z Processor Consumption Analysis, or What Is Consuming All The CPU?

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

Creators of Pivotor®

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

Performance Workshops Available

During these workshops you will be analyzing your own data!

- WLM Performance and Re-evaluating of Goals
  - Instructor: Peter Enrico
  - September 15 – 19, 2014 - Kansas City, Missouri, USA
  - October 20 - 24, 2014 - Munich, Germany
- Parallel Sysplex and z/OS Performance Tuning (Web / Internet Based!)
  - Instructor: Peter Enrico
  - July 29 – 31, 2014 (Web)
- Essential z/OS Performance Tuning Workshop
  - Instructors: Peter Enrico and Tom Beretvas
- z/OS Capacity Planning and Performance Analysis
  - Instructor: Ray Wicks

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 and Reports Offer

- **Abstract**
  - The first step to any processor analysis is to understand your processor configuration and settings. The second step is to understand what workloads, address spaces, and transactions are consuming the fixed processor resource. It is only after understanding what and how the processor is being consumed can you conduct any sort of processor tuning or optimization exercise.
  - During this presentation Peter Enrico will show you how to conduct a processor resource consumption analysis. You will be provided with a top down approach to better understand processor measurements available to help you gain a drilldown insight into how the CPU resource is being consumed, and by what LPARs, Workloads, and transactions. Shown is what is known as a drill down approach for a processor performance analysis.

Presentation Overview

- Many areas need to be examined when decomposing CPU consumption
  - This presentation just discussions some of the many areas

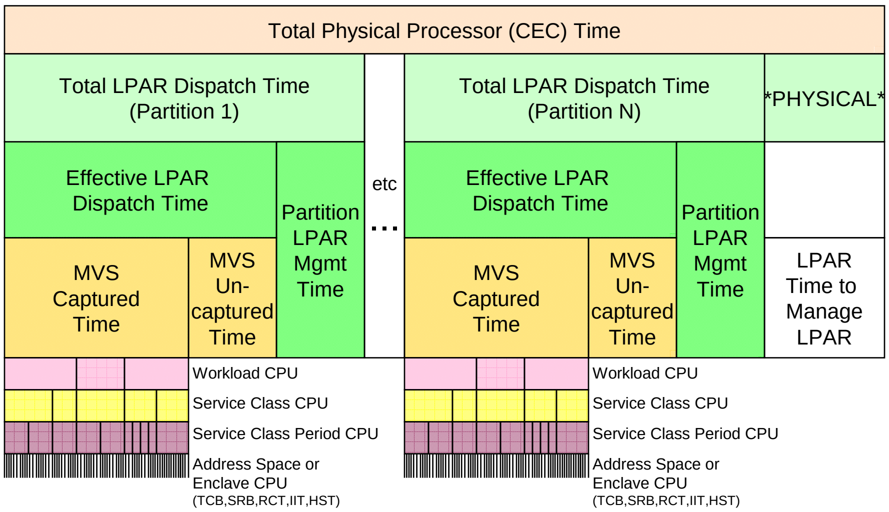
- Basic Processor Consumption Analysis
  - Decomposing CPU Consumption
    - By importance level
    - Displaced workloads
    - By Service Class and Report Class
  - Looking at CPU Dispatching Priorities
  - Looking at Latent Demand
  - New SMF 30 instruction counts (z/OS R1V2)
  - Workload Activity Analysis
    - CPU time during promotion

CPU 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 coupling facility based reports (charts and tables)
    - CPU - Machine Level Analysis
    - CPU - LPAR Level Analysis
    - CPU - HiperDispatch CPU Activity
    - CPU - SMF 113 Processor Counters
    - WLM - Workload Utilization Analysis
    - Coupling Facility Host Effect
    - And much more!
    - One-on-one phone call to explain your coupling facility measurements

Breakdown of General Purpose Processor

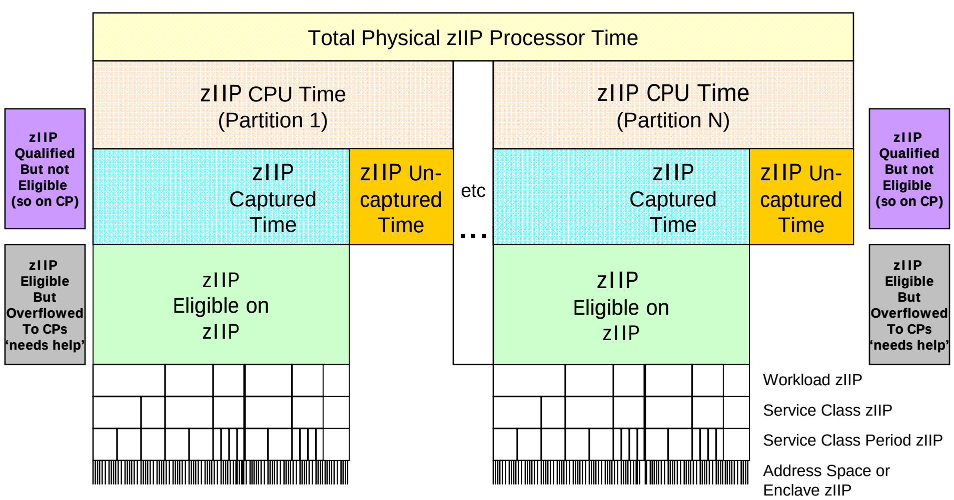
- We always needed to understand the break down of CP CPU consumption



Peter Enrico : www.epstrategies.com

Breakdown of zIIP Engine Time

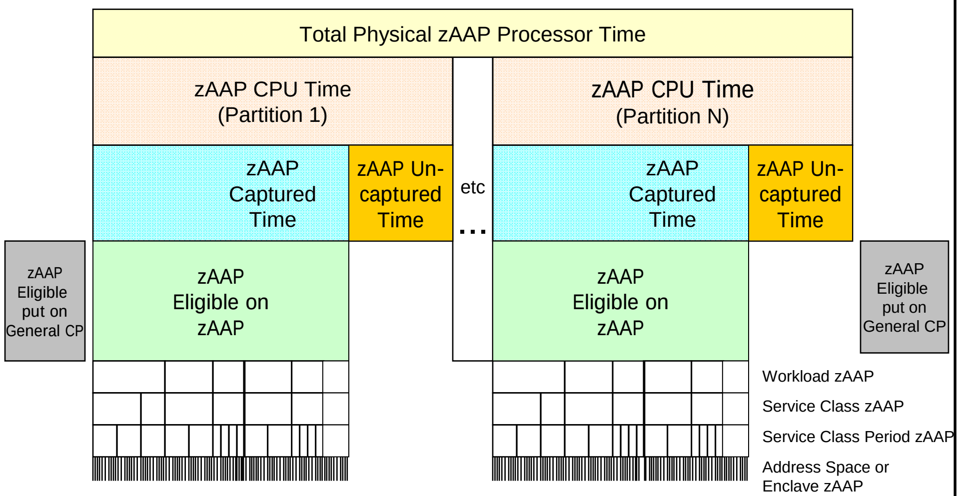
- We need to understand how PR/SM allocates the zIIP processor resource
  - In all measurements zIIPs



