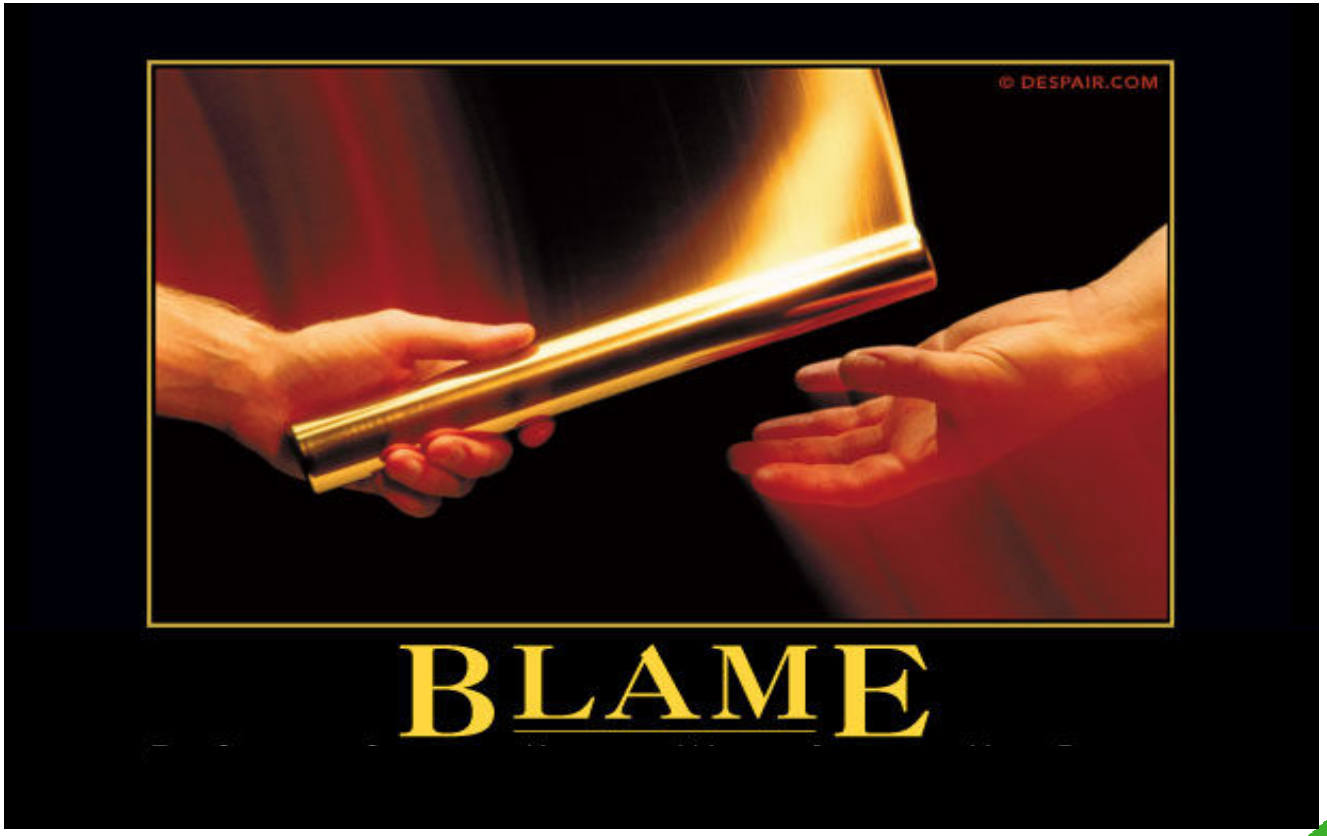
Avoiding Performance SRs



Adrian Burke
DB2 SWAT Team SVL
agburke@us.ibm.com

Agenda

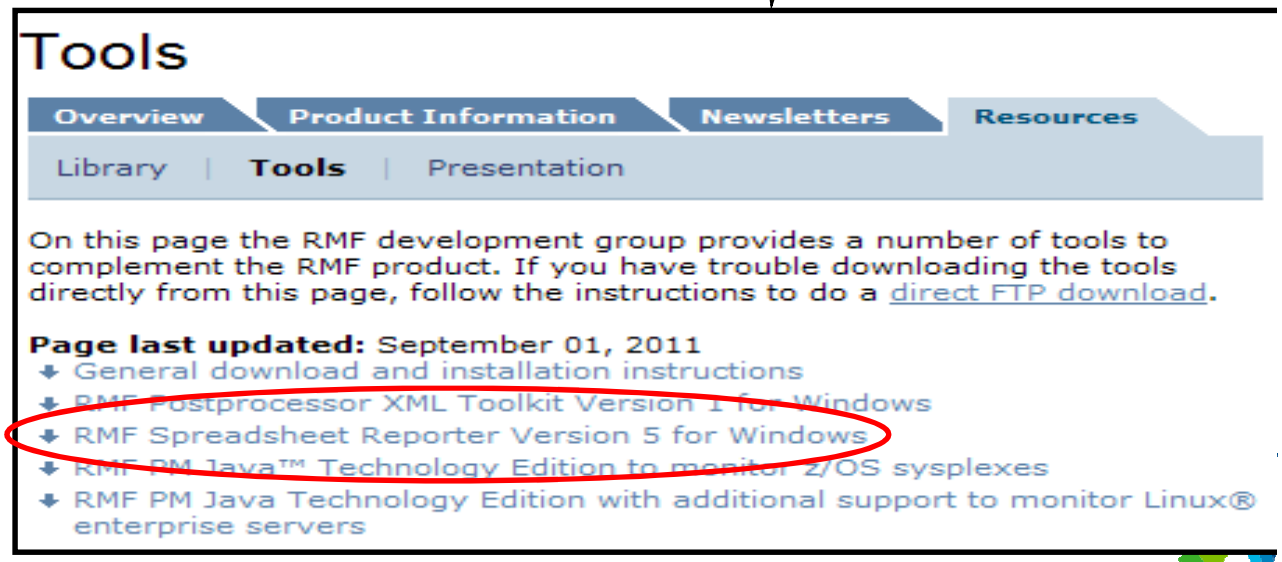
- Premise
- Overview of RMF and the Spreadsheet Reporter
  - As a means of visualizing z/OS performance problems
- CPU and WLM Concerns
- I/O and DASD Subsystem
- zIIP
- Real and Virtual Storage Analysis
- Class 3 Suspense Time
- Resources

Premise

- Holistic approach → top down = RMF → SMF → Traces/Dumps
  - If SWAT team or L2 performance team is involved this is where we start
  - Get the big picture then drive to root cause
  - Often problems are intermittent and require an understanding of the entire environment
- If there is a perceived DB2 subsystem, or data sharing group performance issue
  - Rule out Sysplex/ CF/ CEC/ LPAR constraints first
- If there is a workload or period of the day suffering
  - WLM/ CPU constraint/ DB2 internals
- If there is a single job, or group of transactions suffering
  - Object contention
  - Access path or DB2 component
  - Storage subsystem

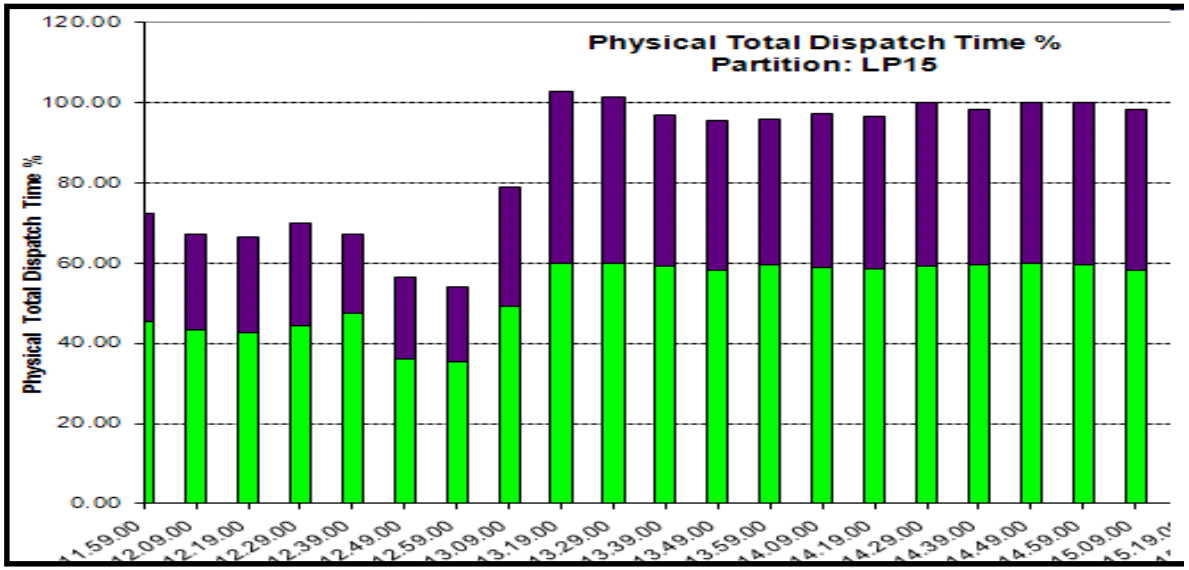
RMF Spreadsheet Reporter

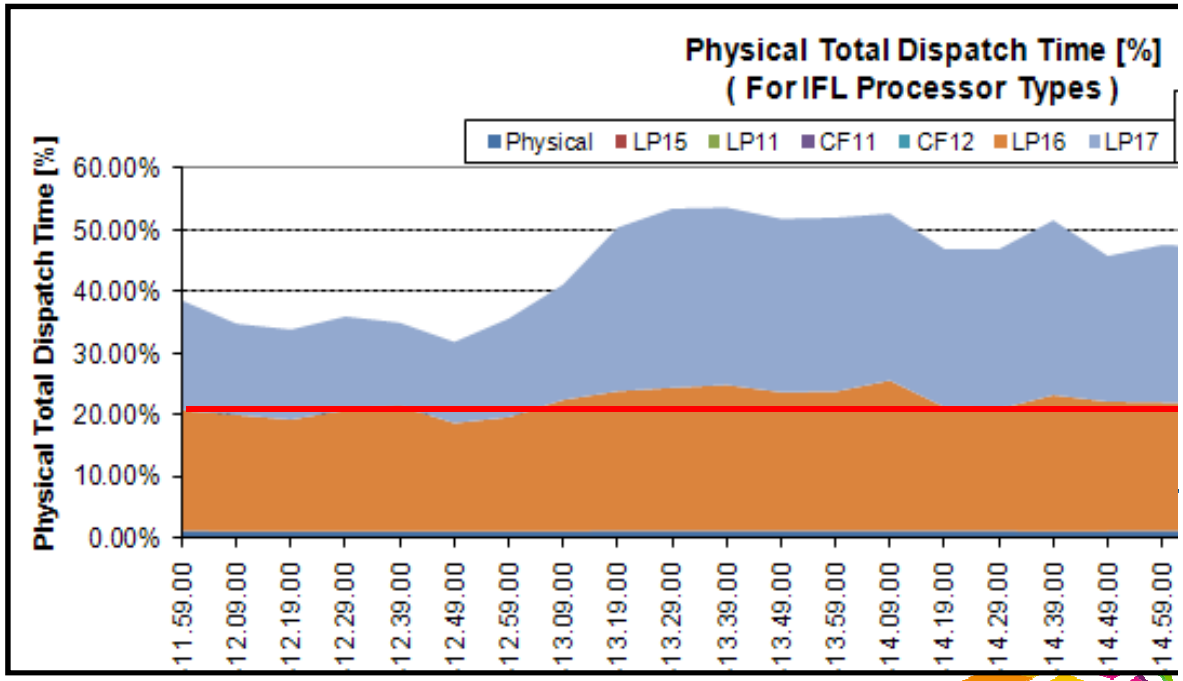
- … A tool to create, post-process and analyze RMF (Resource Measurement Facility) reports in the form of Excel Spreadsheets: a graph is worth a 1,000 words, especially if it has Red in it
- SMF (System Management Facility) records you need: reports can be run from tool, or MVS then pulled down and post-processed
  - **70-1: % CPU, zIIP busy, weightings, number of MSU's consumed, WLM capping, physical and logical busy**
  - 70-2: crypto HW busy
  - 71: paging activity
  - **72-3: workload activity**
  - 73: channel path activity
  - **74-1: device activity**
  - **74-4: coupling facility**
  - 74-10: SCM - *new*
  - 74-5: cache activity
  - 74-8: disc systems
  - 75: page data sets
  - 78-2: virtual storage
  - 78-3: I/O Queueing



