WLM and SMF 99.2 – Period Measurements Deeper Dive

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:

Enterprise Performance Strategies, Inc. presentation materials contain trademarks and registered trademarks of several companies.

The following are trademarks of Enterprise Performance Strategies, Inc.: **Health Check®, Reductions®, Pivotor®**

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries: IBM®, z/OS®, zSeries® WebSphere®, CICS®, DB2®, S390®, WebSphere Application Server®, and many others.

Other trademarks and registered trademarks may exist in this presentation

Abstract

- **WLM and SMF 99.2 – System Measurements Deeper Dive**

- The SMF 99.2 record has been around since the first availability of WLM. It contains service class period measurements by the WLM policy and resource adjustment algorithms.

- During this webinar, Peter Enrico will walk through the measurements of the SMF 99.2 record and show how these measurements can be used. During this webinar, the attendees will also gain insights into the WLM policy adjustment algorithms and how they work.

EPS: We do z/OS performance…

- **Pivotor** – z/OS performance reporting and analysis software and services
  - Not just SMF reporting, but analysis-based reporting based on expertise
  - www.pivotor.com
- **Education and instruction**
  - We teach our z/OS performance workshops all over the world
  - Want a workshop in your area? Just contact me.
- **z/OS Performance War Rooms**
  - Intense, concentrated, and highly productive on-site performance group discussions, analysis and education
  - Amazing feedback from dozens of past clients
- **MSU Reduction Exercises**
  - The goal is to reduce the MSU consumption of your applications and environment
- **Information**
  - We present around the world and participate in online forums
  - https://www.pivotor.com/content.html
    https://www.pivotor.com/webinar.html

z/OS Performance workshops available

During these workshops you will be analyzing your own data!

- WLM Performance and Re-evaluating Goals
  - February 19-23, 2024
- Parallel Sysplex and z/OS Performance Tuning
  - August 20-21, 2024
- Essential z/OS Performance Tuning
  - October 7-11, 2024
- Also… please make sure you are signed up for our free monthly z/OS educational webinars! (email contact@epstrategies.com)

Like what you see?

- Free z/OS Performance Educational webinars!
  - The titles for our Summer / Fall 2024 webinars are as follows:
    - ✓ *What a z/OS Guy Learned About AWS in 10 Years*
    - ✓ *Advantages of Multiple Period Service Classes*
    - ✓ *Understanding z/OS Connect Measurements*
    - ✓ *WLM and SMF 99.1 – System Measurements Deeper Dive*
    - *WLM and SMF 99.2 – Service Class Period Measurements Deeper Dive*
    - *Optimizing Performance at the Speed of Light: Why I/O Avoidance is Even More Important Today*
    - *Understanding MVS Busy % versus LPAR Busy % versus Physical Busy %*
    - *Rethinking IBM Software Cost Management Under Tailored Fit Pricing*
    - *Understanding Page Faults and Their Influence on Uncaptured Time*
    - *Response Time Goals: Average or Percentiles?*
    - *Understanding and Using Enclave*

- If you want a free cursory review of your environment, let us know!
  - We're always happy to process a day's worth of data and show you the results
  - See also: http://pivotor.com/cursoryReview.html

Like what you see?

- The z/OS Performance Graphs you see here come from Pivotor

- If you don't see them in your performance reporting tool, or you just want a free cursory performance review of your environment, let us know!
  - We're always happy to process a day's worth of data and show you the results
  - See also: http://pivotor.com/cursoryReview.html

- We also have a free Pivotor offering available as well
  - 1 System, SMF 70-72 only, 7 Day retention
  - That still encompasses over 100 reports!

History of the SMF 99 Record

- When WLM was first being developed, lots of trace records were embedded in the code so the WLM algorithm developers could understand and debug WLM actions
  - Before WLM was released, all these trace record were turned off

- When WLM was shipped in 1994, it was realized that that there would be a need for these trace records to assist in understanding and debug of customer situations
  - So, at the last minute, via an APAR, the SMF 99 records were coded and introduced

  - Originally only consisted of SMF 99 subtypes 1 through 6
  - Later releases of MVS introduced additional subtypes

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)
  - *And very voluminous!*
- 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 99.2 Record

Note: This presentation is meant to give a taste of the contents of the SMF 99.2 record. In no way is it a complete lesson, and in no way are all fields and value understood

Key Lesson about the SMF 99.2 Records

- The original intention of the SMF 99 records was to help IBM debug WLM algorithm decisions
  - These records are cut every 10 seconds
- While practical use has been found by regularly processing and analyzing certain subtypes:
  - Such as the SMF 99.6, SMF 99.12, SMF 99.14
- The other SMF 99 subtypes, such as SMF 99.1 and SMF 99.2…
  - Are cryptic, not well documented
  - Contain values that are only meaningful to WLM algorithms
  - Are not very useful for our day-to-day use by general performance analyst
- At EPS, we use the SMF 99.1 and SMF 99.2 by searching for unusual patterns during periods of time customers call us asking to analyze a situation

Overview of SMF 99, subtype 2 record

- Subtype 2
  - Contains service class level data used as input to make WLM algorithm decisions.

- Triplets include:
  - Class data section
  - Period data section
  - Cross memory entry section
  - Server data entry section
  - Server sample data entry section
  - Queue server data entry section
  - Remote queue server data entry section
  - Address space expanded storage access policy section

SMF 99 – PA and RA Decisions

- 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

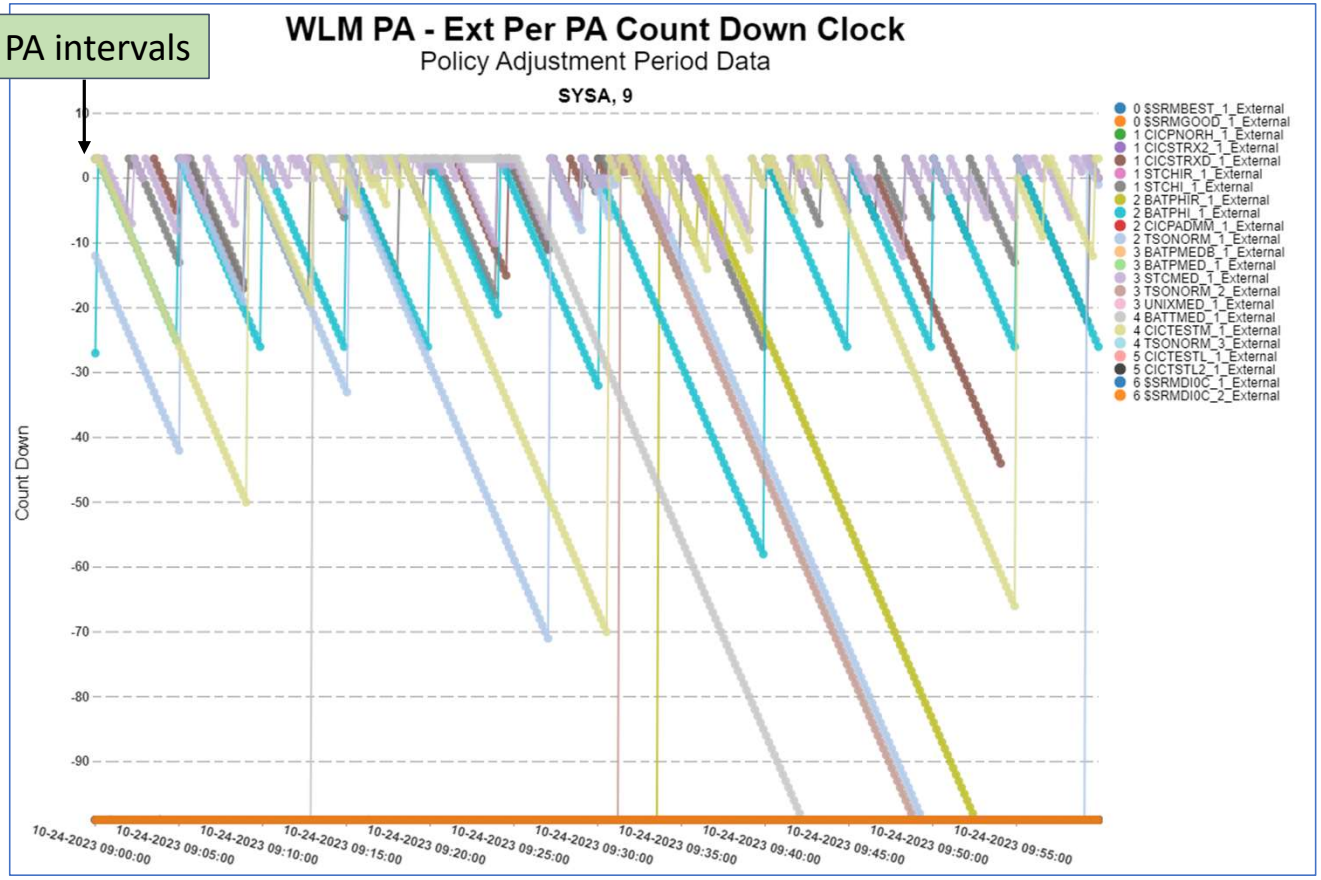
WLM Policy Adjustment – 'The Loop'

