Putting SMT to Work in a z/VM Environment

August 10, 2020 - Version 11

Bill Bitner
z/VM Client Focus and Care

bitnerb@us.ibm.com

Trademarks

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

BladeCenter*	FICON*	OMEGAMON*	RACF*	System z9*	zSecure
DB2*	GDPS*	Performance Toolkit for VM	Storwize*	System z10*	z/VM*
DS6000*	HiperSockets	Power*	System Storage*	Tivoli*	z Systems*
DS8000*	HyperSwap	PowerVM	System x*	zEnterprise*	
ECKD	IBM z13*	PR/SM	System z*	z/OS*	

* 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.
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.
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.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.
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.
Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
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
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.
OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.
TEALEAF is a registered trademark of Tealeaf, an IBM Company.
Windows Server and the Windows logo are trademarks of the Microsoft group of countries.
Worklight is a trademark or registered trademark of Worklight, an IBM Company.
UNIX is a registered trademark of The Open Group in the United States and other countries.

* Other product and service names might be trademarks of IBM or other 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.

This information 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) ("SEs"). IBM authorizes customers to use IBM SE 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 SE 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.

Agenda

- Key Concepts
- Terminology and Basic Concepts
- Determining Application Performance
- Variability Factors
- Metrics
- Performance Measurement Results
- Summary

Acknowledgements

- **The following people contributed charts or information to this presentation:**
  - **Bill Bitner**
  - **John Franciscovich**
  - **Emily Hugenbruch**
  - **Damian Osisek**
  - **Dan Rosa**
  - **Xenia Tkatschow**
  - **Brian Wade**
  - **Charles Webb**
  - **Romney White**
  - **Don Wilton**

Key Concepts

- New Terminology
  - Core ≠ Processor ≠ IFL ≠ Thread ≠ CPU ≠ Engine

- Need to capture Response Time and Application Throughput
  - Difficult to provide actual throughput or response time of applications, especially applications that span virtual machines/address spaces or systems

- The relationships among capacity, utilization, and productivity become even more variable with SMT
  - As soon as components of the system started being shared (e.g., cache), variability was introduced
  - SMT increases how much is shared and therefore adds to the variability of performance

- New SMT metrics
  - No one metric describes the environment
  - Use various metrics and their relationships to gain understanding

Terminology and Basic Concepts

Enabling SMT

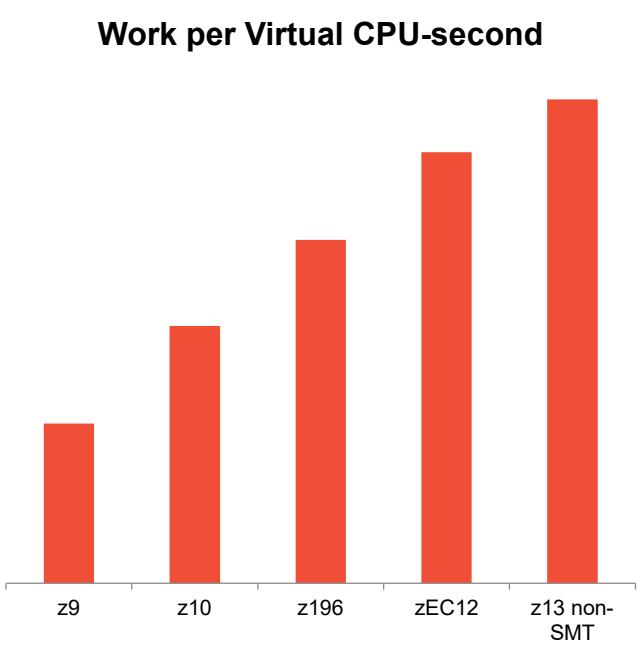
- SMT usage, i.e. setting number of threads per core, is controlled by enabling/disabling SMT
  – You can be enabled for SMT, and still run with 1 thread per core
- System Configuration File MULTITHREADING statement

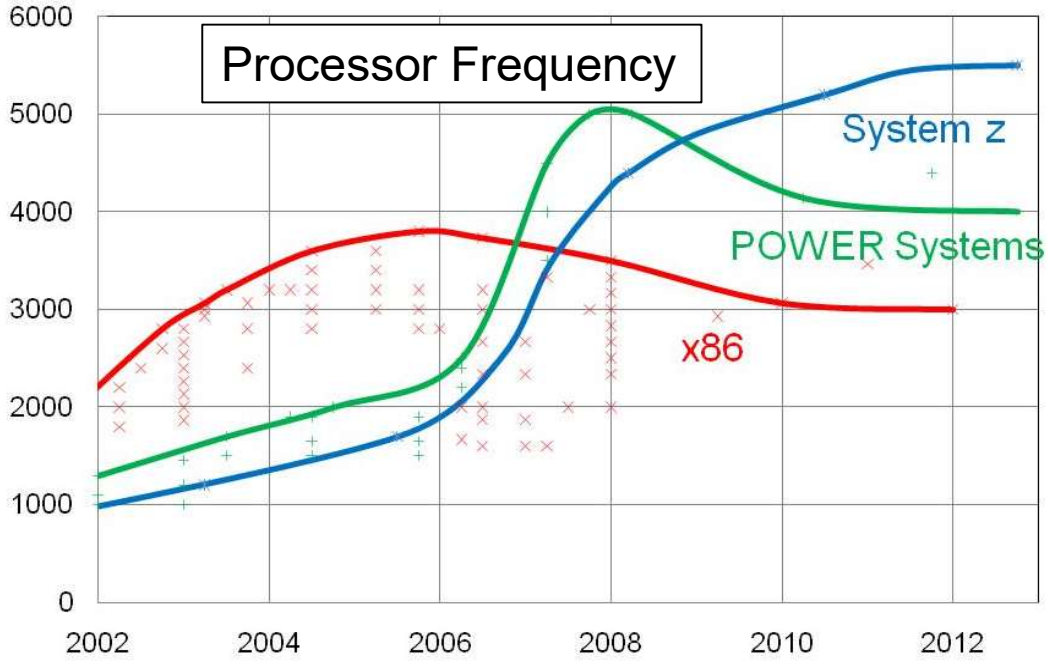
	MULTITHREADING Statement	SMT	Threads per Core
SMT-0	DISABLE	Disabled	n/a
SMT-1	ENABLE TYPE ALL 1	Enabled	1
SMT-2	ENABLE TYPE ALL 2	Enabled	2

- If SMT is “Enabled” then can change the number of threads per core vis CP command SET MT TYPE ALL *n*

Why Simultaneous Multithreading?

