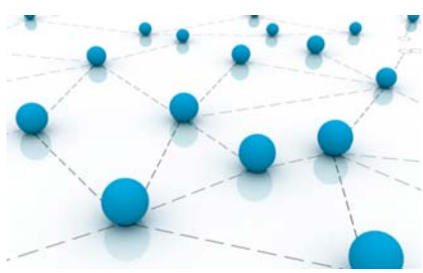
Glenn Anderson, IBM Lab Services and Training

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

September 2015

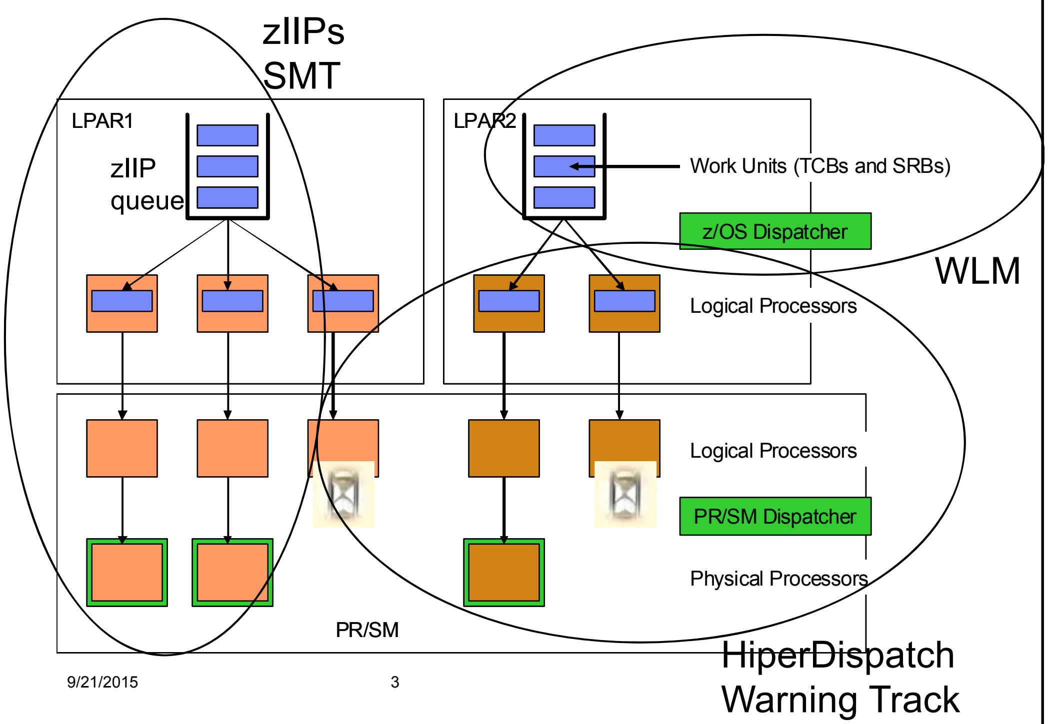
What I hope to cover......

