The Hidden Gold in the SMF 99s

Peter Enrico

Email: Peter.Enrico@EPStrategies.com

Enterprise Performance Strategies, Inc.
3457-53rd Avenue North, #145
Bradenton, FL 34210
http://www.epstrategies.com
http://www.pivotor.com

Voice: 813-435-2297
Mobile: 941-685-6789

z/OS Performance
Education, Software, and
Managed Service Providers

Creators of Pivotor®

Contact, Copyright, and Trademark Notices

Questions?

Send email to Peter at Peter.Enrico@EPStrategies.com, or visit our website at http://www.epstrategies.com or http://www.pivotor.com.

Copyright Notice:

Trademarks:

Abstract

- The Hidden Gold in the SMF 99s

  - The SMF 99 records contain a wealth of information related to WLM algorithm decisions. They were originally developed to trace WLM decisions, but over the years they have been expanded to provide insights into HiperDispatch, Capping, Group Capacity Limits, machine topology, and more. Most customers have the SMF 99 WLM decision records turned off due to their high volume. However, there are many reasons to turn these records on during performance debugging and analysis.

  - During this presentation, Peter Enrico will provide an introduction to the SMF 99 records, as well as show some very practical uses and a number of performance insights that these records provide.

Overview of SMF 99 Subtypes

- Subtype 1
  - System level measurement data used for decision input
  - Trace of WLM actions
  - Written every 10 seconds (i.e. policy adjustment interval)
- Subtype 2
  - Service class period measurement data used for decision input
  - Written every 10 seconds (i.e. policy adjustment interval)
- Subtype 3
  - Service class period plot data
  - Written every 10 seconds (i.e. policy adjustment interval)
- Subtype 4
  - Service class device cluster information
  - Written every 10 seconds (i.e. policy adjustment interval)
- Subtype 5
  - Data about monitored address spaces
  - Written every 10 seconds (i.e. policy adjustment interval)

Overview of SMF 99 Subtypes cont…

- Subtype 6
  - Service class period settings and measurements
  - Written every 10 seconds (i.e. policy adjustment interval)
- Subtype 7
  - Enterprise Storage Server ® (ESS) with Parallel Access Volumes (PAVs)
  - Written every 30 seconds (i.e. 3 policy adjustment intervals)
- Subtype 8
  - Information about LPAR CPU management
  - Written every 10 seconds (i.e. policy adjustment interval)
- Subtype 9
  - Information about dynamic channel path management
  - Written every 10 seconds (i.e. policy adjustment interval)
- Subtype 10
  - Information about dynamic processor speed changes
  - Written when speed changes

Overview of SMF 99 Subtypes cont…

- Subtype 11
  - Information about Group Capacity Limits
  - Written every 5 minutes
- Subtype 12
  - HiperDispatch interval data
  - Written every 2 seconds (i.e. policy adjustment interval)
- Subtype 13
  - HiperDispatch IBM internal use only (so undocumented)
- Subtype 14
  - HiperDispatch topology data
  - Written every 5 minutes

SMF 99 Recommendations

- Consider regularly collecting the following SMF 99 subtypes
  - Subtype 6 - Service class period settings and measurements
  - Subtype 11 - Information about Group Capacity Limits
  - Subtype 12 - HiperDispatch interval data
  - Subtype 14 - HiperDispatch topology data

  - Collectively these records typically produce about 40MiB/system/day
  - They contain the most interesting and useful data of the 99s

- Records to collect for problem periods of time, or when doing a study to better understand WLM decision making
  - Subtype 1 - System level measurement and trace data used for decisions
  - Subtype 2 - Service class period measurement data used for decision input
  - Subtype 3 - Service class period plot data
  - Subtype 5 - Data about monitored address spaces

  - Then call Peter Enrico and / or Scott Chapman to process with Pivotor

SMF 99.6

SMF 99.6 Overview

- Subtype 6
  - Service class period settings and measurements
  - Written every 10 seconds (i.e. policy adjustment interval)
  - The purpose of this subtype is to record the WLM controls that are set for for each service class period

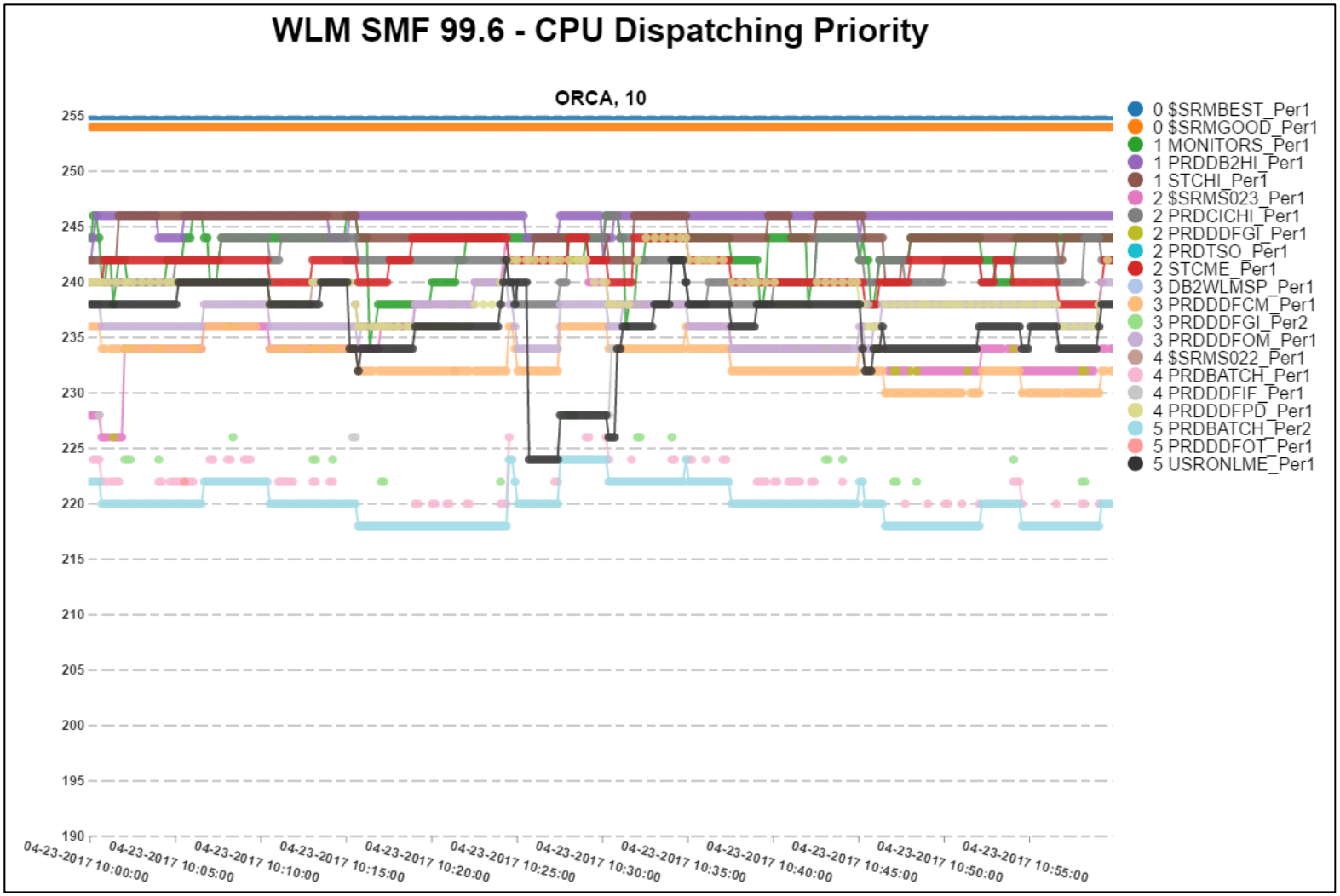
- It is recommended that SMF 99.6 record be turned on
  - Typically about 40MiB/system/day

- Key data in the SMF 99.6 includes
    - Service class, service class period, and goal information
    - Performance Indexes – both local and Sysplex PIs
    - CPU and I/O dispatching priorities
    - CPU service consumption (CP / zIIP / zAAP)
    - MPL in-targets and out-targets
    - Storage isolation and protection
    - For $SRMSxxx periods – the external service class period(s) served

