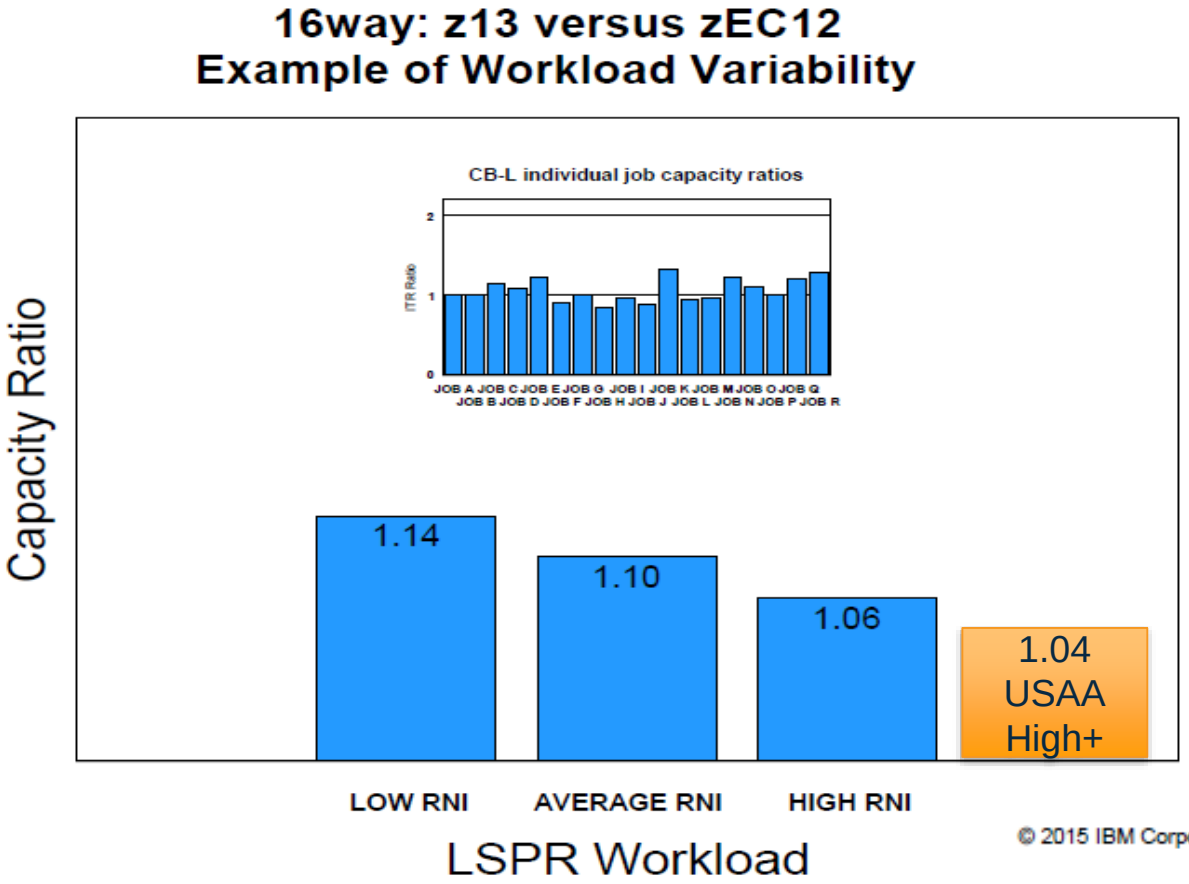
L1MP: 3.8%

RNI: 1.7

LSPR Workload: HIGH+

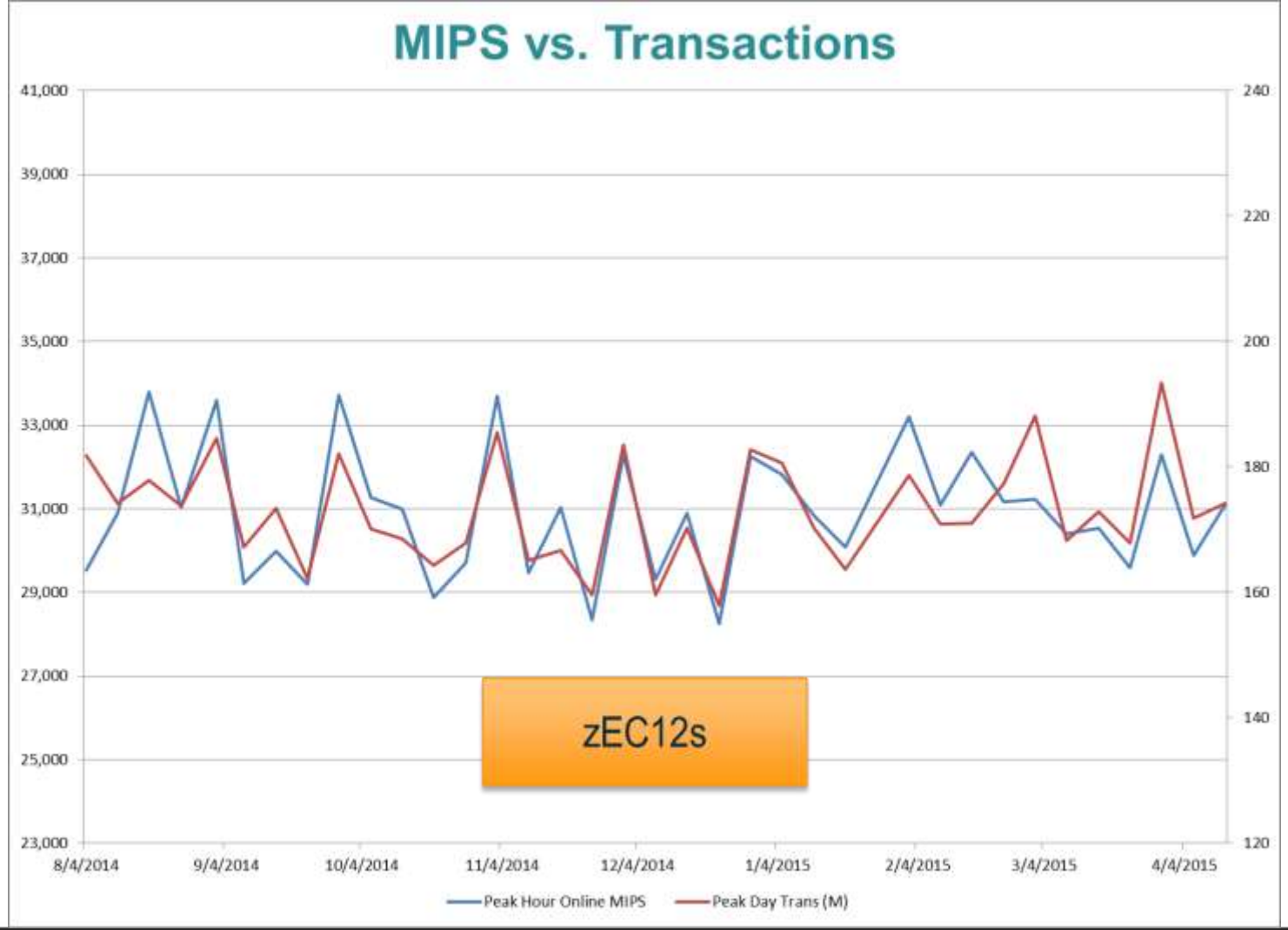
Current table applies to z10 EC, z10 BC, z196, z114, zEC12, zBC12 and z13 CPU MF data