- 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

Service Class and Period Data Sections

- Contains service class period data used as input to WLM decisions
  - Data is cut every 10 seconds, so much more granular than SMF 72.3 RMF/CMF data
  - *HOWEVER: Data will not match SMF 72.3 or the SMF 99.6*
    - *BECAUSE: data is based on histories, and not what just happened in last 10 seconds*
    - *BEWARE!*
- Data can be logically broken up into the following areas of interest
  - PA clock counts and history rows used
  - Period CPU data (based on history)
  - Response time statistics data
    - Response time goal periods
    - Non-response time goal periods
  - PA samples details
  - PA storage measurements
  - Misc processor measurements
  - PA miscellaneous period stats

Does WLM stop trying? No!



If WLM cannot meet the goal for work, it never stops ‘trying’.

There are cases when WLM may ‘ignore’ work for some period-of-time.

Usually, the minimum countdown is 3 PA intervals (which equates to 30 seconds).

This chart shows the countdown chart for each period.

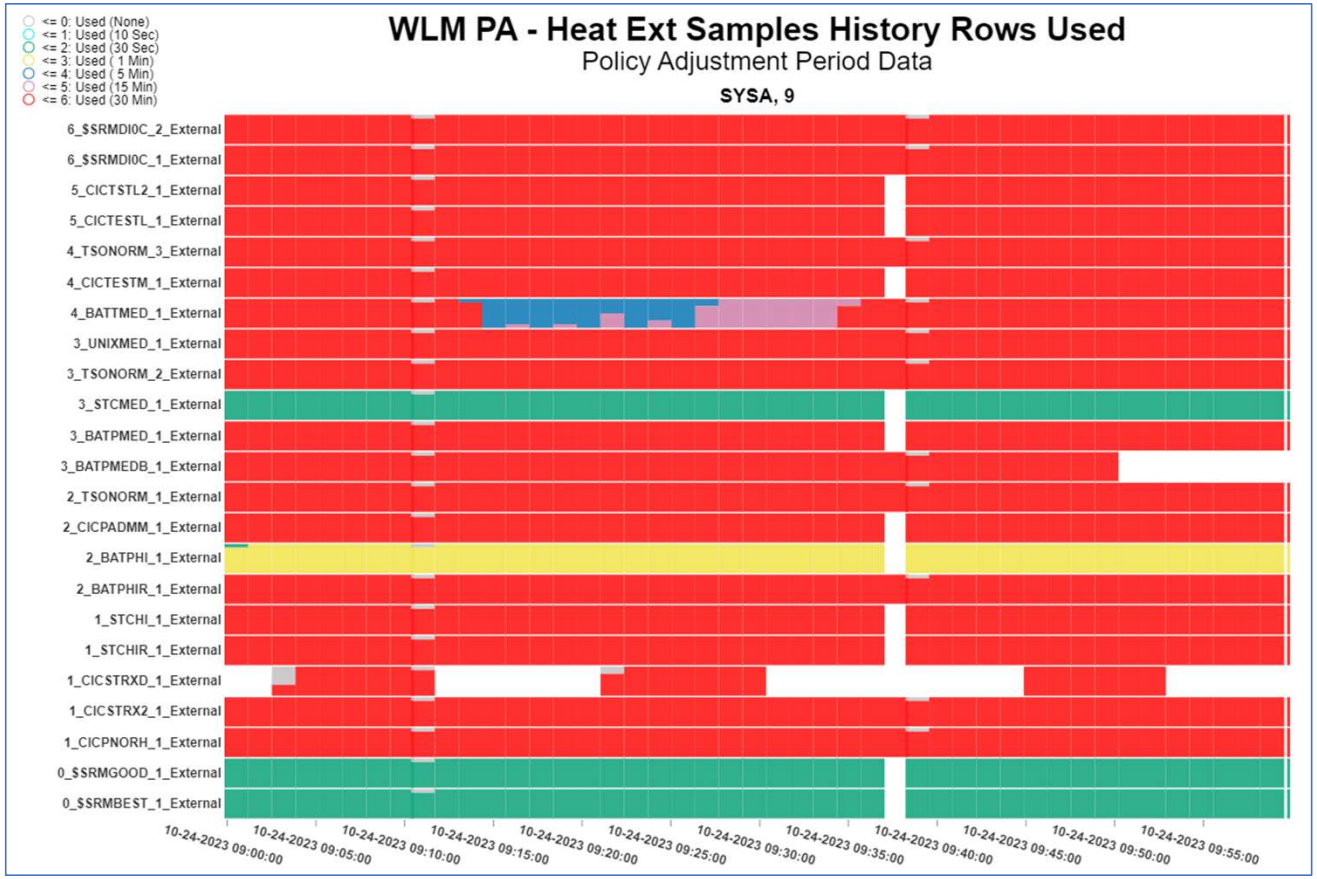
WLM maintains histories

- Histories used to examine data over a period of time
  - Allows WLM to have a controlled way to go back in time for enough representative data points for accurate decisions
  - For example, a history *may* have the following structure

Last 10 Seconds
Last 30 seconds
Last 60 seconds
Last 5 minutes
Last 15 minutes
Previous 15 minutes

  - New data put in row one
  - After some number of intervals, the data is rolled forward
  - Each successive row represents data that was collected further in the past and over a longer period of time

Examples of Data Histories Usage



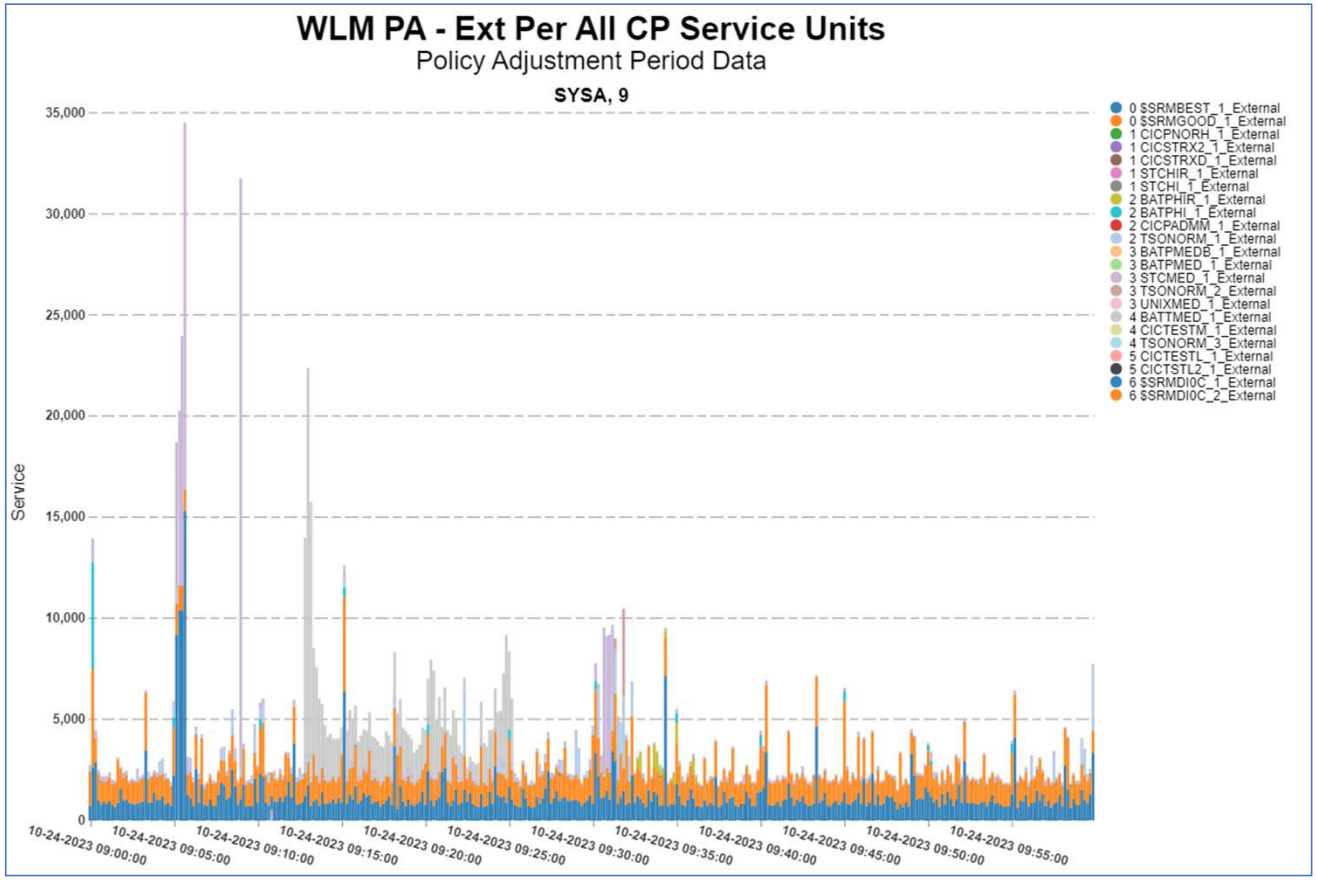
This heat chart highlights how many 'rows' of data history are used when making decisions.

