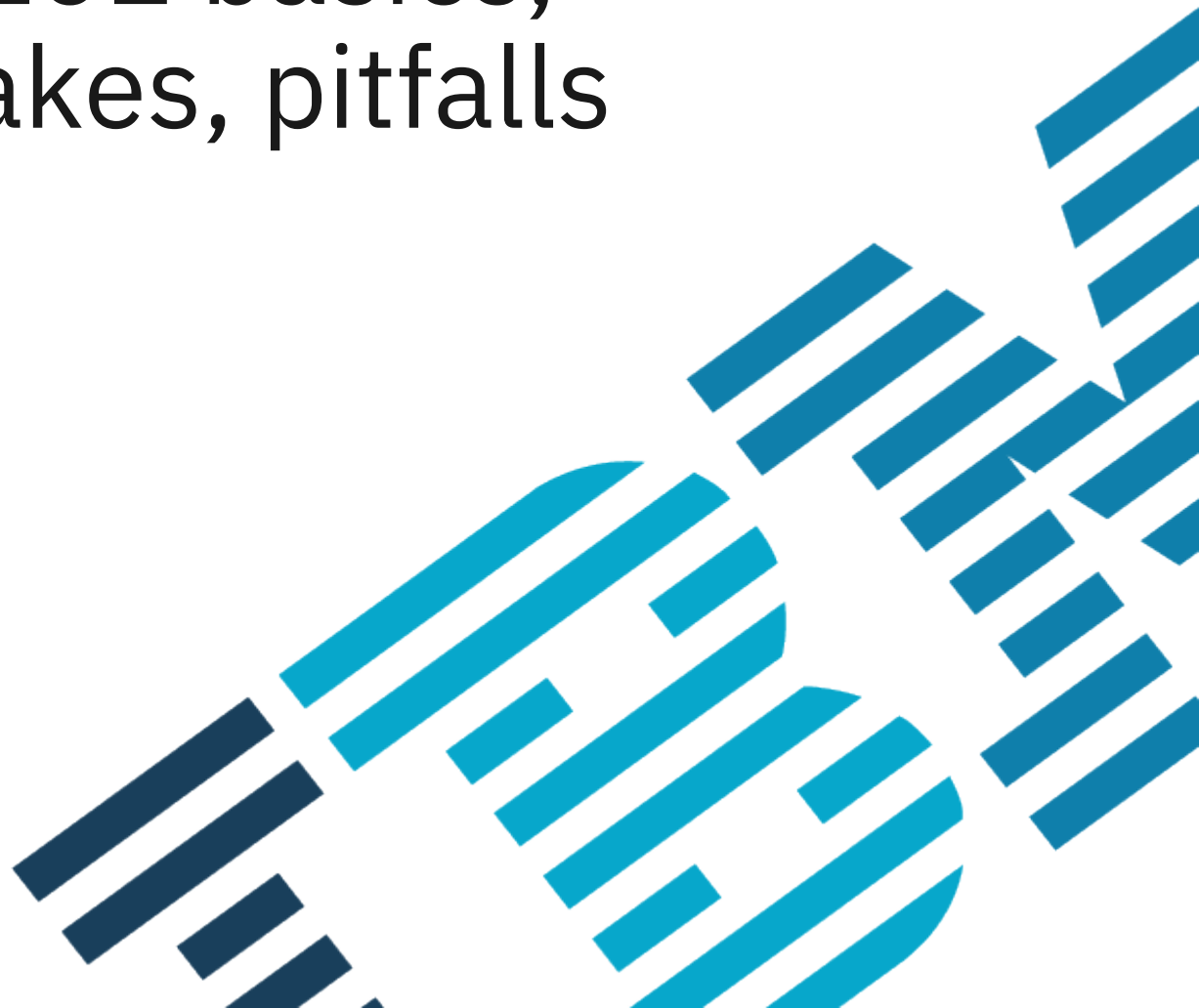
Db2 z/OS meets WLM - 101 basics, stories of common mistakes, pitfalls and failures



New England Users Group, 21 March 2024

Michał Białecki, michal.bialecki@pl.ibm.com
IBM Db2 z/OS SWAT team

Agenda

- **WLM policy setup for Db2 z/OS**
- **Common problems shown by examples**

Common problems

- **Systems are run at very high CPU utilisation for elongated periods of time**
- **Little or no non-Db2 work to pre-empt when the system becomes 100% busy i.e., no non-Db2 work that can be sacrificed**
- **Sporadic slowdowns or hangs of a Db2 subsystem or an entire Db2 data sharing group as a result of poorly managed WLM policy settings**
  - Db2 applications workloads are allowed to ‘fall asleep’
    - If a thread is starved while holding a major Db2 internal latch or other vital resource, this can cause an entire Db2 subsystem or data sharing group to stall
      - ✓ Especially if other work running on the LPAR has latent demand for CPU
  - Db2 system address spaces are not protected from being pre-empted under severe load
    - Any delays in critical system tasks as a result of a Db2 system address space being pre-empted can lead to slowdowns and potential ‘sympathy sickness’ across the data sharing group

WLM policy setup for Db2 z/OS

WLM policy setup

- **Recommendation**
  - Configure the WLM policy defensively to deal with situations when the system is over-committed
- **General guidelines**
  - VTAM, IRLM, and RRS must be mapped to Service Class SYSSTC
    - not classified Started Tasks run in SYSSTC by default, so be sure to set rules for all other STCs
  - All Db2 system address spaces should be isolated into a unique user-defined Service Class defined with Importance 1 and a very high velocity goal (e.g. 85-90)
    - MSTR, DBM1, DIST, WLM-managed stored procedure address spaces
    - **Set CPU CRITICAL** for Db2 system Address Spaces to protect against stealing of CPU from lower priority work
    - Do not generally recommend putting Db2 Address Spaces into SYSSTC because of risk of Db2 misclassifying incoming work
      - ✓ Special case: recurring Db2 slowdowns when running at very high CPU utilisation / elongated periods to help with Problem Determination

WLM policy setup …

Why protect Db2 system address spaces from being pre-empted?

- **Answer: These address spaces are critical for efficient system operation and should be defined with an aggressive goal and a very high importance**
  - MSTR contains the Db2 system monitor task
    - Requires an aggressive WLM goal so it can monitor CPU stalls and virtual storage constraints
  - DBM1 manages Db2 threads and is critical for both local Db2 latch and cross-system locking negotiation
    - Any delay in negotiating a critical system or application resource (e.g. P-lock on a space map page) can lead to a slowdown of the whole Db2 data sharing group
  - DIST and WLM-managed stored procedure AS only run the Db2 service tasks i.e. work performed for Db2 not attributable to a single user
    - Classification of incoming workload, scheduling of external stored procedures, etc.
    - Typically means these address spaces place a minimal CPU load on the system
      - ✓ BUT… they do require minimal CPU delay to ensure good system wide performance
    - DDF and/or SP workloads are controlled by the WLM Service Class definitions for the DDF enclave workloads or the other workloads calling the SP and not “inherit” service class from DIST/WLM address spaces.

WLM policy setup ...

- **General guidelines ...**
  - Reserve Importance 1 for critical server address spaces:
    - Db2 Address Spaces, IMS control region, CICS TOR
  - Use Importance 2-5 for all business applications starting in Importance 2 for high priority ones
  - Try to have non-Db2 work ready to be pre-empted, with lower importance
    - classify for important / less-important workload
  - Remove Discretionary definitions for any Db2-related work
    - unclassified jobs run in SYSOTHER (DISCRETIONARY) by default, make sure you classify all jobs
  - Do not use WLM resource group capping (e.g. to control CPU usage for batch workloads)
    - Very risky if caps are aggressive and workloads accessing Db2 data get penalised
    - If workload will stall on important resource latch it may block whole Db2 and other workloads/members, eg:
      - ✓ page split on index latch will queue other splits on same index pageset
      - ✓ Log output buffer latch will stop writing to logs, etc

WLM policy setup ...