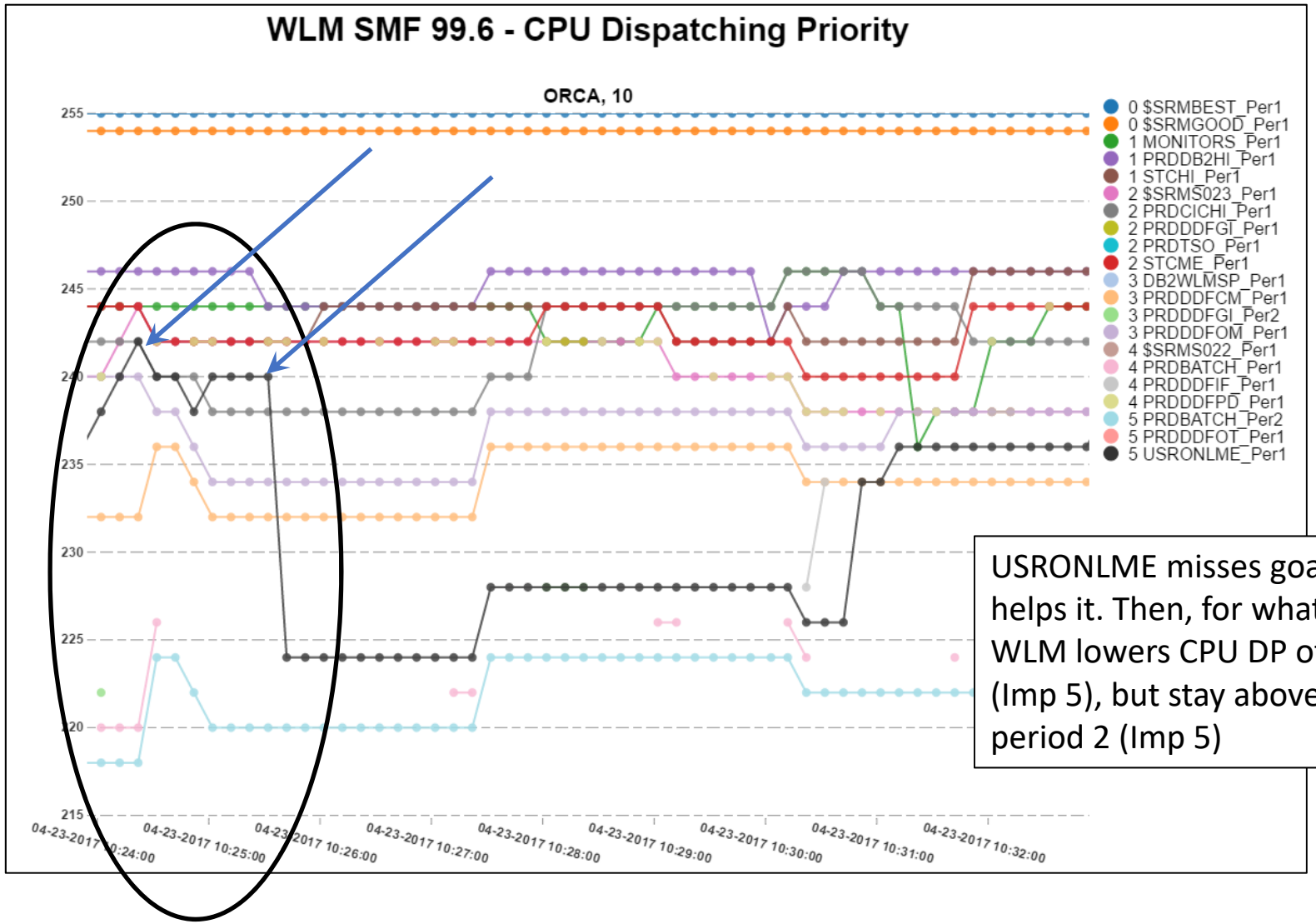
Using the SMF 99.6 record

- The SMF 99.6 record is helpful for answering the following questions:
  - Over time, what is the assigned dispatching priorities of each service class period?
  - How do the priorities change over time?
  - Relative to the goal value and importance level, is the assign priority as desired?
  - How much service is accumulated by each period every 10 seconds?
  - How much service is accumulated at CPU priorities above, below, and at the priority of the service class period being studied?
  - What is the relationship between the local PI and the Sysplex PI?
    - Is the Sysplex PI delaying WLM from helping a period missing its local PI?
  - What is server / served relationship between external periods and internal periods?
  - What types of protections are in place for large storage intensive workloads?

SMF 99.6 CPU Dispatching Priority – Every 10 Seconds

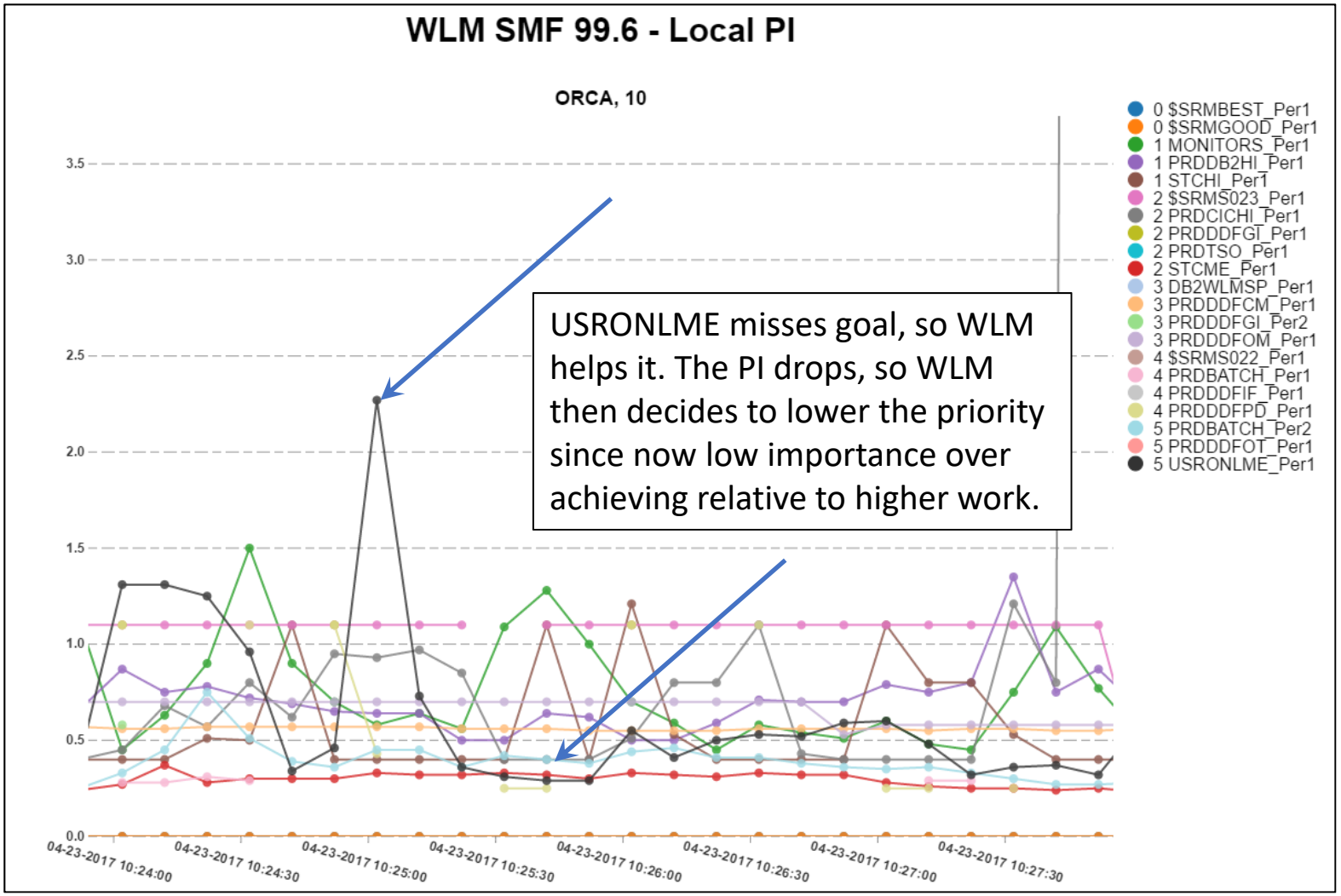


SMF 99.6 CPU Dispatching Priority – Every 10 Seconds

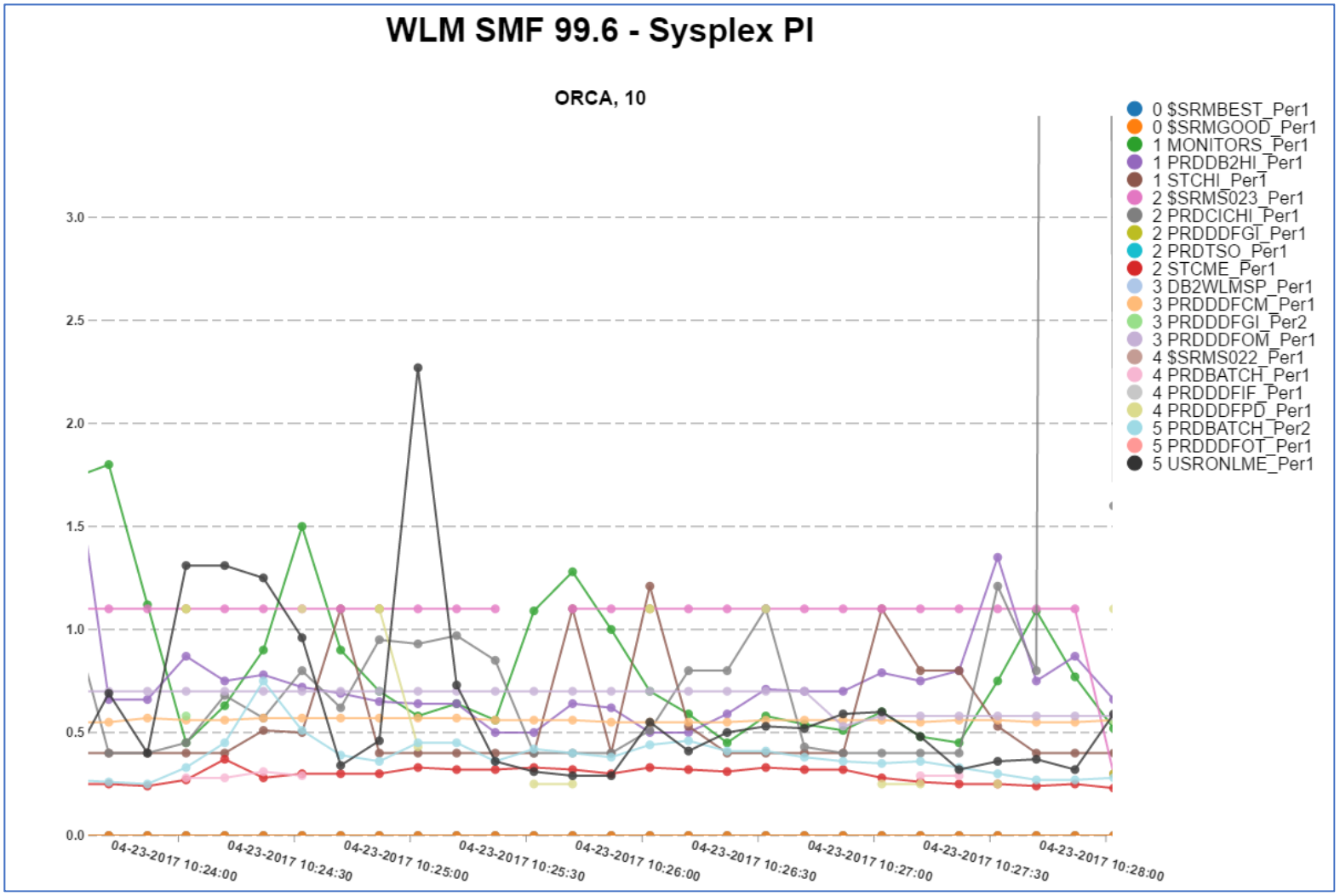


USRONLME misses goal, so WLM helps it. Then, for what ever reason, WLM lowers CPU DP of USRONLME (Imp 5), but stay above PRDBATCH, period 2 (Imp 5)

