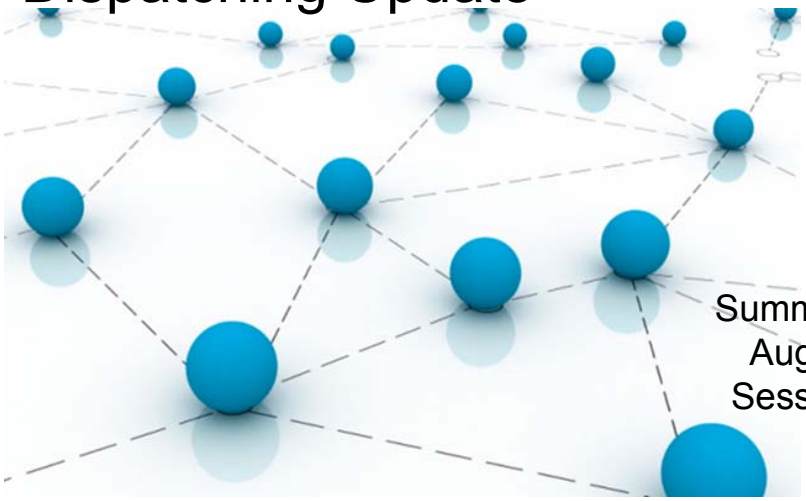
Glenn Anderson, IBM Lab Services and Training

IBM

Connect the Dots: A z13 and z/OS Dispatching Update

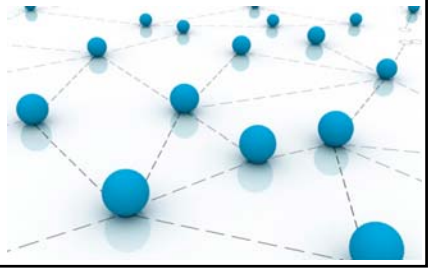


Summer SHARE
August 2015
Session 17705

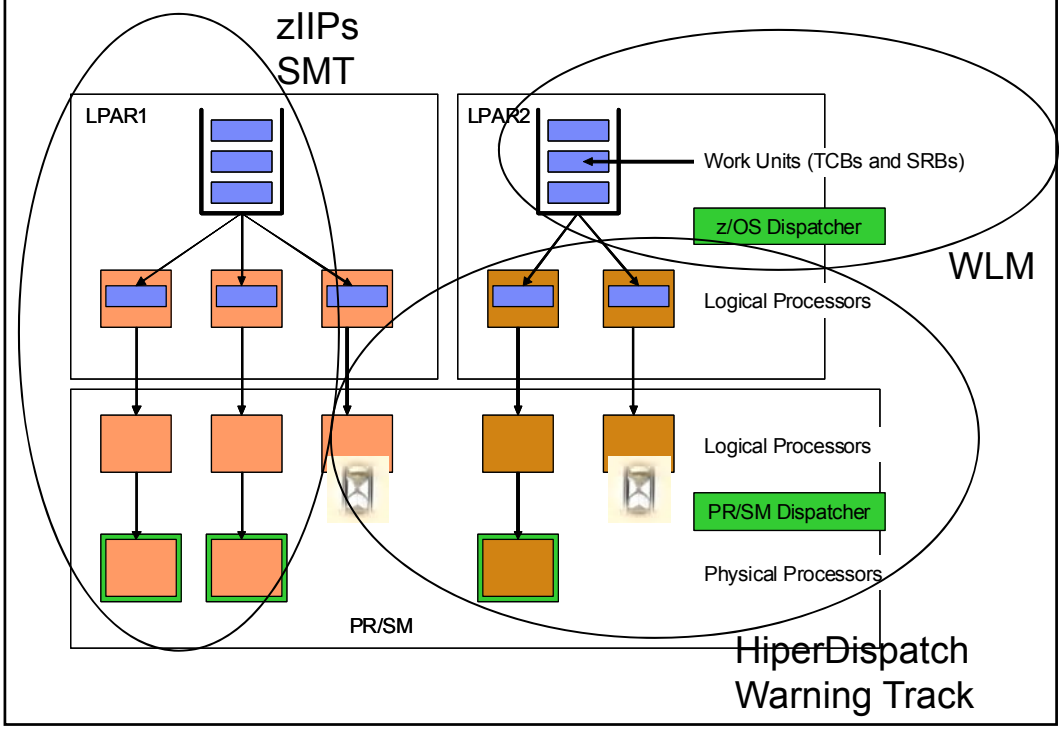
What I hope to cover.....

IBM

- What are dispatchable units of work on z/OS
- How WLM manages dispatchable units of work
- The role of HiperDispatch and Warning Track
- Dispatching work to zIIP engines
- z13 Simultaneous Multithreading (SMT)