Peter Enrico : www.epstrategies.com

Breakdown of zAAP Engine Time

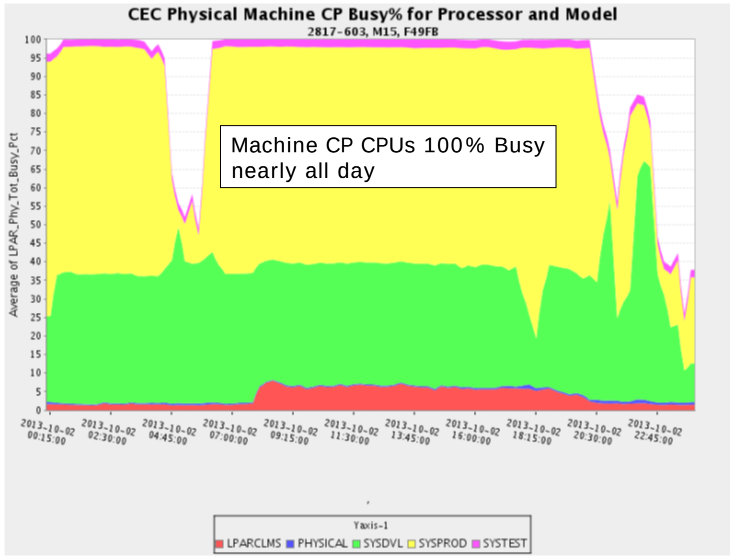
- We now need to understand where the zAAP CPU time is consumed



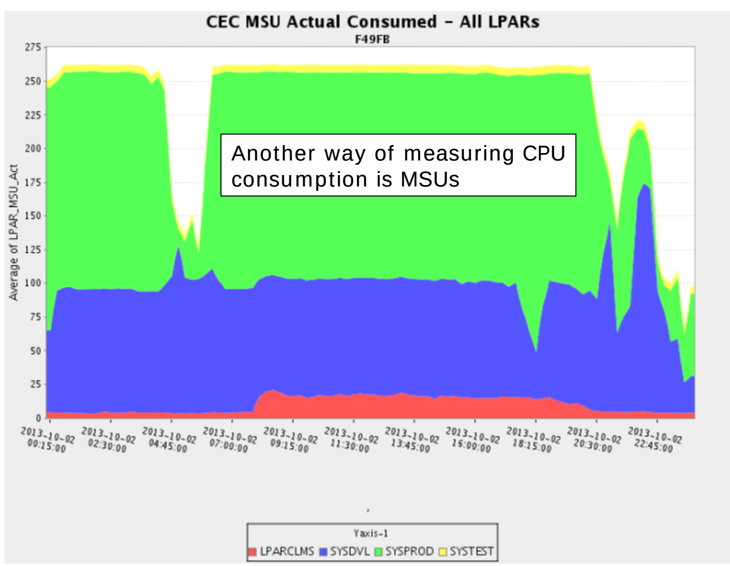
Decomposing CPU Consumption - WLM Workload Level Analysis

- CPU Consumption at the importance level
- CPU Consumption at the WLM Service Class and Service Class Period Level
- Commentary about Report Classes
- Other CPU consumption measurements
  - CPU consumed at promotion
  - Did lower importance work not consume CPU due to lack of demand or due to lack of CPU?

Machine Busy – CP Percent Busy

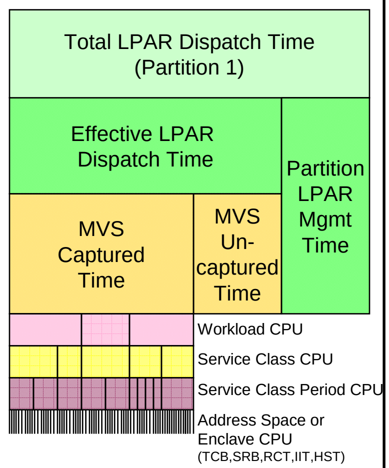


Could measure CPU Consumption in MSUs



z/OS CPU Times

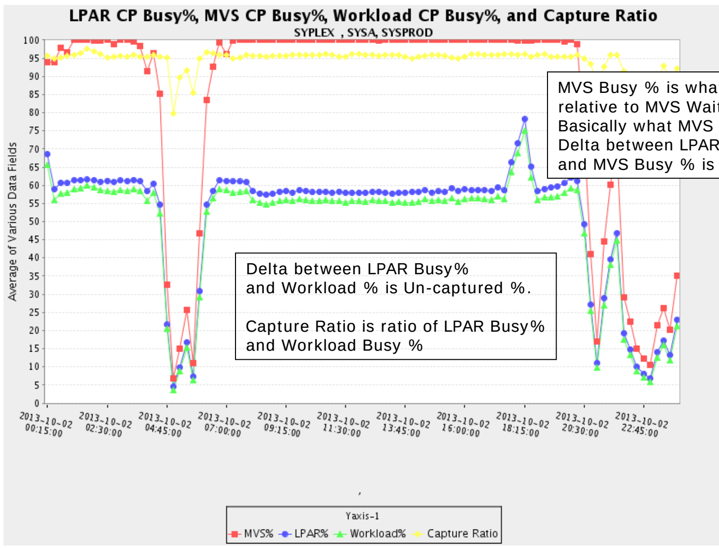
- Capture Ratios used to understand the stability and cost of system overhead
- Effective Dispatch Time
  - Time that the z/OS and the workloads were executing on the CPU
- MVS Capture Time
  - Time that can be accounted for towards specific workloads
- MVS Un-captured Time
  - System overhead
- Capture Ratio
  - Ratio of MVS Capture Time to Effective Dispatch Time



Causes for Uncaptured Time

- Many causes for uncaptured time. Common causes are as follows:
  - High page fault rates
  - Full preemption
  - Suspense lock contention
  - Spin lock contention
  - Getmain/Freemain activity (recommend cell pools)
  - SRM time-slice processing
  - Interrupts
  - SLIP processing
  - Long queues being processed in uncaptured processing
  - Excessive swap-out and swap-in activity
  - Affinity processing (such as need for a specific CPU or crypto facility)

LPAR Busy%, Workload%, MVS%, and Capture Ratio



MVS Busy % is what utilization relative to MVS Waits.
Basically what MVS wanted.
Delta between LPAR Busy % and MVS Busy % is usually due to capping

Delta between LPAR Busy% and Workload % is Un-captured %.

Capture Ratio is ratio of LPAR Busy% and Workload Busy %

Understanding Dispatching to Gain Insight to MVS Busy %

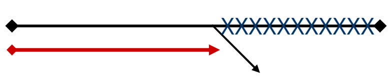
- Dispatch Time
  - Time logical processor is associated with a physical processor

Dispatch Interval

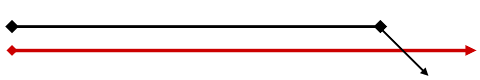
Logical Processor associated with a physical processor

Logical Processor disassociated with a physical processor

- MVS Time
  - Time z/OS was busy before voluntarily giving up a processor

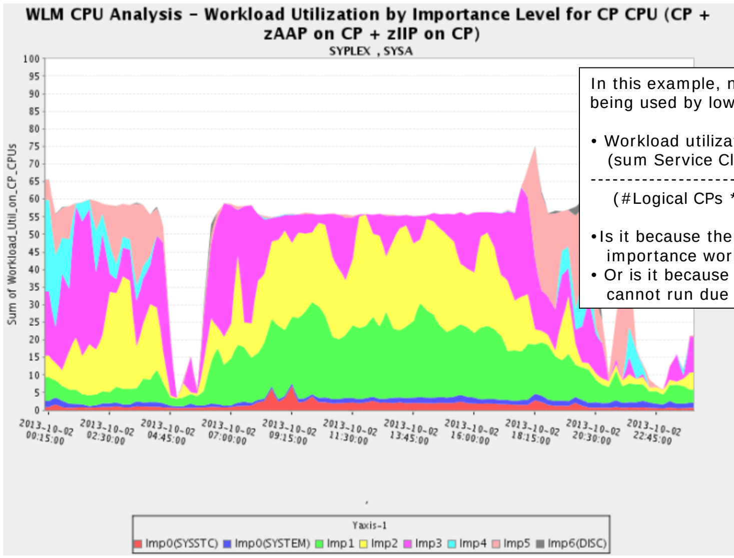


Voluntary Wait
- z/OS voluntarily gives up the processor
- MVS time equals dispatch time



Involuntary Wait
- z/OS does not give up the processor voluntarily
- Instead PR/SM un-dispatches the partition
- MVS time will be greater than dispatch time

Workload Utilization by Importance Level



In this example, notice that little CPU is being used by low importance work.