- 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

IBM



9/21/2015 3

z/OS dispatchable units

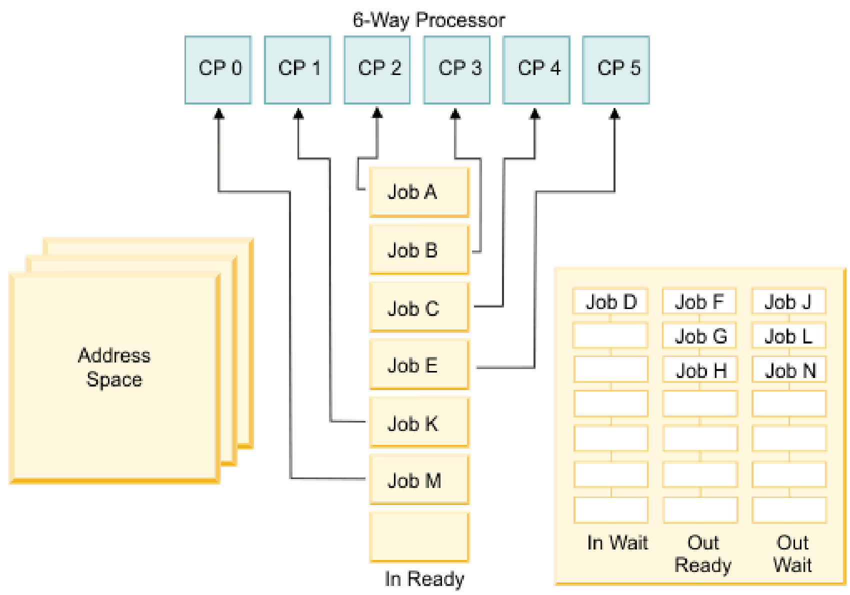
IBM

- **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

9/21/2015 4

z/OS dispatching work

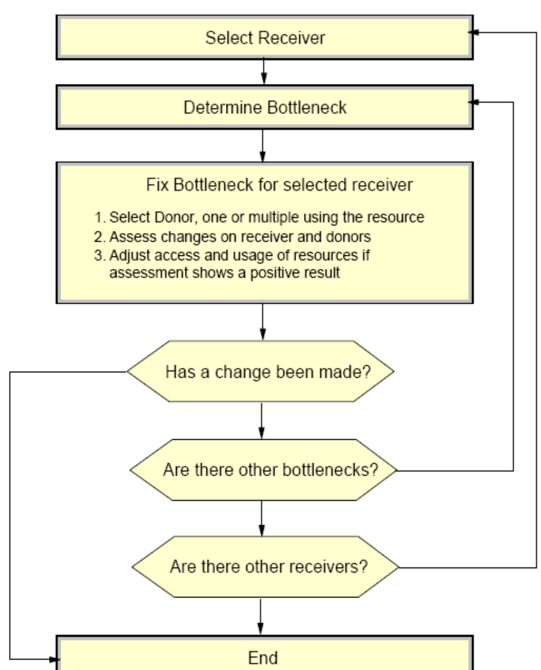
IBM



9/21/2015 5

WLM policy adjustment algorithm

IBM

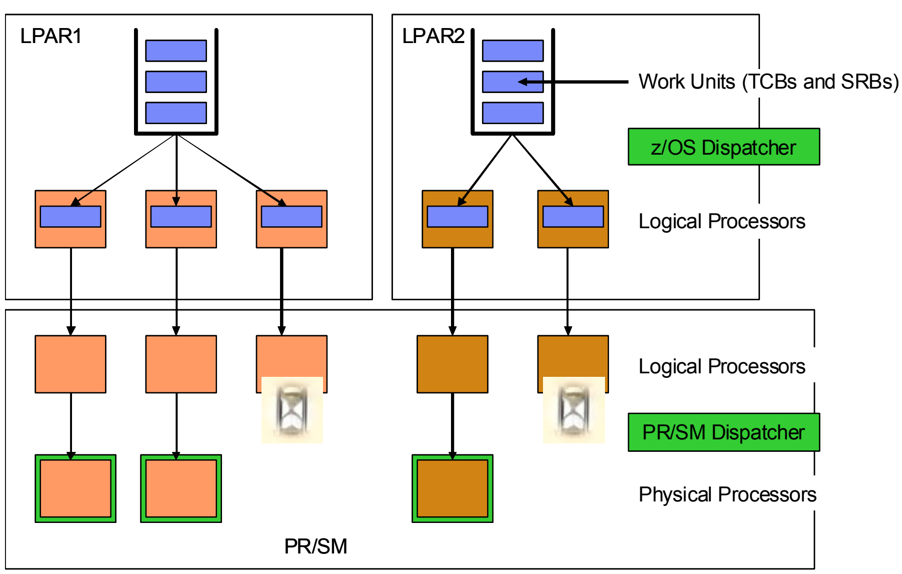


9/21/2015 6

WLM dispatching priority usage

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

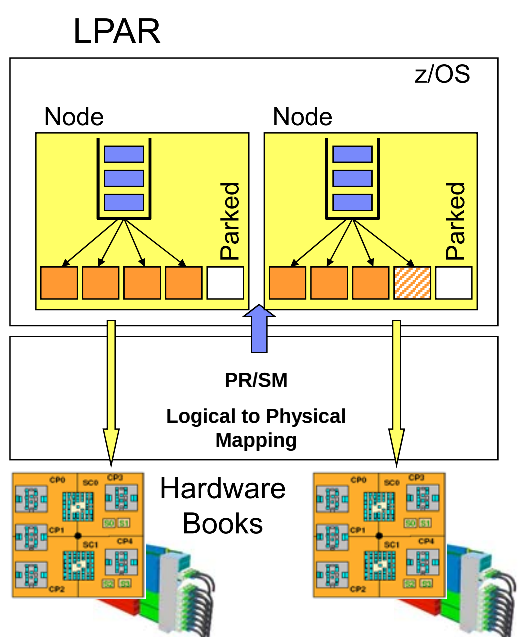


HiperDispatch mode

- 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 high share CPs in each node
  - Each node has TCBs and SRBs assigned to the node
  - Optimizes the execution of work units on z/OS



(z13 – Drawer / Nodes / PU Chips)

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
```

RMF CPU activity report

```
---------------- TIME % ----------------          LOG PROC
