Db2 and the zIIP Processor: Exploitation not Abuse

—

Adrian Burke
Db2 for z/OS SWAT Team
IBM (agburke@us.ibm.com)

TechU

2018 IBM Systems Technical University

Orlando

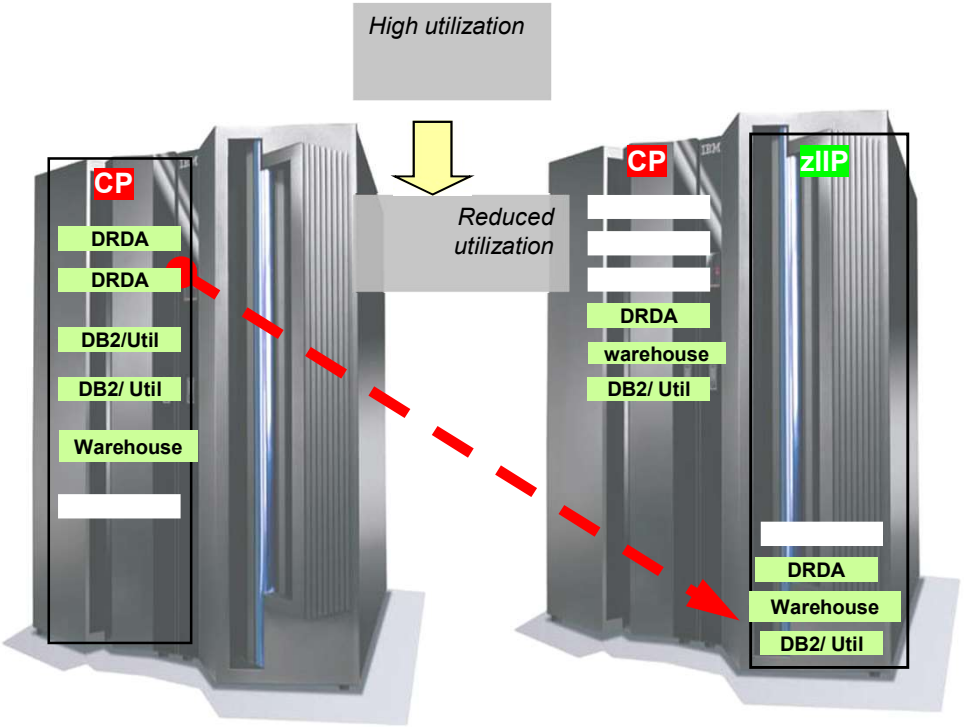
IBM

Please note

- IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.
- Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.
- The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.
- The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
- Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

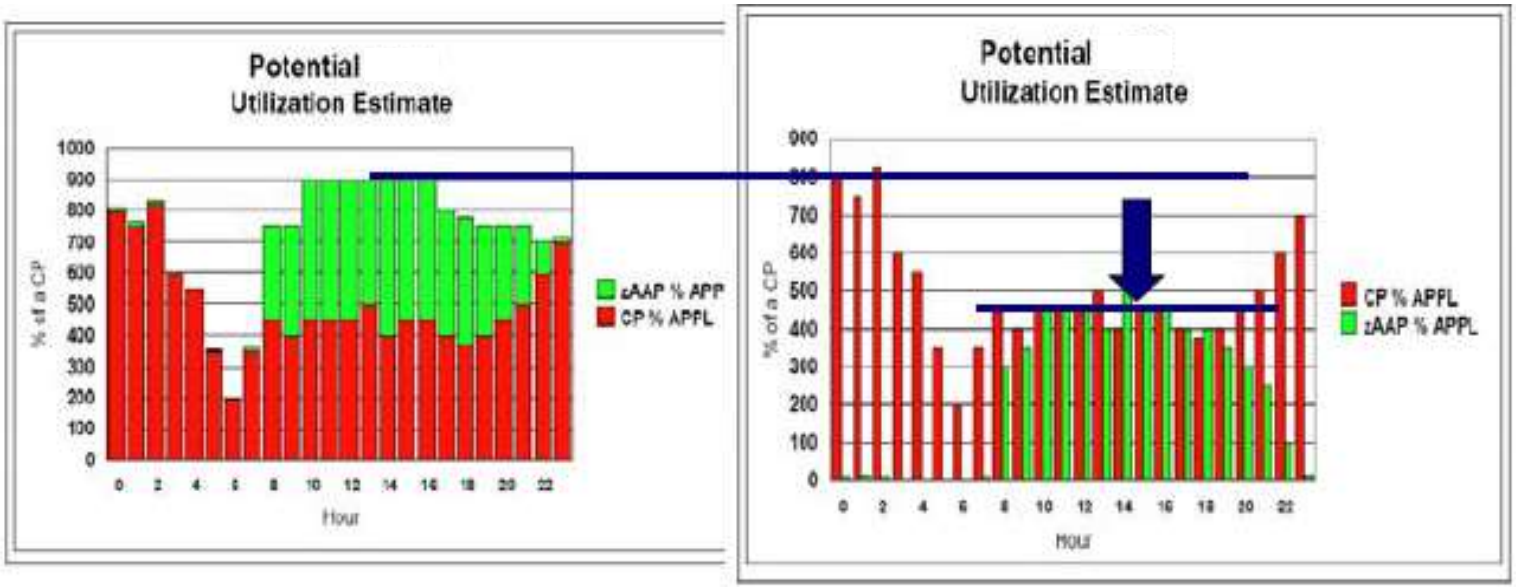
Agenda

- Background
  - Distributed access
- Capacity
  - Un-zIIPed work
- Eligibility
  - Recent enhancements
- Exploitation
  - What can I control?



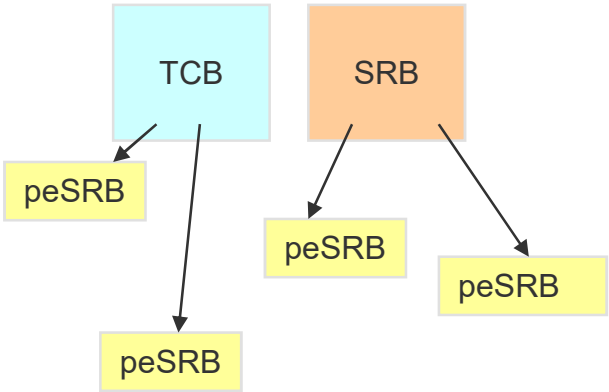
How can specialty engines help me?

- Software costs: MSU units, generally increase with the # of general processors and/or their utilization; while neither zIIP processors, nor their utilization add to the total MSU count
- Hardware costs: move work from GP to zIIP (zAAP), higher cost to lower cost processors, possibly postpone an upgrade
  - Specialty engines run at full rated speed of processor, so it could be the fastest one on the CEC
- BUT/AND…. it can also result in latent demand processing so processor utilization remains constant



Work is dispatched

- There are four types of dispatchable units in z/OS:
  - Preemptible Task Control Block (TCB)
  - Non-preemptible Service Request Block (SRB)
  - Preemptible Client Service Request Block (client SRB)
  - Preemptible Enclave Service Request Block (enclave SRB)
- Some are zIIP eligible
  - IBM moves TCB and preemptible SRB work to enclaves as a way to increase offload



DBM1 SRBs (* means data sharing)
- Asynchronous I/O (enclave SRB in V10)
- Memory management
- Prefetch (enclave SRB in V10)
- Real time stats
- *Castout* (V11)
- *P-lock negotiation*
- *GBP checkpoints*
- Backout (preemptible V10)

DBM1 TCBs
- Open/close
- Pre-format/ extend
- Full system contraction

Why not SNA?

- **If DB2 for z/OS workload comes over TCP/IP and is DRDA compliant, a portion of that DB2 workload is eligible to be redirected to the zIIP**
- Many customers still use DRDA over SNA for DB2 z/OS to DB2 z/OS calls
  - As of DB2 9 SNA incurs overhead due to DIST going to 64-bit addressing
  - Look in the statistics long report and compare the SRB times in the DDF Address space CPU
    - The PREEMPT IIP SRB time should be => PREEMPT SRB if the DRDA work is coming in over TCP/IP and thus zIIP eligible
- This customer migrated from SNA to TCP/IP and measured a 24 hour period before and after
  - 58% of the CPU used by DIST address space was offloaded to the zIIP
  - The CPU per commit was reduced by 66% due to running in 64-bit mode
    - Watch out if you use INBOUND AUTHID translation in SNA, not there in TCP/IP

