IntelliMagic

Processor Reporting: From Capture Ratio to RNI

CMG Computer Measurement Group
imPACt
INTERNET • MOBILE • PERFORMANCE & CAPACITY • CLOUD • TECHNOLOGY
NOVEMBER 6-9, 2017 | LOEWS HOTEL | NEW ORLEANS, LA

Dr. Gilbert Houtekamer, IntelliMagic BV

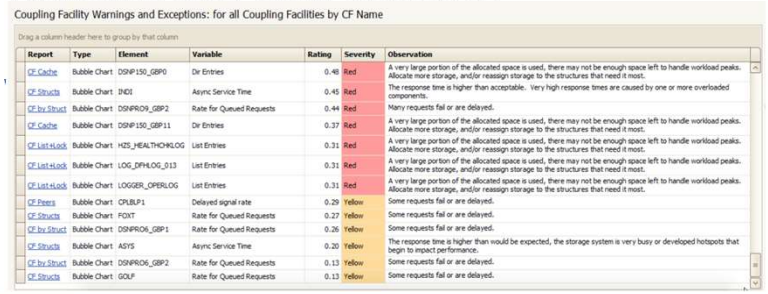
Availability Intelligence

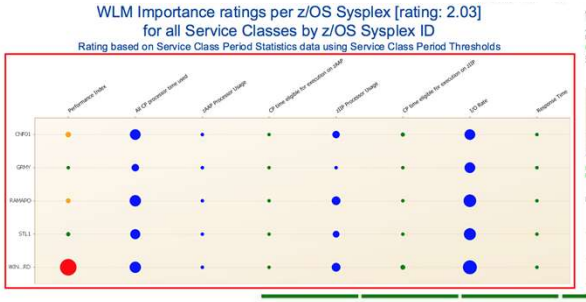
About IntelliMagic

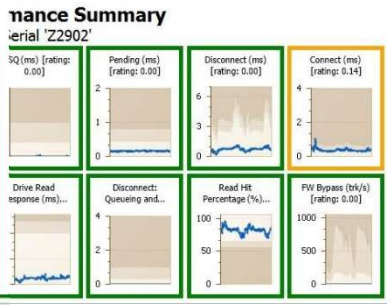
- Availability Intelligence for z/OS and SAN
  - z/OS infrastructure: processors, DASD, CF, XCF, TCP/IP, …
  - SAN, Fabric up to VMware
- On-premise and Cloud delivery
- Headquarters in Leiden, NL
  - German office in Munich
  - US office in Dallas, TX

Availability Intelligence for z/OS Infrastructure

- What risky conditions exist right now across our entire environment?
  (exception charts, rated metrics)
- What related metrics are relevant to the context of this issue?
  (side-by-side mini-charts of related metrics that are clickable)
- Where to go next to see root causes?
  (intelligent drill-downs)
-







Objective

- Convince you to rethink how you report on performance

Abstract

The processor remains the most expensive resource in the data center, both directly with the hardware costs and the indirectly with the MSU based software changes.

All installations use some form of processor reporting, but in many cases this reporting goes back many, many years. Capture Ratios are still interesting, but there is a lot more now.

Today processor performance relies more and more on efficient use of processor cache and it's also important to look at new metrics like the SMF 113 Hardware Counters and the SMF 99.14 topology data.

In this presentation, we will cover processor reporting from Capture Ratio to RNI (Relative Nest Intensity), and show you how understanding these metrics can help you tune your system.

Note on Charts

- The presentation contains charts from the IntelliMagic Vision product
- However, all charts are based on RMF/SMF data and could be recreated through other methods

Agenda

- Utilization & Workload Analysis
- Capping and Capacity Groups
- Capture Ratio & LPAR Mgmt Overhead
- zIIP Processing
- Processor Cache Concepts & Metrics

Utilization & Workload Analysis

CPU Consumption – Hardware Perspective

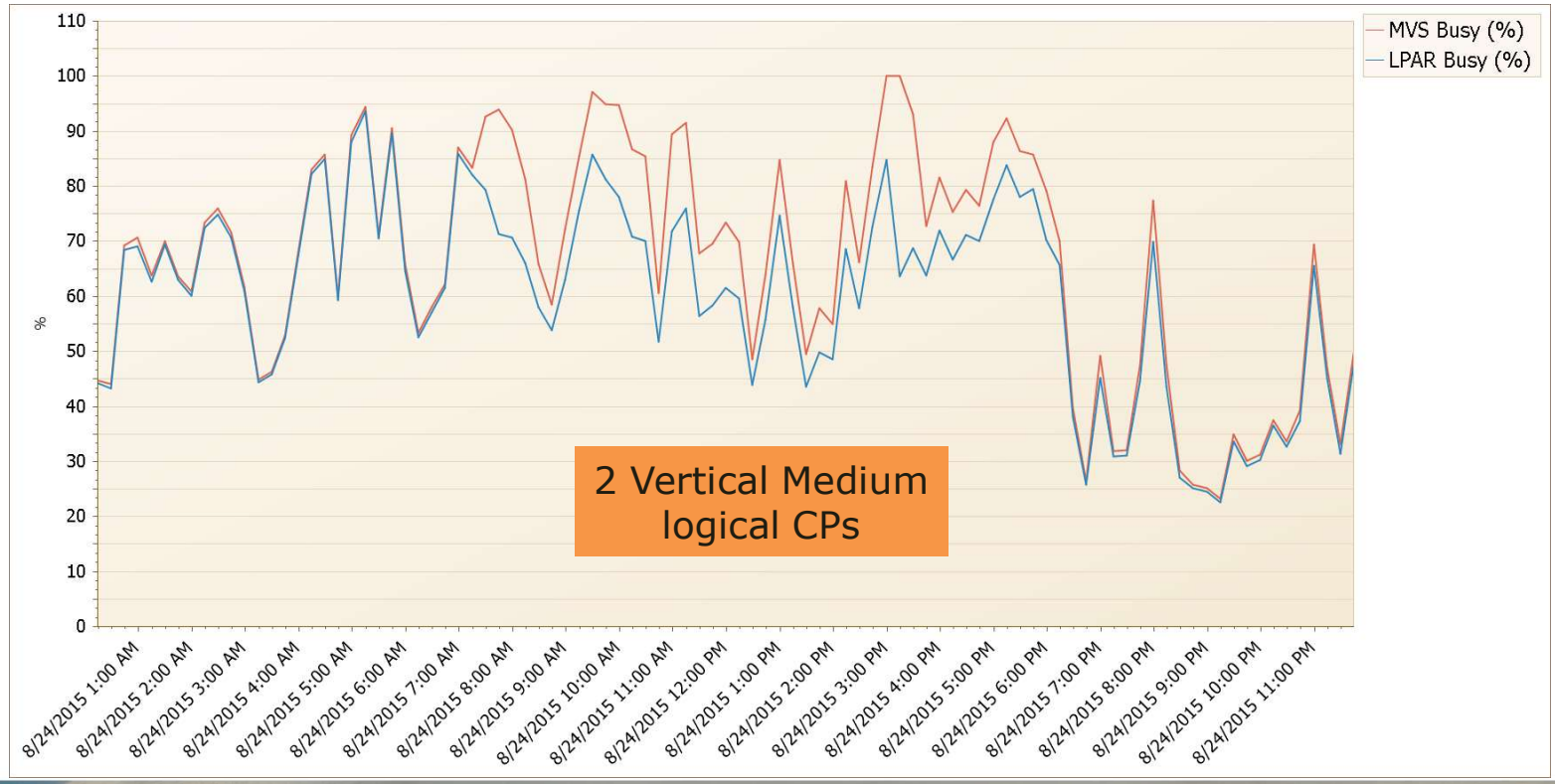
- By Processor (CEC)
  - Basis for MLC reporting
  - CEC typically shared by Sysplexes
- By System (LPAR)
  - Size of system can be managed by PR/SM or Capping
- By Sysplex (LPAR group)
  - Typically across CECs

Processor Utilization

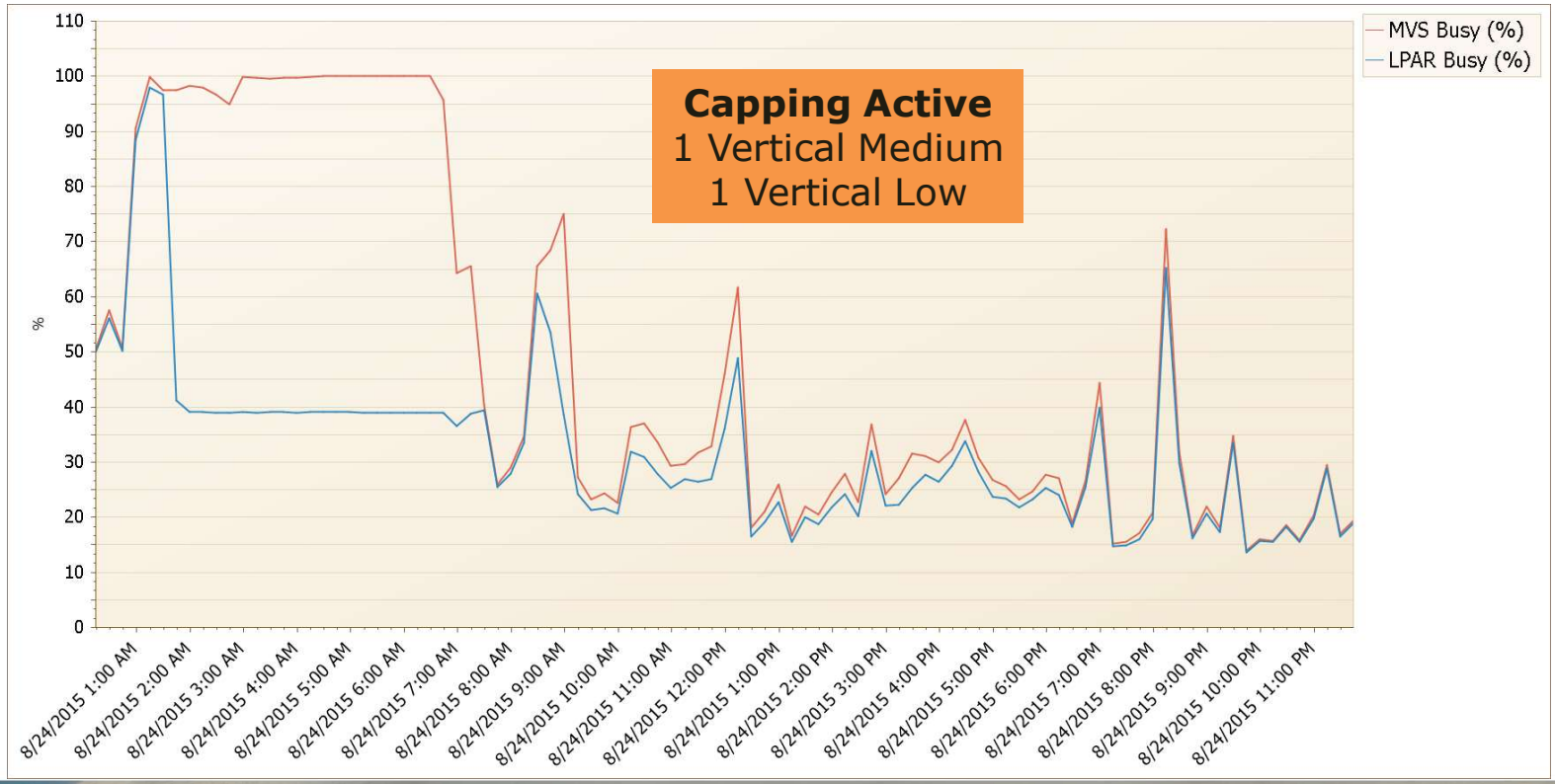
MVS Busy & LPAR Busy

- MVS % Busy – z/OS has dispatched work on a logical CP eligible to be executed
- LPAR % Busy – PR/SM has dispatched work on a physical CP so that it can be executed
- MVS Busy > LPAR Busy when workload exceeds LPAR weight & surplus CPU unavailable from other LPARs, e.g. because of (soft)capping

