IBM z Systems z13 Simultaneous Multi-Threading (R)Evolution

Daniel Rosa
IBM Poughkeepsie
dvrosa@us.ibm.com

#SHAREorg

SHARE is an independent volunteer-run information technology association that provides education, professional networking and industry influence.

SHARE in Seattle 2015

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

IBM*
IBM Logo*

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon 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 throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Agenda

- SMT Overview
- General Industry SMT Exploitation
- z Systems SMT Exploitation
- z/OS SMT Exploitation with Runtime Capacity and Accounting
- z/OS SMT Control Implications
- References

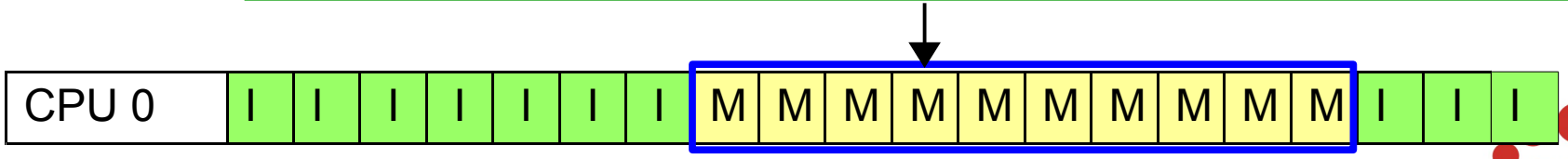
z Systems pre-z13 Background

APPL
CPU

- A “ready” CPU executes a “ready” instruction (not resolving a cache miss) immediately
- A “not ready” CPU generally originates from a “not ready” instruction (resolving a cache miss)
  - CPU unproductive until instruction “ready”
  - z Systems workloads generally experience many cache misses

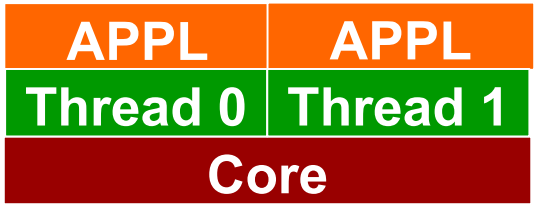
Legend: M=Resolve Cache Miss, I=Execute Ready Instruction

1 not ready instruction resolving a cache miss. The CPU is unproductive.
If the CPU had a ready instruction, it could execute during these cycles.



CPU Cycles (10 instructions in 20 cycles)

z Systems New z13 Machine



- A core contains multiple threads resembling CPUs
  - Each thread has its own state (PSW, regs, CPU Timer, etc)
  - "Ready" threads compete for use of shared core execution resources
    - Every cycle core arbitrates "ready" threads fairly, winner executes
  - "Not ready" threads do not compete, still unproductive
- z13 supports SMT-2 (2 threads per core)
  - Core has "ready" instructions more frequently, leads to higher core throughput

Legend: M=Resolve Cache Miss, F=Execute Ready Floating Point Instruction
O=Execute Ready Other Instruction, o=Defer Ready Other Instruction

Core arbitrates such that
1 ready instr executes
1 ready instr deferred

1 ready instr executes,
1 not ready instr resolving miss

2 not ready instr resolving misses
core unproductive

2 ready instr execute

Thread 0	O	o	O	o	O	O	O	M	M	M	M	M	M	M	M	M	M	O	O	O
Thread 1	o	O	o	O	M	M	M	F	F	M	M	M	M	M	M	M	M	O	O	F

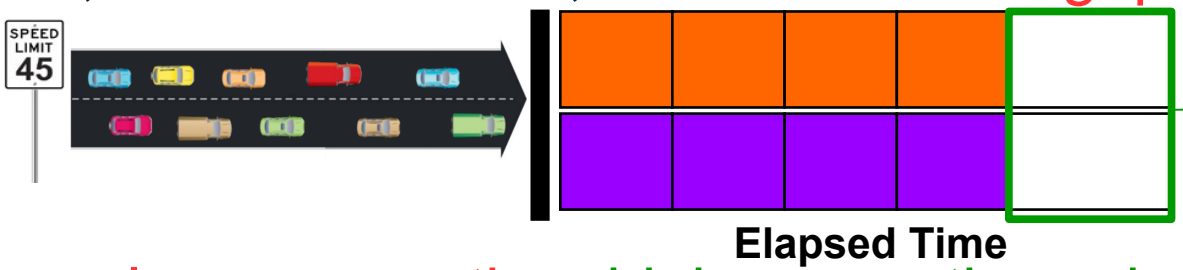
Core Cycles (14 instructions in 20 cycles, 1.4x throughput gain)