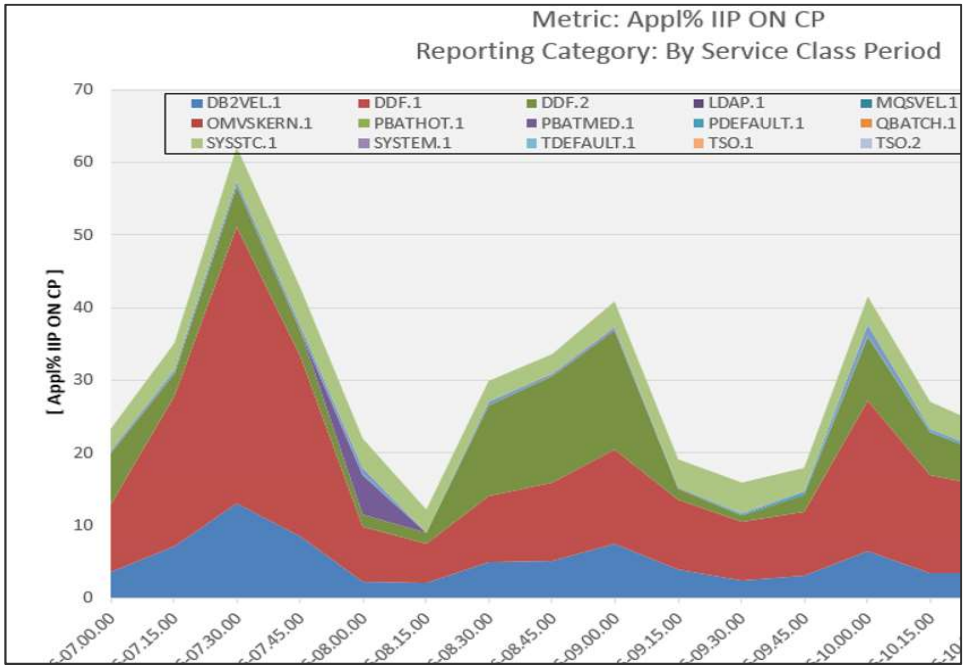
	CPU TIMES	TCB TIME	PREEMPT SRB	NONPREEMPT SRB	CP CPU TIME	PREEMPT IIP SRB	CP CPU /COMMIT
SNA	DDF ADDRESS SPACE	15.759816	11:40:48.726492	3:25:35.999739	15:06:40.486046	9:45.940413	0.008500
TCP/IP	DDF ADDRESS SPACE	14.758614	6:14:38.618730	22:31.612655	6:37:24.989999	9:06:58.739546	0.002866

- **Also with Db2 11 and Db2 12 Db2 Native REST services are supported with the same zIIP offload as DRDA compliant work**
  - **http://www-01.ibm.com/support/docview.wss?uid=isg1II14827**
  - Whether using z/OS Connect EE or not, Db2 runs REST services in Enclave SRB mode and the work runs on a DBAT just like any other distributed work

CAPACITY

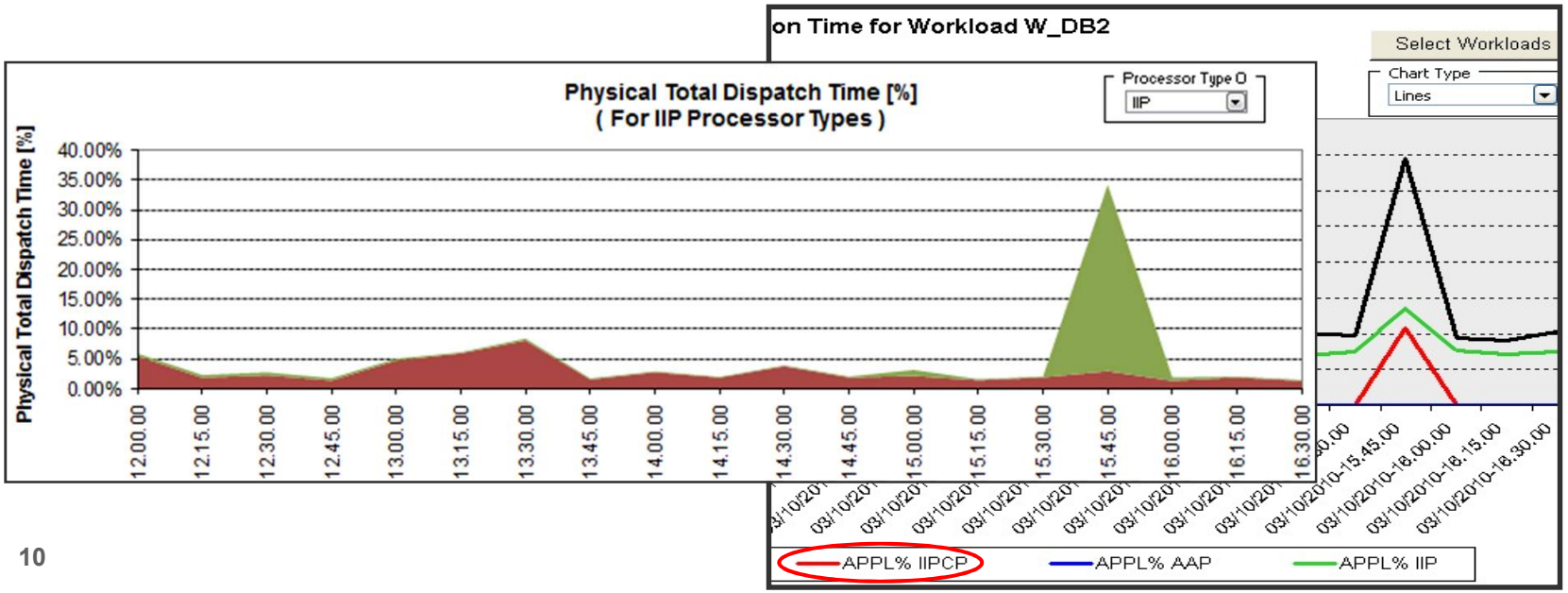
Measuring zIIP overflow

- Capacity planning should monitor zIIP overflow to the GCPs, ***not*** absolute utilization
- The WLM activity report (SMF 72-3) records zIIP eligible work that ran on a GCP as a % of utilization of a single GCP (APPL% IIPCP CPU)
  - Broken down by WLM service/report class in RMF
- zIIPs always run at speed of a 701 processor so this workload may only need 30% of a zIIP if it is 2x speed of the GCPs on the box
- zIIP redirect means work waited in a queue for a zIIP, as well as aggravating RNI of the LPAR
  - Relative Nesting Intensity (RNI) affects MIP consumption → no L1, L2, or L3 CPU cache hit if work moves from a zIIP to GCP
  - 20% RNI savings could be 10% MIPS savings



zIIP overflow

- How many zIIPs do you need (this scenario 12:1 ratio CP to zIIP)
  - zIIP eligible work went to CP either because zIIP is overloaded - Red line on graph (APPL% IIPCP) – missed opportunity for savings
  - **Needs Help algorithm** ensures work does not pile up waiting on zIIP
  - Must have enough capacity to absorb spikes, not just typical offload
    - **Size the zIIP for the spikes, it doesn't matter if it is only 10% utilized outside of the 4 hour rolling average window**
- Law of probability for many CPs vs. zIIPs (next slide)



zIIP Overflow

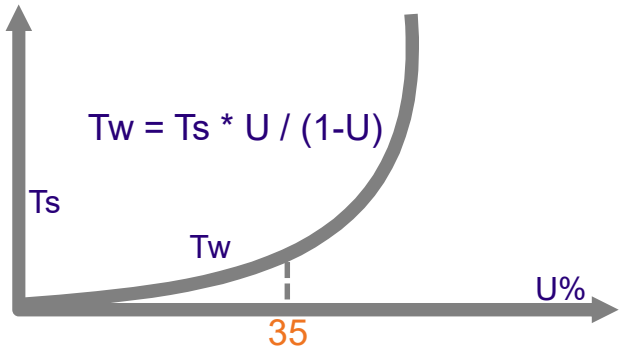
- If 12 CPs are 65% (0.65^12) utilized then each CP is 0.5% instantaneously busy
  - If 1 zIIP is 35% busy then 35% of the time it is 100% busy
  - So with 'needs help' algorithm it is likely some zIIP eligible work could fall back to a CP
    - See IIPHONORPRIORITY, later slide
- Markov's Equation is based on 1 server (CP) in steady state
  - As Utilization approaches 100% wait time approaches ∞
  - This will cause more work to overflow to a CP starting at around 35% utilization of a single zIIP processor
  - More zIIPs = more offload