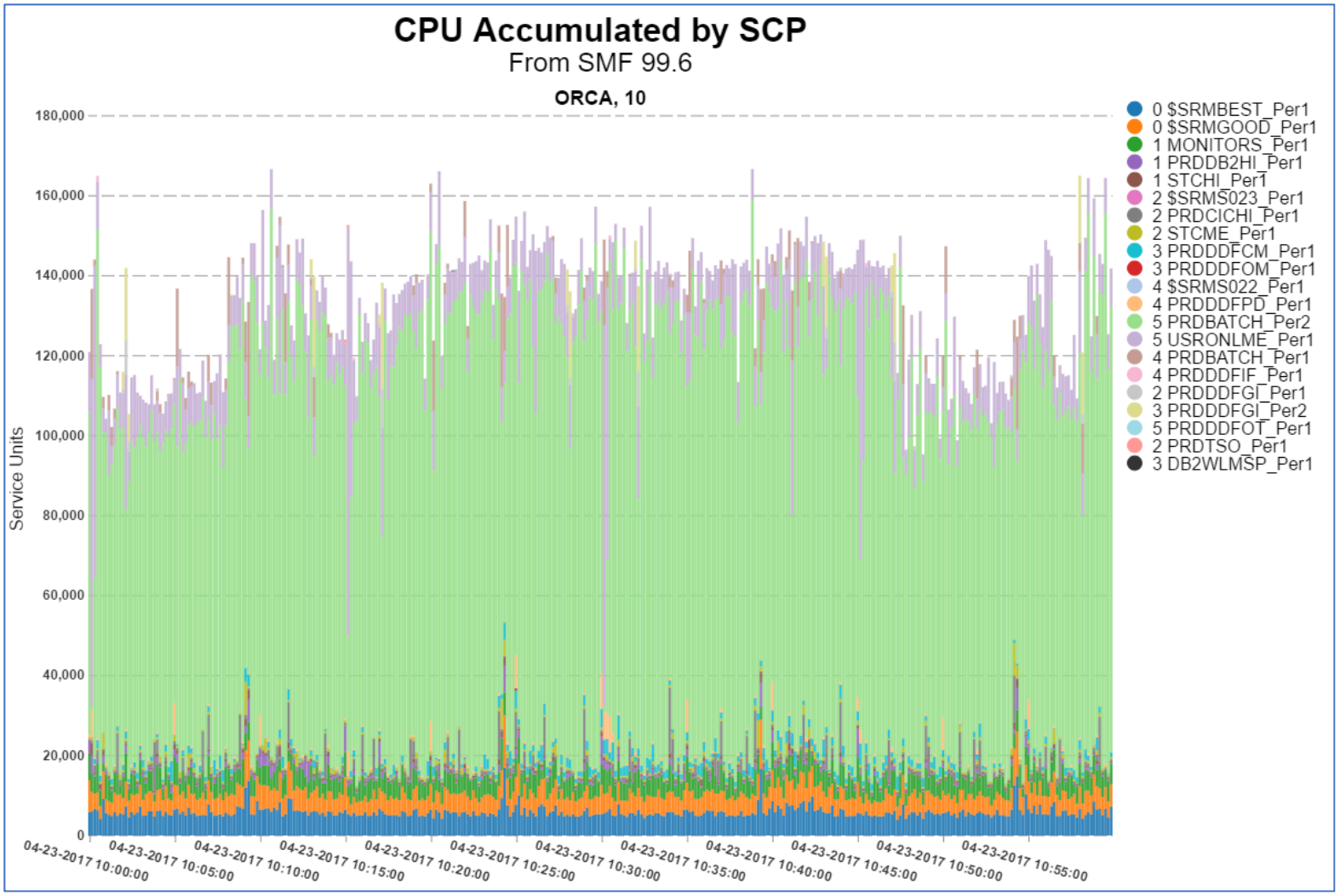
SMF 99.6 Local PI – Every 10 Seconds



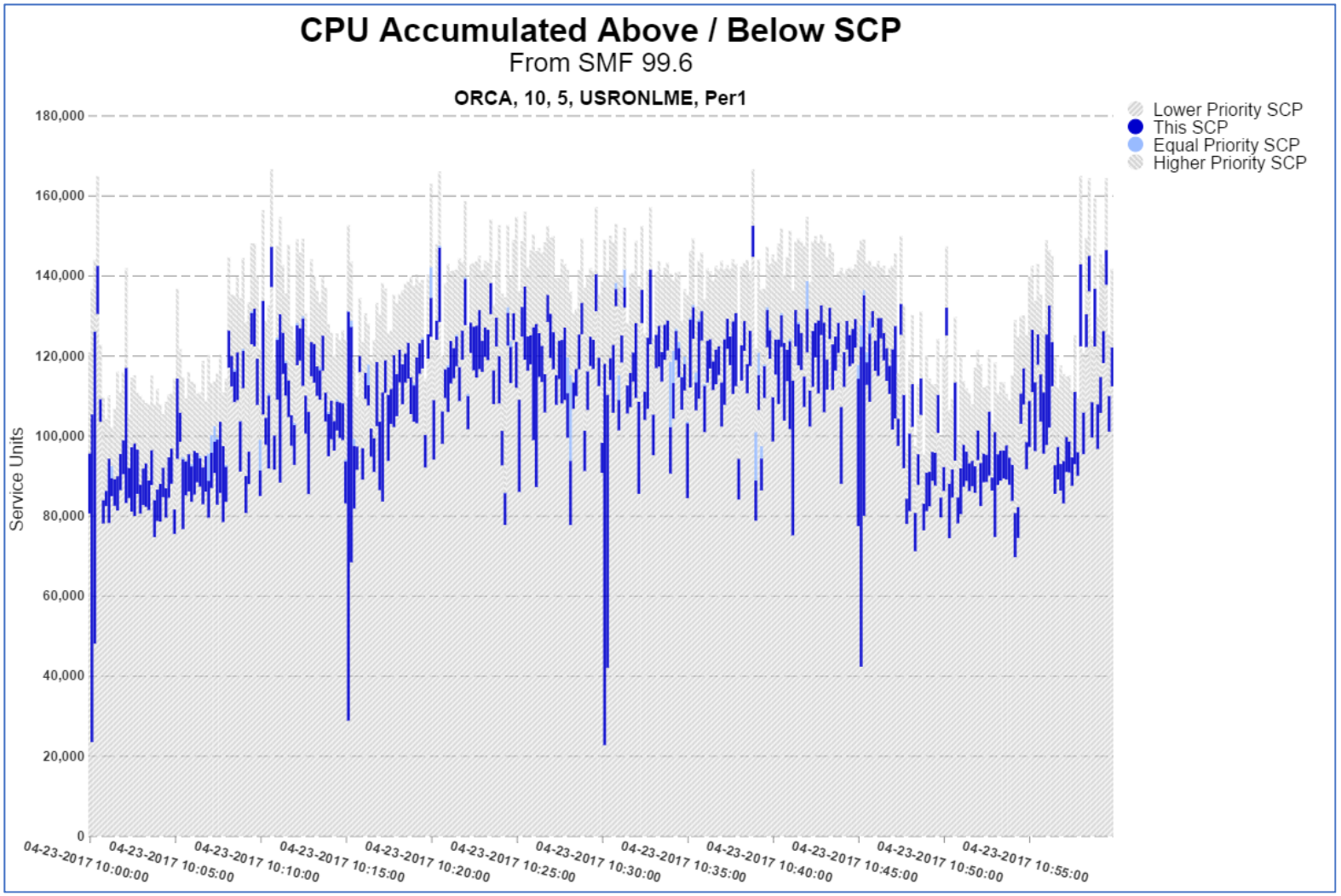
SMF 99.6 Sysplex PI – Every 10 Seconds



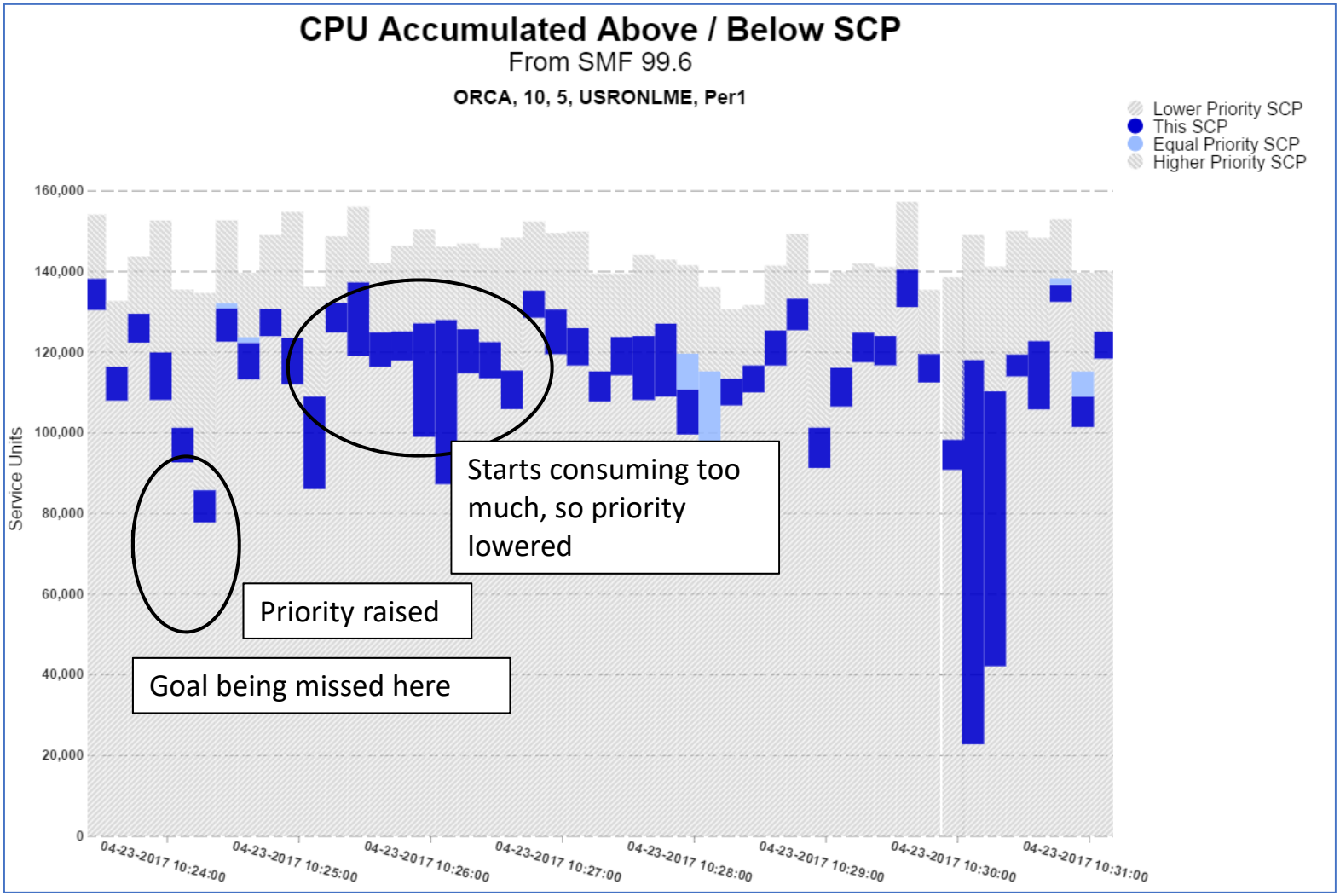
SMF 99.6 Service Consumed by Period – Every 10 Seconds