Effects of SMT

SMT-1 with 2 work units:

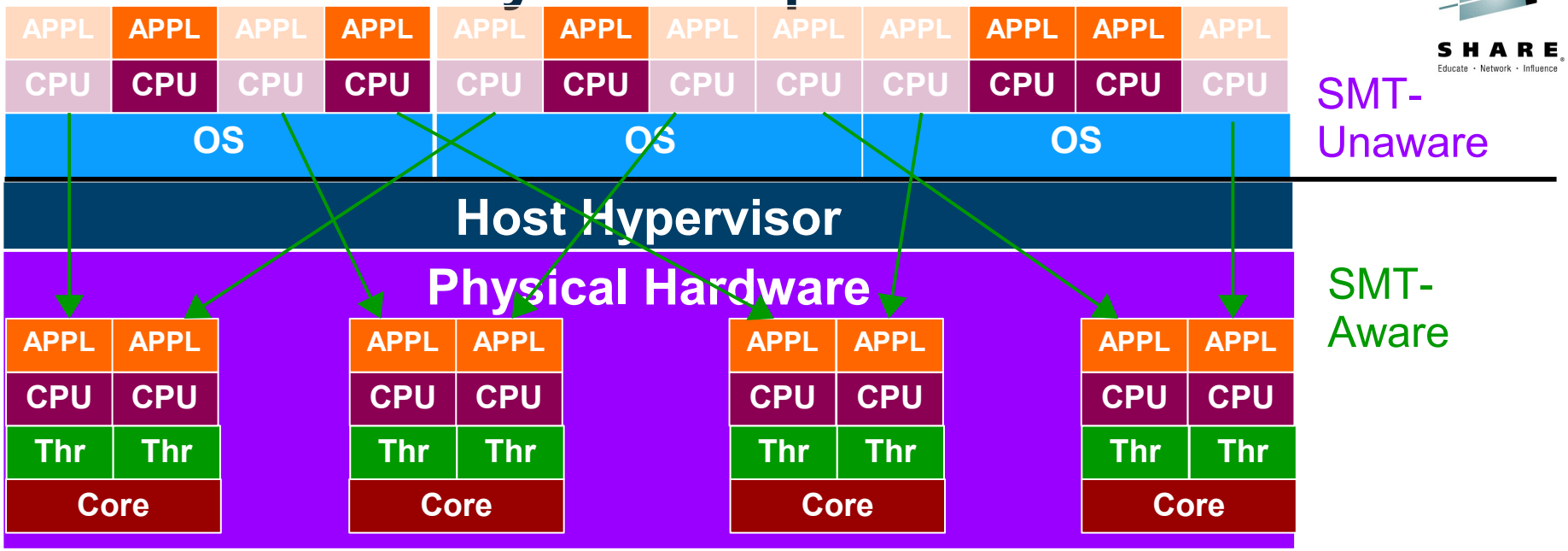
- Slower access, faster execution, lower core throughput

SMT-2 with 2 work units:



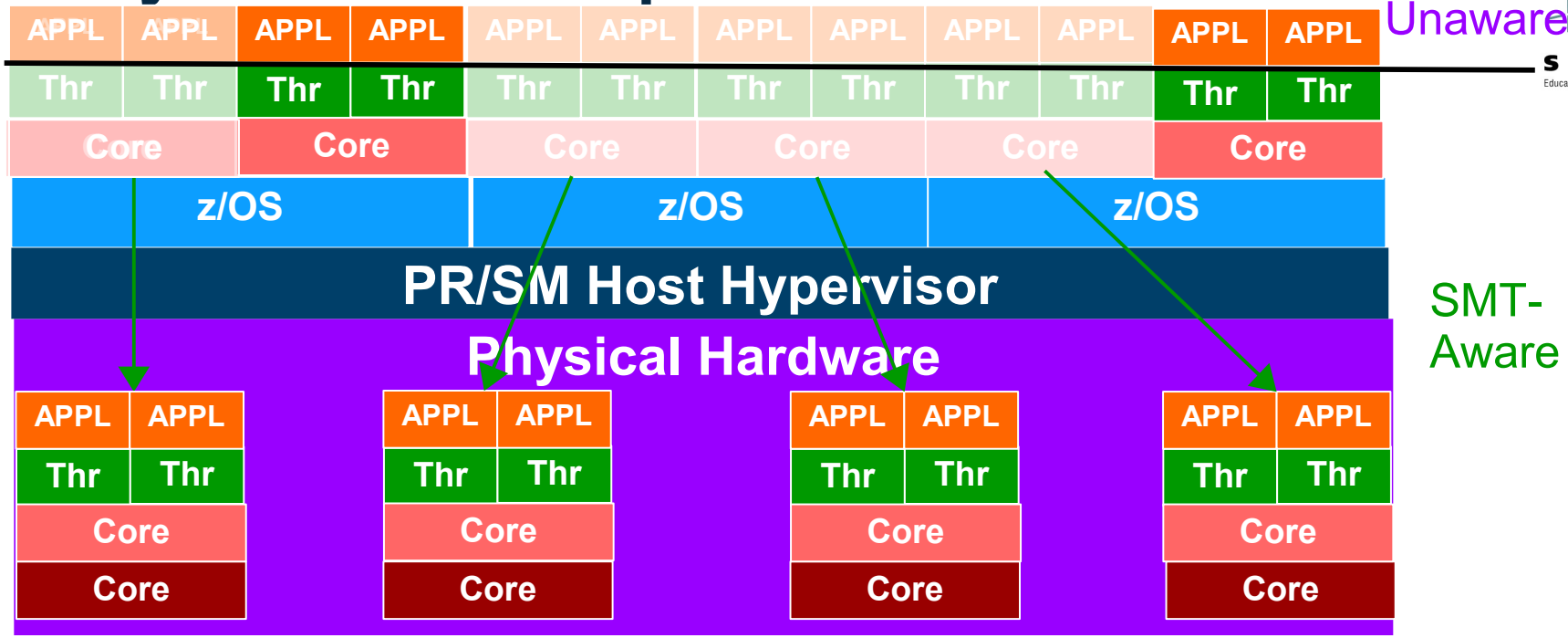
- Faster access, slower execution, higher core throughput
- Thread Density=2 (2 non-waiting threads, abbreviated TD=2), Workload dependent:
  - Capacity gain (benchmarks observe +10% to +40%)
  - Speed loss (benchmarks observe -30% to -40% <like z196>)
- Thread Density=1 (1 waiting & 1 non-waiting thread, abbr TD=1)
  - Performs like SMT-1: no capacity gain, no speed loss
  - Workload dependent capacity in use, free
- Core Waiting (all threads waiting)

General Industry SMT Exploitation



- Generally focuses on maximizing core throughput
- Hypervisor provides SMT transparently to SMT-unaware OS
  - Dispatches OS logical CPU to physical thread, OS receives:
    - Less predictable capacity (core floats b/w TD=1, TD=2)
    - Less meaningful accounting (disconnected from capacity)
    - More latency, response time variability

z Systems SMT Exploitation



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

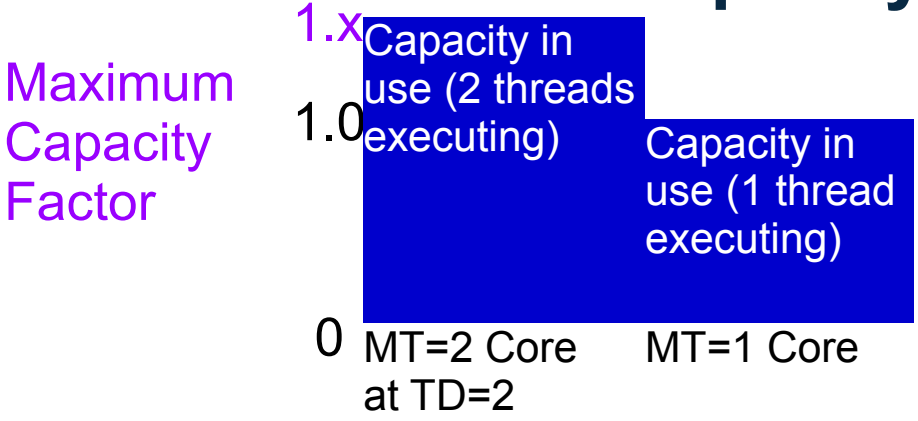
z/OS MT Exploitation with z13

- Architecture and z/OS refer to MT, synonymous with SMT
- MT available via PTFs on z/OS V2R1
- Define LOADxx PROCessor VIEW for the life of the IPL
  - PROCVIEW CORE – MT environment, controls
    - Pre-z13, must use MT controls (common look and feel)
    - z13 becomes MT-2 config (defines 2 threads per core)
  - PROCVIEW CPU – Pre-z13 environment, controls
- Without IPL, IEAOPTxx supports operator changing MT Mode=x (active threads per online core, abbreviated MT=x) by core type:
  - MT_ZIIP_MODE=1|2 – active threads per online zIIP core
    - Enables MT=2 always, or MT=2 (OLTP) and MT=1 (batch)