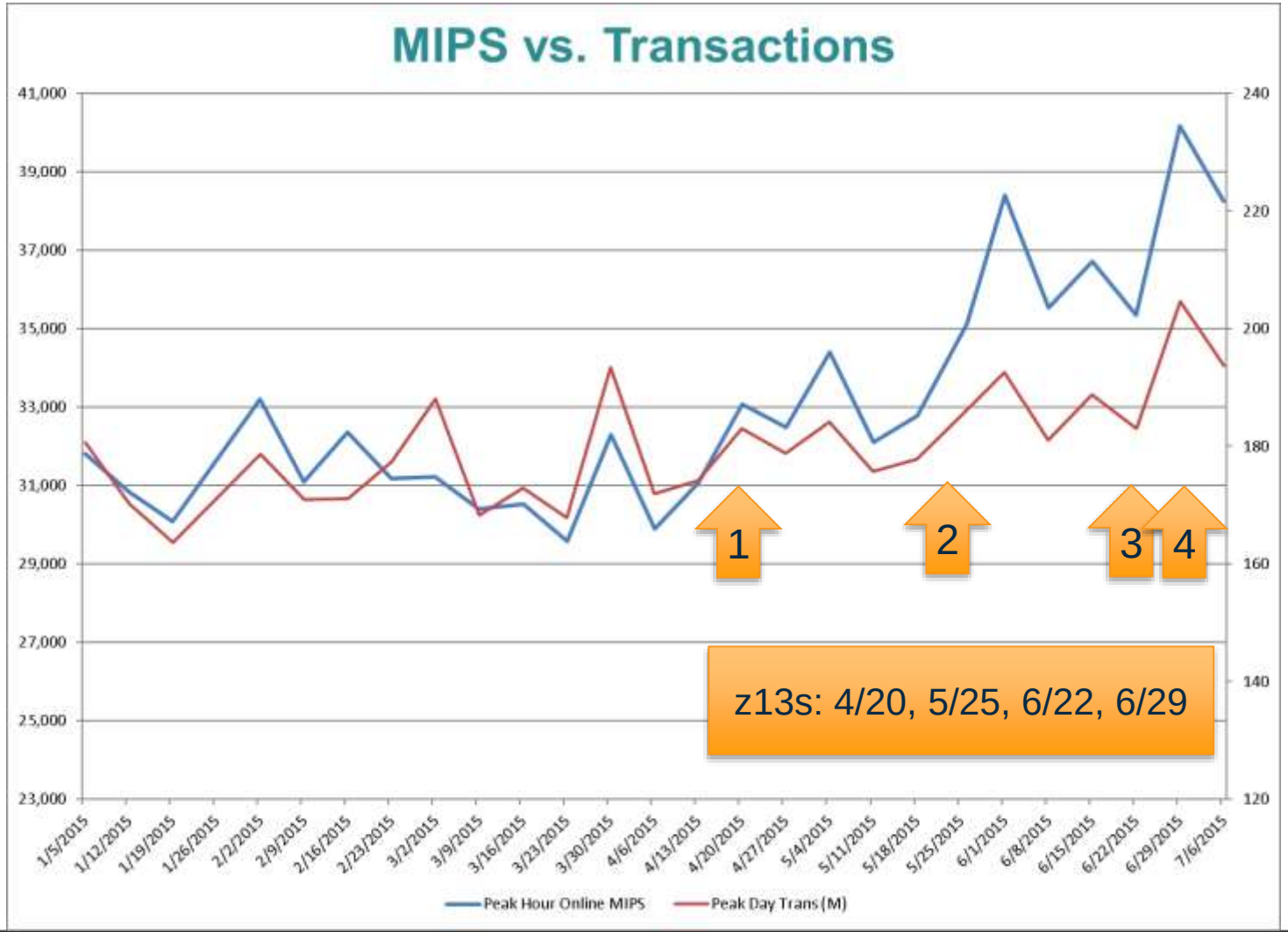
LSPR Single Image Capacity Ratios



711s	MSUs/CP
zEC12	144.8
z13	160.4
Delta	**10.7%**

MIPS vs. Transactions





+4K MIPS vs. zEC12

Key Metrics – L1MP (SMF 113 Records)

- L1 Cache Misses per 100 Instructions
  - Frequency that required data is not present in L1 cache (making processor wait)

zEC12-711	z13-711
4.47	3.81

Key Metrics – Relative Nest Intensity (RNI)

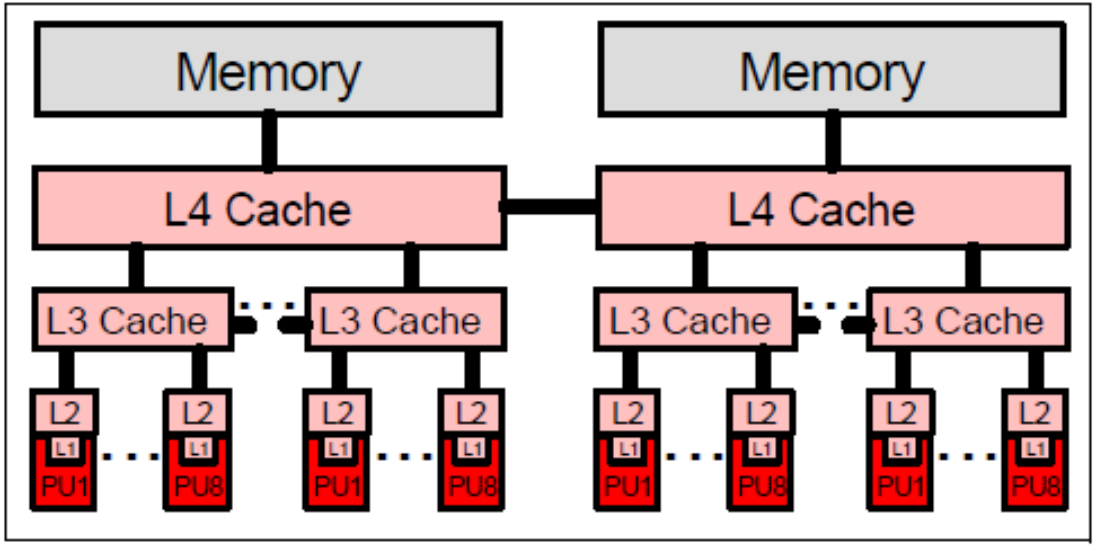
- How deep into the shared cache and memory hierarchy (“nest”) the processor must go to retrieve data
- Access time increases significantly with each additional level (increasing processor wait time)
- 2.6*(0.4*L3P + 1.6*L4LP + 3.5*L4RP + 7.5*MEMP) / 100

zEC12-711	z13-711
1.41	1.67

Key Concepts – 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



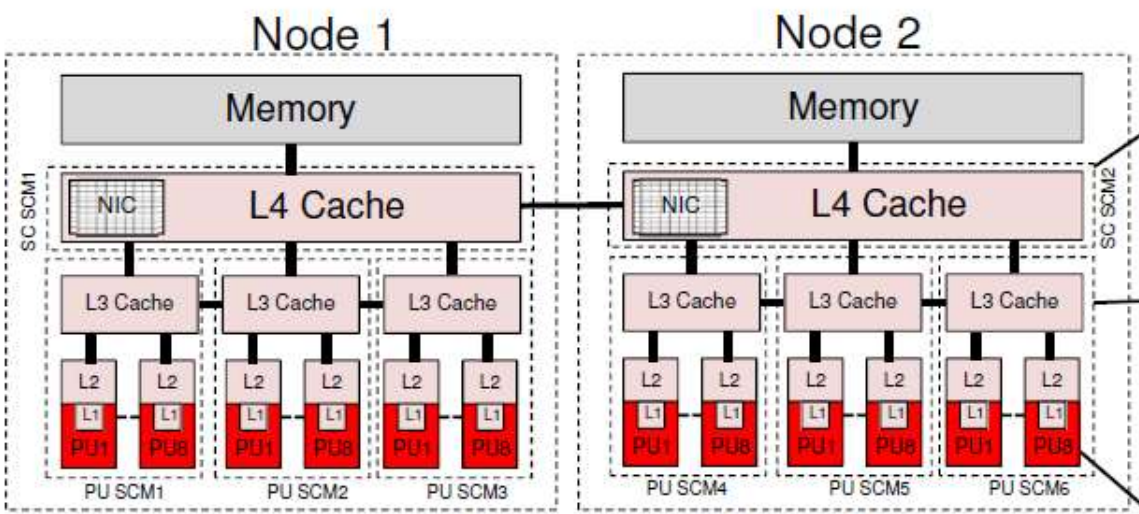
(© IBM)

Vertical CP Configuration

- Based on LPAR weights PR/SM categorizes 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)
- Work running on VHs will use same L3 cache and usually same L1 & L2 cache
- Work running on VMs & VLs is subject to being dispatched on various CPs