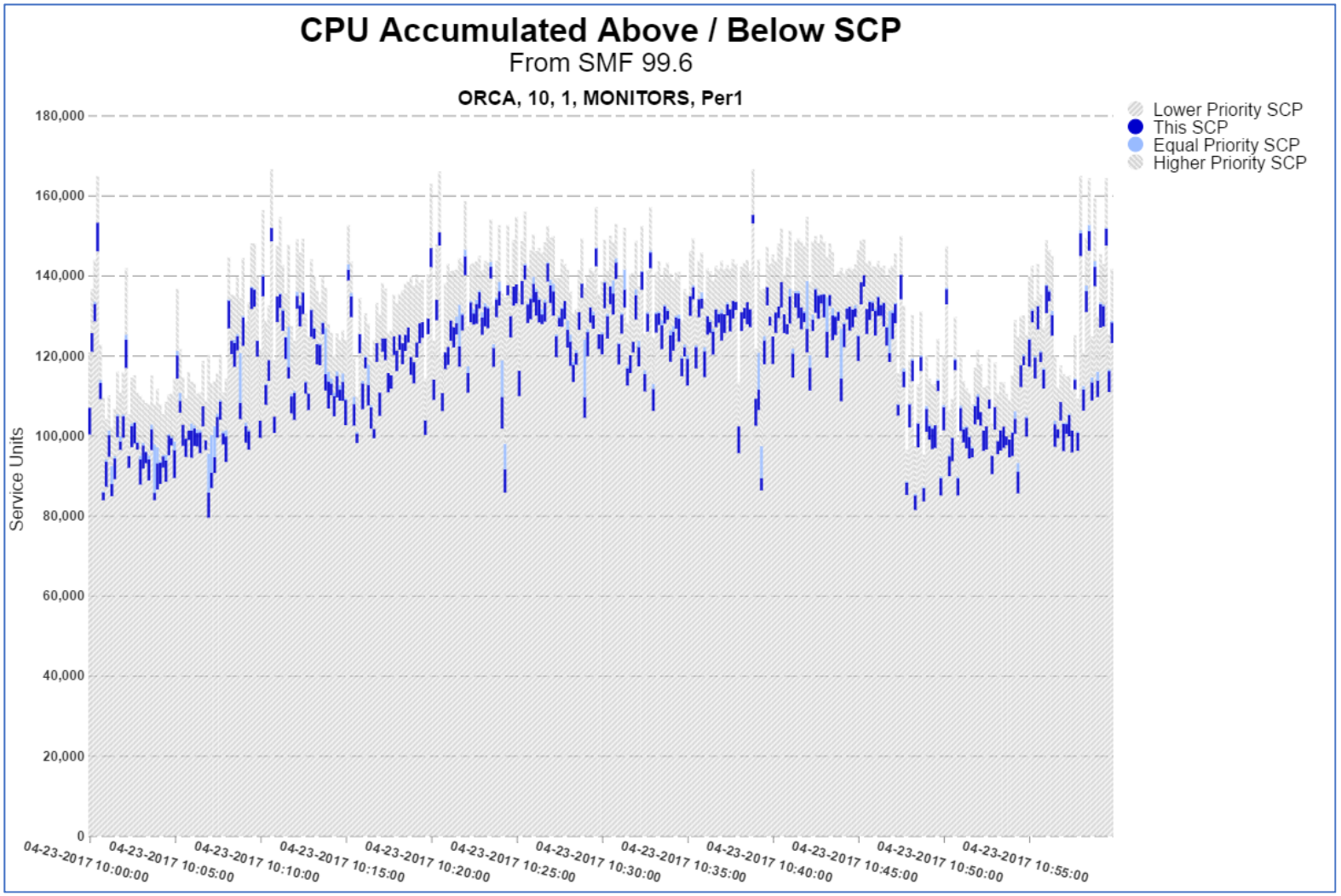
SMF 99.6 Service Consumed above / below – Every 10 Seconds



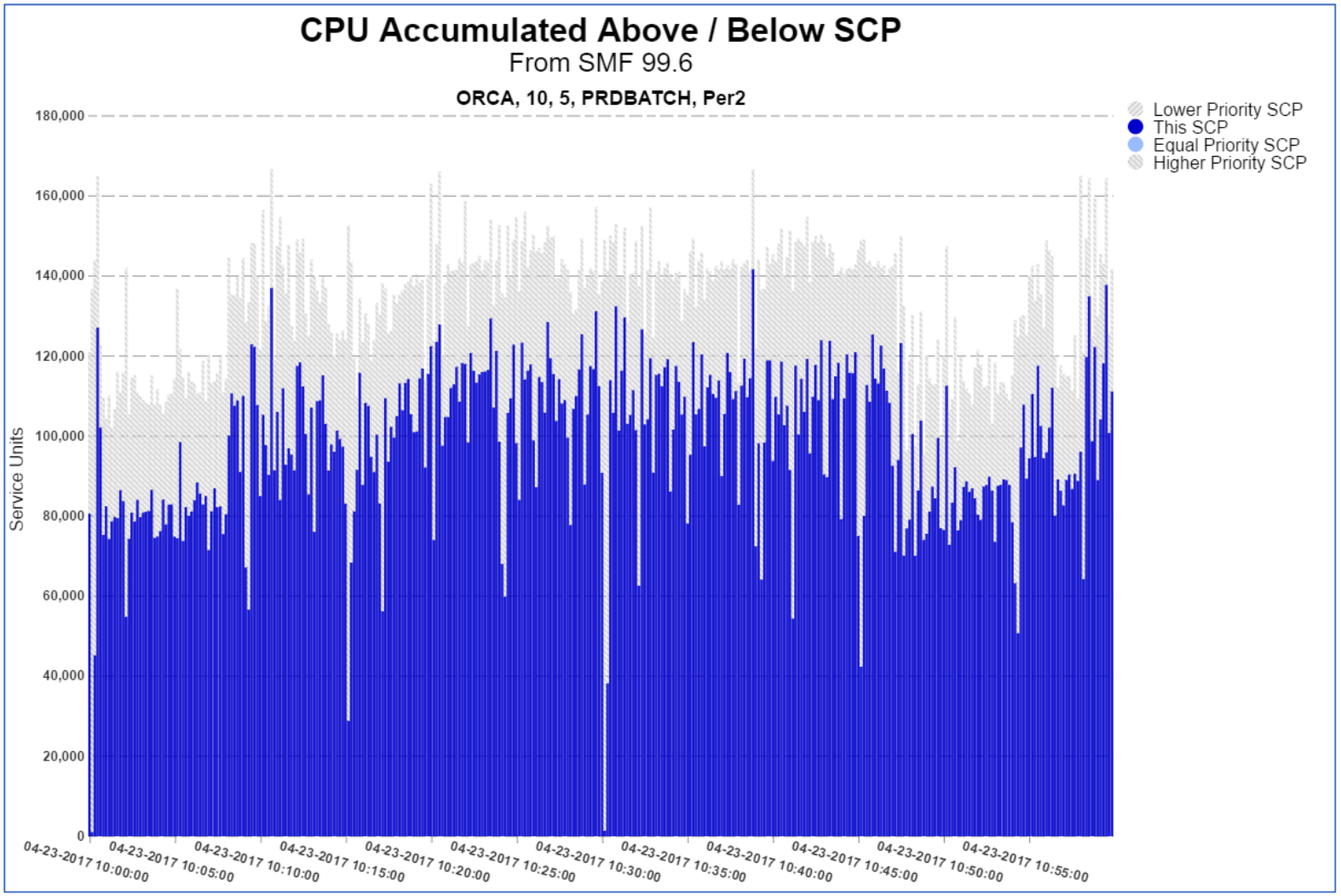
SMF 99.6 Service Consumed above / below – Every 10 Seconds



SMF 99.6 Service Consumed above / below – Every 10 Seconds



SMF 99.6 Service Consumed above / below – Every 10 Seconds



SMF 99.12

SMF 99.12 Overview

- Subtype 12
  - HiperDispatch interval data
  - Written every 2 seconds (i.e. HiperDispatch interval)
  - The purpose of this subtype is to record the factors that influence HiperDispatch parking and un-parking of processors