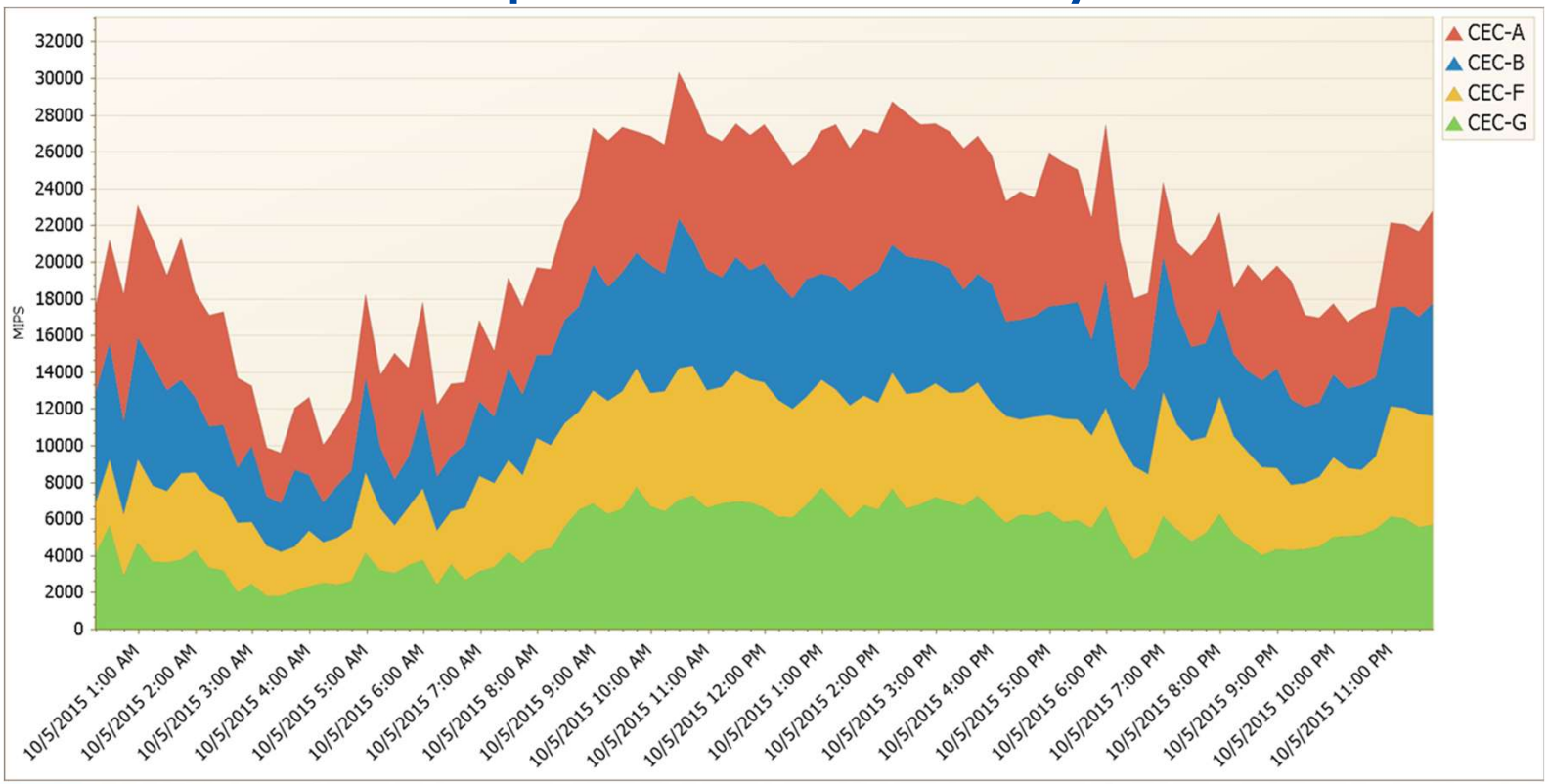
MVS & LPAR Busy (%) – 1



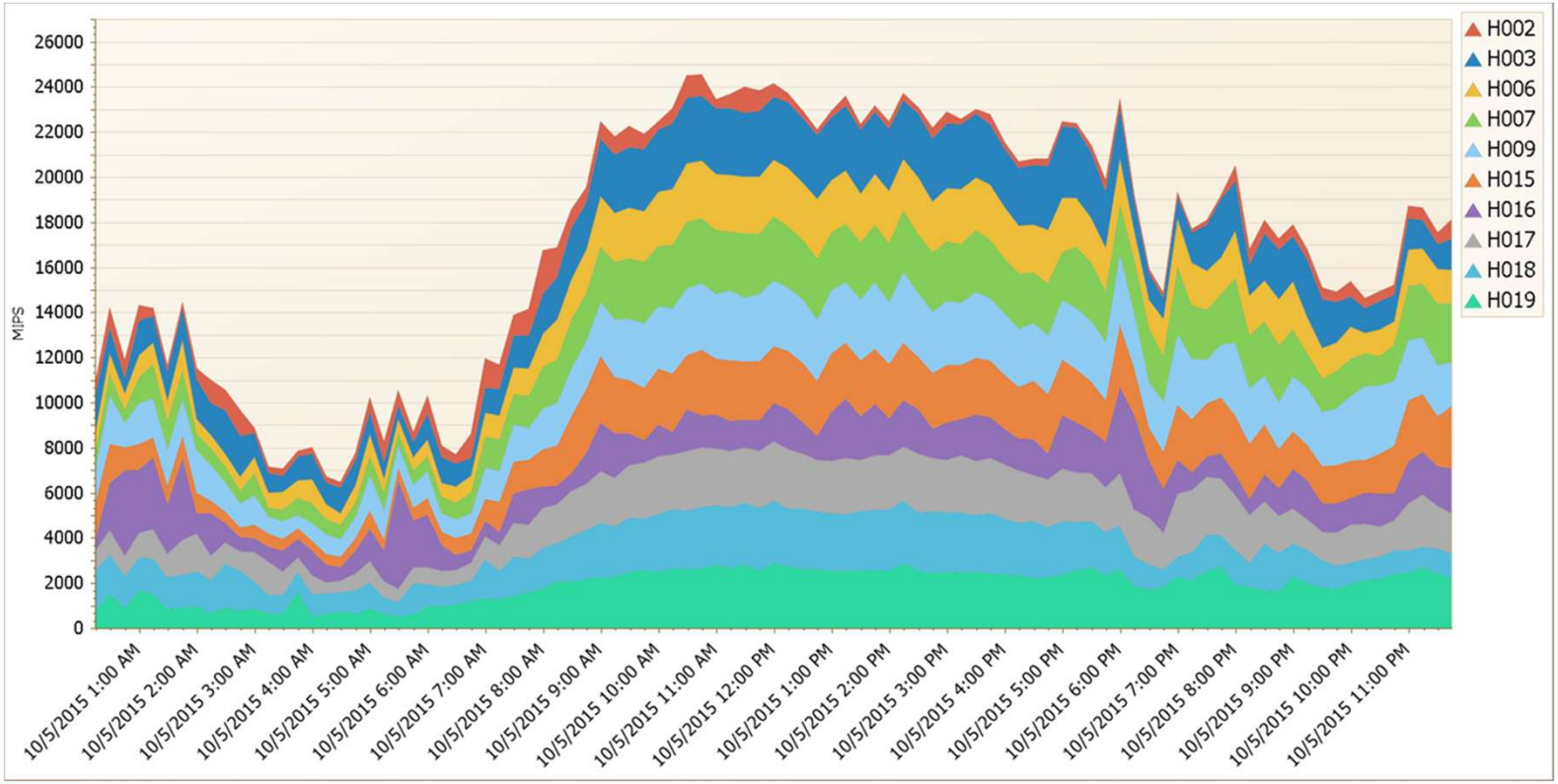
MVS & LPAR Busy (%) – 2



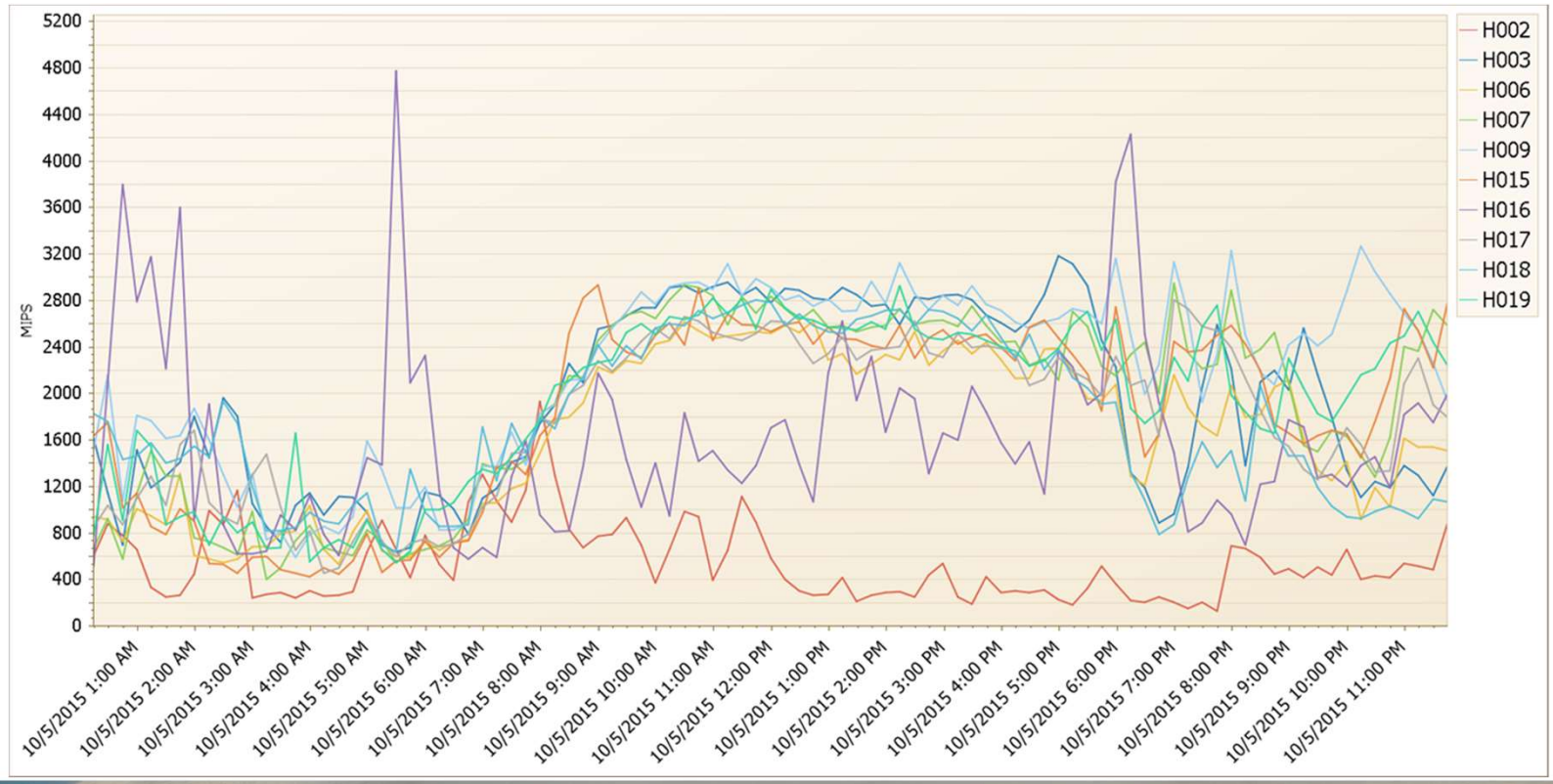
CP Dispatch Time – By CEC



CP Dispatch Time – By System



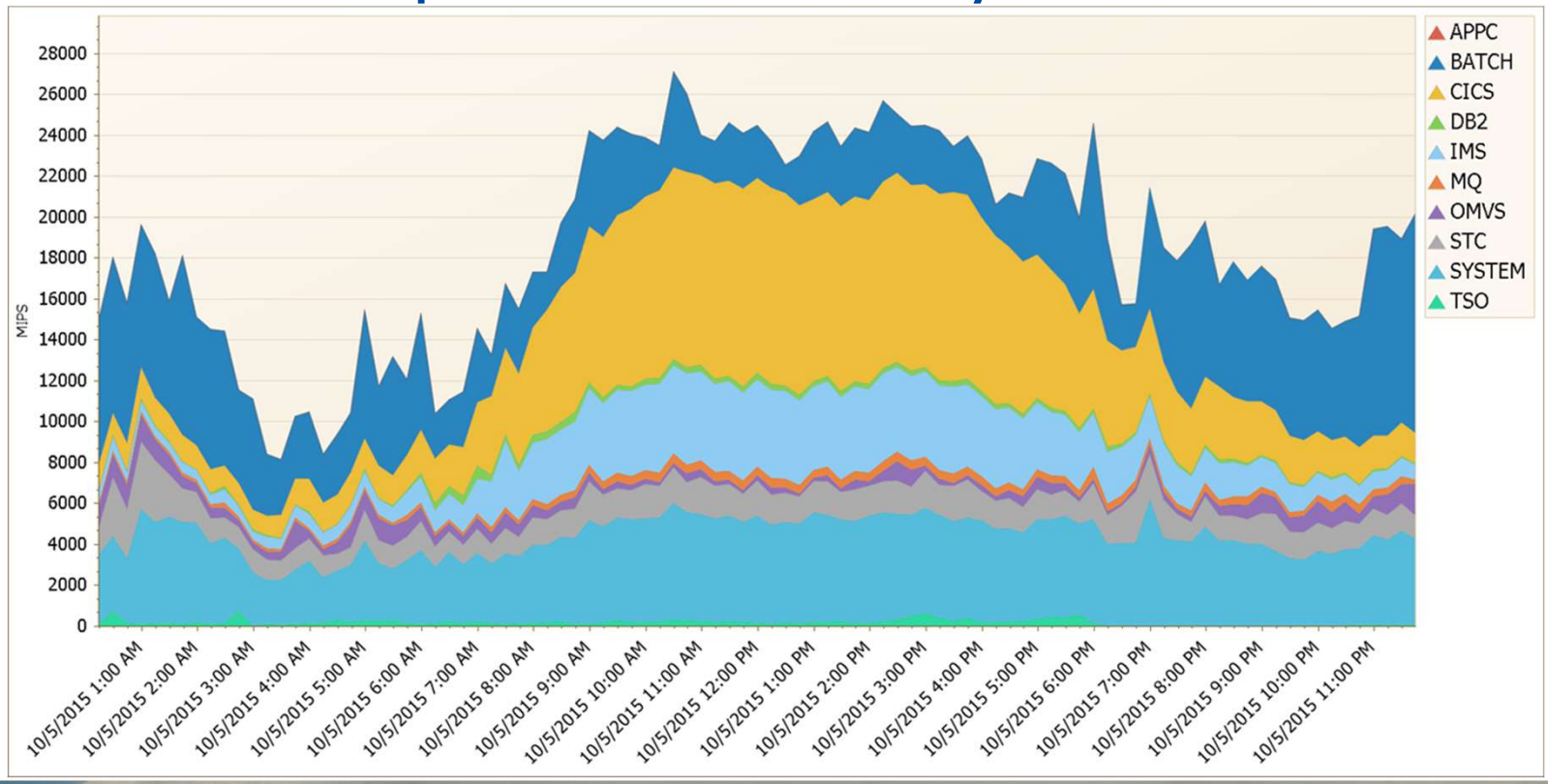
CP Dispatch Time – By System



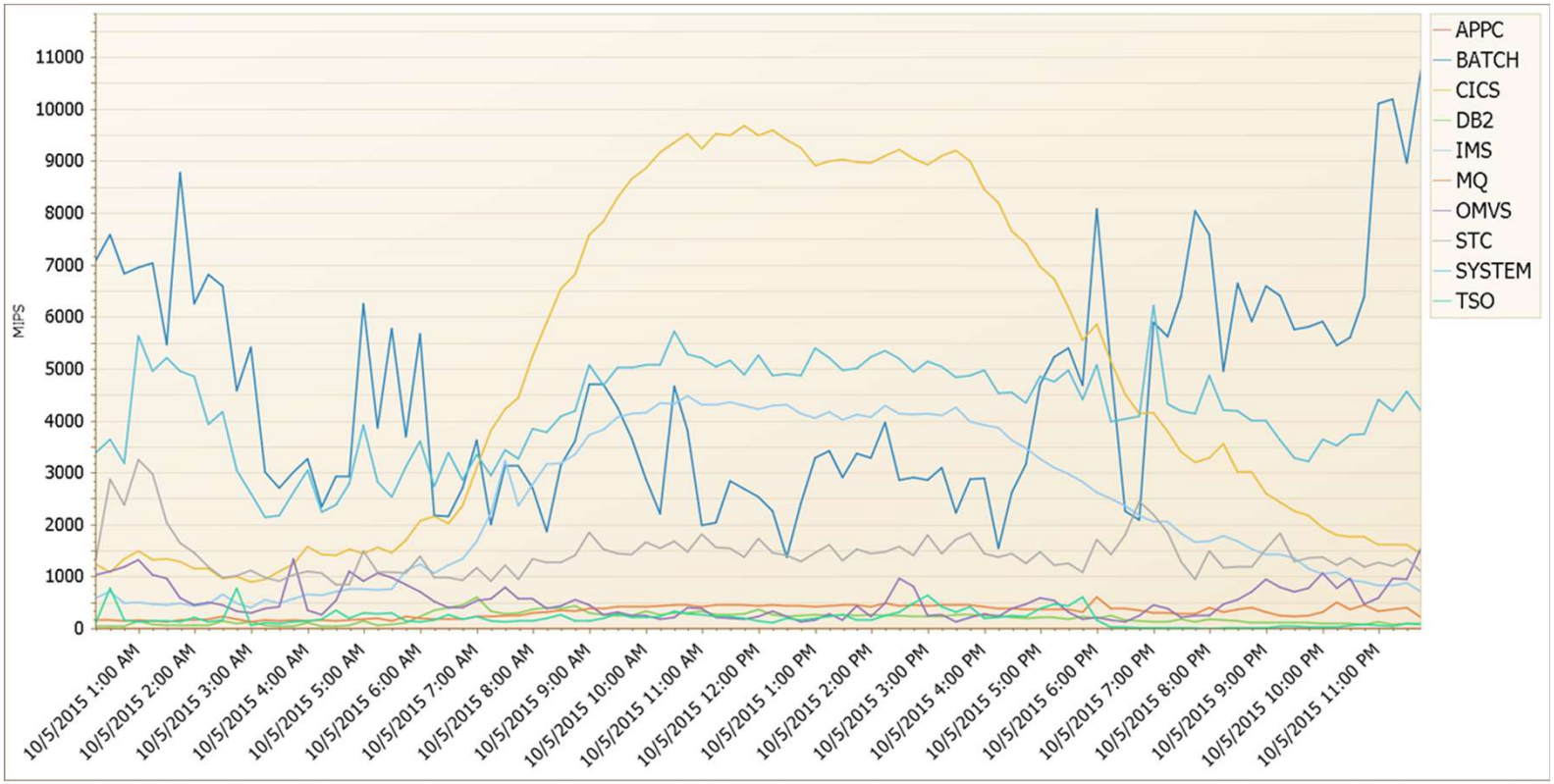
CPU Consumption – z/OS & WLM Perspective

- By WLM Workload
  - Good across LPARs and CECs, covers all work
- By WLM Service Class
  - Good across LPARs and CECs, covers all work
- By WLM Report Class
  - Coverage depends on your definitions
  - Now also for CICS, IMS and DB/2 transactions
    - z/OS 2.1 with OA47042; CICS 5.3; IMS 14 with PI46933 and PI51948
- By Address Space
  - SMF 30 Job records as used for chargeback/showback