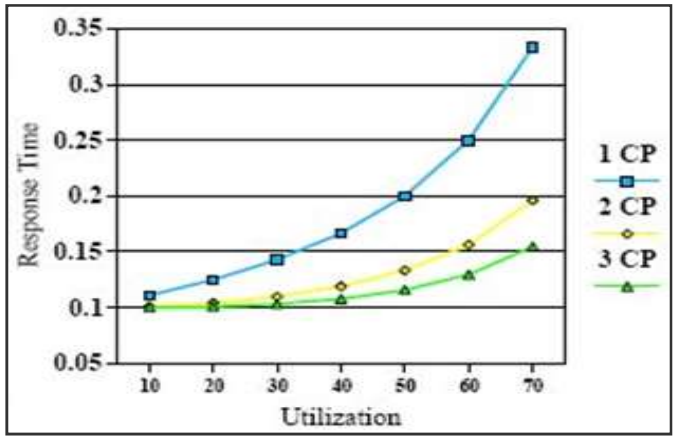
Tw = wait time of transaction

Ts = service time of transaction

U= utilization

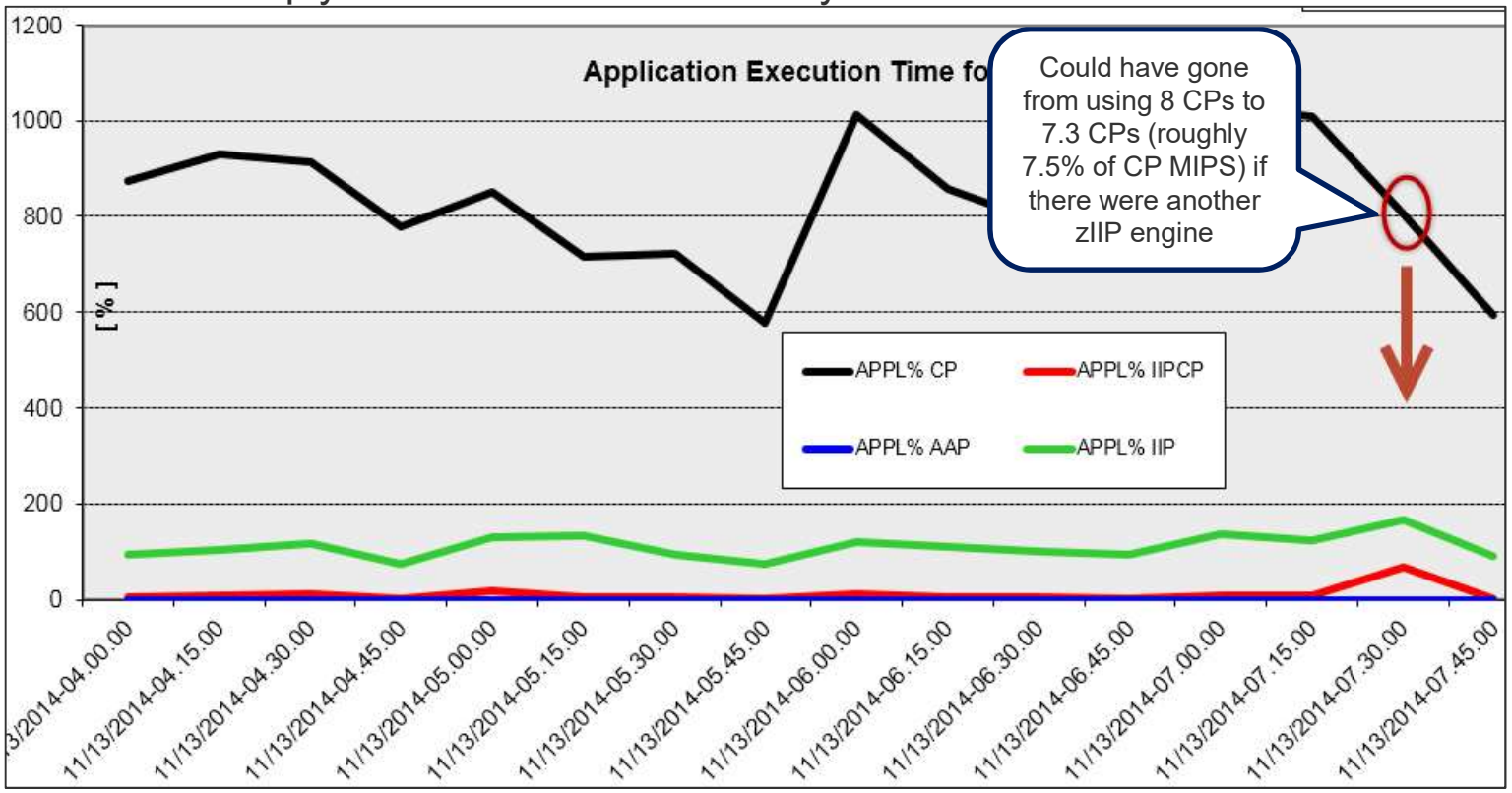
The knee of the curve occurs at 35% for 1 processor, thereafter Tw increases drastically





System zIIP Shortage…

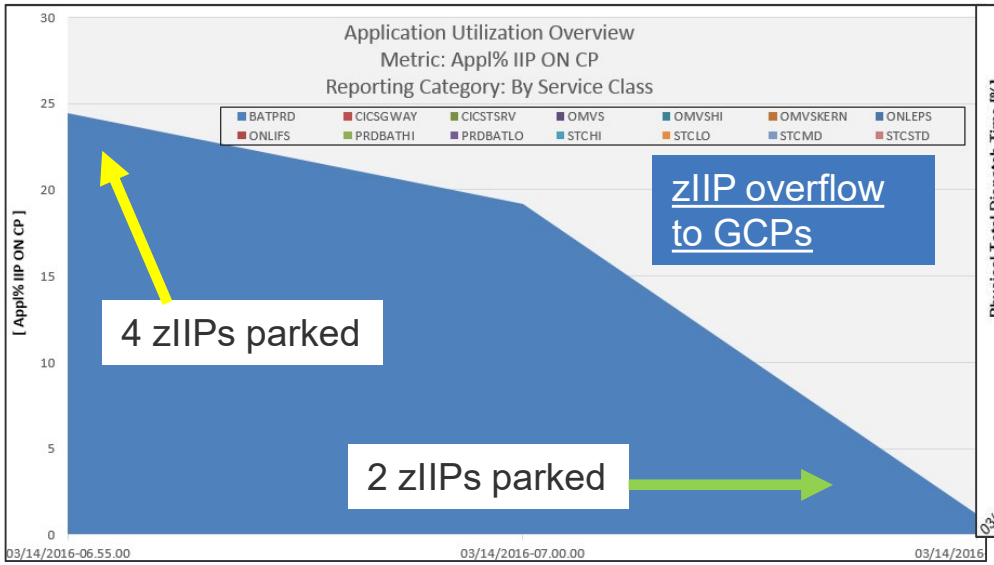
- Looking at a Customer's 4 hour rolling average peak for the month, there are just over 10 GCPs in use and 2 zIIPs available
  - When zIIP eligible work ran on the GPs it represented about 7.5% of the chargeable MIPs for DB2 on the system during that interval, which could affect the MLC bill
  - In this case the CPs were full-speed, but if they were knee-capped you would need to multiply the APPL% IIPCP CPU by the MSU ratio difference

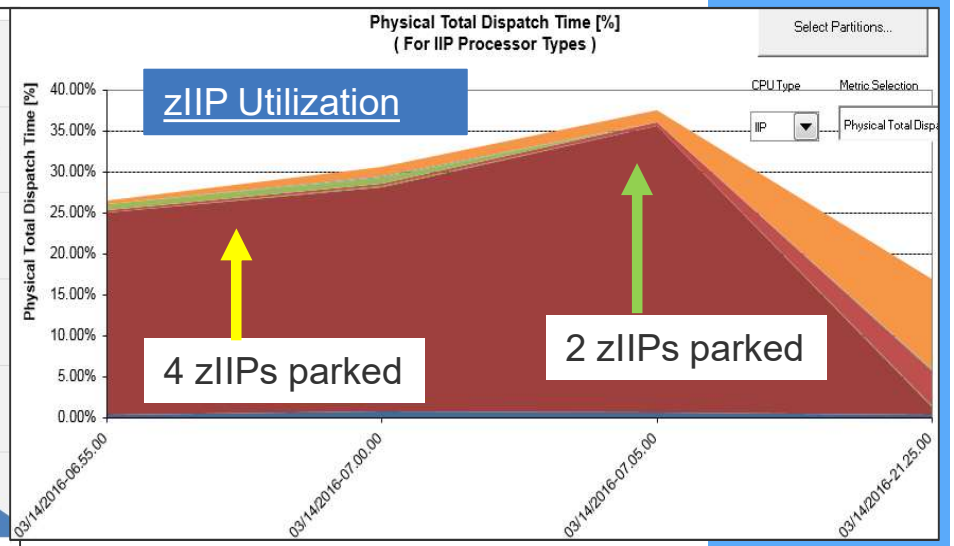


zIIP Overflow…LPAR Weights

- Hiperdispatch is VERY sensitive to the relative LPAR weights (Vertical HIGH/MED/LOW polarity)
  - 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 in CPU activity
  - Rush of DRDA requests, Utilities, or SQL CPU parallelism leads to overflow
- ** Need enough Dedicated zIIPs (VH's) to handle peaks

---------NUMBER OF WORK UNITS----------			
CPU TYPES	MIN	MAX	AVG
CP	1	51	9.5
IIP	0	306	2.1





zIIP Overflow…

- Aside from simple queue theory there are other reason's you could be seeing zIIP work spill-over to the GCPs
- z/OS local lock contention and I/O interrupt CPU for zIIP eligible tasks is reported as zIIP eligible time in RMF, but cannot in-fact run on a zIIP processor
  - This local lock is used for storage acquisition and there is only 1 per address space
- If you have single digit % of IIPCP CPU could be due to lock contention or I/O
  - Evidence of contention in PROMOTED for lock (LCK) field in WLM Activity Report
  - This shows CPU used to promote waiters <u>not</u> IIPCP CPU time, but implies relative use of the local lock, and if the task was zIIP eligible this time would be reported in IIPCP
  - Do the math to determine if CPU time = % IIPCP
- In order to prove that was causing the IIPCP CPU % you could turn IIPHONORPRIORITY= NO
  - **BUT** then you loose system agent offload in V11!!
  - With z/OS 2.1 + OA50845 Honor Priority can be done at the WLM Service Class level
    - Leave Honor Priority=DEFAULT (default) for DB2 address spaces
- If it is in-fact lock contention there is no way to tune this away