Periods with lots of work and accumulating lots of samples will use fewer rows of data history.

This chart can provide insights:

- As to why the SMF 99.2 CPU usage data does not match the SMF 99.6 CPU usage data
- If periods could benefit from consolidation

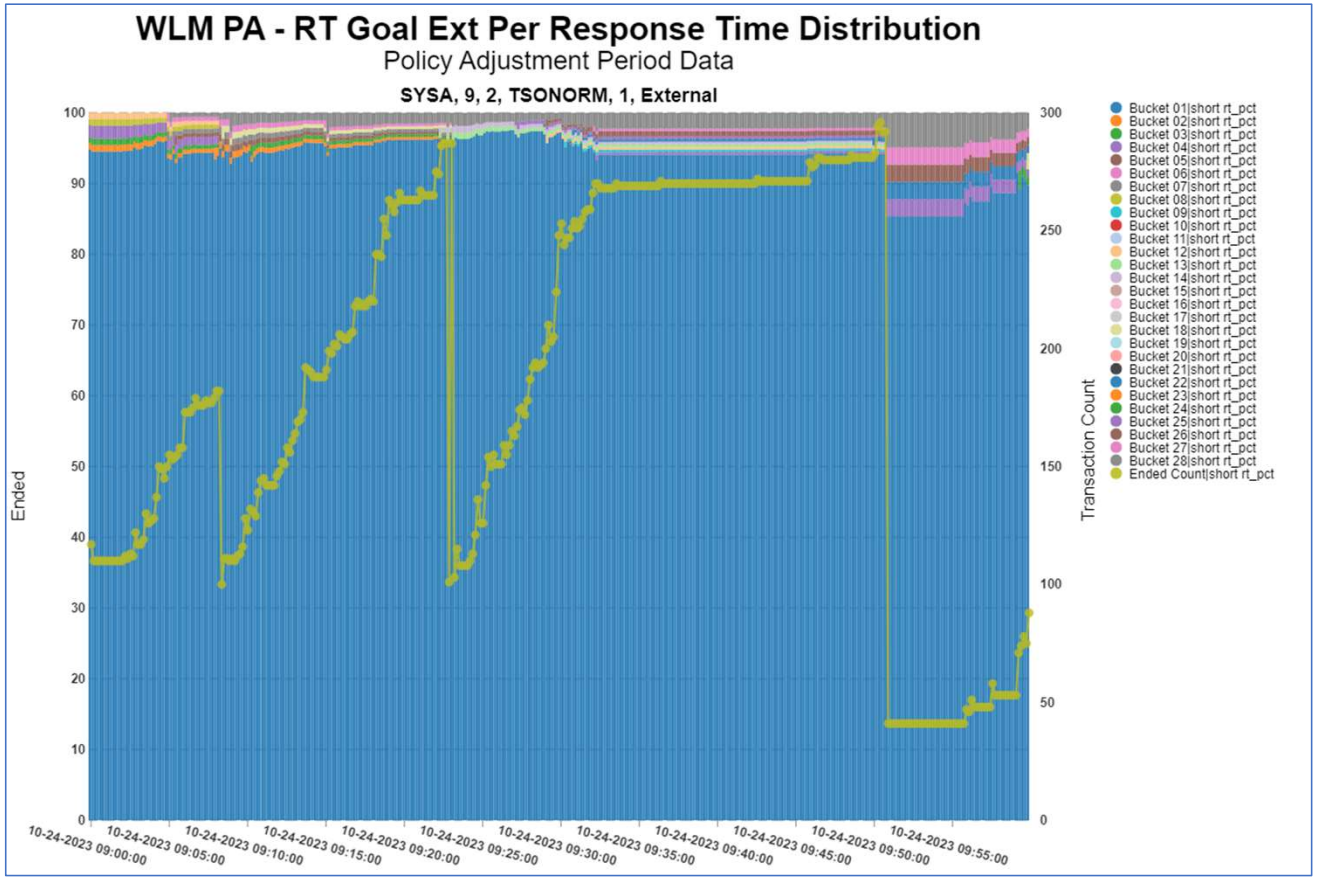
Example of CPU Service Usage



This charge shows the amount of CP CPU service consumed broken down by service class period.

This is an example of the CP CPU usage used by the WLM algorithms based on histories.