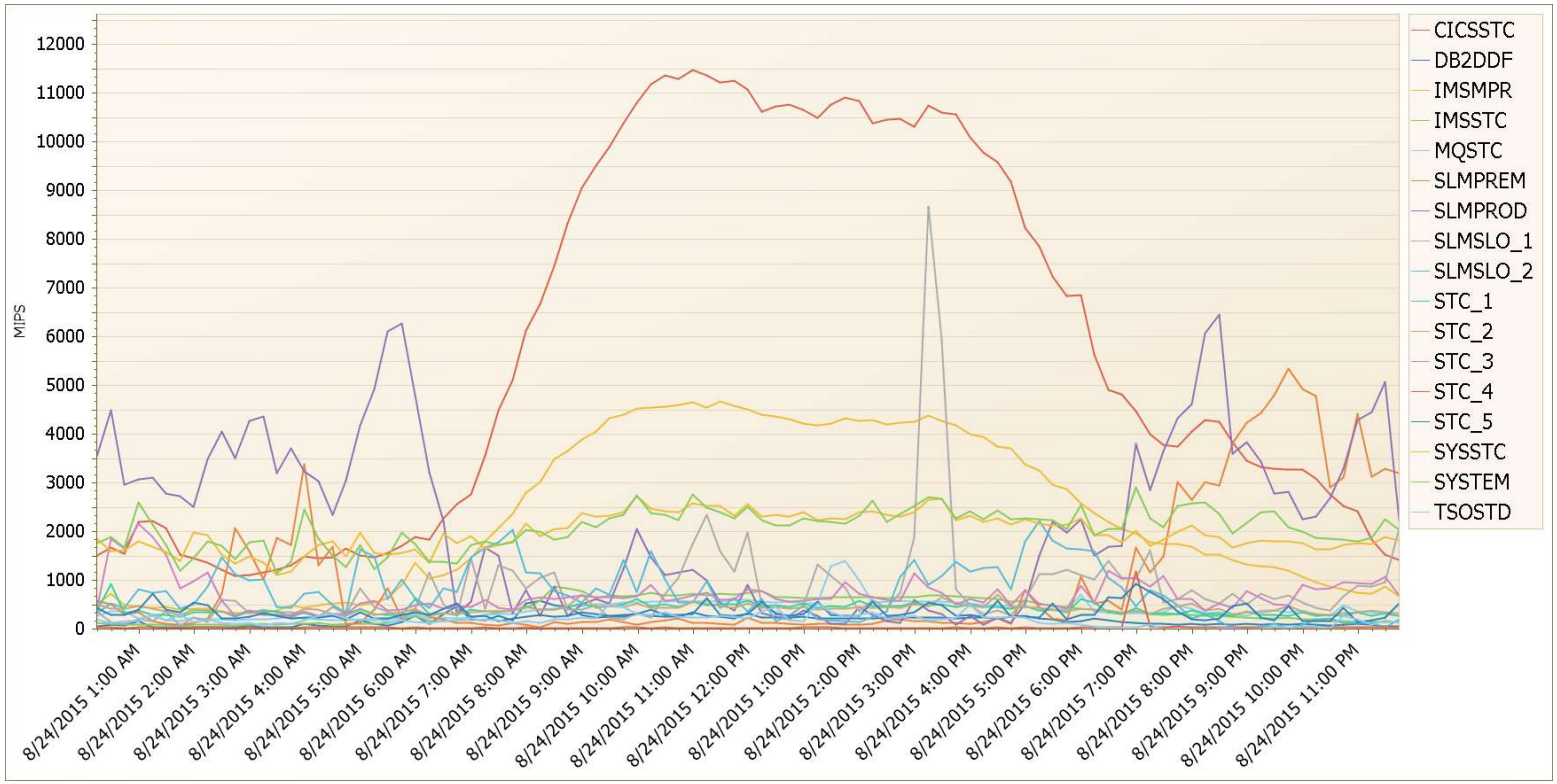
CP Dispatch Time – By Workload



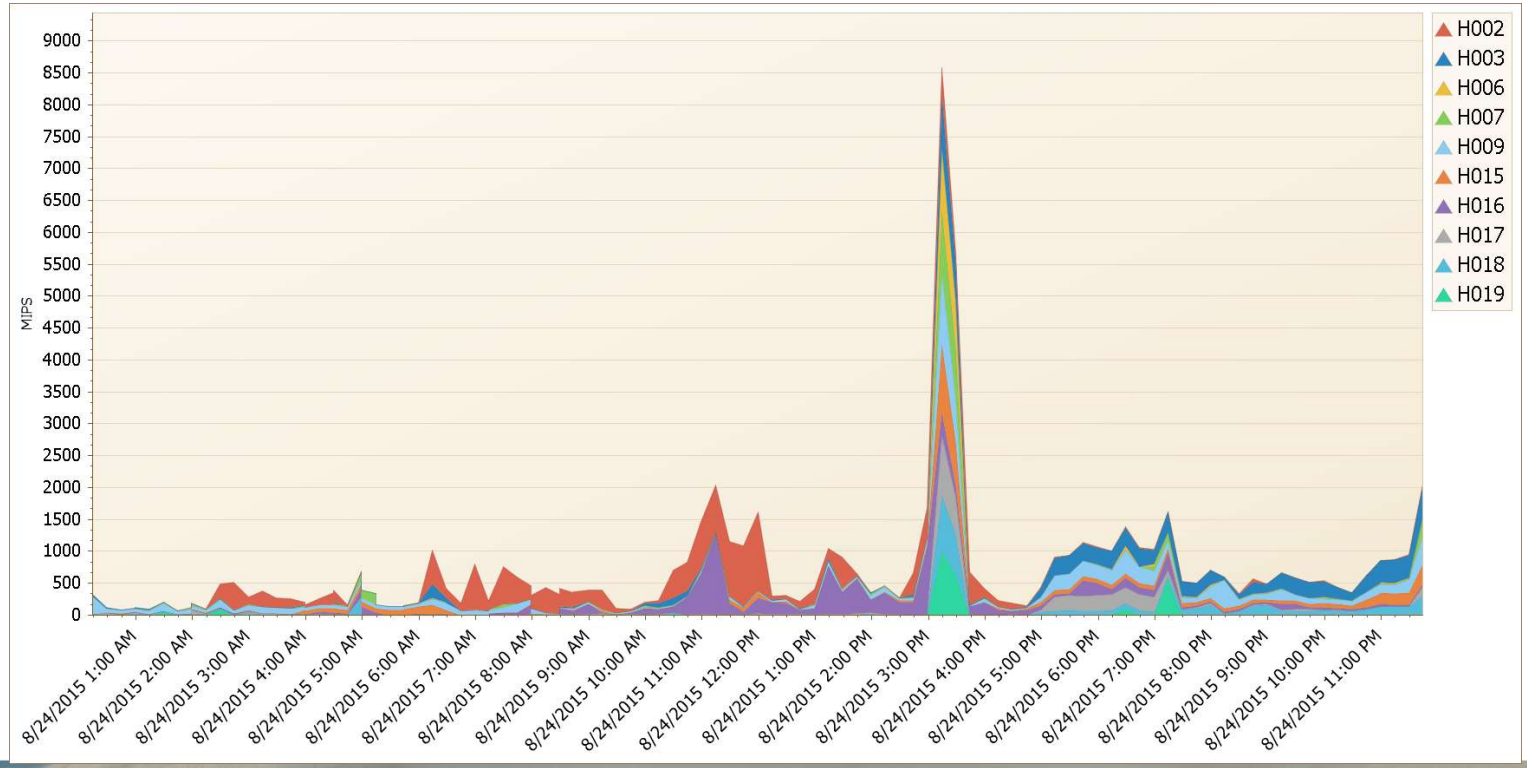
CP Dispatch Time – By Workload



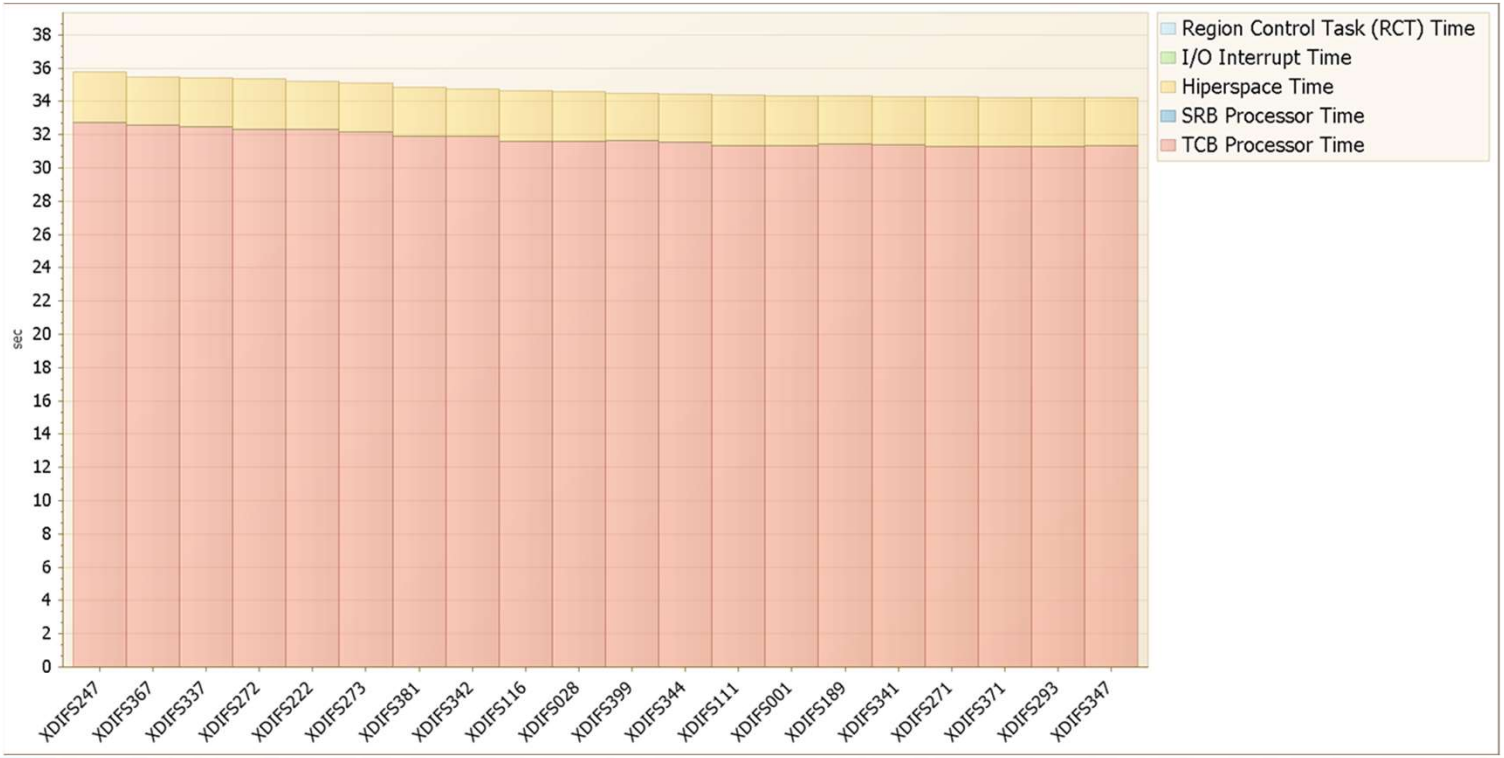
CP Dispatch Time – By Service Class



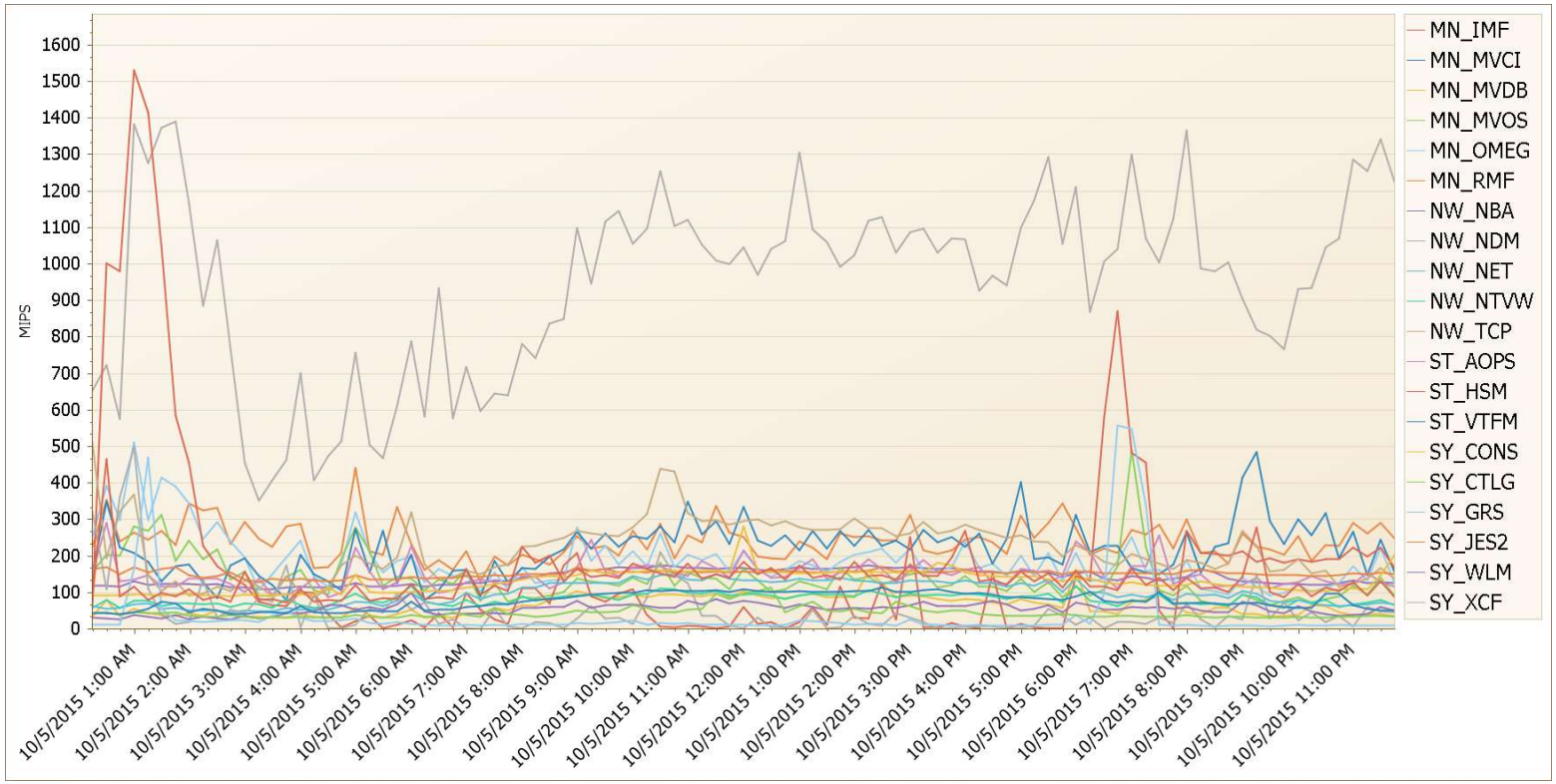
CP Dispatch Time –
For Service Class SLMSLO_1 by System