ONLINE     LPAR BUSY     MVS BUSY    PARKED        SHARE %
100.00     96.60         96.74         0.00        100.0  HIGH
100.00     97.51         97.69         0.00        100.0  HIGH
100.00     96.02         96.23         0.00         96.0  MED
100.00     39.26         80.81        51.23          0.0  LOW
100.00     48.71         79.90        38.77          0.0  LOW
100.00     41.06         79.34        48.01          0.0  LOW
100.00     12.42         78.35        84.11          0.0  LOW
100.00      0.00         -----       100.00          0.0  LOW
100.00      0.00         -----       100.00          0.0  LOW
100.00     33.05         80.34        58.68          0.0  LOW
```

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
```

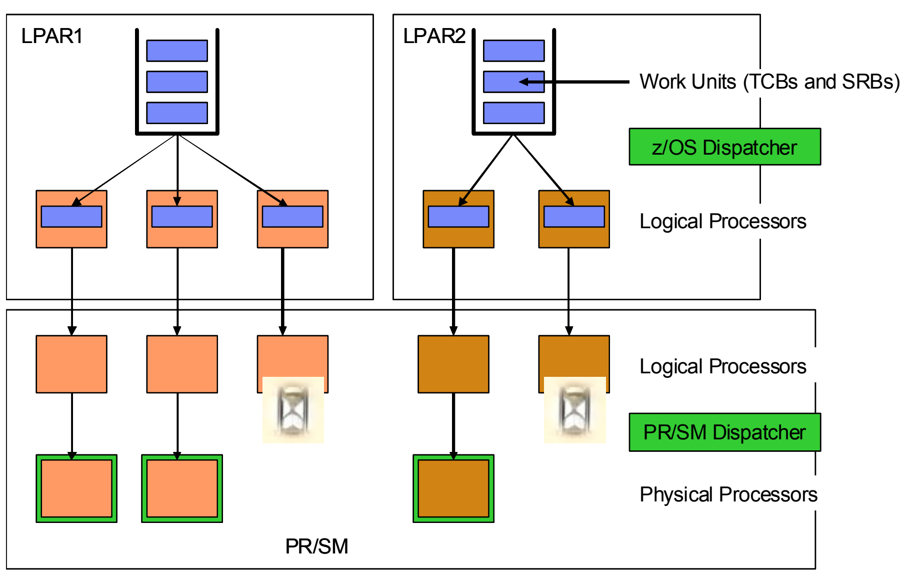
Total LPAR weight = 1000
LPAR1 494/1000 = .494 * 53 CPs = 26.18 CPs
LPAR2 446/1000 = .446 * 53 CPs = 23.64 CPs