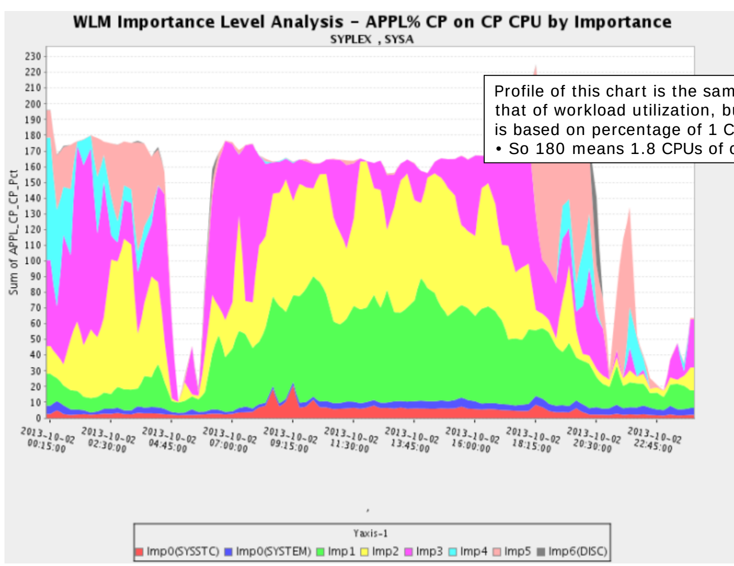
- Workload utilization is calculated as (sum Service Class CPU Tim)

(#Logical CPs * Interval Time)

- Is it because there is little to no low importance work?
- Or is it because low importance work cannot run due to lack of capacity?

Peter Enrico : www.epstrategies.com © Enterprise Performance Strategies, Inc. z/OS Processor Analysis - 17

APPL% by Importance Level

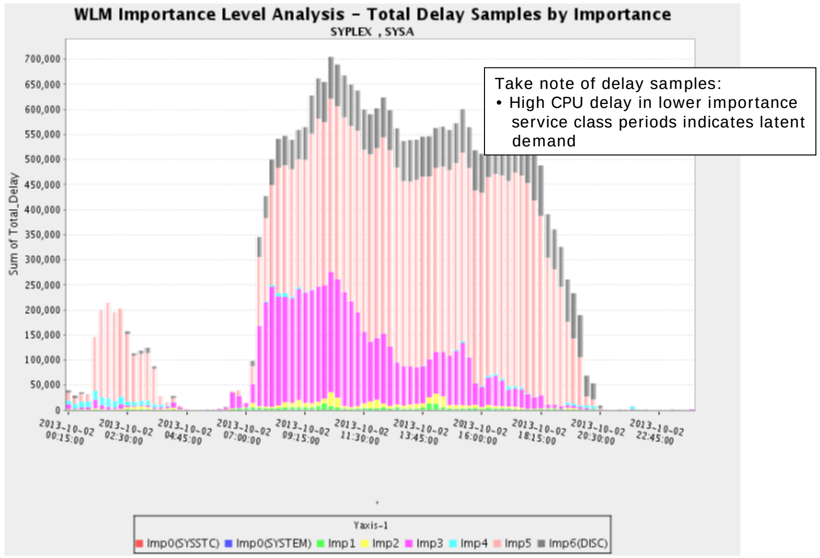


Profile of this chart is the same as that of workload utilization, but scale is based on percentage of 1 CPU.

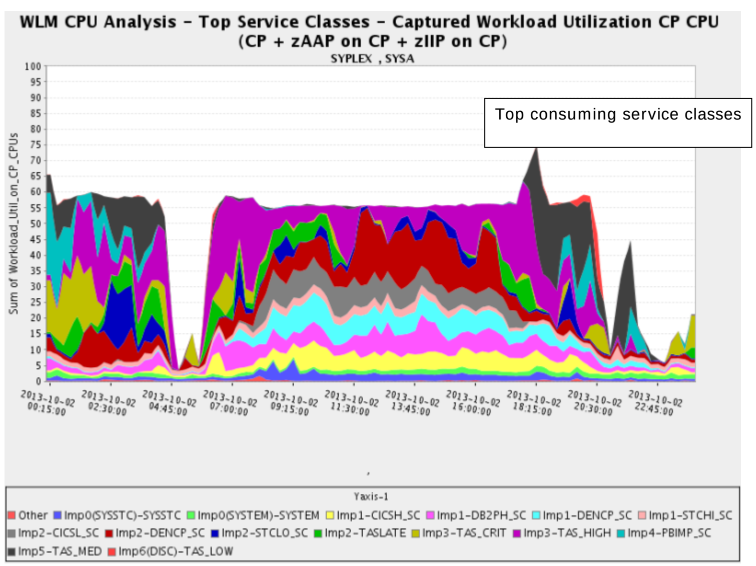
- So 180 means 1.8 CPUs of capacity.

Peter Enrico : www.epstrategies.com © Enterprise Performance Strategies, Inc. z/OS Processor Analysis - 18

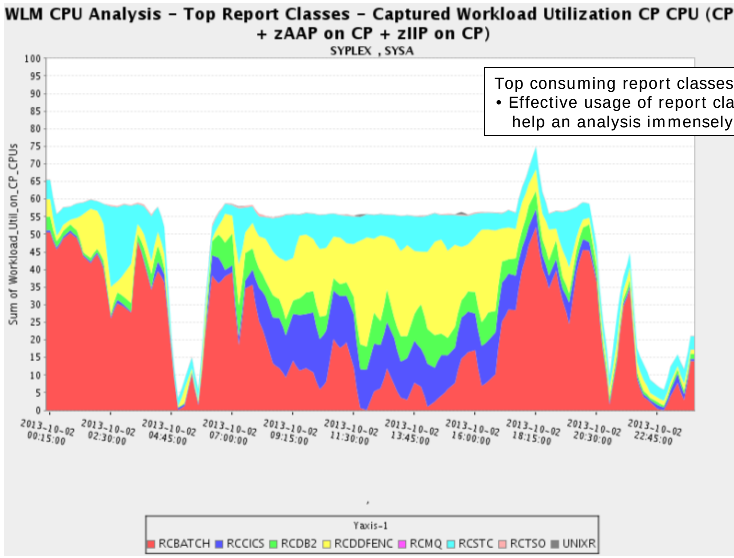
Delay Samples by Importance Level



Workload Utilization by Service Class Period



Workload Utilization for Top Report Classes

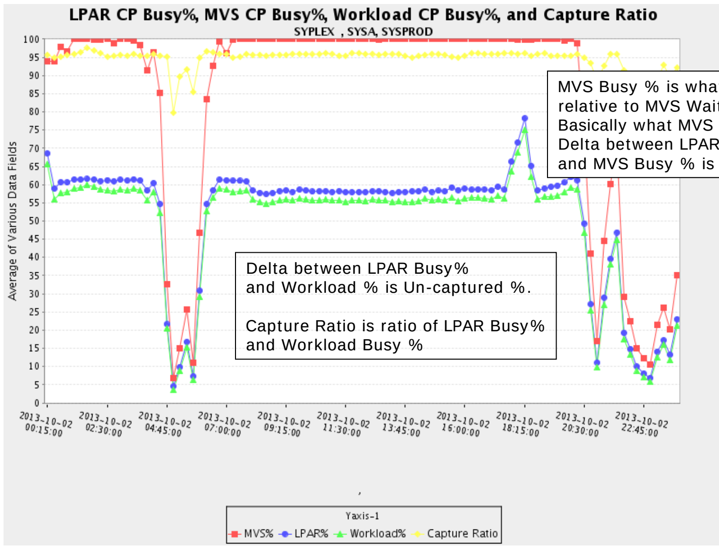


Top consuming report classes
- Effective usage of report classes can help an analysis immensely.