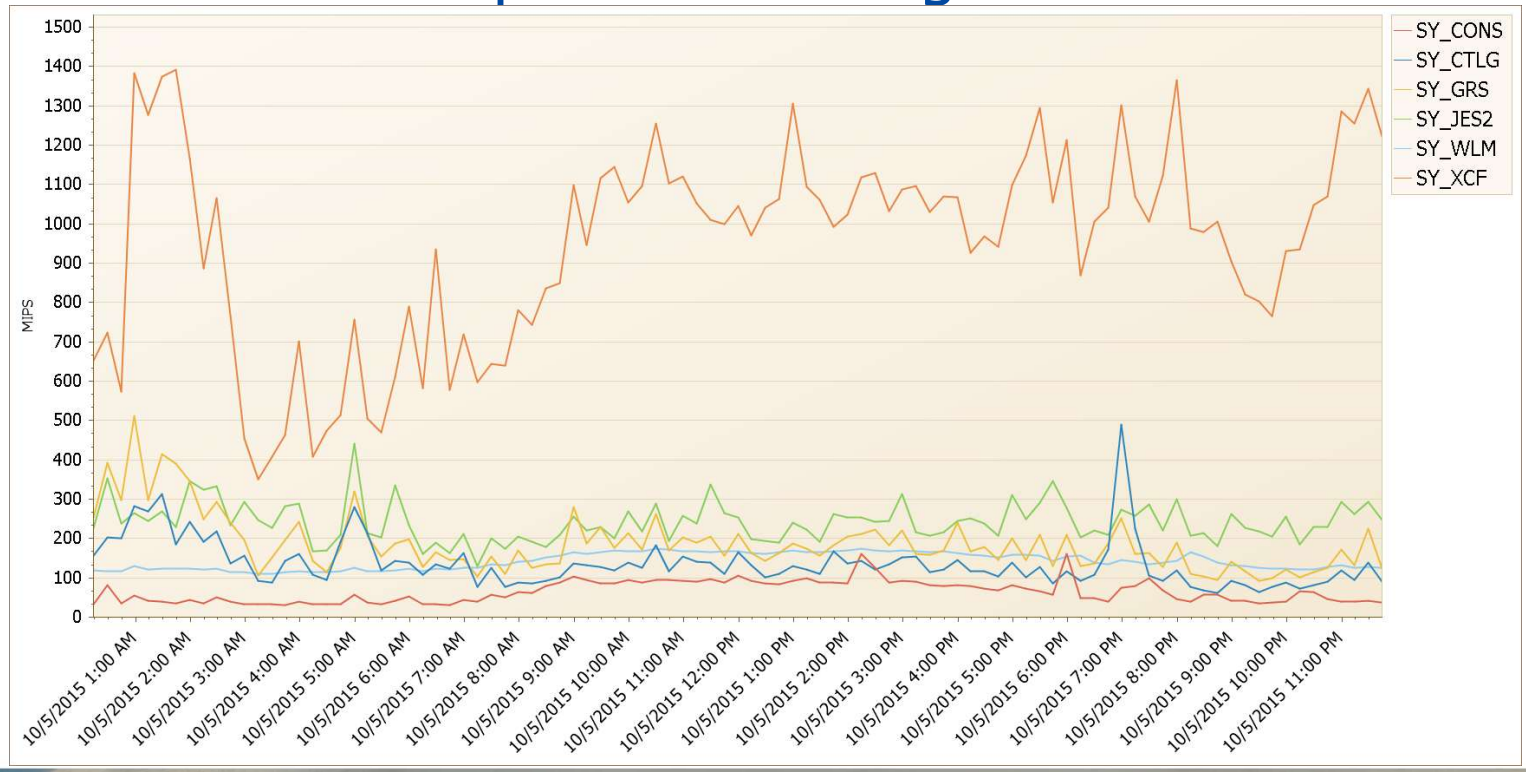
CP Dispatch Time (Top 20) – For Service Class SLMSLO_1 for System H007 by Address Space Name



CP Dispatch Time – By Report Class



CP Dispatch Time – By Report Class Where Report Class Begins with 'SY'



Capping and Capacity Groups

Capping and Capacity Groups

- Capping
  - Reduce resources to reduce cost (and performance)
  - Many flavors available, from IBM and 3rd parties
  - Objective is to limit resource usage

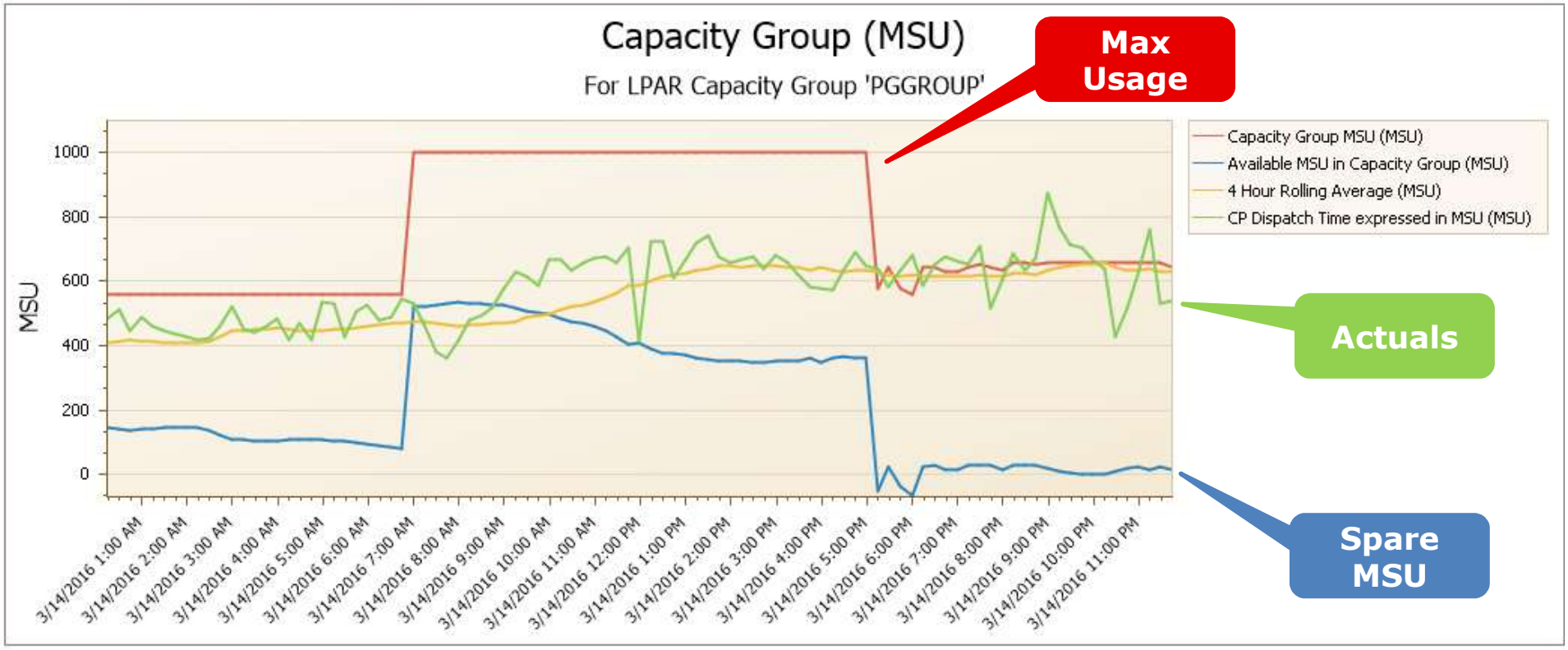
- Capacity Groups
  - Reduce resources across group of LPARs on CEC

GH2

SCRT and Usage Data ▸ Capping and Cap Groups ▸ Cap Grp MSU ▸ Capacity Group

GH1

Drill down to: ○ Systems ◉ Importance ○ Identify



Max Usage

Actuals

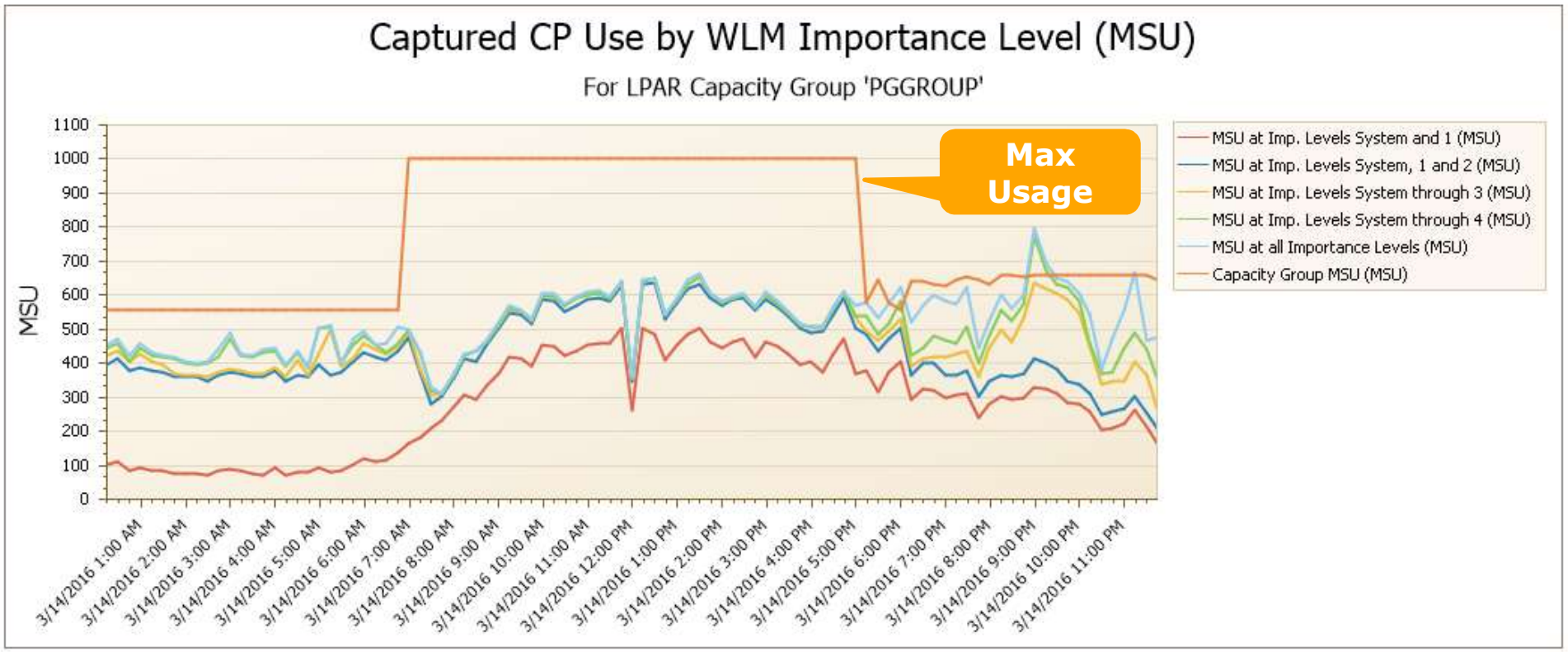
Spare MSU

Slide 27

GH1 Gilbert Houtekamer, 4/18/2017

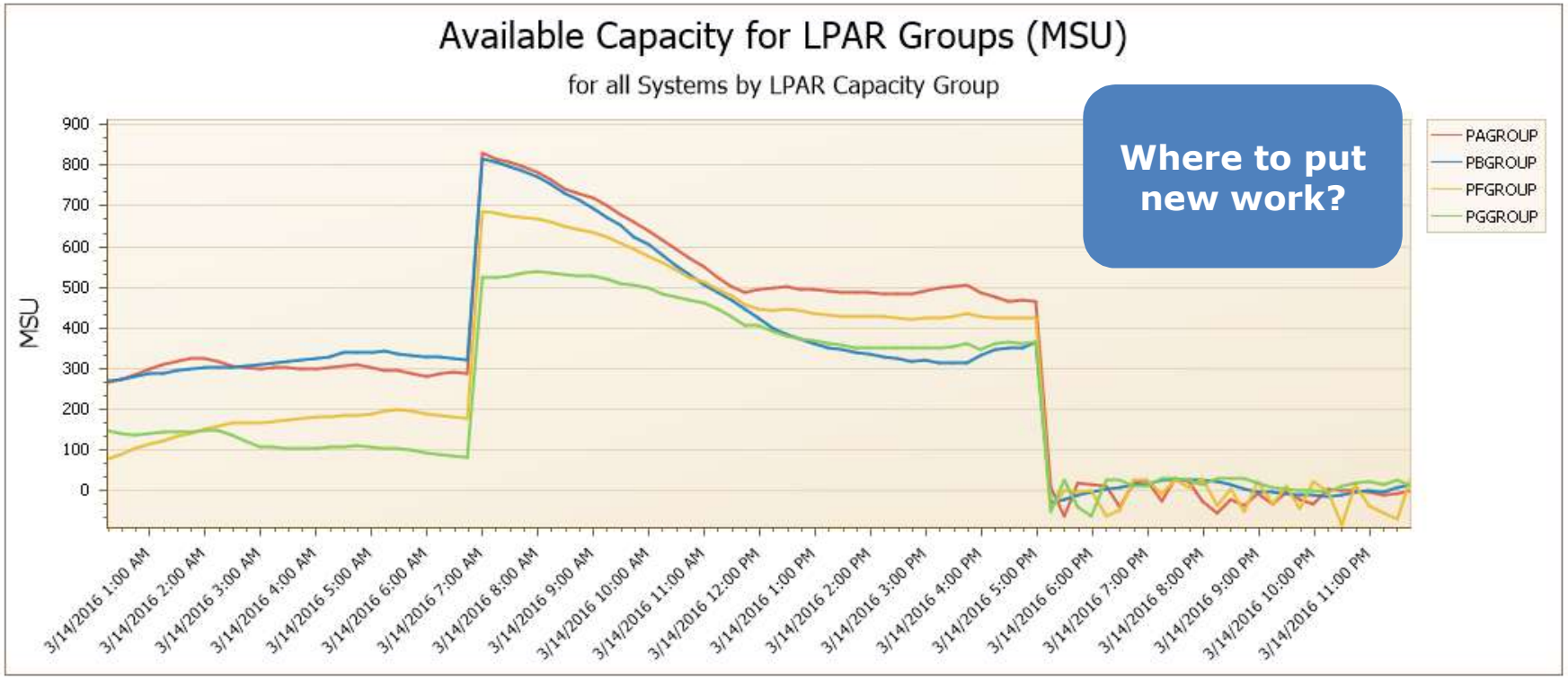
GH2 Gilbert Houtekamer, 4/18/2017

Drill down to: ◉ Identify



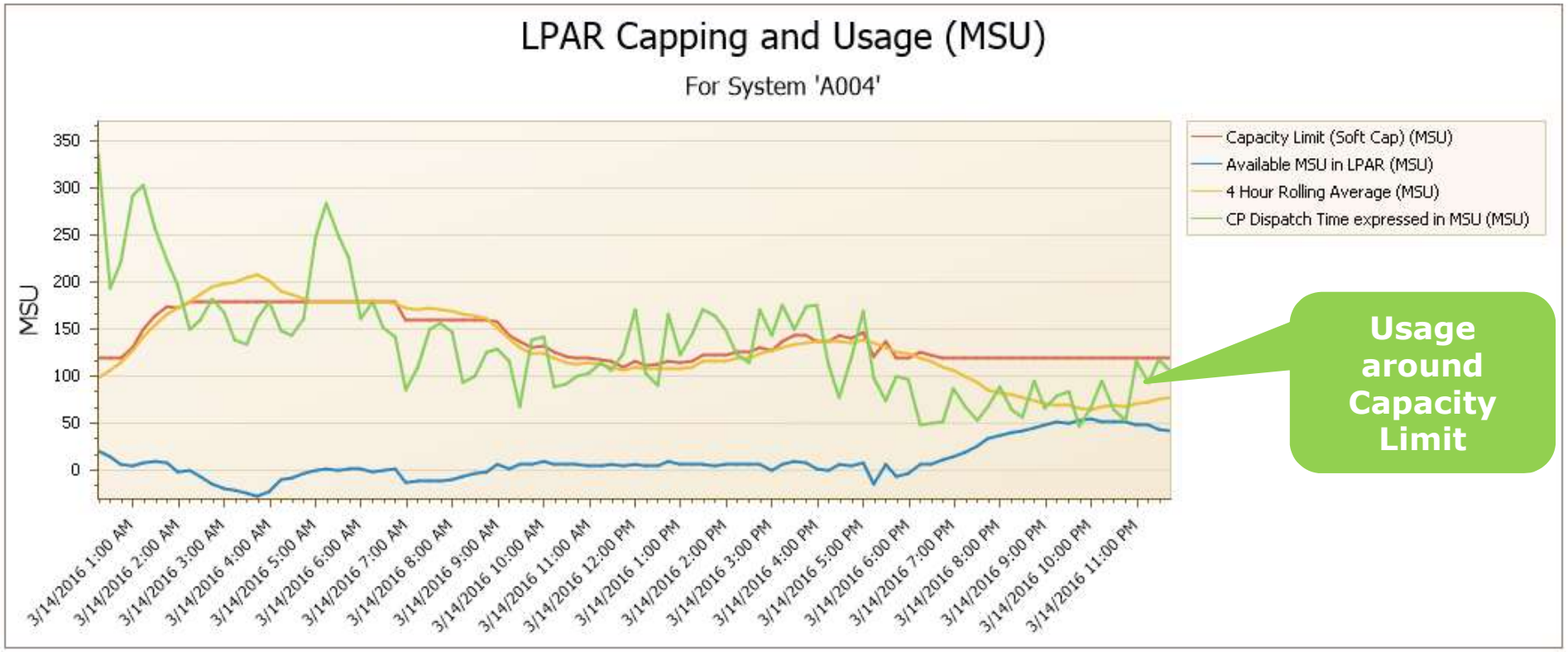
Processor Reporting ▸ Capping and Cap Groups ▸ Avail Grp MSU

Drill down to: ◉ Capacity Group ○ Soft Cap ○ Identify ○ Month ○ By date ○ Average by day of week ○ Isolate