Capacity and Accounting Overview

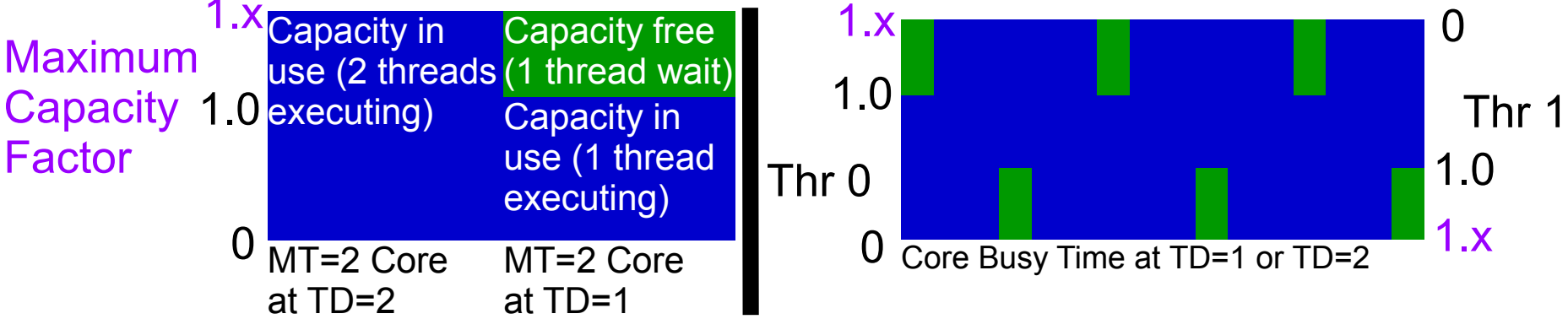
- Apply existing concepts to MT environment by:
  - Reinterpreting existing fields, using new MT metrics
- Generally requires the following over some time interval:
  - Transactions complete (External Throughput Rate)
  - Capacity max, free, in use (utilization) by job (accounting)
- Capacity in use per transaction, different for every workload
  - MT Mode=1, capacity max known, free & in use calculable
  - MT Mode=2, capacity max, free, in use workload dependent
    - Measures workload behavior at TD=1, TD=2, delivers meaningful capacity and accounting **at runtime**

New MT Metric: Maximum Capacity Factor



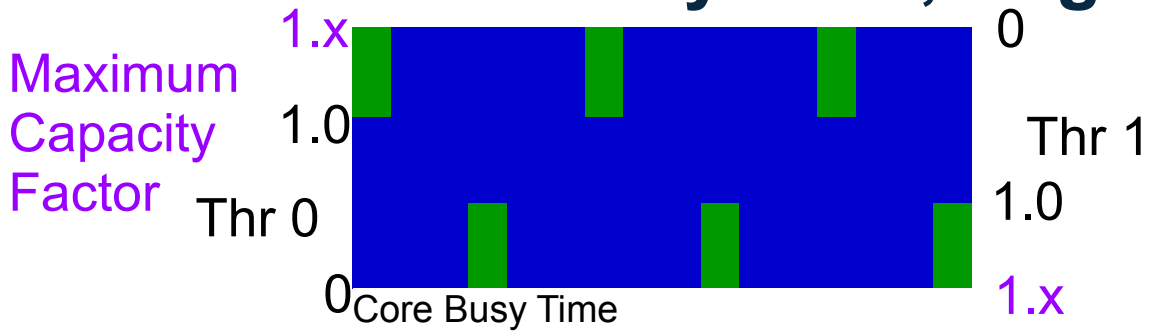
- Max Capacity Factor: How much work a core can complete
- MT=1 Max Capacity Factor is 1.0 (100%)
- MT=2 Max Capacity Factor represents how much work a MT=2 core at TD=2 completes relative to a MT=1 core
- Expect Max Cap Factor between 1.1-1.4 (110%-140%)
  - Theoretical highest Max Capacity Factor=2.0, can get < 1.0