RMF Reports - CPU

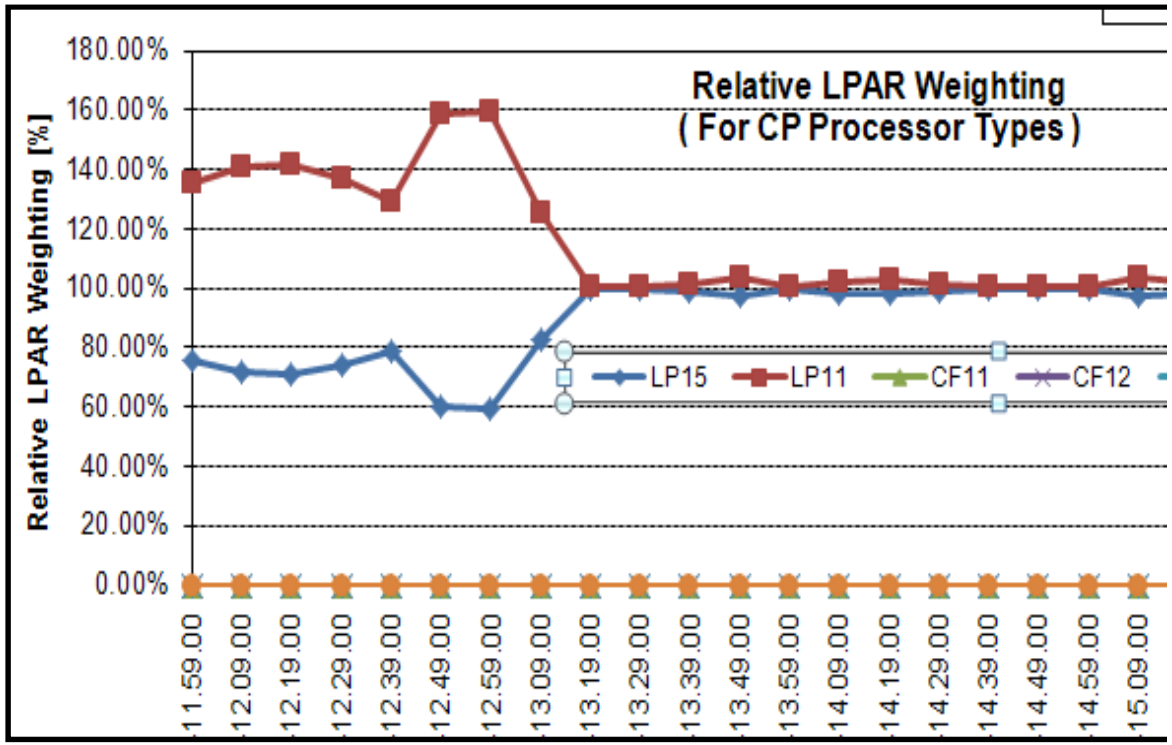
- LPAR Trend report
  - REPORTS(CPU)
- Can see stacked picture of single LPAR (GP/zIIP/IFL)
  - This is useful to get an idea of the CEC utilization across processors
- Look at CEC's CPU trend over the time period with GP and specialty engines
  - You can superimpose the max CPU % the LPAR will achieve based on weightings
  - Also see entire CEC saturation





RMF Reports - CPU

- LPAR Trend report
  - REPORTS(CPU)
- Can see relative weights between LPARs to determine if one is exceeding its share
  - i.e. who will be punished when a CPU constraint occurs
- LPAR Design tool very helpful in getting the right mix of vertical High/Med/Low processors



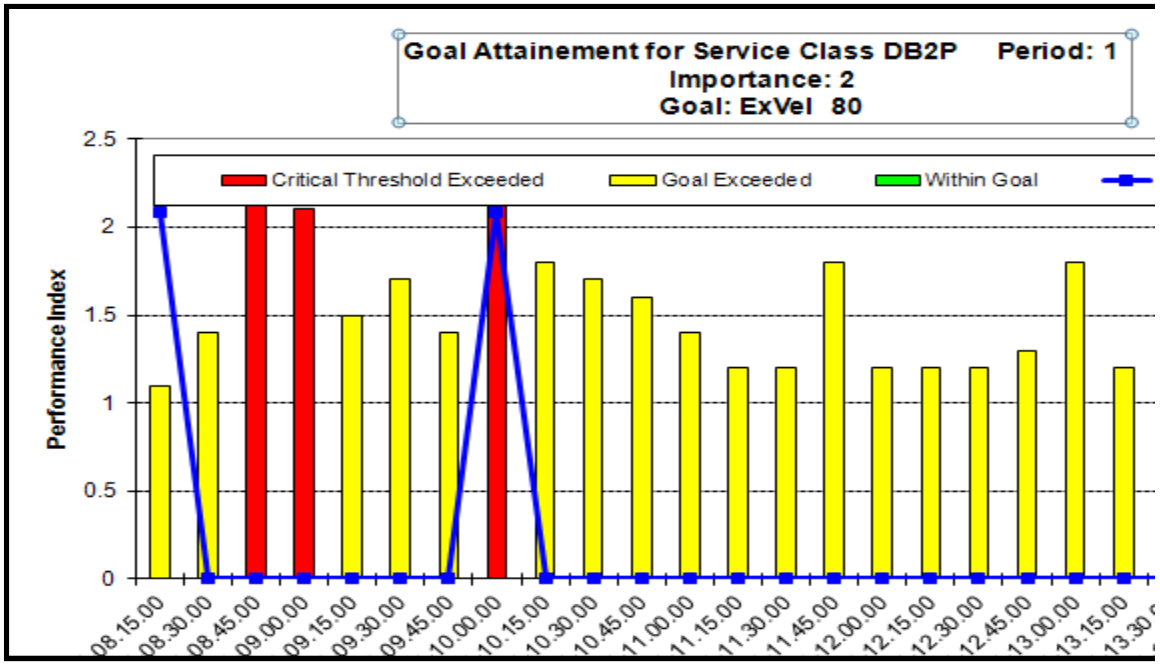
<---------- Weight ---------->				<---------- Average Processor Utilization ---------->				
	Percent	<----- % of CEC ----->		< Logical Processor % >		<------ Physical Processors % ------>		
Actual	of Total	Guaranteed	Max	Effective	LPARMgmt	Effective	LPARMgmt	of Weight
600	60.0	60.0	100.0	58.66	0.12	58.66	0.13	98%
400	40.0	40.0	100.0	40.97	0.02	40.97	0.02	102%

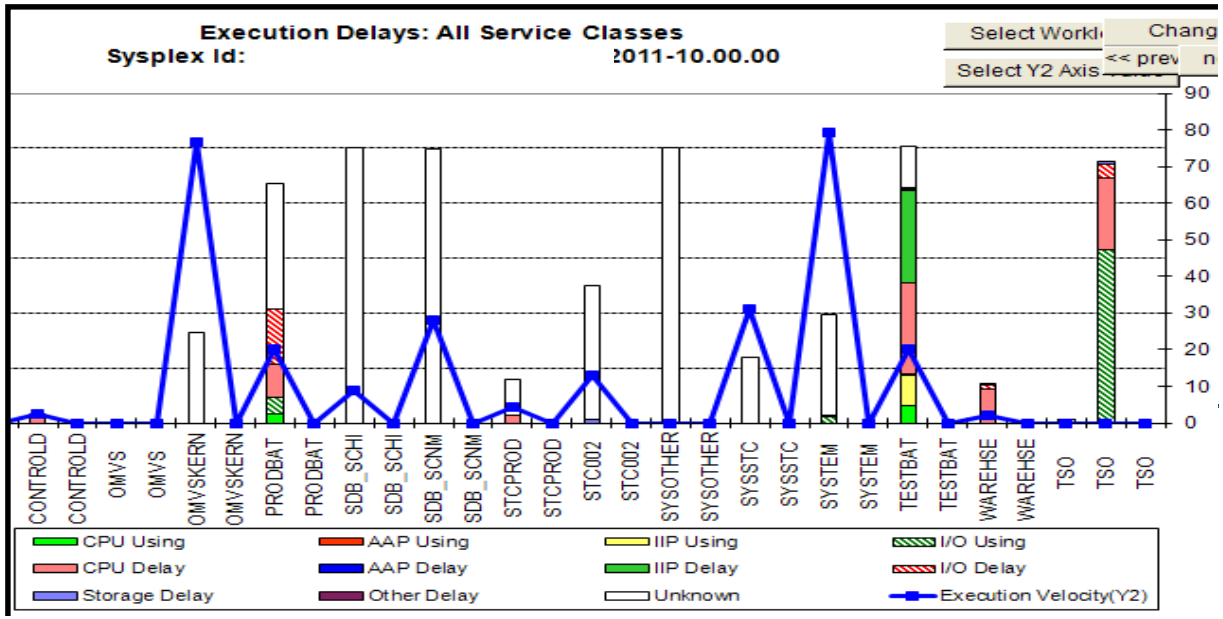
RMF Reports - WLM

- WLM activity report

  SYSRPTS(WLMGL(POLICY, WGROUP,SCLASS,SCPER, RCLASS,RCPER,SYSNAM(S WCN)))

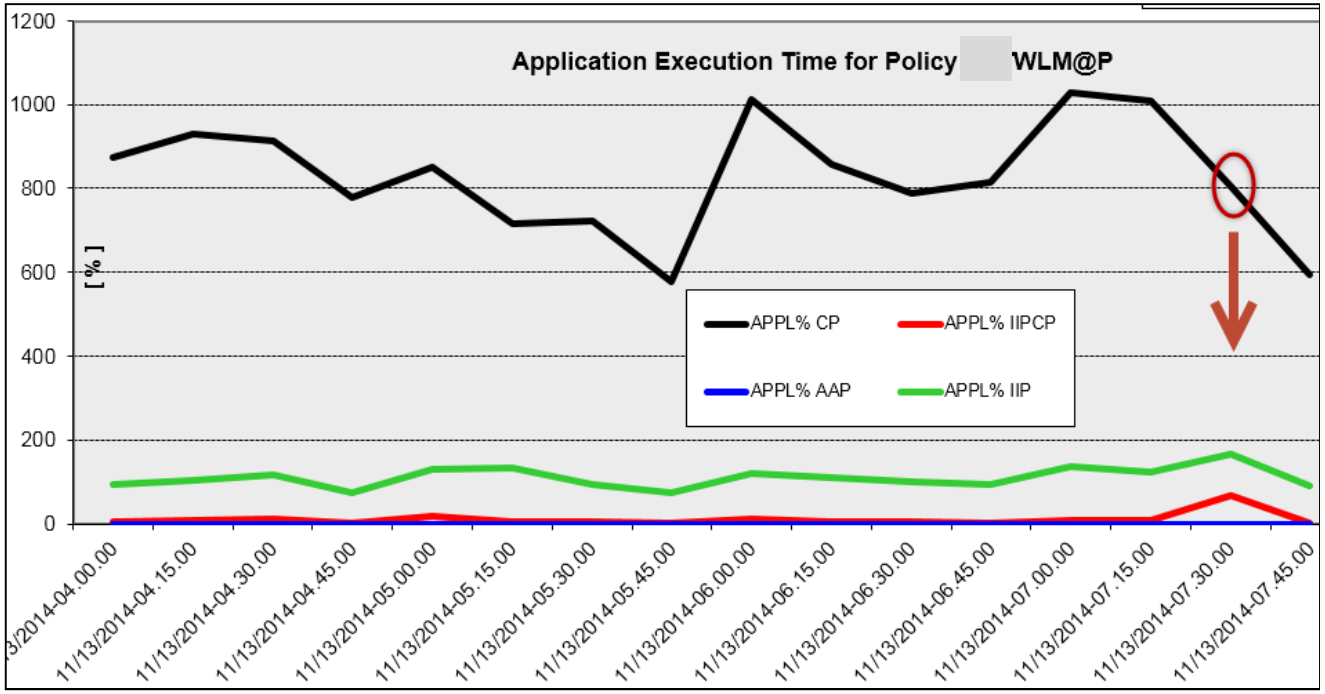
- Look at all service classes during a certain interval or 1 class over the course of several intervals
  - Yellow missed its goal, Red is a PI of >2
- See reason for delays across all service classes in an interval
  - I/O, CPU, zIIP
- Looking at raw report is tedious, could be hundreds of MBs of data