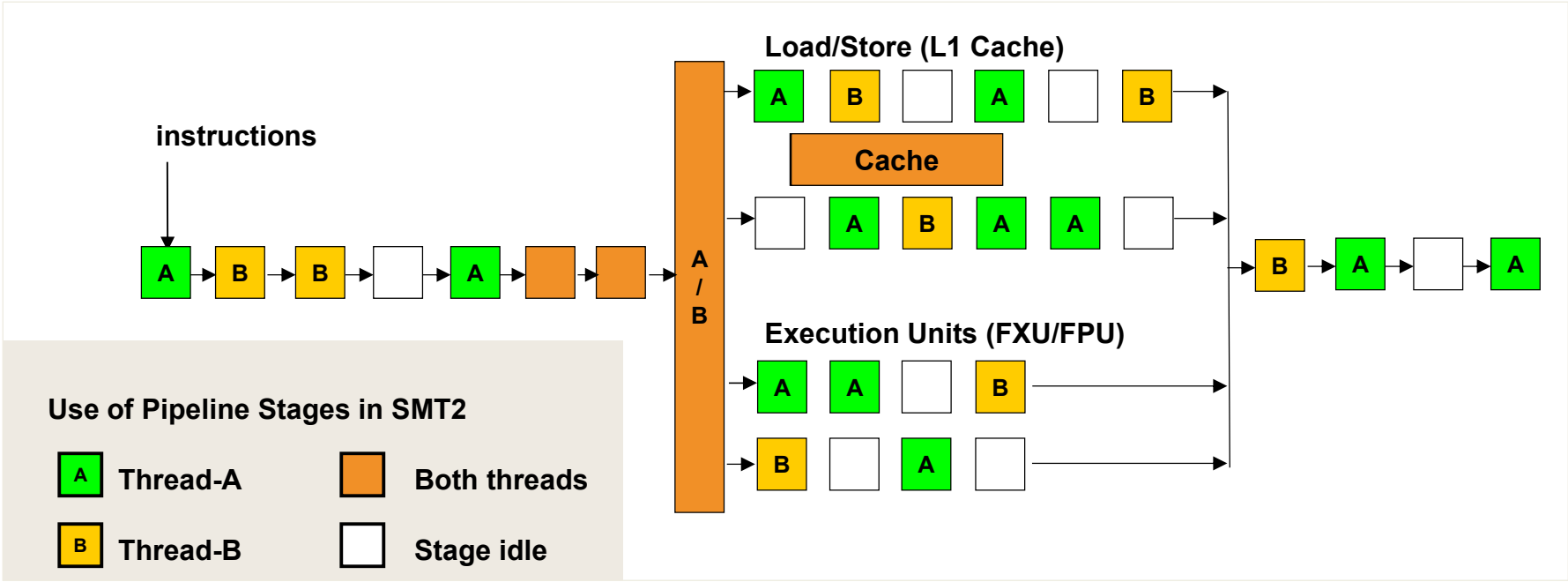
- Other architectures are already doing it.
- We're reaching the physical limits of the machine; we can't just keep making chips smaller and faster.
- We need now to look at ways to use the chip resources more efficiently.





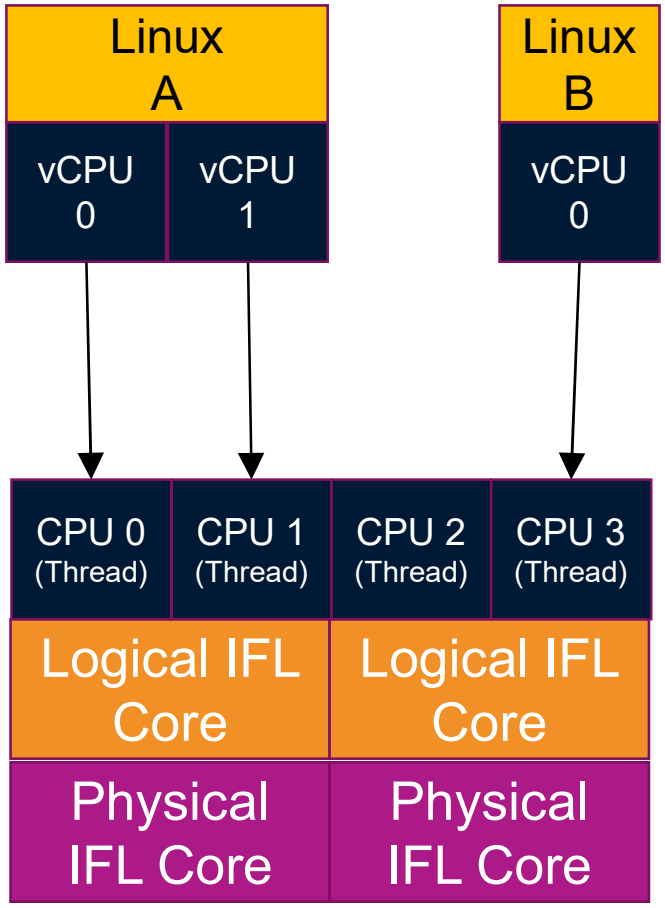
SMT Technology

- Multiple programs (software threads) run on same processor core
- Active threads share core resources
  - In space: e.g, data and instruction caches, TLBs, branch history tables
  - In time: e.g., pipeline slots, execution units, address translator
- Increases overall throughput per core when SMT active
  - Amount of increase varies widely with workload
  - Each thread runs more slowly than a single-thread core