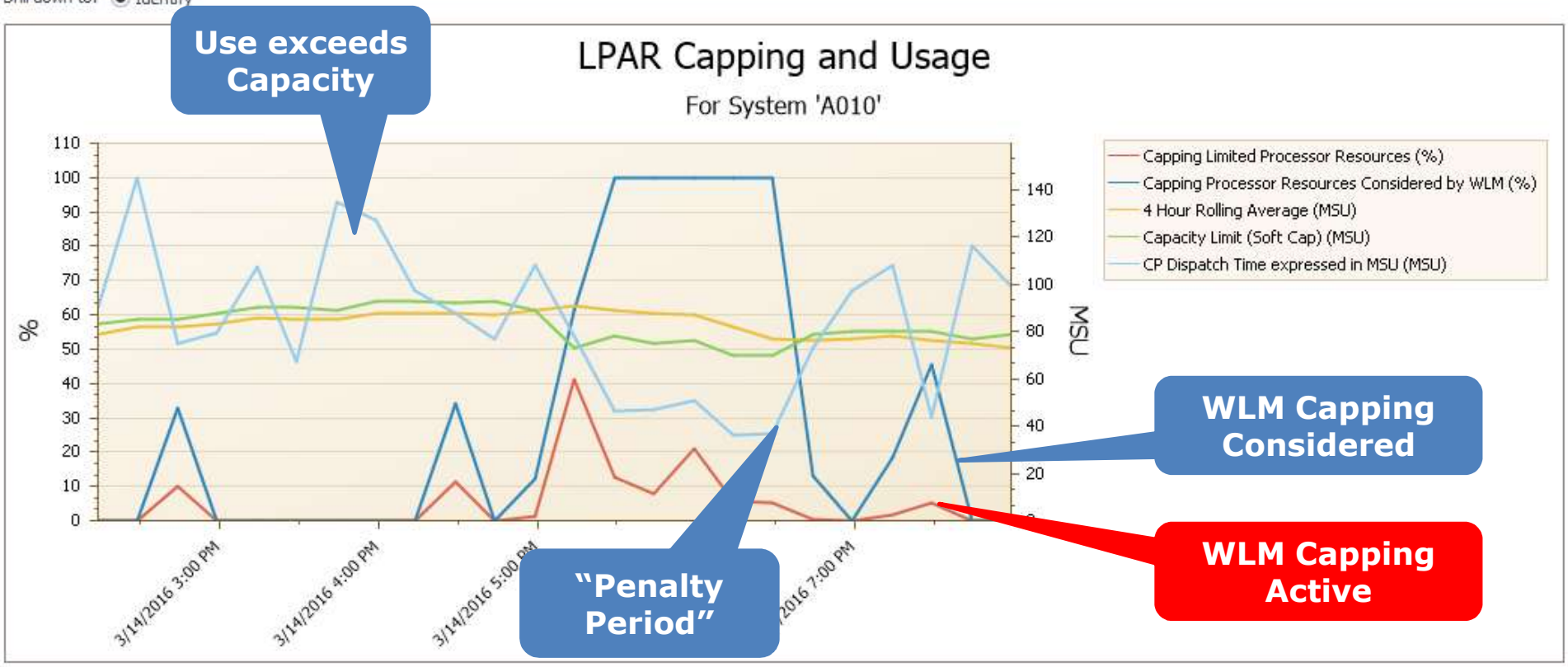
Where to put new work?

Drill down to: (•) Systems ( ) Importance ( ) Identify



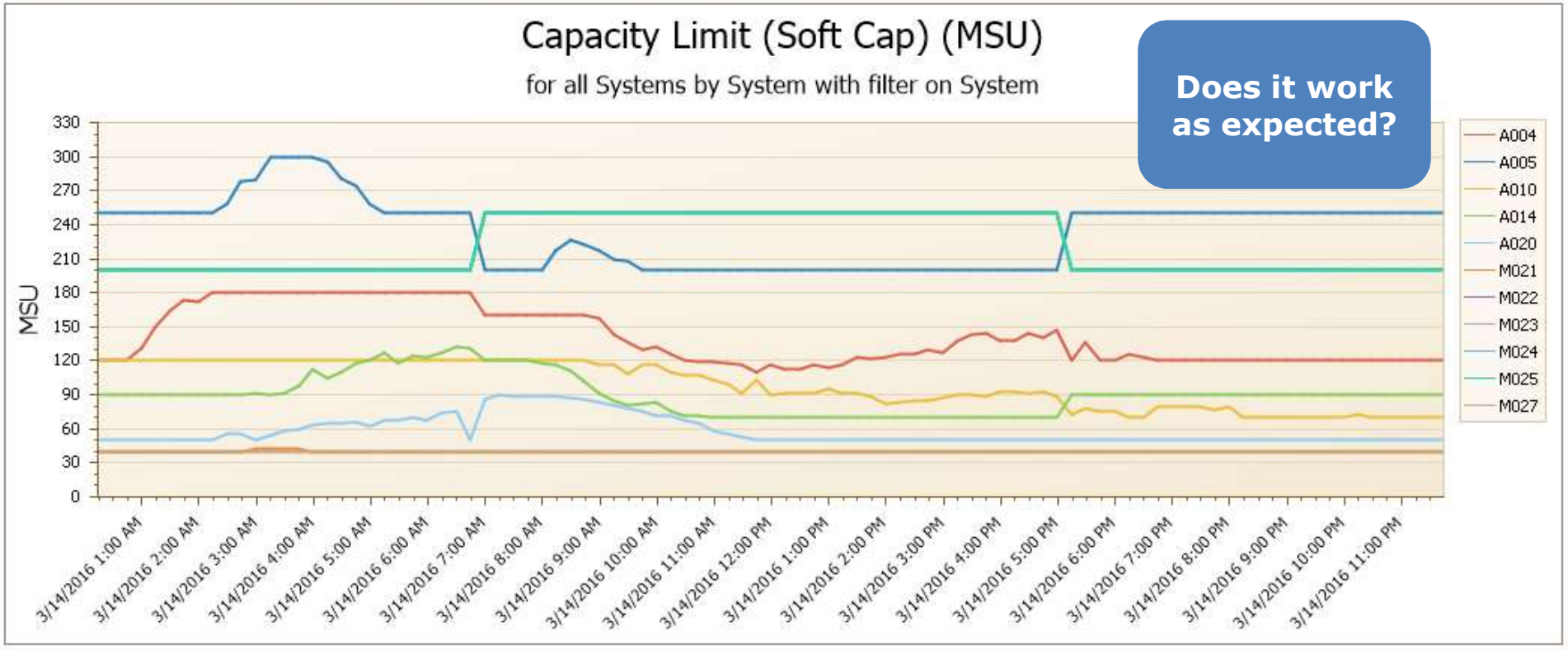
SCRT and Usage Data ▸ Capping and Cap Groups ▸ LPARs ▸ Capping

Drill down to: Identify



SCRT and Usage Data ▸ Capping and Cap Groups ▸ Soft Cap

Drill down to: (•) Capping ( ) Identify ( ) Month ( ) By date ( ) Average by day of week ( ) Isolate



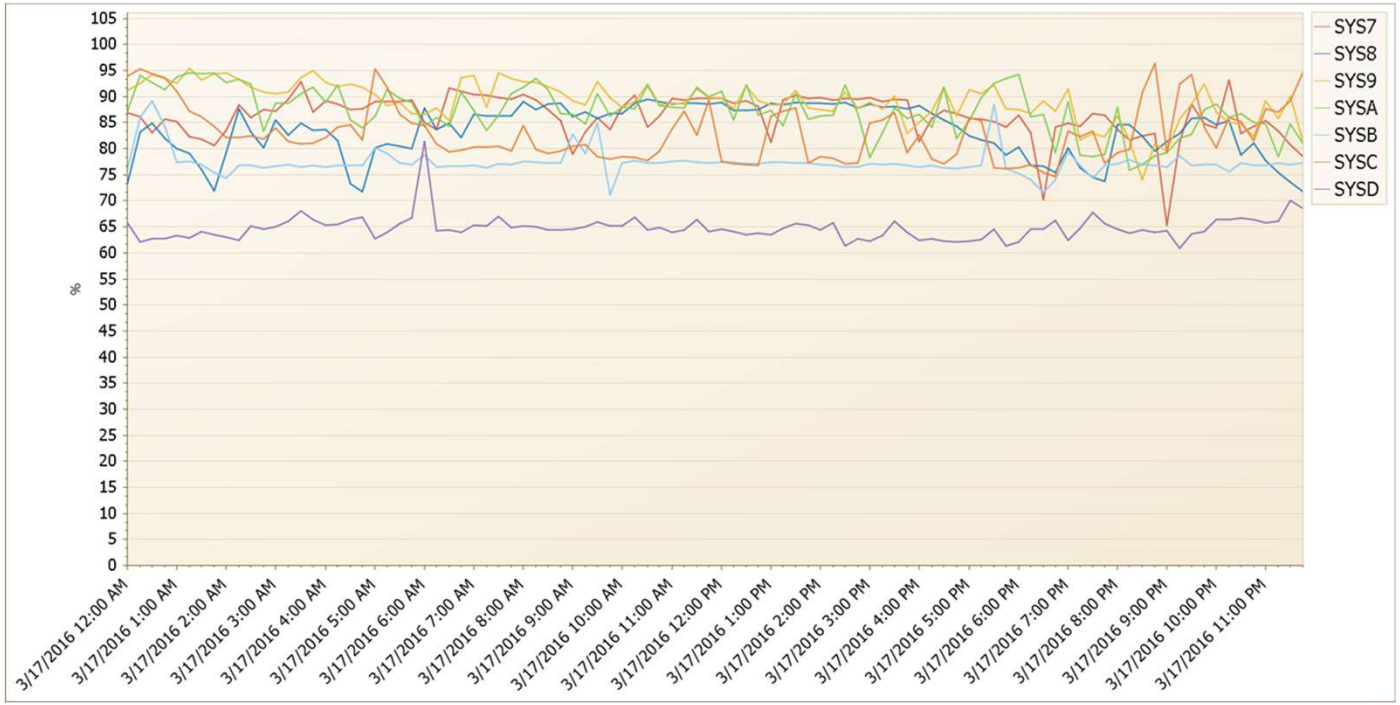
Does it work as expected?

Capture Ratio & LPAR Mgmt Overhead

Capture Ratio

- Processor time that can be attributed to workloads
- Calculation –total service class time from Type 72s (**i.e. useful work**) divided by total processor time from Type 70s (**i.e. what you pay IBM**)
- Factors that can increase uncaptured time
  - Storage-constrained system, Long dispatching queues
  - SLIP PER trace (we saw a case from 90% to 60%)
  - Small LPARs, LPAR management overhead

CP Capture Ratios



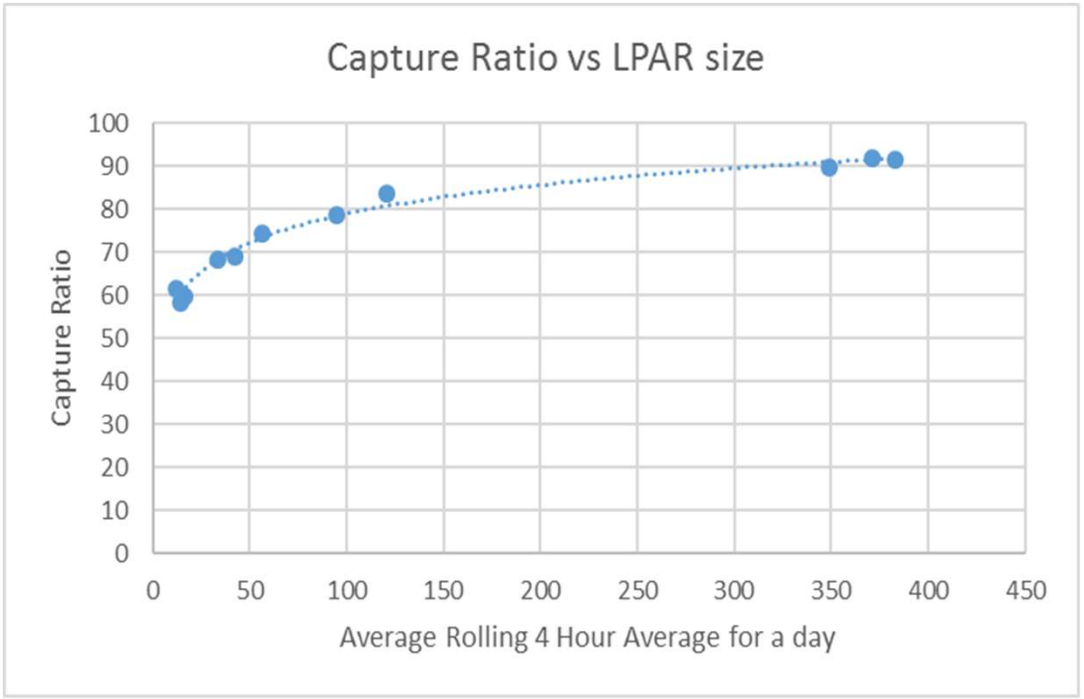
Uncaptured time is part of RMF 70 R4HA, so it is charged to you with MLC pricing.

LPAR Size and Capture Ratio

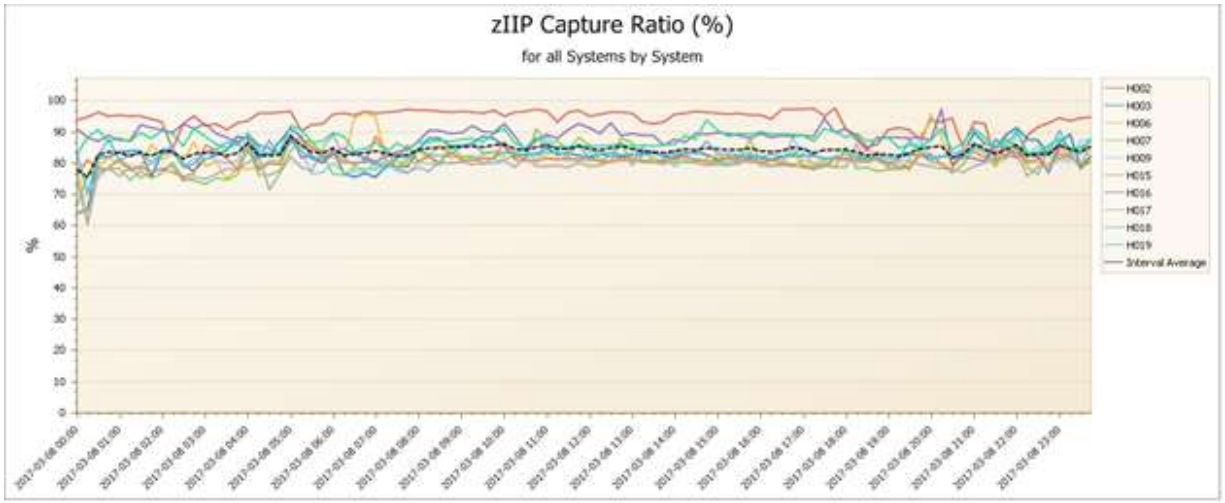
Some old wisdom still applies:

- Small LPARs (1 or 2 CPs) are expensive in terms of z/OS overhead too
- Uncaptured MSUs are MLC-charged just like captured MSUs

Consolidating LPARs will result in higher capture ratio, and lower MSU charges.