New MT Metric: Capacity Factor



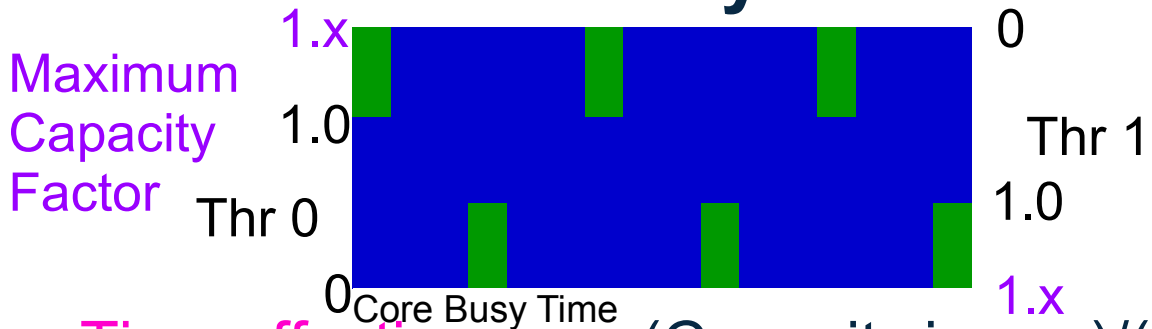
- Capacity Factor: How much work core actually completes
- MT=1 Capacity Factor is 1.0 (100%)
- MT=2 Capacity Factor represents how much work a MT=2 core is actually completing relative to a MT=1 core. Considers:
  - MT=1, MT=2 Max Cap Factors. Time at TD=1, TD=2
- Expect Cap Factor 1.0-1.4 (100%-140%), can be < 1.0 (100%)
  - More TD=1 nears 1.0, more TD=2 nears max Cap Factor

New MT Metrics: Core Busy Time, Avg TD



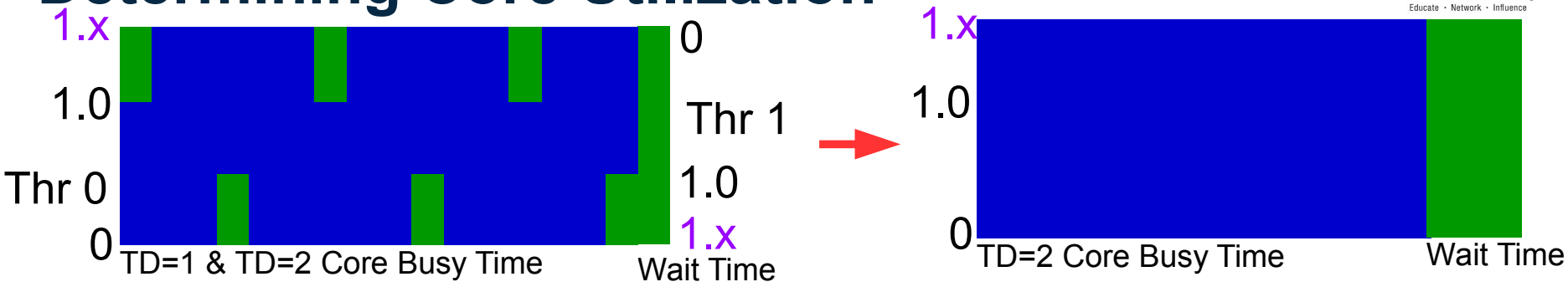
- Core Busy Time: Core execution time, includes:
  - Time any thread on the core is executing
  - Time core executes at TD=1, TD=2
- Average Thread Density: Average executing threads during Core Busy Time
  - MT=1 Average Thread Density is 1.0 (100%)
  - MT=2 Average Thread Density is 1.0-2.0 (100%-200%)

New MT Metric: Productivity



- Core Busy Time effectiveness: (Capacity in use)/(Capacity max)
  - MT=1 all capacity in use. Productivity is 1.0 (100%)
  - MT=2 some capacity in use (some at TD=1, all at TD=2)
- MT=2 Productivity is capacity in use (Cap Factor) relative to capacity max (Max Cap Factor) during Core Busy Time
  - Considers Max Cap Factor, Cap Factor, Average TD
- w/ Max Cap Factor=1.4, expect productivity 0.7-1.0 (70%-100%)
  - More TD=1 nears 1.0/1.4 = 0.7 (70%)
  - More TD=2 nears 1.4/1.4 = 1.0 (100%)
- If TD=1 2s, TD=2 2s, max Cap Factor=1.25
  - Productivity = [2 * 1.0 + 2 * 1.25] / [4 * 1.25] = 90%