RMF Reports - WLM

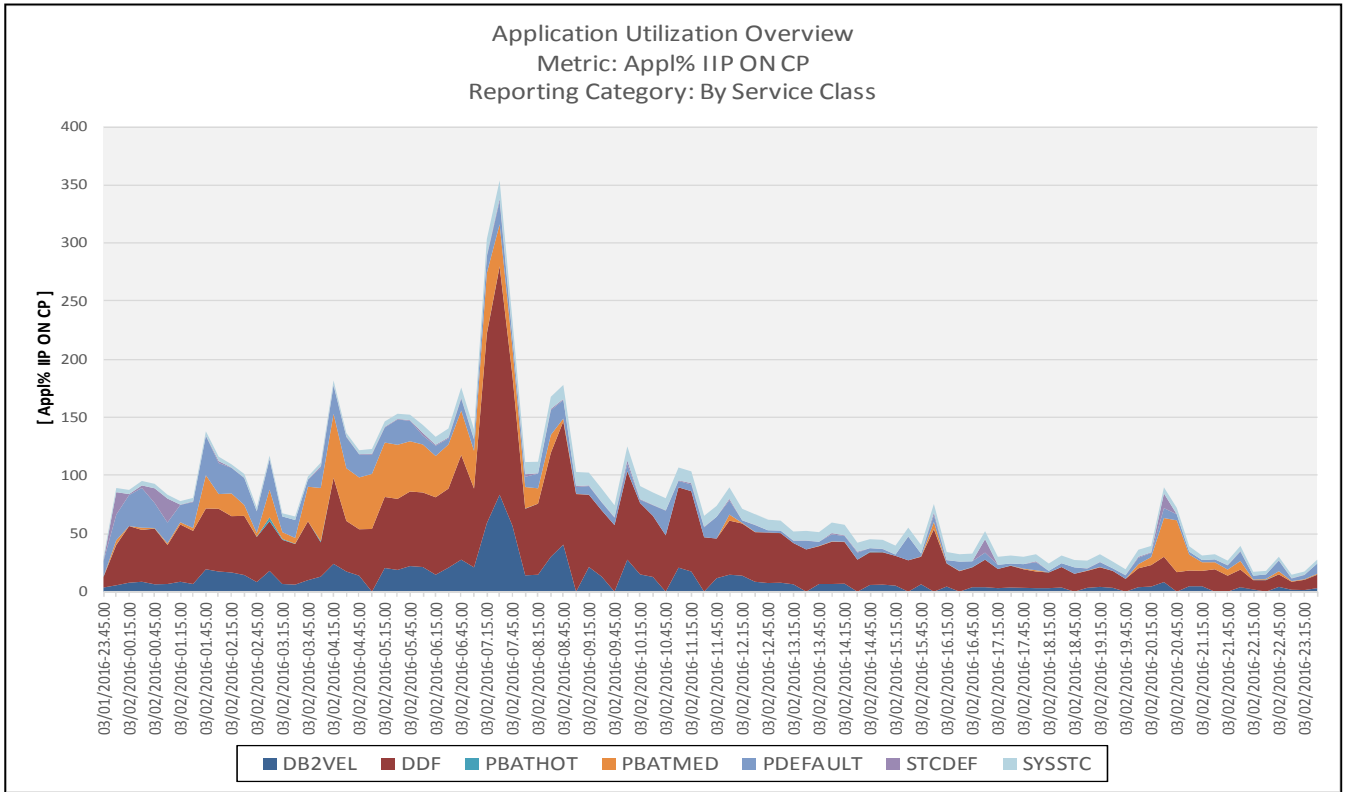
- Look for potential zIIP offload that landed on a GP
  - AAPL% IIPCP
    - **Red line**
  - See what % (not normalized) of a processor the workload consumed
- Response times can be seen and charted as well
  - Actual average execution time



8	RESCGRP	PERIOD	IMPORTN	<---------	---------	----TRANS-ACTIONS-		---------	--------->	<---------	-TRANSACTION	-TIMES----
9				AVG	MPL	ENDED	END/SEC	SWAPS	EXECUTD	ACTUAL	EXECUTION	QUEUED
42	DDFTHRDS	1	2	3.11	3.11	140089	233.48	0	0	0.013	0.012	0
43	DDFTHRDS	2	3	0.35	0.35	38	0.06	0	0	5.604	5.604000092	0
44	DDFTHRDS	3	4	11.02	11.02	3	0.01	0	0	22.781	22.78000069	0
45	DDFTHRDS			14.49	14.49	140130	233.55	0	0	0.015	0.014	0
46	DDFTHRDS	1	3	8.14	8.14	12392	20.65	0	0	0.103	0.103	0

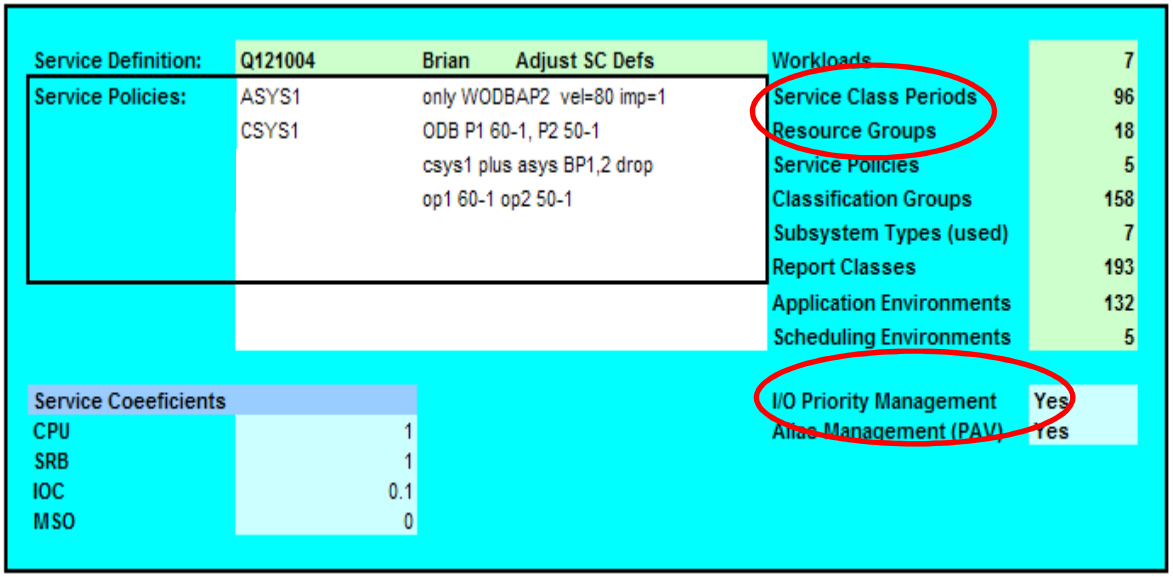
Overview Records

- This will show the CPU/zIIP Utilization broken down by service/ report class as well as CPU utilized by zIIP eligible work during the intervals
- It can show you when certain workloads collide and WLM goals are missed: who is driving the CPU % through the roof
  - By using RMF Spreadsheet reporter you can generate the Overview Records
    - Then create and run the Overview Report from your desktop
- ApplOvwTrd tab now included in spreadsheet, no need to create WLM OVWs



Reports – WLM service definition formatter

- FTP down your WLM policy in .txt format
- Import the WLM policy into a spreadsheet to analyze and filter
- Overview of total classes, periods, resource groups**
  - Resource group capping results in unpredictable response times
  - ROT is 30-40 service class periods running at once
- Policy itself can be filtered
  - So why do we have 9 Imp 1 Velocity 60 service classes?
  - This is redundant work for WLM to monitor and manage these identical classes
- Easy to search through rules to determine what work is in what service class



Service Definition:	Q121004	Brian	Adjust SC Defs
Service Policies:	ASYS1	only WODBAP2 vel=80 imp=1	
	CSYS1	ODB P1 60-1, P2 50-1	
		csys1 plus asys BP1,2 drop	
		op1 60-1 op2 50-1	

Workloads	7
Service Class Periods	96
Resource Groups	18
Service Policies	5
Classification Groups	158
Subsystem Types (used)	7
Report Classes	193
Application Environments	132
Scheduling Environments	5
I/O Priority Management	Yes
Alias Management (PAV)	Yes

Service Coeficients	
CPU	1
SRB	1
IOC	0.1
MSO	0

Service Policy

Policy	Workload	Service Class	Per	Dur	Imp	Goal Type	Goal Pct
ASYS1	W_BATCH		1		1	ExVel	60
ASYS1	W_BATCH		1		1	ExVel	60
ASYS1	W_BATCH		1		1	ExVel	60
ASYS1	W_BATCH		1		1	ExVel	60
ASYS1	W_BATCH		1		1	ExVel	60
ASYS1	W_BATCH		1		1	ExVel	60
ASYS1	W_BATCH		1		1	ExVel	60
ASYS1	W_BATCH		1		1	ExVel	60
ASYS1	W_STC	STC1	1		1	ExVel	60

RMF Reports - DASD

- DASD Activity Report
  - REPORTS(DEVICE(DASD))
- Gives you overview of top 5 Logical Control Units
  - See what volumes are on there, and what DB2 data is on those volumes
- Device Summary Top10 Shows top 10 volumes based on criteria you specify and you can manipulate graphs

RMF DASD System Summary

System Id:	ESA1	Operating System:	z/OS V1R10 -	Report Range(seconds):	900
Reporting Date:	02/23/2011	Reporting Time:	08.30.00	Report Range(hh.mm.ss):	15.00.000

System Summary

#of LCUs	# of DASD	I/O Intens.	ST Intens.	Path Int.	Act. Rt.	Resp. Tm	Serv. Tm	IOSQ Tm	Pend. Tm	Disc. Tm
9	694	94249.73	5221.12	3793.31	1538.43	61.26	3.39	57.57	0.30	0.93

LCU Summary

	LCU	I/O Intens.	ST Intens.	Path Int.	Act. Rt.	Resp. Tm	Serv. Tm	IOSQ Tm	Pend. Tm	Disc. Tm
Top 5	0035	85827.90	923.02	814.34	146.46	585.63	6.30	579.00	0.33	0.74
	0036	4210.37	1464.61	1401.45	300.74	[illegible]	4.87	8.81	0.31	0.21
	0030	2933.42	1893.53	1090.33	388.02	7.56	4.88	2.38	0.30	2.07
	002F	620.17	431.46	174.53	442.98	1.40	0.97	0.14	0.29	0.58
	0034	326.09	236.88	71.78	51.27	6.36	4.62	1.39	0.35	3.22

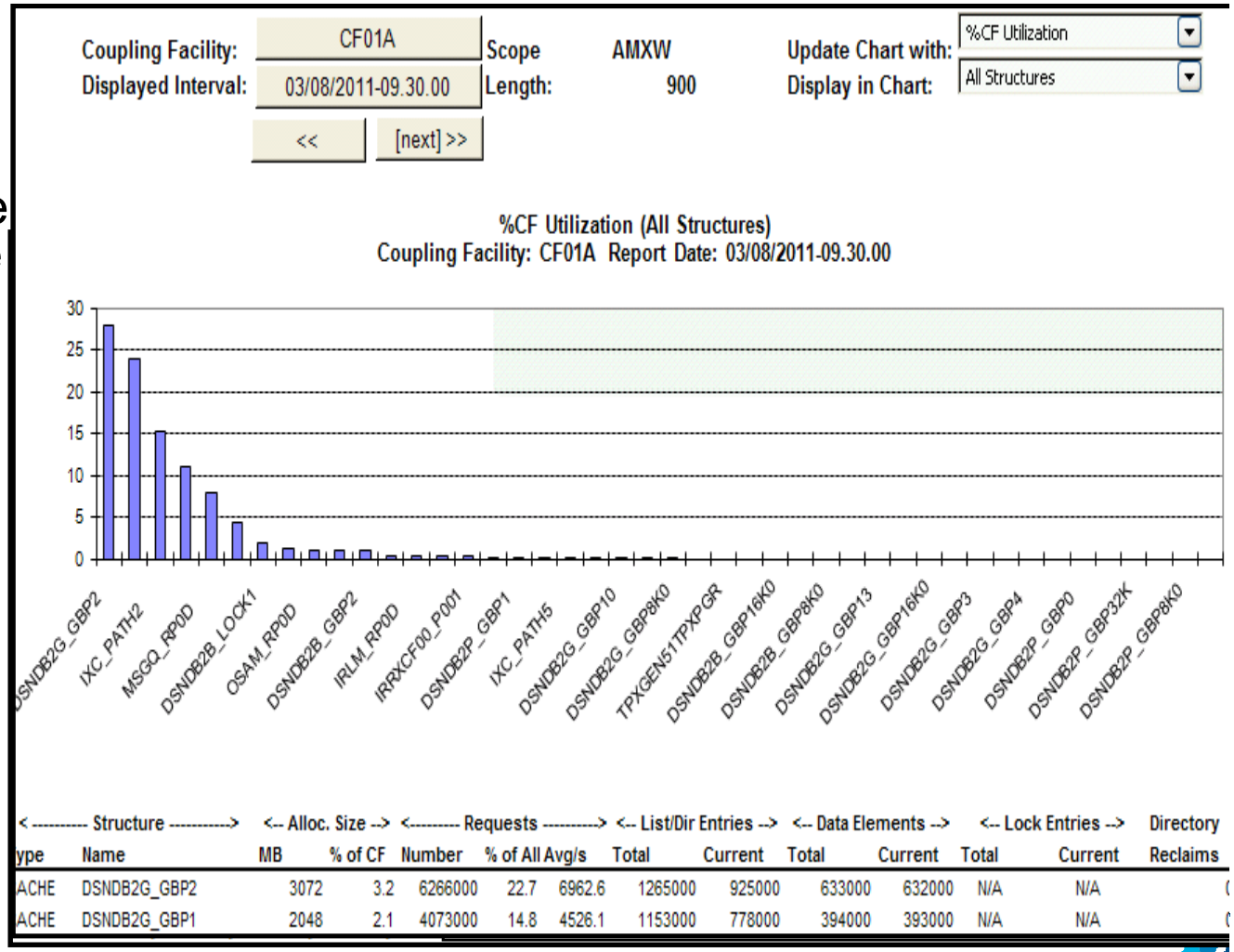
sorted by I/O Intensity

Device Summary Top 10

LCU	VolSer	I/O Intens.	ST Intens.	Path Int.	Act. Rt.	Resp. Tm	Serv. Tm	IOSQ Tm	Pend. Tm	Disc. Tm
0035	DBS001	84992.57	600.32	516.51	91.10	932.93	6.59	926.00	0.34	0.92
0036	DB2WC1	1842.65	71.56	68.10	16.02	114.80	4.47	110.00	0.33	0.22

RMF Reports - CF

- CF activity report
  - SYSRPTS(CF)
- Look at CPU/storage utilization over entire day
- See comparison of sync vs. async across intervals
- Look for delays due to sub-channels being unavailable
- Look for directory reclaims
- Look at all metrics for all structures during a single interval