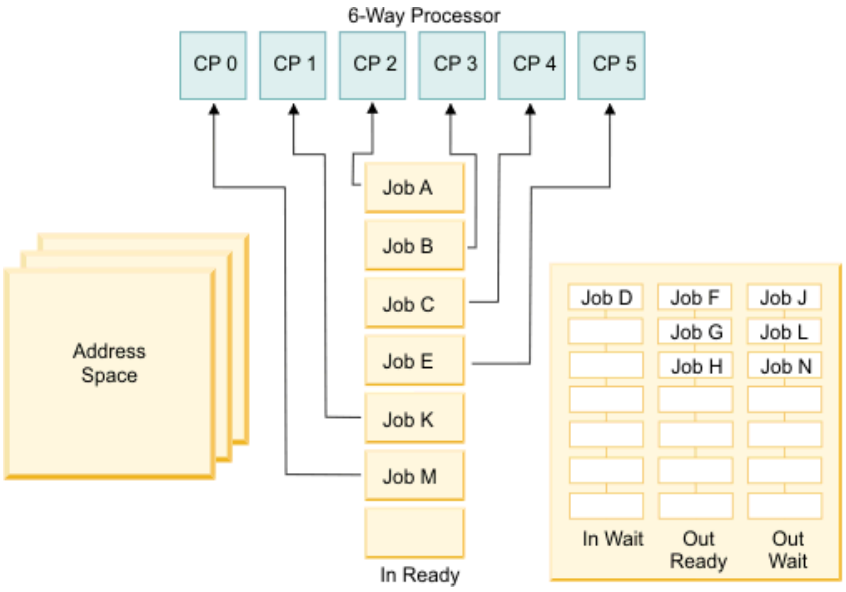
z13 and z/OS dispatching



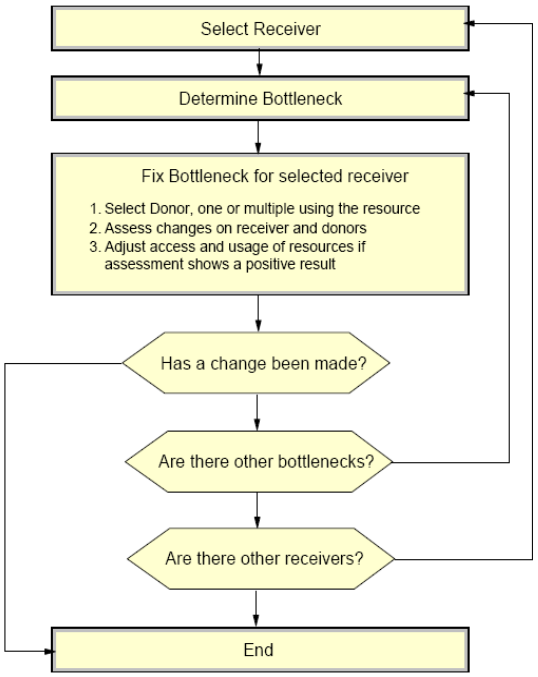
z/OS dispatchable units

- **There are different types of Dispatchable Units (DU's) in z/OS**
  - Preemptible Task (TCB)
  - Non Preemptible Service Request (SRB)
  - Preemptible Enclave Service Request (enclave SRB)
    - Independent - a new transaction
    - Dependent – extend existing address space
    - Work-dependent – extend existing independent enclave

z/OS dispatching work



WLM policy adjustment algorithm



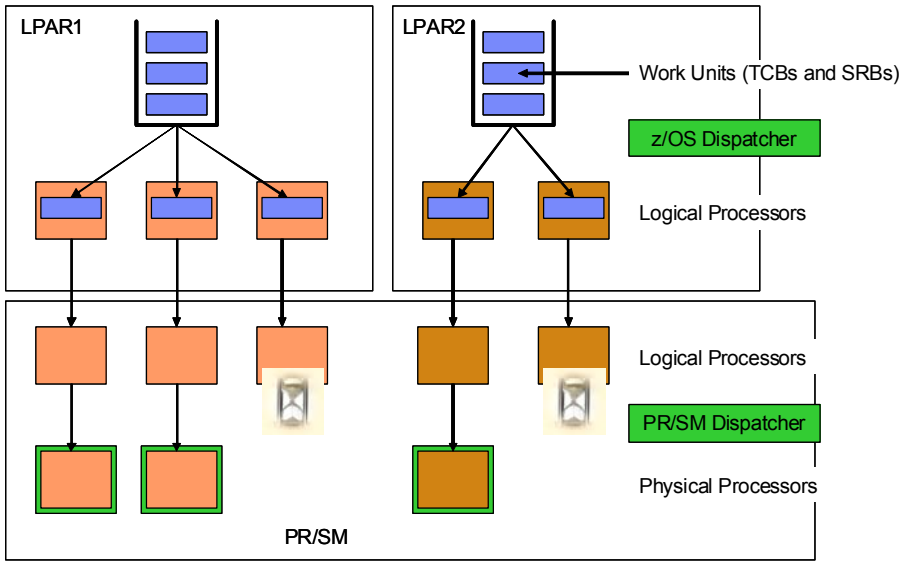
WLM dispatching priority usage

IBM

255	SYSTEM
254	SYSSTC
253	*Small Consumer*
252 208	Priorities for dynamic policy adjustment
207 202	Not used
201 192	Discretionary work Mean Time to wit algorithm

Dispatching in an LPAR environment

IBM

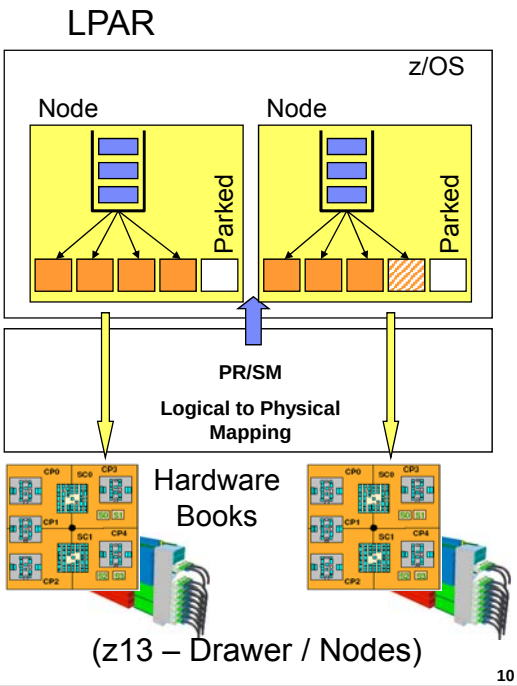


HiperDispatch mode

- PR/SM
  - Supplies topology information/updates to z/OS
  - Ties *high priority* logicals to physicals (gives 100% share)
  - Distributes remaining share to *medium priority* logicals
  - Distributes any additional service to unparked *low priority* logicals
- z/OS
  - Ties tasks to small subsets of logical processors
  - Dispatches work to *high priority* subset of logicals
  - Parks *low priority* processors that are not need or will not get service
- **Hardware cache optimization occurs when a given unit of work is consistently dispatched on the same physical CPU**

HiperDispatch: z/OS part

- z/OS obtains the logical to physical processor mapping in Hiperdispatch mode
  - Whether a logical processor has high, medium or low share
  - On which book and chip the logical processor is located
- z/OS creates dispatch nodes
  - The idea is to have 4 high share CPs in one node
  - Each node has TCBs and SRBs assigned to the node
  - Optimizes the execution of work units on z/OS