- **General guidelines ...**
  - For CICS, IMS and DDF, favour 80th or 90th percentile Response Time goals and not Average Response Time goals:
    - Transactions are typically non-uniform
    - Response time goals must be practically achievable
      - ✓ Use RMF Workload Activity Report to validate
  - Online workloads that share Db2 objects should have similar WLM performance goals to prevent interference slowdowns
    - Example: An online DDF enclave-based workload classified much lower in importance than an online CICS workload → If they share any of the Db2 objects, a low running DDF enclave may end up causing the CICS online workload to become stalled
  - Online transactions shall have achievable goal and Performance Index (PI) in ranges of 0.81-1.2
    - If they overachieve (PI = 0.5) and if there will be workload spike, WLM will not react quick enough, till PI > 1
    - PI of 0.5 is lowest possible value for Response Time goal, and may mean far more over-achieved goal
      - ✓ we can’t say as there is one bucket for PI =< 0.5
    - PI of 4 is the highest possible value for Response Time goal, and may mean, far more under-achieved goal
      - ✓ we can’t say as there is one bucket for PI >= 4

WLM policy setup …

- **General guidelines …**
  - For BATCH workload give enough Importance and Execution Velocity not to exceed SLAs / deadline – usually Online day start (eg 8am)
  - Don’t use more than 35 Service Class periods - recommendation by IBM Washington Systems Center:
    - Every 10 seconds WLM can adjust dispatching priority ONLY one most suffering Service Class
    - It may take a while to get to all suffering Service Classes for WLM to make priority adjustments if there are too many
  - Starting with z/OS V2.5, the IBM service coefficients will be the hardcoded values to CPU=1, MSO=0, IOC=0 and SRB=1 and can no longer be modified.
    - Did you recalculate all Service Classes definition periods before migration to v2.5+?
  - Actively monitor / review how the goals are met
    - if not met, adjustments are needed or more capacity ?
    - if not met due to transaction nature, such goal may need to be loosen to be achievable/realistic
      - ✓ eg 99% Response time goal is normally unrealistic
    - WLM will abandon to help “hopeless” service classes
  - Enable defensive mechanisms to help with stalled workloads
    - -DISPLAY THREAD(*) SERVICE(WAIT) command
    - z/OS WLM Blocked Workload Support

-DIS THREAD(*) SERVICE(WAIT) SCOPE(GROUP)

- **Identifies allied agents and distributed DBATs that have been suspended for more than x seconds**
  - x = MAX(60, 2x IRLM resource timeout interval)
- **If the thread is suspended due to IRLM resource contention or Db2 latch contention, additional information is displayed to help identify the problem**

```
19:23:01.31  STC42353   DSNV401I  -DT45 DISPLAY THREAD REPORT FOLLOWS -
                        DSNV402I  -DT45 ACTIVE THREADS -
                        NAME     ST A   REQ ID           AUTHID   PLAN     ASID TOKEN
                        CT45A    T  * 12635 ENTRU1030014 ADMF010  REPN603  00C6 90847
                        V490-SUSPENDED 04061-19:20:04.62 DSNTLSUS +00000546 14.21
```

- **Note that DISTSERV may return false positives for DBATs that have not been in use for more than the x seconds**

-DIS THREAD(*) SERVICE(WAIT) SCOPE(GROUP)

- **Dynamically boost priority for any latch holder that appears to be stuck → targeted boost via WLM services**
- **Automatically driven every minute by the internal Db2 System Monitor**
  - Diagnostic info is written out to MSTR log and syslog (APAR PI29671 (2020))
- **Recommendation:**
  - Save away the output as diagnostics
    - Check DSNV507I message on -DISPLAY THREAD(*) TYPE(SYSTEM) output
    - And DSNV523I to determine which threads are being boosted

```
V507-ACTIVE MONITOR, INTERVALS=1235, STG=47%, BOOSTS=1, HEALTH=100
     REGION=1633M, AVAIL=1521M, CUSHION=375M

LATCH WAITERS CAUSING A BOOST
  DSNV523I  -D3P7 DSNVMON - AGENT 1:  513
  NAME     ST A   REQ ID           AUTHID   PLAN     ASID TOKEN
  ----     -- -   --- --           ------   ----     ---- -----
  DB2A     N  *     0 020.TLPLKP1E SYSOPR            00E8     0
```

z/OS WLM Blocked Workload Support

- **Gives small amounts of CPU to stalled dispatchable units of work on the system ready queue**
  - Not specific to Db2 workloads
  - Allows even frozen discretionary jobs to get some small amount of CPU
  - **But will not help jobs that are stalled because of WLM resource group capping**
- **Controlled by two parameters in IEAOPTxx parmlib member**
  - BLWLINTHD – Threshold time interval for which a blocked address space or enclave must wait before being considered for promotion
    - Default: 20 seconds (may be too long)
  - BLWLTRPCT – How much of the CPU capacity on the LPAR is to be used to promote blocked workloads
    - Default: 5 (i.e. 0.5%)

z/OS WLM Blocked Workload Support …

- **Recommendations**
  - All customers should run with this function ENABLED
  - Changing BLWLINTHD to 2-5 seconds may provide better overall system throughput at very high CPU utilisation
  - Regularly review the statistics on this function provided in RMF CPU Activity and Workload Activity reports :

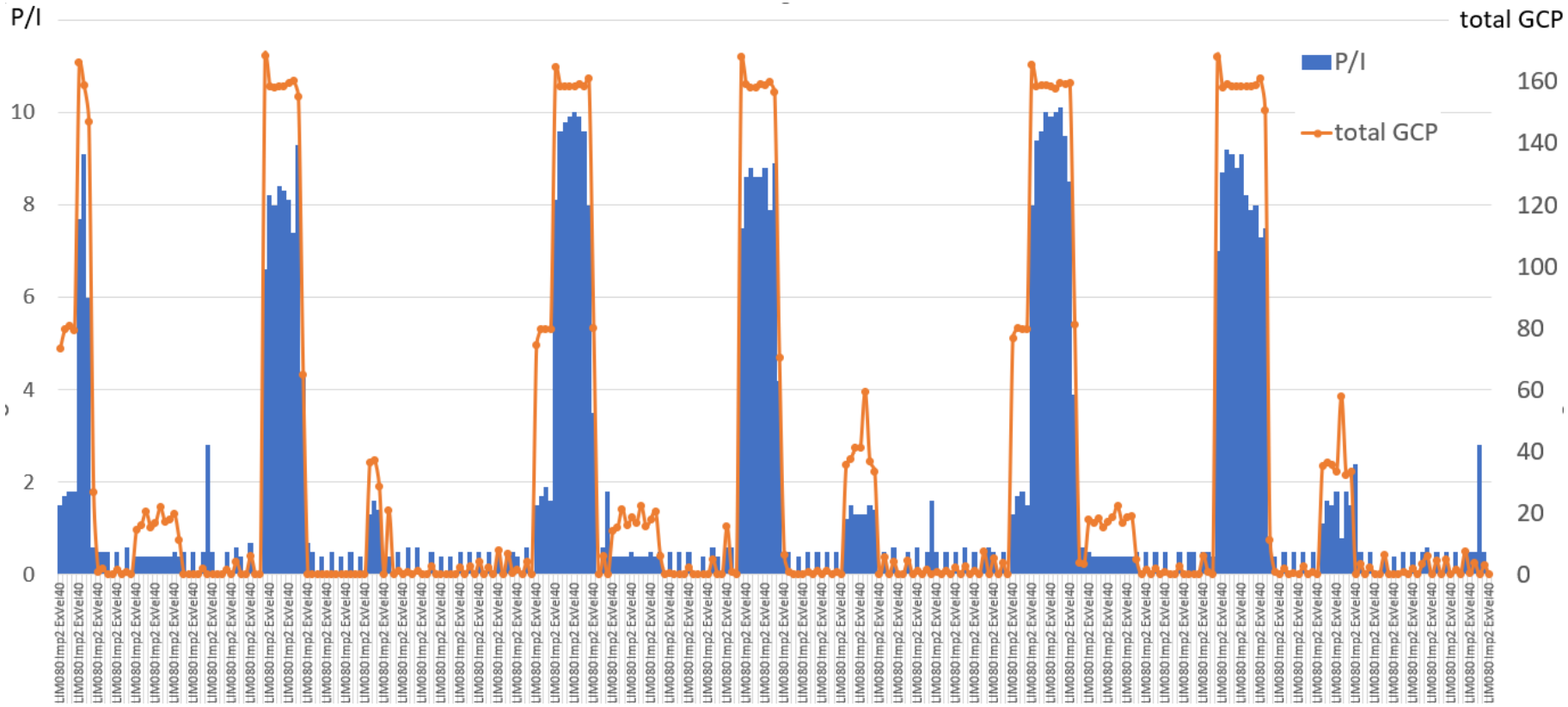
```
BLOCKED WORKLOAD ANALYSIS
OPT PARAMETERS: BLWLTRPCT (%)    0.5  PROMOTE RATE:  DEFINED      13  WAITERS FOR PROMOTE:  AVG 0.010
                BLWLINTHD         20                 USED (%)     4                        PEAK    1


                                SERVICE CLASS=BATLOW      PERIOD=1 IMPORTANCE=5
---SERVICE---   SERVICE TIME  ---APPL %---  --PROMOTED--  ----STORAGE----
IOC     60439K  CPU 11303.13  CP    294.32  BLK    7.498  AVG    17647.61
CPU    664890K  SRB  176.656  AAPCP   0.00  ENQ    0.000  TOTAL  359251.4
MSO          0  RCT    0.715  IIPCP   2.68  CRM    0.000  SHARED   208.28
SRB     10392K  IIT   73.298                LCK   46.300
TOT    735720K  HST    0.013  AAP      N/A  SUP    0.000  -PAGE-IN RATES-
/SEC   204367   AAP      N/A  IIP    26.62                 SINGLE      0.0
                IIP  958.208                               BLOCK       0.0
```