zIIP Capture Ratio

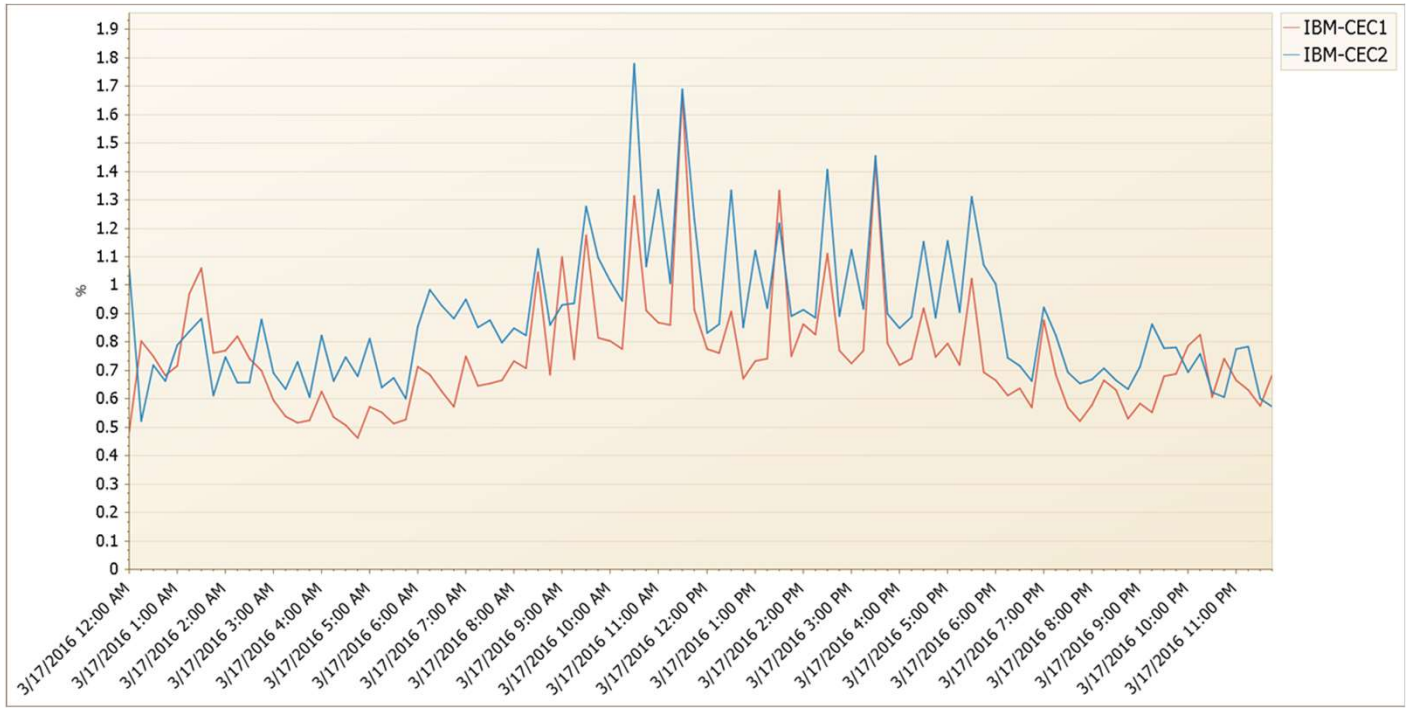


- Same concept as for CP time
- Less relevant from chargeback perspective

LPAR Management Time

- CPU consumed by PR/SM LPAR management overhead, **not part of MLC prices**
- RMF assigns this overhead to LPARs when possible
  - Time that cannot be attributed to an LPAR is reported by RMF in *PHYSICAL* LPAR
- Factors contributing to LPAR overhead include the number of LPARs & number of logical CPs defined
  - Rule of Thumb is Logical CP to Physical CP ratio of 2:1

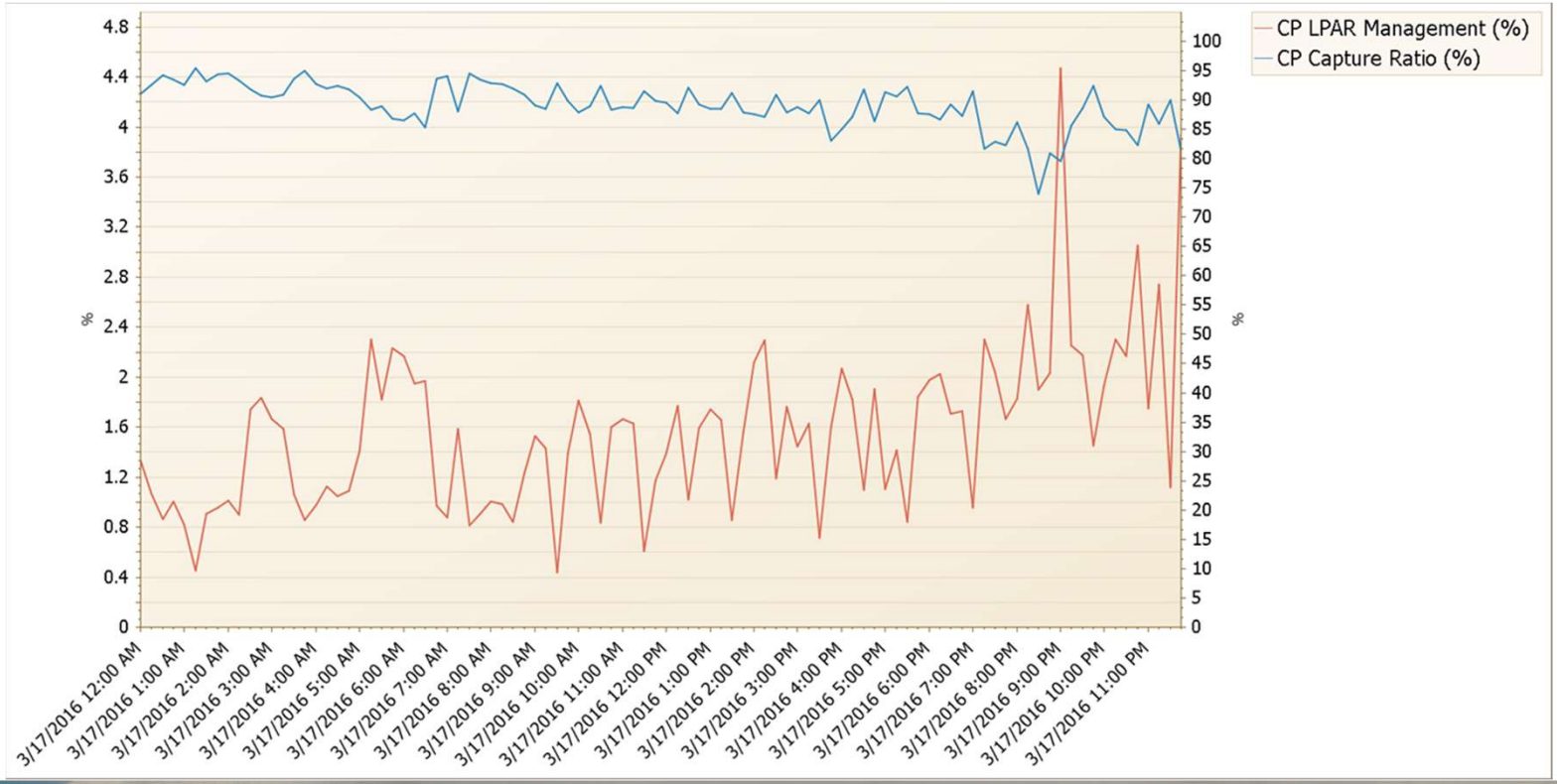
Unattributed LPAR Overhead - *PHYSICAL*



LPAR management is **not** part of RMF 70 R4HA, so it is **not** charged to you.

It is also very small.

CP LPAR Management & Capture Ratio



zIIP Processing

zIIPs

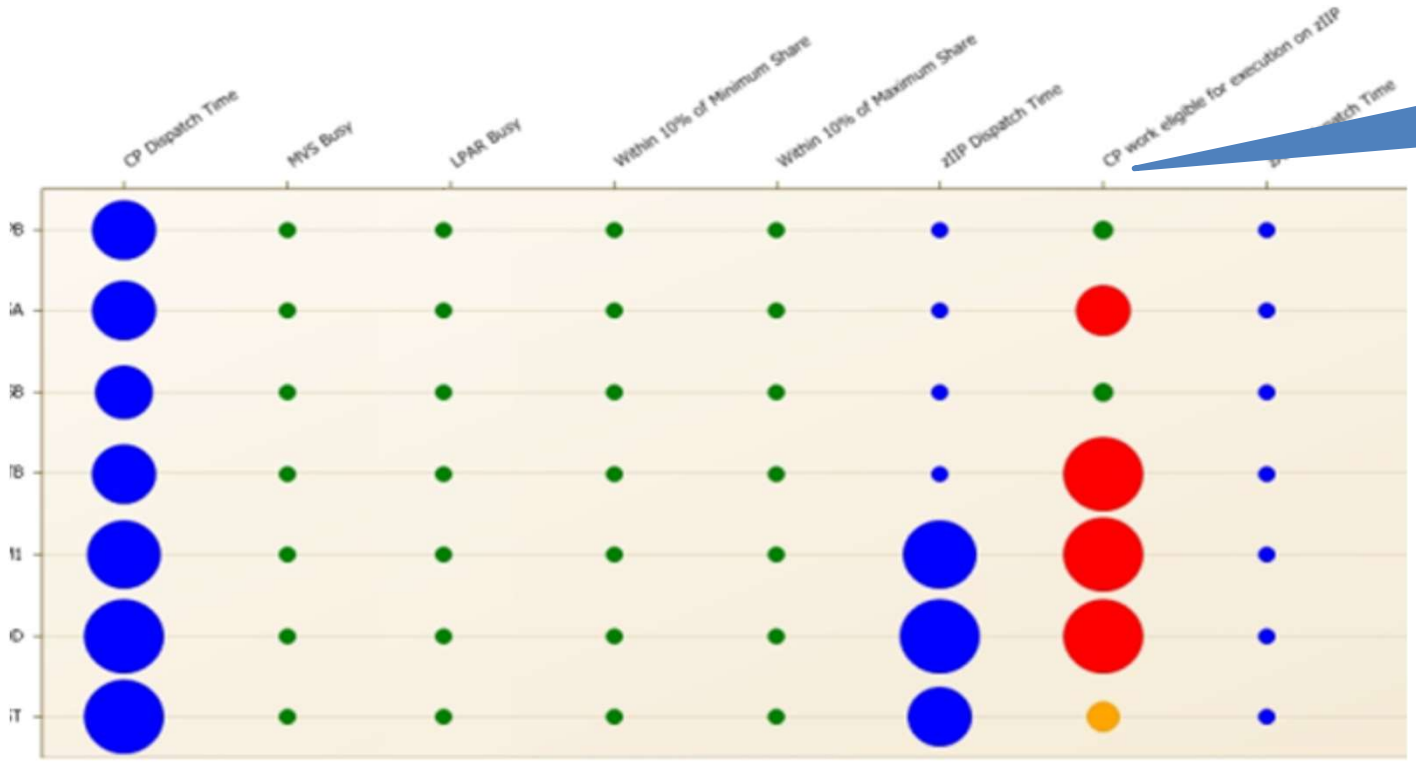
- Separate hardware to execute zIIP-eligible workloads without incurring software costs
- zIIPs always execute at full rated speed (even on subcapacity models)
- zIIP-eligible workloads can include
  - Java
  - Distributed DB2
  - Various ISV product functions
  - Note: consolidates work formerly executing on zAAPs

% zIIP Utilization

zIIP Eligible Work on GCPs

- zIIP eligible work can also overflow to execute on general purpose CPs if zIIP CPs are fully utilized
- This "overflow" work can drive software expense, so awareness is important
- IEAOPTxx parameter IIPHONORPRIORITY=NO prevents this overflow
  - IF set to NO, DB2 will not utilize zIIPs for system tasks

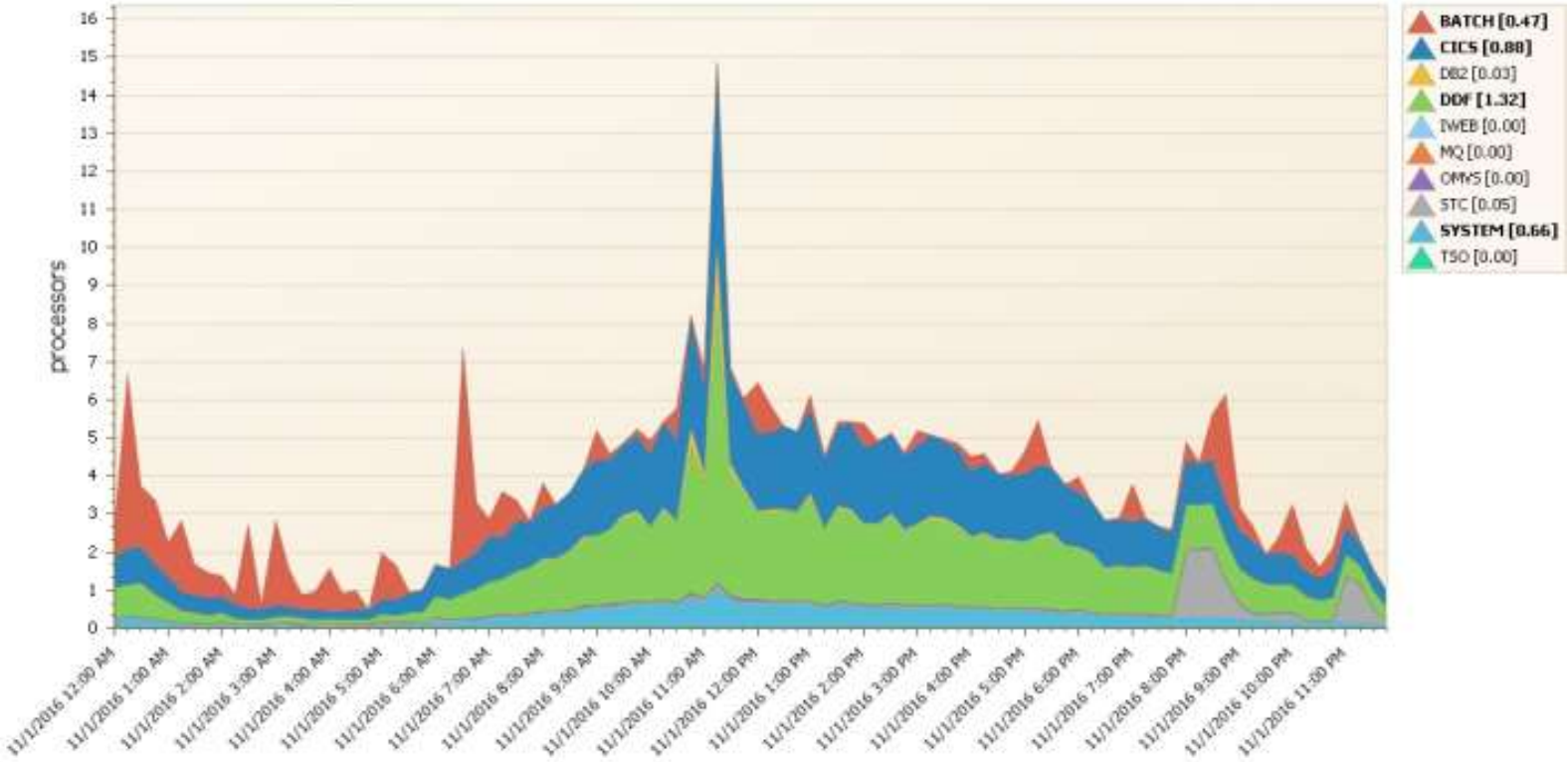
zIIP-eligible Work on GCPs



zIIP Eligible

Note that zIIP overflow to CP is only **really bad** during MSU peak hours.

zIIP-eligible Work on GCPs



Simultaneous MultiThreading (SMT)