10

RMF CPU activity report

```
1                                          C P U  A C T I V I T Y

            z/OS V1R11              SYSTEM ID Z2              START 09/11/2009-02.30.00  INTERVAL 000.30.00
                                    RPT VERSION V1R11 RMF     END   09/11/2009-03.00.00  CYCLE 0.100 SECONDS
-CPU  2097   MODEL  737   H/W MODEL  E56  SEQUENCE CODE 00000000000699FF    HIPERDISPATCH=YES
0---CPU---    ---------------- TIME % ----------------     LOG PROC      --I/O INTERRUPTS--
 NUM  TYPE    ONLINE    LPAR BUSY    MVS BUSY   PARKED     SHARE %       RATE     % VIA TPI
  0    CP     100.00      96.60        96.74      0.00      100.0  HIGH    1593     2.64
  1    CP     100.00      97.51        97.69      0.00      100.0  HIGH    1607     2.73
  2    CP     100.00      96.02        96.23      0.00       96.0  MED     5.12    29.30
  3    CP     100.00      39.26        80.81     51.23        0.0  LOW     0.00     0.00
  4    CP     100.00      48.71        79.90     38.77        0.0  LOW     0.00     0.00
  5    CP     100.00      41.06        79.34     48.01        0.0  LOW     0.00     0.00
  6    CP     100.00      12.42        78.35     84.11        0.0  LOW     0.00     0.00
  7    CP     100.00       0.00        -----    100.00        0.0  LOW     0.00     0.00
  8    CP     100.00       0.00        -----    100.00        0.0  LOW     0.00     0.00
  9    CP     100.00      33.05        80.34     58.68        0.0  LOW    199.6     1.01
 TOTAL/AVERAGE            46.46        89.73               296.0          3405     2.62
0 A    AAP    100.00      57.35        88.68      0.00       32.0  MED
  B    AAP    100.00      46.71        92.85     17.56        0.0  LOW
  C    AAP    100.00      45.27        90.82     17.79        0.0  LOW
  D    AAP    100.00      53.81        85.00      0.00        0.0  LOW
 TOTAL/AVERAGE            50.78        89.09                32.0
0 E    IIP    100.00       0.26         0.26      0.00       16.2  MED
  F    IIP    100.00       0.01         0.01      0.00        0.0  LOW
 TOTAL/AVERAGE             0.13         0.13                16.2
```

HiperDispatch and LPAR

```
1                                        P A R T I T I O N  D A T A  R E P O R T                                  PAGE    3

            z/OS V1R10              SYSTEM ID LPAR1           DATE 04/29/2011          INTERVAL 14.59.998
                                CONVERTED TO z/OS V1R12 RMF   TIME 19.28.00            CYCLE 1.000 SECONDS
-
 MVS PARTITION NAME                LPAR1          NUMBER OF PHYSICAL PROCESSORS         55              GROUP NAME       N/A
 IMAGE CAPACITY                     3165                        CP                       53              LIMIT            N/A
 NUMBER OF CONFIGURED PARTITIONS       4                        IIP                       2              AVAILABLE        N/A
 WAIT COMPLETION                      NO
 DISPATCH INTERVAL                DYNAMIC
-
--------- PARTITION DATA ----------------  -- LOGICAL PARTITION PROCESSOR DATA --   -- AVERAGE PROCESSOR UTILIZATION PERCENTAGES --
                  ----MSU----  -CAPPING--  PROCESSOR-  ----DISPATCH TIME DATA----   LOGICAL PROCESSORS  --- PHYSICAL PROCESSORS ---
 NAME       S   WGT  DEF    ACT  DEF   WLM%  NUM   TYPE   EFFECTIVE       TOTAL      EFFECTIVE    TOTAL  LPAR MGMT  EFFECTIVE  TOTAL
 LPAR1      A   494    0    582  NO     0.0 32.0   CP    02.17.24.319  02.20.44.154       28.63    29.32       0.44      17.96  18.40
 LPAR2      A   446    0    762  NO     0.0 32.0   CP    03.01.28.607  03.04.05.167       37.81    38.35       0.34      23.72  24.06
 LPAR3      A    59    0      0  NO     0.0  3.0   CP    00.00.00.000  00.00.00.000        0.00     0.00       0.00       0.00   0.00
 LPAR5      A     1    0      0  NO     0.0  1.0   CP    00.00.00.000  00.00.00.000        0.00     0.00       0.00       0.00   0.00
 *PHYSICAL*                                             00.10.58.833                                            1.44                1.44
                                                        ------------  ------------                              -----      -----  -----
   TOTAL                                                05.18.52.927  05.35.48.155                               2.21      41.68  43.90
```

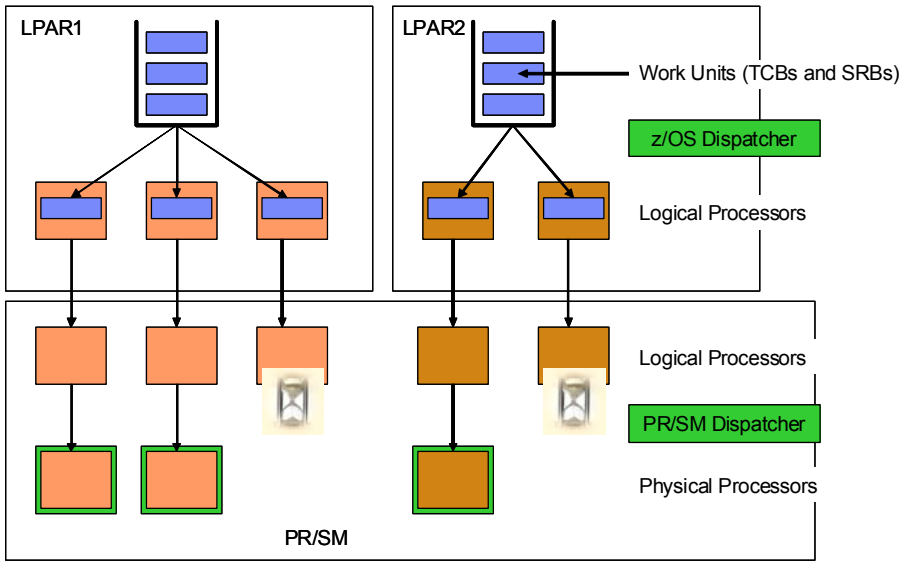
Total LPAR weight = 1000
LPAR1 494/1000 = .494 * 53 CPs = 26.18 CPs
LPAR2 446/1000 = .446 * 53 CPs = 23.64 CPs