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



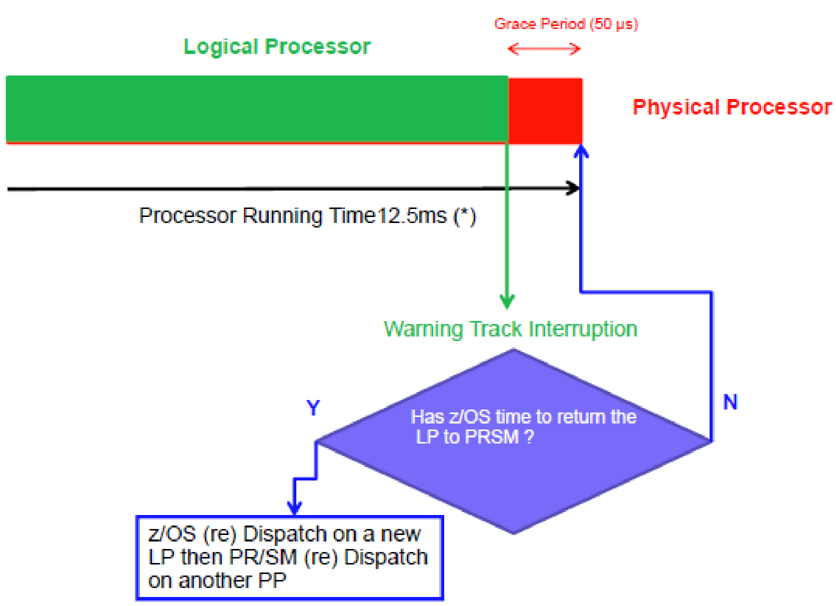
Warning track

IBM

- 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

IBM



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

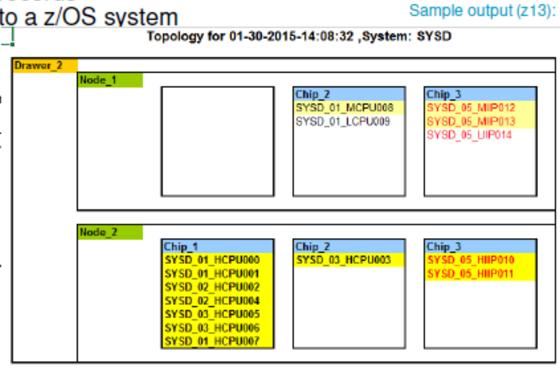
Name	Qualifier	Description
WTRKCP (WTRKAAP) (WTRKIIP)	cpu-id	The percentage of times PR/SM issued a warning-track interruption to a processor and z/OS was able to return it to PR/SM within the grace period.
WTRKTCP (WTRKTAAP) (WTRKTIIP)	cpu-id	Time in microseconds that a purpose processor was yielded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

WLM Topology Report Tool

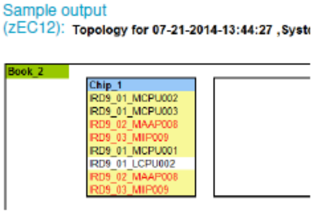
IBM

- 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 (z13):