Common problems shown by examples

WLM Resource Group Capping

- **Definition example: Resource Group - LIMIT 80% of an engine in LPAR (2x)**



TAPE archive/migration started task effectively was capped / slowed down

May be causing storage pool overflow and DB2 logs not being migrated, hold of writing logs / can stop Db2 workload

WLM CPUCRIT / CPU Critical attribute

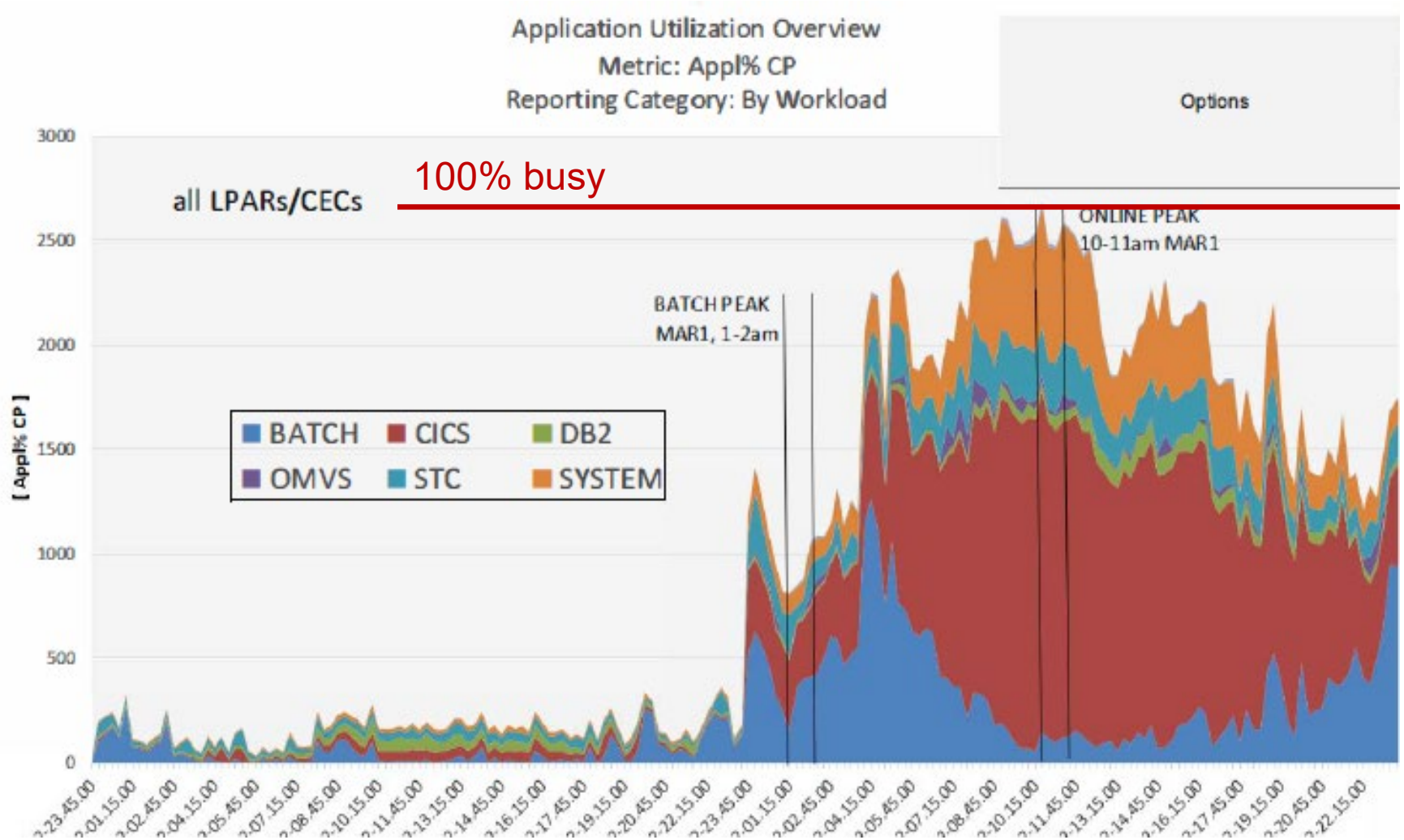
- **CPU Critical only protects that work from lower importance work, no protection from work at same or higher importance**
- **What is wrong here:**

Service Class	Imp	goal	CPUCRIT	description
DB2ASIDS	1	ExVel80	YES	DB2 DBM1, MSTR, DIST
DDFHIGH	1	RspTime90 0.06	YES	hot DDF queries
CICSHIGH	1	RspTime85 0.3		hot CICS transactions

1. Transaction workload (DDF queries/CICS tx) Service Classes shall be lower importance than Db2 address space Service Class.
2. CPUCRIT=YES for DB2ASIDS Service Class would protect only from lower importance classes, not from equal importance classes
3. Transactions Service Classes DDFHIGH and CICSHIGH when not meeting PI, WLM can set higher dispatching priority than DB2ASIDS, and steal CPU from Db2 and slow down Db2 and those transactions will suffer even more.

WLM service execution review

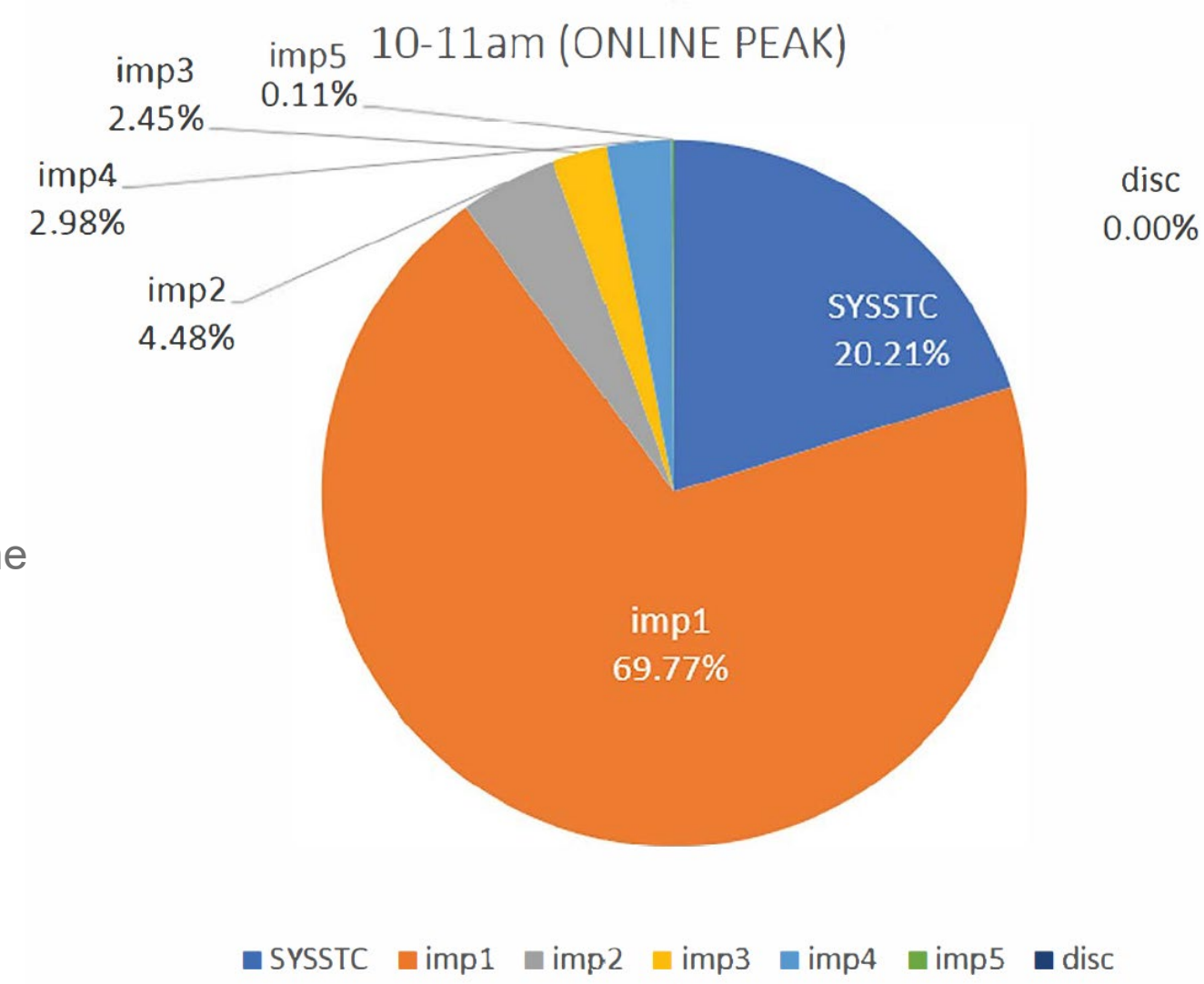
- **Know how your workload looks like, and what are major contributors, with peak days, importance's/vel goals**