LPAR1 = 25 VH and 2 VM at 59% share (27 logicals unparked)
LPAR2 = 23 VH and 1 VM at 64% share (24 logicals unparked)

51 logicals unparked

53 physicals

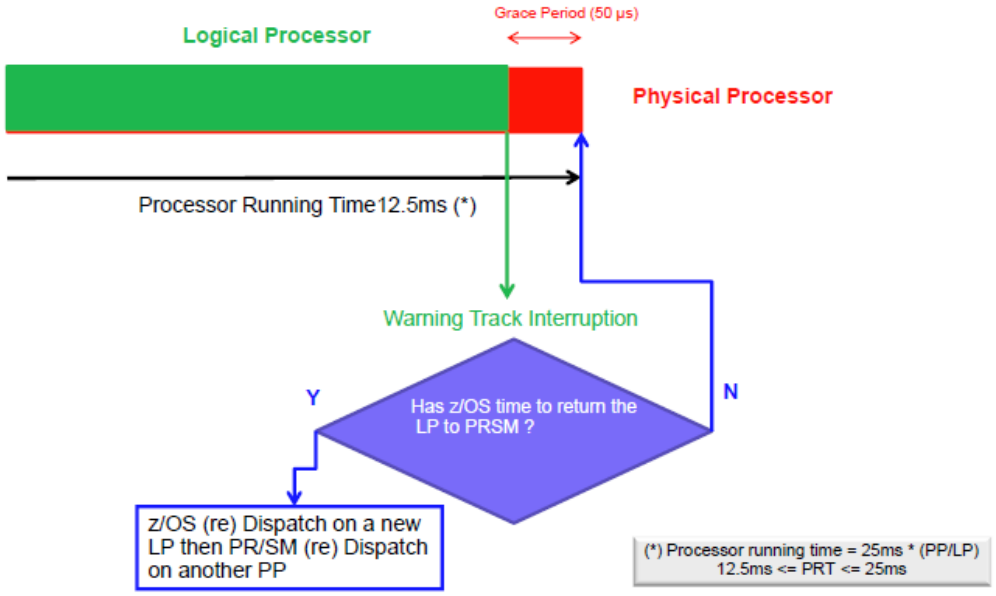
Dispatching in an LPAR environment



Warning track

- In a PR/SM™ environment the LPAR hypervisor assigns physical engines to logical engines accordingly to the weighting factors of the partitions.
- Once the time slice for a logical engine is expired the currently executing work is suspended until a physical engine is assigned to the logical engine again.
- The Warning Track Interruption Facility notifies the operating system that PR/SM™ will undispatch a certain logical processor within the next 50 microseconds (grace period).
- z/OS is now able to save status for the running unit of work and re-dispatch the work unit on a different logical processor within the grace period.
- z/OS now signals to PR/SM via Diagnose x'9C' that the logical processor can be un-dispatched.
- Warning Track processing is only supported in HyperDispatch=YES environments.
- A high benefit can be achieved for Low Share processors which might be parked by WLM.

Warning track



Latent demand: LPAR Busy vs MVS Busy

CPU 2097 CPC CAPACITY 1451
MODEL 719 CHANGE REASON=N/A HIPERDISPATCH=YES

---CPU--- NUM	TYPE	ONLINE	LPAR BUSY	MVS BUSY	PARKED	LOG PROC SHARE %	
0	CP	100.00	96.77	96.80	0.00	100.0	HIGH
1	CP	100.00	94.91	94.95	0.00	100.0	HIGH
2	CP	100.00	96.72	96.74	0.00	100.0	HIGH
3	CP	100.00	95.07	95.10	0.00	100.0	HIGH
4	CP	100.00	**50.18**	**93.55**	0.00	66.0	MED
5	CP	100.00	50.15	93.56	0.00	66.0	MED
6	CP	100.00	**20.30**	**89.09**	**56.00**	0.0	LOW
7	CP	100.00	11.40	90.19	72.00	0.0	LOW
8	CP	100.00	22.12	88.49	50.79	0.0	LOW
9	CP	100.00	**46.12**	**87.87**	**0.00**	0.0	LOW
A	CP	100.00	45.37	86.74	0.00	0.0	LOW
B	CP	100.00	38.46	86.76	11.21	0.0	LOW
C	CP	100.00	35.08	86.96	19.43	0.0	LOW
D	CP	100.00	19.29	84.13	57.66	0.0	LOW
E	CP	100.00	0.00	-----	100.00	0.0	LOW
F	CP	100.00	0.00	-----	100.00	0.0	LOW
10	CP	100.00	0.00	-----	100.00	0.0	LOW
TOTAL/AVERAGE			**42.47**	**91.45**		**532.0**	

CEC Busy = 98.85

.0115 * 19 CP = .22 CPs available

Weight: 5.32 CPs

Using: 42.47/100 * 17 LCP = 7.22 CPs

Warning track statistics

IBM

- RMF keeps track of the number of times PR/SM issued a warning-track interruption to a logical processor and z/OS was able/unable to return the logical processor within the grace period.
- RMF measures the amount of time in microseconds that a processor was yielded to PR/SM due to Warning-track processing.

SMF record type 70 subtype 1 (CPU Activity) – CPU data section				
Offset	Name	Length	Format	Description
80 x50	SMF70WTS	4	Binary	The number of times PR/SM issued a warning-track interruption to a logical processor and z/OS was able to return the logical processor within the grace period.
84 x54	SMF70WTU	4	Binary	The number of times PR/SM issued a warning-track interruption to a logical processor and z/OS was unable to return the logical processor within the grace period.
88 x58	SMF70WTI	4	Binary	Amount of time in microseconds that a logical processor was yielded to PR/SM due to Warning Track processing.

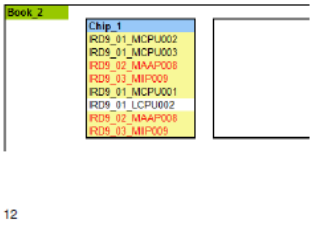
RMF Postprocessor Overview Conditions		
Name	Qualifier	Description
WTRKCP (WTRKAAP) (WTRKIIP)	cpu-id	The percentage of times PR/SM issued a warning-track interruption to a processor and z/OS was able to return it to PR/SM within the grace period.
WTRKTCP (WTRKTAAP) (WTRKTIIP)	cpu-id	Time in microseconds that a purpose processor was yielded to PR/SM due to Warning Track processing.