Example of Response Time Distributions

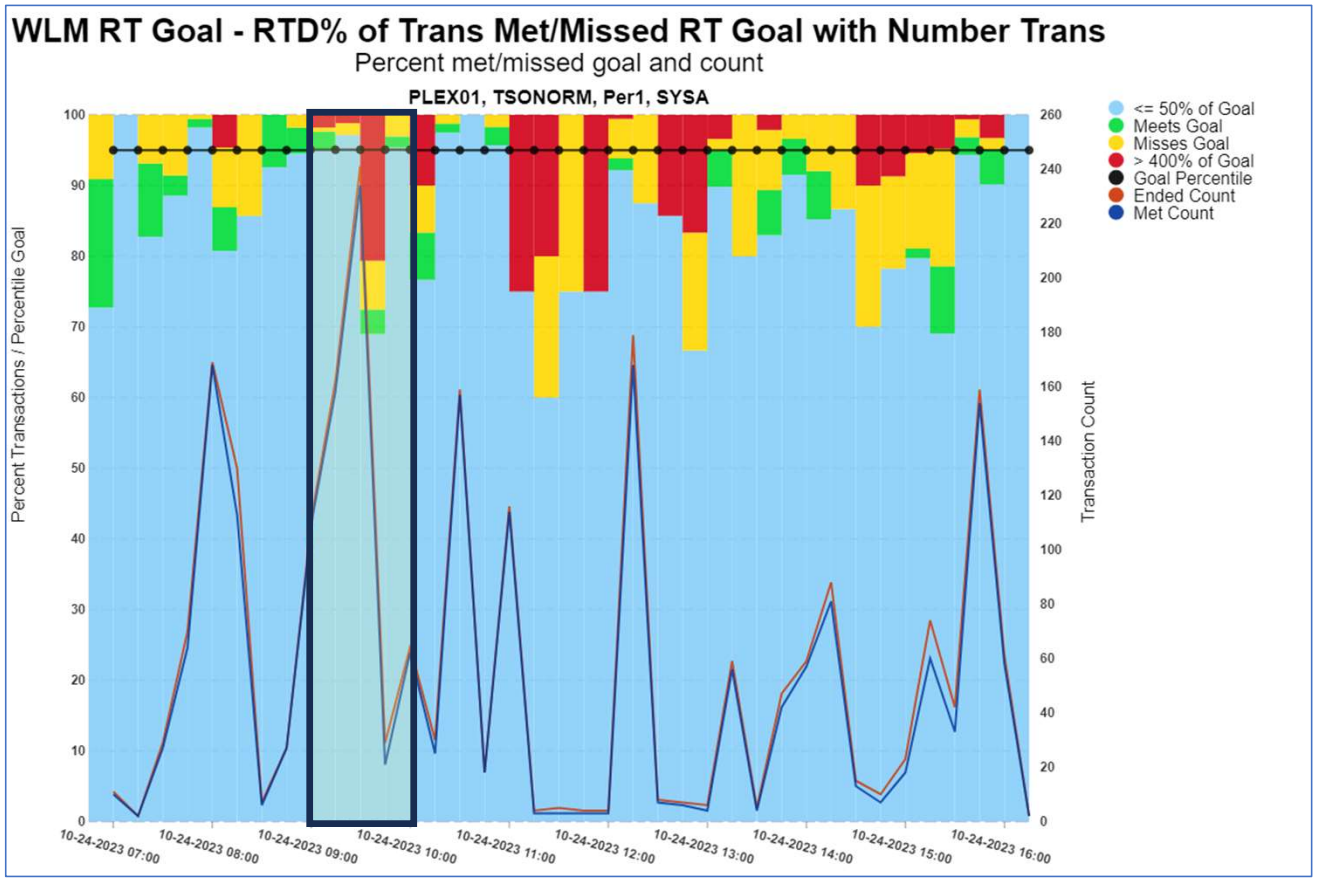


This shows the response time distribution for TSONORM periods 1 with a 10 second measurement interval.

Notice there are 28 RTD buckets rather than the SMF 72.3 14 RTD buckets.

This distribution is based on 6 rows of history (or 30 minutes).

Example of Response Time Distributions



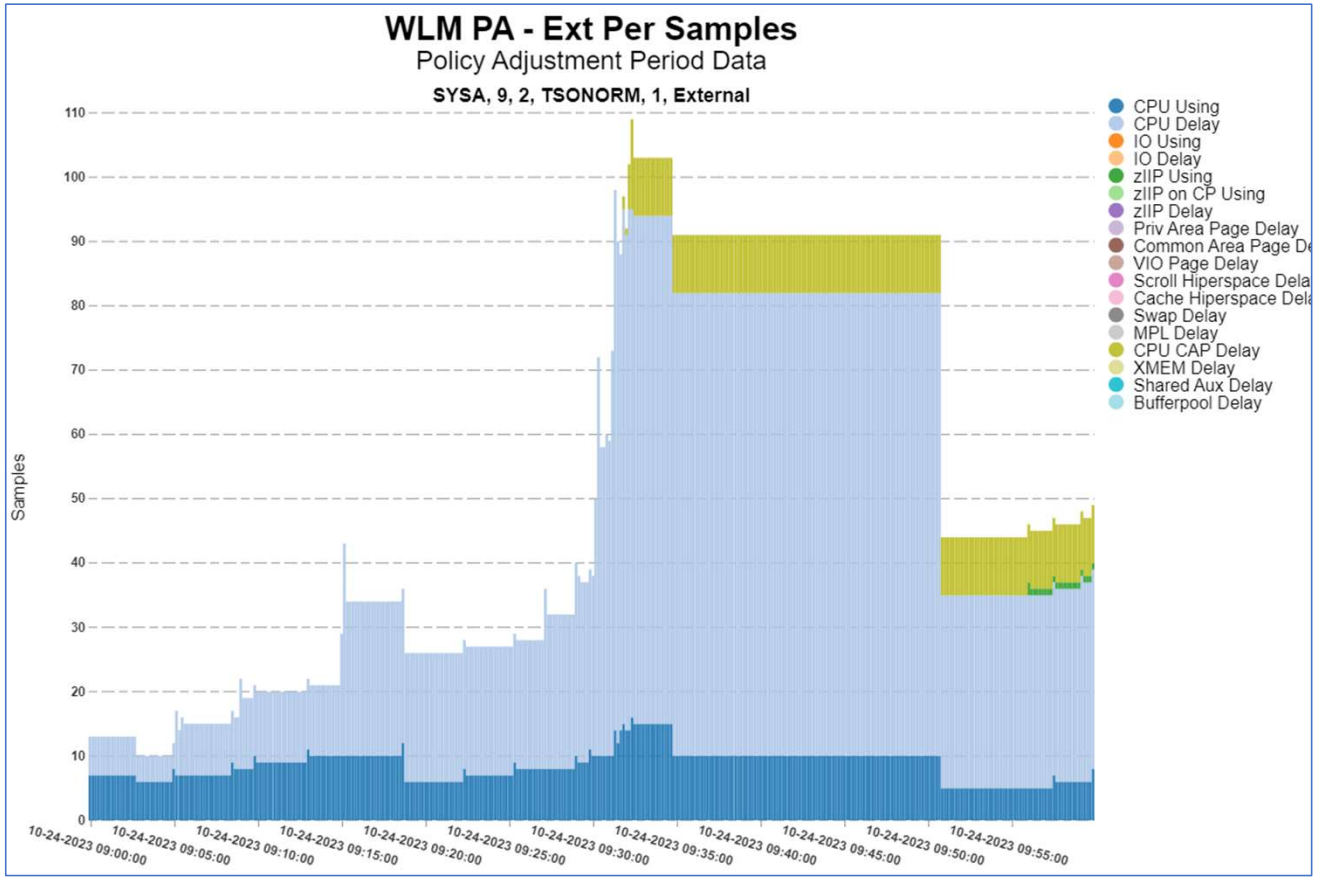
This chart is an example of a response time distribution based on the SMF 72.3 RMF/CMF records cut every 15 minutes.

The SMF 72.3 contains the RTD buckets broken down into 14 buckets. This report consolidates the distribution into 4 buckets.

The highlighted box is the 9:00am hour.

This distribution is based on what happened each 15-minute interval.

Example of Samples

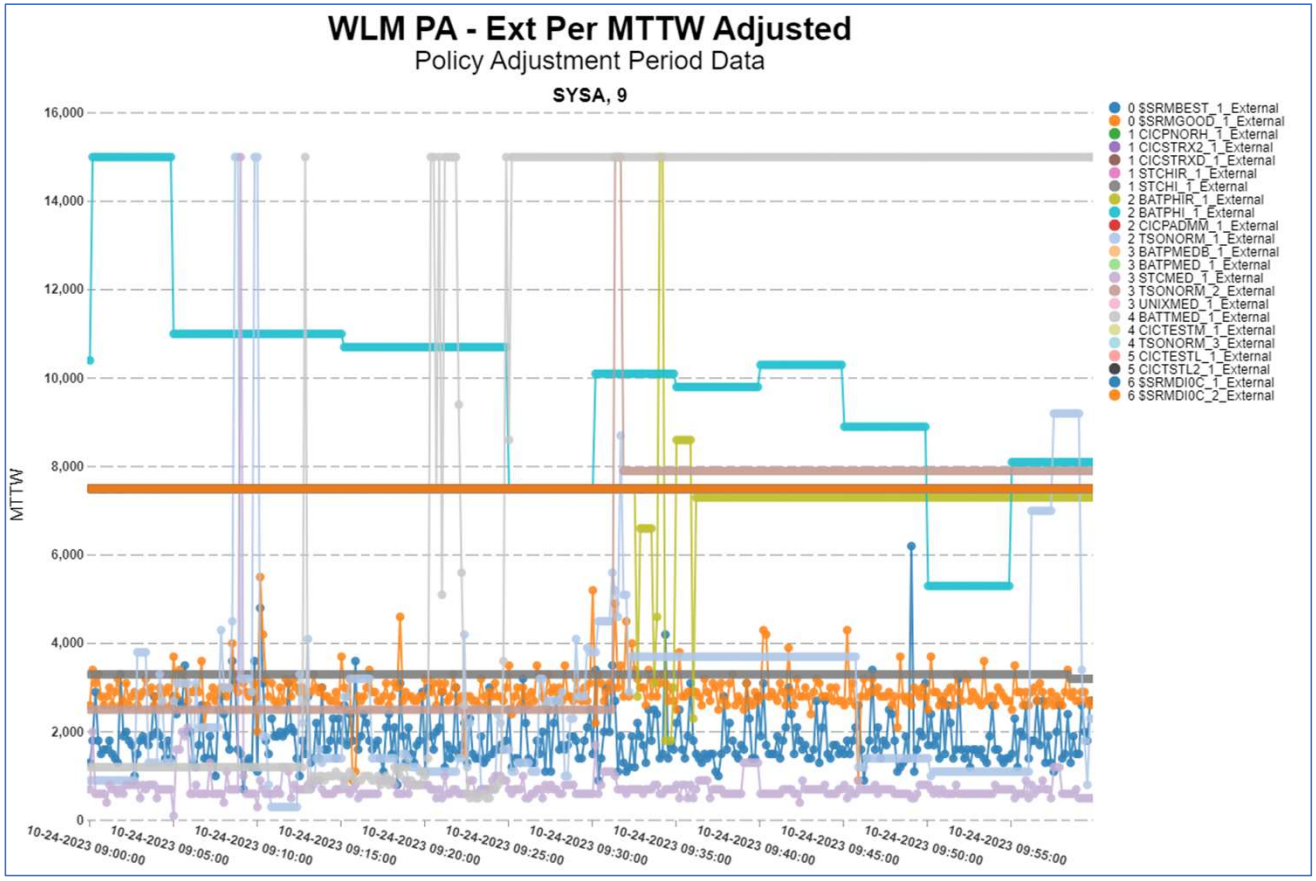


This chart shows an example of WLM using and delay samples every 10 seconds for TSONORM periods 1.