WLM service execution review

What's Wrong

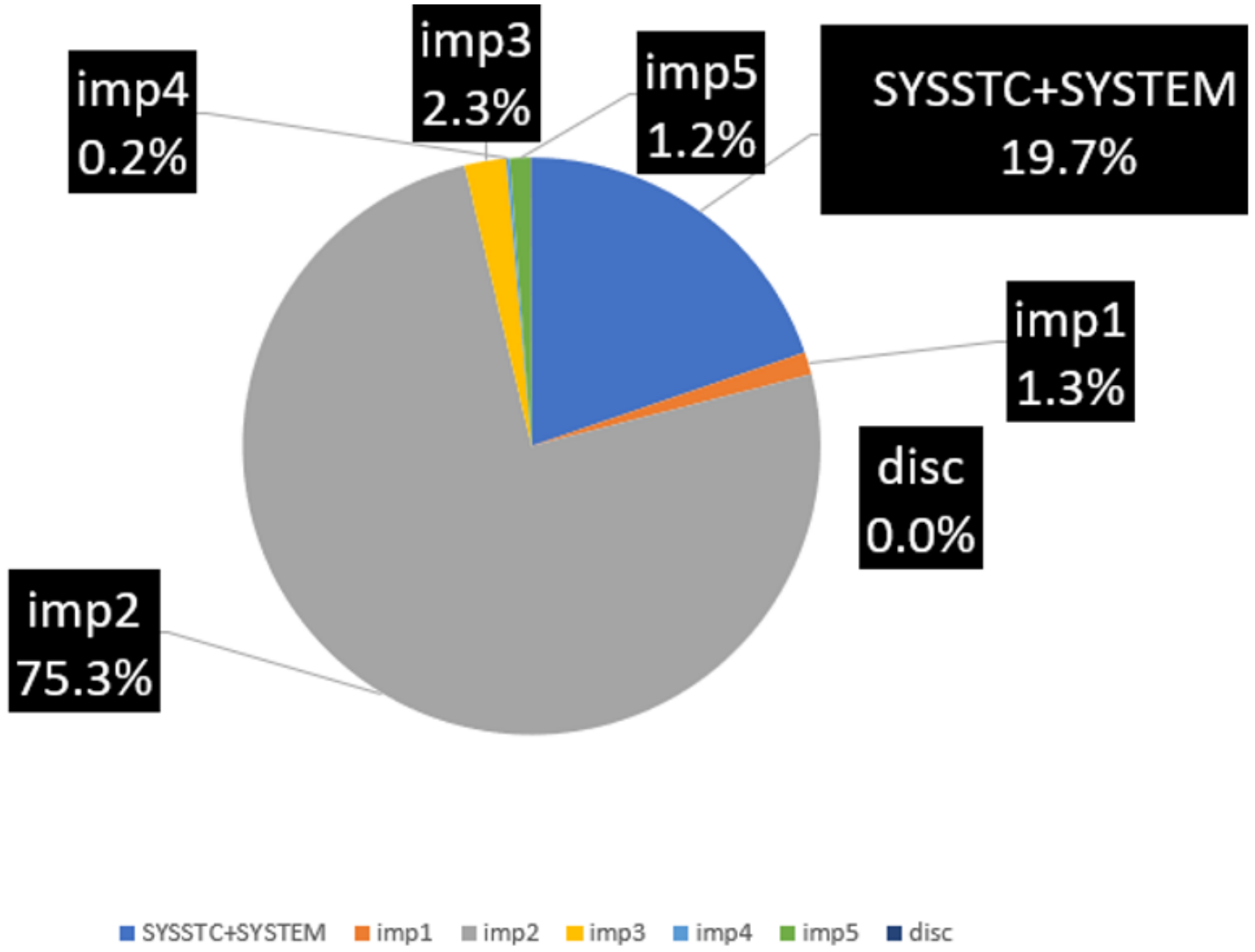
- 90% of workload is SYSSTC and Importance 1
- not much to pre-empt if there is increased workload demand
- almost everything is of equal importance, no one can sacrifice here



WLM service execution review …

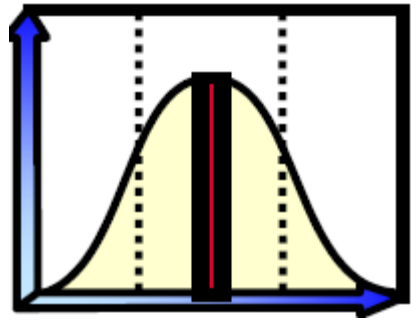
- **Importance1+SYSSTC is 20%.**
  So it looks better now ?

Not really – importance1 is tiny - almost not used and again 95% of workload is SYSSTC and importance 2, not much to pre-empt

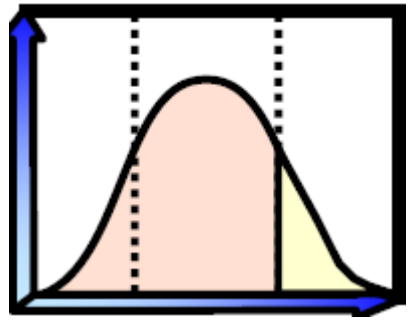


Response Time goals

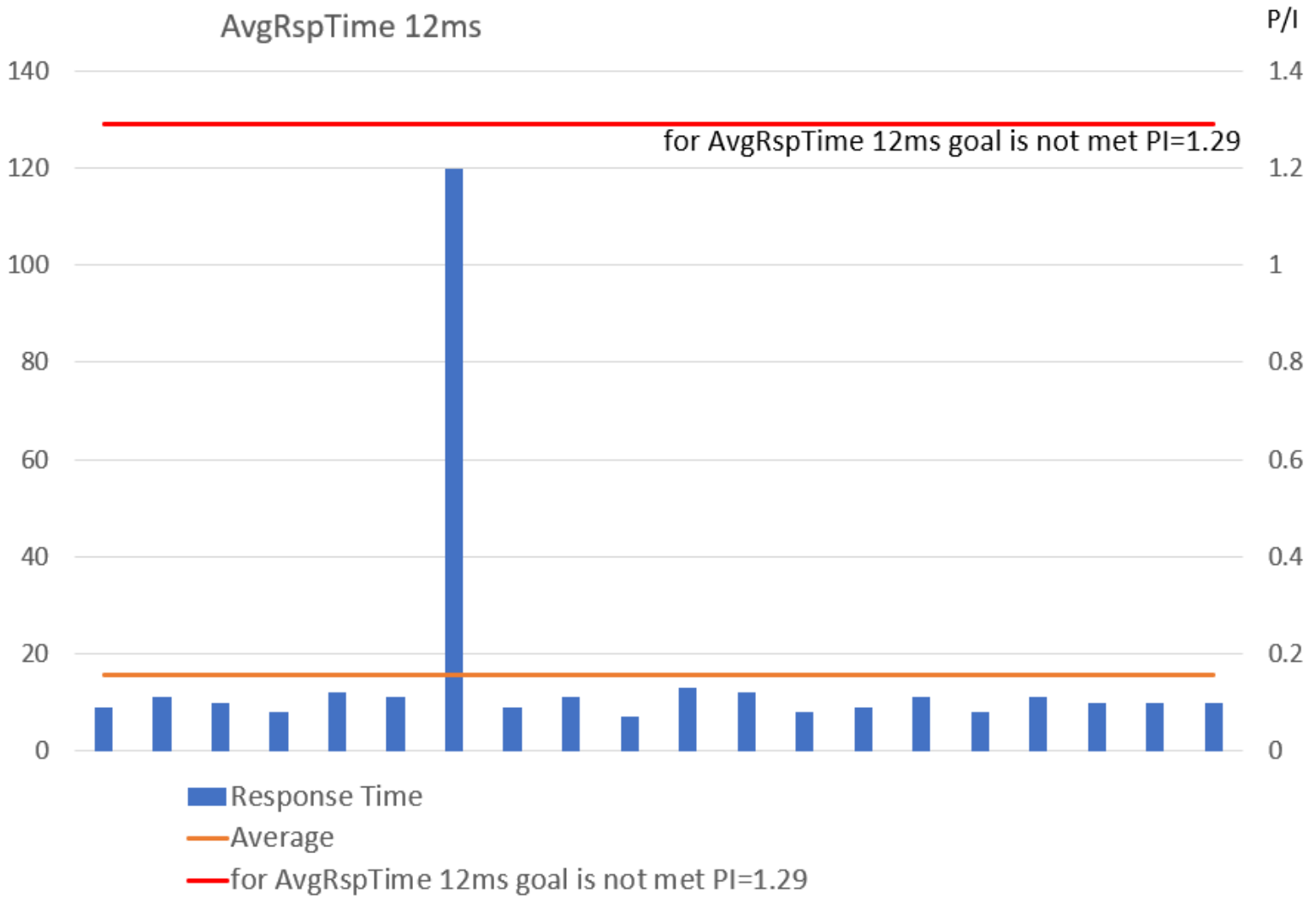
- **Average Response Time goal (AvgRspTime)**
  - Defines the average transaction response time for all ended transactions
    = Sum of elapsed time for ended transactions / Number of ended transactions
  - Example: Average response time = 1 second