WLM Topology Report Tool

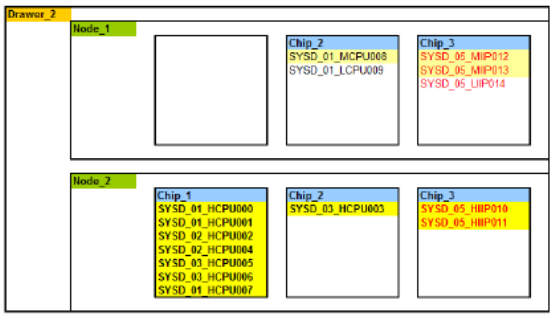
IBM

- New ***as-is*** tool available for download from the WLM homepage
  - http://www.ibm.com/systems/z/os/zos/features/wlm/WLM_Further_Info_Tools.html#Topology
- Visualizes mapping of HiperDispatch affinity nodes to physical structure
- Supports IBM zEC10 and later
- To use:
  1. Download from above location
  2. Run installer
  3. Collect SMF99.14 records
  4. Upload Host code to a z/OS system

Sample output (zEC12): Topology for 07-21-2014-13:44:27 ,Syst

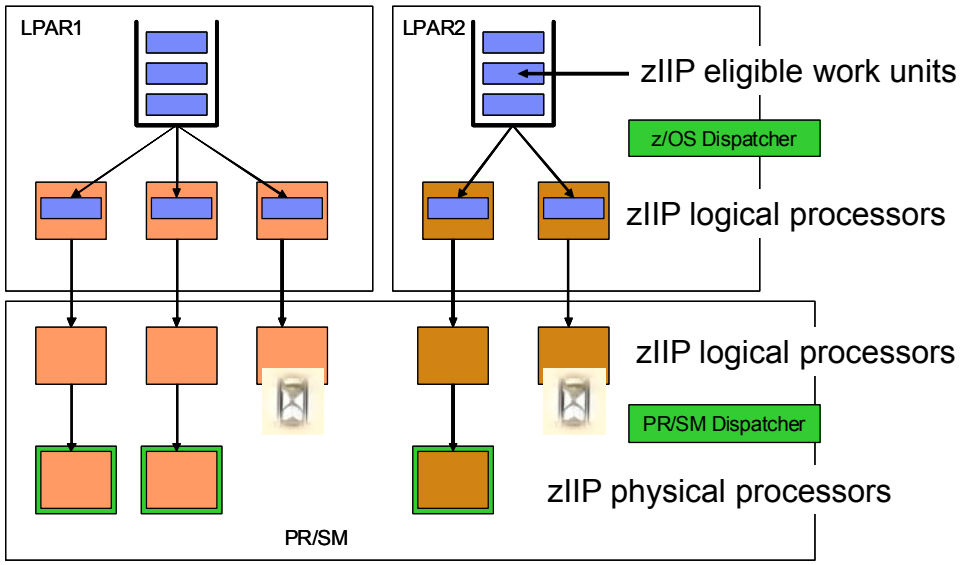


Sample output (z13): Topology for 01-30-2015-14:08:32 ,System: SYSD



12

Dispatching zIIP eligible work



IBM z Integrated Information Processor (zIIP) on the z13

- **The IBM z13 continues to support the z Integrated Information Processor (zIIP) which can take advantage of the optional simultaneous multithreading (SMT) technology capability. SMT allows up to two active instruction streams per core, each dynamically sharing the core's execution resources.**
  - With the multithreading function enabled, the performance capacity of the zIIP processor is expected to be up to 1.4 times the capacity of these processors on the zEC12
- **The rule for the CP to zIIP purchase ratio is that for every CP purchased, up to two zIIPs may be purchased**
- **zAAP eligible workloads such as Java and XML, can run on zIIPs using zAAP on zIIP processing**
- **zAAPs are no longer supported on the z13**

Current IBM exploitation of zAAPs and zIIPs

Specialty CP	Eligible	Major Users
zAAP or zIIP on z13	Any Java Execution	Websphere CICS Native apps XMLSS
zIIP	Enclave SRBs	DRDA over TCPIP DB2 Parallel Query DB2 Utilities Load, Reorg, Rebuild DB2 V9 z/OS remote native SQL procedures TCPIP - IPSEC XMLSS zIIP Assisted HiperSockets Multiple Write Virtual Tape Facility Mainframe (VTFM) Software z/OS Global Mirror (XRC), System Data Mover (SDM) z/OS CIM Server RMF Mon III OMEGAMON on z/OS and DB2 IMS Ver 8

What is "Needs Help?"

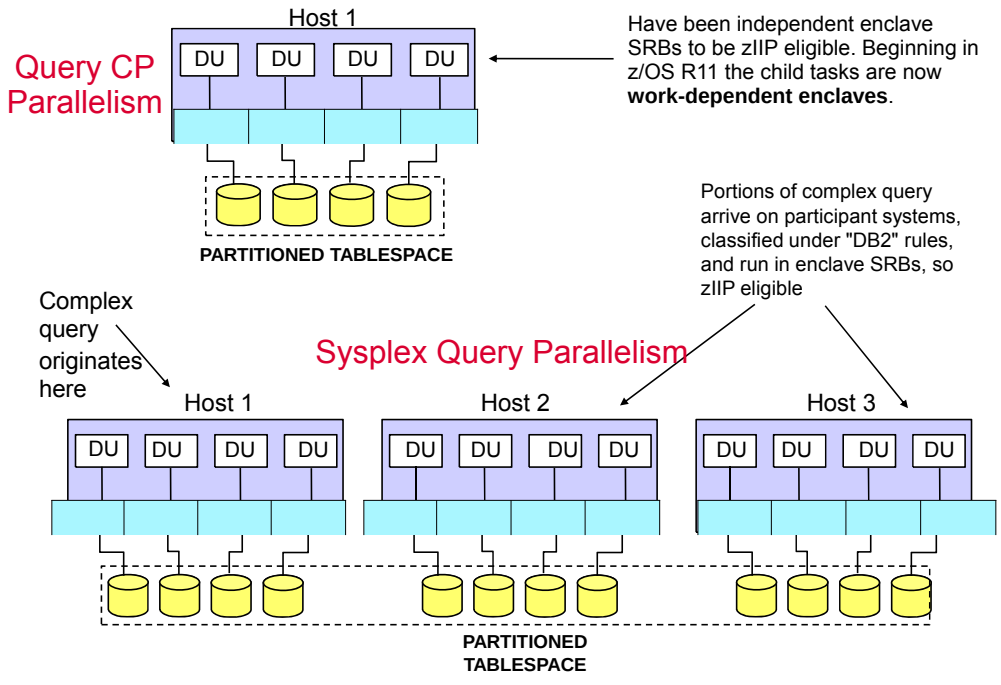
- Determination zIIP or zAAP work is being delayed and additional resources should help process the work
  - Requires xxPHONORPRIORITY=YES to be set
- If help is required:
  - The zxxP CP signals waiting zxxP to help
  - When all zxxP CPs are busy the zxxP asks for help from the GCP
    - All available speciality engines (of the same type) must be busy before help is asked of the GCPs
      - IF the zxxPs needs help and all zxxPs are busy help is obtained from 1 GCP
      - IF zxxPs continue to need help additional CPs may be asked to help
  - Help is always provided in dispatch priority order