z13 CPC Drawer Cache Hierarchy Detail

Single CPC Drawer View (N30 Model) – 2 Nodes



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

z13-711s

z13-716s

z13-726s

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

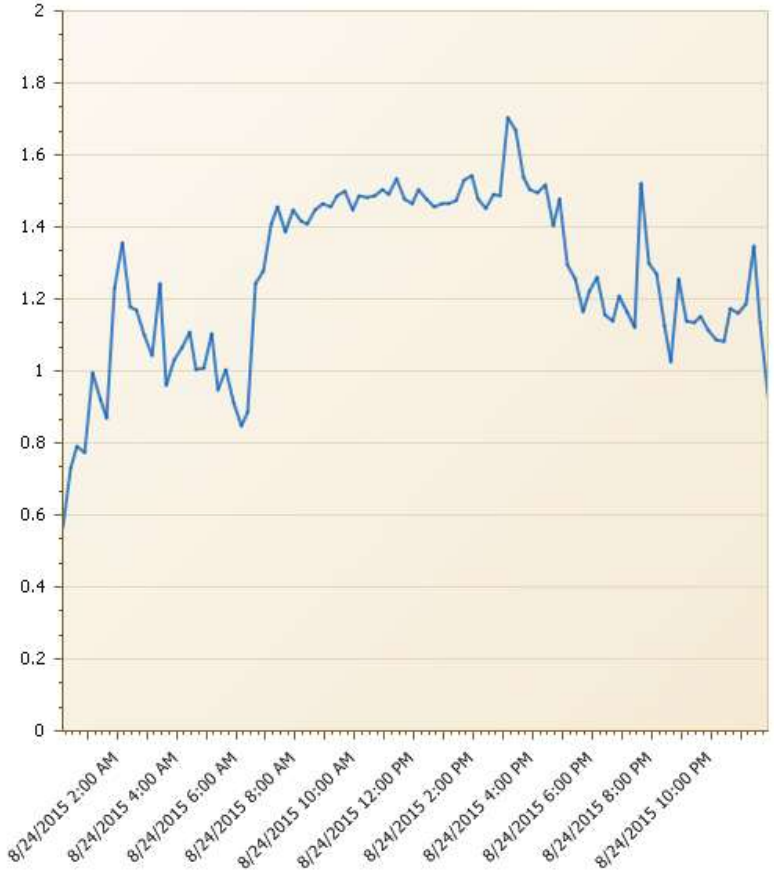
USAA Vertical CP Configurations

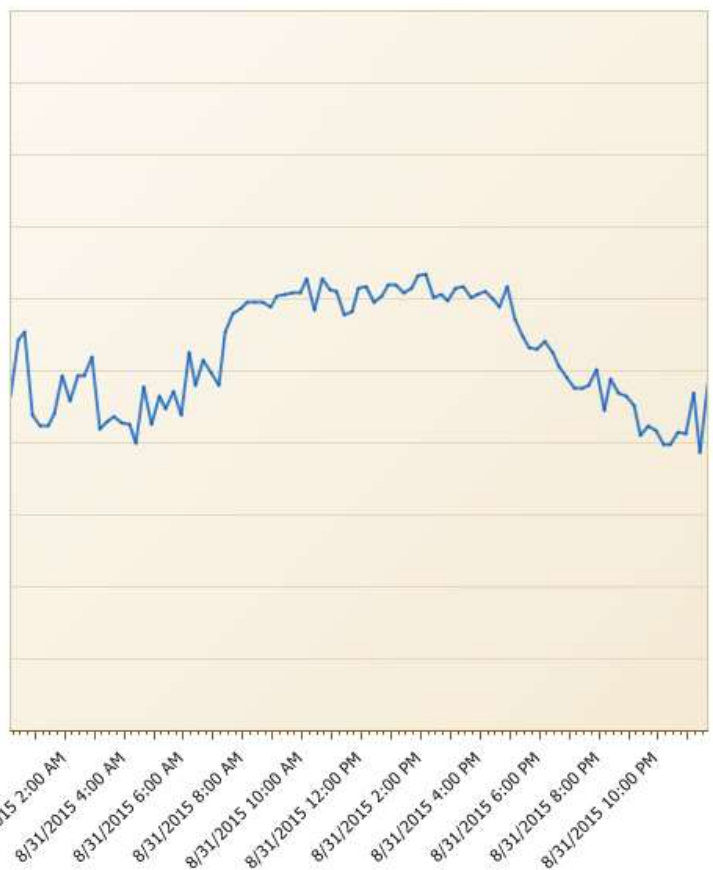
	Prod Insurance			
	zEC12	711	716	726
Highs	3	3	5	6
Mediums	1	1	0	0
Lows	2	2	0	0

Impact of Vertical High CPs on RNI

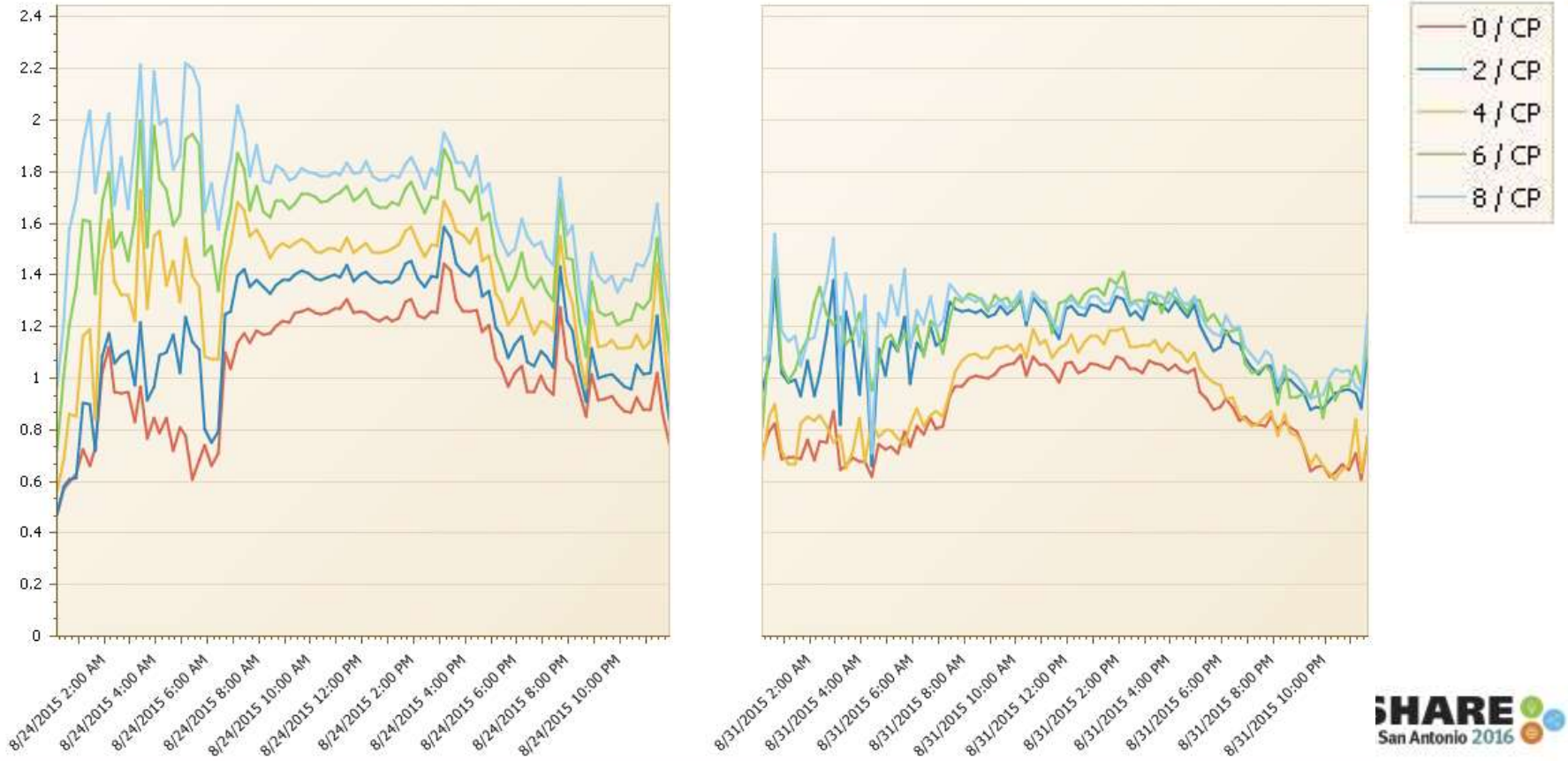
	Prod Insurance			
	zEC12	711	716	726
Highs	3	3	5	6
Mediums	1	1	0	0
Lows	2	2	0	0
RNI	1.41	1.56	1.26	1.24

RNI Impact: 711 to 716 (H003)