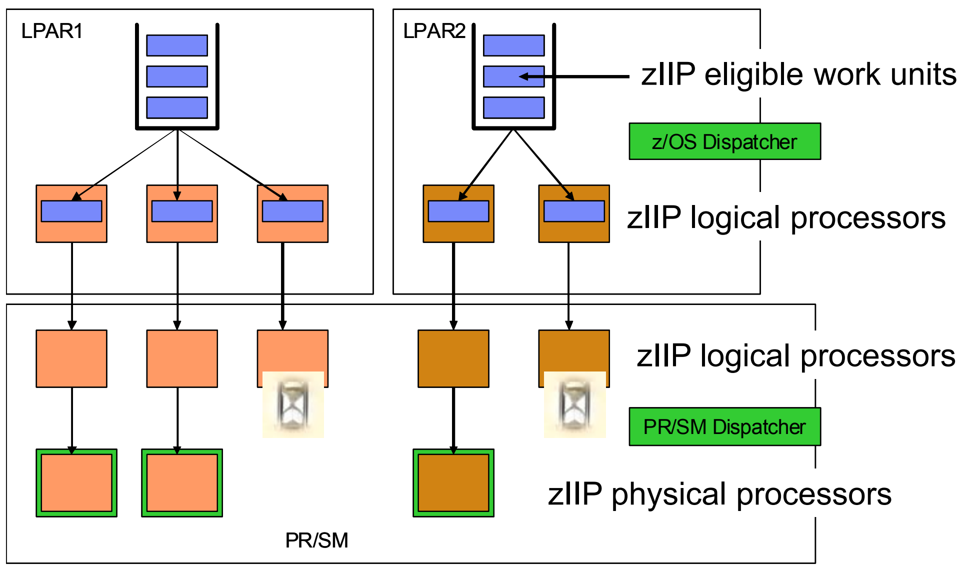


Sample output (zEC12):



12

Dispatching zIIP eligible work



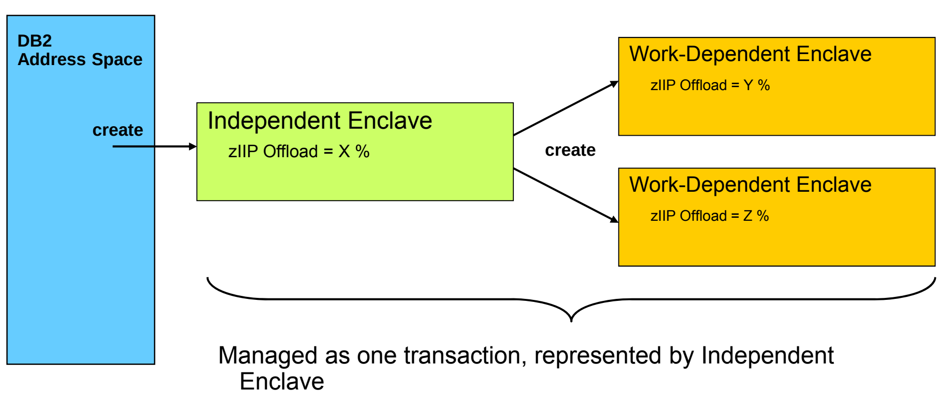
IBM z Integrated Information Processor (zIIP) on the z13