- **Percentile Response Time goal (RspTime)**
  - Defines the number of transactions ending with a response time lower than or equal to the time value
    = Number of transactions ended with time <= goal / Number of ended transactions
  - Example: goal = 90% in less than 1 sec, we don't care about rest of transactions

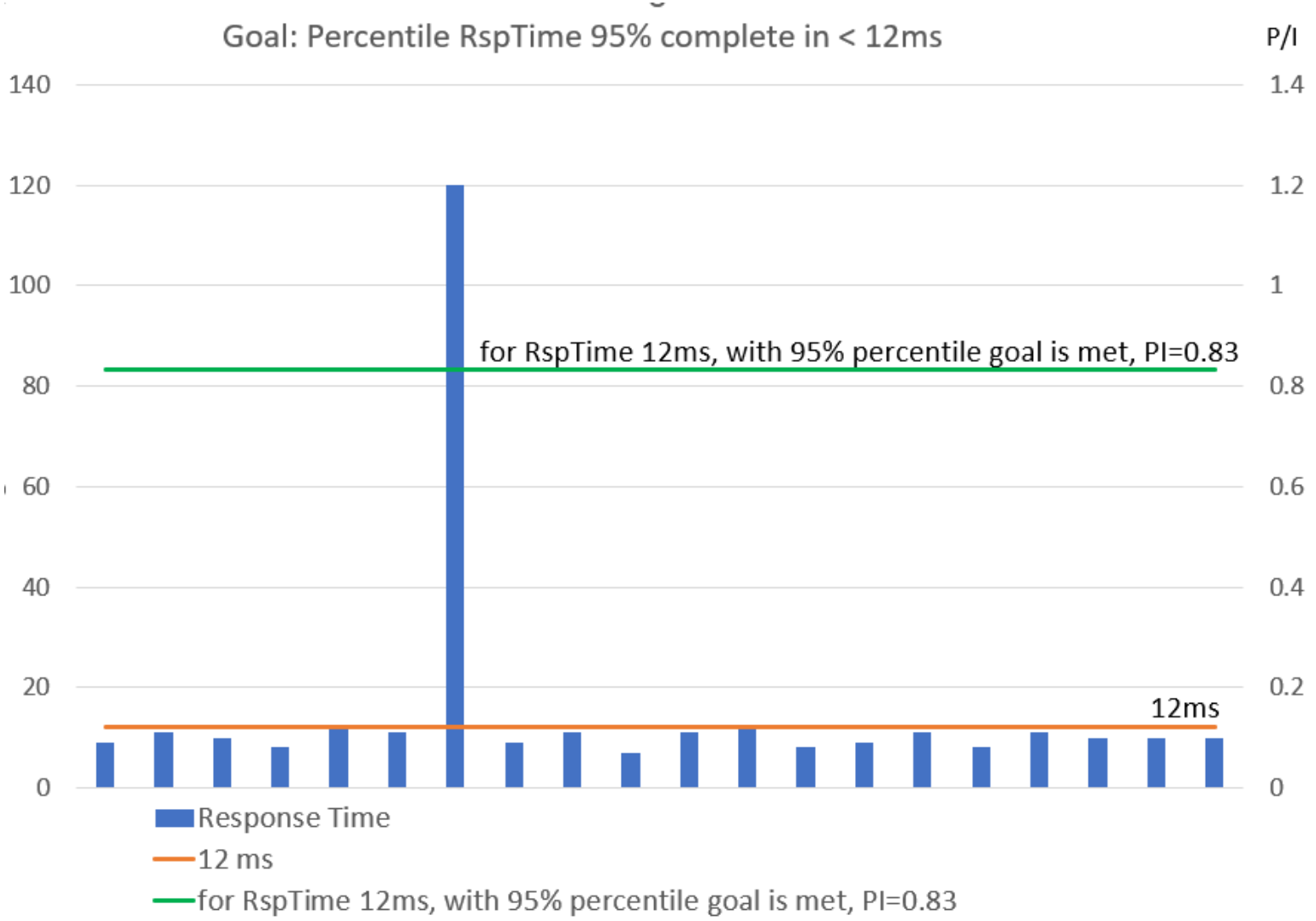


Average Response Time goals



- Service class defined as Average response time of 00:00:00.012, not meeting PI due to **ONE** outlier transaction,
- WLM will try to get more CPU to this class also for transactions that do not need it. This class will be given CPU as goal is missed (in reality 95% completes OK)
- This goal is fine ONLY if transactions are fairly equal – so there are no outliers
  - very rare case