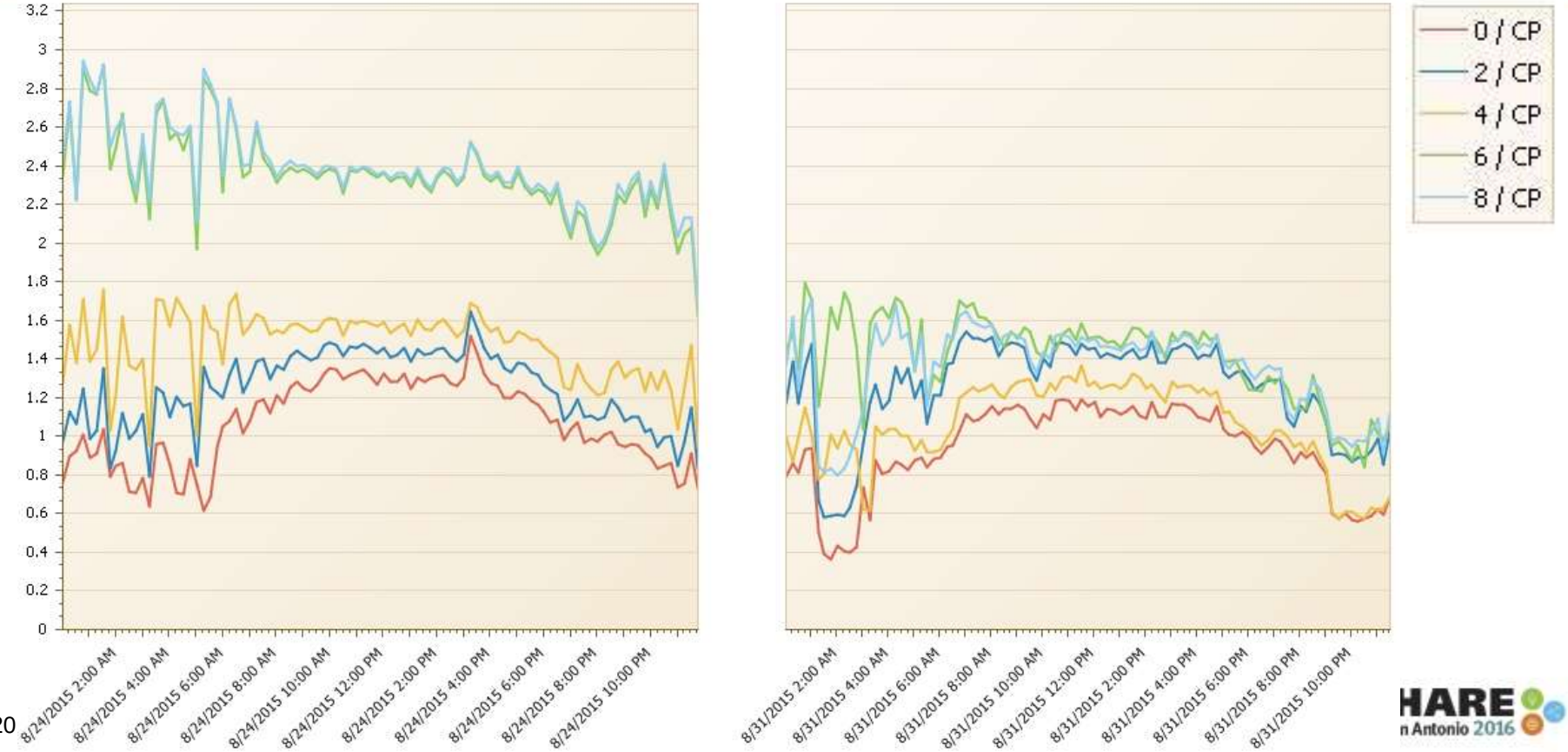




RNI Impact: 711 to 716 by CP (H003)



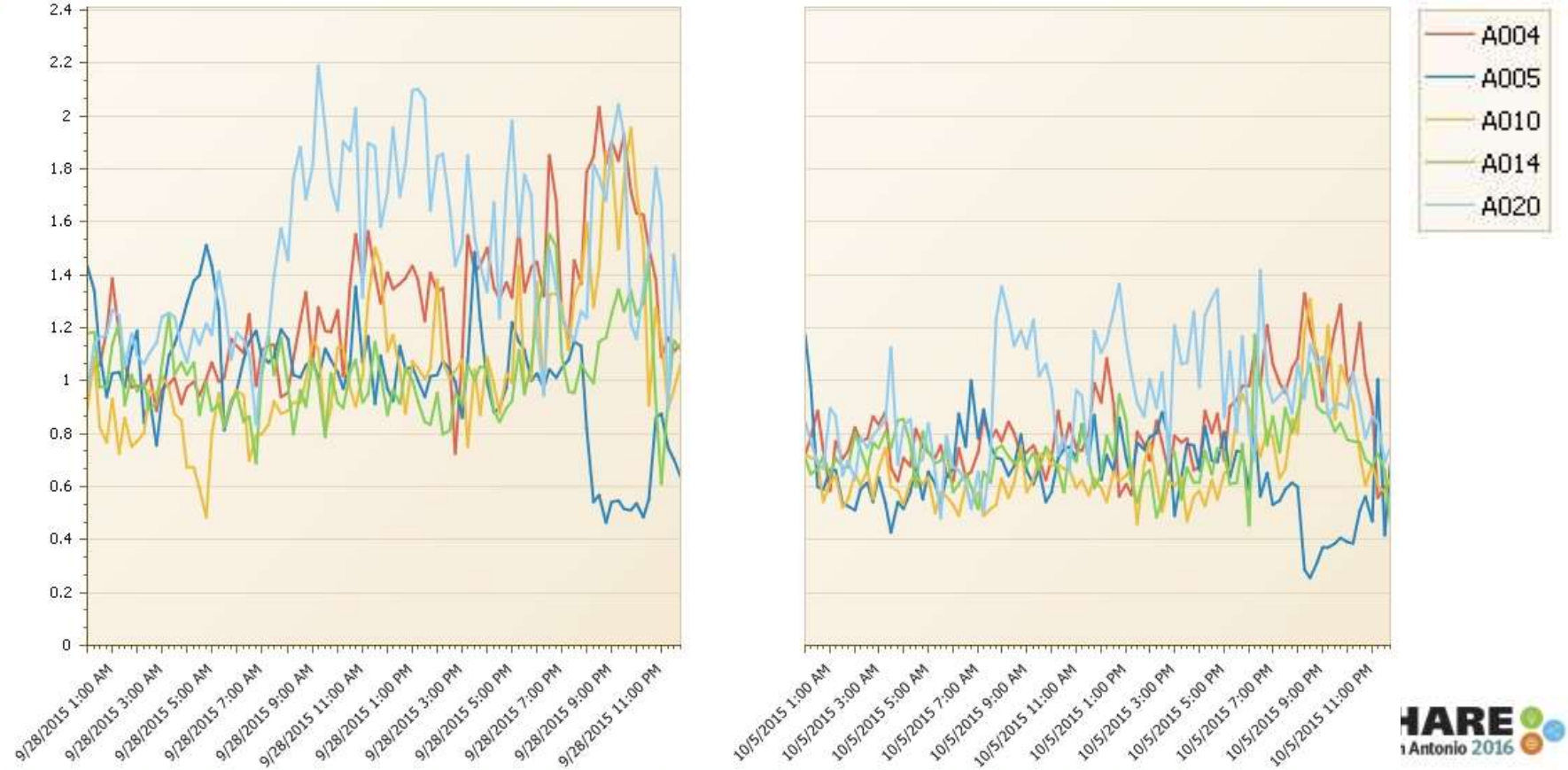
RNI Impact: 711 to 716 by CP (H018)



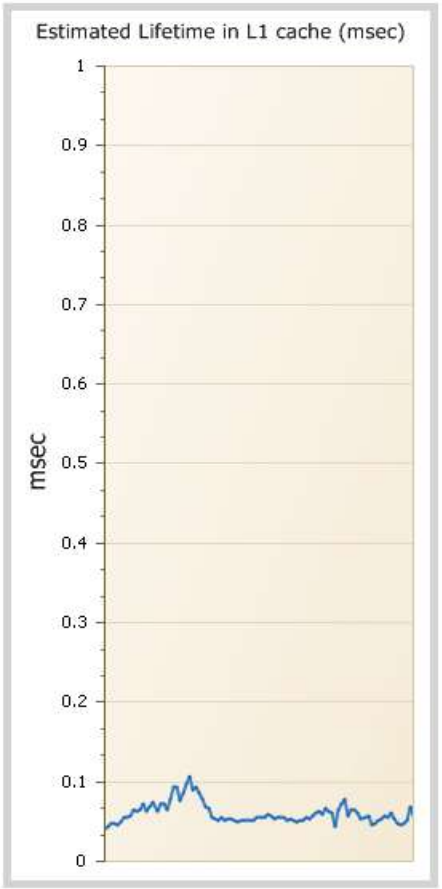
Vertical CP Configs & RNI – All Sysplexes

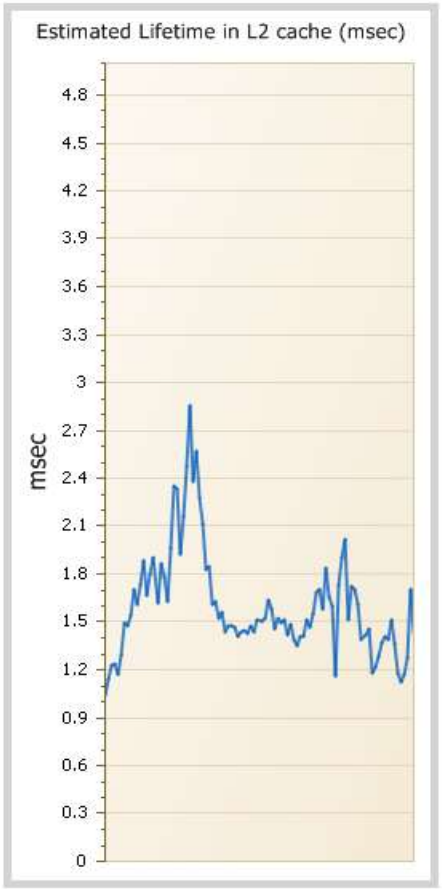
	Prod Insurance			Prod Banking			Dev/Test		
	711	716	726	711	716	726	711	716	726
Highs	3	5	6	0	0	4	0	0	3
Mediums	1	0	0	2	2	0	1	2	0
Lows	2	0	0	1	1	0	1	0	0
RNI	1.56	1.26	1.24	1.35	0.87	0.71	1.24	1.16	0.73