These numbers may not match the SMF 72.3 data because these values are influenced by the number of rows of history used to accumulate the samples.

We can still get an idea of bursts of activity.

Example of Samples



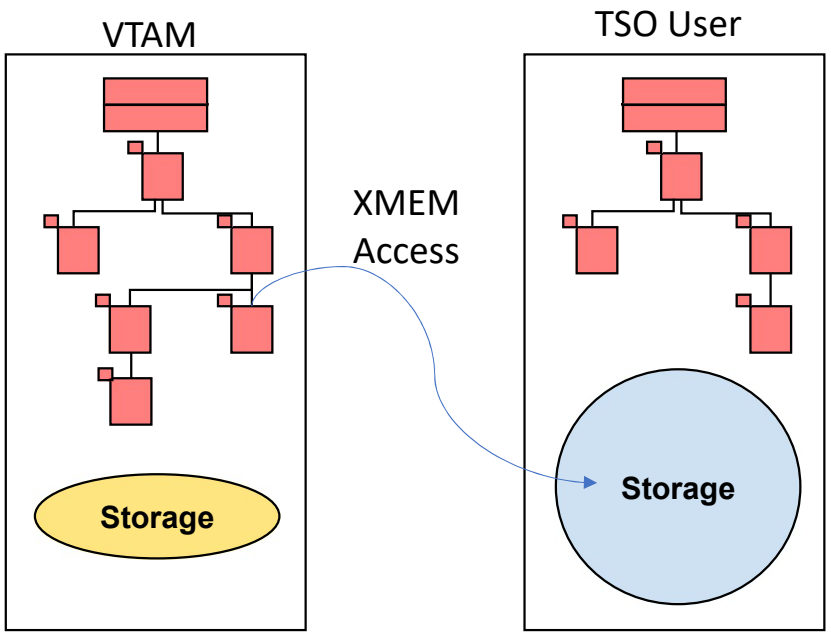
This chart shows an example of the Mean-Time-To-Wait (MTTW) for the various WLM service class periods.

Low MTTWs are usually indicative of a I/O intensive workload

High MTTWs are usually indicative of CPU intensive workloads

Cross Memory Data Entry Section

- Contains the names of address spaces causing cross memory delays
  - As well as the number of samples of cross memory delay



For example:

- Say VTAM does a cross-memory access to a TSO user, and that TSO user takes a page fault.
- VTAM would be delayed because the TSO user took a page fault.
- VTAM incur cross-memory delays
- This section of the record contains the name of the TSO user and the number of delay samples it caused
- To help VTAM, WLM needs to storage isolate the TSO user

Server Entry Data Section

- Contains data to help gain insights into the server / served relationship
  - Useful to gain insight into the server topologies
  - Would be fantastic if the server address space names were included, but only the period data is given

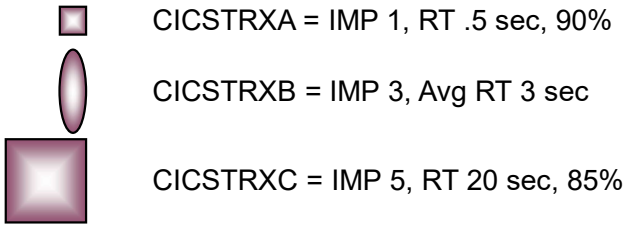
- Data includes:
  - Served service class names
  - Server service class names
  - Server and served counts

Service Class	Period	Goal Type	Server/Served Class	Period	Observations	Class Type
CICSTRXA	1	short_rt_avg	$SRMS02D	1	0	INT
CICSTRXB	1	short_rt_pct	$SRMS02E	1	0	INT
CICSTRXB	1	short_rt_pct	$SRMS02C	1	134	INT
$SRMS02C	1	internal	CICSTRXB	1	134	EXT
$SRMS02D	1	internal	CICSTRXA	1	0	EXT
$SRMS02E	1	internal	CICSTRXB	1	0	EXT

- Note: Topologies are re-evaluated regularly, but an internal period can be inactive for up to X (30?) minutes before it is deleted

Example of Server Topology

- WLM sampler samples PBs, tallies Report and Notifies and gains an understanding
  - Which transactions are being processed by which regions
  - Frequency that each transaction was processed by each region

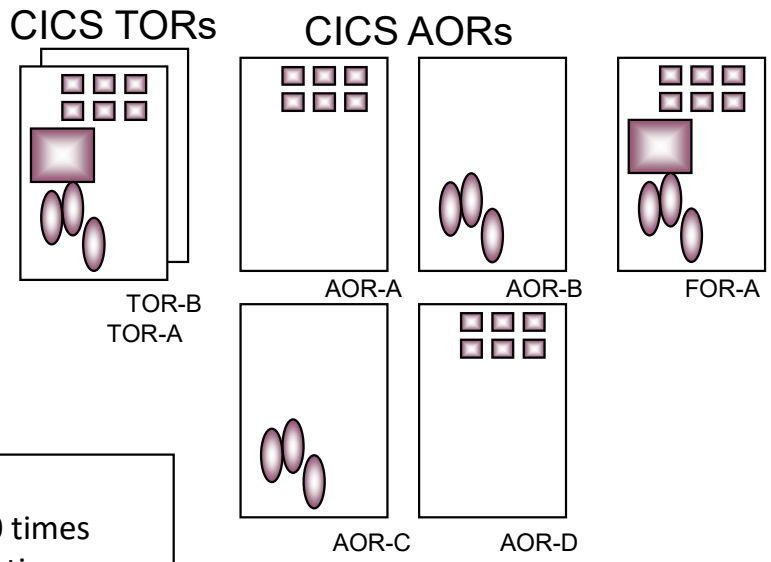


Service Class **Goal**

CICSTRXA = IMP 1, RT .5 sec, 90%

CICSTRXB = IMP 3, Avg RT 3 sec

CICSTRXC = IMP 5, RT 20 sec, 85%



Server Histories – One per server

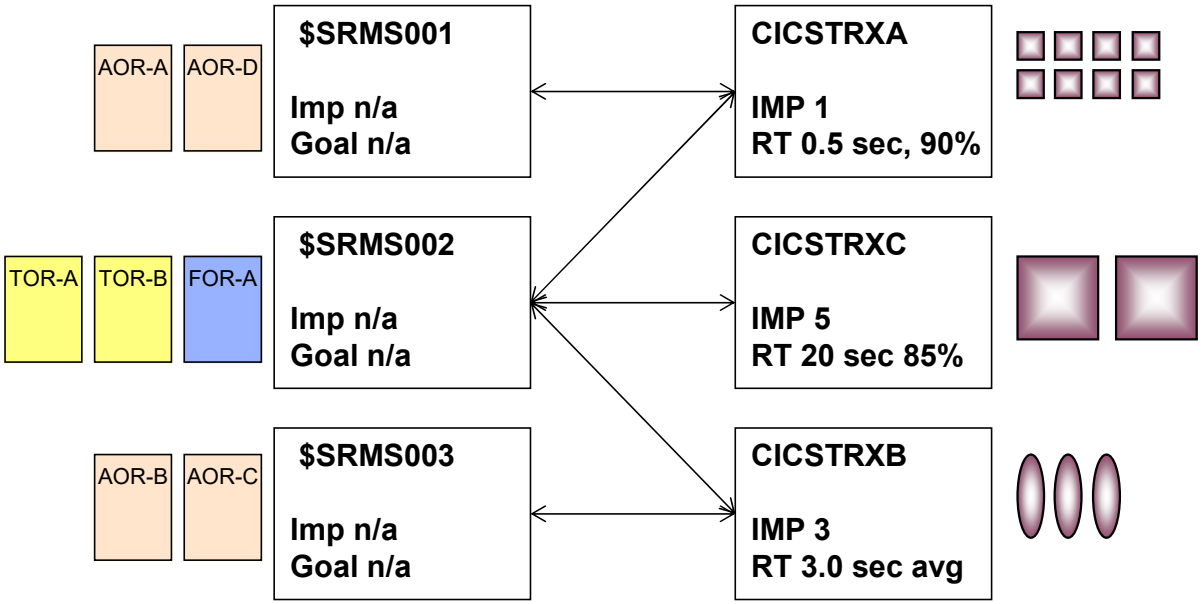
TOR-A served:
- CICSTRXA = 1000 times
- CICSTRXB = 400 times
- CICSTRXC = 50 times

AOR-A served:
- CICSTRXA = 400 times
- CICSTRXB = 0 times
- CICSTRXC = 0 times

Etc..