Percentile Response Time goals



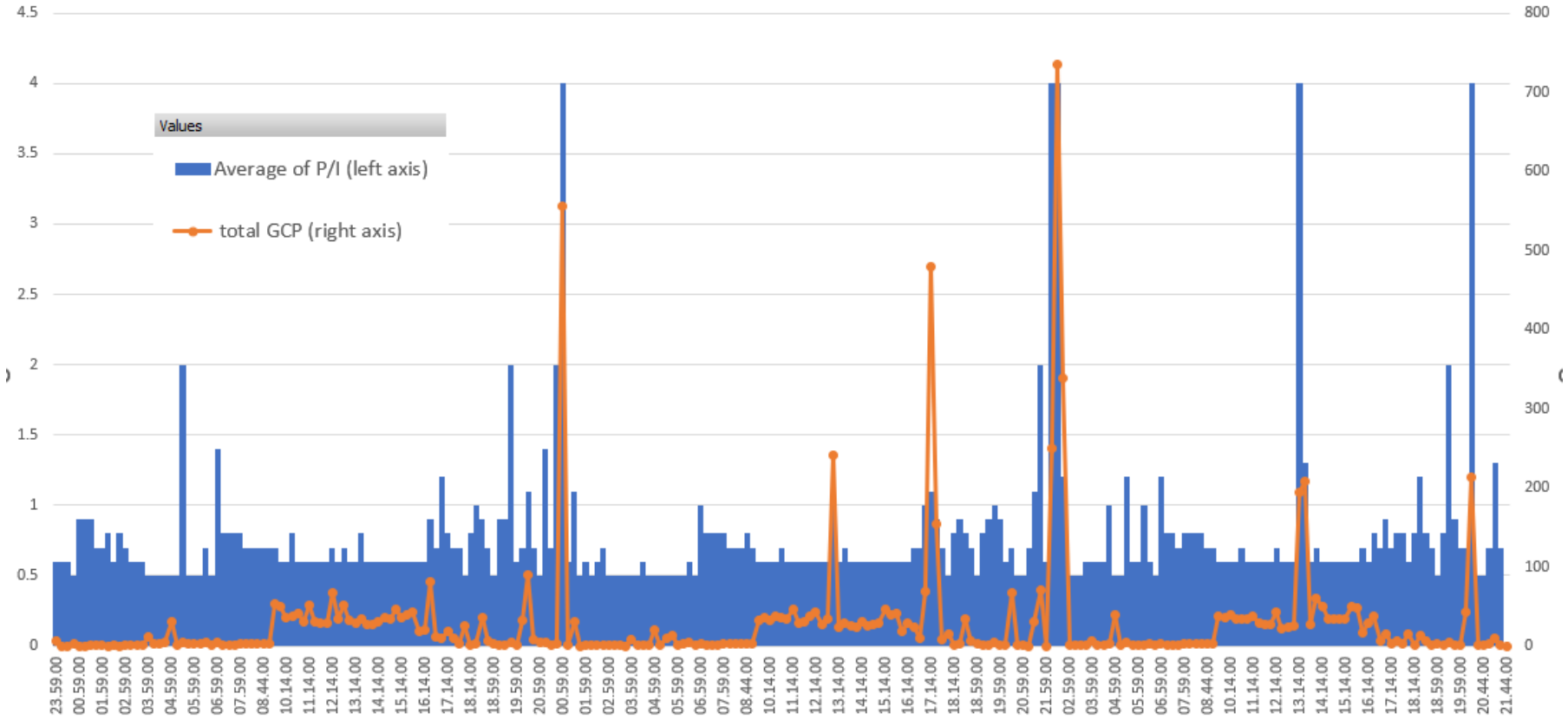
What's Wrong

Nothing wrong

- WLM ignores this outlier
- PI is met, no priority adjustments done by WLM

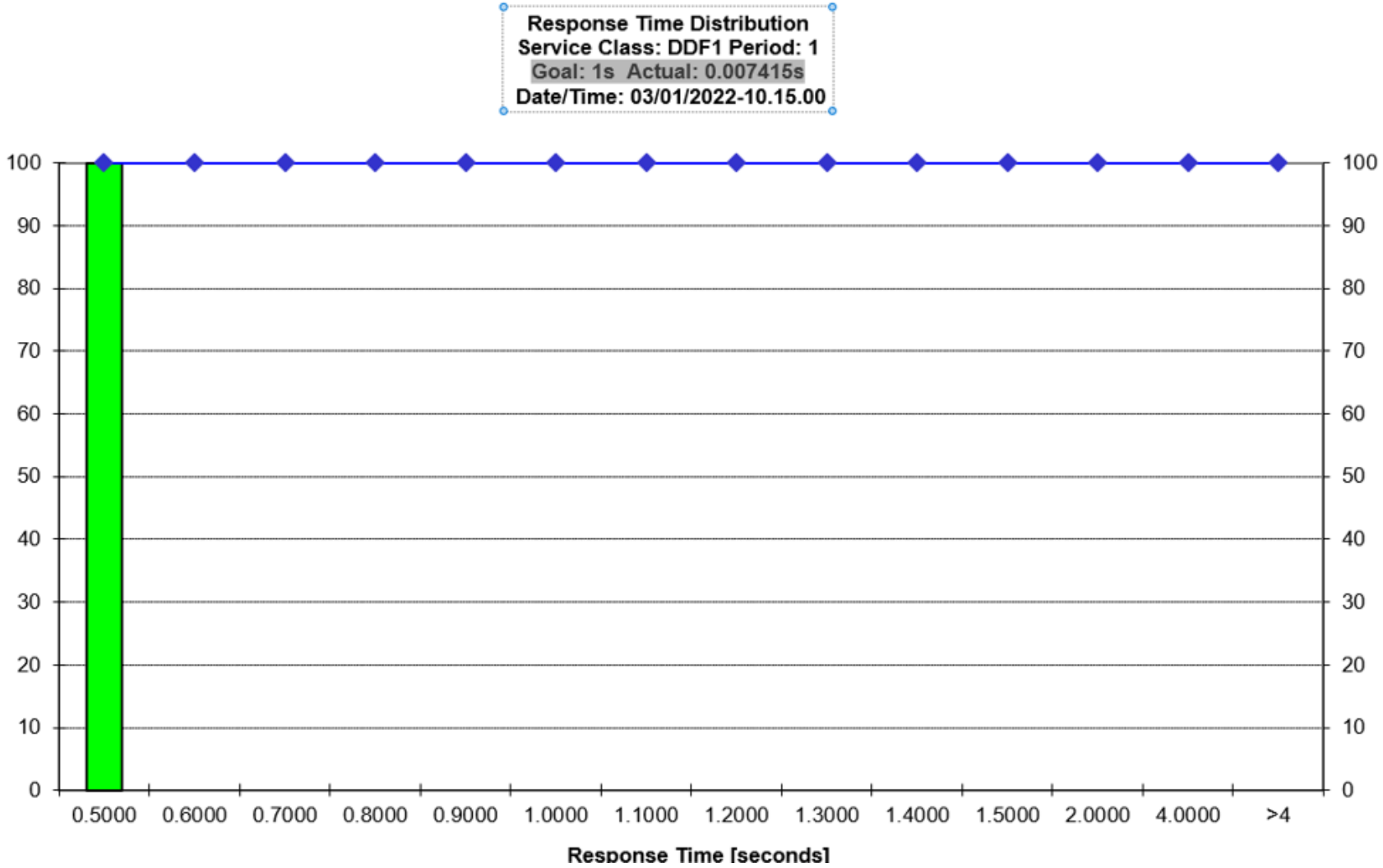
Percentile Response Time goals

- **If >90% of transactions complete in less than ½ of their goal, the goal should be adjusted tighter**
  - If Performance Index is <0.81 WLM will not pay attention ➔ ideally PI should be ~1.2 at peak processing
  - Avoid violent swings in response time under CPU constraint, allowing WLM to react more quickly



- PI is too loose (<0.81)
- Response time suffers in periods with increased workload

Percentile Response Time goals ...

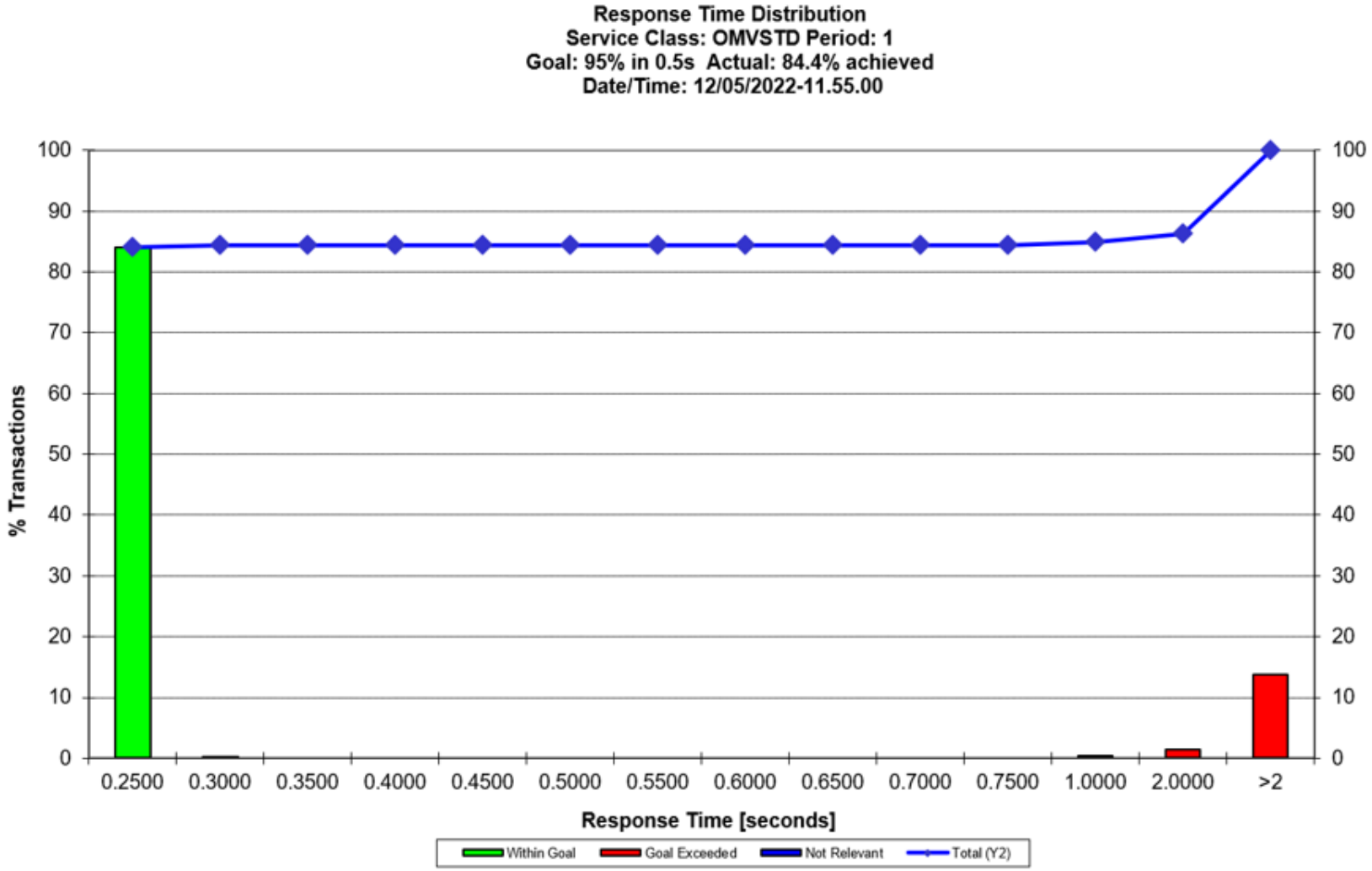


What's Wrong

- The transactions are finishing in 7.4 milliseconds
- Defined goal of 1 second is way too loose
- goal shall be set to 1.2*7.4 = 10 milliseconds so the service level can be maintained
- The WLM goals should align with the reasonable business goals / agreed SLAs

Percentile Response Time goals …

- **Are all transactions fitting well here to this goal?**



Goal is not met for 16% of transactions.