Determining Core Utilization



- Core Utilization = (Core Busy Time) / (Core Available Time)
  - MT=1 busy all capacity in use, available time is busy+wait
  - MT=2 busy some capacity in use and some capacity free
    - Multiplying by productivity yields “effective” Core Busy
    - Core Utilization=(“effective” Core Busy Time) / (Core Available Time)
    - If TD=1 2s, TD=2 2s, wait time=6s, Productivity = 90% (previous chart)
      - “Effective” Core Busy Time = (2s+2s)*(0.9) = 3.6s
      - Core Utilization = 3.6s / (2s+2s+6s) = 36%

z/OS Accounting With MT Mode=2

- Similar capacity use should yield similar accounting
- MT Mode=1, ready instructions execute immediately
  - Account for capacity use with CPU Timer delta
- MT Mode=2, ready instructions compete to execute
  - CPU Timer decrements on both threads each cycle win or lose
  - Convert CPU Timer delta to MT=1 Equivalent Time
    - (MT=1 Core Capacity use) = (Capacity Factor)*(Core Busy Time)
      - Use Avg Thread Density to distribute over threads
      - If Cap Factor=1.25, Core Busy Time=4s, Avg TD=2.0, core used 1.25 * 4s = 5s of MT=1 Equivalent Time and each thread used 5s / 2.0=2.5s of MT=1 Equivalent Time
  - All accounting fields (SMF 30, 72, etc), services (TimeUsed) transparently become MT=1 Equivalent Times

HISMT Service

- Provides MT metrics between current, previous call. Requires:
  - Authorization: Supervisor State key 0
  - Dispatchable Unit: Task or SRB
  - Cross Memory Mode: Any HASN, any PASN, any SASN
  - Amode: 31-bit
  - ASC Mode: Primary
  - Interrupts: Enabled/Disabled for I/O & external interrupts
  - Control Parameters, save area must be addressable from primary and be in Disabled REFerence or Fixed storage
- Callable for any PROCVIEW, any MT Mode, HIS address space can be active/inactive

HISMT Syntax

```
HISMT   INTVAREA=xintvarea
        ,INTVAREALEN=xintvarealen
       [,PRODCLASS=NO|YES]
       [,PRODCORE=NO|YES]
       [,CAPCLASS=NO|YES]
       [,MAXCAPCLASS=NO|YES]
       [,COREBUSYTIME=NO|YES] ← requires z13 PROCVIEW CORE
                                 (see CVT macro field
                                 CvtMultiCpusPerCore)
       [,AVGTDCLASS=NO|YES]
       [,RETCODE=xretcode]
       [,RSNCODE=xrsncode]
```

MT=1 RMF Cpu Activity Report

```
--CPU--- ------------ TIME % ------------ --- MT % ---    LOG PROC
NUM TYPE ONLINE LPAR BUSY MVS BUSY PARKED PROD   UTIL     SHARE %
...
 4   IIP 100.00 70.47     70.32      0.00 100.00 70.47    100.0
 5   IIP 100.00 55.40     55.32      0.00 100.00 55.40    100.0
 6   IIP 100.00 71.62     71.49      0.00 100.00 71.62    100.0
 7   IIP 100.00 57.77     57.71      0.00 100.00 57.77    100.0
TOTAL/AVERAGE   63.81     63.71           100.00 63.81    400.0
------------ MULTI-THREADING ANALYSIS ---------------
 CPU TYPE    MODE     MAX CF           CF       AVG TD
      CP        1      1.000        1.000        1.000
      IIP       1      1.000        1.000        1.000
```

- LPAR Busy = PR/SM dispatching logical CPU to physical CPU
- MVS Busy = Unparked logical CPU not waiting
- Parked = Logical CPU parked
- CPU utilization = (sum of CPUs' LPAR Busy)
- Available CPU capacity = (Total Log Proc Share %) - (Sum of CPUs' LPAR Busy)

MT=2 RMF Cpu Activity Report

```
--CPU--- ------------ TIME % ------------ --- MT % -- LOG PROC
NUM TYPE ONLINE LPAR BUSY MVS BUSY PARKED PROD  UTIL  SHARE %
...
 4   IIP 100.00 78.23     67.24       0.00 87.30 68.29 100.0
                          58.40       0.00
 5   IIP 100.00 59.46     50.57       0.00 85.64 50.92 100.0
                          41.88       0.00
 6   IIP 100.00 80.77     70.34       0.00 88.38 71.38 100.0
                          62.20       0.00
 7   IIP 100.00 63.67     55.08       0.00 86.43 55.03 100.0
                          45.52       0.00
TOTAL/AVERAGE   70.53     56.41            86.94 61.41 400.0
------------ MULTI-THREADING ANALYSIS ---------------
 CPU TYPE    MODE     MAX CF           CF      AVG TD
      CP        1      1.000        1.000       1.000
      IIP       2      1.473        1.283       1.600
```

- LPAR Busy = PR/SM dispatching logical core to physical core
- MVS Busy = Unparked logical CPU not waiting
- Parked = Logical CPU parked