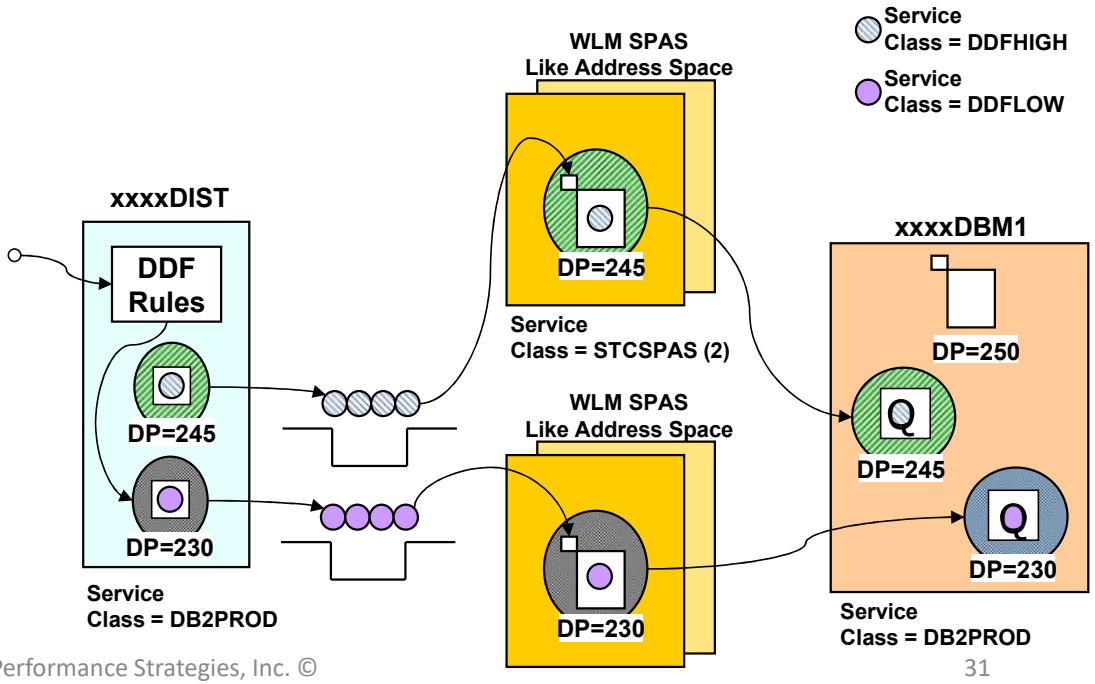
Example of Server Topology

- With server histories WLM is able to map out the server topology
  - Map of which servers are processing which transactions
  - Map of which transactions are being processed by which servers
- WLM creates internal periods and manages the work manages in these periods
  - Assign goal is ignored

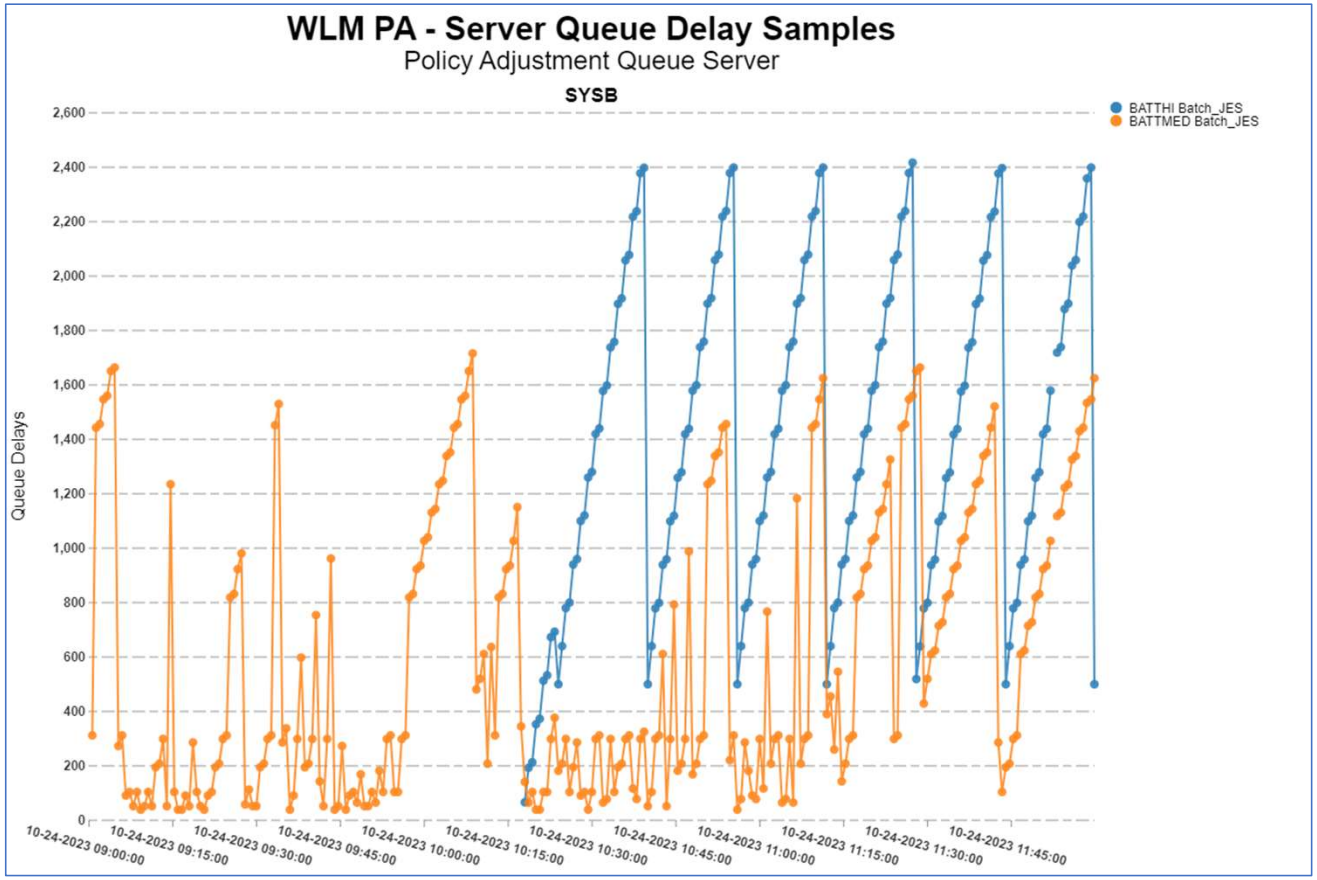


Server Sample Data Entry Section

- Contains data to gain insights into enclave work unit delays incurred by served periods
  - Work queue delays
  - Storage delays (Aux private, aux VIO, MPL, and other storage delays).
- Data include:
  - Delay samples incurred by:
    - Enclave servers
    - Queue servers
    - Batch work servers



Example Server Sample Data



This chart shows an example of the queue delays encountered by a queue server.