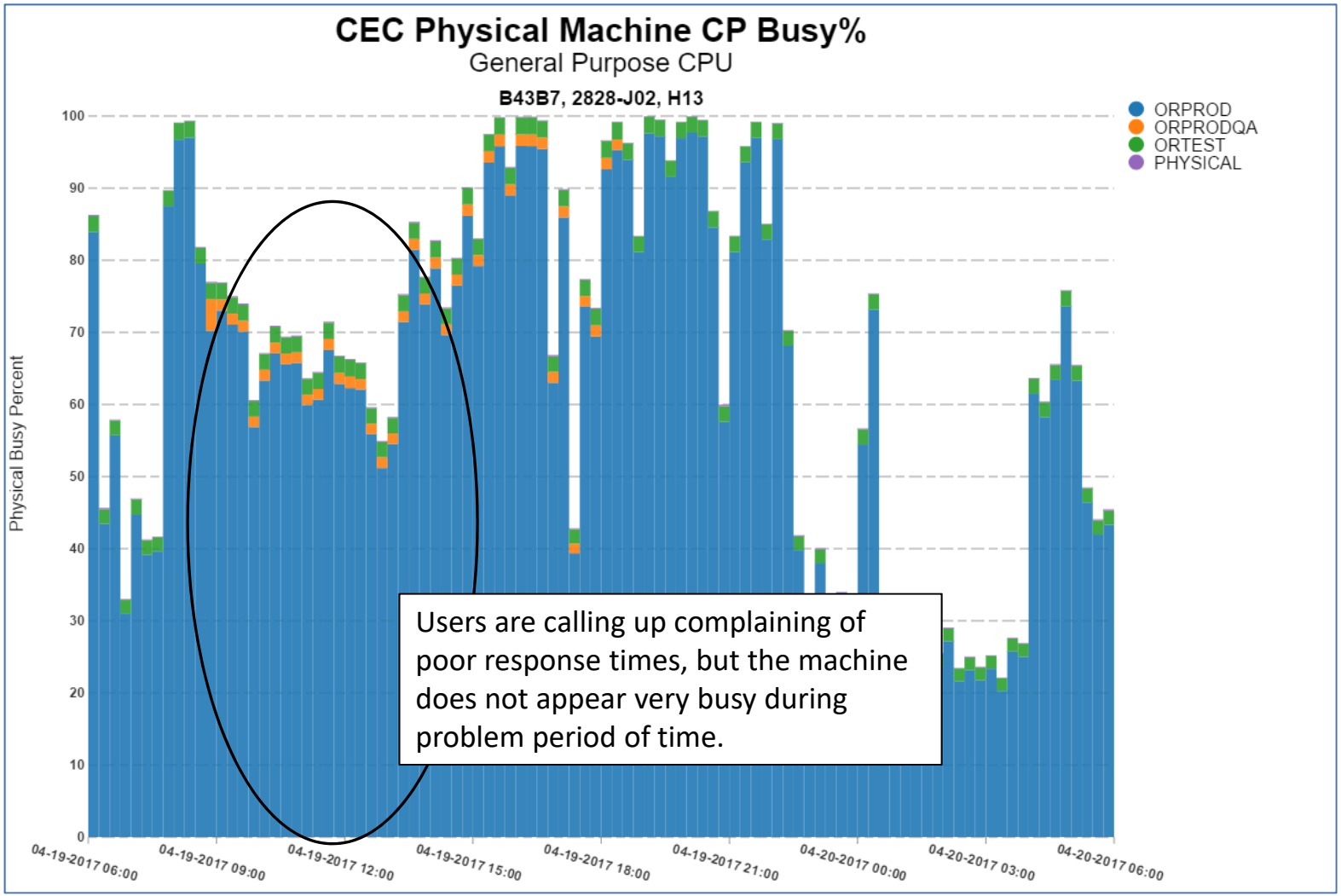
- It is recommended that SMF 99.12 record be turned on

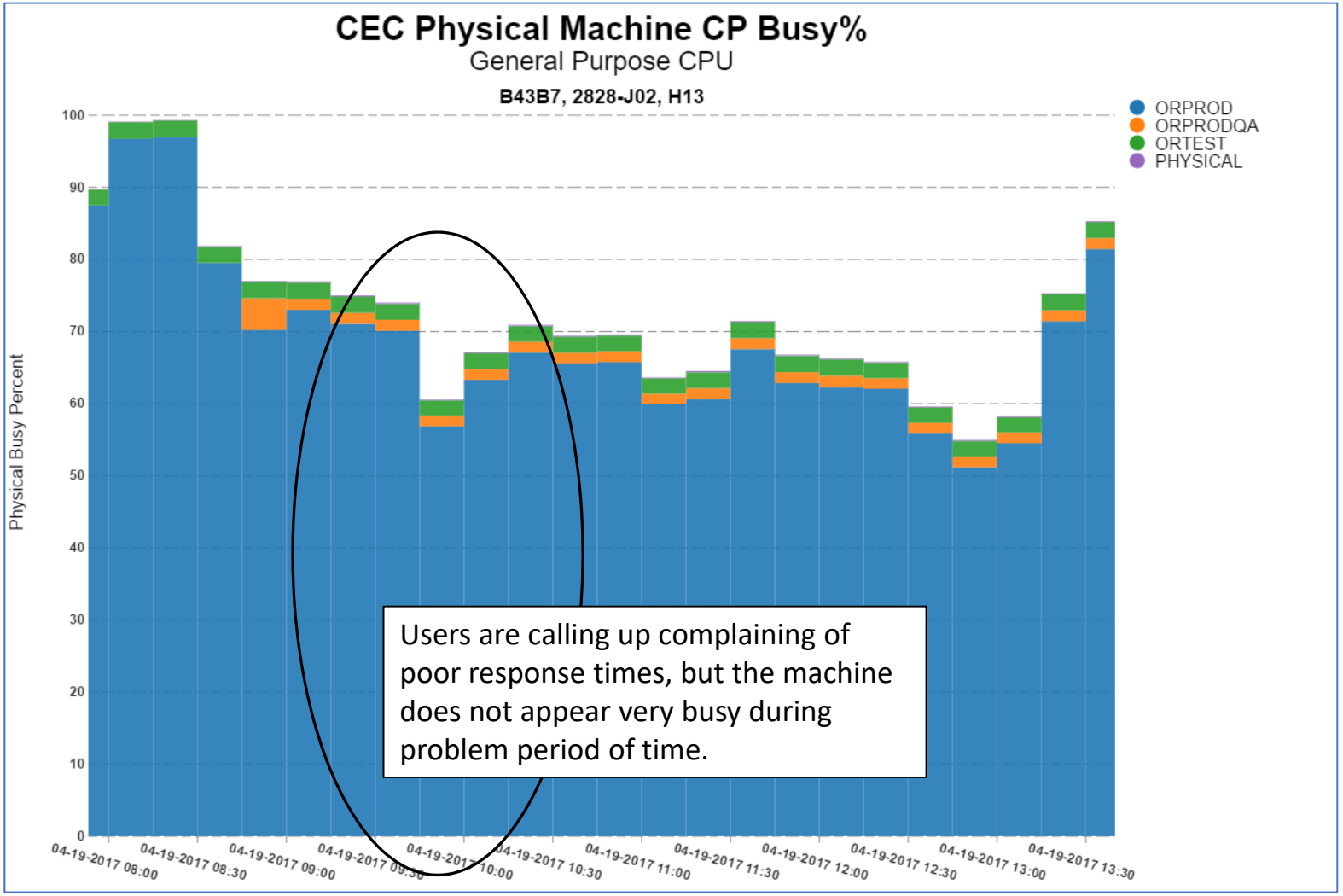
- Key data in the SMF 99.12 includes
  - LPAR level configuration information relevant to HiperDispatch
    - Example: LPAR share, LPAR capacities, SMT enablement, etc.
  - Processor utilizations (current and projected)
  - Pooling of Vertical Highs, Vertical Mediums, Vertical Lows
  - Capacity used / available to each pool (VHs, VMs, VLs)
  - Guaranteed shares to VHs, VMs, VLs
  - CPU displaced by parking and un-parking