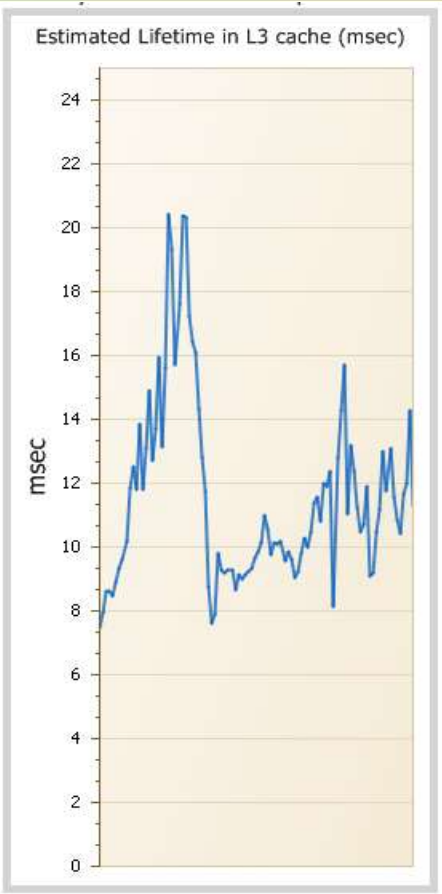
RNI Impact: 716 to 726 – Dev/Test Sysplex

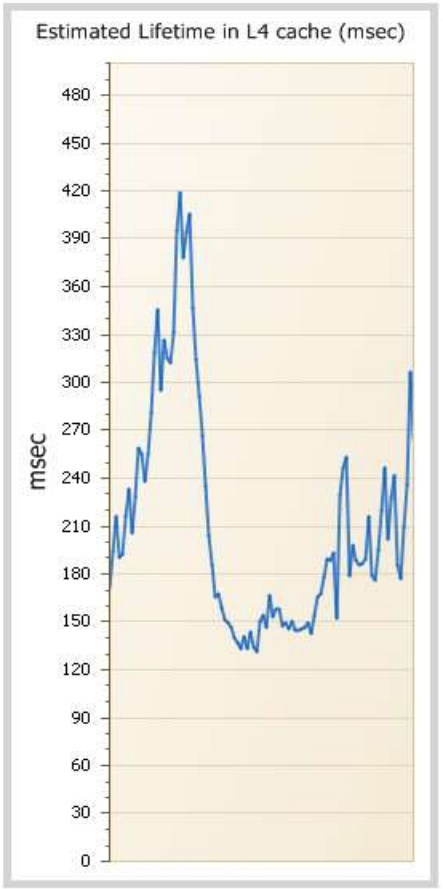


Cache Data Lifetime



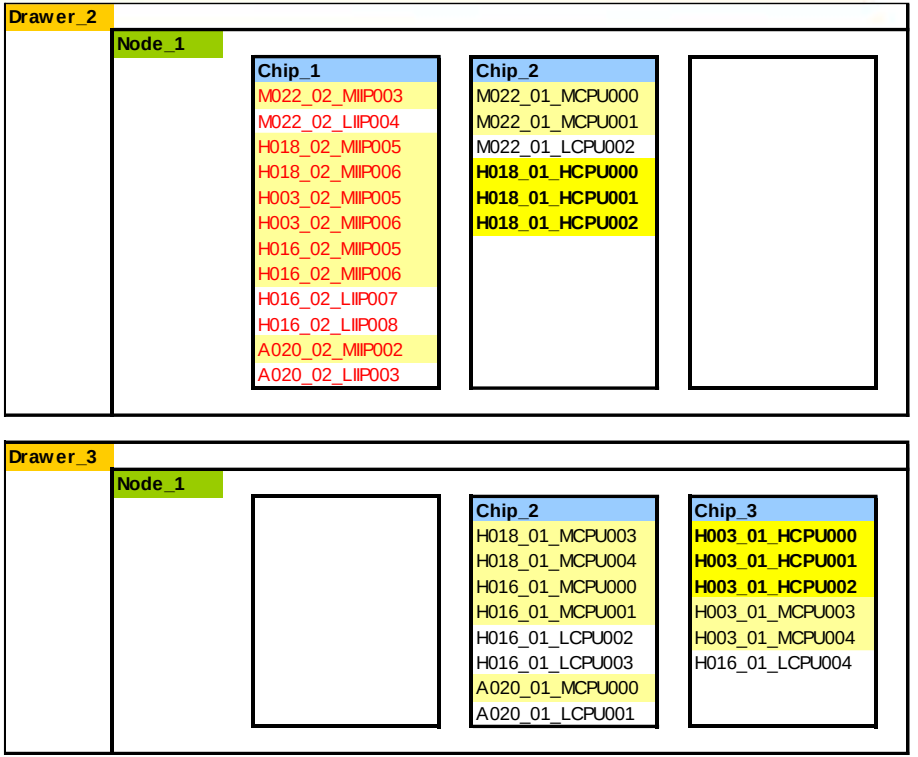






WLM Topology Report – SMF 99.14

- PR/SM dynamically assigns LPAR CPs and memory to HW chips, nodes and drawers to optimize cache efficiency
- Topology can have big impact on performance
- IBM provides tool to display graphically in Excel

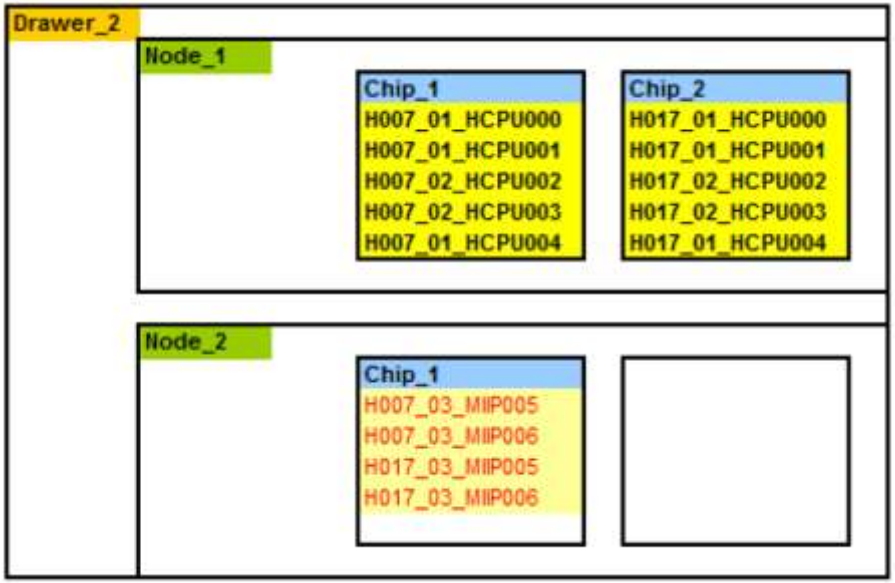


http://www-03.ibm.com/systems/z/os/zos/features/wlm/WLM_Further_Info_Tools.html#Topology

Topology Change

- IBM identified opportunity after initial 716 upgrade; LPAR memory increase forced PR/SM to distribute VHs across drawers