DDF enclaves

--PROMOTED--	
BLK	0.000
ENQ	0.000
CRM	0.000
LCK	0.659
SUP	0.000

DBM1 address space

--PROMOTED--	
BLK	0.000
ENQ	0.000
CRM	0.000
LCK	8.699
SUP	0.000

ELIGIBILITY

IBM

zIIP Eligibility

Release	Function	Amount Redirected	Pre-reqs
DB2 10	1. All of DB2 v8 and 9 offload++ 2. RUNSTATS 3. Prefetch and deferred write processing including index compress/decrompress 4. Parallelism enhancements 5. multi-version XML clean-up	1. BUILD phase, Remote Native SQL procs, parallelism, 60% DRDA requests 2. Basic RUNSTATS for table, NO Histogram, DSTATS, COLGROUP… BUT index stats almost all offloaded (not DPSIs) 3. 100% (roughly 80% of DBM1 SRB time) 4. Parallelism more likely (80% of child tasks) 5. All of it	1. DB2 10/ z/OS 1.10 2. Run RUNSTATS, no inline STATS 3. Shows up in DBM1 SRB time 4. V10 NFM with rebind 5. PM72526
Other stuff	1. IPSec 2. Global Mirror for z/OS (formerly Extended Remote Copy) 3. HiperSockets for Large messages 4. DFSORT 5. zAAP on zIIP	1. Encryption processing, header processing and crypto validation (93% for bulk data movement) 2. Most System Data Mover processing 3. Handles large outbound messages (multiple channel paths given to SRBs) 4. Sorting of fixed length rows (10-40% Utility), memory object work file sorts 5. zAAP eligible work can move to zIIP	1. N/A 2. N/A 3. GLOBALCONFIG ZIIP IQDIOMULTIWRITE 4. PM62824 and z/OS 1.12 5. z/OS 1.11 base or 1.9 or 1.10 w/ APAR OA27495 / OA38829 if both

zIIP Eligibility

Release	Function	Amount Redirected	Pre-reqs
DB2 11	1. More RUNSTATS 2. LOAD REPLACE with dummy input 3. Most of the system engines (GBP write, castout, log write/ prefetch,) 4. Index pseudo delete clean-up 5. PARAMDEG_DPSI	1. COLCARD, FREQVAL, HISTOGRAM statistics, including inline stats (80%, possibly more) 2. 100% of delete processing eligible 3. 100% eligible 4. 100% eligible 5. 100%	1. N/A 2. N/A 3. N/A 4. INDEX_CLEANUP_THREADS >0 5. Parallel query access through DPSI parts
DB2 12	1. Parallel child tasks 2. DRDA fast load 3. RELOAD phase of REORG and LOAD 4. Fast Traversal Block for buffer pools 5. GRECP/LPL retry agents	1. 100% 2. 100% movement of data blocks to LOAD 3. ~59% for REORG, ~99% for LOAD 4. 100% of parent daemon 5. 100%	1. CPU query parallelism 2. Fast load from client 3. N/A 4. Enable FTBs 5. N/A
Other stuff	1. Db2 Native REST Services 2. z/OS Connect Adaptor 3. DFSORT…DB2SORT 4. ZAAP on zIIP	1. ~60% (same as any DBAT work) 2. 100% 3. Sorting of fixed length rows (10-40% Utility), memory object work file sorts… 10-20% for DB2SORT 4. 100%	1. PI70652 2. JSON API access to DB2 z/OS data 3. PM62824 4. zAAP support removed with z13

IBM

Results of zIIP maintenance

- Pre- PM12256 – some trans run on both CP and zIIP

```
CLASS 2 TIME DISTRIBUTION
-------------------------------------------------
CPU     |=========> 19%
SECPU   |=> 3%
```

- After – PM12256 – trans run either on CP or zIIP

```
CLASS 2 TIME DISTRIBUTION
-------------------------------------
CPU     |
SECPU   |======> 12%
```

- After – PM28626 – longer running trans run on both CP and zIIP
  - Less noticeable elapsed time difference for customers with knee-capped general CPs

```
CLASS 2 TIME DISTRIBUTION
---------------------------------------------
CPU     |===================> 37%
SECPU   |============================> 55%
```

- **THOUGHT...
  - Heuristics for swapping from zIIP back to GCP aligns closely with CPU query parallelism.. So if you have some tasks which in a UOW run on both GCP and zIIP… parallelism could save MIPS (processor cache coherency and give you more zIIP offload avoiding the switch

IBM

Asynchronous I/O (V10+)

- In DB2 10 prefetch and deferred write are zIIP eligible
  - Increase due to index I/O parallelism/ index list prefetch for disorganized indexes/ access path changes/ more dynamic prefetch in V9,V10

```
DB2 VERSION: V8                                    SCOPE: MEMBER                                                    TO: 09/10

---- HIGHLIGHTS ----------------------------------------------------------------------------------------------------
INTERVAL START  : 09/09/11 05:30:01.83    SAMPLING START: 09/09/11 05:30:01.83    TOTAL THREADS      :    90.00
INTERVAL END    : 09/10/11 05:00:02.70    SAMPLING END  : 09/10/11 05:00:02.70    TOTAL COMMITS      :  6328.8K
INTERVAL ELAPSED:      23:30:00.864709    OUTAGE ELAPSED:              0.000000    DATA SHARING MEMBER:      N/A

CPU TIMES                                   TCB TIME      PREEMPT SRB   NONPREEMPT SRB       TOTAL TIME  PREEMPT IIP SRB
------------------------------  ---------------  ---------------  ---------------  ---------------  ---------------
SYSTEM SERVICES ADDRESS SPACE        1:39.995961         0.000000      3:25.079924      5:05.075886              N/A
DATABASE SERVICES ADDRESS SPACE      1:31.822012         0.000000  12:28:38.995808  12:30:10.817820         0.000000
IRLM                                    0.456105         0.000000      3:02.893287      3:03.349391              N/A
DDF ADDRESS SPACE                       2.730084  20:28:36.142998     30:35.615420  20:59:14.488502  19:33:32.868978

TOTAL                                3:15.004163  20:28:36.142998  13:05:42.584438  1 09:37:33.7316  19:33:32.868978
```

```
DB2 VERSION: V10                                   SCOPE: MEMBER                                                    TO: 11/11

---- HIGHLIGHTS ----------------------------------------------------------------------------------------------------
INTERVAL START  : 11/10/11 06:09:00.00    SAMPLING START: 11/10/11 06:09:00.00    TOTAL THREADS      :   290.00
INTERVAL END    : 11/11/11 06:06:00.00    SAMPLING END  : 11/11/11 06:06:00.00    TOTAL COMMITS      : 10749.2K
INTERVAL ELAPSED:      23:57:00.000072    OUTAGE ELAPSED:              0.000000    DATA SHARING MEMBER:      N/A

CPU TIMES                                   TCB TIME      PREEMPT SRB   NONPREEMPT SRB       TOTAL TIME  PREEMPT IIP SRB
------------------------------  ---------------  ---------------  ---------------  ---------------  ---------------
SYSTEM SERVICES ADDRESS SPACE        2:26.595613         2:14.698997        13.547515      4:54.842125              N/A
DATABASE SERVICES ADDRESS SPACE      1:04.360185   5:49:17.448125        11.274434   5:50:33.082744   4:25:03.509555
IRLM                                    0.032864         0.000000      3:39.871402      3:39.904266              N/A
DDF ADDRESS SPACE                       6.096981  2 22:30:18.7722     56:23.794572  2 23:26:48.6638  1 11:39:09.8193

TOTAL                                3:37.085643  3 04:21:50.9193   1:00:28.487923  3 05:25:56.4929  1 16:04:13.3288
```

Asynchronous I/O (V10+)…

- Index I/O Parallelism for updates
  - If there are more than 2 indexes on a table (clustering index does not count) or 2 if the table is defined with APPEND, HASH, or MEMBER CLUSTER
    - DB2 detects an I/O delay we use sequential prefetch engine to do the I/O for each index leaf page in parallel
  - You will see S.PRF.PAGES READ/S.PRF.READ = 1.00 in the statistics report for index buffer pools
    - Use IFCID 357-358 to trace it
    - zParm INDEX_IO_PARALLELISM
      - =YES (default)
    - VPPSEQT or VPSEQT = 0**
      - Disables it at BP level
      - PREF.DISABLED-NO BUFFER
        - » Will be non-0

SEQUENTIAL PREFETCH REQUEST	22308.00
SEQUENTIAL PREFETCH READS	0.00
PREF.DISABLED-NO BUFFER	22308.00

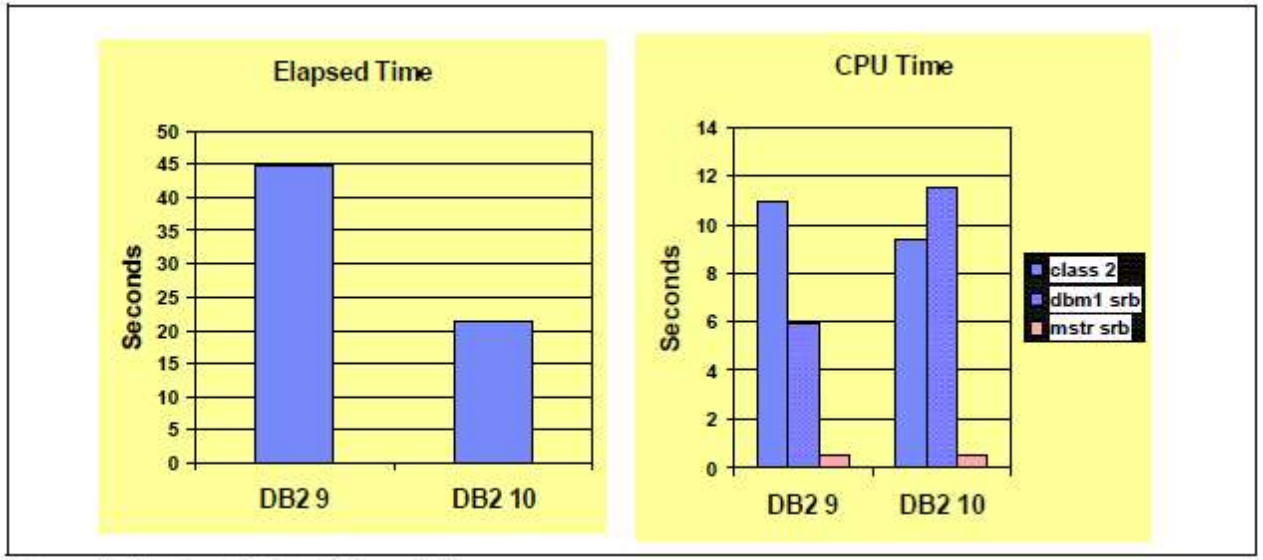


Figure 2-14 Insert index I/O parallelism

Asynchronous I/O (V10+)…

- What happens if Engines are starved of zIIP?
  - Other Read / Write I/O events and time per event will increase
  - PREF. DISABLED – NO READ ENG could increase
  - SYNC I/O SEQ / Sync Write
- Customers have seen batch programs miss their timing windows
- Even if prefetch <u>is not</u> used, DB2 may try to schedule it, and app still sees delays with BP hit and no I/Os
  - Prefetch delayed waiting on zIIP
  - Increased elapsed time/CPU