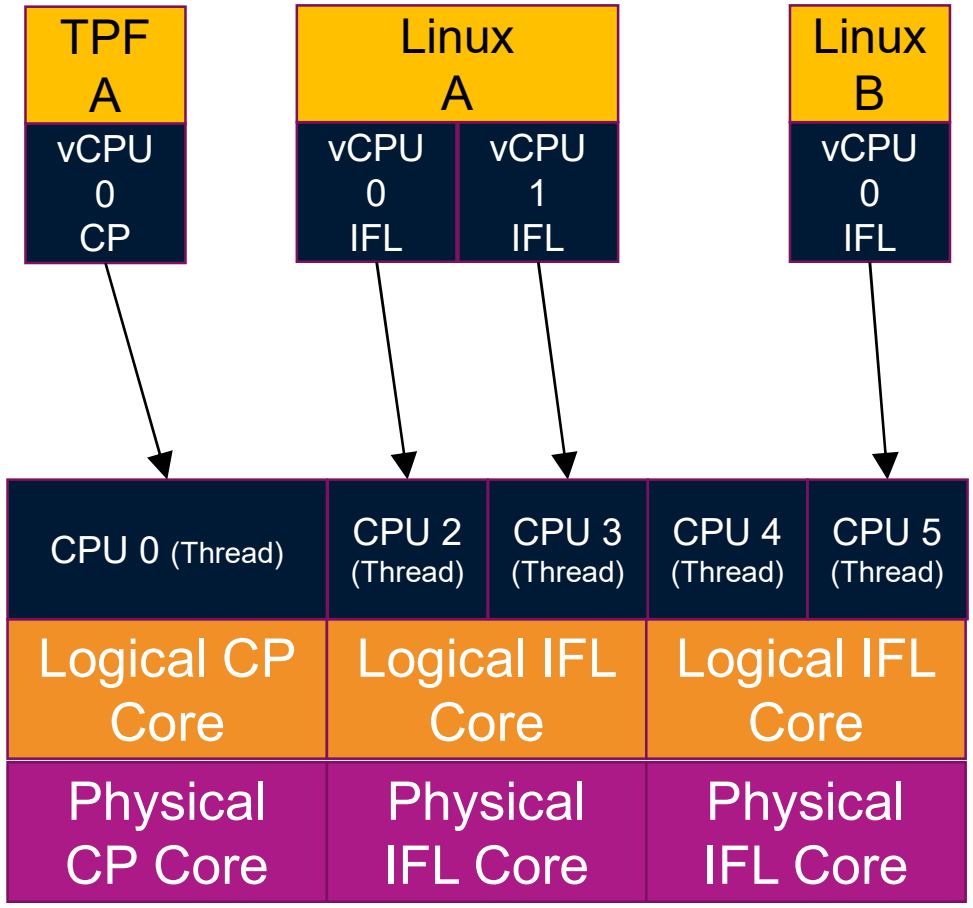
SMT in z/VM

- **Physical IFL Cores (you purchase these) with SMT allow up to two threads to be used**
- **Logical IFL Cores are presented to z/VM as in the past (you define these in the logical partition profile on the HMC)**
- **z/VM creates a CPU or logical processor associated with each thread (reflected in commands like QUERY PROCESSORS)**
- **The virtual CPUs of guests can then be dispatched on different threads intelligently, based on topology information, sometimes referenced as virtual IFL**



SMT in z/VM – Mixed Engine Environment

- **In a mixed-engine environment, general purpose processors are not enabled for threading, but a second CPU address is consumed (CPU 1 in this example) when SMT is enabled for the partition**
- **Virtual IFL CPUs are dispatched on the logical IFLs and virtual CP CPUs are dispatched on the logical CPs**



SMT in z/OS

- **SMT exploitation for zIIPs**

- **IEAOPTxx PROCVIEW CORE enables SMT for life of IPL**
  - **Operator can set MT_ZIIP_MODE=1|2 to change number of active threads per online core dynamically**
    - **E.g., MT_ZIIP_MODE=2 for daytime OLTP, MT_ZIIP_MODE=1 for overnight batch**

- **Same terminology, metrics, considerations, and issues for z/OS and z/VM**

Determining Application Performance

Throughput

- Given that SMT's objective is to increase system throughput, it makes sense to measure it
- Applications can span virtual machines, z/VM systems, and even include components outside of IBM Z or LinuxONE
- An accurate view of throughput is available only from the application itself or externally
- Some poor alternatives from z/VM data can be used
  - Virtual I/O rate: assumes for a given transaction the virtual I/O will remain constant (could be true)
  - Network Input or Output rate: for workloads driven by network requests
- These measures can be for particular virtual machines or groups, depending on the structure of the application

Response Time

- Multiple threads that are slower than a single thread may potentially degrade response time
- An accurate view of response time is available only from the application itself or externally
- Unlike throughput, there are no alternatives from z/VM data that can map to response time, making it more important to have application or external data for this analysis
- z/VM State Sampling information can be helpful from a different perspective
  - z/VM does virtual CPU state sampling by default every two seconds
  - Captures ‘state’ of virtual CPU (e.g., Running, CPU wait, Page wait)
  - Delays due to waiting for (as opposed to using) a resource can be a major influence on response time
  - All systems wait at the same speed (wall clock)

Variability Factors

Work and Processor Time

- Old Myth: “Virtual CPU time should be constant for a given workload”
  - True on much older machines where
    - Resources were not shared with other instruction streams
    - More sophisticated optimizations (e.g., pipelining) were not used
  - True when competition or demand for shared resources was constant
  - True when fewer instruction streams were competing for those resources
  - True when memory was closer and uniform, in relative terms

- The variability has been there for a long time; it is just more noticeable today

- Key to recognize
  - Performance of a system, or a single part of a system, needs to be measured in that system to get an accurate picture
  - Some things vary that you can control
    - E.g., number of guests, number of logical partitions

z/VM implemented the following for SMT enabled LPARs

- Thread Affinity – An effort is made to run the virtual CPU on the same thread as long as the virtual CPU remains on the core's dispatch vector
  - Reduces L1, L2, and TLB penalties

- Preemption Disabled – To give the current CPU more time on the thread
  - Reduces L1, L2, and TLB penalties

- Minor Time Slice Increased – Allow the virtual processor to benefit from build up of cache L1, L2 and TLB

- Time Slice Early – If the virtual CPU loads a wait PSW, and certain conditions are true, CP ends the virtual CPU minor time slice early (helps assure the virtual CPU is not holding a guest spin-lock at the end of time slice)

More details: http://www.vm.ibm.com/perf/reports/zvm/html/1q5smt.html

Metrics

Need to Rethink Metrics

CAPACITY ≠ UTILIZATION ≠ PRODUCTIVITY

Processor Time Reporting

- **Raw time** (the old way, but with new implications)
  - Amount of time each virtual CPU is run on a thread
  - This is the only kind of time measurement available when SMT is disabled
  - Used to compute dispatcher time slice and scheduler priority
- **MT-1 equivalent time** (new)
  - Used when SMT is enabled
  - Approximates what the raw time would have been if the virtual CPU had run on the core all by itself
    - Adjusted downward (decreased) from raw time
  - Intended to be used for chargeback
- **Pro-rated core time**
  - Used when SMT is enabled
  - "Discounts" raw time proportionally when core is shared between active threads
    - Full time charged while a virtual CPU runs alongside an idle thread
    - Half time charged while vCPU is dispatched beside another active thread
  - Suitable for core-based software license metrics

A Word About Metrics

- Estimate vs. Measure
  - Some metrics are actually measured by z/VM or firmware, others are estimated by either z/VM or the firmware

- Full wall clock vs. Core dispatched interval
  - Some metrics cover all of time, i.e. Wall Clock
  - Other metrics are based on the span of time in which z/VM's logical core was dispatched

Metrics: Core Busy and Thread Density

Two Threads on a Core

- Core Busy = 4 ÷ 5 = 80%

Measure
Over wall clock time

- Thread Density = Average(1,1,2,1) = 1.25

Measure
Over core-busy time

Metrics: Productivity

Estimate
Over core-busy time

Two Threads on a Core

Numbers indicate “work” completed metric (illustrative purposes here)

Compared to extrapolation of keeping both threads busy whenever core is in use

- Productivity = Ratio of actual “work” completed compared to estimated “work” that could be completed if threads kept busy
- Productivity = (9+9+7+7+9) ÷ (4 x (7+7)) = 41 ÷ 56 = 73%