Before: 1 node,
480 MB L4L cache



After: 2 nodes,
960 MB L4L cache

Topology Impact

- Improved % L1 misses sourced from L4 local cache -> RNI

  2.6*(0.4*L3P + 1.6*L4LP + 3.5*L4RP + 7.5*MEMP) / 100

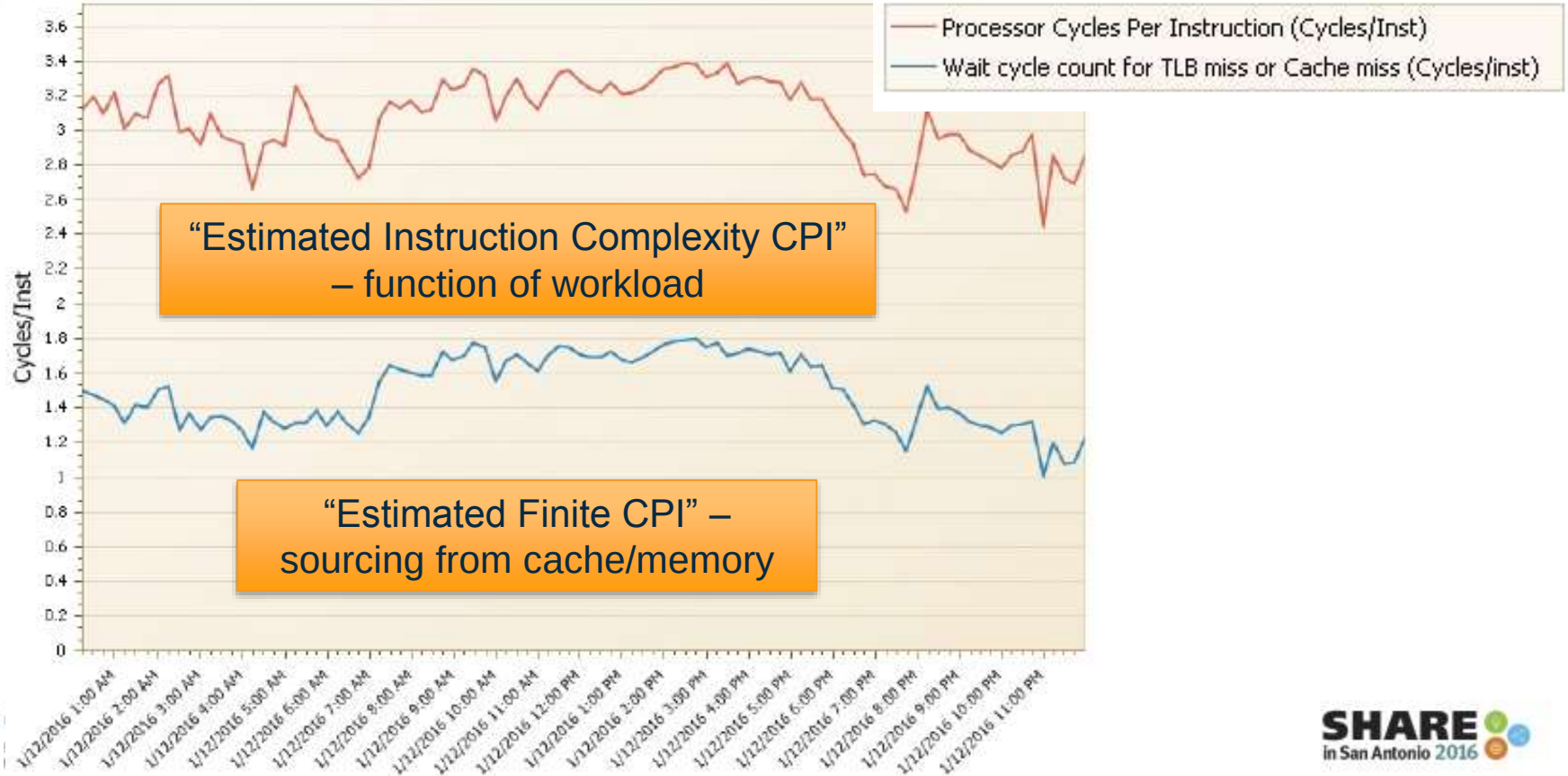
- IBM measured 6% capacity improvement

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

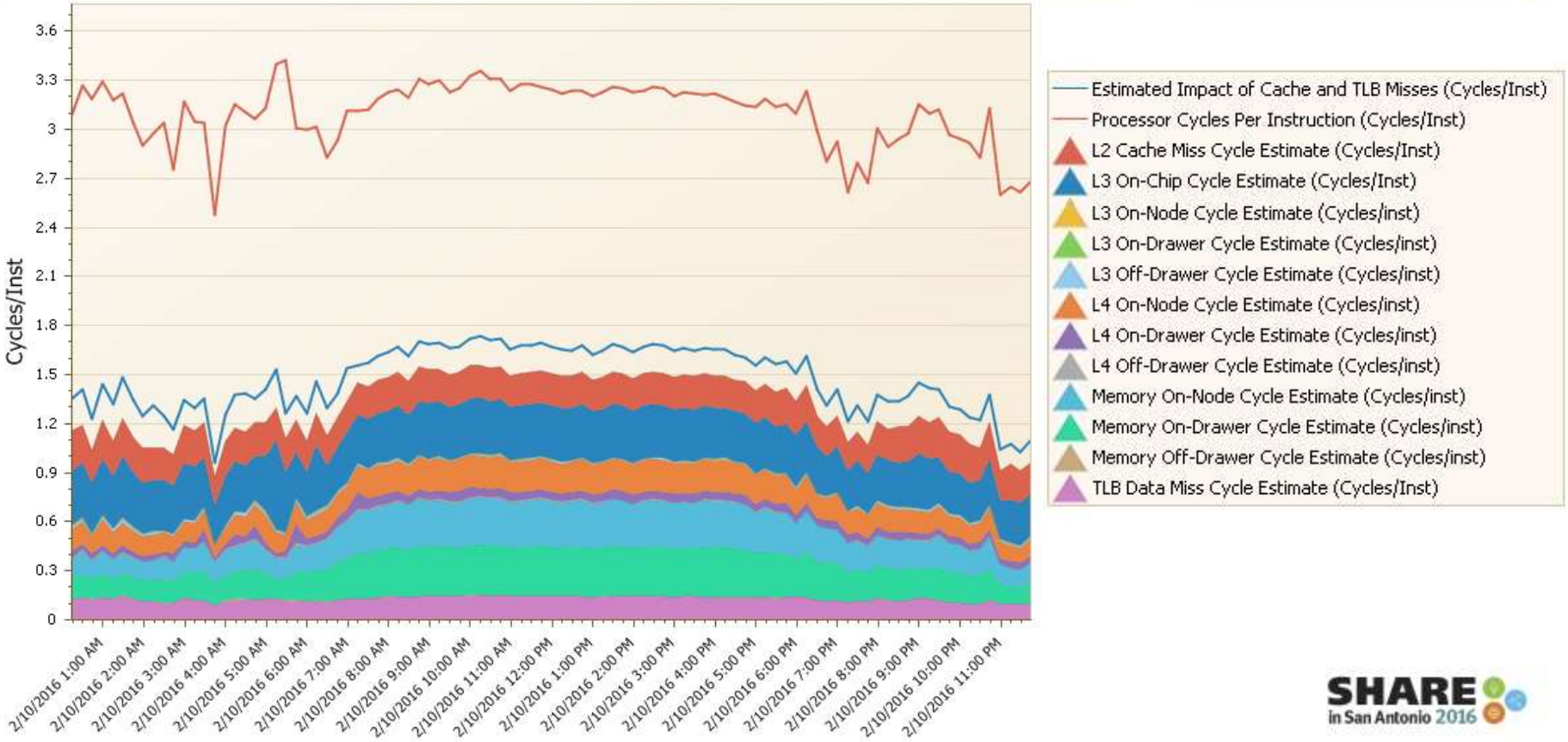
Key Metrics – Cycles per Instruction (CPI)

- Processor cycles are spent
  - Productively – executing instructions (complex instr. > 1 cycle)
  - Unproductively – waiting to stage data (L1 cache or TLB miss)
- "Waiting" does not always mean waiting
  - Enhanced Out Of Order (OOO) execution
  - Other pipeline enhancements

Cycles per Instruction & Cache Miss Impact



Estimated Impact Cache & TLB Misses



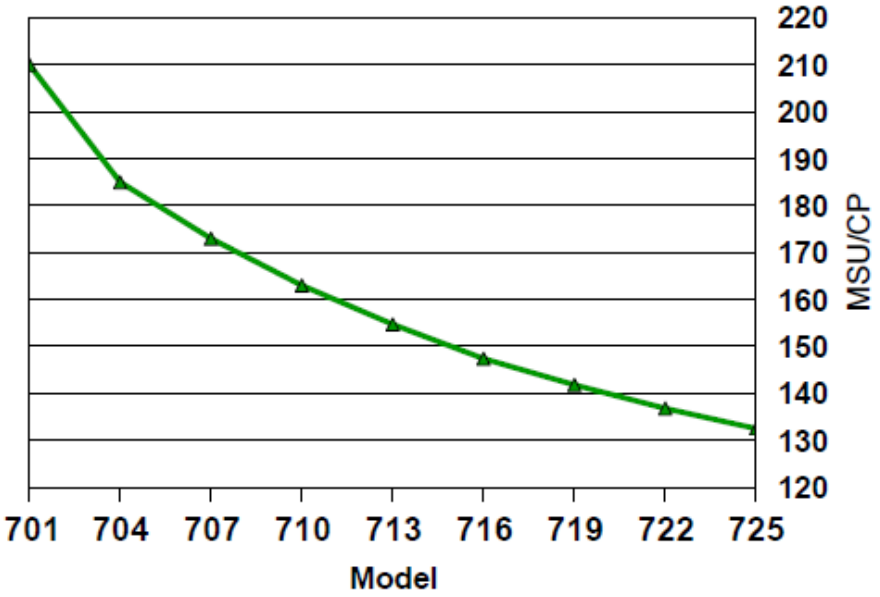
CPI Components and Impact

- Estimated Finite CPI – variable and correlates with RNI

	Prod Insurance			Prod Banking		
	711	716	726	711	716	726
Vertical High	3	5	6	0	0	4
Vertical Med	1	0	0	2	2	0
Vertical Low	2	0	0	1	1	0
RNI	1.56	1.26	1.24	1.35	0.87	0.71
CPI	4.00	3.52	3.46	4.89	3.92	3.48
Est Finite CPI	2.32	1.84	1.79	2.82	1.83	1.47
Est Instr Cmplx CPI	1.68	1.70	1.68	2.07	2.10	2.02