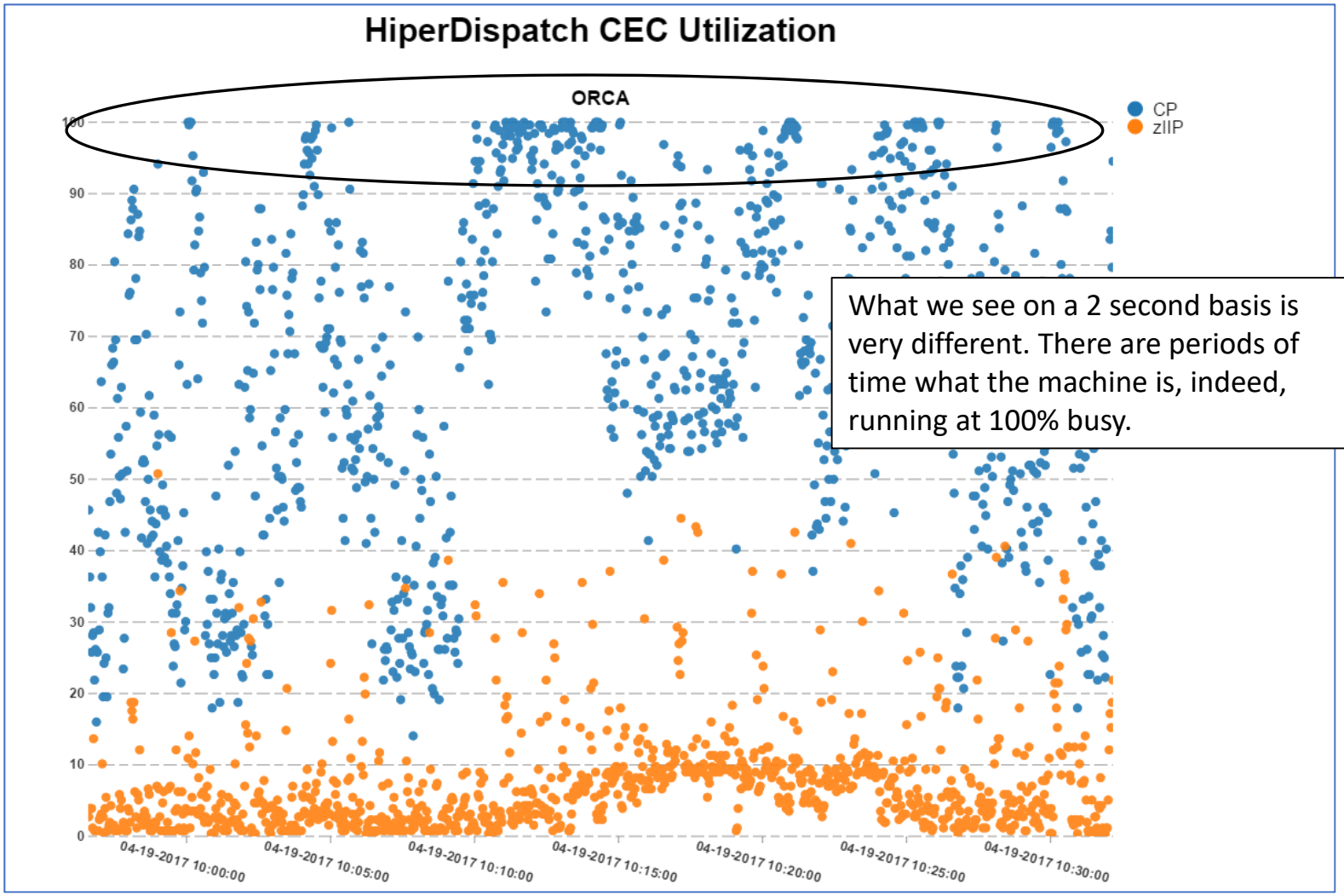
Using the SMF 99.12 record

- The SMF 99.12 record is helpful for answering the following questions:
  - Over time, what is the LPAR configuration and did it change?
  - What is the logical processor pooling for the LPAR?
  - Did the pooling change due to a configuration change or due to capping?
  - What is the parking and un-parking of the logical processors?
  - What is the utilization of the processors?
    - Remember, this is every 2 seconds, so much more granularity than SMF 70 data.
  - What may be inhibiting the un-parking of a processor?
  - What are the effects of capping on the decisions of parking and un-parking processors?

SMF 70 – A look at physical machine utilization




CEC Physical Machine CP Busy%
General Purpose CPU
B43B7, 2828-J02, H13
ORPROD
ORPRODQA
ORTEST
PHYSICAL
Physical Busy Percent
100
90
80
70
60
50
40
30
20
10
0
04-19-2017 08:00
04-19-2017 08:30
04-19-2017 09:00
04-19-2017 09:30
04-19-2017 10:00
04-19-2017 10:30
04-19-2017 11:00
04-19-2017 11:30
04-19-2017 12:00
04-19-2017 12:30
04-19-2017 13:00
04-19-2017 13:30
Users are calling up complaining of poor response times, but the machine does not appear very busy during problem period of time.


HiperDispatch CEC Utilization
ORCA
CP
zIIP
100
90
80
70
60
50
40
30
20
10
0
04-19-2017 10:00:00
04-19-2017 10:05:00
04-19-2017 10:10:00
04-19-2017 10:15:00
04-19-2017 10:20:00
04-19-2017 10:25:00
04-19-2017 10:30:00
What we see on a 2 second basis is very different. There are periods of time what the machine is, indeed, running at 100% busy.

SMF 99.1

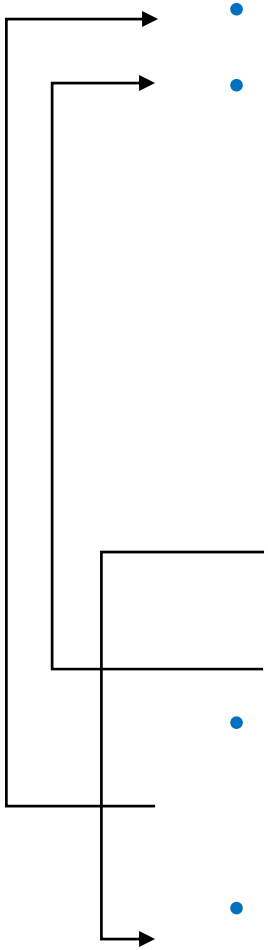
WLM Algorithm Phases

- There are two primary phases of WLM algorithms

- Policy Adjustment (PA)
  - Done approximately every 10 seconds (AKA 'PA interval')
  - Objectives include:
    - Summarize state of system and resources
    - Help work meet goals by setting resource controls
    - Housekeep resource controls that may be out of date

- Resource Adjustment (RA)
  - Done approximately every 2 seconds (AKA 'RA interval')
  - Objectives include:
    - improve efficiency of system resources
    - avoided if at the expense of goals

WLM Policy Adjustment – 'The Loop'



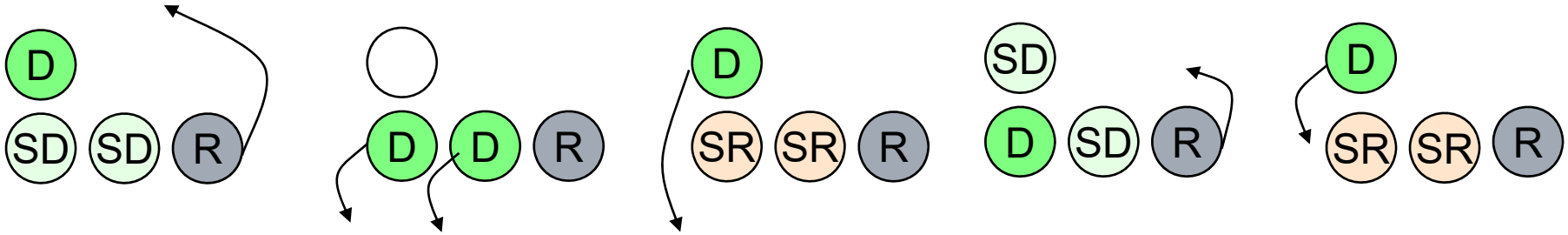
- Summarize data for state of the system and workloads
- Select a receiver period (highest importance missing goal the most)
- Find the receiver's largest bottleneck
  - Determine fix for receiver's bottleneck
    - Determine if needed resources can be gotten from unused resources
    - Find donor(s) of resource that receiver needs
    - Assess effect of reallocating resources from donor(s) to receivers
    - If allocation has both net and receiver value
      Then commit change
      Else don't make change
  - If reallocation was done
    then jump to Exit and allow change to be absorbed
  - If reallocation was not done
    then try to fix receiver's next largest bottleneck
- If cannot help receiver
  then look for next receiver (highest importance missing goal the most)
- Exit
  - Housekeep current set of controls

Receivers and Donors

- Receiver
  - Service class period to potentially 'receive' resources
  - WLM will help only one receiver during each policy adjustment interval
    - Goal Receiver - Period with goal that needs help
    - Resource Receiver - Period to give the resources to in order to help the goal receiver
    - Secondary Receiver - Period helped indirectly due to an action to help the goal receiver
- Donor
  - Service class period to potentially 'donate' resources to help receiver
  - WLM may take from multiple donors during each policy adjustment interval
    - Goal Donor - Period whose goals may be impacted by resource donation
    - Resource Donor - Period to donate resources
    - Secondary Donor - Period that donates indirectly when receiver is helped

Policy Adjustment Actions - CPU

- Dispatching priority adjustments
  - Objective: Increase Receiver's CPU using, or decrease Receiver's CPU delay
  - Interesting concepts:
    - Wait-to-Using ratio - ratio of CPU delay samples to CPU using samples (change in ratio used to determine change in CPU delay)
    - Maximum demand
      - Theoretical maximum percentage of total processor time a period can consume if it had no CPU delay
    - Achievable maximum demand
      - Percentage of total processor time a service period is projected to consume, taking into account demand of all higher work
  - Some possible actions