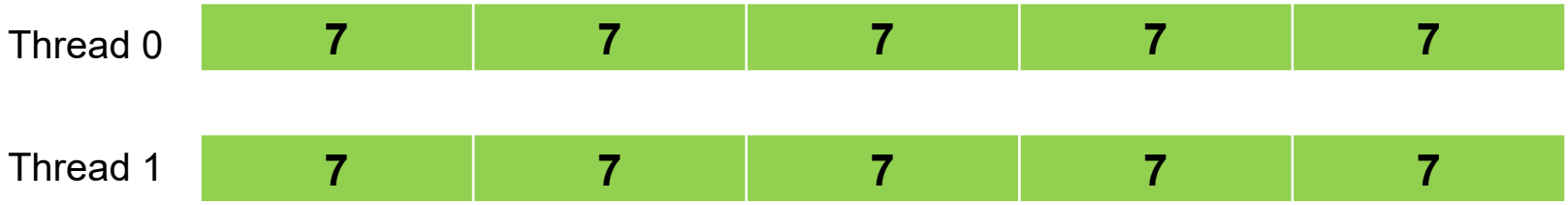
Metrics: MT Utilization

Estimate
Over wall clock time

Two Threads on a Core

Numbers indicate “work” completed metric (illustrative purposes here)

Extrapolated to behavior with two threads always busy



- MT Utilization = Ratio of actual “work” completed to “work” estimated that could have completed if the core was 100% busy with thread density 2. A view of how close the workload is to saturating the core.
- MT Utilization = (9+9+7+7+9) ÷ (5 x (7+7)) = 41 ÷ 70 = 59%
  - Observation: MT Utilization ~= Core Productivity * Core Busy

Metrics: Capacity Factor & Max Capacity Factor

Measure
Over core-busy time

Two Threads on a Core

- Capacity Factor = TotalWorkRate / SingleThreadWorkRate =
  - (41÷4) ÷ (27÷3) = 10.25 ÷ 9 = 1.13 = 113%

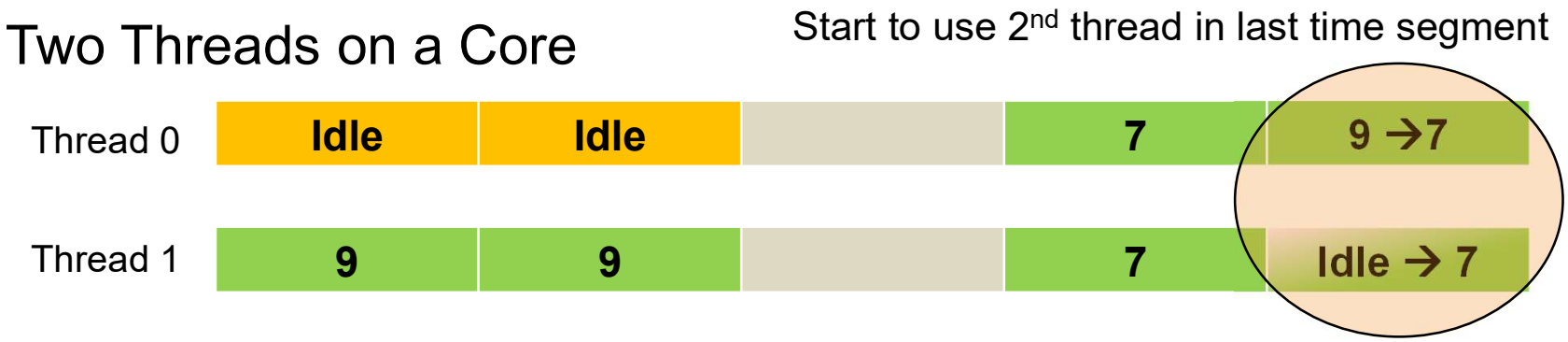
- Max Capacity Factor: Projected upper bound
  - TwoThreadWorkRate / SingleThreadWorkRate =
  - (14÷1) / (27÷3) = 14 ÷ 9 = 1.56 = 156%

Measures: Putting It All Together

Two Threads on a Core

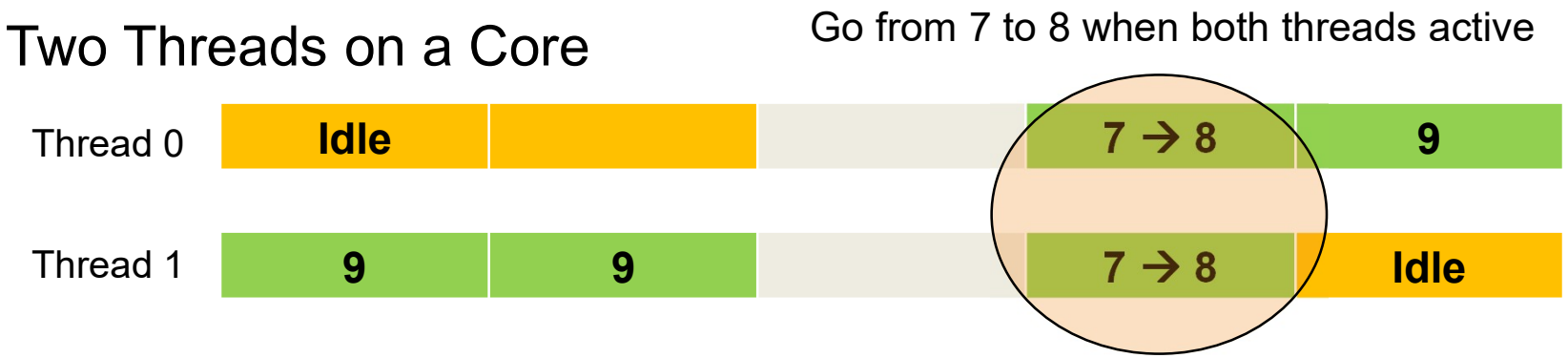
Metrics	Example
Core Busy	80%
Thread Density	1.25
Productivity	73%
MT Utilization	59%
Capacity Factor	113%
Max Capacity Factor	156%