< Structure >		<-- Alloc. Size -->		<-- Requests -->			<-- List/Dir Entries -->		<-- Data Elements -->		<-- Lock Entries -->		Directory
ype	Name	MB	% of CF	Number	% of All	Avg/s	Total	Current	Total	Current	Total	Current	Reclaims
ACHE	DSNDB2G_GBP2	3072	3.2	6266000	22.7	6962.6	1265000	925000	633000	632000	N/A	N/A	
ACHE	DSNDB2G_GBP1	2048	2.1	4073000	14.8	4526.1	1153000	778000	394000	393000	N/A	N/A	

RMF Summary Report

- RMF Summary
  - Look at CPU Busy (remember this is usually a 15 minute interval though)
  - DASD response taking into account the rate, a very low rate could show increased response time due to missing cache, etc.
  - Demand paging
    - **Now-a-days we don't want to see paging at all as storage gets cheaper and the price paid by the online applications in response time not proportional to the 'paging rate'**
    - **z/OS measured the CPU cost of a sync I/O at 20-70us**

R M F S U M M A R Y R E P O R T

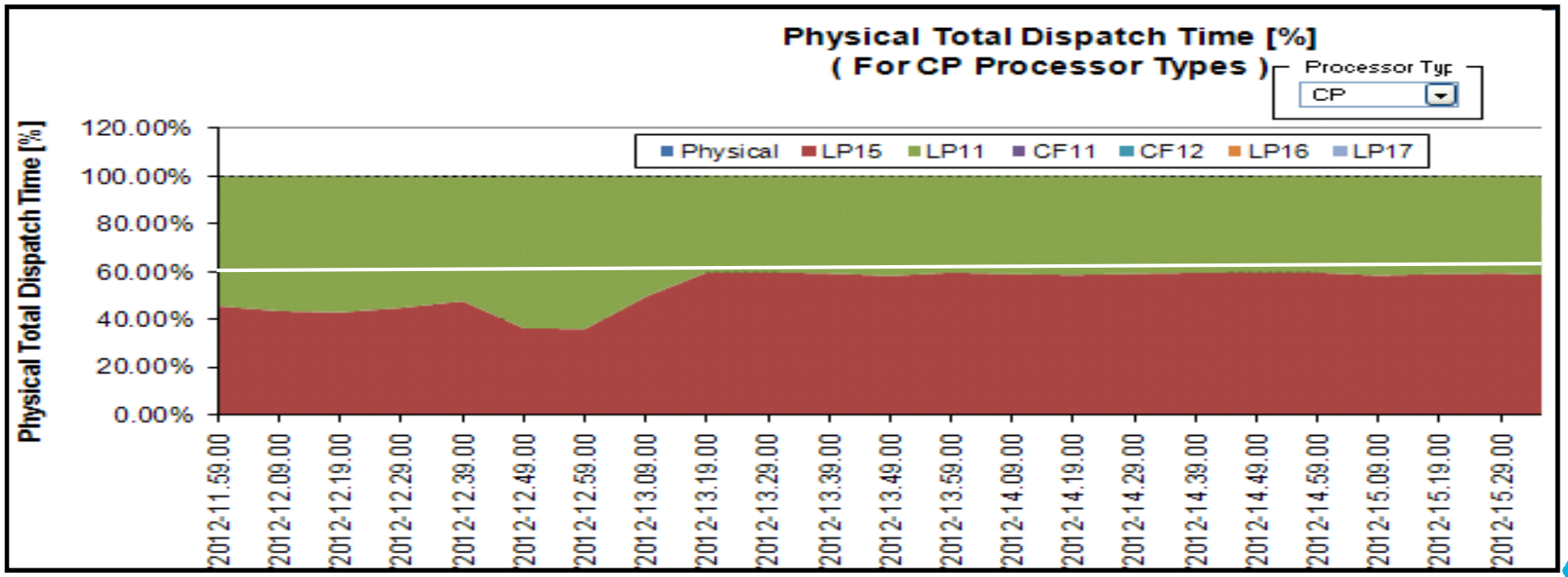
2 SYSTEM ID 2D11 START 09/25/2012-11.59.00 INTERVAL 00.09.59
CONVERTED TO z/OS V1R13 RMF END 09/25/2012-16.59.00 CYCLE 1.000 SECONDS

TOTAL LENGTH OF INTERVALS 04.59.44

CPU BUSY	DASD RESP	DASD RATE	TAPE RATE	JOB MAX	JOB AVE	TSO MAX	TSO AVE	STC MAX	STC AVE	ASCH MAX	ASCH AVE	OMVS MAX	OMVS AVE	SWAP RATE	DEMAND PAGING
58.3	1.7	1848	124.0	75	72	6	6	181	178	0	0	5	5	0.00	0.05
55.5	1.8	1589	27.1	73	71	6	6	180	178	0	0	5	5	0.00	0.02

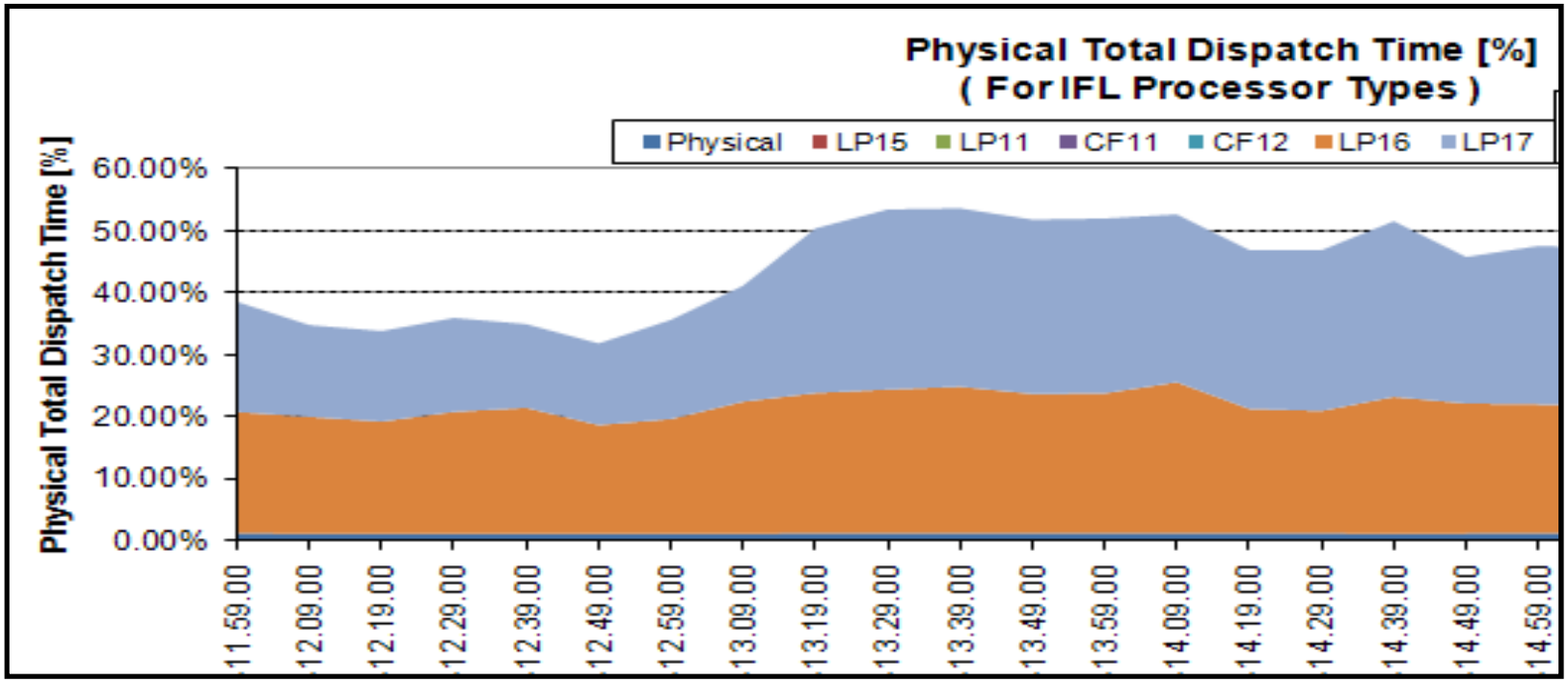
CPU constraints (1)

- These 2 LPARs LP11 and LP15 are consuming every MIP on the box, borrowing back and forth
  - This was meant to be a load test, and you can see where the test LPAR (Green) ran out of steam as the production LPAR took the CPU cycles
- In internal benchmarks maximum throughput is achieved between 92-94% - determining root cause almost impossible at 100%, no consistency



CPU Constraints (2)

- LP11 and LP15 saturate the **2 out of 2 CPs** during the day, trading off resources while at the same time Portal is driving **2.5 out of 5 IFLs**
  - The GCPs on the previous slide is already fully utilized, and the Portal workload here has 50% of its capacity left, so it appears DB2 is the bottleneck
  - So it is the CPU capacity...as well as



WLM

- ROT: DB2 threads should not end up in a service class which uses WLM resource group capping
  - Resource group capping will ensure that this workload does not get over ‘x’ Service Units a second, and this includes all the DB2 subsystems in the plex,
  - Blocked workload support cannot help these capped transactions, so if there is a serious CPU constraint all DDF work could be starved, and could be suspended while holding important DB2 locks/latches
    - In general we suggest avoiding resource group capping in favor of lowering the priority of the work
    - The CAP delay is the % of delays due to resource group capping

	RESPONSE TIME	EX	PERF	AVG	--EXEC USING%--				----------			EXEC DELAYS %		
SYSTEM	HHH.MM.SS.TTT	VEL%	INDX	ADRSP	CPU	AAP	IIP	I/O	TOT	CPU	CAP	IIP	Q MPL	I/O
*ALL	000.00.00.027	15.4	0.0	10.6	4.9	N/A	1.0	0.4	35	22	11	1.3	0.1	0.1
1E10	000.00.00.015	25.3	0.0	4.0	9.3	N/A	2.6	0.1	35	17	15	3.4	0.0	0.0
2D11	000.00.00.169	7.6	0.0	6.6	2.2	N/A	N/A	0.7	34	25	9.0	0.0	0.2	0.1