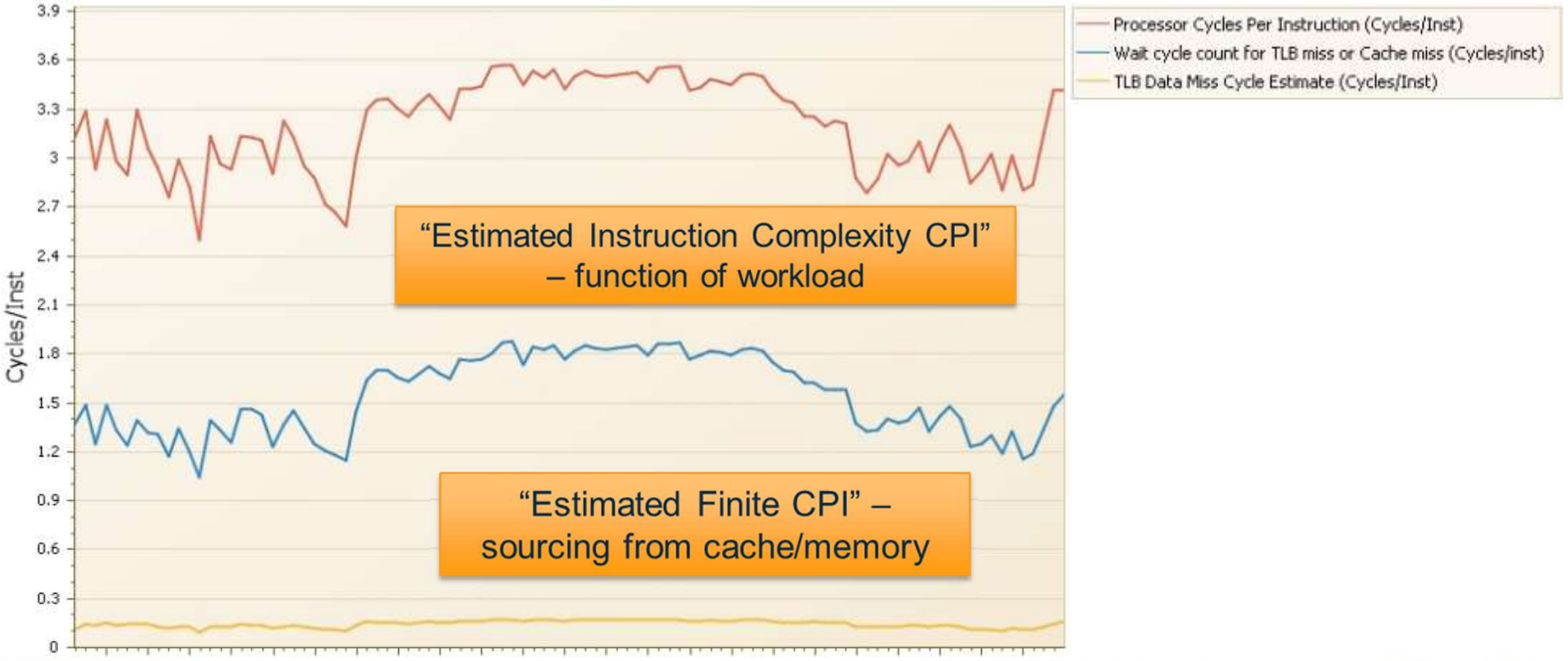
- Two threads (units of work) can execute simultaneously on one core (CP)
- Supported for zIIPs (not general purpose CPs)
- When to use
  - Sizable multi-threading workloads on zIIPs (e.g., Java)
  - Need to expand zIIP capacity
  - May not help compute-bound workloads
- Introduces significant amount of new RMF 70 instrumentation (discussed in Performance Insights)

Processor Cache Concepts & Metrics

Key Metrics – Cycles Per Instruction (CPI)

- Number of processor cycles spent per completed instruction
- Processor cycles are spent either
  - Productively – executing instructions
  - Unproductively – waiting to stage data (L1 cache miss)

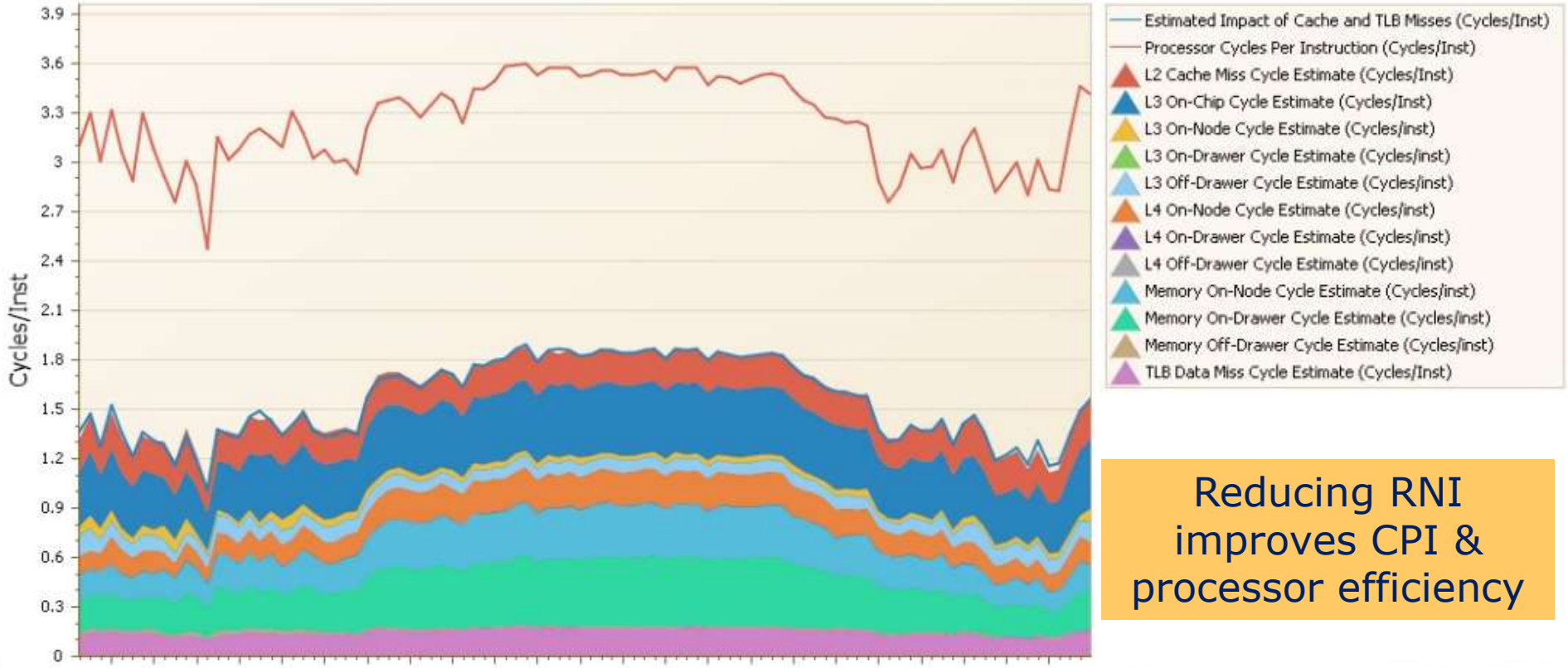
Cycles Per Instruction



Key Metrics – Relative Nest Intensity (RNI)

- How deep into the shared cache and memory hierarchy ("nest") the processor must to go retrieve data not present in L1 cache
- Access time increases significantly for each level of cache (increasing processor wait time)
- 2.3*(0.4*L3P+1.6*L4LP+3.5*L4RP + 7.5*MEMP) / 100
- Reducing RNI improves processor efficiency

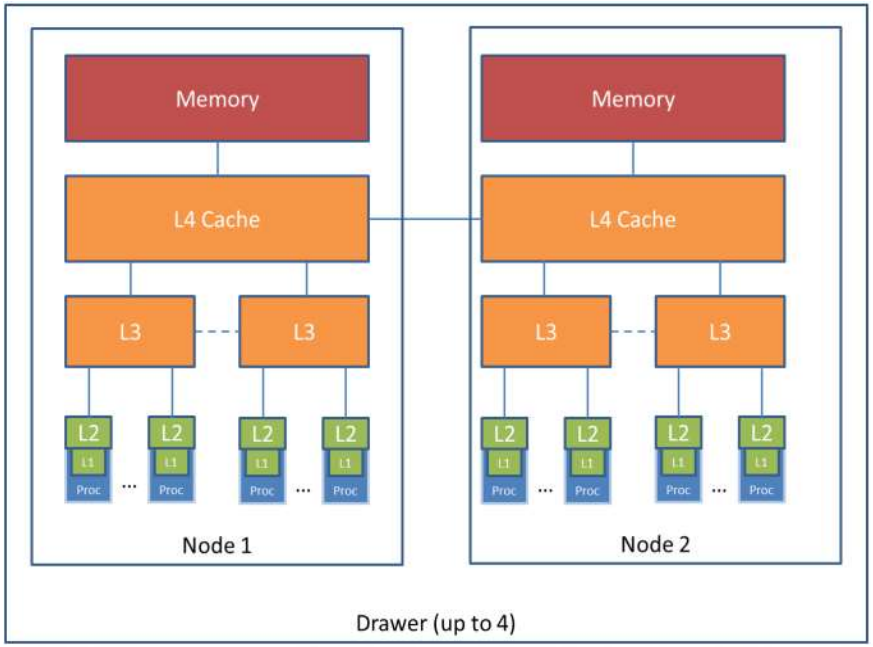
CPI with Components of RNI



Reducing RNI improves CPI & processor efficiency

HiperDispatch

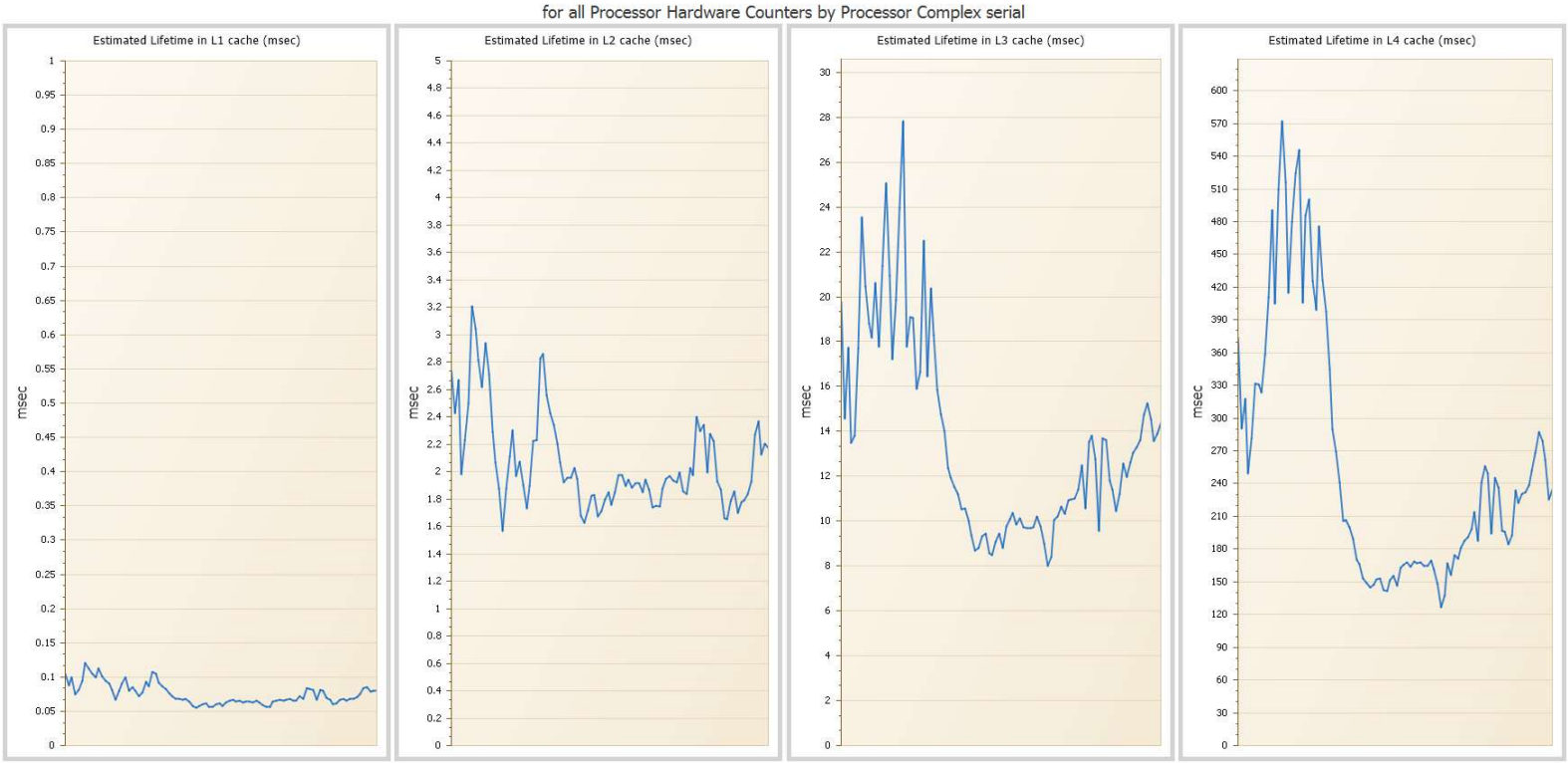
- Interfaces with PR/SM & z/OS Dispatchers to align work to logical processors (LPs) & align LPs to physical CPs
- Repeatedly dispatching the same work to the same or nearby CP is vital to optimizing processor cache hits