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

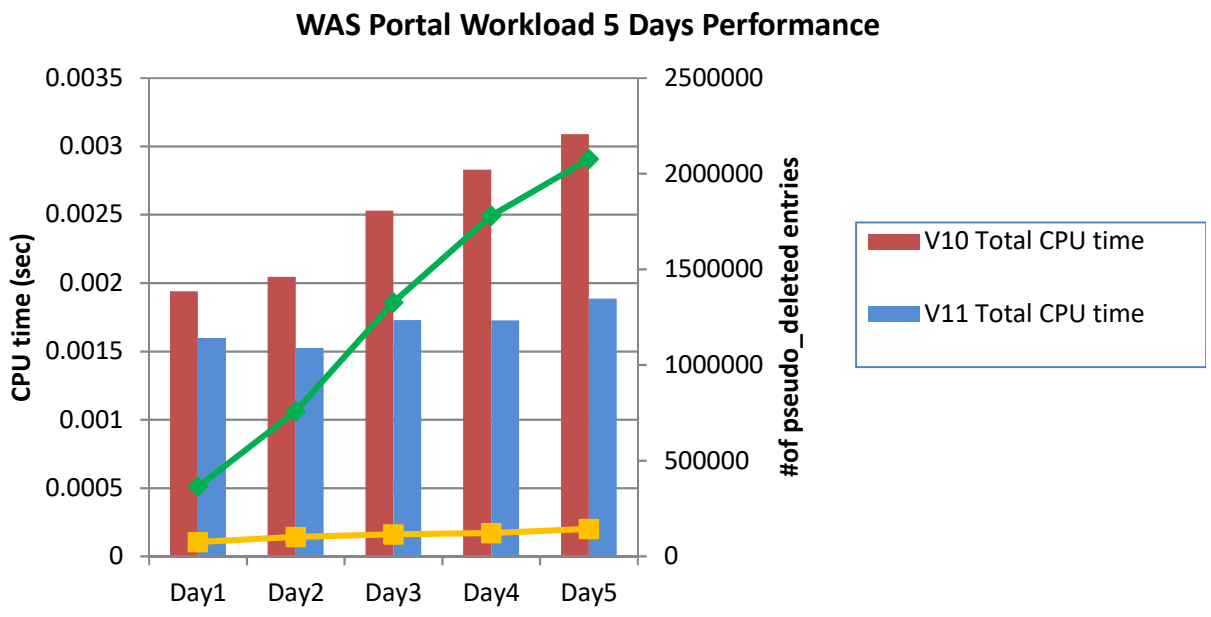
More zIIP in DB2 11

- Finally the majority of RUNSTATS (DSTATS), as well as INLINE stats
- 100% of delete processing with LOAD REPLACE
- Q REP: decompress and decode operations of capture process
- Index pseudo delete child task time will show up under the DBM1 SRB (PREEMPT IIP SRB)
- DPSI parallelism agents (PARAMDEG_DPSI)
- Log write I/O and log prefetch (MSTR) all go to the zIIP Roughly 10-20% of MSTR SRB
  - DBM1 saw another 10-15% additional zIIP offload (larger for heavy data sharing)
    - GBP castout (300), GBP writes (300)
    - Already had prefetch engines (600), deferred write engines (300)

CPU TIMES	TCB TIME	PREEMPT SRB	NONPREEMPT SRB	CP CPU TIME	PREEMPT IIP SRB
SYSTEM SERVICES ADDRESS SPACE	7:53.094670	1:38:43.086689	3:18.683030	1:49:54.864388	13:22.349651
DATABASE SERVICES ADDRESS SPACE	2:04.784117	41:56.094613	13:20.062439	57:20.941169	15:10:25.381584
IRLM	0.119303	0.000007	32:05.263272	32:05.382582	0.000000
DDF ADDRESS SPACE	1:08.358258	2:08:39.436318	1:11.055976	2:10:58.850552	2:11:30.752914

Automatic Pseudo Deleted Index Clean-up…

- Up to 39% DB2 CPU reduction per transaction in DB2 11 compared to DB2 10
- Up to 93% reduction in Pseudo deleted entries in DB2 11
- Consistent performance and less need of REORG in DB2 11
- Avoid possible wasted…
  - Getpages
  - I/Os
  - Prefetch
  - Deadlocks on insert trying to reuse deleted RID



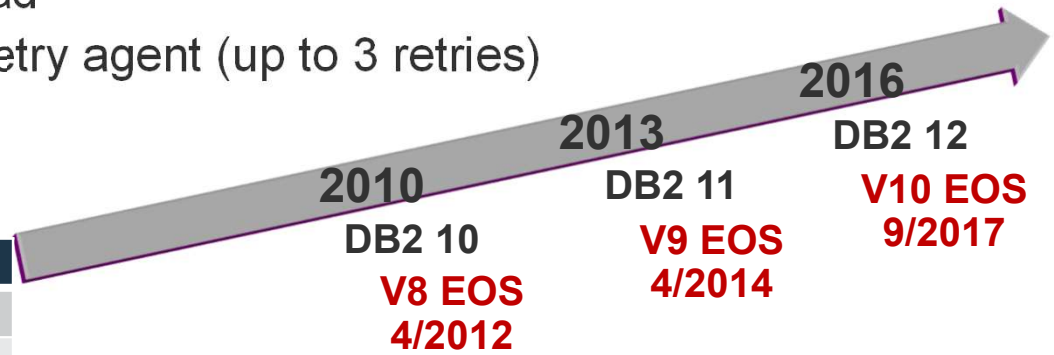
Automatic Pseudo Deleted Index Clean-up (V11)

- Autonomic solution provided in CM and turned on automatically for all indexes
  - Automatic clean-up of pseudo-deleted index entries in index leaf pages
  - Automatic clean-up of pseudo-empty index pages
  - Designed to have minimal or no disruption to concurrent DB2 work
  - Clean-up is done under system tasks, which run as enclave SRBs and are zIIP eligible
    - Parent thread (one per DB2 member) loops through RTS to find candidate indexes
    - Child clean-up threads only clean up an index if it already is opened for INSERT, UPDATE or DELETE on the DB2 member
- Clean-up is customizable
  - Can control the number of concurrent clean-up threads or disable the function using zparm INDEX_CLEANUP_THREADS
    - 0=Disable, 1-128, 10 is default
    - Monitor with IFCID 377
  - Entries in new Catalog table SYSIBM.SYSINDEXCLEANUP
    - Define when / which objects are to be considered in a generic way

IBM

DB2 12 and zIIP

- Expand use for parallelism – 100% of child tasks
- Increase portions of utility processing (RELOAD)
  - About 99% for LOAD and 59% for REORG
- z/OS Connect allows more mobile apps access
  - z/OS Connect runs in WebSphere Liberty Profile
  - Native restful interface for DIST does not need z/OS Connect (V11 & 12)
- Less wasteful prefetch, less chance of prefetch disabled due to no engine
- In-memory bufferpool enhancements (FTB parent task)
- DRDA FastLoad
- GRECP/LPL retry agent (up to 3 retries)



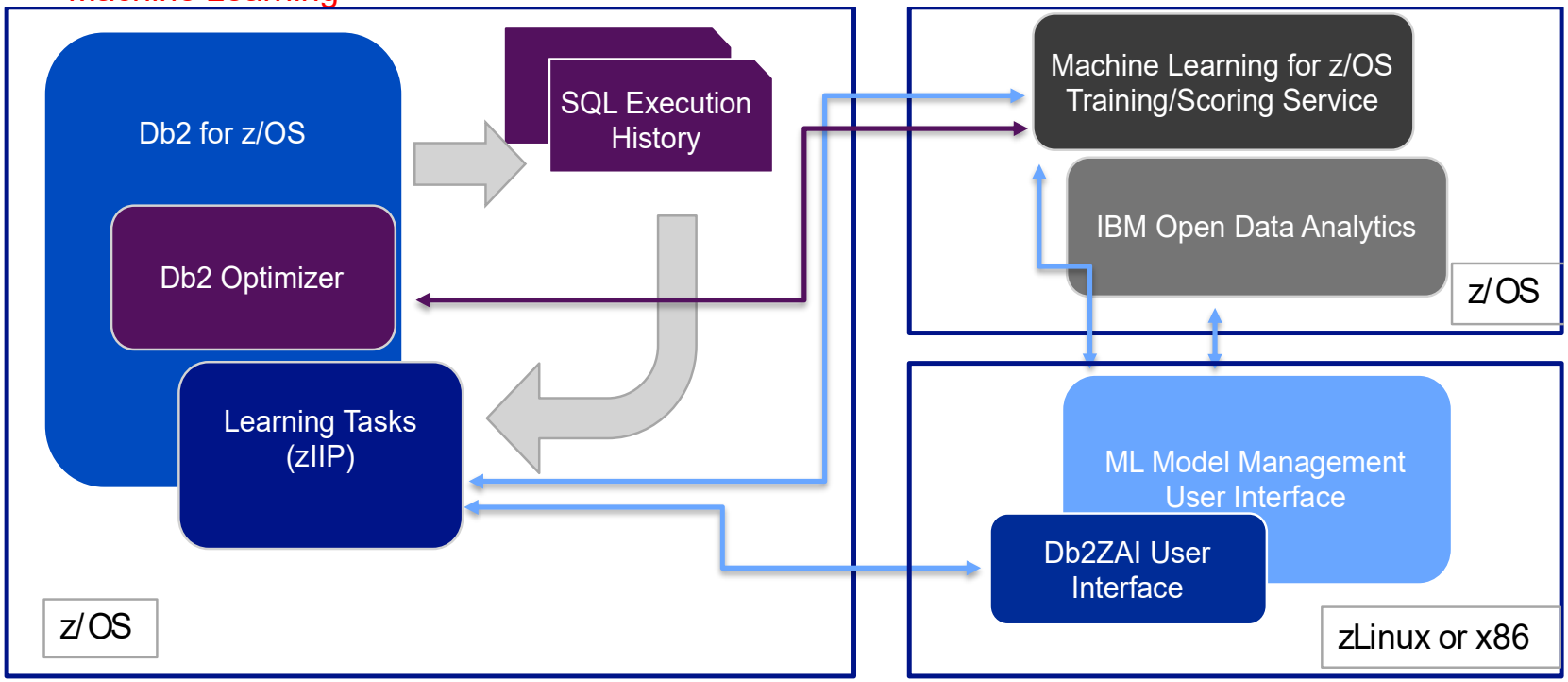
Version	GA
V7	3/2001
V8	3/2004
V9	3/2007
V10	10/2010
V11	10/2013