Frustrating CPU Time Problem

Sometimes you may see more CPU time consumed by your workloads than the CPU time that PR/SM is dispatching to the LPAR.

LPAR Busy%, Workload%, MVS%, and Capture Ratio

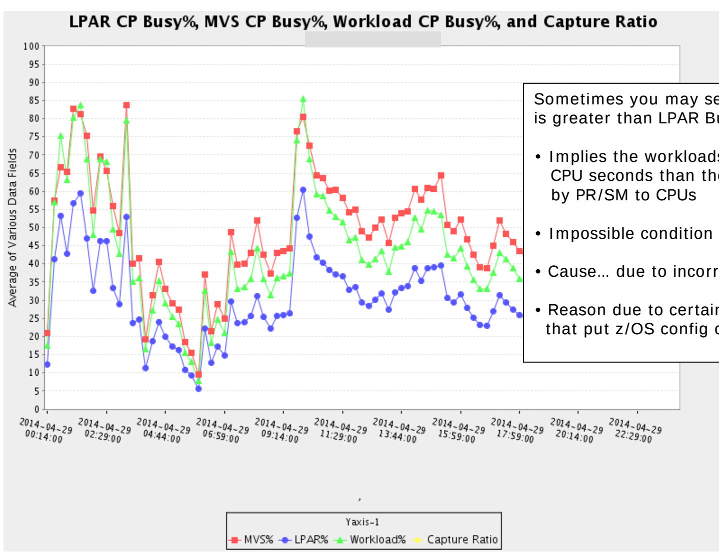


MVS Busy % is what utilization relative to MVS Waits. Basically what MVS wanted. Delta between LPAR Busy % and MVS Busy % is usually due to capping

Delta between LPAR Busy% and Workload % is Un-captured %.

Capture Ratio is ratio of LPAR Busy% and Workload Busy %

Beware... sometimes workload CPU can be greater than dispatch CPU. (???????)

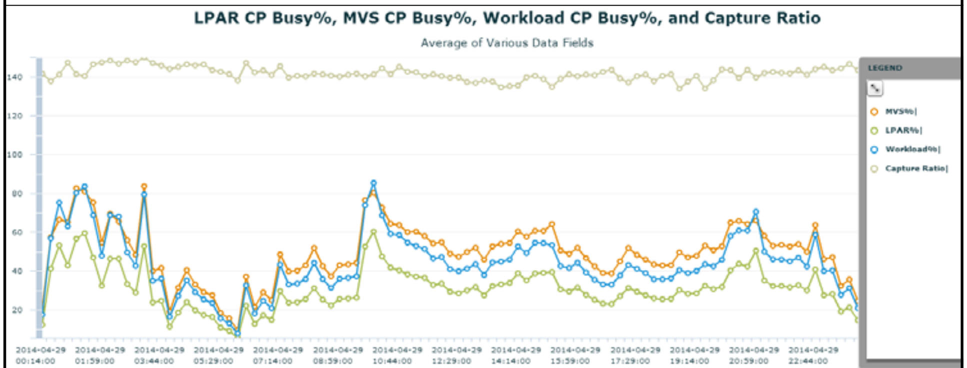


Sometimes you may see that Workload % is greater than LPAR Busy %

- Implies the workloads consumed more CPU seconds than the system was dispatched by PR/SM to CPUs
- Impossible condition
- Cause... due to incorrect SU/Sec constant
- Reason due to certain configuration changes that put z/OS config out of sync with PR/SM

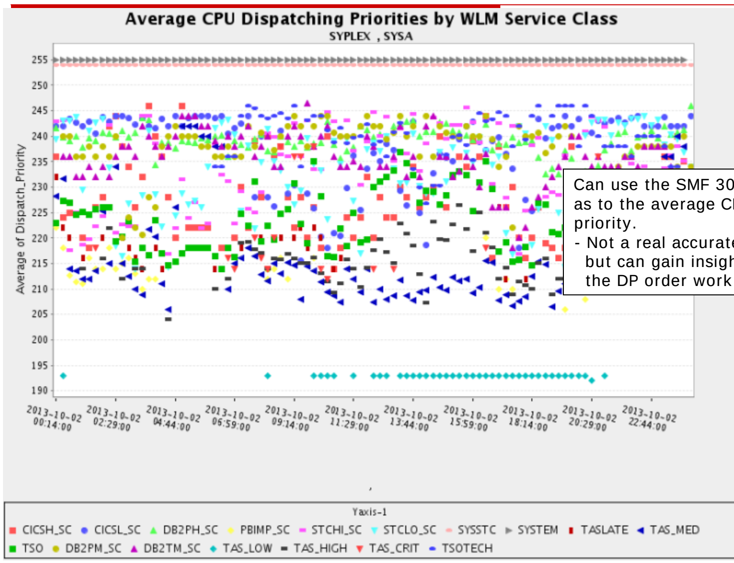
Beware… sometimes workload CPU can be greater than dispatch CPU. (???????)

- This is the same chart as previous slide only showing the capture ratios of greater than 100% since Workload Time > LPAR Dispatch Time
- Results in higher than expected CPU time per transactions… screws up CPU



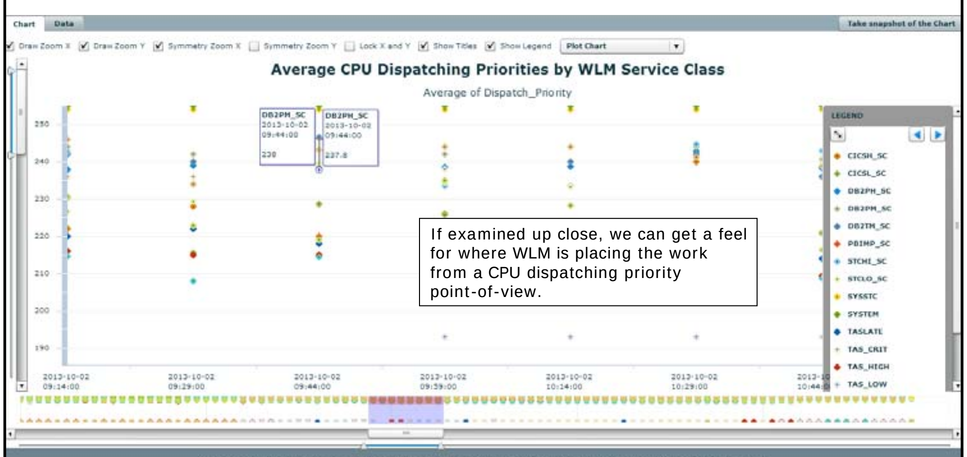
Looking at CPU Dispatching Priorities (an approximation)

Average CPU Dispatching Priorities for Address Spaces



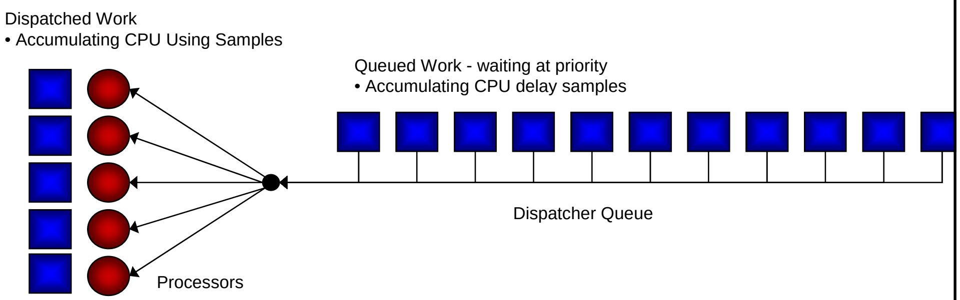
Can use the SMF 30 to gain insights as to the average CPU dispatching priority.
- Not a real accurate measurement, but can gain insight into where in the DP order work is.