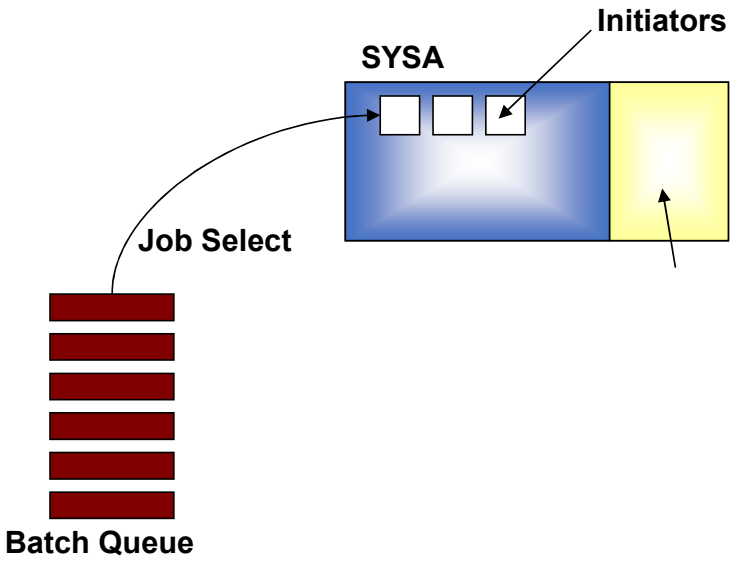
This is a batch example.

This report is super useful to better understand the queue delays for not only WLM managed initiator classes, but also for queue server classes for WebSphere servant regions and DB2 stored procedure regions.

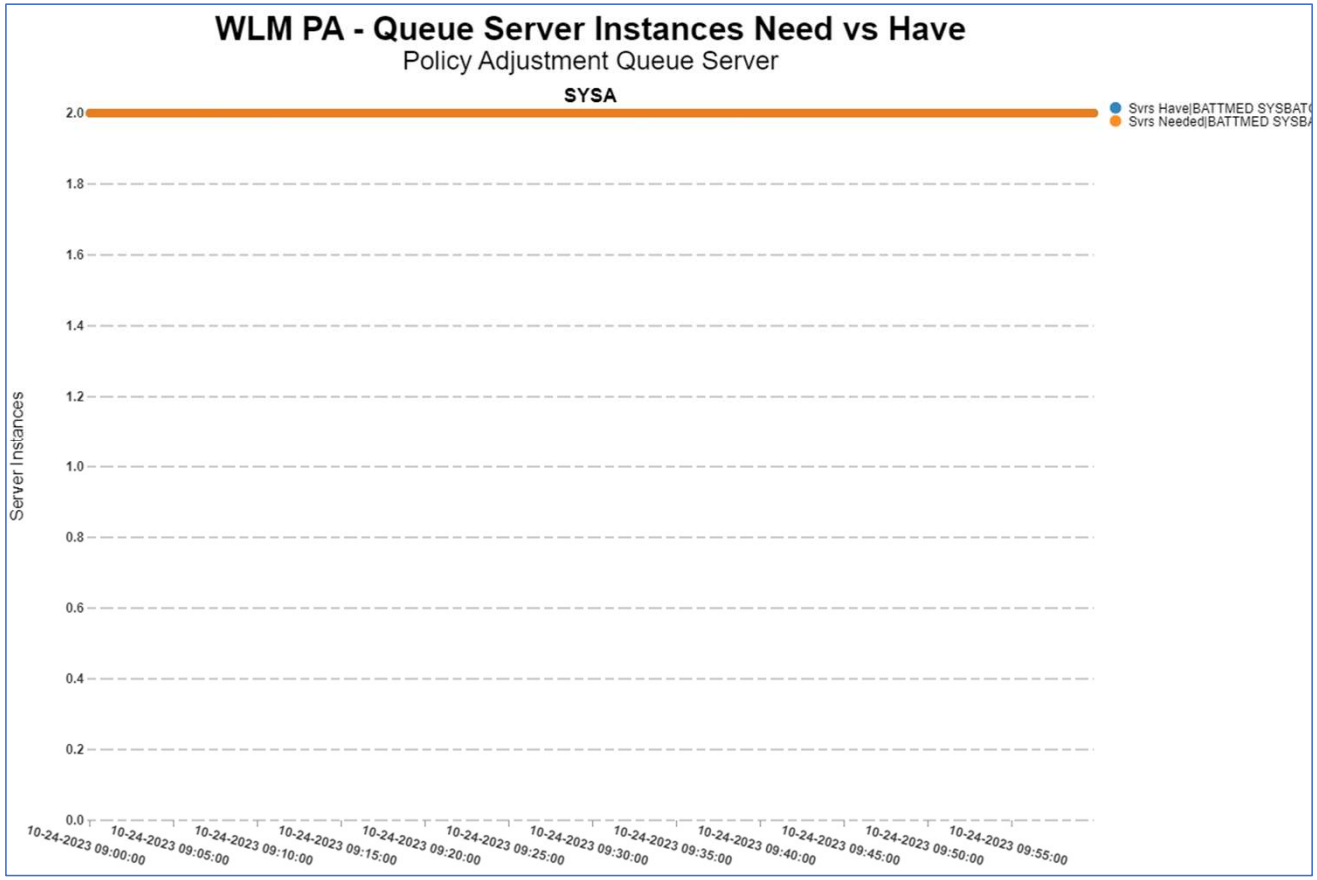
Queue Server Data Entry Section

- Contains data to gain insights application environments and scalable servers
  - Queue managers for application environments
  - Batch for WLM managed initiators

- Data include:
  - Information needed to gain insights into queuing of work
  - Work unit information
  - Queue types
  - Address space information



Example Queue Server Data



This chart shows an example of the queue server data for WLM batch initiator queues for a particular system.

This particular report is not very interesting since the values are static and over lay each other.

It shows that this particular batch service class needs 2 initiators and it has 2 initiators.

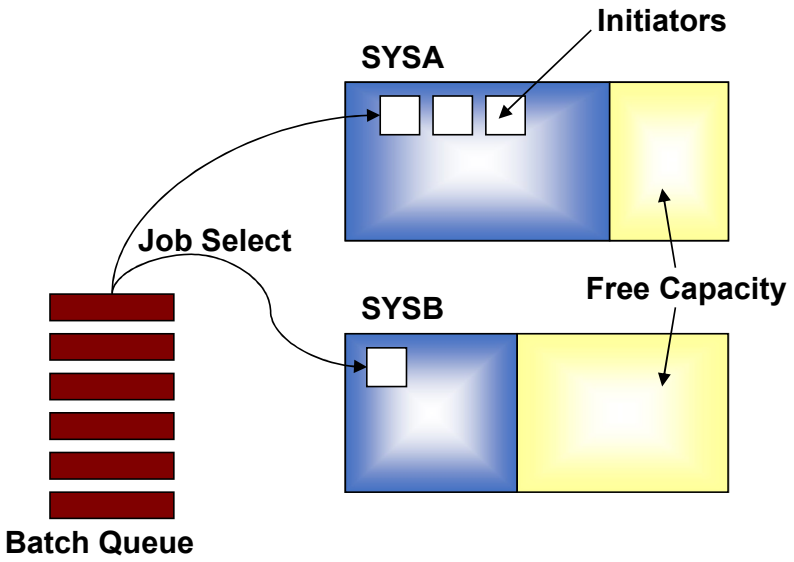
A more interesting report is when the need count is greater than the have count.

This report is super useful to better understand the queues for not only WLM managed initiator classes, but also for queue server classes for WebSphere servant regions and DB2 stored procedure regions.