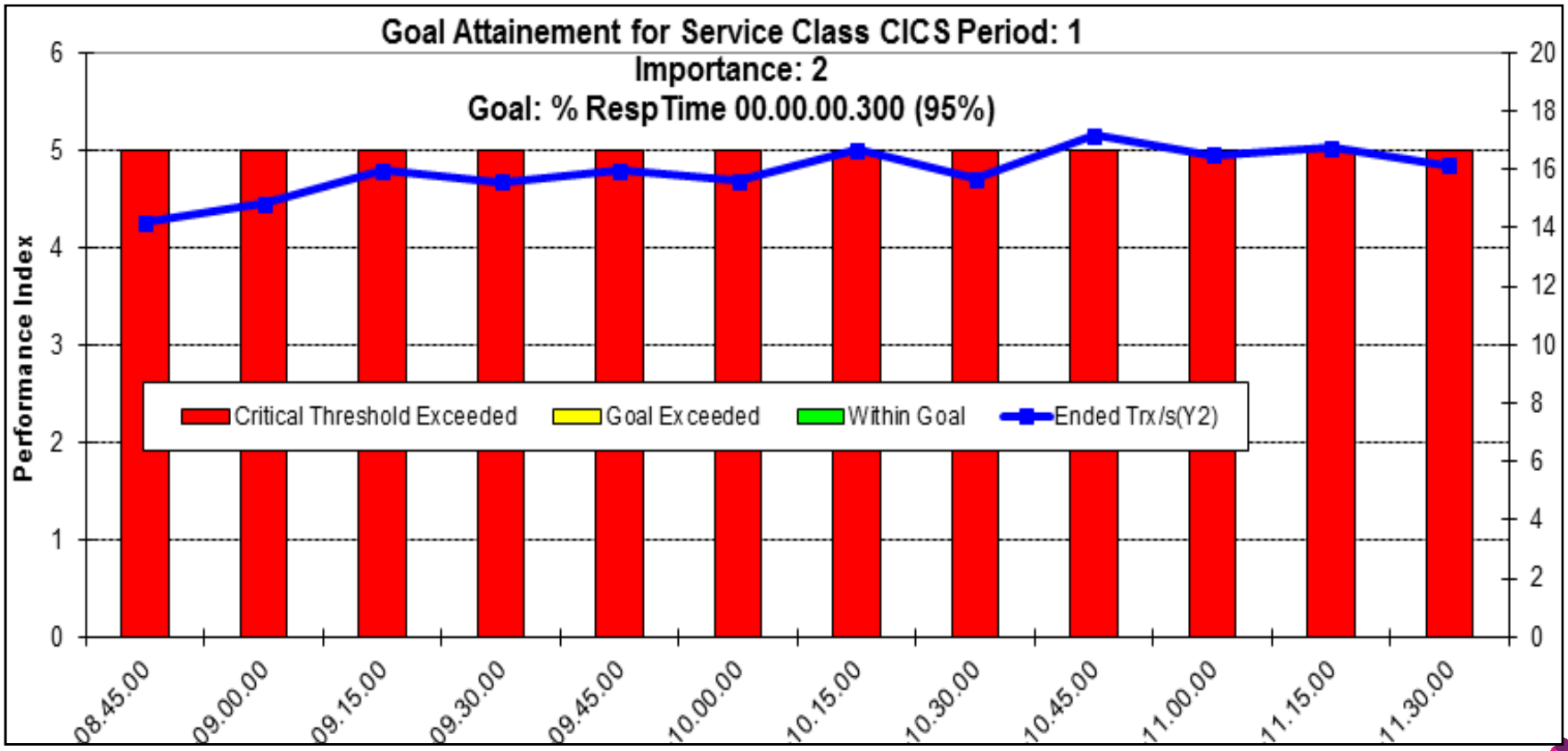
Response time goals… too loose

- We do not want the goals to be too loose: if >90% of transactions complete in less than ½ of their goal, the goal should be adjusted tighter, to avoid violent swings in response time under CPU constraint
  - The goal here is 10 seconds for a Portal application that must render its page in 3 seconds, and the transactions are finishing in 4 milliseconds
    - The WLM goals should align with the business goals/ SLAs
  - Make the goal around 20 milliseconds so the service level can be maintained

```
GOAL: RESPONSE TIME 000.00.10.000 AVG

         RESPONSE TIME EX    PERF  AVG    --EXEC USING%--  -------------- EXEC DELAYS % -----------  -USING%-  --- DELAY % ---    %
SYSTEM   HHH.MM.SS.TTT VEL%  INDX ADRSP   CPU AAP IIP I/O  TOT CPU IIP                                CRY CNT  UNK IDL CRY CNT  QUI

*ALL     000.00.00.004 67.3   0.0   0.4    27 N/A 7.4 0.0   17  15 1.3                                0.0 0.0   51 0.0 0.0 0.0  0.0
1E10     000.00.00.003 72.1   0.0   0.3    31 N/A  10 0.0   16  14 1.7                                0.0 0.0   45 0.0 0.0 0.0  0.0
2D11     000.00.00.009 43.2   0.0   0.1    14 N/A N/A 0.0   19  19 0.0                                0.0 0.0   67 0.0 0.0 0.0  0.0

                                          ----------RESPONSE TIME DISTRIBUTION----------
   ----TIME----    --NUMBER OF TRANSACTIONS--    -------PERCENT-------  0    10   20   30   40   50   60   70   80   90   100
   HH.MM.SS.TTT    CUM TOTAL        IN BUCKET    CUM TOTAL   IN BUCKET  |....|....|....|....|....|....|....|....|....|....|
<  00.00.05.000        80322            80322          100         100  >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
<= 00.00.06.000        80322                0          100         0.0  >
<= 00.00.07.000        80322                0          100         0.0  >
<= 00.00.08.000        80322                0          100         0.0  >
<= 00.00.09.000        80322                0          100         0.0  >
<= 00.00.10.000        80322                0          100         0.0  >
```

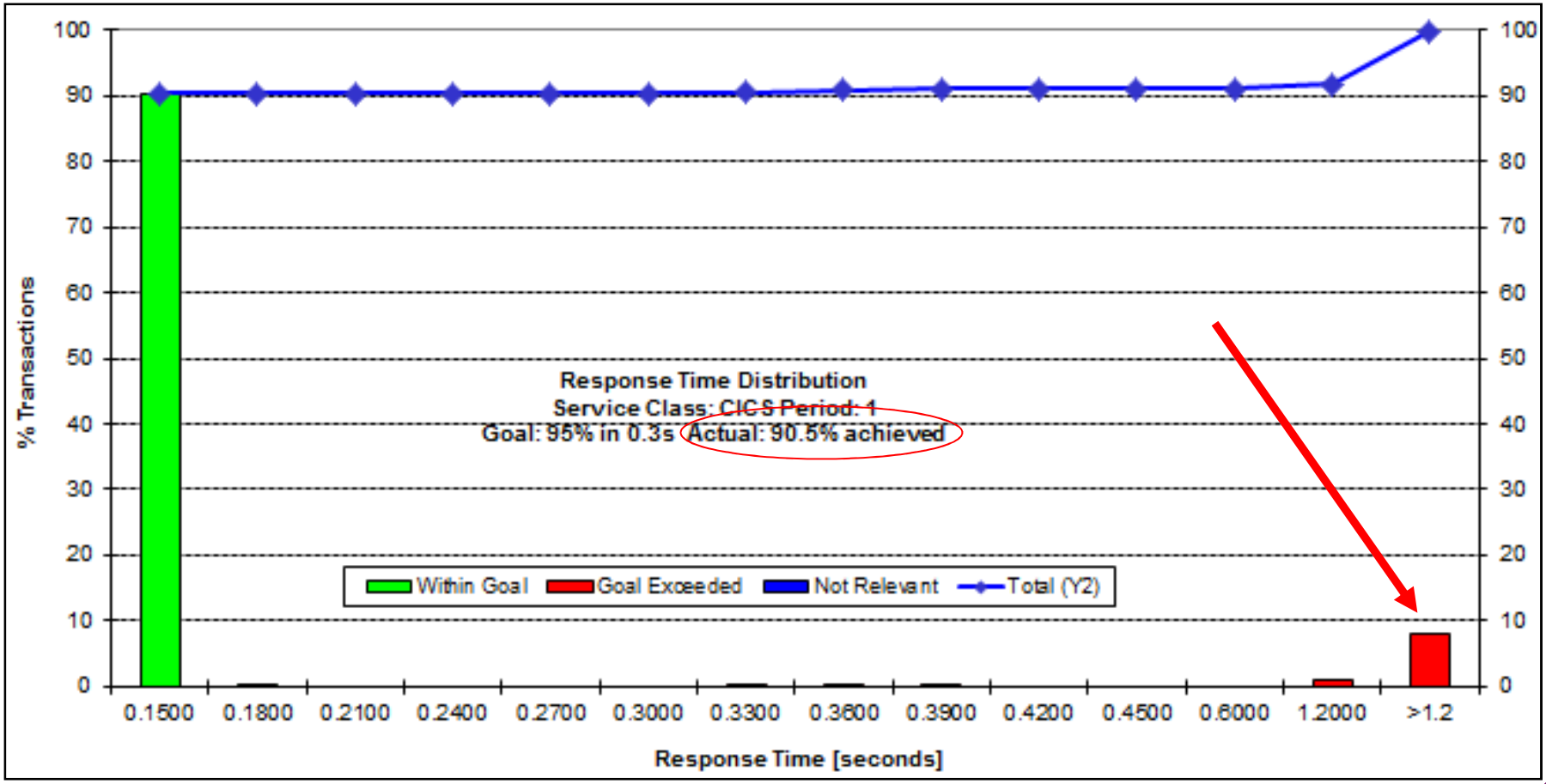
Response time goals… too stringent

- The goals need to be reasonable, i.e. attainable by the workload
  - WLM cannot shorten the response time to something lower than the CPU time needed for the transaction to complete
  - With a performance index of 5 all day long this workload could be skip clocked (ignored) if there were CPU constraints



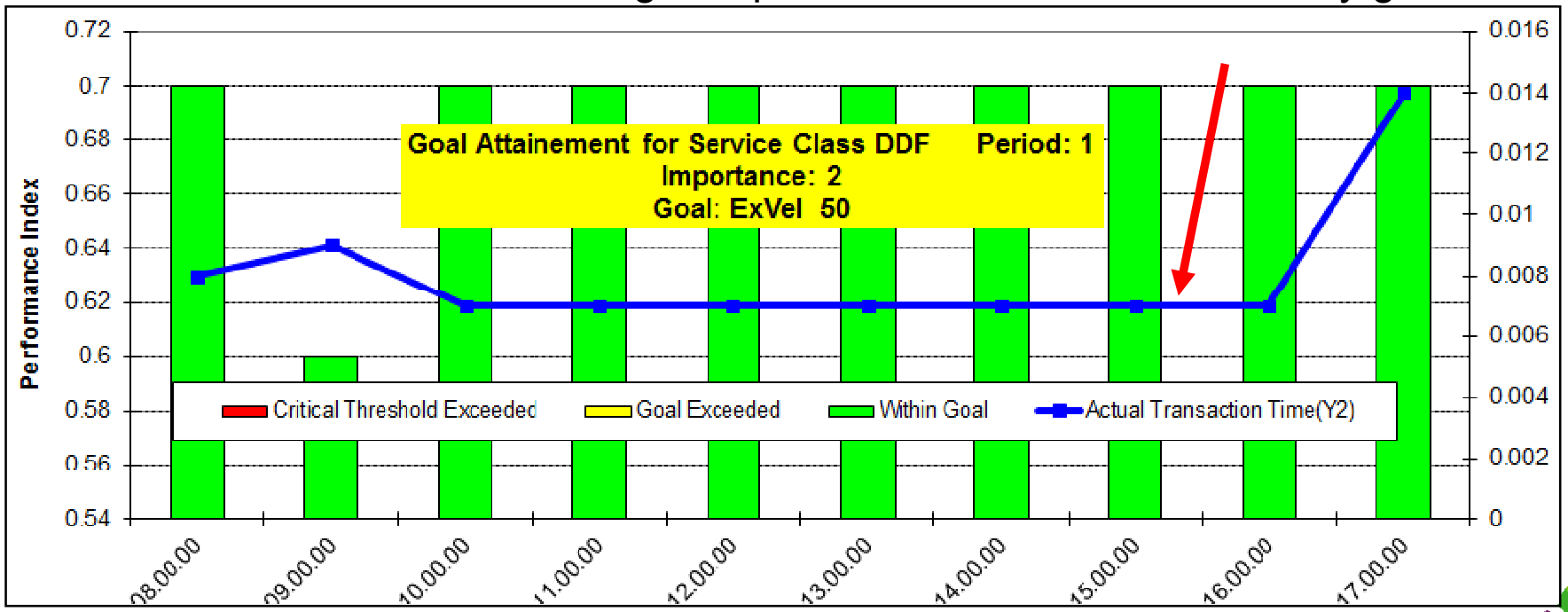
WLM Buckets

- Look at the response time buckets in WLM activity report to gauge reality
- No amount of CPU could bring these transactions back in line with the others
  - The goal is 95%, but only 90% complete in time, so take these outlying trans and break them out into another service class, or adjust the goal to 90%



Response time goals vs. velocity goals

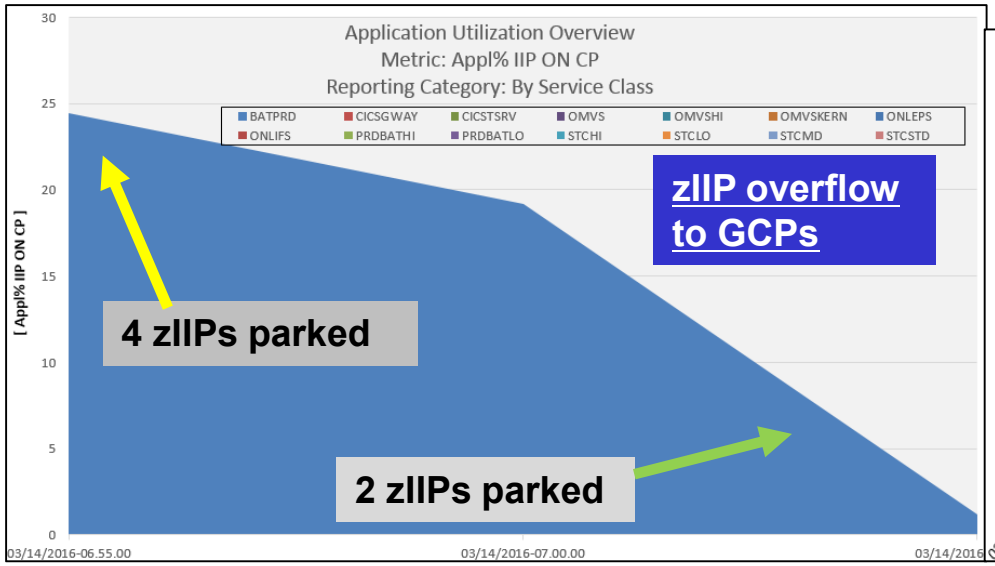
- For transactions and most business processes a Response time goal is much more effective/predictable during times of CPU constraint than velocity goals – percentile or average goal?
  - Transaction classes with outliers of 2x the average or more should be percentile goals
- When determining a good response time goal you need to trend it out
  - Determine where the business goal is in relevance to what it is achieving
  - z/OS 1.13 includes average response time info even for velocity goals