- **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

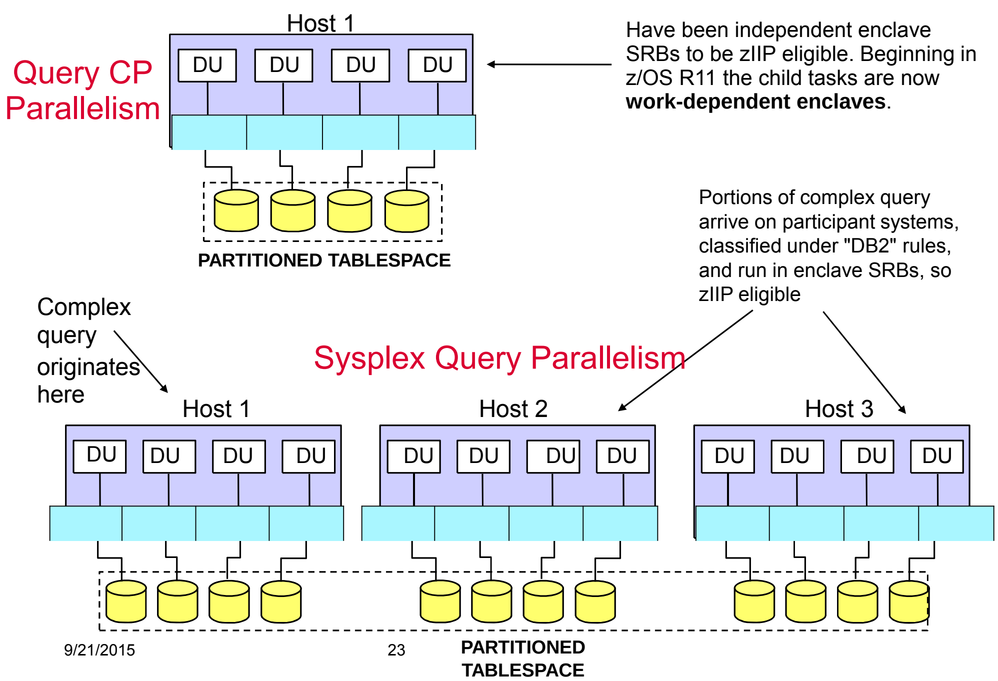
Specialty CP	Eligible	Major Users
zAAP or zIIP on z13	Any Java Execution	Websphere CICS Native apps XMLSS
zIIP	Enclave SRBs	DRDA over TCPIP DB2 Parallel Query DB2 Utilities Load, Reorg, Rebuild DB2 V9 z/OS remote native SQL procedures TCPIP - IPSEC XMLSS zIIP Assisted HiperSockets Multiple Write Virtual 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 SDSF (V2.2)

Work-Dependent enclaves



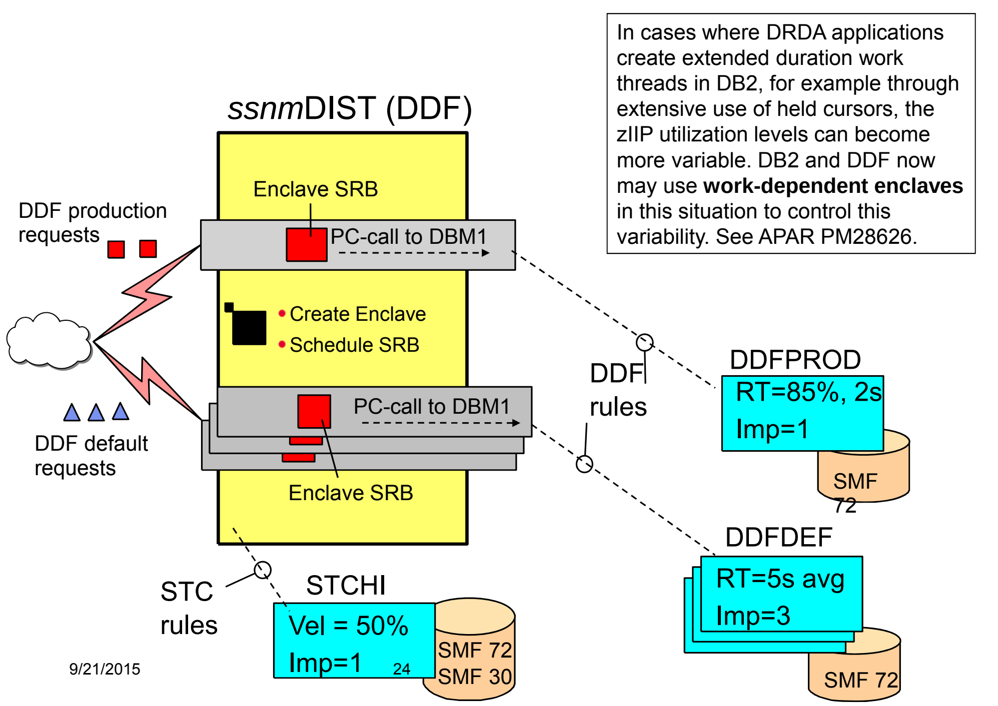
Implement a new type of enclave named “Work-Dependent” as an extension of an Independent Enclave. A Work-Dependent enclave becomes part of the Independent Enclave’s transaction but allows to have its own set of attributes (including zIIP offload percentage)