Average CPU Dispatching Priorities for Address Spaces

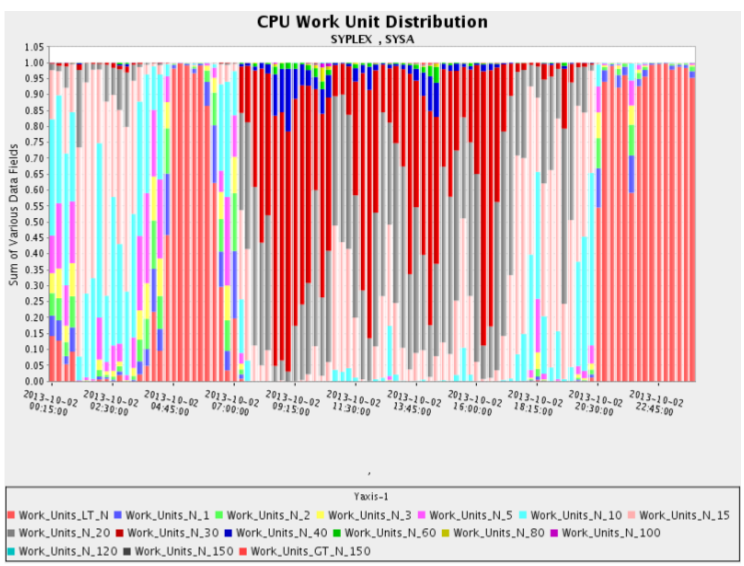


If examined up close, we can get a feel for where WLM is placing the work from a CPU dispatching priority point-of-view.

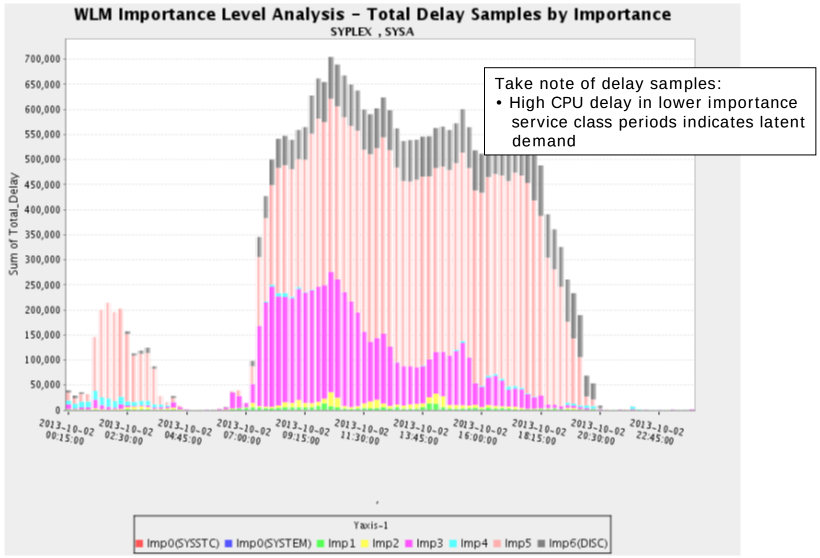
Insights into Latent Demand



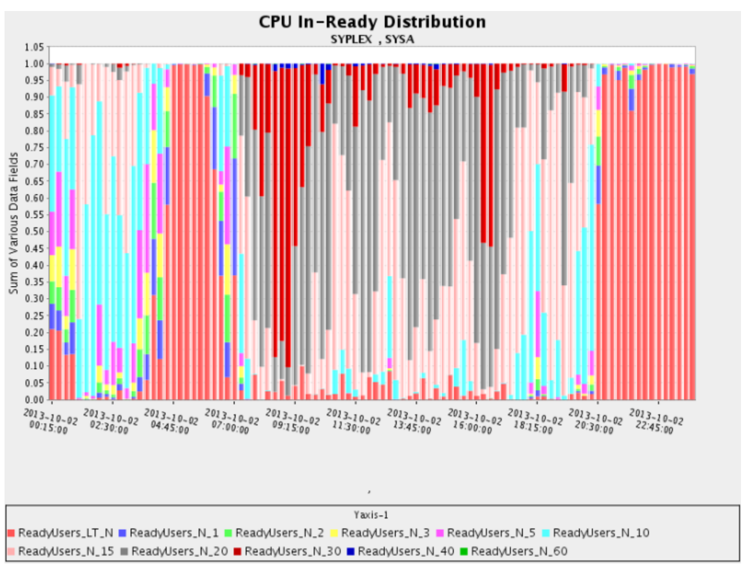
Work Unit Distribution Showing Latent Demand



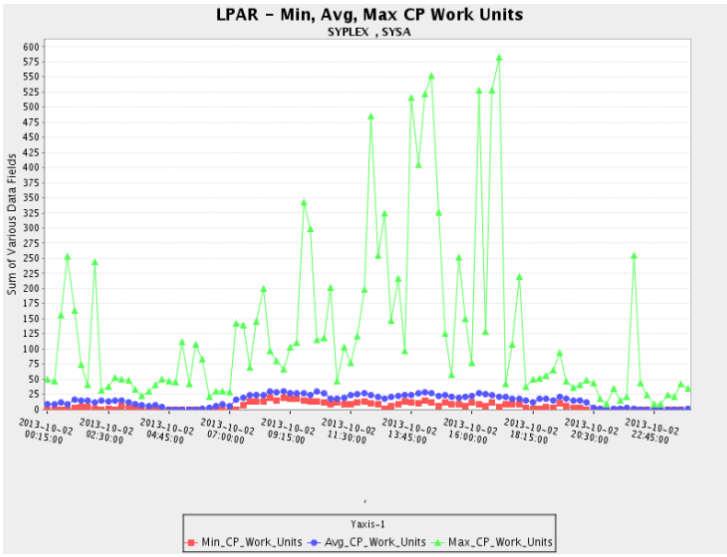
Delay Samples by Importance Level



Older Style In-Ready Distribution – Less Accurate Latent Demand



Min / Max / Avg Work Unit Queuing



Big z/OS V2R1 Update

Instruction Counts added to SMF30 Record

Path Length versus Memory Hierarchy

- Chargeback usually based on CPU time, but CPU time can be influenced by memory hierarchy
  - Path length of the code executed by customer applications and transactions
    - When move from one processor to another this generally does not change much for a specific customer
  - Memory Hierarchy
    - Cycles consumed include cycles to resolve processor cache misses
    - Heavily influenced by key factors result potentially wide variations in realized capacity
- Currently for chargeback the SMF 30 measurements include CPU time values which can be heavily influenced by the memory hierarchy
  - CPU Time for chargeback
- New corresponding path length measurements will be added to the SMF 30
  - Instruction for chargeback

General Concept of SMF30 Instruction Count Fields

- As a reminder, the two primary ways processor cycles are spent include
  - Running the workload
  - Resolving the cache misses
- Thus, CPU chargeback could vary based on cycles consumed to resolve the CPU misses through the memory hierarchy.
- Is it always fair to a customer to charge based on a variable factor outside the customer's control?
- The new instruction count fields will provide insights into what the customer is actually doing, and variability will be minimized since path length generally remains constant regardless of configuration and usage of the memory hierarchy.