Solution:

- Response percentile shall be lowered,

OR

- added 2nd period for outliers

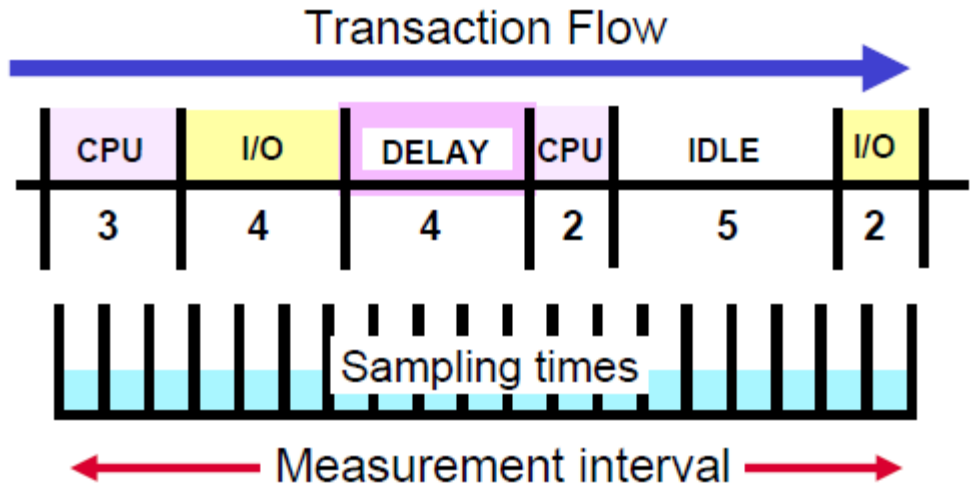
OR

- outliers shall be in separate Service Class

Velocity goals

- **Velocity goal is a measure of acceptable delay (percentage)**

  = (CPU Using + I/O Using) x 100 / (CPU Using + I/O Using + CPU Delay + I/O Delay + Paging Delay + MPL Delay + A/S Delay)



Execution Velocity =
(3+4+2+2) / (3+4+**4**+2+2) *100% = 73%

- **request to WLM that states a percentage of 'time' when the work wants to run, it is able to run, and it is not delayed for lack of WLM managed resources**
- **measured by resources WLM can control (eg does not control IDLE time, waiting for user input)**

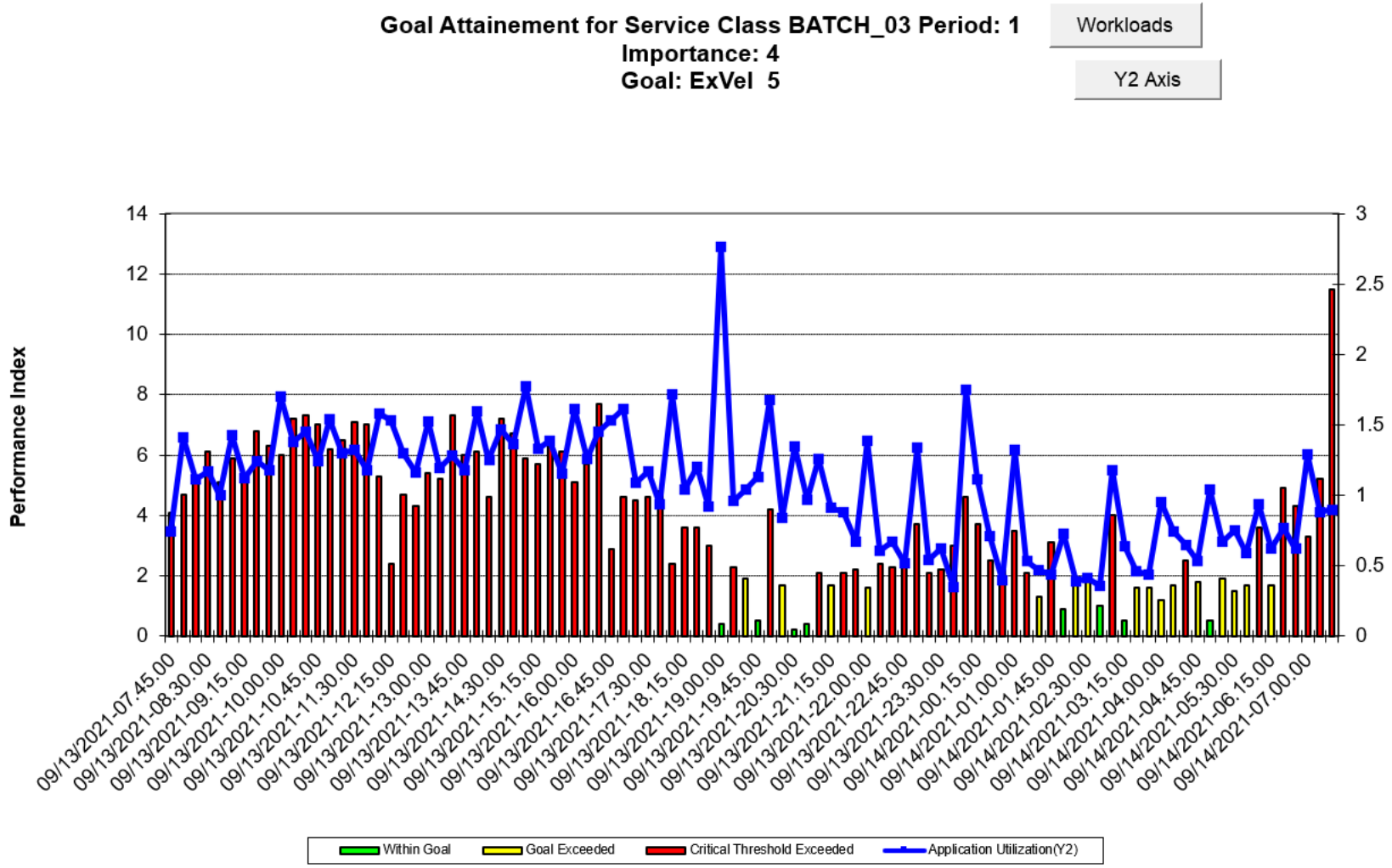
Velocity goals for online transactions

				Goal								
Service Class	Per	Dur	Imp	Type	Pct	Value	ResGrp	CPU Crit	I/O Prior	# Rule	Printed	Comment
SDDF	1	12000	2	ExVel	70			NO	NORMAL	9	Execution velocity of 70	DDF work
SDDF	2	13000	3	ExVel	60			NO	NORMAL	9	Execution velocity of 60	DDF work
SDDF	3		4	ExVel	50			NO	NORMAL	9	Execution velocity of 50	DDF work

- **If there are online transactions running when the CPU is constrained their response time will increase dramatically but WLM may not act at all**
  - velocity goal may still be met, and dispatching priority not increased
- **Hard to match SLA response time for online transactions with Velocity goal**
- **Exception is for HPDBATs – then velocity goal is the only reasonable choice**
  - A high-performance DBAT is a database access thread **that stays** associated with a remote connection at transaction boundaries, rather than being pooled:
    - effective response time of the enclave will be longer than the response time to process a single transaction when using high performance DBATs
      - ✓ we have many transactions seen as ONE transaction by WLM
    - Reference https://www.ibm.com/support/pages/apar/PH34080
  - PH41024 is changing reporting part provided to WLM (OA61811) / RMF about number of transactions and their response times use in enclave by Db2 https://www.ibm.com/support/pages/apar/PH41024

Reviewing how goals are met / not met

- **It is crucial to review if goals are realistically set and met**



- Very low velocity goal is almost never met
- If this is Db2 workload, may have some delays on resources that are accessed by other higher importance workload
- It may be ok for some “sacrificial” batch workload, eg reporting, not accessing “online/transactional data”