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



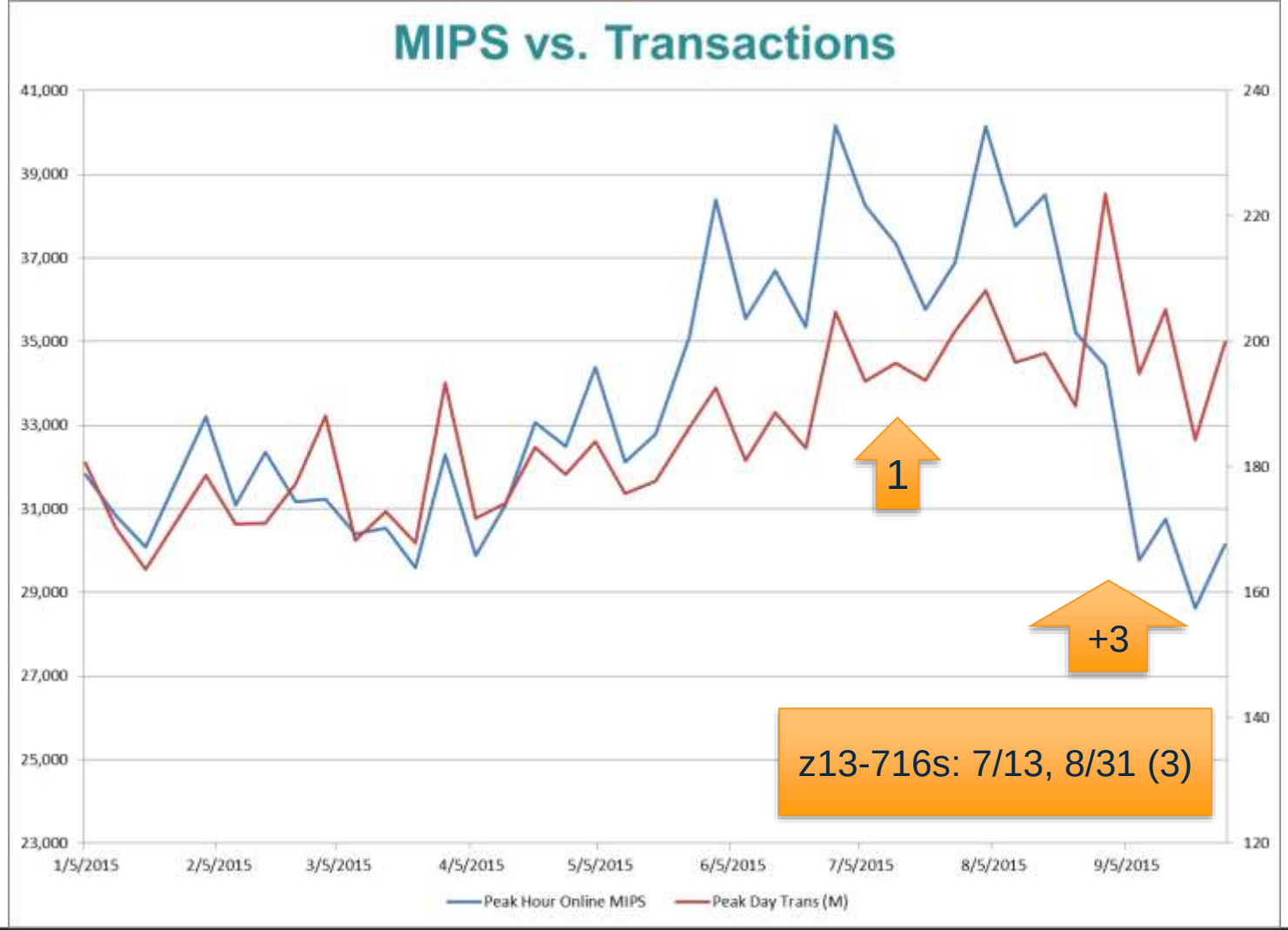
Savings from MSU/CP Ratings

- 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%

Two-fold Benefits of Deploying Capacity

- Consume less CPU due to operating efficiencies
  - Vertical High CPs
  - Improved LPAR Topology
  - Evidenced by RNI and CPI metrics
- CPU that is consumed translates into fewer MSUs / MIPS
  - Lower processor MSU/CP ratings

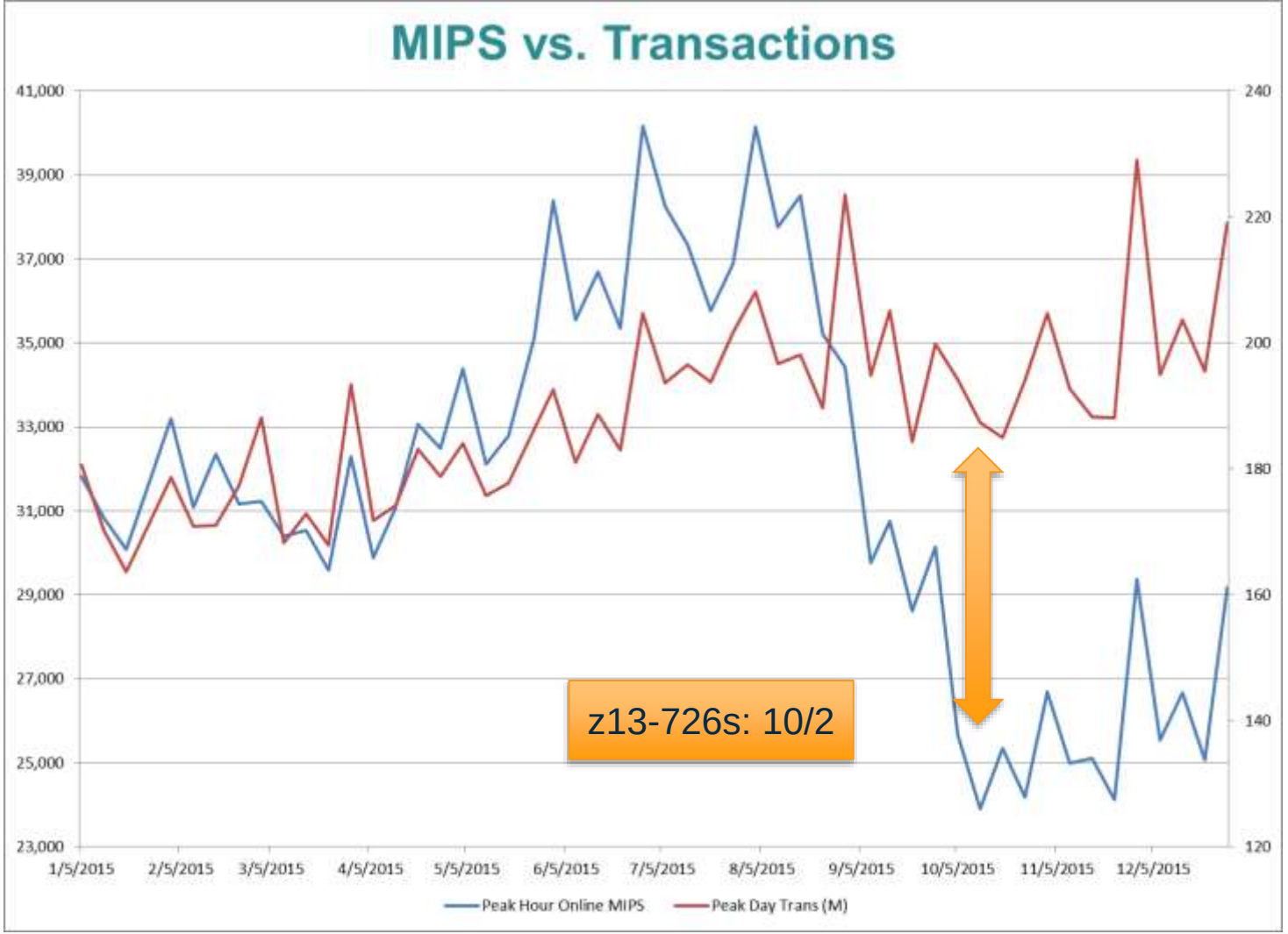


SHARE
EDUCATE • NETWORK • INFLUENCE

-5K MIPS vs. zEC12

-9K MIPS vs. 711

SHARE in San Antonio 2016



SHARE
EDUCATE · NETWORK · INFLUENCE

-9K MIPS vs. zEC12

-13K MIPS vs. 711

-4K MIPS vs. 716

SHARE in San Antonio 2016

Potential Financial Obstacles with ISVs

- Previously converted many ISV licenses from capacity-based to usage-based
- Approached remaining ISVs and explained value proposition for requiring a usage-based agreement
  - Most ISVs were very receptive
  - Two created their first ever usage-based agreements
  - All others ultimately agreed to usage-based terms
- Experience confirmed our previous perceptions about which ISVs were truly “partners”