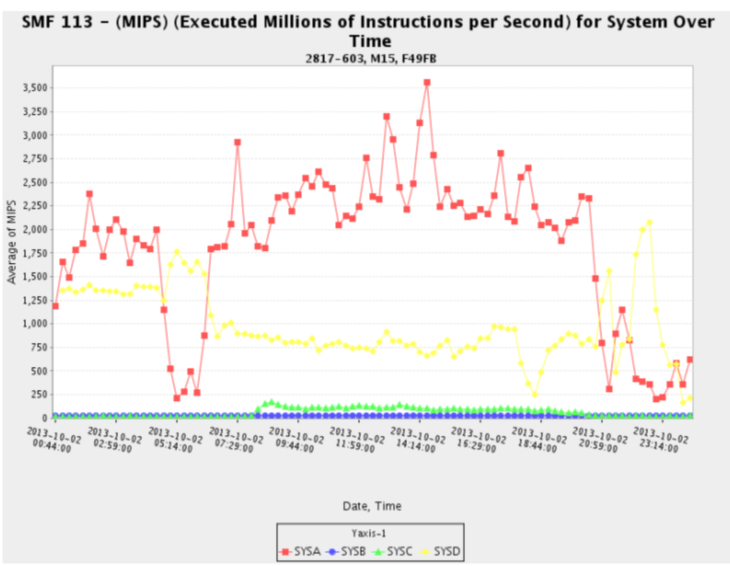
Models for Understanding Usage

- **CPU Time Based**
  - Provider Cost Model
  - Data center has only so many CPU seconds to sell
  - Service based on the number of CPU seconds used is basing the price on provider costs

- **Instruction / Path Length Based**
  - Buyer Value Model
  - Transaction has value to the user
  - Number of instructions executed by a program is a measure how value the user is getting

Peter Enrico : www.epstrategies.com © Enterprise Performance Strategies, Inc. z/OS Processor Analysis - 37

Executed MIPS as per the SMF 113 Record



Peter Enrico : www.epstrategies.com © Enterprise Performance Strategies, Inc. z/OS Processor Analysis - 38

New z/OS 2.1 SMF 30 fields

- Need SMF113 HIS data collector to run to get these new fields
- New SMF30 Instruction Count Fields
  - SMF30_Inst_CP_Task
  - SMF30_Inst_CP_NonPreemptSRB
  - SMF30_Inst_CP_PreemptSRB
  - SMF30_Inst_Offload
  - SMF30_Inst_OffloadOnCP
  - SMF30_Inst_CP_Enclave
  - SMF30_Inst_Offload_Enclave
  - SMF30_Inst_OffloadOnCP_Enclave
  - SMF30_Inst_CP_DepEnc
  - SMF30_Inst_Offload_DepEnc
  - SMF30_Inst_OffloadOnCP_DepEnc
  - SMF30_InstCaptDisruption
  - SMF30_InstCaptLimited
  - Missing some Enclave and DepEnc instruction counts

Relating SMF30
CPU Time fields to new Instruction Count fields

- SMF30CPT –
  - All standard CPU step time. Includes enclave time, preemptable class SRB time, client SRB time. Also includes time consumed by zAAP or zIIP eligible work running on a standard processor.
  - Comparable to sum of the following instruction counts fields:

    SMF30_Inst_CP_Task +
    SMF30_Inst_CP_PreemptSRB +
    SMF30_Inst_OffloadOnCP +
    SMF30_Inst_CP_Enclave +
    SMF30_Inst_OffloadOnCP_Enclave +
    SMF30_Inst_CP_DepEnc +
    SMF30_Inst_OffloadOnCP_DepEnc
- SMF30CPS –
  - Step CPU time under the service request block (SRB) non preemptable.
  - Comparable to the following instruction count fields:

    SMF30_Inst_CP_NonPreemptSRB

Mapping of Legacy SMF 30 CPU Time Fields to SMF 30 Instruction Count Fields

Legacy SMF 30 CPU Time Fields	New SMF30 Instr Count Fields
SMF30CPT	SMF30_Inst_CP_Task + SMF30_Inst_CP_PreemptSRB + SMF30_Inst_OffloadOnCP + SMF30_Inst_CP_Enclave + SMF30_Inst_OffloadOnCP_Enclave + SMF30_Inst_CP_DepEnc + SMF30_Inst_OffloadOnCP_DepEnc
SMF30CPS	SMF30_Inst_CP_NonPreemptSRB
SMF30ASR	SMF30_Inst_CP_PreemptSRB
SMF30ENC	SMF30_Inst_CP_Enclave + SMF30_Inst_OffloadOnCP_Enclave
SMF30DET	SMF30_Inst_Offload_DepEnc + SMF30_Inst_OffloadOnCP_DepEnc
SMF30_TIME_ON_IFA + SMF30_TIME_ON_SUP	SMF30_Inst_Offload
	SMF30_Inst_Offload_Enclave SMF30_InstCaptDisruption SMF30_InstCaptLimited Missing some Enclave and DepEnc instruction counts

Relating SMF30 CPU Time fields to new Instruction Count fields

- SMF30ASR –
  - Additional CPU time accumulated by the preemptable and client SRBs for this job.
  - This value included in SMF30CPT
  - Comparable to the following instruction count fields:
    SMF30_Inst_CP_PreemptSRB
- SMF30ENC –
  - CPU time used by the independent enclave, but only when in the WLM enclave.
  - Note that independent enclave time on an IFA is not included.
  - This value included in SMF30CPT
  - Comparable to the following instruction count fields:
    - SMF30_Inst_CP_Enclave
    - SMF30_Inst_OffloadOnCP_Enclave

Relating SMF30
CPU Time fields to new Instruction Count fields

- SMF30DET –
  - CPU time used by the dependent enclave, but only when in the WLM enclave.
  - Note that dependent enclave time on an IFA is not included.
  - This value included in SMF30CPT
  - Comparable to the following instruction count fields:
    - SMF30_Inst_CP_DepEnc
    - SMF30_Inst_OffloadOnCP_DepEnc
- SMF30_TIME_ON_IFA and SMF30_TIME_ON_SUP
  - Accumulation of CPU time spent on zAAP.
  - Accumulation of CPU time spent on zIIP.
  - Comparable to the following instruction count fields:
    - SMF30_Inst_Offload

Reports / SMF 30 Instruction Counts Processing/Discussion Offer !!!

- Study underway
  - I am in the process of studying SMF 30 instruction counts, and their uses
- Request for Data!
  - I am looking for a collection of SMF 30, 70, 72, and 113 data to better understand the instruction based measurements
  - Please contact me, Peter Enrico for instructions for sending raw SMF data
    - Send an email to peter.enrico@epstrategies.com
  - Deliverable:
    - Dozens of SMF reports (charts and tables)
      - Summary by system
      - Summary by CPU
      - Before / After comparison reports
      - Raw counter reports
      - Much more…
    - One-on-one phone call to explain your measurements

Top Address Spaces Consuming CPU

Top 20 Address Spaces Consuming Most CPU in 24 Hours

SC_Name	RC_Name	Job_Name	AS_Type	SYS1	SYS2	SYS3	SYS4	Sum	Machine%
DB2PH_SC	DB2R	DSNDIST	STC	35,883.6				35,883.6	13.8%
CICSH_SC	CICSR	CICSHADP	STC	13,921.3				13,921.3	5.4%
CICSL_SC	CICSR	CICSH81P	STC	10,527.0				10,527.0	4.1%
DB2PH_SC	DB2R	DSNDBM1	STC	10,127.9				10,127.9	3.9%
STCLO_SC	STCR	DFHSM	STC	7,964.1			214.6	8,178.7	3.2%
CICSH_SC	CICSR	CICSH11P	STC			5,797.9		5,797.9	2.2%
STCHI_SC	STCR	OMEGDSST	STC	1,622.2	1,019.7	1,146.9	1,827.2	5,616.0	2.2%
SYSTEM	STCR	WLM	SYS	535.2	342.7	211.5	1,890.2	2,979.6	1.1%
HPS_HIGH	BATCHR	HM026D03	JOB	2,376.6				2,376.6	0.9%
SYSSTC	STCR	NET	STC	1,005.0	44.8	485.3	749.8	2,285.0	0.9%
HPS_HIGH	BATCHR	IT110D01	JOB	2,145.8				2,145.8	0.8%
SYSTEM	STCR	CATALOG	SYS	1,540.7	11.3	14.7	572.2	2,138.9	0.8%
SYSSTC	STCR	TCPIP	STC	1,476.8	98.9	118.5	374.9	2,069.1	0.8%
TBATL_SC	BATCHR	DB2HRWS0	JOB				1,924.4	1,924.4	0.7%
CICSH_SC	CICSR	CICSMG1P	STC	1,735.9				1,735.9	0.7%
TBATL_SC	BATCHR	SITH085U	JOB				1,685.2	1,685.2	0.7%
DB2TH_SC	DB2R	HPDQDIST	STC				1,683.1	1,683.1	0.6%
DB2TH_SC	DB2R	HPDQDBM1	STC				1,551.3	1,551.3	0.6%
PBIMP_SC	BATCHR	HPSVSAM1	JOB	1,302.6				1,302.6	0.5%
HPS_HIGH	BATCHR	HM026D01	JOB	1,296.8				1,296.8	0.5%

Objective of WLM Management of CICS & IMS

- Allow assignment of goals to the transactions and let the WLM determine which regions need the resources to meet these goals.