Vertical CP Configuration

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

Cache Data Lifetime

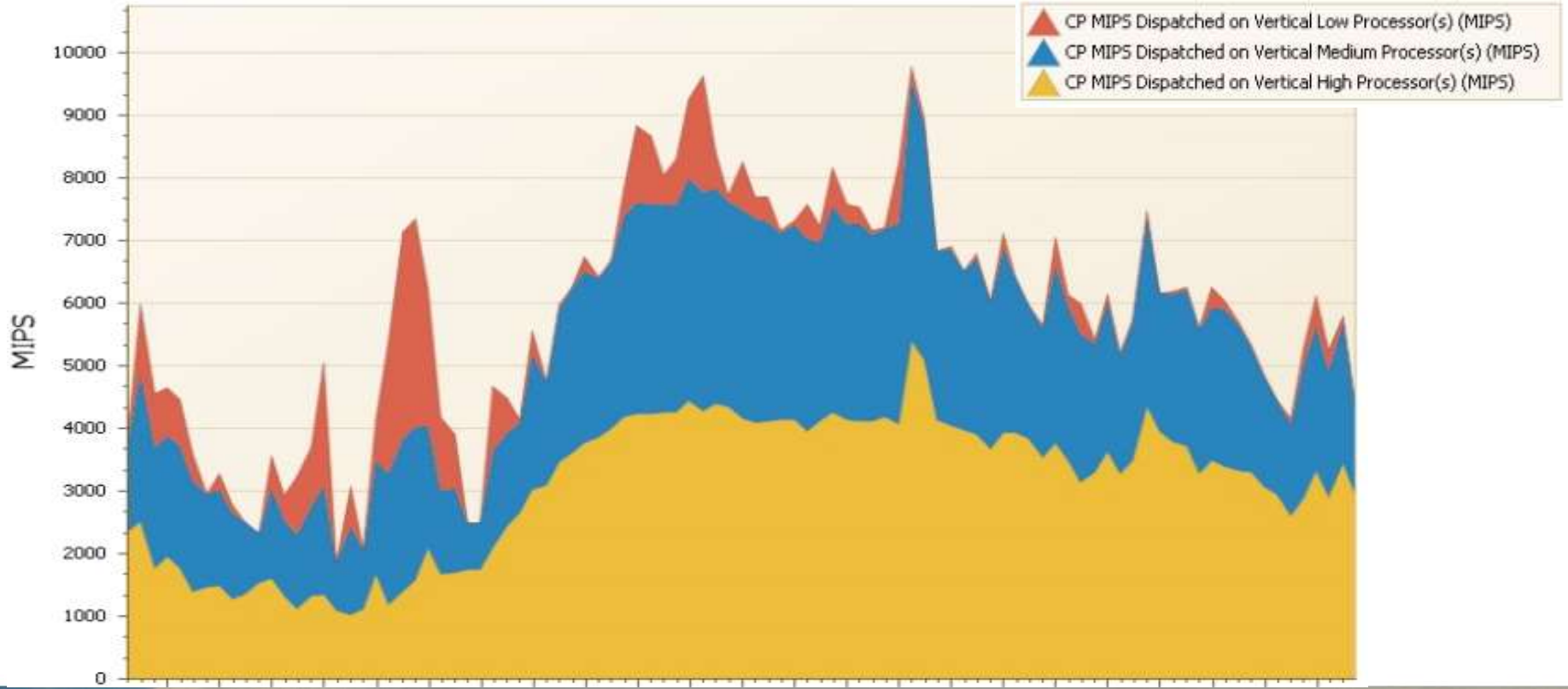


RNI Impact of Executing More Work on VHs

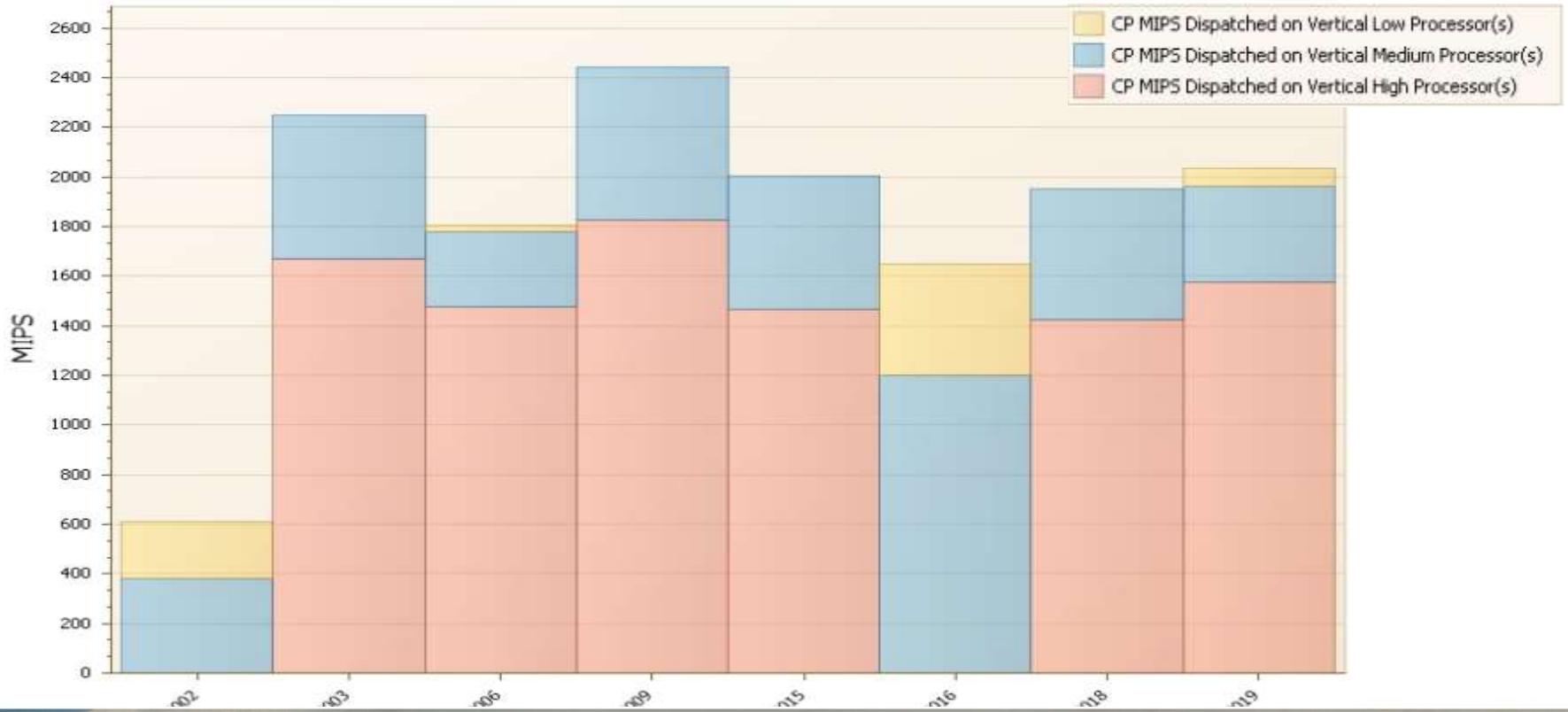
RNI by Logical CP

% CP Time Dispatched on VHs by CEC

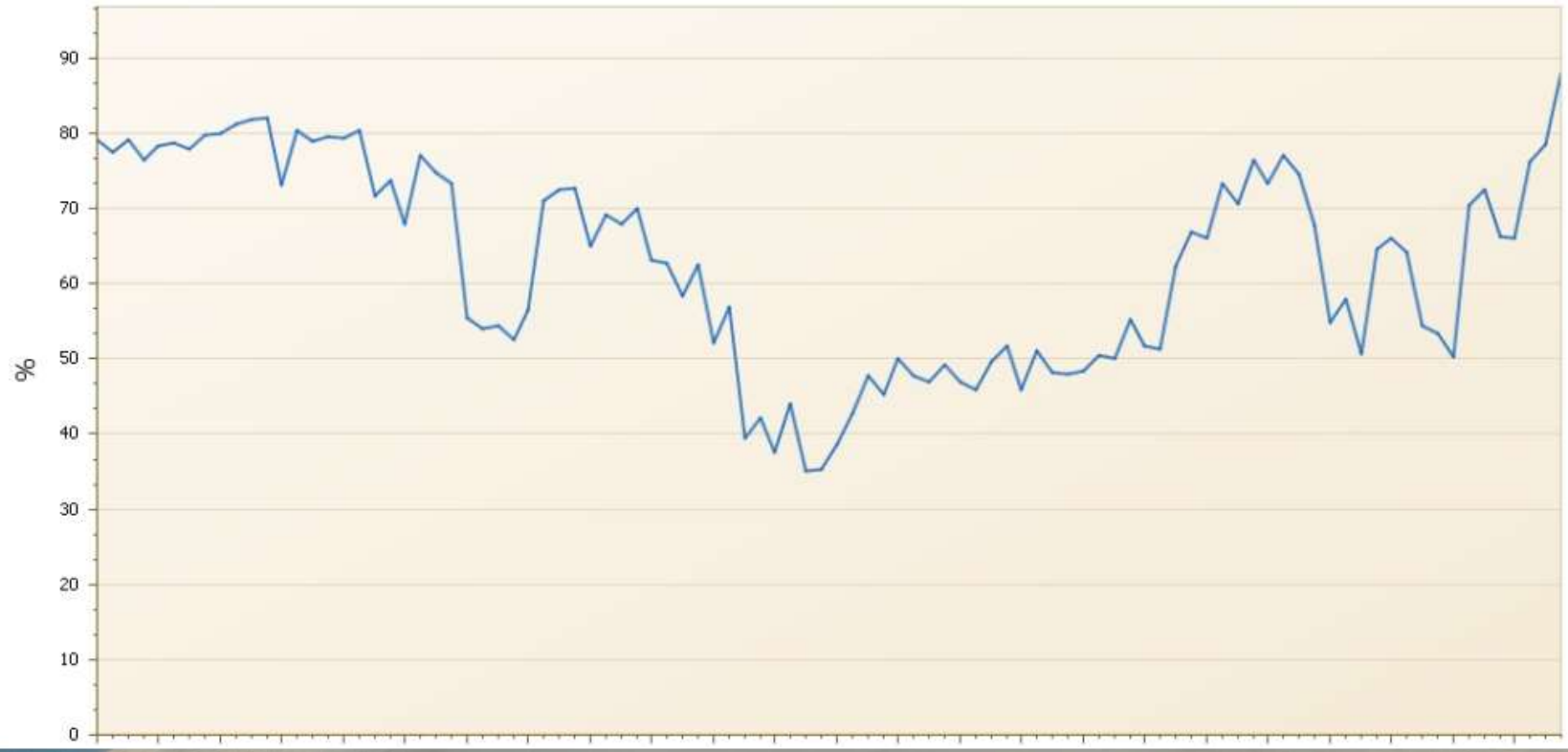
Dispatched MIPS by CEC by Polarity



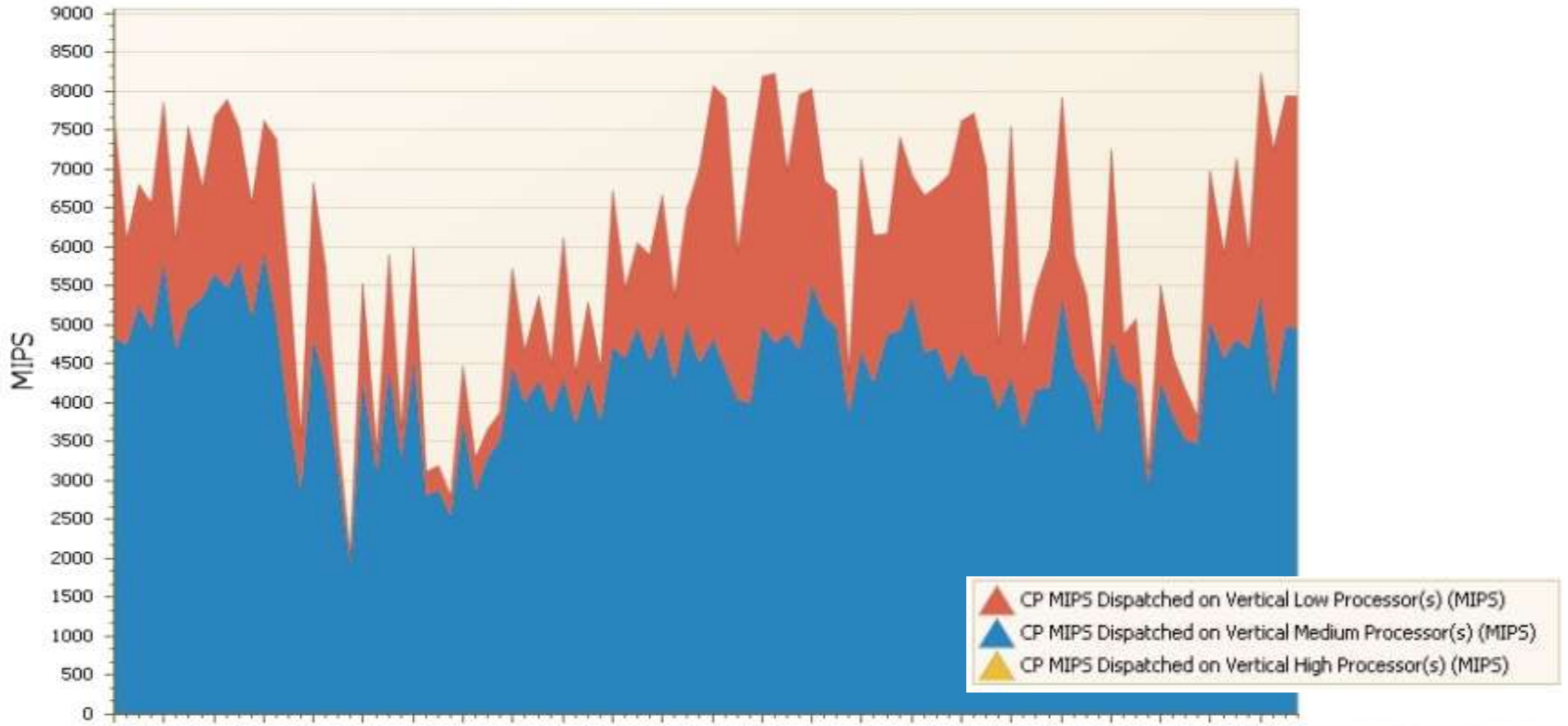
Dispatched MIPS by LPAR by Polarity



% CP Time Dispatched on VHs by CEC



Dispatched zIIP MIPS by CEC by Polarity



Ways to Increase Work Executing on VH CPs

- Adjust LPAR weights
  - Increase weights for high CPU LPARs
  - **Customize weights by shift**
  - Configure fewer, larger LPARs
- Increase the number of physical CPs
  - Install additional hardware capacity
  - Utilize sub-capacity hardware models

LPAR Topology

- PR/SM dynamically assigns LPAR CPs and memory to the hardware (chips, nodes and drawers) to optimize cache efficiency
- This topology can have a significant impact on performance
- Topology data is captured in the SMF 99.14 records

LPAR Topology

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

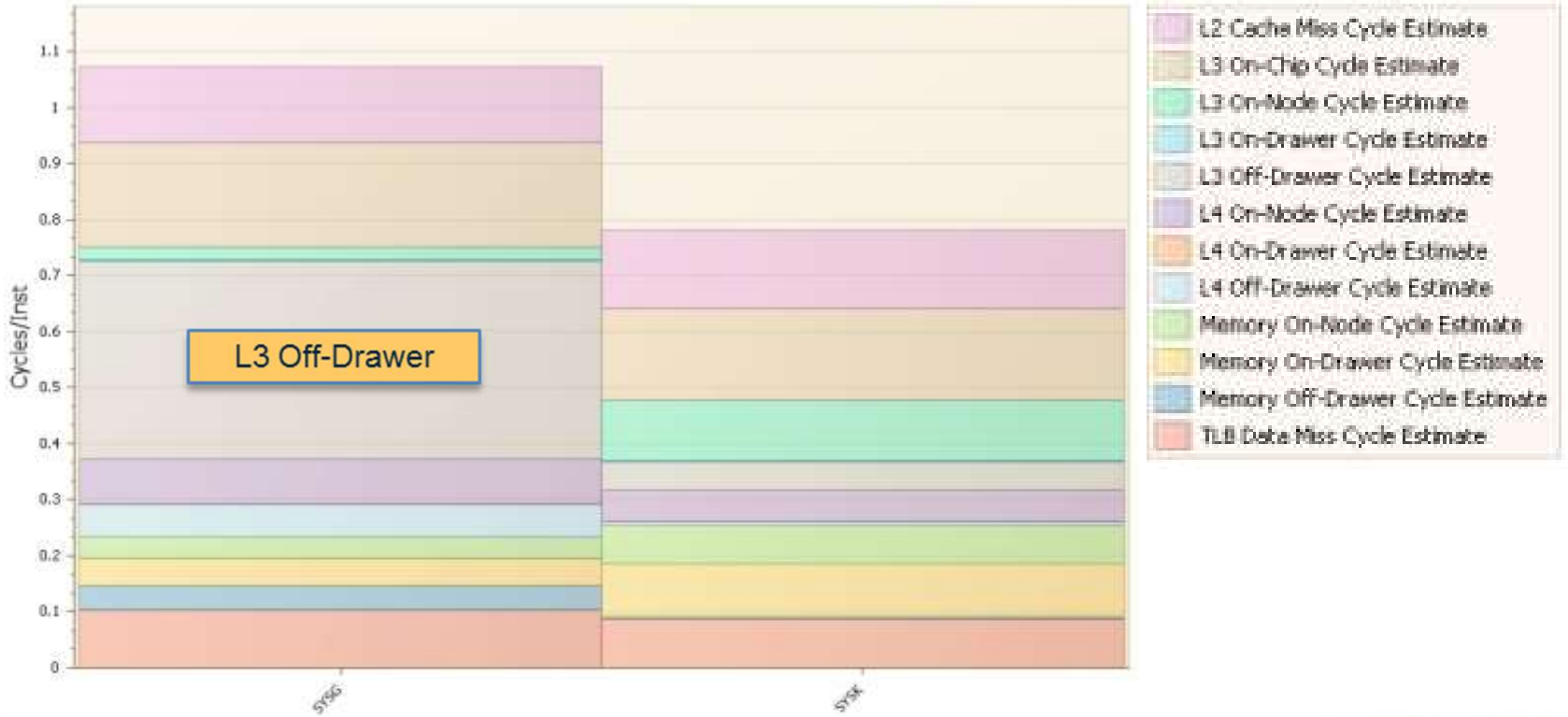
System	WLM Node	Proces...	Vert...	Vert...	Vert...	WLM Node Flags	WLM ...	Chip...	Nod...	Draw...
SYS2	0001	CP	2	1	7	CPUs/cores on this node are boundary crossing		0	1	2
SYS2	0002	zIIP	0	1	3	CPUs/cores on this node are boundary crossing		0	1	2

LPAR Topology

Drawer 3 Node 1	Chip 1 (3P)	Chip 2 (1P)	Chip 3 (5P)
	SYSG VM02	SYSG VM03	SYSK VH00
	SYSK VM04	SYSK VL05	SYSK VH01
	SYSV VH00	SYSK VL06	SYSK VH02
	SYSV VH01	SYSK VL07	SYSK VH03
		SYSG VLx2 *	SYSG VLx4 *
Drawer 4 Node 1		Chip 2 (3P)	
		SYSG VH00	
		SYSG VH01	
		SYSK VLx2 *	
* - not executing any work			

Estimated Impact Cache and TLB Misses for CPs

for all Processor Hardware Counters by System



Conclusions

- IBM has kept up with instrumentation for processor reporting in RMF and SMF 30.
- Classic concepts like Capture Ratio are still relevant
- Processor Polarization (High/Medium/Low) has a very significant impact on MIPS/MSU achieved (>10%)

Questions?


CMG
Computer Measurement Group
imPACt
INTERNET • MOBILE • PERFORMANCE & CAPACITY • CLOUD • TECHNOLOGY
NOVEMBER 6-9, 2017 | LOEWS HOTEL | NEW ORLEANS, LA