Subcapacity Alternative

- If single engine speeds or other considerations do not require full capacity models, subcapacity models could be selected to add CPs without increasing hardware capacity
- Improve Vertical CP configurations and likely RNI
- From z13-710 subcapacity models could add 7-15 CPs

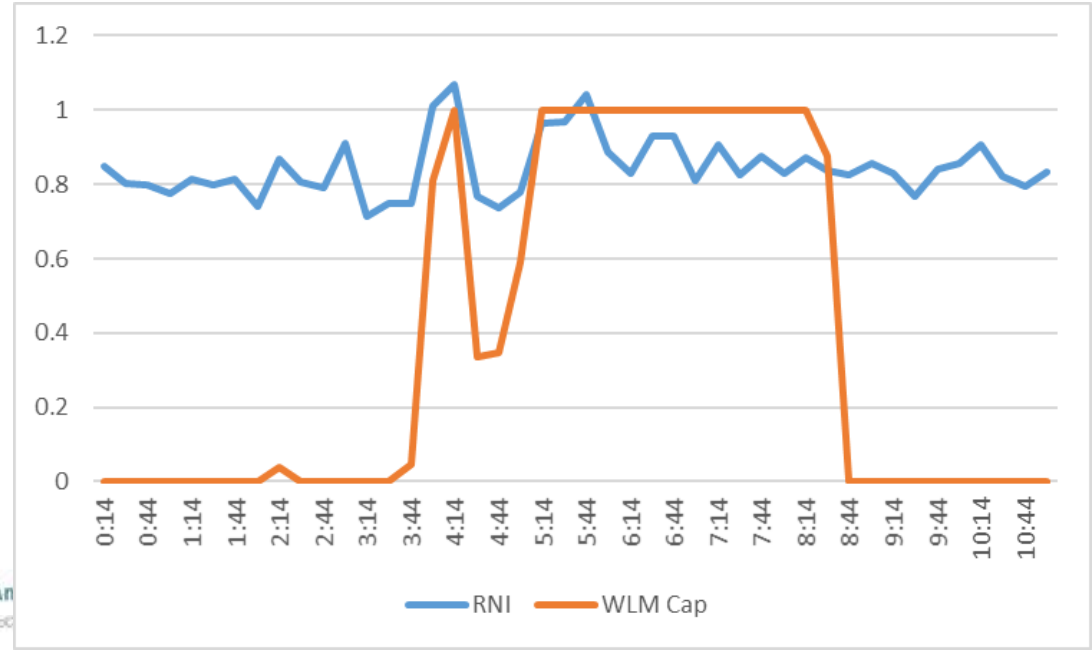
Total	zEC12-711	z13-710	z13-617	z13-525
MSUs	1593	1632	1610	1603

Perspectives on HW Capacity

- Review LPAR topology and system multiprogramming level for potential opportunities
- Current economics of software (largest expense, recurring) vs. hardware (smaller expense, one-time) may justify acquisition of additional hardware capacity
  - Similar framework as case IBM is seeking to make for memory
- Consider availability benefit provided by capacity cushion

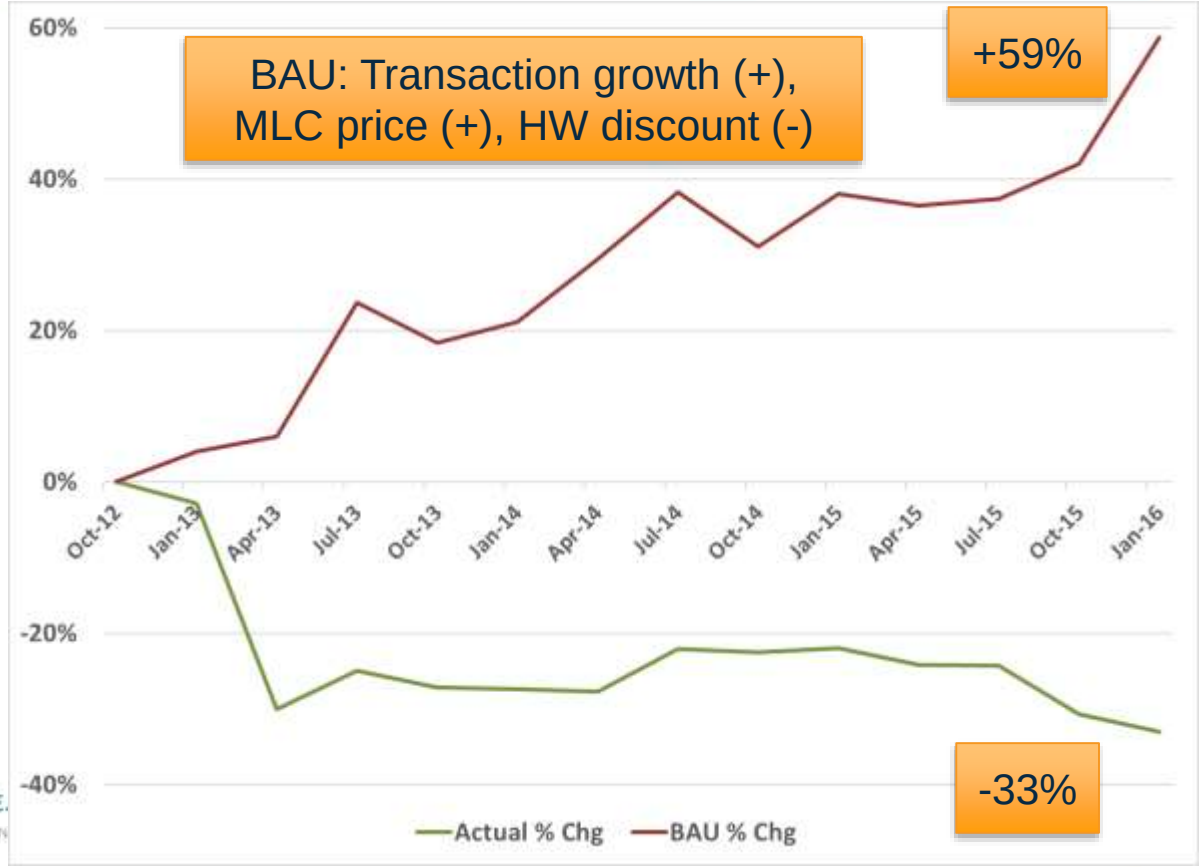
Capping Impact on Vertical CP Configs

- WLM (group or soft) capping is enforced by PR/SM changes to Vertical CP configuration
  - 6 VHs → 2 VMs, 4 VLs
- Concern about impact on delivered capacity
  - Observed small RNI increases (limited analysis)



MLC Expense Reduction Journey

- Sysplex aggregation
- Application tuning
- LPAR SW stack
- LPAR wkld mgmt
- Capping off-peak
- Batch mgmt
- Deploy HW capacity
- SHARE Boston 2013



Sources

- John Burg, “2015 CPU MF Counters Update”, Edge 2015 **
- Peter Enrico, “SMF 113 Processor Cache Counter Measurements”, Edge 2015
- Gilbert Houtekamer, “How to get more MIPS out of your z13”, white paper at www.intellimagic.com
- Gary King, “The Relatively New LSPR and IBM z13 Performance Brief” and “To MIPS or Not to MIPS”, SHARE March 2015 **
- Frank Kyne, “HiperDispatch Q&A” and “A Holistic Approach to Capacity Planning”, Cheryl Watson’s Tuning Letter 2015 No. 4

Related Sessions at this SHARE

- Wed. 11:15, Gary King, “To MIPS or Not to MIPS”, 304A
- Thur. 10:00, Gary King, “The Relatively New LSPR and the latest zSystems Processors”, 304A
- Thur. 11:15, John Burg, “2016 CPU MF Counters Update”, 304A
- Fri. 10:00, Cheryl Watson & Frank Kyne, “The Cheryl and Frank zRoadshow”, 303B

Special Thanks

- John Burg and Gary King, IBM
- Gilbert Houtekamer, IntelliMagic
- Frank Kyne, Watson & Walker

- Please complete your session evaluation

- Questions?

Complete your session evaluations online at SHARE.org/SanAntonio-Eval