MT=2 RMF Cpu Activity Report

```
--CPU--- ------------ TIME % ------------ --- MT % -- LOG PROC
NUM TYPE ONLINE LPAR BUSY MVS BUSY PARKED PROD  UTIL  SHARE %
...
 4   IIP 100.00 78.23     67.24       0.00 87.30 68.29 100.0
                          58.40       0.00
 5   IIP 100.00 59.46     50.57       0.00 85.64 50.92 100.0
                          41.88       0.00
 6   IIP 100.00 80.77     70.34       0.00 88.38 71.38 100.0
                          62.20       0.00
 7   IIP 100.00 63.67     55.08       0.00 86.43 55.03 100.0
                          45.52       0.00
TOTAL/AVERAGE   70.53     56.41            86.94 61.41 400.0
------------ MULTI-THREADING ANALYSIS ---------------
 CPU TYPE    MODE     MAX CF           CF      AVG TD
      CP        1      1.000        1.000       1.000
      IIP       2      1.473        1.283       1.600
```

- Core utilization (% MT UTIL) = [(LPAR Busy)*(Productivity)]
  - Total zIIP (MT=2) core utilization (% MT UTIL): 245.62%
- Available core capacity = (Total Log Proc Share %) - (Sum of cores' MT % UTIL)
  - Total zIIP (MT=2) available: 400%-245.62%=154.38%

MT=2 Workload Activity Report

```
                W O R K L O A D   A C T I V I T Y
z/OS V2R1 SYSPLEX PATPLX29 DATE 01/21/2015 INTERVAL 10.14.053
       RPT VERSION V2R1 RMF        TIME 21.46.55
REPORT BY: POLICY=PATPLEX   WORKLOAD=WASWKLD   SERVICE CLASS=WASTRANS
       RESOURCE GROUP=*NONE         CRITICAL=CPU
-TRANSACTIONS-   ... SERVICE TIME   ---APPL %---
AVG       12.69       CPU 3545.743  CP    252.44
MPL       12.69       SRB    0.000  AAPCP   0.00
ENDED   5502452       RCT    0.000  IIPCP   0.43
END/S   8960.87       IIT    0.000
#SWAPS         0      HST    0.000  AAP      N/A
EXCTD          0      AAP      N/A  IIP    220.69
AVG ENC   12.69       IIP 1995.640
```

- Service Times are in MT=1 Equivalent Time units
- APPL % is % of core relative to its max Capacity Factor
  - IIP APPL %=(1995.64s)*100/ [(614s)*(1.473)] = 220.65 %
  - CP APPL % = (3545.743s-1995.640s) * 100 / [(614s) * (1.0)] = 252.46 %

MT=2 Workload Activity Report

```
                W O R K L O A D   A C T I V I T Y
z/OS V2R1 SYSPLEX PATPLX29 DATE 01/21/2015 INTERVAL 10.14.053
       RPT VERSION V2R1 RMF        TIME 21.46.55
REPORT BY: POLICY=PATPLEX   WORKLOAD=WASWKLD   SERVICE CLASS=WASTRANS
       RESOURCE GROUP=*NONE        CRITICAL=CPU
-TRANSACTIONS-  ... SERVICE TIME  ---APPL %---
AVG       12.69       CPU 3545.743  CP    252.44
MPL       12.69       SRB    0.000  AAPCP   0.00
ENDED   5502452       RCT    0.000  IIPCP   0.43
END/S   8960.87       IIT    0.000
#SWAPS         0      HST    0.000  AAP      N/A
EXCTD          0      AAP      N/A  IIP    220.69
AVG ENC   12.69       IIP 1995.640
```

- zIIP Capture Ratio = Sum(service classes' APPL %) * 100 / Sum(cores' MT % UTIL). If WASTRANS is the only service class using zIIPs:
  - zIIP Capture Ratio: 220.69 * 100 / 245.62 = 89.9%

Capacity and Accounting Observations

- Production zIIP transactional workloads 20-90% Busy (Washington Systems Center recommends 1-2 zIIPs run <=50% Busy, >=3 zIIPs run <=70% busy) generally receive optimal metric results
  - Yields workload representative sample at TD=1, TD=2
- Cases where MT metrics may yield suboptimal results:
  - High/low utilization can yield insufficient samples at TD=1/2
  - Non-production, non-transactional, or unfriendly MT workloads like compute intensive, small memory footprint
- Indicators of suboptimal metric results may include:
  - Frozen metrics (typically insufficient samples at TD=1/2)
  - Max Cap Factor, Cap Factor at min (0.5) or max (2.0)