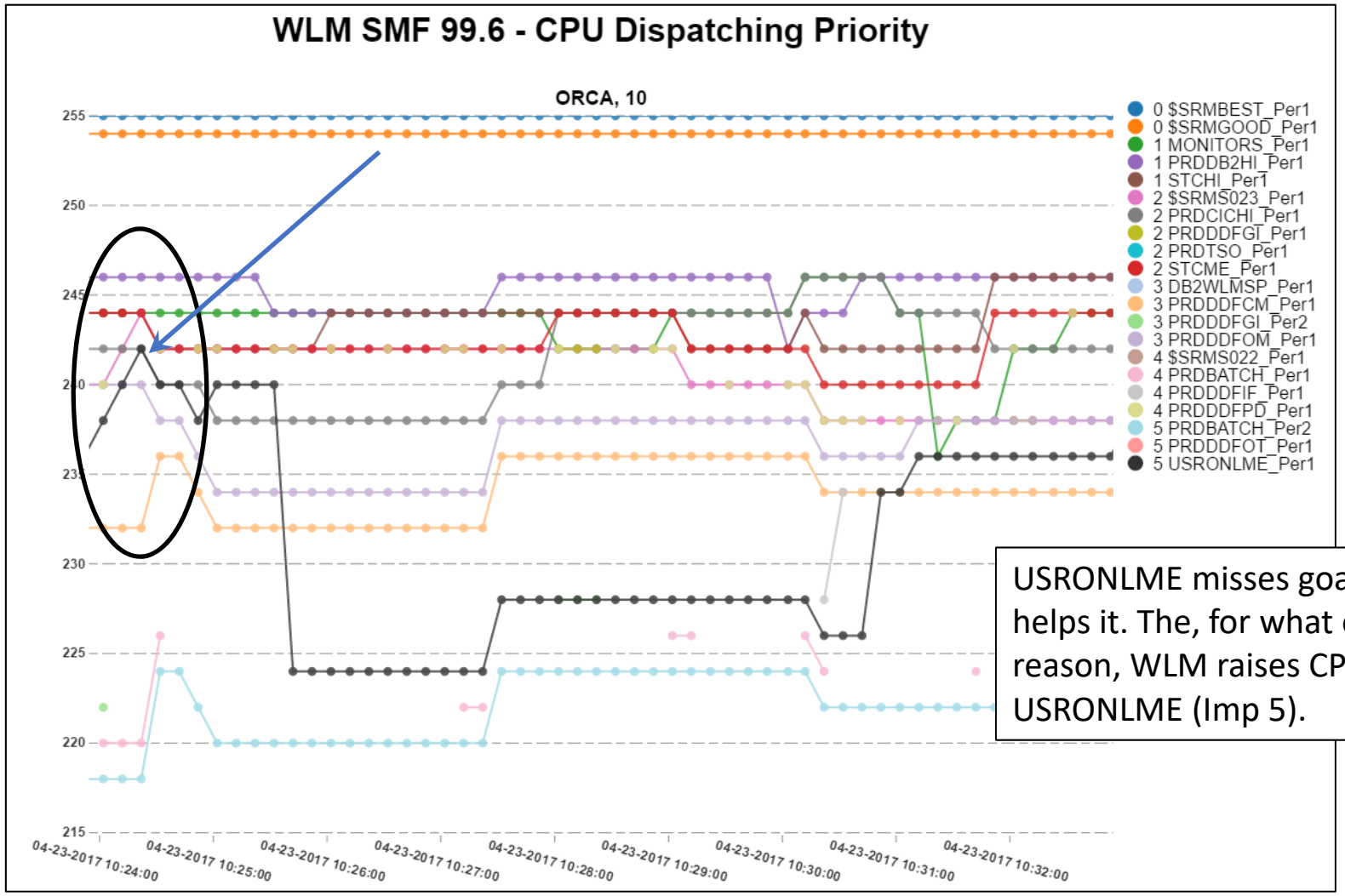
PA Loop: Receiver Value Check

- Receiver Value
  - Receiver helped only if there is projected to be sufficient *receiver value*
    - Designed to reject 'small or marginal improvements'
    - Allows WLM to get on to addressing larger problems for other periods
  - Minimum projected improvement to make change worth the effort
    - Projected PI improvement
    - or projected minimum group service increase
    - or some other projected minimum criteria
- Guideline:
  - Projected PI improvement is the larger of (10% of the PI change to meet goal) or (0.05)
  - Or Reduction in delay samples is at least half of the largest delay
- Example:
  - PRODTSO period 1 PI = 3.5
  - WLM algorithms suggest improvements can bring PI to 3.46
  - Don't take action

PA Loop: Net Value Check

- Receiver is only helped by a specific donor if there is projected to be sufficient *net value*
  - Designed to reject changes that will harm the donor more than the projected improvement to the receiver
  - Allows WLM to assess taking from other donors
- All external service policy specifications are considered for both primary and secondary donors
    - goals
    - importance
    - resource group minimums and maximums
- Example
  - PRODBAT PI = 4.0
  - WLM algorithms suggest improvements can bring PI to 3.0
  - Change hurts donor more then helps receiver

SMF 99.6 CPU Dispatching Priority – Every 10 Seconds

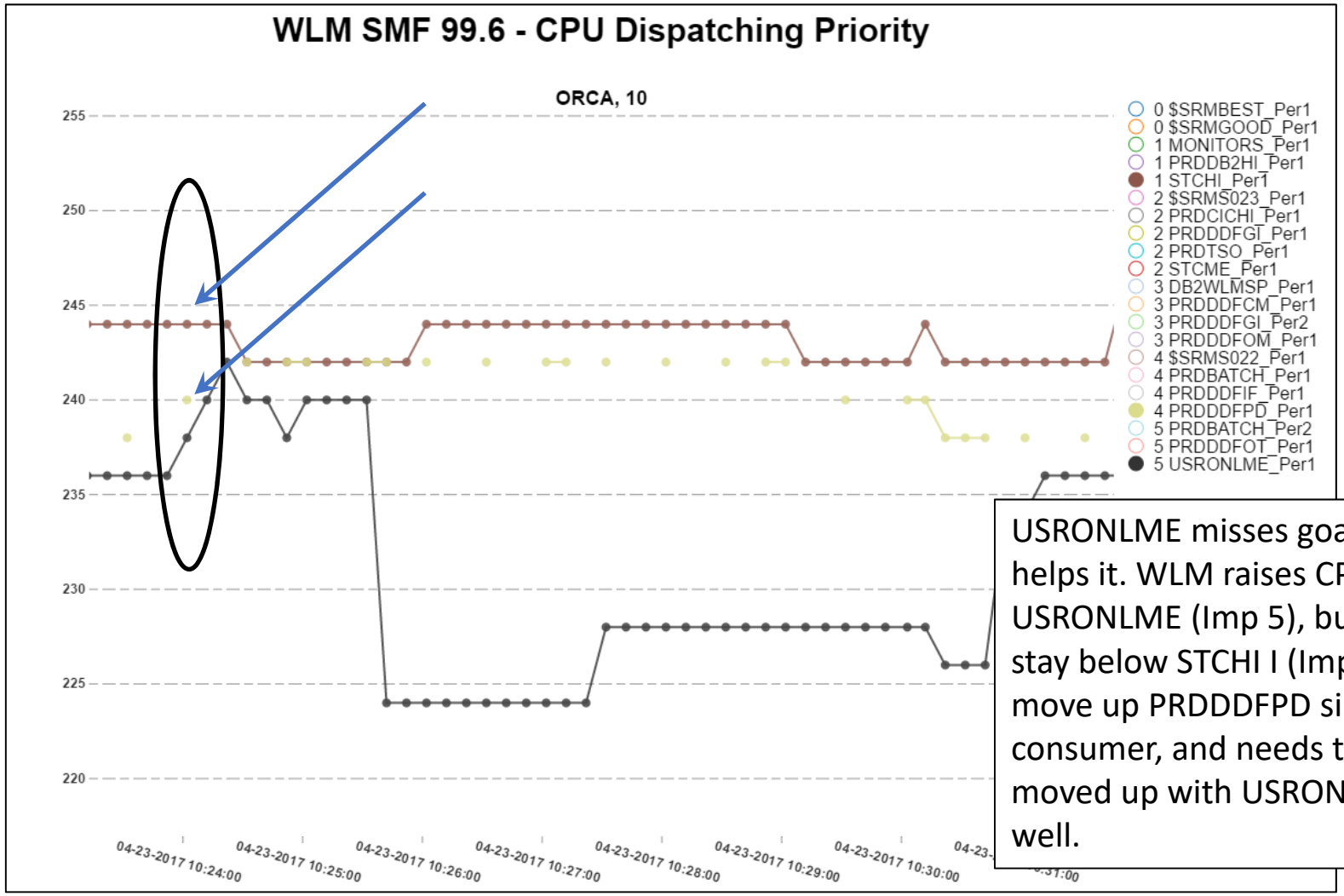


USRONLME misses goal, so WLM helps it. The, for what ever reason, WLM raises CPU DP of USRONLME (Imp 5).

SMF 99.1
- Example of WLM Actions Trace

SMFDateTime	PA Inteval	RA Interval	Trace Code	Code	Job	Local PI	Sysplex PI	Service Class	Period
4/23/17 10:24:03 AM	175	124	270	PA_REC_CAND		131	131	USRONLME	1
4/23/17 10:24:03 AM	175	124	975	PA_SDO_DONFAIL_SPC		110	110	PRDDDFGI	1
4/23/17 10:24:03 AM	175	124	975	PA_SDO_DONFAIL_SPC		70	70	PRDDDFOM	1
4/23/17 10:24:03 AM	175	124	975	PA_SDO_DONFAIL_SPC		27	27	STCME	1
4/23/17 10:24:03 AM	175	124	975	PA_SDO_DONFAIL_SPC		110	110	PRDDDFPD	1
4/23/17 10:24:03 AM	175	124	308	PA_DONOR_PERIOD		40	40	STCHI	1
4/23/17 10:24:03 AM	175	124	880	PA_PRO_RDON_CAND		40	40	STCHI	1
4/23/17 10:24:03 AM	175	124	620	PA_PMUO_REC		131	131	USRONLME	1
4/23/17 10:24:03 AM	175	124	620	PA_PMUO_REC		131	131	USRONLME	1
4/23/17 10:24:03 AM	175	124	620	PA_PMUO_REC		131	131	USRONLME	1
4/23/17 10:24:03 AM	175	124	651	PA_PMU_SPC_NXT_DP		110	110	PRDDDFPD	1
4/23/17 10:24:03 AM	175	124	940	PA_PRO_UNC_DON		40	40	STCHI	1
4/23/17 10:24:03 AM	175	124	940	PA_PRO_UNC_DON		40	40	STCHI	1
4/23/17 10:24:03 AM	175	124	940	PA_PRO_UNC_DON		40	40	STCHI	1
4/23/17 10:24:03 AM	175	124	740	PA_PRO_INCP_DON		110	110	PRDDDFPD	1
4/23/17 10:24:03 AM	175	124	740	PA_PRO_INCP_DON		110	110	PRDDDFPD	1
4/23/17 10:24:03 AM	175	124	740	PA_PRO_INCP_DON		110	110	PRDDDFPD	1
4/23/17 10:24:03 AM	175	124	780	PA_PRO_INCP_SC		110	110	PRDDDFPD	1
4/23/17 10:24:03 AM	175	124	780	PA_PRO_INCP_SC		110	110	PRDDDFPD	1
4/23/17 10:24:03 AM	175	124	780	PA_PRO_INCP_SC		110	110	PRDDDFPD	1
4/23/17 10:24:03 AM	175	124	750	PA_PRO_INCP_REC		113	113	USRONLME	1
4/23/17 10:24:03 AM	175	124	750	PA_PRO_INCP_REC		113	113	USRONLME	1
4/23/17 10:24:03 AM	175	124	750	PA_PRO_INCP_REC		113	113	USRONLME	1