Region Goals

IMP 1, Velocity 60

Transaction Goals

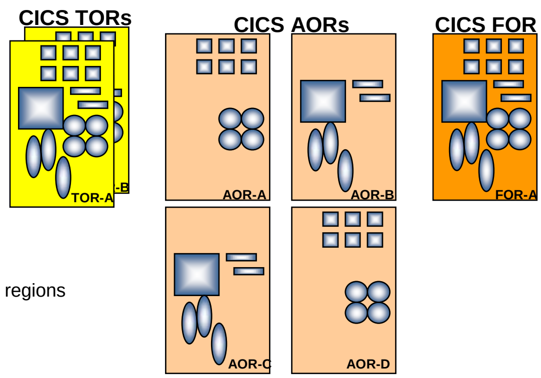
IMP 1, RT .5 sec, 90%

IMP 1, RT .75 sec, 90%

IMP3, RT 2 sec, 90%

IMP 3, Avg RT 3 sec

IMP 5, RT 20 sec, 85%

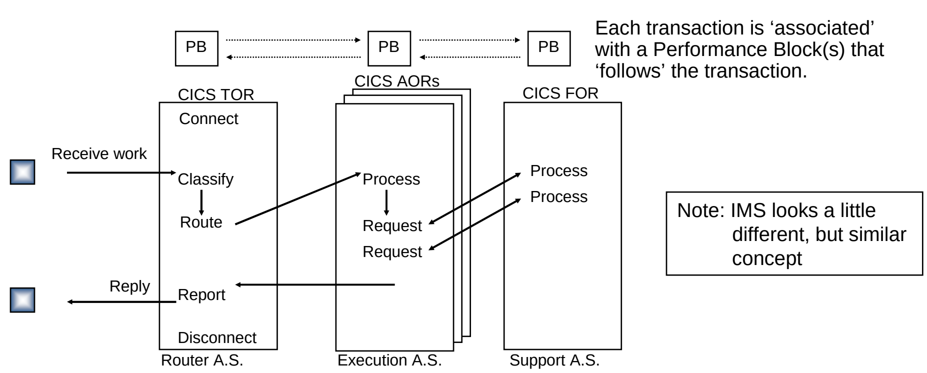


To meet the RT goals of ▢ the following regions must be managed:

- CICS TOR-A, TOR-B
- CICS AOR-A, AOR-D
- CICS FOR-A

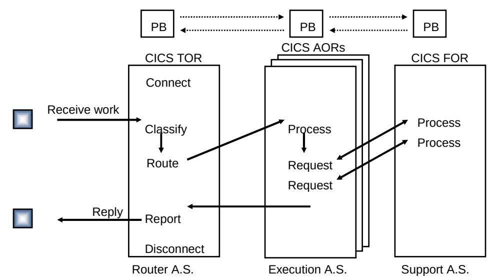
WLM needs an awareness of which regions are processing which transactions, and how often

- CICS and IMS exploit WLM Work Manager services
  - Regions '***Connect***' (ie 'register') to WLM during startup & obtain current service policy
  - At transaction startup, region uses WLM '***Classify***' to associate incoming transaction with a service class
  - At transaction end, region uses WLM '***Report***' to signal end and report response time
  - Other important services to make this all work



WLM Sampling and CICS MAXTASK Parameter

- Beware of excess sampling overhead due to CICS MAXTASK parameter!
  - In a CICS environment, one PB is pre-allocated for each possible task as set by the CICS MAXTASK parameter
- All PBs are sampled every 1/4 second
  - Could cause lots of WLM sampling overhead!
  - Check CICS MAXTASK parameter to make sure it is not set unnecessarily high
    - Set to your system's true high water mark
  - Mostly resolved, but still watch MAXTASK

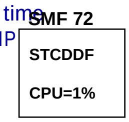


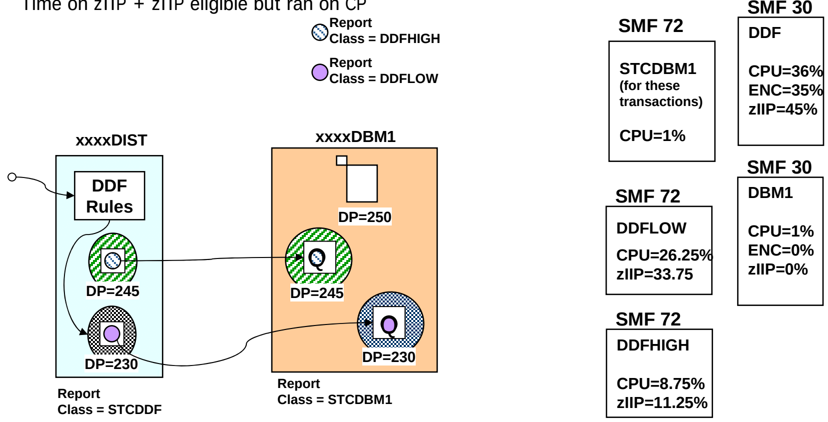
Top 20 Address Spaces Consuming Most CPU in 24 Hours

SC_Name	RC_Name	Job_Name	AS_Type	SYS1	SYS2	SYS3	SYS4	Sum	Machine%
DB2PH_SC	DB2R	DSNDIST	STC	38,655.5				38,655.5	14.9%
CICSH_SC	CICSPRHR	CICSHADP	STC	14,269.1				14,269.1	5.5%
DB2PH_SC	DB2R	DSNDBM1	STC	7,147.9				7,147.9	2.8%
CICSH_SC	CICSPRHR	CICSH81P	STC	5,032.1				5,032.1	1.9%
STCLO_SC	OMEGAMON	OMEGDSST	STC	1,403.0	807.8	928.4	1,340.4	4,479.6	1.7%
CICSH_SC	CICSPRHR	CICSH11P	STC			3,662.2		3,662.2	1.4%
STCLO_SC	DFHSMR	DFHSM	STC	2,929.9			295.1	3,225.0	1.2%
DB2TM_SC	DB2R	DB2JDIST	STC				2,839.3	2,839.3	1.1%
SYSTEM	STCR	WLM	SYS	483.3	304.8	192.9	1,314.1	2,295.1	0.9%
SYSSTC	STCR	RMFGAT	STC	414.7	644.8	376.3	858.9	2,294.7	0.9%
SYSSTC	STCR	TCPIP	STC	1,319.6	85.6	99.4	541.3	2,045.9	0.8%
PSTD_SC	BATSTDR	DB105M00	JOB	2,007.3				2,007.3	0.8%
PMED_SC	BATMEDR	HPSVSMTH	JOB	1,939.5				1,939.5	0.7%
PHIGH_SC	BATHIGHR	IT110D01	JOB	1,860.6				1,860.6	0.7%
TBATL_SC	BATTSTR	DSNLRW00	JOB				1,717.5	1,717.5	0.7%
SYSSTC	STCR	NET	STC	728.5	35.7	364.5	507.5	1,636.2	0.6%
SYSTEM	STCR	CATALOG	SYS	1,185.5	11.2	13.6	389.6	1,600.0	0.6%
CICSH_SC	CICSPRHR	CICSMG1P	STC	1,326.8				1,326.8	0.5%
PMED_SC	BATMEDR	HPSVSMA1	JOB	1,200.9				1,200.9	0.5%
PMED_SC	BATMEDR	DB2REOF1	JOB	1,120.8				1,120.8	0.4%