Heterogenous Report Classes

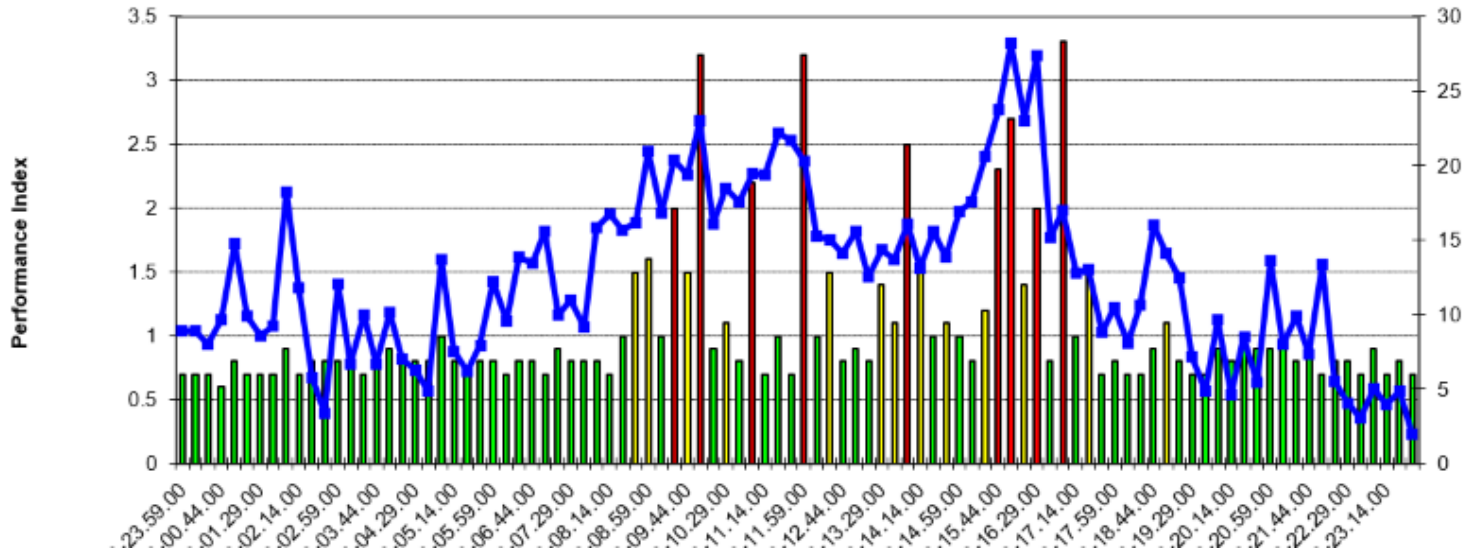
Classification Rules

	Qualifier					
Subsys	Level	Type	Name	Pos	Service Class	Report Class
STC	1	TN	DB%SPAS		STCMED	STCDB2
STC	1	TN	DB%%IRLM			STCDB2
STC	1	TN	DB%%MSTR		STCHI	STCDB2

Goal Attainement for Report Class STCDB2 Period: 1

Workloads

Y2 Axis



- Heterogeneous report classes can cause incorrect performance data, since the data collected is based on different goals, importance, or duration.

Reference:
https://www.ibm.com/docs/en/zos/2.2.0?topic=management-defining-report-classes

Velocity goals with multiple periods

What's Wrong

Service Class	Per	Dur	Imp	Type	Pct	Value	ResGrp	CPU Crit	I/O Prior	# Rule	Printed
BATCH_00	1		3	ExVel	65			NO	NORMAL	0	Execution velocity of 65
BATCH_01	1		4	ExVel	40			NO	NORMAL	5	Execution velocity of 40
BATCH_02	1		5	ExVel	30			NO	NORMAL	2	Execution velocity of 30
BATCH_03	1	75	4	ExVel	5			NO	NORMAL	24	Execution velocity of 5
BATCH_03	2			Disc				NO	NORMAL	24	Discretionary

```
POLICY=DAY    WORKLOAD=BATCH      SERVICE CLASS=BATCH_03     RESOURCE GROUP=*NONE       PERIOD=1 IMPORTANCE=4
-TRANSACTIONS--  TRANS-TIME HHH.MM.SS.FFFFFF  TRANS-APPL%-----CP-IIPCP/AAPCP-IIP/AAP  ---ENCLAVES---
 AVG         0.04  ACTUAL              3.349320  TOTAL          1.01        0.00    0.00  AVG ENC    0.00
 MPL         0.04  EXECUTION             707892  MOBILE         0.00        0.00    0.00  REM ENC    0.00
 ENDED        112  QUEUED              2.613340  CATEGORYA      0.00        0.00    0.00  MS ENC     0.00

POLICY=DAY    WORKLOAD=BATCH      SERVICE CLASS=BATCH_03     RESOURCE GROUP=*NONE       PERIOD=2 IMPORTANCE=DISC
-TRANSACTIONS--  TRANS-TIME HHH.MM.SS.FFFFFF  TRANS-APPL%-----CP-IIPCP/AAPCP-IIP/AAP  ---ENCLAVES---
 AVG       461.49  ACTUAL           1.35.462215  TOTAL         10933        0.34  679.14  AVG ENC    0.00
 MPL       461.49  EXECUTION        1.32.005216  MOBILE         0.00        0.00    0.00  REM ENC    0.00
 ENDED       6449  QUEUED              3.316190  CATEGORYA      0.00        0.00    0.00  MS ENC     0.00
```

- BATCH_03:
  - Duration of Period 1 is 75 which is equal to 0.019 CP seconds on this machine.
  - most of BATCH_03 executes/finishes as discretionary in Period 2
  - Both periods shall be merged

DISCRETIONARY goal

What's Wrong

Service Class	Per	Dur	Imp	Type	Pct	Value	ResGrp	CPU Crit	I/O Prior	# Rule	Printed
BATCH_00	1		3	ExVel	65			NO	NORMAL	0	Execution velocity of 65
BATCH_01	1		4	ExVel	40			NO	NORMAL	5	Execution velocity of 40
BATCH_02	1		5	ExVel	30			NO	NORMAL	2	Execution velocity of 30
BATCH_03	1	75	4	ExVel	5			NO	NORMAL	24	Execution velocity of 5
BATCH_03	2			Disc				NO	NORMAL	24	Discretionary

```
POLICY=DAY     WORKLOAD=BATCH       SERVICE CLASS=BATCH_03     RESOURCE GROUP=*NONE       PERIOD=1 IMPORTANCE=4
-TRANSACTIONS--  TRANS-TIME HHH.MM.SS.FFFFFF  TRANS-APPL%-----CP-IIPCP/AAPCP-IIP/AAP  ---ENCLAVES---
 AVG         0.04  ACTUAL              3.349320  TOTAL          1.01        0.00    0.00  AVG ENC    0.00
 MPL         0.04  EXECUTION             707892  MOBILE         0.00        0.00    0.00  REM ENC    0.00
 ENDED        112  QUEUED              2.613340  CATEGORYA      0.00        0.00    0.00  MS ENC     0.00

POLICY=DAY     WORKLOAD=BATCH       SERVICE CLASS=BATCH_03     RESOURCE GROUP=*NONE       PERIOD=2 IMPORTANCE=DISC
-TRANSACTIONS--  TRANS-TIME HHH.MM.SS.FFFFFF  TRANS-APPL%-----CP-IIPCP/AAPCP-IIP/AAP  ---ENCLAVES---
 AVG       461.49  ACTUAL           1.35.462215  TOTAL         10933        0.34  679.14  AVG ENC    0.00
 MPL       461.49  EXECUTION        1.32.005216  MOBILE         0.00        0.00    0.00  REM ENC    0.00
 ENDED       6449  QUEUED              3.316190  CATEGORYA      0.00        0.00    0.00  MS ENC     0.00
----SERVICE----    SERVICE TIME  ---APPL %---   --PROMOTED--   --DASD I/O---   ----STORAGE----   -PAGE-IN RATES-
 IOC         1155M  CPU 100059.6  CP      11282  BLK     0.000  SSCHRT  488.3K AVG    173684.2  SINGLE       0.0
 CPU         6930M  SRB 4453.766  IIPCP    0.34  ENQ   540.018  RESP      0.6  TOTAL  80153880  BLOCK        0.0
 MSO            0   RCT    1.845  IIP    679.14  CRM     0.000  CONN      0.5  SHARED  5415.14  SHARED       0.0
```

- This Service Class runs on zIIPs **- a Discretionary zIIP eligible workload won't get help from General CPs, even if IIPHONORPRIORITY is set to YES**

Summary

- **WLM needs to be carefully planned to suit business needs / agreed SLAs**
- **Making all Service Classes high importance is not helpful, there have to be some lower Service Classes that can be pre-empted**
- **WLM is not one time setup, it shall follow workload changes.**
  - Execution monitoring results must be continuously reviewed and actioned by:
    - Re-adjusting the goals
    - Adding capacity

THANK YOU !

Speaker: Michal Bialecki

Company: IBM

Email Address: michal.bialecki@pl.ibm.com

Interested in more knowledge?

https://ibm.biz/masterclass2024