Relationship: As Thread Density increases



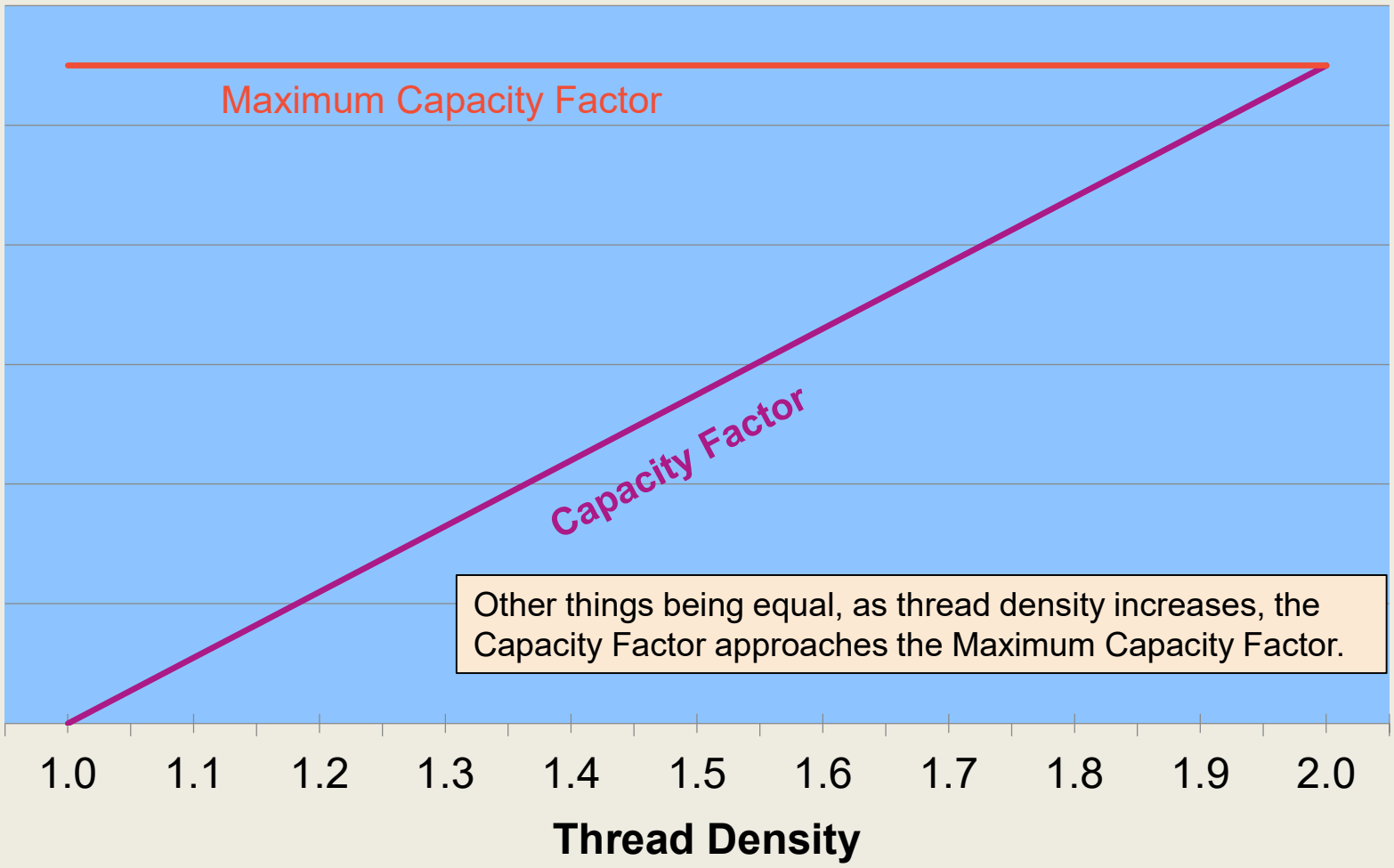
Metrics	Example	w/ Increased TD	
Core Busy	80%	80%	No effect
Thread Density	1.25	1.5	1.25 → 1.5
Productivity	73%	82%	Increases
MT Utilization	59%	66%	Increases
Capacity Factor	113%	127%	Increases
Max Capacity Factor	156%	156%	No effect

Relationship: More efficient threading

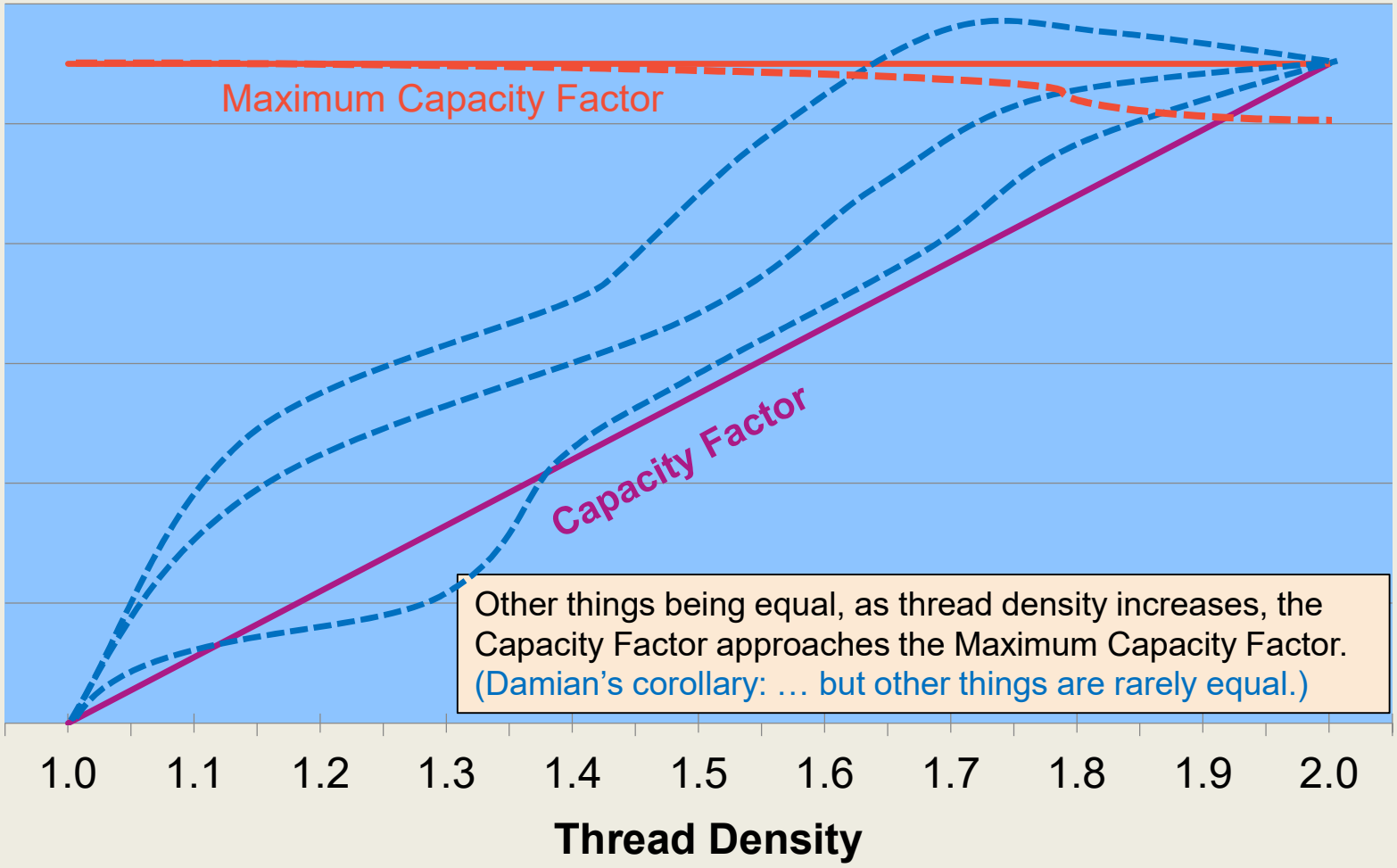


Metrics	Example	w/ more Efficient Threading	
Core Busy	80%	80%	No effect
Thread Density	1.25	1.25	No effect
Productivity	73%	67%	Decreases
MT Utilization	59%	54%	Decreases
Capacity Factor	113%	119%	Increases
Max Capacity Factor	156%	178%	Increases