DDF and Independent Enclaves with zIIP Engines

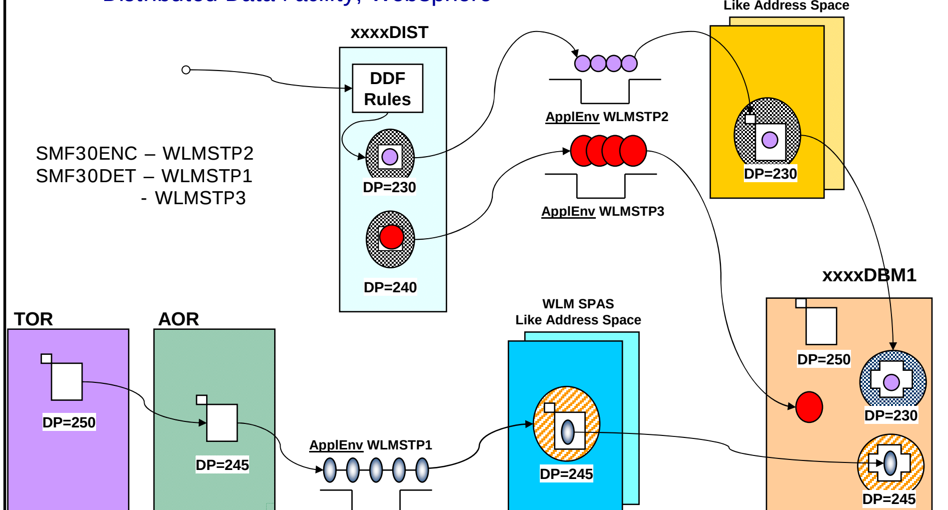
- When zIIP engines are configured SMF30ENC is the enclave CPU time that was qualified for zIIP but was prevented from running on zIIP
  - Ran on CP engines
- zIIP time is zIIP eligible time and not included in SMF30ENC
  - Time on zIIP + zIIP eligible but ran on CP





Stored Procedures After Enclaves

- Stored procedure client could be IMS, CICS, TSO, Batch, Started Tasks, Distributed Data Facility, WebSphere



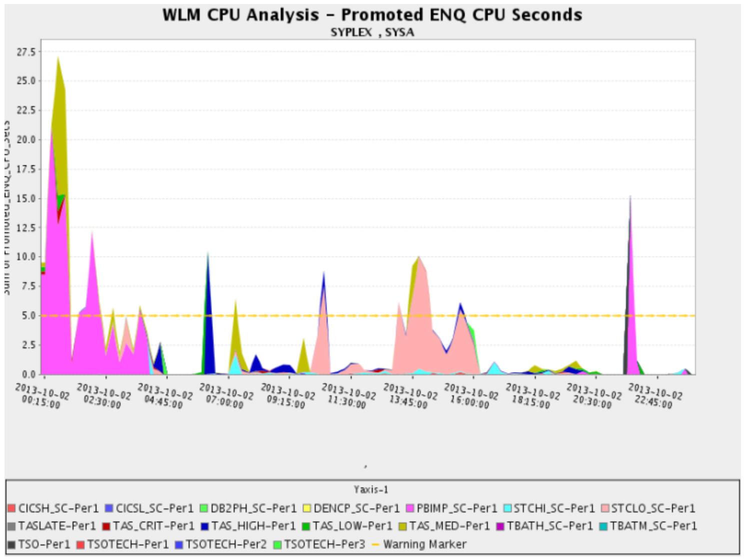
CPU Time for Promoted Work

How many CPU seconds did work consume while promoted?

- BLK
  - CPU seconds consumed while promoted to help blocked workloads
- ENQ
  - CPU seconds consumed while promoted by enqueue management because the work held a resource that other work needed.
- CRM
  - CPU seconds consumed while promoted by chronic resource contention management because the work held a resource that other work needed
- LCK (HiperDispatch mode only)
  - CPU seconds consumed while promoted to shorten the lock hold time of a local suspend lock held by the work unit
- SUP
  - CPU seconds consumed while the dispatching priority for a work unit was temporarily raised by the z/OS supervisor to a higher dispatching priority than assigned by WLM

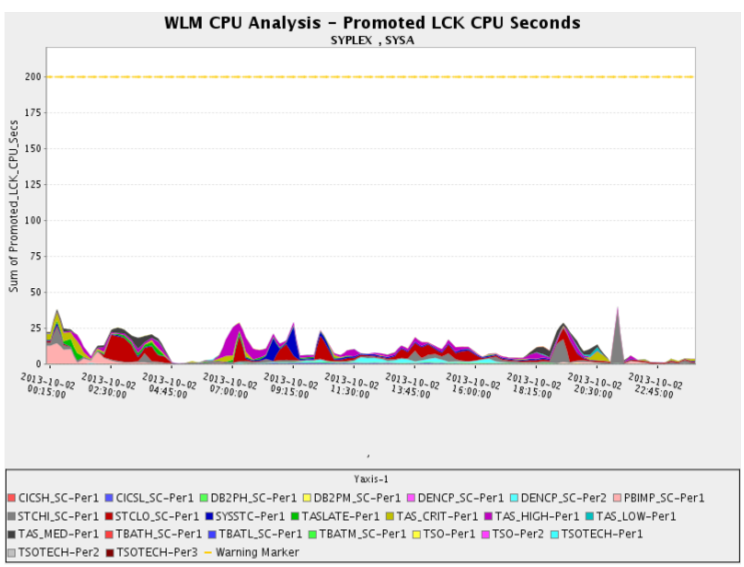
```
---SERVICE---   SERVICE TIME  ---APPL %---  --PROMOTED--
IOC    981747   CPU  873.159  CP     98.96  BLK    0.000
CPU     28865K  SRB   14.712  AAPCP   0.00  ENQ    0.000
MSO     89995K  RCT    0.081  IIPCP   0.00  CRM    0.000
SRB    486359   IIT    2.371                LCK    0.263
TOT    120328K  HST    0.408  AAP      N/A  SUP    0.000
/SEC   133694   AAP      N/A  IIP     0.01
                IIP    0.051
```

Promoted CPU Time by Service Class
- Promotion for Enqueue



Peter Enrico : www.epstrategies.com © Enterprise Performance Strategies, Inc. z/OS Processor Analysis - 55

Promoted CPU Time by Service Class
- Promotion for Lock



Peter Enrico : www.epstrategies.com © Enterprise Performance Strategies, Inc. z/OS Processor Analysis - 56

Performance Workshops Available

During these workshops you will be analyzing your own data!

- WLM Performance and Re-evaluating of Goals
  - Instructor: Peter Enrico
  - September 15 – 19, 2014 - Kansas City, Missouri, USA
  - October 20 - 24, 2014 - Munich, Germany
- Parallel Sysplex and z/OS Performance Tuning
  (Web / Internet Based!)
  - Instructor: Peter Enrico
  - July 29 – 31, 2014 (Web)
- Essential z/OS Performance Tuning Workshop
  - Instructors: Peter Enrico and Tom Beretvas
- z/OS Capacity Planning and Performance Analysis
  - Instructor: Ray Wicks

Peter Enrico : www.epstrategies.com