DB2 parallel query, enclave SRBs and zIIPs



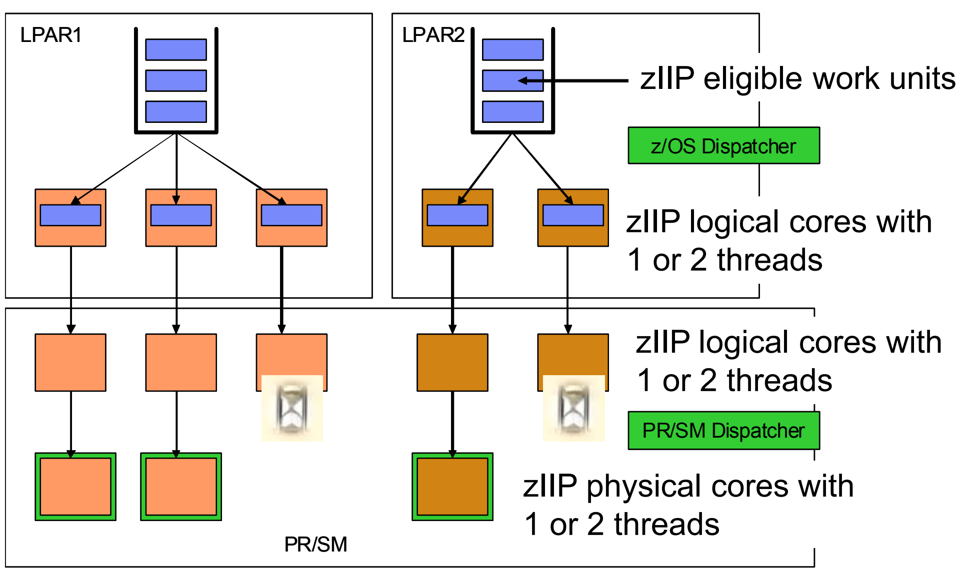
DDF and work-dependent enclaves

IBM



Work-dependent enclaves in SDSF

zIIP processors and simultaneous multithreading

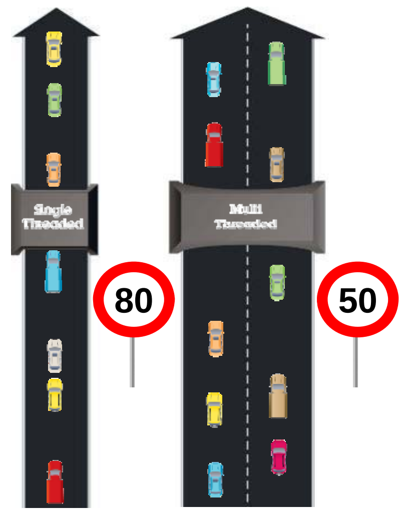


9/21/2015 26

z13 - Simultaneous Multithreading (SMT)