Thread Density Effect on Capacity Factor



Thread Density Effect on Capacity Factor



New Metrics on INDICATE command & monitor

- **Indicate Load** will still show information by processor, which means by individual thread on multithreaded cores.
  - The percent-busy is thread-busy aka logical CPU-busy

- A new command, **INDicate MULTITHread (MT)** will show you the per type information, giving you an idea of how much capacity you have left for each type. The utilization shown is an average of the utilization of the cores of that type.

```
indicate multith
Multithreading is enabled.
Statistics from the interval 12:00:53 - 12:01:23
Core Type CP    Busy    8%    TD  1.00 of  1   Prod 100%   Util   8%
  CF   100%    MaxCF  100%
Core Type IFL   Busy    1%    TD  1.50 of  2   Prod  90%   Util   1%
  CF   113%    MaxCF  125%
Core Type ZIIP Busy    0%    TD  1.00 of  1   Prod 100%   Util   0%
  CF   100%    MaxCF  100%
Ready;
```

SMTMET Display Tool

- A CMS EXEC that extracts and prints MT metrics from MonDomain 0 Record 2.
- Available on z/VM download library: http://www.vm.ibm.com/download/packages/
- SMTMET Documentation: http://www.vm.ibm.com/perf/tips/smtmet.html

The process for reducing the SMTMET counters is the following:

1. Start with a MONWRITE file that contains D0 R2 records.

2. Command Syntax from CMS prompt:

   `SMTMET` *`filename`* `MONDATA` *`filemode`*

3. Resultant file from CMS prompt:

   filename $SMTMET *filemode*

SMTMET Output File Sample: Per-Core-Type Report

D0R2 Per-Core-type Report for file: AMPDGLD1 MONDATA

Interval Ended	Core Type	Secs	Sampled Cores	Pct Core Prodctvity	Pct Cap Factor	Pct Max Cap Fct	Pct MT Utilztion	Average Thread Den
>>Mean>>	IFL	120.0	4.0	93.6	156.4	167.1	86.0	1.83
21:32:02	IFL	120.0	4.0	93.6	159.2	170.1	74.5	1.84
21:32:32	IFL	120.0	4.0	93.6	158.7	169.6	89.7	1.84
21:33:02	IFL	120.0	4.0	93.2	157.7	169.2	89.3	1.83
21:33:32	IFL	119.6	4.0	93.4	158.9	170.1	89.3	1.84
21:34:02	IFL	120.0	4.0	93.3	159.2	170.5	89.4	1.84
21:34:32	IFL	120.0	4.0	93.5	158.9	169.9	89.8	1.84
21:35:02	IFL	120.0	4.0	94.1	161.1	171.1	91.2	1.86
21:35:32	IFL	120.0	4.0	93.4	159.0	170.1	89.8	1.84
21:36:02	IFL	120.0	4.0	93.8	159.5	170.0	90.5	1.85
21:36:32	IFL	120.0	4.0	93.0	158.5	170.4	88.7	1.83
21:37:02	IFL	120.0	4.0	93.4	159.1	170.3	89.7	1.84
21:37:32	IFL	120.0	4.0	93.7	159.4	170.1	90.2	1.85
21:38:02	IFL	120.0	4.0	93.7	159.0	169.6	90.4	1.85

SMTMET Output File Sample: Per-Core Report

D0R2 Per-Core Report for file: AMPDGLD1 MONDATA

Interval Ended	Core ID	Core Type	Secs	Pct Core Prodctvity	Pct MT Utilztion	Average Thread Den	Pct Core Busy
>>Mean>>	00	IFL	30.0	93.6	86.0	1.83	92.06
>>Mean>>	01	IFL	30.0	93.5	86.0	1.83	91.92
>>Mean>>	02	IFL	30.0	93.7	86.3	1.83	92.20
>>Mean>>	03	IFL	30.0	93.6	85.9	1.84	91.86
21:32:02	00	IFL	30.0	93.4	74.0	1.84	79.26
21:32:02	01	IFL	30.0	93.1	74.4	1.83	79.91
21:32:02	02	IFL	30.0	93.8	74.2	1.85	79.11
21:32:02	03	IFL	30.0	93.8	75.4	1.85	80.39
21:32:32	00	IFL	30.0	94.1	91.2	1.86	96.86
21:32:32	01	IFL	30.0	93.3	88.8	1.84	95.16
21:32:32	02	IFL	30.0	92.7	88.3	1.82	95.28
21:32:32	03	IFL	30.0	94.0	90.7	1.86	96.42

SMTMET Per-Core Report with Extremes

D0R2 Per-Core Report for file: IDLESYS MONDATA

Interval __Ended_	Core _ID_	Core Type	___Secs___	Pct Core Prodctvity	Pct MT Utilztion_	Average Thread Den	Pct Core ___Busy___
>>Mean>>	00	IFL	30.0	71.7	0.1	1.20	0.05
>>Mean>>	01	IFL	30.0	78.8	0.1	1.39	0.11
>>Mean>>	02	IFL	30.0	0.0	0.0	1.38	0.01
>>Mean>>	03	IFL	30.0	0.0	0.0	1.40	0.01
13:52:42	00	IFL	30.0	.....	.....	1.24	0.02
13:52:42	01	IFL	30.0	80.8	0.1	1.55	0.12
13:52:42	02	IFL	29.9	.....	.....	1.40	0.01
13:52:42	03	IFL	30.0	.....	.....	1.40	0.01
13:53:12	00	IFL	30.0	.....	.....	1.25	0.02
13:53:12	01	IFL	30.0	80.3	0.1	1.55	0.12
13:53:12	02	IFL	30.1	.....	.....	1.39	0.01
13:53:12	03	IFL	30.0	.....	.....	1.40	0.01

- **At very high and very low core utilizations, z/VM may not be able to calculate Productivity and Utilization values**
- **Represented as '.....' in report and not included in the >>Mean>> calculation**

SMTMET Per-Core-Type Report with Low MT Utilization

D0R2 Per-Core-type Report for file: IDLESYS MONDATA

Interval __Ended_	Core Type	Sampled ___Secs___	Sampled ___Cores__	Pct Core Prodctvity	Pct Cap __Factor__	Pct Max _Cap Fct__	Pct MT Utilztion_	Average Thread Den
>>Mean>>	IFL	90.0	3.0	76.7	124.1	172.1	0.0	1.36
13:51:42	IFL	119.8	4.0	80.6	137.3	175.7	0.0	1.49
13:52:12	IFL	120.1	4.0	80.9	136.1	173.4	0.0	1.49
13:52:42	IFL	120.0	4.0	80.8	135.8	173.7	0.0	1.49
13:53:12	IFL	120.1	4.0	80.3	137.2	177.3	0.0	1.50
13:53:42	IFL	120.0	4.0	82.1	132.1	164.1	0.0	1.48
13:54:12	IFL	60.0	2.0	81.2	134.0	171.9	0.1	1.49
13:54:42	IFL	60.0	2.0	70.0	106.5	167.1	0.1	1.18
13:55:12	IFL	60.0	2.0	59.2	103.2	200.0	0.1	1.12
13:55:42	IFL	60.0	2.0	86.0	101.0	118.2	0.1	1.09
13:56:12	IFL	60.0	2.0	66.0	118.0	200.0	0.1	1.26

When looking at this data, first look at MT Utilization; the cores are practically idle.