Specialty CP work running in a WLM service class

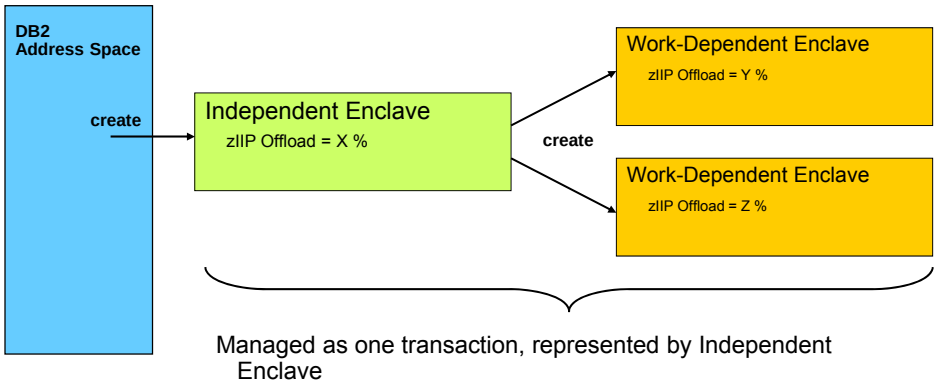
```
REPORT BY: POLICY=WLMPOL     WORKLOAD=BAT_WKL    SERVICE CLASS=BATSPEC    RESOURCE GROUP=BATMAXRG
TRANSACTIONS    TRANS-TIME HHH.MM.SS.TTT  --DASD I/O--  ---SERVICE----  SERVICE TIMES  ---APPL %---
AVG      0.98   ACTUAL            6.520   SSCHRT  11.5  IOC      8326   CPU      24.7  CP       0.97
MPL      0.98   EXECUTION         6.128   RESP     7.0  CPU    662386   SRB       0.0  AAPCP    0.01
ENDED      10   QUEUED              391   CONN     6.9  MSO         0   RCT       0.0  IIPCP    0.00
END/S    0.17   R/S AFFIN             0   DISC     0.0  SRB       965   IIT       0.0
#SWAPS      0   INELIGIBLE            0   Q+PEND   0.1  TOT    671677   HST       0.0  AAP     40.27
EXCTD       0   CONVERSION            0   IOSQ     0.0  /SEC    11195   AAP      24.2  IIP      0.00
AVG ENC  0.00   STD DEV               0                                 IIP       0.0

 GOAL: EXECUTION VELOCITY 35.0%     VELOCITY MIGRATION:   I/O MGMT  99.2%     INIT MGMT 92.2%

         RESPONSE TIME EX   PERF  AVG    ------ USING% ----- ------------ EXECUTION DELAYS % -------
-
SYSTEM                  VEL% INDX ADRSP   CPU  AAP  IIP  I/O  TOT  CPU

SYSD        --N/A--     99.2  0.4   1.0   0.8 45.9  0.0  3.9  0.4  0.4
```