DB2 12 ESP started in March, GA'ed on Oct 21st, 2016

Db2ZAI-V1.1 and V1.1.0.1

- Optimizer utilizes Machine Learning to improve access paths

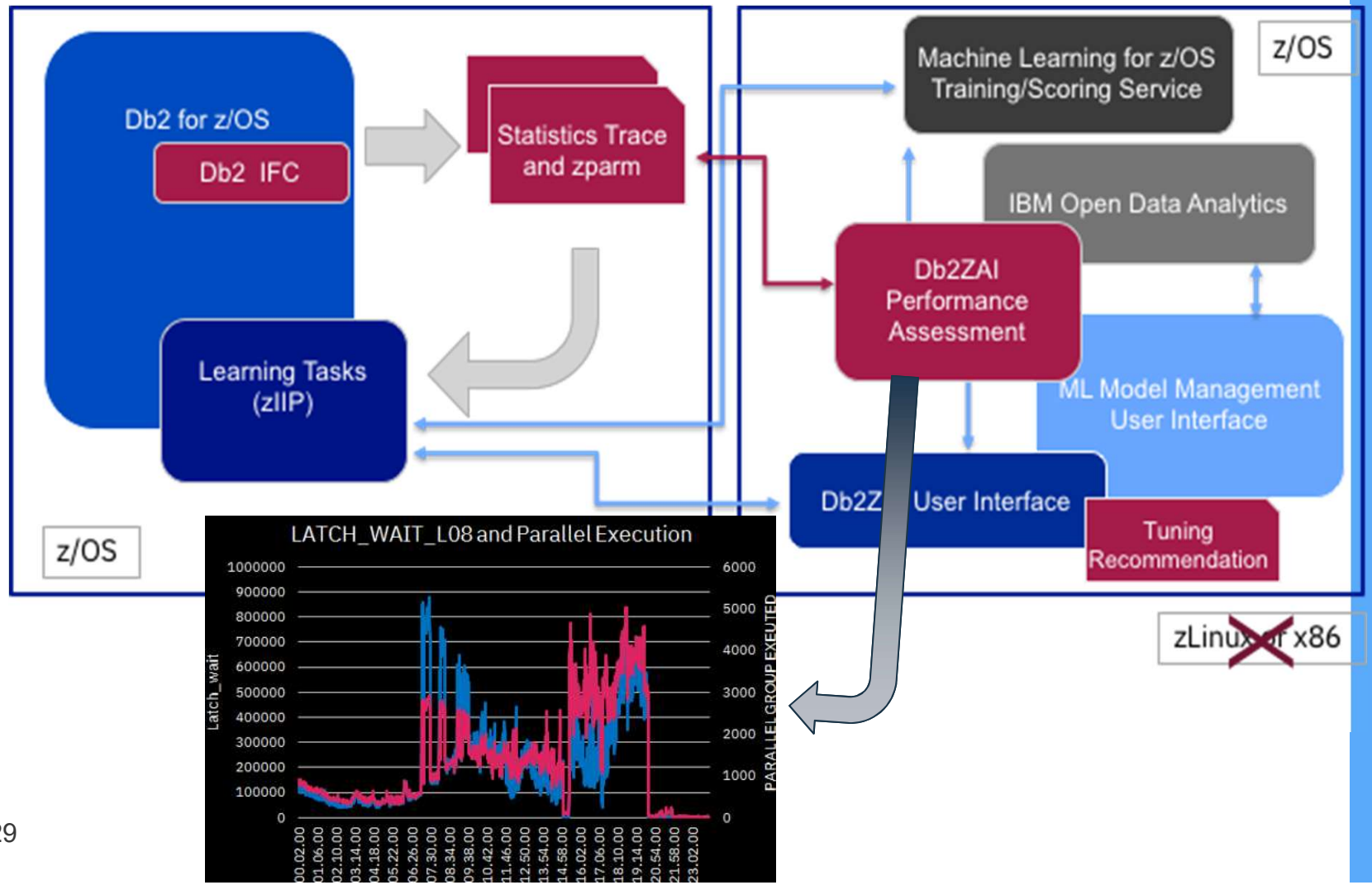
Better Access Path Selection for SQL statement with improved prediction / estimate thru Machine Learning



Announcing… Db2ZAI-V1.2!

- Db2 zAI pulls IFCIDs to monitor and improve performance

Exception analysis and tuning recommendation based on your workload thru Machine Learning



Asynchronous I/O (V12)…

- More prefetch engines are available for use
  - Moved from 600 to 900 engines per DB2 subsystem
    - Hidden ZPARM SPRMRDU controls the number
    - Still uses ESQA and some below the bar storage so don't go crazy
- Remove unnecessary prefetch scheduling in V12
  - Tracks Dyanamic Prefetch failures
    - If last 3 prefetch requests did not result in prefetch I/O
      - Disable dynamic prefetch in the pool
    - First Synchronous sequential I/O detected
      - Dynamic prefetch is re-enabled
- Saves
  - zIIP cycles
  - Unnecessary other READ I/O class 3 delays
  - Prefetch disabled NO READ ENGINE
  - LC24 contention caused by multiple prefetch requests against the same page

Customer has a 70:1 ratio of requests vs. scheduled prefetch

Dsnb414i	Dsnb414i
Dynamic Prefetch Requests	Dynamic Prefetch I/O
36,973,390	550,642

Ex. LC24 contention at 250k per second caused DBM1 zIIP spike

DSAS - SRB TIME	DSAS - PREEMPT SRB	DSAS - PREEMPT IIP SRB
0.010091	0.008489	0.133704
1.400117	1.398301	8.189153

Encryption… Function Level 502

- Encryption of data in motion
  - Encryption of data due to deferred write, castout is zIIP eligible for DBM1 agents
  - Encryption of log records is zIIP eligible in MSTR address space
- Decryption of data in motion
  - Neither synchronous reads nor prefetch is zIIP eligible..Why?
  - Decryption cost is accounted for in I/O interrupt time (IIT) and NOT zIIP eligible (neither sync I/O nor prefetch)
    - Stats shown with PI92652
- zIIP offload also shows up in SMF113 for CPACF
- NEW SMF 70 record for 4-hour rolling avg impact-OA54404

Overall	Service Time (s)							
Duration (s)	RCLASS=IDAALOAD		RCLASS=TCPIP		RCLASS=SYSLOGD		TOTAL	
10068.88	JES2 R$LOAD%%		TCPIP,PAGENT,CSF,TRMD,NSSD,IKED		SYSLOGD			
	CP	6269.99	CP	12898.70	CP	3218.02	CP	22386.71
	IIPCP	0.00	IIPCP	12476.84	IIPCP	0.00	IIPCP	12476.84
			Offload%	96.73%			Offload%	55.73%

Figure 10-19 Service time of load with encryption of data in motion

EXPLOITATION

Stored Procedures with zIIPs

- If invoked remotely portions of stored procedures are zIIP eligible
  - Native stored procedures represent the most efficient offload
  - Internal tests showed a remote call to an NSP was cheaper/faster than a straight dynamic JDBC call!!