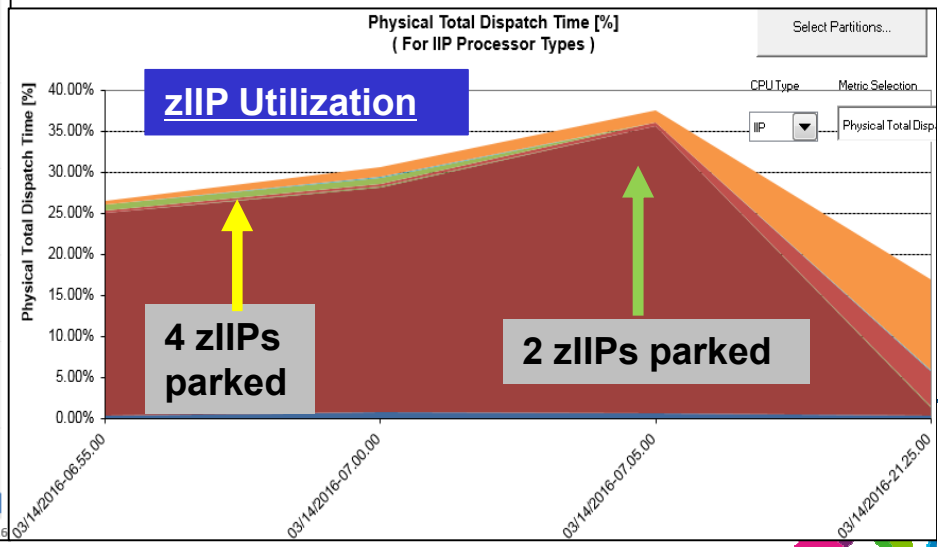


zIIP and LPAR Weights

- For capacity planning monitor zIIP redirect to CP, not absolute Utilization
- Correct technical solution is to add more zIIP capacity to avoid zIIP eligible work running on a CP (APPL% IIPCP CPU in WLM activity report SMF 72-3)
  - zIIPs are assist processors and not intended to be run as hard as GCPs
  - Using the RMF Spreadsheet reporter you can see the service/report class spilling over
  - On z13 proper LPAR weightings are key
- Hiperdispatch is VERY sensitive to the relative LPAR weights (HIGH/MED/LOW)
- Key is to apportion weights based on actual utilization – not share zIIPs with everyone
  - Otherwise engines will remain parked causing work to spill over to the GCPs
- Many zIIP eligible workloads are 'spikey' in nature: look at Work Units in CPU activity
  - Parallelism from SQL, Utility, or Sort work → → →

---------NUMBER OF WORK UNITS----------			
CPU TYPES	MIN	MAX	AVG
CP	0	24	2.1
IIP	0	185	1.0

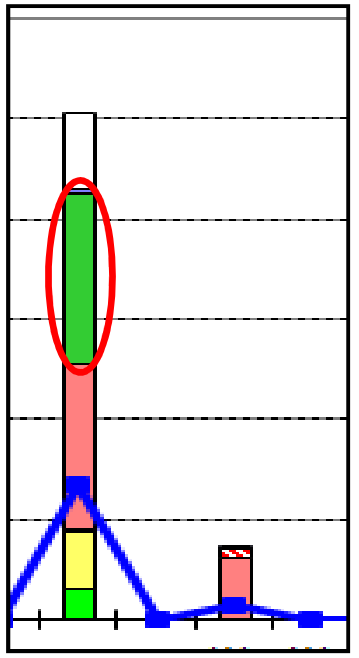




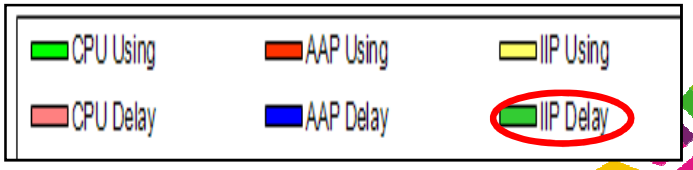
zIIP Shortages

- *What if I have lots of not accounted for time?*
  - OMPE accounting report (parallel tasks on zIIP)
- RMF Spreadsheet Reporter response delay report
  - Part of WLM activity trend report
- SYS1.PARMLIB (IEAOPTxx) setting
  - IIPHONORPRIORITY = **NO** **(not recommended)**
    - Meaning all zIIP eligible work will queue waiting for a zIIP
      - Normally 'Needs Help' algorithm re-dispatches work to a GCP
      - Parallel tasks are 80% zIIP eligible, so PARAMDEG should be influenced by the number of zIIPs you have
    - **Important in v10 and v11, and if you have zAAP on zIIP – no zAAP on z13**
      - V10 includes prefetch, deferred writes,
      - V11 includes GBP writes, castout/notify, log prefetch/write
- Discretionary work always waits on the zIIP

```
CLASS 2 TIME DISTRIBUTION
-------------------------------
CPU      |=====> 11%
SECPU    |
NOTACC   |==================> 37%
SUSP     |=========> 19%
```



CPU delay at about 33%, and the zIIP delay is at 34%.



zIIPs and Prefetch

- What happens if Prefetch Engines are starved of zIIP?
  - Other Read I/O events and time per event will increase
  - PREF. DISABLED – NO READ ENG could increase
- Customers have seen batch programs miss their window
  - solution is to add zIIP capacity
- Prefetch may be scheduled even if all the pages are resident, so app still sees delays with 100% BP hit ratio and no I/Os
  - Increased elapsed time

CLASS 3 SUSPENSIONS	AVERAGE TIME	AV.EVENT
LOCK/LATCH(DB2+IRLM)	0.060293	48.65
IRLM LOCK+LATCH	0.000465	0.10
DB2 LATCH	0.059829	48.54
SYNCHRON. I/O	28.298614	69721.17
DATABASE I/O	28.298426	69720.92
LOG WRITE I/O	0.000188	0.25
OTHER READ I/O	5.036911	4802.06
OTHER WRTE I/O	0.000000	0.00

TOT4K READ OPERATIONS	QUANTITY	/SECOND	/THREAD	/COMMIT
SEQUENTIAL PREFETCH READS	4472.3K	311.88	12.35	0.55
LIST PREFETCH REQUESTS	1874.3K	130.70	5.18	0.23
LIST PREFETCH READS	745.1K	51.96	2.06	0.09
DYNAMIC PREFETCH REQUESTED	119.0M	8301.34	328.82	14.74
DYNAMIC PREFETCH READS	16325.1K	1138.43	45.09	2.02
PREF.DISABLED-NO BUFFER	285.00	0.02	0.00	0.00
PREF.DISABLED-NO READ ENG	656.00	0.05	0.00	0.00
PAGE-INS REQUIRED FOR READ	811.9K	56.62	2.24	0.10

DASD response time

- Sometimes you need the entire picture when going after response time issues
  - After migration to DB2 10 customer's applications were experiencing 'good' and 'bad' days
  - Some access path regressions… but was this related?
- Here are two top 5 logical control unit report from the same time each day
  - Activity rate is quite close (same work going on)
  - Where does the increase in response time come from? – DISC (disconnect time)
  - Synchronous remote copy (Metro Mirror) where the target cannot keep up, and asynchronous copy with write pacing (XRC) can cause high DISC time

LCU Summary

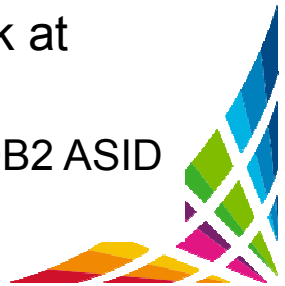
	LCU	I/O Intens.	ST Intens.	Path Int.	Act. Rt.	Resp. Tm	Serv. Tm	IOSQ Tm	Pend. Tm	Disc. Tm
Top 5	004C	1135.07	1005.99	206.60	346.06	3.28	2.91	0.11	0.27	2.31
	004E	442.14	399.44	100.49	83.74	5.28	4.77	0.16	0.35	3.57

LCU Summary

	LCU	I/O Intens.	ST Intens.	Path Int.	Act. Rt.	Resp. Tm	Serv. Tm	IOSQ Tm	Pend. Tm	Disc. Tm
Top 5	004C	11836.20	9407.79	446.53	305.85	38.71	30.76	7.51	0.44	29.30
	0078	4055.60	1636.81	276.85	242.85	16.69	6.74	9.52	0.43	5.60

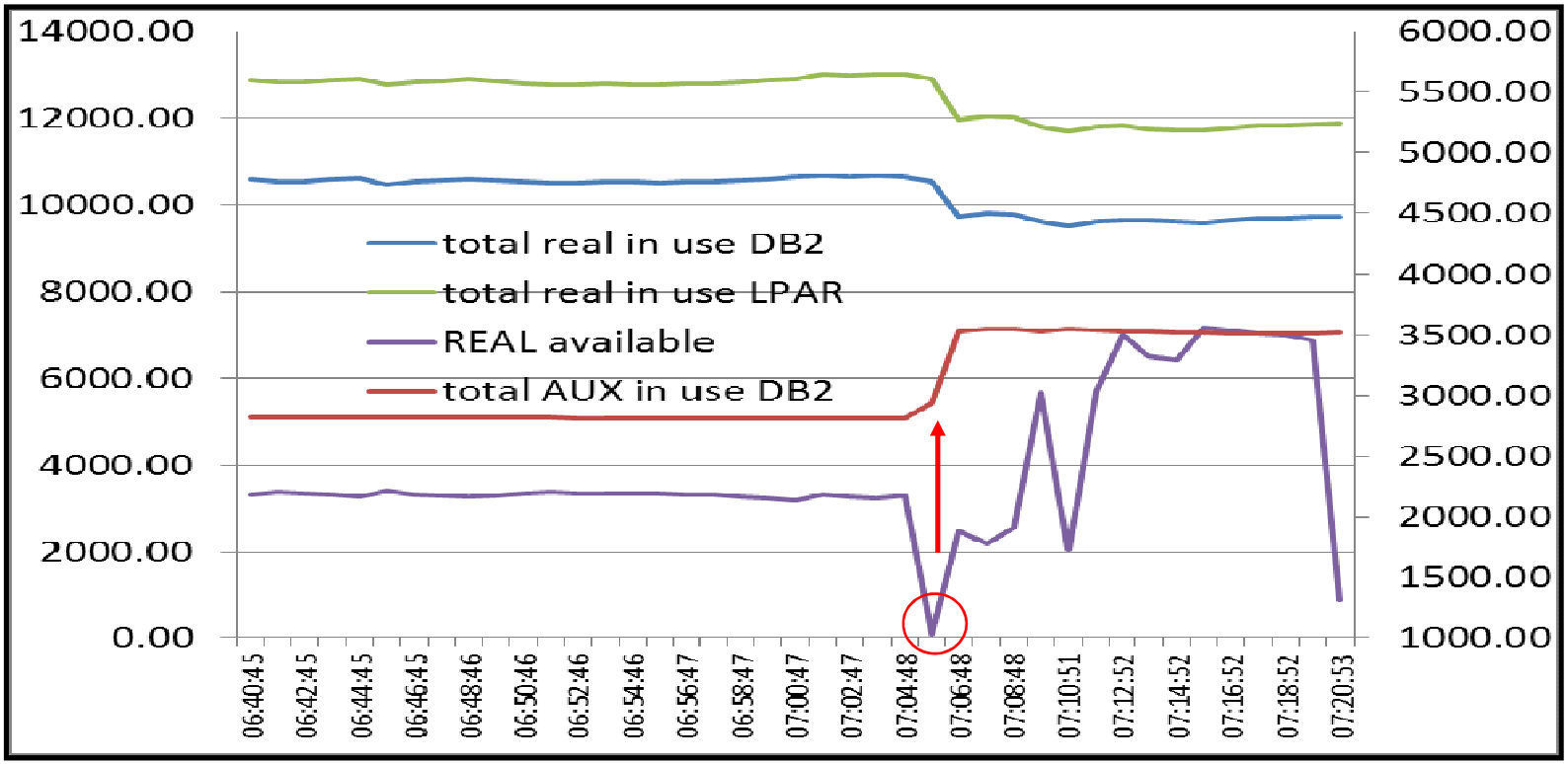
DB2 and Storage

- What is an acceptable paging rate? → 0 for DB2 storage
- REAL storage is a one time charge which will save CPU cycles, which you pay for on a monthly basis
  - z/OS measured the CPU cost of a sync I/O to be 20us→70us
- Even if you want to wait until V11 to tune your buffer pools with the simulation capabilities you can save CPU today by avoiding paging
- Impact customers have seen from being short on REAL storage
  - Transaction times begin to climb, customers see sub-second trans take 10's of seconds (buffer pool hit might require a page-in from AUX)
  - # of concurrent threads in DB2 begin to climb, CTHREAD/MAXDBAT might be hit
  - SYSPROG and DBA perception is of a system slowdown
    - If SVCDUMP occurs (SDUMP,Q) workload may be non-dispatchable until dump finishes
- If however you are real storage rich (i.e. > MAXSPACE in reserve) , look at turning REALSTORAGE_MANAGEMENT = AUTO → OFF
  - Have seen cases with large memory and high thread deallocations where DB2 ASID CPU could be saved