IBM

- "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



Two lanes process more traffic overall

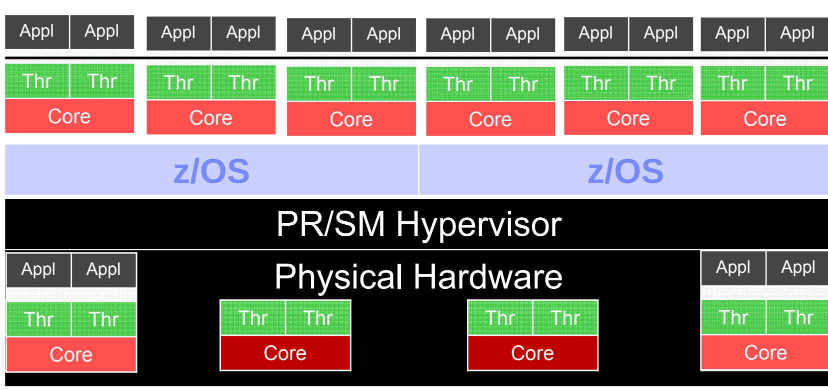
Note: Speed limit signs for illustration only

**Wikipedia®*

9/21/2015 27

z13 - SMT Exploitation

IBM



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

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.

z13 – z/OS SMT Metrics

- Capacity Factor (CF)
  - How much work core actually completes for a given workload mix at current utilization - relative to single thread
  - MT-1 Capacity Factor is 1.0 (100%)
  - MT-2 Capacity Factor is workload dependent
- Maximum Capacity Factor (mCF)
  - How much work a core can complete for a given workload mix at most
- Core Busy Time
  - Time any thread on the core is executing instructions when core is dispatched to physical core
- Average Thread Density
  - Average number of executing threads during **Core Busy Time** (Range: 1.0 - 2.0)
- Productivity
  - Core Busy Time Utilization (percentage of used capacity) for a given workload mix
  - Productivity represents capacity in use (**CF**) relative to capacity total (**mCF**) during **Core Busy Time**.
- Core Utilization
  - Capacity in use relative to capacity total over some time interval
  - Calculated as **Core Busy Time** x **Productivity**

Actual MT-2 Efficiency

Estimated max MT-2 Efficiency

% Used MT-2 Core Capacity during Core Busy Time

% Used MT-2 Core Capacity during Measurement Interval

 30

z13 – SMT: Postprocessor CPU Activity Report

- PP CPU activity report provides new metrics when SMT is active
  - MT Productivity and Utilization of each logical core
  - MT Multi-Threading Analysis section displays MT Mode, MT Capacity Factors and average Thread Density
- One data line in PP CPU activity report represents one thread (CPU)
  - CPU NUM designates the logical core
- Some metrics like TIME % ONLINE and LPAR BUSY provided at core granularity only

```
                                          C P U   A C T I V I T Y
z/OS V2R1                 SYSTEM ID CB8B              DATE 02/02/2015              INTERVAL 15.00.004
                          RPT VERSION V2R1 RMF        TIME 11.00.00                CYCLE 1.000 SECONDS
---CPU---     --------------- TIME % ------------   --- MT % ----  LOG PROC       --I/O INTERRUPTS–
NUM  TYPE     ONLINE    LPAR BUSY  MVS BUSY PARKED  PROD    UTIL   SHARE %        RATE      % VIA TPI
 0    CP      100.00    68.07      67.94      0.00  100.00  68.07  100.0  HIGH    370.1     13.90
 1    CP      100.00    46.78      46.78      0.00  100.00  46.78   52.9  MED       5.29    16.93
...
TOTAL/AVERAGE            8.66      54.17            100.00   8.66  152.9          375.3     13.95
 A    IIP     100.00    48.15      41.70      0.00   85.84  41.33  100.0  HIGH
                                   35.66      0.00
 B    IIP     100.00    38.50      32.81      0.00   85.94  33.09  100.0  HIGH
                                   26.47      0.00
...
TOTAL/AVERAGE           29.48      23.23             86.47  25.39  386.7
------------ MULTI-THREADING ANALYSIS --------------
 CPU TYPE     MODE      MAX CF         CF        AVG TD
      CP        1       1.000        1.000       1.000
      IIP       2       1.485        1.279       1.576
```

MT-2 core capacity used

Productivity of logical core while dispatched to physical core

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

What I hope I covered......

- 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)