SMF 99.6 CPU Dispatching Priority – Every 10 Seconds



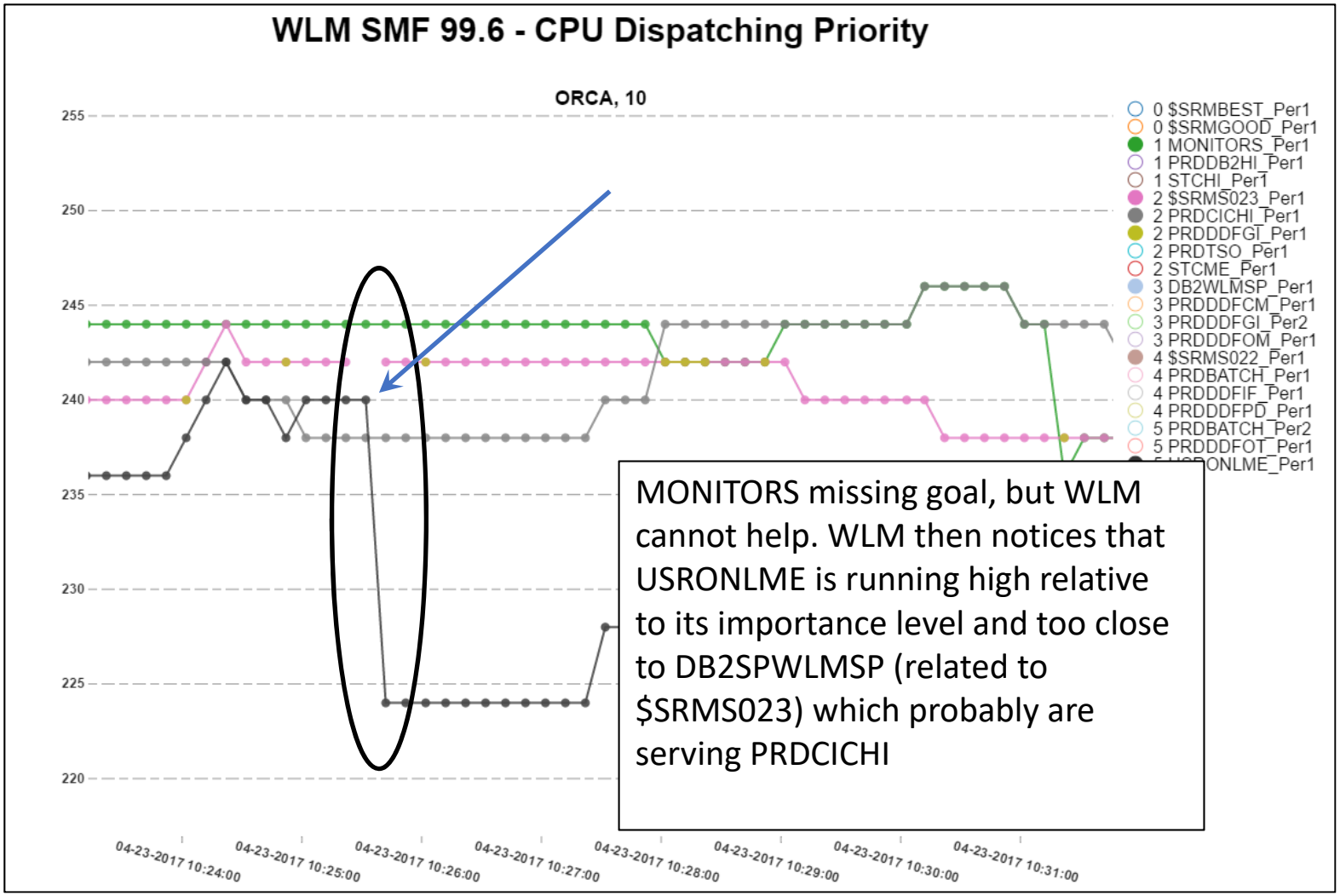
USRONLME misses goal, so WLM helps it. WLM raises CPU DP of USRONLME (Imp 5), but also to stay below STCHI I (Imp 2). Also move up PRDDDFPD since small consumer, and needs to be moved up with USRONLMD as well.

SMF 99.1
- Example of WLM Actions Trace

SMFDateTime	PA Inteval	RA Interval	Trace Code	Code	Local PI	Sysplex PI	Service Class	Period
4/23/2017 10:25:43	185	174	270	PA_REC_CAND	128	128	MONITORS	1
4/23/2017 10:25:43	185	174	975	PA_SDO_DONFAIL_SPC	64	64	PRDDB2HI	1
4/23/2017 10:25:43	185	174	850	PA_PRO_RDON_CAND	128	128	MONITORS	1
4/23/2017 10:25:43	185	174	9348	PA_LMP_SKIPPED	128	128	MONITORS	1
4/23/2017 10:25:43	185	174	9301	PA_PPP_POT_REC	3000	3000	DB2WLMSP	1
4/23/2017 10:25:43	185	174	9301	PA_PPP_POT_REC	3000	3000	DB2WLMSP	1
4/23/2017 10:25:43	185	174	9301	PA_PPP_POT_REC	3000	3000	DB2WLMSP	1
4/23/2017 10:25:43	185	174	9300	PA_PPP_DECP_DON	29	29	USRONLME	1
4/23/2017 10:25:43	185	174	9300	PA_PPP_DECP_DON	29	29	USRONLME	1
4/23/2017 10:25:43	185	174	9300	PA_PPP_DECP_DON	29	29	USRONLME	1

- Select receiver - MONITORS
- Select donor fail since a small CPU consumer – PRDDB2HI
- Keep looking for a resource donor
- Attempt LPAR weight management, but failed since some condition was not met
- Severe delays noted for DB2WLMSP
- Detected and lower priority of USRONLME since low importance and severe delays

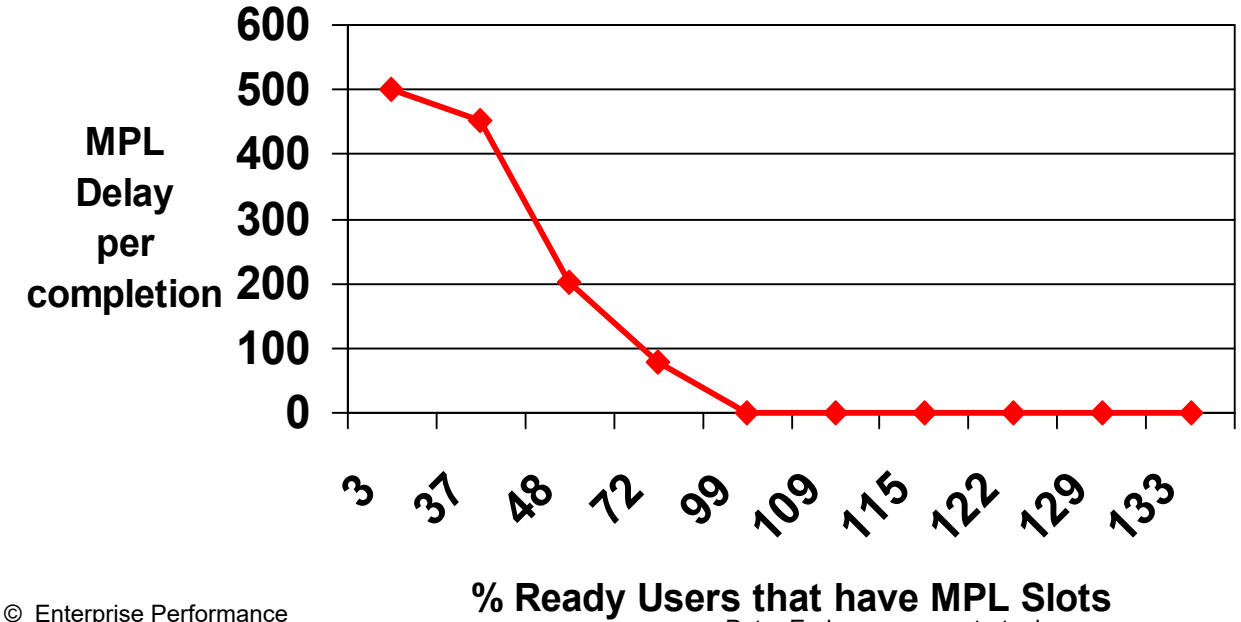
SMF 99.6 CPU Dispatching Priority – Every 10 Seconds



SMF 99.3

Plots

- Plots used to track how well work is being processed
- Some of the plots include
  - System Paging Delay Plot
  - Period MPL Delay Plot
  - Period Ready User Average Plot
  - Queue Delay Plot
  - Period Paging Rate Plot
  - Proportionate Aggregate Speed Plot
  - Address Space Paging Plot



MPL Plot Example:

- shows how response time may improve by increasing MPL slots
- shows how response time may degrade by reducing MPL slots

SMF 99.5

Individual Address Space Monitoring

- The granularity of most WLM controls are at the service class period level
  - But some address spaces may have individual storage policies
  - That is, an address space may have a storage policy separate and different from all other address spaces in its own period
- For an address space to have an individual address space storage policy it must be monitored separately for a period of time by WLM
  - WLM only monitors address space if finds 'interesting'
    - Example: if an address space is using a lot of storage
- Address spaces eligible for monitoring and individual storage policies include
  - Address spaces assigned a velocity goal
  - Address space assigned a discretionary goal
  - Address space is found to be a server
    - Example: CICS transaction management turned on, so a CICS address space is eligible
  - Address space is assigned a response time goal of greater than 20 seconds

SMF 99.4

Policy Adjustment Actions - I/O Priority

- I/O Priority
  - Set similar to the way CPU dispatch priority is set
  - Donor must be competing with receiver for at least some of devices or action will have no effect
  - Device Clustering
    - WLM needs to be aware of periods competing for same devices
    - Device clustering is used to determine this relationship
      - Each class associated with a single cluster

Service Class	Dev 200	Dev 201	Dev 202	Dev 500	Dev 501	Dev 502	Dev 503
Class 1	100	150	150	0	0	0	0
Class 2	0	90	100	0	0	0	0
Class 3	0	100	100	5	0	0	0
Class 4	0	0	0	100	100	100	100
Class 5	0	0	0	0	150	0	150

Device Clusters:
- Cluster 1 = 1,2,3
- Cluster 2 = 4,5

WLM Measurement Reports Processing/Discussion Offer !!!

- Special Reports Offer!
  - See your Coupling Facility records in chart and table format
  - Please contact me, Peter Enrico for instructions for sending raw SMF data
    - Send an email to peter.enrico@epstrategies.com
  - Deliverable: Dozens of WLM based reports (charts and tables)
    - Period setup
    - Performance Index analysis
    - Velocity goal analysis
    - Response time goal analysis
    - Multi-period analysis
    - WLM workload analysis – batch, started tasks, DDF, etc.
    - WLM resource analysis (CPU / storage / I/O)
    - And much more!
  - One-on-one phone call to explain your measurements

www.pivotor.com