CPUMF Display Tool

- A CMS exec that extracts CPU MF counters and other data from a MONWRITE file and then calculates CPU performance metrics such as CPI
- Available on z/VM download library: http://www.vm.ibm.com/download/packages/
- How to collect the CPU MF counters: http://www.vm.ibm.com/perf/tips/cpumfhow.html
- How to interpret the CPU MF report: http://www.vm.ibm.com/perf/tips/cpumf.html

The process for reducing the CPU MF counters is the following:

1. Start with a MONWRITE file containing D5 R13 and other records
2. Command: `CPUMFINT` *`filename`* `MONDATA` *`filemode`*
   (this produces a file called *`filename`* `CPUMFINT` *`filemode`*)
3. Command: `CPUMFLOG` *`filename`* `CPUMFINT` *`filemode`*
   (this produces a file called *`filename`* `$CPUMFLG` *`filemode`*)
4. Your report is in the $CPUMFLG file.

Sample $CPUMFLG Output

IntEnd	LPU	Typ	___L1MP___	___L2P____	___L3P____	___L4LP__
>>Mean>>	0	IFL	2.05	87.22	12.71	0.0
>>Mean>>	1	IFL	2.01	87.27	12.66	0.0
>>Mean>>	2	IFL	2.02	87.13	12.80	0.0
>>Mean>>	3	IFL	2.04	87.06	12.86	0.0
>>Mean>>	4	IFL	2.01	87.25	12.68	0.0
>>Mean>>	5	IFL	2.01	87.21	12.72	0.0
>>MofM>>			2.02	87.19	12.74	0.0
>>AllP>>						
00:46:02	0	IFL	1.99	87.00	12.93	0.0
00:46:02	1	IFL	1.99	87.04	12.91	0.0
00:46:02	2	IFL	1.96	87.01	12.93	0.0
00:46:02	3	IFL	1.96	86.93	13.01	0.0
00:46:02	4	IFL	1.97	86.95	12.98	0.0
00:46:02	5	IFL	1.99	86.96	12.96	0.0

Memory Footprint within the Cache

- 2% of the instructions incur an L1 cache miss. (L1MP)
- 87% of the L1 misses are sourced from the L2 cache (L2P)
- 13% of the L1 misses are sourced from the L3 cache (L3P)

z/OS RMF CPU Activity Report (MT=2)

```
--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 x 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%**

z/OS RMF Workload Activity Report (MT=2)

```
                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 in MT=1 Equivalent Time units
- APPL % is % of core relative to its maximum Capacity Factor
- IIP APPL % = 1995.64 x 100 ÷ (614 x 1.473) = 220.69 %
- CP APPL % = (3545.743 - 1995.640) x 100 ÷ (614 * 1.0) = 252.44 %

Performance Measurement Results

SMT-2 Ideal Application

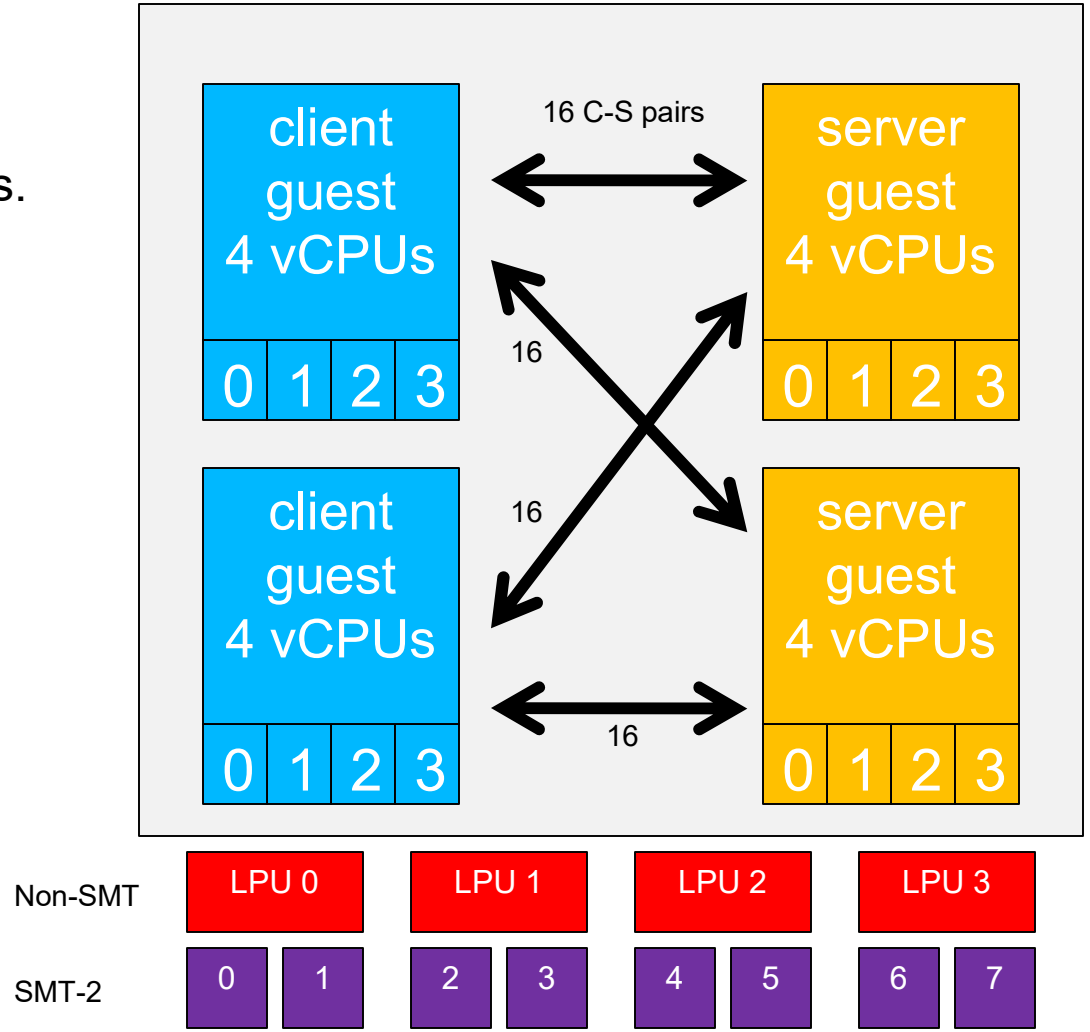
We designed this workload to show SMT-2 benefit:

1. No relationship between the 64 Client-Server pairs.
2. Virtual >>Logical IFLs
3. Almost no Control Program activity

This workload is limited by dispatch parallelism.

If we can increase dispatch parallelism, we hope we will also increase throughput.

This assumes L1, etc. behavior doesn't hurt us.