RMF report is at 1 minute interval

DB2 parallel query, enclave SRBs and zIIPs



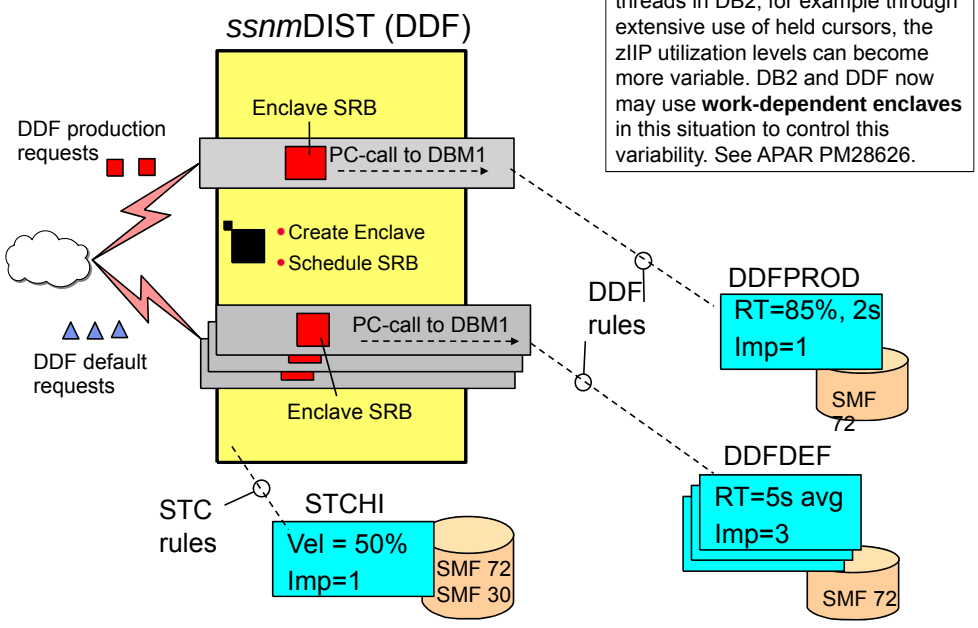
Work-Dependent enclaves



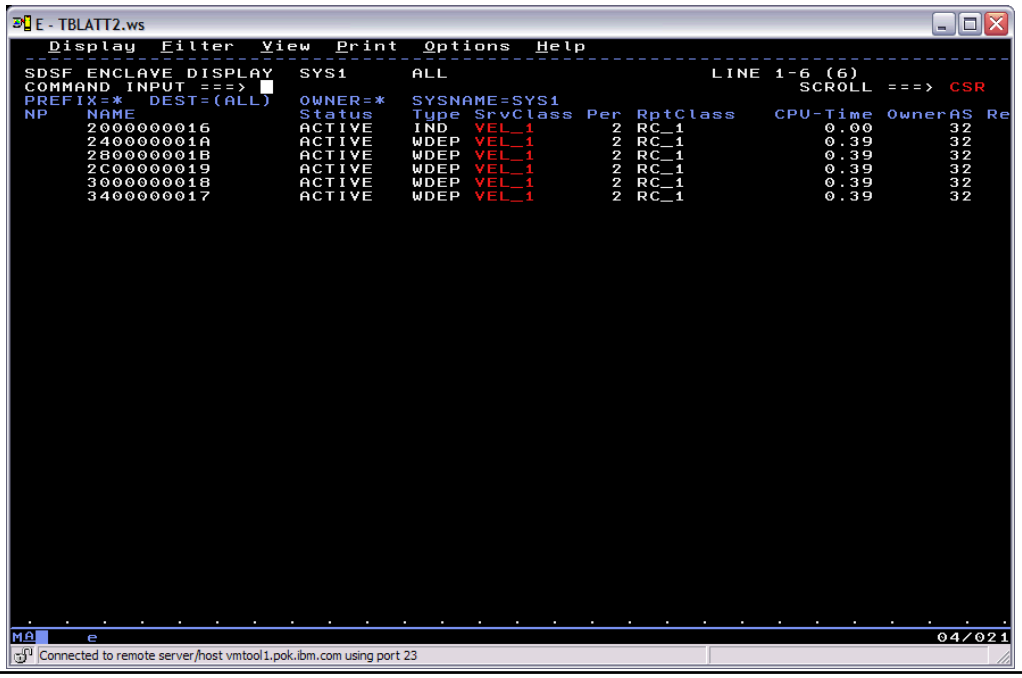
Implement a new type of enclave named “Work-Dependent” as an extension of an Independent Enclave. A Work-Dependent enclave becomes part of the Independent Enclave’s transaction but allows to have its own set of attributes (including zIIP offload percentage)

DDF and work-dependent enclaves

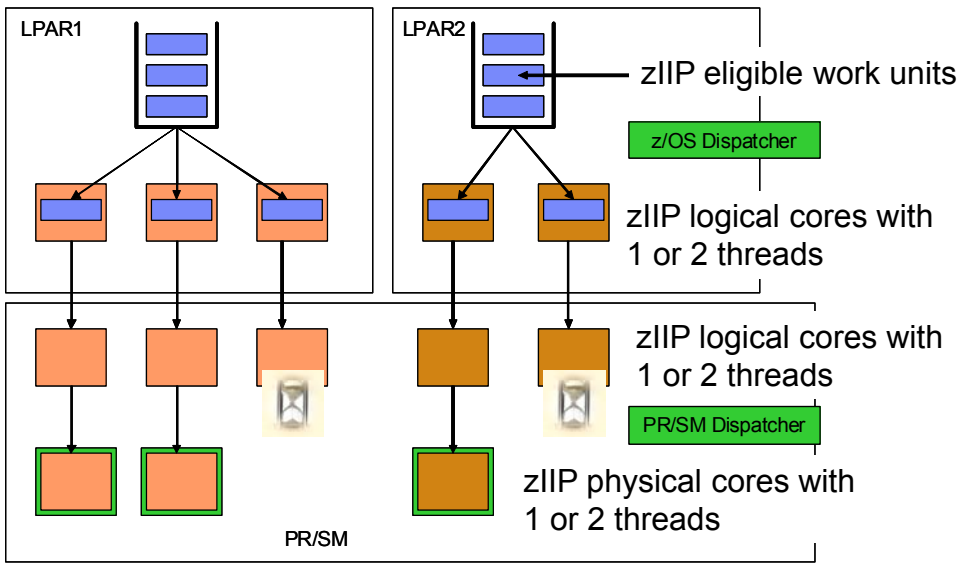
In cases where DRDA applications create extended duration work threads in DB2, for example through extensive use of held cursors, the zIIP utilization levels can become more variable. DB2 and DDF now may use **work-dependent enclaves** in this situation to control this variability. See APAR PM28626.



Work-dependent enclaves in SDSF



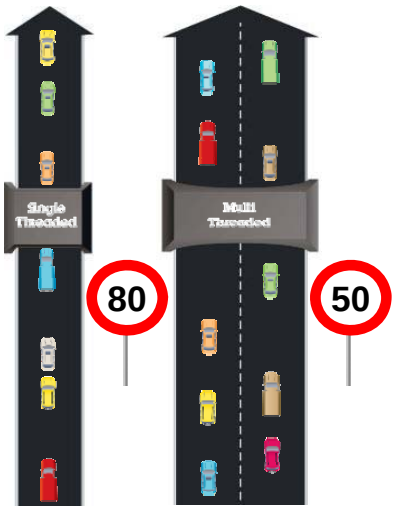
zIIP processors and simultaneous multithreading



z13 - Simultaneous Multithreading (SMT)

- "Simultaneous multithreading (SMT) permits multiple independent threads of execution to better utilize the resources provided by modern processor architectures."*
- With z13, SMT allows up to two instructions per core to run simultaneously to get better overall throughput
- SMT is designed to make better use of processors
- On z/OS, SMT is available for zIIP processing:
  - Two concurrent threads are available per core and can be turned on or off
  - Capacity (throughput) usually increases
  - Performance may in some cases be superior using single threading