Remote Queue Server Data Entry Section

- Like the queue server data entry section
  but contains data to gain insights into the remote queuing of application environments and scalable servers,
  - Queue managers for application environments
  - Batch for WLM managed initiators

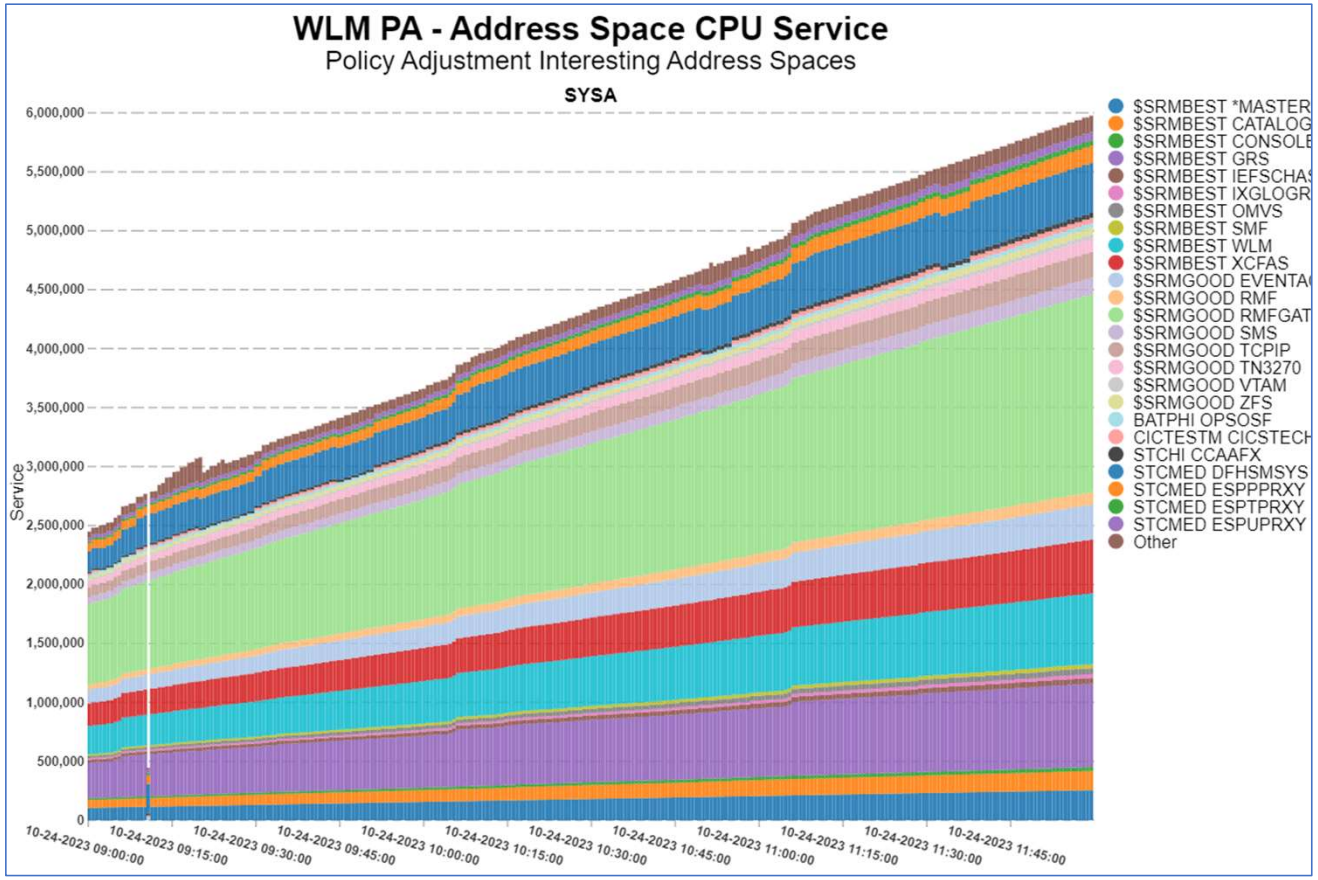
- Data include:
  - Information needed to gain insights into queuing of work
  - Work unit information
  - Queue types
  - Address space information
  - Donor and receiver information



Address Space Expanded Storage Access Policy

- Originally used to manage expanded storage
- Today contains data mostly geared towards interesting address spaces
  - The granularity of most WLM controls are at the service class period level
  - Also, some address spaces may have individual storage policies
- Data includes:
  - Names of interesting address spaces
  - Address space CPU usage information
    - CP, zIIP, zIIP on CP
  - Address Storage usage information

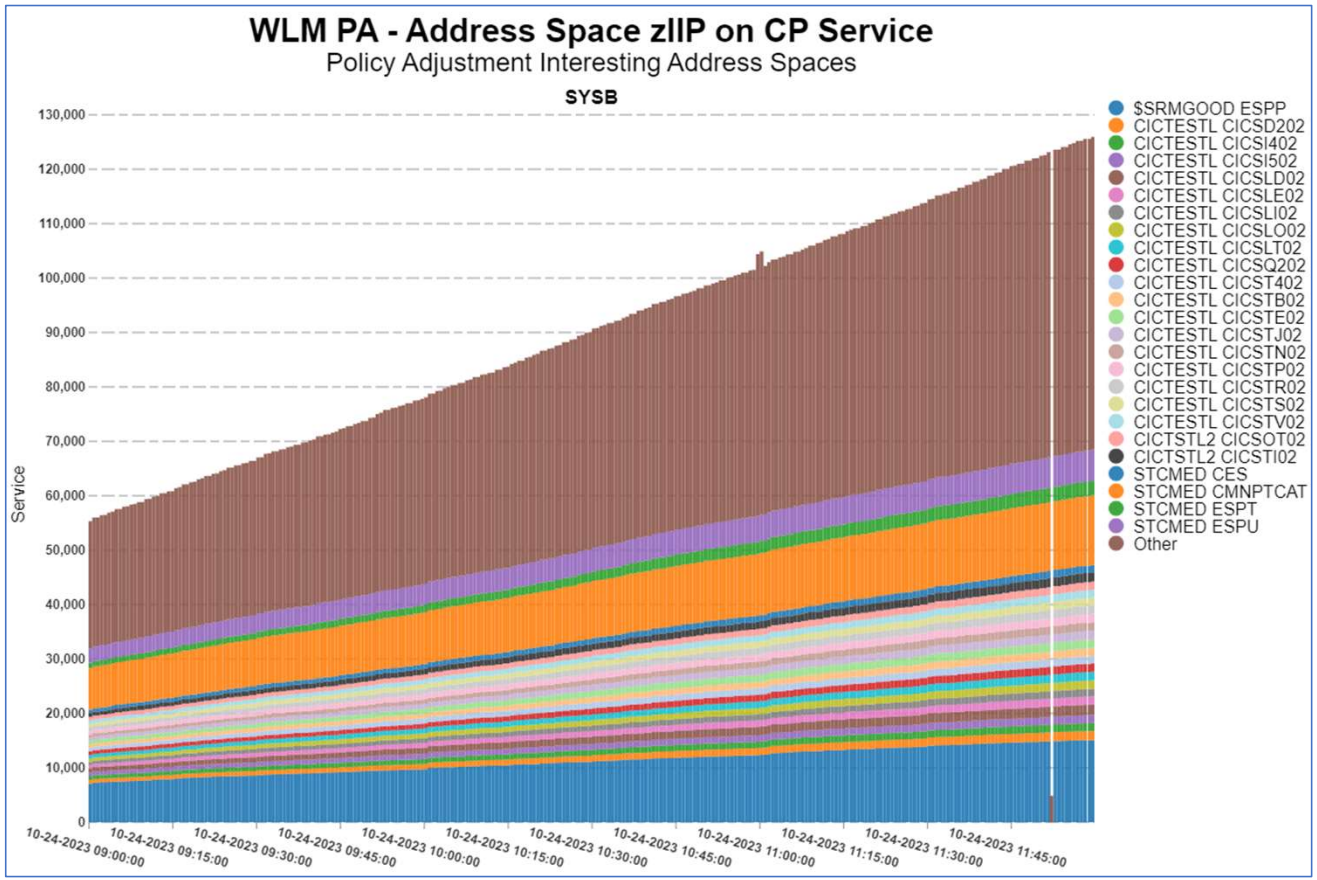
Example Interesting Address Spaces - CPU



This chart shows an example the top CPU service consumes for address spaces that WLM finds interesting.

Sometimes CPU will ‘monitor’ individual address spaces to better understand the contributors to the performance and consumption of service class periods.

Example Interesting Address Spaces - Crossover



This chart shows an example the top zIIP on CP CPU service consumes for address spaces that WLM finds interesting.

Sometimes CPU will 'monitor' individual address spaces to better understand the contributors to the performance and consumption of service class periods.

Key Lesson about the SMF 99.2 Records

- The original intention of the SMF 99 records was to help IBM debug WLM algorithm decisions
  - These records are cut every 10 seconds
- While practical use has been found by regularly processing an analyzing certain subtypes:
  - Such as the SMF 99.6, SMF 99.12, SMF 99.14
- The other SMF 99 subtypes, such as SMF 99.1 and SMF 99.2:
  - Are cryptic, not well documented
  - Contain values that are only meaningful to WLM algorithms
  - Are not very useful for our day-to-day use by general performance analyst
- At EPS, we use the SMF 99.1 and SMF 99.2 by searching for unusual patterns during periods of time customers call us asking to analyze a situation

Questions?