Language	Base Billable Cost	Billable Cost after zIIP and/or zAAP acceleration
COBOL stored proc	1X (Baseline)	.74x
C stored proc	1.02x	.83x
Remote SQLJ	1.78x	1.06x
SQLJ stored proc	1.71x	1.16x (zIIP + zAAP)
JDBC stored proc	2.19x	1.54x (zIIP + zAAP)
External SQL stored proc	1.62x	1.49x
Native SQL stored proc	1.07x	.47x

IBM

Parallelism offload %

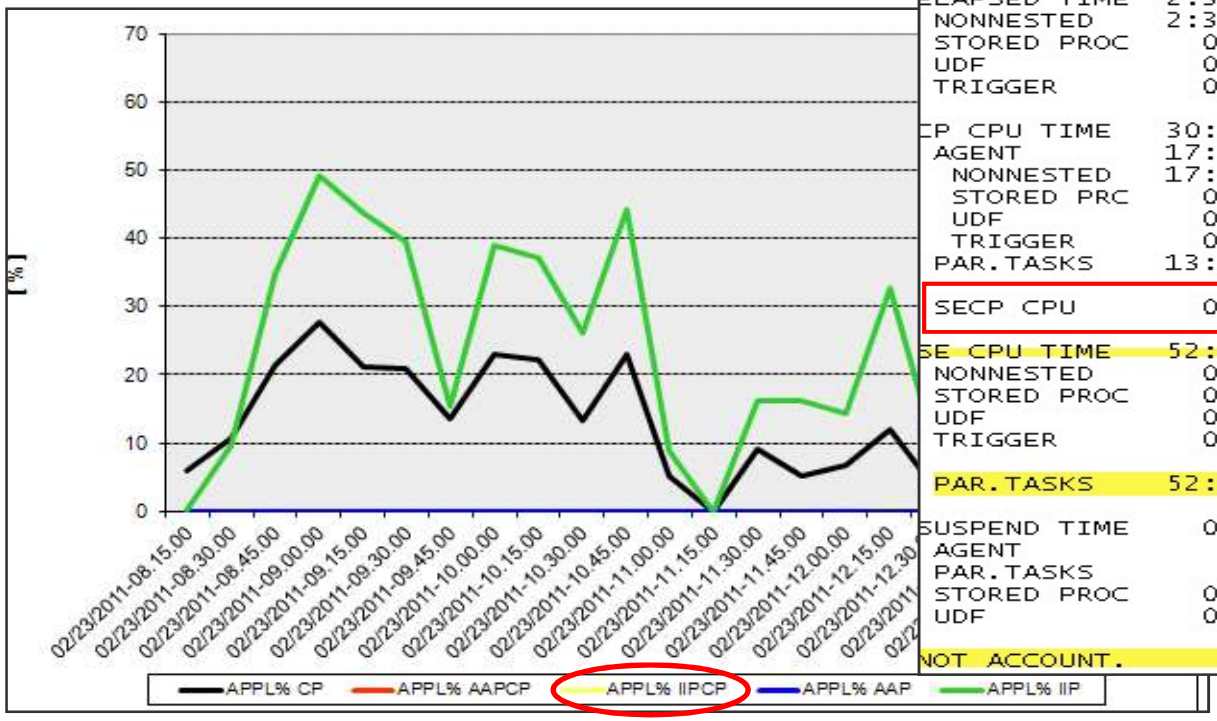
- V8
  - Access path based on serial cost
  - Parallelism cut on first table
  - limited 1x processors
  - 80% of child tasks zIIP eligible
- V9
  - Optimizer costs parallel tasks
  - Parallelism can be cut on inner table
  - Limited by 4x processors
- V10
  - Limited by 2x processors
  - Straw model parallelism
- V11
  - Sysplex Query Parallelism is **removed**
  - DPSI parallelism added
  - System negotiation based on storage
- V12
  - **100%** of parallel child threads eligible
  - I/O parallelism **REMOVED**

If query uses this...	I/O parallelism	CP parallelism
Parallel access through RID list (list prefetch and multiple index access)	Yes	Yes
Materialized views or materialized table expressions at reference time	No	Yes
Security label column on table	Yes	Yes
Parallel access through IN-list	Yes	Yes

IBM

Parallelism in production – COBOL Batch

- 80% of parallel child tasks are zIIP eligible (pre-V12) so it is the best way to affect zIIP Utilization %
  - Here we see there are no zIIP cycles that went to a GCP
  - But customer is complaining of a 3x increase in elapsed time for this batch job
  - However NOT ACCOUNT. For time is a significant portion of the elapsed time
  - 4CPs and 1 zIIP installed



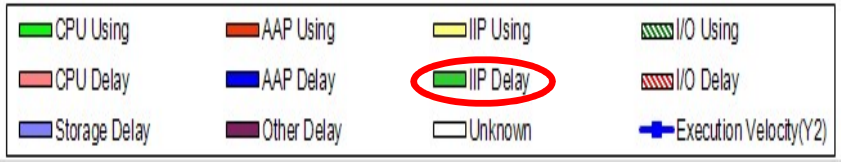
TIMES/EVENTS	APPL(CL.1)	DB2 (CL.2)
ELAPSED TIME	2:37:53.45	2:37:53.19
NONNESTED	2:37:53.45	2:37:53.19
STORED PROC	0.000000	0.000000
UDF	0.000000	0.000000
TRIGGER	0.000000	0.000000
CP CPU TIME	30:44.3617	30:44.3556
AGENT	17:38.9171	17:38.9111
NONNESTED	17:38.9171	17:38.9111
STORED PRC	0.000000	0.000000
UDF	0.000000	0.000000
TRIGGER	0.000000	0.000000
PAR.TASKS	13:05.4446	13:05.4446
SECP CPU	0.000000	N/A
SE CPU TIME	52:07.3400	52:07.3400
NONNESTED	0.000000	0.000000
STORED PROC	0.000000	0.000000
UDF	0.000000	0.000000
TRIGGER	0.000000	0.000000
PAR.TASKS	52:07.3400	52:07.3400
SUSPEND TIME	0.000000	47:44.0115
AGENT	N/A	29:21.9858
PAR.TASKS	N/A	18:22.0257
STORED PROC	0.000000	N/A
UDF	0.000000	N/A
NOT ACCOUNT.	N/A	58:44.9516

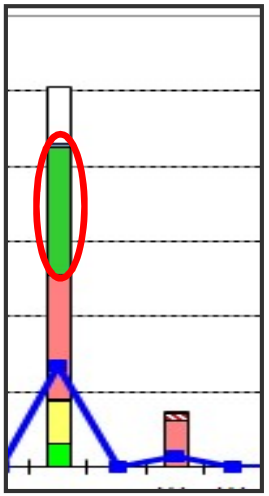
Parallelism Investigation

- RMF Spreadsheet Reporter Response delay report showed delay for zIIPs
  - Needs Help algorithm should redirect zIIP work to GCPs
- Lots of unaccounted for time
  - OMPE accounting
  - Child task class 2 time not reported (normal)
- SYS1.PARMLIB (IEAOPTxx)
  - IIPHONORPRIORITY = **NO**
    - 3 parallel tasks waiting for 1 zIIP (max degree=4)
  - In V11 this will stop system agents from being zIIP enabled

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

QUERY PARALLEL.	TOTAL
MAXIMUM MEMBERS	N/P
MAXIMUM DEGREE	4
GROUPS EXECUTED	78
RAN AS PLANNED	78
RAN REDUCED	0
ONE DB2 COOR=N	0
ONE DB2 ISOLAT	0
ONE DB2 DCL TTABLE	0
SEQ - CURSOR	0
SEQ - NO ESA	0
SEQ - NO BUF	0
SEQ - ENCL.SER	0





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

What you control for parallelism..

- Hidden zParm SPRMPTH – DSN6SPRC
  - Threshold below which parallelism disabled
- PARAMDEG – MAX_DEGREE limits parallel groups
  - Static and dynamic SQL (default ‘0’, unlimited)
- CDSSRDEF – SET CURRENT DEGREE special register for dynamic queries
  - Default =1, ‘ANY’ lets DB2 decide
- DEGREE(ANY) and CURRENTDATA(NO) bind options
  - Or DB2 needs to know if cursor is read-only
- VPPSEQT - % of sequential steal (VPSEQT) for parallel operations
  - Each utility task needs 128 pages in BP
- Star join enabled, number of tables involved
- PARA_EFF - % of cost reduction regarding parallel access path improvement (PM16020)

AccessPath	sequential_cost	parallel_degree	parallel_reduced_cost
AP1	1000	5	400
AP2	2000	20	300

PARMLIB Parameters

- IIPHONORPRIORITY (YES/NO) in IEAOPTxx parmlib member
  - This means if we reach the queue limit and ZIIPAWMT is triggered the dispatcher will route work over to a GP
  - If set to NO in DB2 11 then no system agents will be zIIP eligible
- ZIIPAWMT, ZAAPAWMT – Alternate wait management threshold is how long zIIP will run before checking to see if it needs help from GP
  - Default 12 milliseconds/ 3.2 for Hiperdispatch
  - In Db2 that means system engines may wait 3.2ms
- ZAAPZIIP = YES|NO (IEASYSxx option)
  - Allows zAAP eligible workload to run on a zIIP
- zAAP has other settings not applicable to zIIP
  - IFACrossover – disallow zAAP work on general CP

Ask Level 2 before adjusting!

** Be careful about attempting to FORCE zIIP offload

SMT (z13) Simultaneous Multi-Threading

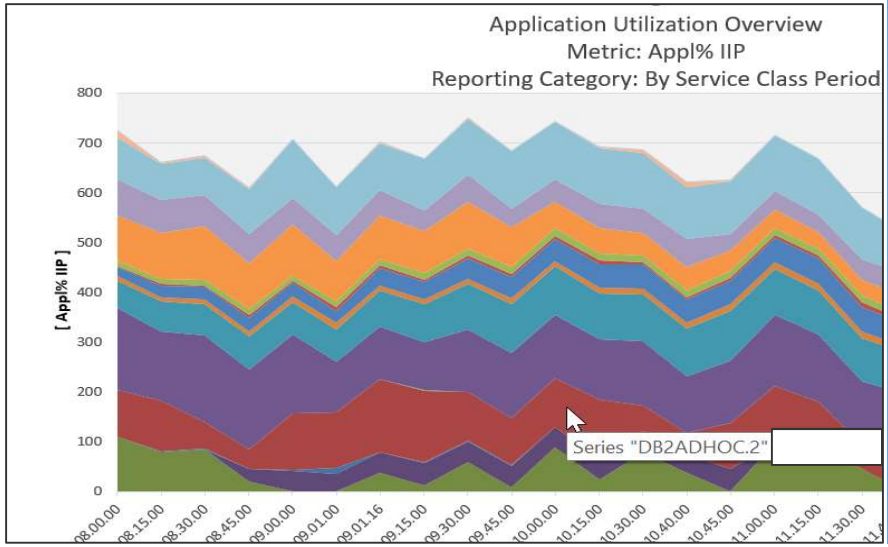
- SMT allows control program to run 1 or 2 threads concurrently on 1 CP
  - Can run parallel threads on 1 zIIP and IFL (not on CPs)
    - Z13 has 8 cores per GCP @ 5 GHz
  - If running parallel each task runs slower, but overall utilization is less
    - **IBM Brokerage OLTP workload showed 20% throughput improvement**
  - V12 has 1,800 system agent engines, hence throughput is key
- New IEAOPTxx parameter to control zIIP SMT mode
- MT_ZIIP_MODE=2 for 2 active threads (the default is 1)
  - Define a LOADxx PROCessor VIEW (PROCVIEW) ***CORE***|CPU for the life of the **IPL**
  - Without an IPL you can change the zIIP processor class MT Mode (the number of active threads per online zIIP) using IEAOPTxx SET OPT=xx
- Requires HyperDispatch=YES
  - Ensure OA51419 is applied to avoid stalls during global recovery