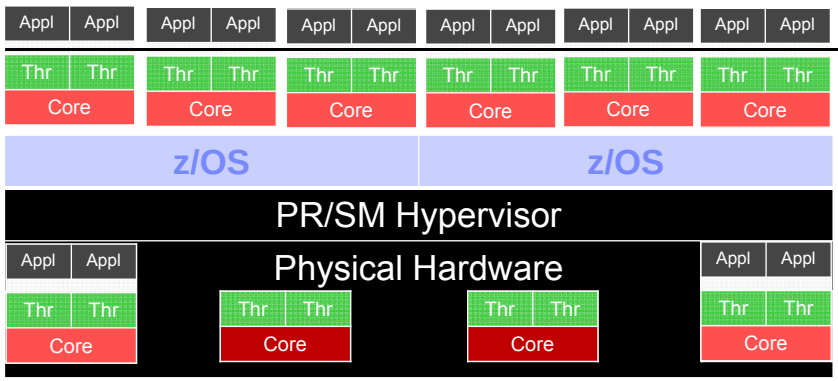
*Wikipedia®



Two lanes process more traffic overall

Note: Speed limit signs for illustration only

z13 - SMT Exploitation



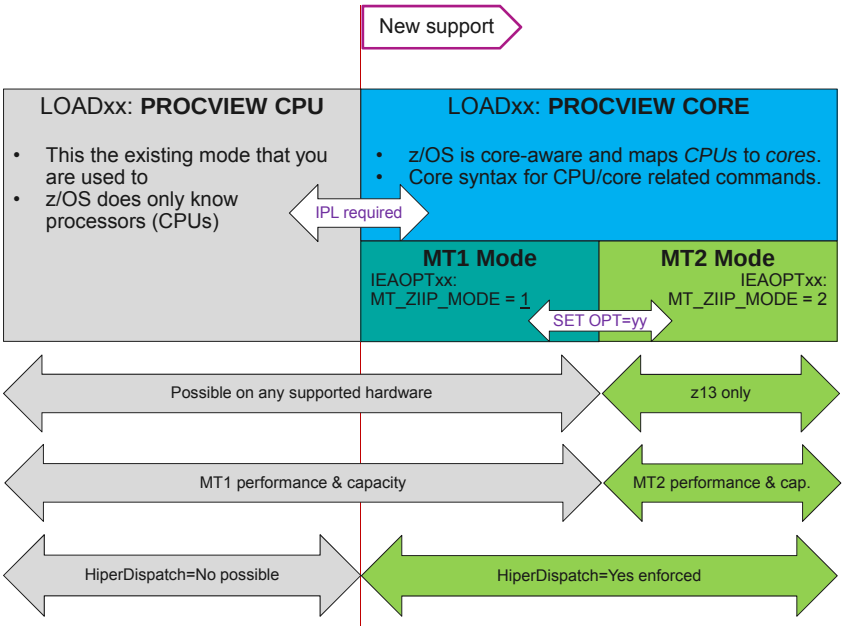
MT Ignorant

MT Aware

- Generally focuses on increasing core throughput predictably and repeatability
- PR/SM supports SMT for SMT aware OS like z/OS via core dispatching
- z/OS controls and manages whole core (all threads) to:
  - Maximize core throughput (fill running cores, spill to waiting cores)
  - Maximize core availability (meet goals using fewest cores)
- Limits SMT variability to a single z/OS workload
  - Makes capacity, accounting, latency, response time more predictable and repeatable

30

What is new with multithreading support?



31

* Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.

New terminology for SMT…

- z/OS logical processor (CPU) ➔ Thread
  - A thread implements (most of) the System z processor architecture
  - z/OS dispatches work units on threads
  - In MT mode two threads are mapped to a logical core
- Processor core ➔ Core
  - PR/SM dispatches logical core on a physical core
    - Thread density 1 (TD1) when only a single thread runs on a core
    - Thread density 2 (TD2) when both threads run on a core
- MT1 Equivalent Time (MT1ET)
  - z/OS CPU times are normalized to the time it would have taken to run same work in MT-1 mode
    - ASCB, ASSB, …, SMF30, SMF32, SMF7x, …
  - You will usually not see the term MT1ET because it is implied
- Several new metrics to describe how efficiently core resources could be utilized…

* Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.

…and several new metrics for SMT…

- New metrics:
  - WLM/RMF: Capacity Factor (CF), Maximum Capacity Factor (mCF)
  - RMF: Average Thread Density, Productivity (PROD)
- How are the new metrics derived?
  - Hardware provides metrics (counters) describing the efficiency of processor (cache use/misses, number instructions when one or two threads were active…)
  - LPAR level counters are made available to the OS
  - MVS HIS component and supervisor collect LPAR level counters. HIS provides HISMT API to compute average metrics between “previous” HISMT invocation and “now” (current HISMT invocation)
  - System components (WLM/SRM, monitors such as RMF) retrieve metrics for management and reporting

* Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.

Transitioning into MT2 mode: WLM considerations (1)

- **Less overflow from zIIP to CPs** may occur because
  - zIIP capacity increases, and
  - number of zIIP CPUs double
- CPU time and CPU service **variability may increase**, because
  - Threads which are running on a core at the same time influence each other
  - Threads may be dispatched at TD1 or TD2
- Sysplex workload routing: routing recommendation may change because
  - zIIP capacity will be adjusted with the mCF to reflect MT2 capacity
  - mCF may change as workload or workload mix changes

* Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.

Transitioning into MT2 mode: WLM Considerations (2)

- **Goals should be verified** for zIIP-intensive work, because
  - The number of zIIP CPUs double and the achieved velocity may change
    - "Chatty" (frequent dispatches) workloads may profit because there is a chance of more timely dispatching
    - More capacity is available
    - Any single thread will effectively run at a reduced speed and the achieved velocity will be lower.
      Affects processor speed bound work, such as single threaded Java batch

* Statements regarding IBM future direction and intent are subject to change or withdrawal, and represent goals and objectives only.

What I hope I covered…..

IBM

- What are dispatchable units of work on z/OS
- How WLM manages dispatchable units of work
- The role of HiperDispatch and Warning Track
- Dispatching work to zIIP engines
- z13 and Simultaneous Multithreading (SMT)

IBM Notice Regarding Specialty Engines

Any information contained in this document regarding Specialty Engines ("SEs") and SE eligible workloads provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIIPs, zAAPs, and IFLs). IBM authorizes customers to use IBM SEs only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the "Authorized Use Table for IBM Machines" provided at:
www.ibm.com/systems/support/machine_warranties/machine_code/aut.html ("AUT").

No other workload processing is authorized for execution on an SE.

IBM offers SEs at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.