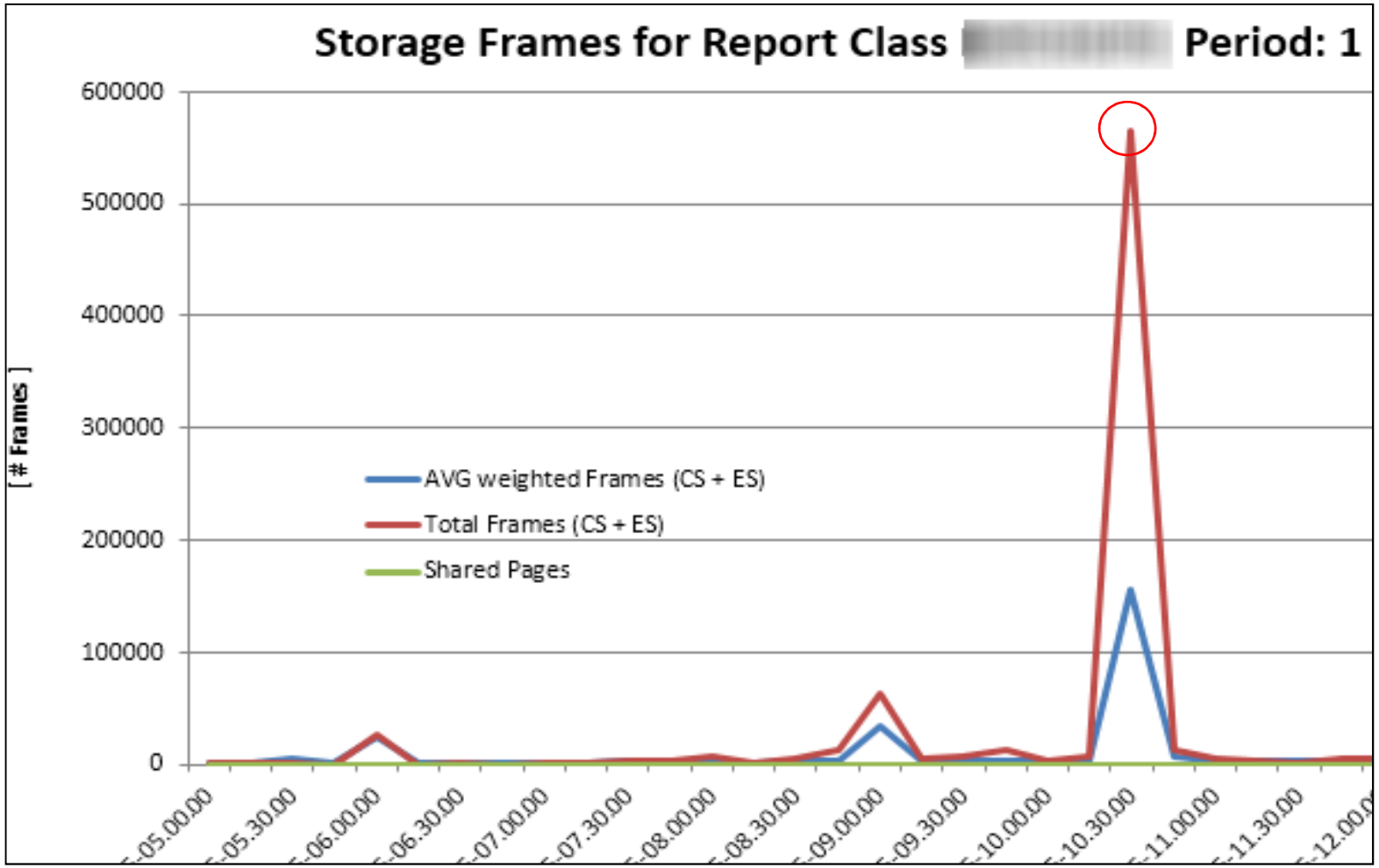
DB2 and Storage

- In the graphic we can see DB2 storage goes out from REAL to AUX when the real available drops to ‘0’ on the LPAR
  - Using IFCID 225 and MEMU2 to look at AUX vs. what’s in REAL
- Worst case in this example to get those pages back in:
  - 700 MB – sync I/O time ~3ms = 0.003*179,200 = 537 seconds
  - MAXSPACE suggestion 16GB… could not be supported here



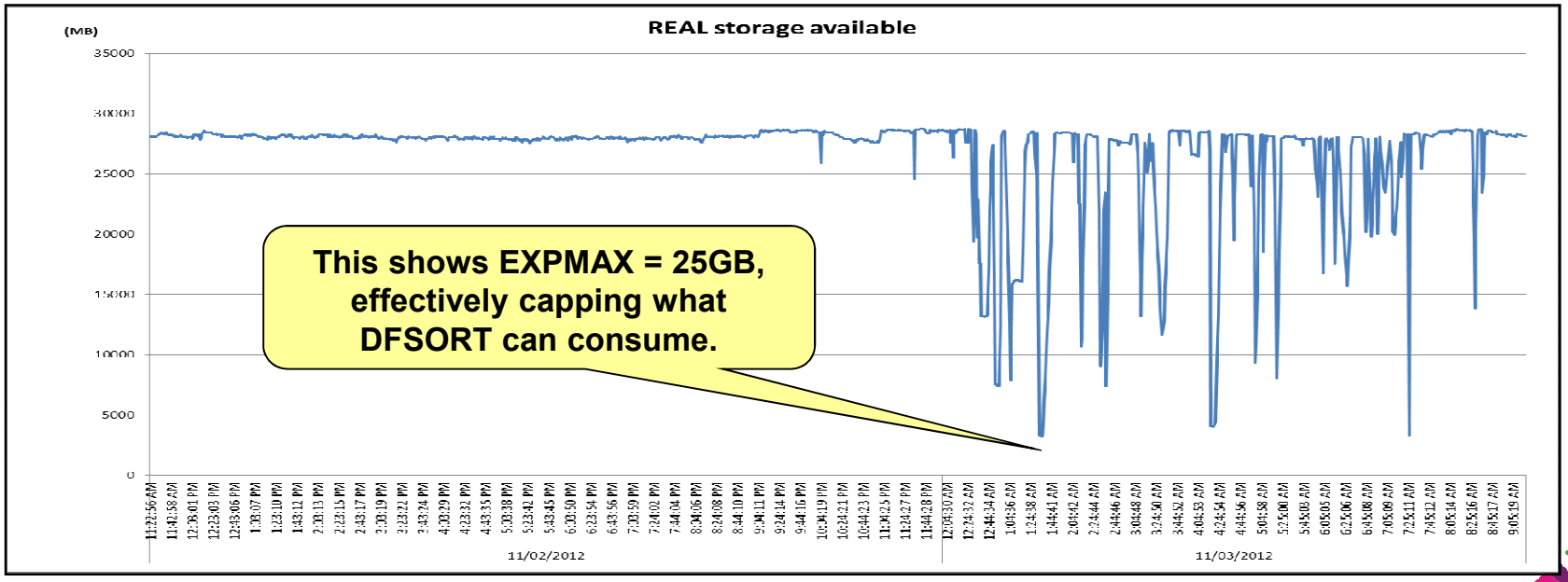
DB2 and Storage

- So who caused me to get paged out??
  - If you run a WLM activity report and look at the Storage Trend graph in the reporter you can see the actual frames used by a service or report class
  - The Page In Rates were also high during this time for DB2 as it recovered from AUX



Real storage and Sort products

- By default DFSORT and other sort products usually take as much storage as they can get, to help performance… but what about everyone else?
- DFSORT parameters affecting storage use (II13495) → means to protect DB2
  - These can be dynamically changed for workloads using ICEPRMxx member
  - EXPMAX=% of storage for memory object and hyperspace sorting, somewhat depends on EXPOLD and EXPRES → how much can you spare
  - EXPOLD = % of storage allowed to be pushed to AUX → 0
  - EXPRES= % of storage to preserve, maybe in case of DUMPSPACE/ MAXSPACE → 16GB min in V10
  - For DB2SORT the PAGEMON parameter limits use of central storage



Sort: 256GB LPAR ~114GB page fixed

- DB2 has over 1GB in AUX - paging of 15 / second
- EXPMAX = 20% so SORT can use ~50GB max
  - Using 45GB which is actually 31% of the available (256-114)
  - EXPMAX looks at total storage configured on the LPAR, regardless of PGFIX
- Look at EXPOLD → 0 and lowering EXPMAX to stop DFSORT from stealing old storage and pushing us out to AUX
- Ensure you do not end up with >70% of the LPAR page fixed
  - If 80% is fixed (IRA400E) and address spaces become non-dispatchable

```
                        RMF V1R13  Storage Frames                  Line 1 of 328
Command ===>                                                  Scroll ===> CSR

Samples: 119      System: LSYS  Date: 10/27/15  Time: 21.00.00  Range: 120   Sec

            Service      -- Frame Occup.--  - Active Frames -  AUX    PGIN
Jobname   C Class    Cr  TOTAL   ACTV  IDLE   WSET FIXED   DIV SLOTS RATE

DP2CDBM1  S S_STCHI      29.4M 29.4M     0 29.4M 27.4M 13169  276K   15
DDL55E1   B S_BATLO      11.5M 11.5M     0 11.5M 48135   572  379K    0
CICSXFB3  S S_CICSHI     3386K 3386K     0 3386K 17185     0  4651    0
DP18DBM1  S S_STCHI      2435K 2435K     0 2435K 1487K 12866  431K    1
CICSXAJ3  S S_CICS1      1552K 1552K     0 1552K  6330     0  2125    0
CICSXAJ1  S S_CICS1      1551K 1551K     0 1551K  6325     0  2201    0
CICSXAJ2  S S_CICS1      1550K 1550K     0 1550K  6318     0  2106    0
DP1BDBM1  S S_STCHI       883K  883K     0  883K  410K 12842  118K    4
```

More on AUX storage

- When 50% of your AUX storage is in use ENF 55 message sent out and all DB2s on that LPAR will enter hard discard mode (to free off 64-bit storage) causing CPU burn (DSNV516I)
- At 70% of AUX used z/OS will mark address spaces which cannot be swapped in real as non-dispatchable
  - Customers see applications starved of CPU
  - -904 for Dynamic Statement Cache
  - IRA200E if shortage of AUX, and IRA260E if you have SCM
    - DASD AUX and SCM (Flash Express memory) used to be combined
- Do not over size LFAREA
  - Large frame (LFAREA) storage is a last resort to be stolen
  - Decomposition and coalescing of 1MB frames in LFAREA to 4k, and back again wastes CPU cycles
    - **Specify INCLUDE1MAFC on LFAREA specification (A42510)**

How do I size LFAREA for DB2?

- LFAREA = 1.04* (sum of VPSIZE from candidate buffer pools)) + 20MB + (OUTBUFF + 31-bit low private for DB2 11)

- **Do not oversize LFAREA**
  - LFAREA used ~ DB2 usage + JAVA heap (verbosegc traces)
  - Can't do anything about it until an IPL, if too small just means there is potential savings you could be missing out on
    - **- IRA127I 100% OF THE LARGE FRAME AREA IS ALLOCATED** = using it all
  - If for any reason RSM denies DB2 request for 1 MB frame, uses 4k instead
- **Decomposing 1MB frames into 4k frames (due to paging w/out FlashExpress) is CPU intensive trying to maintain the LFAREA setting**

  - MAX LFAREA ALLOCATED (4K) = Not a good sign
  - This indicates you do not have enough REAL storage needed by 4k frames, so add more real or make LFAREA smaller

What about LFAREA?

- Useful commands
  - -DISPLAY BUFFERPOOL(BP1) SERVICE(4)
    - Useful command to find out how many 1MB size page frames are being used → *DSNB999I =D2V1 4K PAGES*
      DSNB999I =D2V1 1M PAGES
  - -DISPLAY VIRTSTOR,LFAREA

NOT good, this means we broke down 1MB frames due to a shortage of 4K frames

We reserve 1/8th of real on LPAR for pageable frames, then overflows to LFAREA

IAR019I 14.37.22 DISPLAY VIRTSTOR 735
SOURCE =
TOTAL LFAREA = 4800M , 0G
LFAREA AVAILABLE = 42M , 0G
LFAREA ALLOCATED (1M) = 0M
LFAREA ALLOCATED (4K) = 4628M
MAX LFAREA ALLOCATED (1M) = 6M
MAX LFAREA ALLOCATED (4K) = 4703M
LFAREA ALLOCATED (PAGEABLE1M) = 130M
MAX LFAREA ALLOCATED (PAGEABLE1M) = 130M
LFAREA ALLOCATED NUMBER OF 2G PAGES = 0
MAX LFAREA ALLOCATED NUMBER OF 2G PAGES = 0

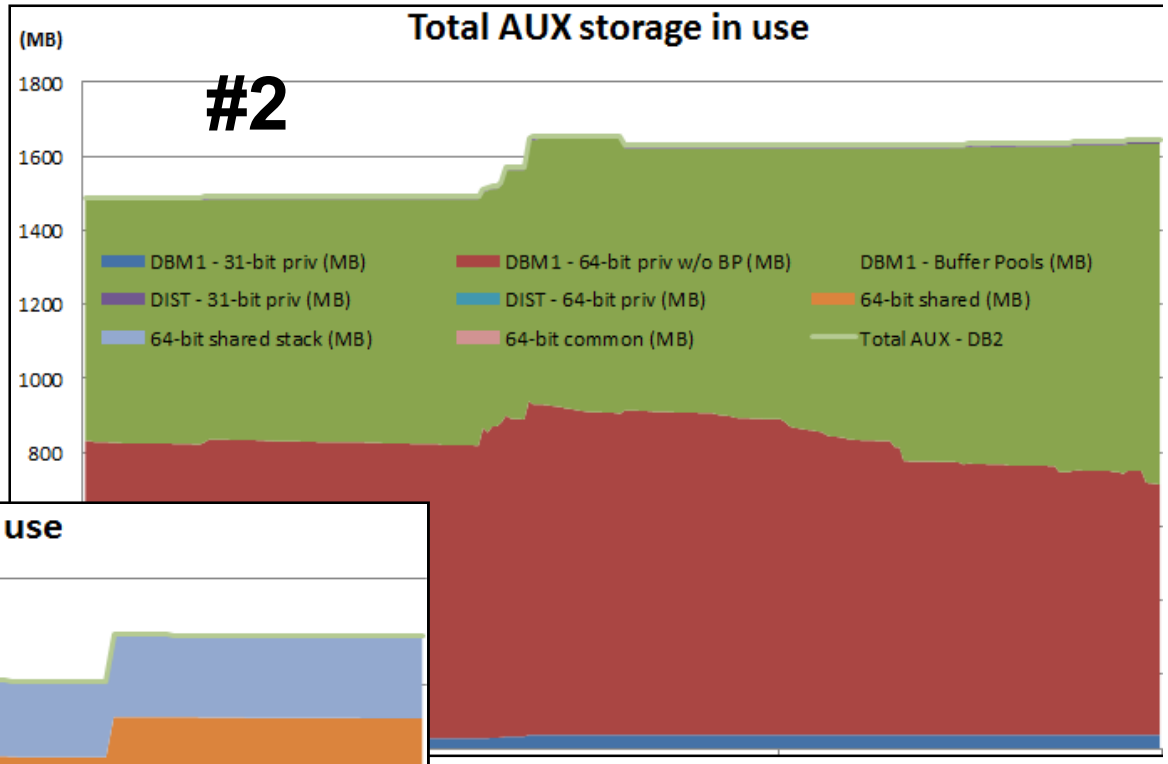
  - Shows total LFAREA, allocation split across 4KB and 1MB size frames, what is available