z13 zIIP capacity:
- *is 38% greater than a zEC12 zIIP*
- *is 72% greater than a z196 zIIP*

z14
- *IBM SO saw 20% redux in zIIP busy time*

zIIP work and WLM

- Any workload which lands in a Discretionary service class AND is zIIP eligible will not be redirected to a GCP
  - Even with IIPHONORPRIORITY=YES Discretionary DDF work will queue for a zIIP and not fall back to a GCP
  - Blocked Workload Support not available for zIIP eligible work
    - Block Workload Support was designed to promote discretionary work which was starved of resources while holding locks/latches needed by higher importance work
    - By default after 20 seconds the stalled work would be eligible for promotion to get it up and out of the way
      - BLWLINTHD=20 (seconds) and BLWLTRPCT (%)=0.5 (% of an engine)
- Recommend: avoid any Db2 dependent workload, especially those that are zIIP eligible from being classified as discretionary (monitor in RMF)



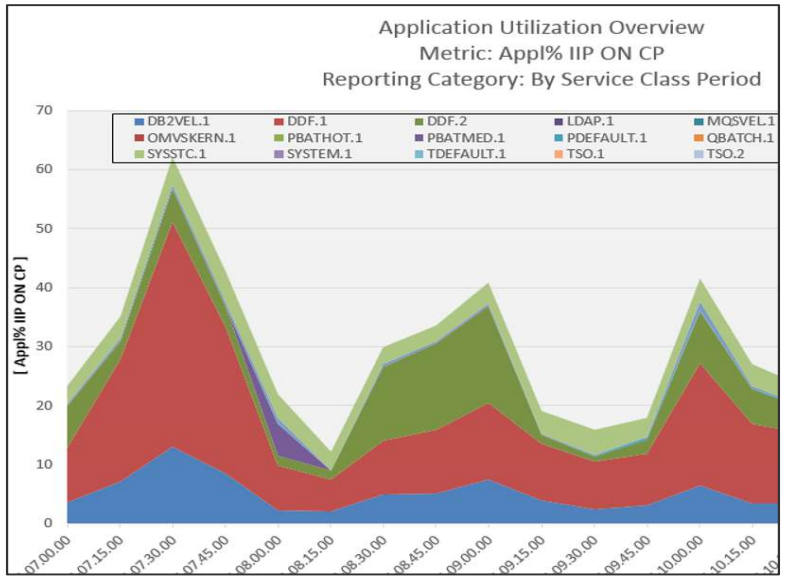
Playing Hardball

- Can now limit zIIP usage as well as CP usage in the Resource Group Definition (originally for Container Based Pricing)
  - OA52132 (z/OS 2.2)
    - This along with the use of Tenant Report Class/Resource Group
  - For instance… S-TAP and Guardium (PI90376)
    - Customer examples of Guardium utilizing multiple zIIPs for several hours
- Can eliminate GCP redirect at the WLM service class level
  - OA50845/OA50953 – adds Honor Priority to WLM service class definition
  - DO NOT attempt to use this against Db2 address spaces
    - System agent hang can impact entire data sharing group
  - Going forward need to monitor WLM % Delays for zIIP engine since there will be no redirect to report
- PROCXCOST on the VIPADISTRIBUTE statement influences DRDA work towards LPARs with more zIIP capacity, and less redirect to the GCPs
  - CAUTION: DRDA work is roughly 60% zIIP eligible, so there must be GCP capacity as well

IBM

Summary

- Monitor zIIP overflow/redirect for capacity planning… not absolute utilization
- If zIIP peak and 4-hour rolling average collide… every MSU counts
- Use RMF Spreadsheet reporter to determine BY SERVICE CLASS which workloads are being hindered
  - ApplOvwTrd tab now included in spreadsheet
- Fewer faster zIIPs on an upgrade is NOT a good idea – aim for the most Vertical High zIIPs assigned to an LPAR
- Review LPAR weightings to determine if zIIPs are parked during times of zIIP redirect
- Test SMT and monitor the zIIP redirect



Db2 SWAT team engagements

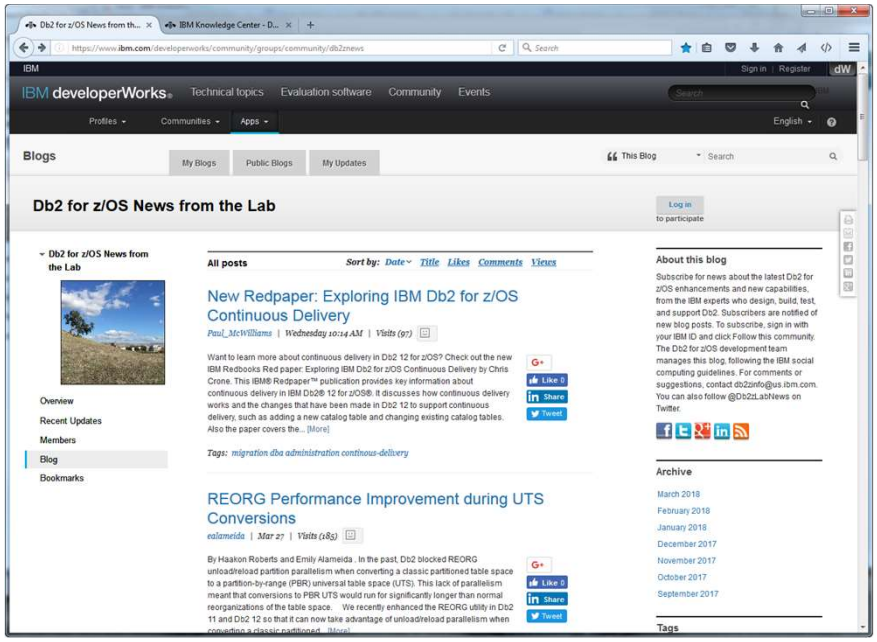
- **Db2 Master Class**– held twice a year, one in the US and one in the UK
  - https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Wc05a3bbc003d_44bf_8673_d5dd7683d239/page/Db2%20for%20zOS%20Master%20Class%202019%20-%20Workshop%20Announcement
  - Hursley Lab – the week of 06/24
  - Silicon Valley Lab – San Jose the week of 09/23
    - Spend a week with John Campbell and the Db2 SWAT team covering performance and availability topics including how to analyze statistics and accounting data

It's all about robustness.

- **Db2 360 Degree Continuous Availability Assessment Study**
  - Comprehensive study performed by the Db2 SWAT team aimed at discovering exposures in continuous availability, performance, and speed of recovery
- *Please contact me or Chunyang Xia (cxia@us.ibm.com)*

Db2 for z/OS News from the Lab blog

http://ibm.biz/db2znews



Get the latest news from the IBMers who design and build Db2!

- New capabilities in Db2 12 for z/OS continuous delivery
- Enhancements in Db2 11 for z/OS
- Helpful tips and best practices from Db2 for z/OS development

– Join the conversation
  - Subscribe to follow the blog
  - Become a member to comment
  - Follow us on Twitter: @Db2zLabNews

Reference material

- II14219 - zIIP Exploitation
- 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
- PI73882 – zIIP enablement of RELOAD for LOAD and REORG
- RMF Spreadsheet Reporting Tool
  - http://www-03.ibm.com/systems/z/os/zos/features/rmf/tools/rmftools.html
- Getting Started Resources
  - http://www-03.ibm.com/systems/z/hardware/features/ziip/resources.html
- Link to article on PARMLIB settings
  - https://www.ibm.com/developerworks/mydeveloperworks/blogs/22586cb0-8817-4d2c-ae74-0ddcc2a409bc/entry/december_17_2012_6_07_am3?lang=en
- World of DB2
  - www.worldofdb2.com

Other Processes

- IPSEC PROFILE.TCPIP settings
  - IPCONFIG IPSECURITY
  - GLOBALCONFIG → ZIIP → IPSECURITY
    - Displaces related CPU cycles to the zIIP (default is NOIPSECURITY)
  - Netstat STATS/-S command will show 'Packets Handled by zIIP'

- Hipersocket multi-write operations, must be 32k in size
  - Usually related to XML, file transfers, SOAP
  - IQDIOMULTIWRITE
    - GLOBALCONFIG → ZIIP → IQDIOMULTIWRITE
      - Default is NOIQDIOMULTIWRITE

- XRC (Global Mirror DFSMS SDM [system data mover]) startup member of PARMLIB
  - ANTAS000, ANTAS0nn, and ANTCL0nn address spaces enabled for zIIP processing
  - ANTXIN00 parmlib member
    - ZiipEnable(FULL) – gives you the max, otherwise YES allows everything but I/O operations)
      - Address space and SRB components offload

- OMPE
  - CICS SLA report builder, up to 73%
  - DB2 Near Term History, all of processing for normalizing the raw SMF
  - DASD UCB sampling 2-10% CPU savings

IBM