HIS Support Of (Core) MT-Diagnostic Counter Set

- Hardware actives, enables on z13 w/ PROCVIEW CORE
- Modify HIS command updates for MT-Diagnostic Counter Set:
  - CTRSET=(MTD or MTDIAG or MTDIAGNOSTIC) collects it
  - CTRSET=HDWR or CTRSET=COMPLETE includes it
- SMF Type 113 Record Changes
  - Subtype 1 counter data sections are CPU or core specific
    - MTDiag deltas in subtype 1 for 1st online CPU per core
    - MTDiag counters not in subtype 2
  - Counter data sections for existing counter sets for Subtype 1 or 2 remain CPU specific
- USS output file
  - New counter data section for MT-diagnostic counter set for first online CPU per core

PROCVIEW CORE Control Implications

- System programmers manage cores (including MT Mode)
  - z/OS manages threads accordingly
- Parmlib changes:
  - CONFIGxx – must specify CORE keyword e.g.:
    - CORE(0),ONLINE,STANDARD
  - LOADxx DYNCPADD nnnn – cores z/OS prepares to dynamically add
- Stand Alone Dump:
  - Continue to follow the same procedure
  - Includes appropriate CPU/thread status for any PROCVIEW

PROCVIEW CORE Control Implications

- MODE command controls machine check type recovery action via:
  - RECORD=ALL (just record error to LOGREC)
  - RECORD=n,INTERVAL=s, If n machine checks within s seconds:
    - Perform ACR or Timer recovery depending on machine check type
- With PROCVIEW CORE on any hardware:
  - ACR eligible machine check types now RECORD=ALL,CPU=ALL for:
    - PD (Instruction processing Damage)
    - SD (System Damage)
    - IV (InValid PSW or registers)
    - TC (Tod Clock damage)
    - PT (Processor Timer damage)
    - CC (Clock Comparator damage)
    - PS (Primary clock Synchronization)
    - AD (ETR attachment)
    - SL (Switch to Local synchronization)
  - Specifying RECORD=n or INTERVAL=s results in:
    - IGF960I MODE COMMAND REJECTED. PROCVIEW CORE IS IN EFFECT

PROCVIEW CORE Control Implications

- Command changes requiring CORE keyword:
  - Config Core(x),Online – Configs core online for MT Mode
    - MT Mode=1, 1st thread online, 2nd thread offline
    - MT Mode=2, both threads online
  - Config Core(x),Offline – Configs all threads on core offline
  - Config Member=xx – Configs cores according to CONFIGxx
  - Config Online or Config Offline – Lists eligible cores to config
    - Reply to IEE522D accepts CORE(x) to configure
  - Display Matrix=Core – Displays core status (new message)
  - Display Matrix=Config(xx) – CONFIGxx vs system differences

Display Matrix=Core Sample Output

```
IEE174I 22.34.15 DISPLAY M 226
CORE STATUS: HD=Y   MT=2 MODE:  CP=1  zIIP=2
ID    ST   ID RANGE   VP  ISCM  CPU THREAD STATUS
0000   +   0000-0001  H   FC00  +N
0001   +   0002-0003  M   FC00  +N
0002   +   0004-0005  M   FC00  +N
0003   +   0006-0007  LP  FC00  +N
0004   +   0008-0009  L   FC00  +N
0005   +I  000A-000B  M   0200  ++
0006   +I  000C-000D  M   0200  ++
0007   +I  000E-000F  LP  0200  ++
0008   +I  0010-0011  L   0200  ++
```

- CP cores 0-4 online, MT=1, thread 0 online, thread 1 offline (N means offline and cannot be brought online)
- ZIIP cores 5-8 online, MT=2, both threads online

z/OS MT Support Summary

- z/OS V2R1 MT support requires applying the following PTFs:
  - UA90753 (OA43366) - Supervisor, SADump, HIS, Reconfig, IPL/NIP, LoadWait
  - UA90762 (OA43622) - SRM / WLM
  - UA76154 (OA44439) – XCF

- z/OS V2R1 MT optional PTFs include:
  - UA76026 (OA44101) – RMF which will require:
    - UA90772 (OA44624) - USS

References

- z/OS MVS Authorized Assembler Services Reference EDT-IXG
- z/OS MVS Initialization and Tuning Reference
- z/OS MVS System Codes
- z/OS MVS System Commands
- z/OS MVS System Management Facilities (SMF)
- z/OS MVS System Messages, Vol 7 (IEB-IEE)
- z/OS RMF Report Analysis
-
- IBM Redbooks Point-Of-View Publication: z Systems Simultaneous Multithreading Revolution

Questions?

Complete your session evaluations online at www.SHARE.org/Seattle-Eval