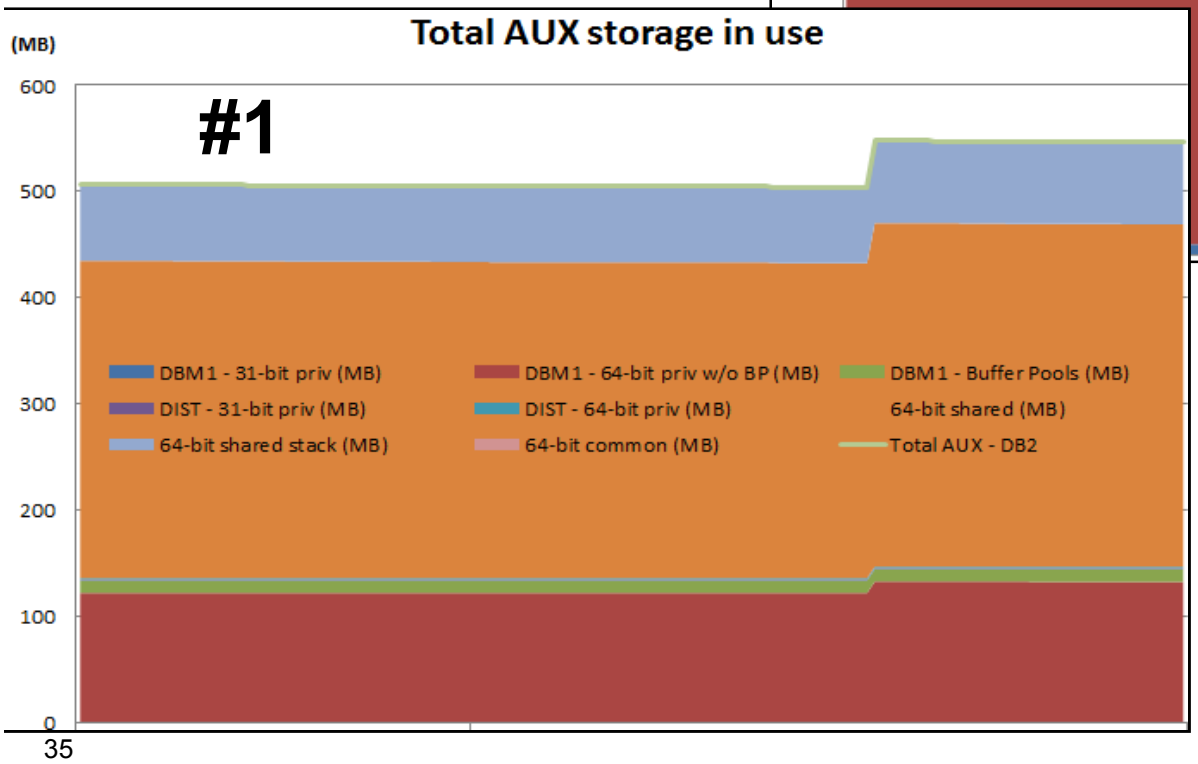
XCF Critical Paging – avoid page faults during HyperSwap

- CRITICALPAGING is a z/OS function designed to help avoid situations where a page needed for HyperSwap is paged out to AUX to a device that has been suspended
- The downside of this is a massive amount of fixed storage to include the following:
  - 31- bit common storage (both above and below 16M)
  - Address spaces that are defined as critical for paging
  - All data spaces associated with those address spaces that are critical for paging (unless CRITICALPAGING=NO was specified on the DSPSERV CREATE)
  - Pageable link pack area (PLPA)
  - Shared pages
  - All HVCOMMON objects
  - All HVSHARED objects
    - In DB2 the 64-bit SHARED houses thread working storage, statement cache, SKCT/SKPT
- Apply z/OS APAR OA44913
  - Allows z/OS to reclaim DB2 64-bit SHARED KEEPREAL=YES frames

What is in AUX now?

With CRITICALPAGING =YES HVSHARE becomes non-pageable so that leaves buffer pools and PRIVATE storage to be sacrificed





Buffer pools are not paged out in customer #1's environment, but they are in #2 causing **1 I/O for the price of 2**, no prefetch, and could have 100% buffer hit but 100% I/Os → can't trust the stats here

Buffer Pool sizing considerations

- Starting in DB2 10 the root pages of the indexes are ‘fixed’ in the buffer pool
  - How many indexes/parts do you have in your index buffer pool?
- This would affect DWQT threshold
  - (ex.) 10,000 buffers, DWQT of 30%
    - With 1,000 indexes you have basically made the DWQT threshold 20%
    - Watch for DWQT being hit multiple times per second and LC23 being elevated
      - Customer saw DWQT threshold being hit 80 times a second and LC23 at 40,000 a second
      - Application response times were significantly impacted due to being I/O bound, elapsed times increased 2-3x
  - Ideally VDWQT should be used over DWQT for efficiency of writes and avoiding latch contention

Sync I/O

- DB2 10 added a mechanism to avoid local buffer pool scans when objects go from GBP dependent to non-GBP dependent
  - This saves DBM1 SRB time, and application elapsed time
  - But depending on the amount of pseudo closes you have it can increase synch I/O for some applications that bounce in and out of GBP dependency
  - V11 APAR PI59168 addresses some XI conditions

DB2 9

GROUP BP7	AVERAGE	TOTAL
GBP-DEPEND GETPAGES	343.4K	16481954
READ(XI)-DATA RETUR	30.67	1472
READ(XI)-NO DATA RT	2.50	120
READ(NF)-DATA RETUR	190.04	9122
READ(NF)-NO DATA RT	12379.02	594193

GROUP BP12	AVERAGE
GBP-DEPEND GETPAGES	102.9K
READ(XI)-DATA RETUR	36.23
READ(XI)-NO DATA RT	0.04
READ(NF)-DATA RETUR	10.52
READ(NF)-NO DATA RT	1227.54

DB2 10

GROUP BP7	AVERAGE	TOTAL
GBP-DEPEND GETPAGES	320.4K	15380145
READ(XI)-DATA RETUR	54.67	2624
READ(XI)-NO DATA RT	16597.00	796656
READ(NF)-DATA RETUR	140.60	6749
READ(NF)-NO DATA RT	16620.69	797793

GROUP BP12	AVERAGE
GBP-DEPEND GETPAGES	91613.63
READ(XI)-DATA RETUR	24.73
READ(XI)-NO DATA RT	24369.08
READ(NF)-DATA RETUR	0.29
READ(NF)-NO DATA RT	3086.73

XI No Data RT means the page was cross invalidated in the local pool, but was not found in the GBP

PCLOSEN/PCLOSET and Synch I/O

- The default in DB2 10 is PCLOSEN=5, PCLOSET=10
  - The customer saw a 20% increase in Synch I/O after migration
  - They had moved from PCLOSET=30 → PCLOSET=10 so every 10 minutes objects without inter R/W interest would pseudo close
  - When the objects moved out of GBP dependency the local buffers would be cross invalidated
    - Next execution of the application would require entire index be read back in

OPEN/CLOSE ACTIVITY	QUANTITY	/SECOND	/THREAD	/COMMIT	
DSETS CONVERTED R/W -> R/O	9010.00	0.67	0.03	0.00	<==V9
DSETS CONVERTED R/W -> R/O	24721.00	1.72	0.07	0.00	<==V10

    - **ROT:** R/W → R/O = 10-15 a minute
  - The solution in this situation was to set PCLOSEN=32767 to disable it, and PCLOSET=45 minutes so that the object did not through pseudo close until the application ran again (every 30 minutes)

Log Write I/O

- Log Write I/O time is Class 3 time resulting from the application waiting for DB2 to synchronously write log records to disc
  - Prior to V11 the culprit was often index page splits from heavy inserts
- For GBP dependent objects if update creates an overflow records result is a forced write (synchronous) of Log records and overflow page to GBP
  - Occurs after applying PM82279 in V10
- This can significantly impact Log Write I/O class 3 suspense time if most of the rows increase in size and do not fit on the same page anymore

SYNCHRON. I/O	8:03.402141	173.7K
DATABASE I/O	5.350565	4705.31
LOG WRITE I/O	7:58.051575	169.0K
WRITE AND REGISTER	169.0K	2197415
WRITE & REGISTER MULT	55.77	725
CHANGED PAGES WRITTEN	169.5K	2203021

- Here we see log write delay for every occurrence of a page being written to the GBP (application elapsed time went from <1 minute to > 8 minutes)

Log Write I/O…

- The solution is to ensure there is enough room on the page for updated rows
  - Overflow records cause more getpages, increase elapsed time, degraded prefetch, up to 2x I/Os even without the forced write
  - PCTFREE (V10) and PCTFREE x% FOR UPDATE n% (v11)
  - Maybe even a larger page size (4k → 8k?)
- How do I know I am creating overflow records?
  - The near and far indirect references are tracked in the real time stats tables (REORGNEARINDREF)
  - Monitor the counts here before and after the application runs
  - Determine the percentage of rows overflowing and increase the free space on the pages by that amount

```
SELECT name,partition,(DEC(REORGNEARINDREF)+DEC(REORGFARINDREF))
   /DEC(TOTALROWS) AS OVERFLOW
     FROM SYSIBM.SYSTABLESPACESTATS
     WHERE TOTALROWS>0 and dbname = 'TEST15' and name= 'GLWSEMP'
      WITH UR;
```


With Sync Receive and IDTHTOIN (V11 CM)

- Customer was seeing batch jobs (utilities) timing out and missing their SLAs
  - Saw timeouts and had to manually cancel threads to let batch break in..
  - So did you see IDTHTOIN pop in the log, what about in the previous release?
    - 00D3003B in the log for threads that hit IDTHTOIN
- (DB2 11) - Now when idle thread timeout is hit DDF must issue TCPIP.DROP command to kill the socket associated with the thread
- If threads are remaining in the system longer than on DB2 10, and the idle threads are not being canceled (causing timeouts or contention with other processes), then MVS.VARY.TCPIP.DROP OPERCMDS missing
  - Get DSNL512I -111 RC = 77E800DC (EACCES/JRSAFNotAuthorized)
- Process is described in setting up DDF and UNIX system services section of the installation guide
  - http://www-01.ibm.com/support/knowledgecenter/SSEPEK_11.0.0/com.ibm.db2z11.doc.inst/src/tpc/db2z_enableddf4uss.dita?lang=en
- PI06325 – message DSNL512I is enhanced to show *socket*=EZBNMIF4_DROPCON to alert you that service failed
  - http://www-01.ibm.com/support/docview.wss?uid=swg1PI06325

References

- Techdoc for V10 and V11 MEMU2 with spreadsheet
  - http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS5279
- Subsystem and Transaction Monitoring and Tuning with DB2 11 for z/OS SG24-8182
  - https://www.redbooks.ibm.com/Redbooks.nsf/RedbookAbstracts/sg248182.html?Open
- RMF spreadsheet reporting tool
  - Link to download
    - http://www-03.ibm.com/systems/z/os/zos/features/rmf/tools/
  - InfoCenter link
    - http://pic.dhe.ibm.com/infocenter/zos/v1r11/topic/com.ibm.zos.r11.erbb200/erbzug91105.htm
  - LPAR design tool
    - http://www-03.ibm.com/systems/z/os/zos/features/wlm/WLM_Further_Info_Tools.html#Design
  - Redbook using RMF and the spreadsheet reporter
    - http://www.redbooks.ibm.com/abstracts/sg246645.html

Acknowledgements and Disclaimers

Availability. References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates.

The workshops, sessions and materials have been prepared by IBM or the session speakers and reflect their own views. They are provided for informational purposes only, and are neither intended to, nor shall have the effect of being, legal or other guidance or advice to any participant. While efforts were made to verify the completeness and accuracy of the information contained in this presentation, it is provided AS-IS without warranty of any kind, express or implied. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this presentation or any other materials. Nothing contained in this presentation is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.

Thank You
(??'s)