SMT Ideal Application Results

Multithreading	Disabled	Enabled
SMT Level	SMT-0	SMT-2
Logical Cores	4	4
Logical Processors	4	8
External Throughput Ratio	1.000	**1.362**
Internal Throughput Ratio	1.000	1.539
Response Time Ratio	1.000	**0.743**
Processor Utilization	95.6	84.6
Avg %CPU Wait	40%	32%
SMT Core Busy %	95.6	95.5
SMT Avg Thread Density	na	1.83
Capacity Factor	na	156.4%

Enabling SMT allowed us to push more work through the cores without increasing the number of cores.

The Processor Log Screen for the SMT-0 run:

```
FCX304  Run 2015/03/04 15:15:30          PRCLOG
                                         Processor Activity, by Time
From 2015/02/14 16:04:29
To   2015/02/14 16:14:59
For    630 Secs 00:10:30                 "This is a performance report f
```

Interval End Time	CP U	Type	PPD	Ent.	DVID	Pct Park Time	<--- Percent Busy ----> Total	User	Syst	Emul	<-- Rates Inst Siml
>>Mean>>	0	IFL	VhD	100	0000	0	95.7	95.5	.2	88.2	38153
>>Mean>>	1	IFL	VhD	100	0001	0	95.7	95.5	.2	88.2	37536
>>Mean>>	2	IFL	VhD	100	0002	0	95.6	95.4	.2	88.0	38178
>>Mean>>	3	IFL	VhD	100	0003	0	95.5	95.3	.2	87.8	38532
>>Total>	4	IFL	VhD	400	MIX	0	382.5	381.6	.9	352.1	152k

The Processor Log Screen for the SMT-2 run:

```
FCX304  Run 2015/03/04 15:16:28          PRCLOG
                                         Processor Activity, by Time
From 2015/02/14 16:31:32
To   2015/02/14 16:42:02
For    630 Secs 00:10:30                 "This is a performance repo
```

							<--- Percent Busy ---->			
Interval End Time	CP U	Type	PPD	Ent.	DVID	Pct Park Time	Total	User	Syst	Emul
>>Mean>>	0	IFL	VhD	100	0000	0	84.7	84.5	.2	77.0
>>Mean>>	1	IFL	VhD	100	0000	0	84.3	84.1	.2	76.8
>>Mean>>	2	IFL	VhD	100	0001	0	84.5	84.4	.2	76.8
>>Mean>>	3	IFL	VhD	100	0001	0	84.6	84.4	.2	77.0
>>Mean>>	4	IFL	VhD	100	0002	0	84.5	84.3	.2	77.0
>>Mean>>	5	IFL	VhD	100	0002	0	84.9	84.7	.2	77.5
>>Mean>>	6	IFL	VhD	100	0003	0	84.8	84.6	.2	77.3
>>Mean>>	7	IFL	VhD	100	0003	0	84.7	84.5	.2	77.3
>>Total>	8	IFL	VhD	800	MIX	0	677.0	675.5	1.5	616.6

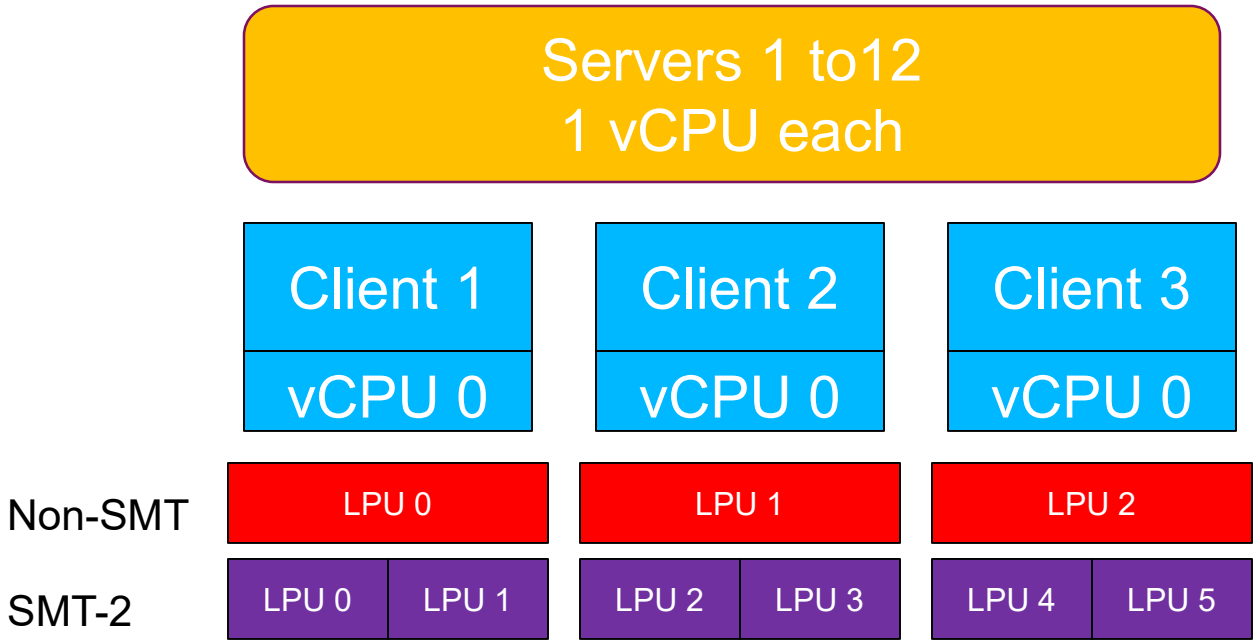
$SMTMET Resultant File

D0R2 Per-Core Report for file: AMPDGLD1 MONDATA

Interval __Ended_	Core _ID_	Core Type	___Secs___	Pct Core Prodctvity	Pct MT Utilztion_	Average Thread Den	Pct Core ___Busy___
>>Mean>>	00	IFL	30.0	93.6	86.0	1.83	92.06
>>Mean>>	01	IFL	30.0	93.5	86.0	1.83	91.92
>>Mean>>	02	IFL	30.0	93.7	86.3	1.83	92.20
>>Mean>>	03	IFL	30.0	93.6	85.9	1.84	91.86
21:32:02	00	IFL	30.0	93.4	74.0	1.84	79.26
21:32:02	01	IFL	30.0	93.1	74.4	1.83	79.91
21:32:02	02	IFL	30.0	93.8	74.2	1.85	79.11
21:32:02	03	IFL	30.0	93.8	75.4	1.85	80.39
21:32:32	00	IFL	30.0	94.1	91.2	1.86	96.86
21:32:32	01	IFL	30.0	93.3	88.8	1.84	95.16
21:32:32	02	IFL	30.0	92.7	88.3	1.82	95.28
21:32:32	03	IFL	30.0	94.0	90.7	1.86	96.42

SMT Single Processor Serialization Application

- Similar to the “Ideal” workload in terms of using clients and servers
- Different in that there are just 3 client virtual machines each with 1 virtual processor
- This configuration uses 3 IFL cores; total of 15 virtual CPUs

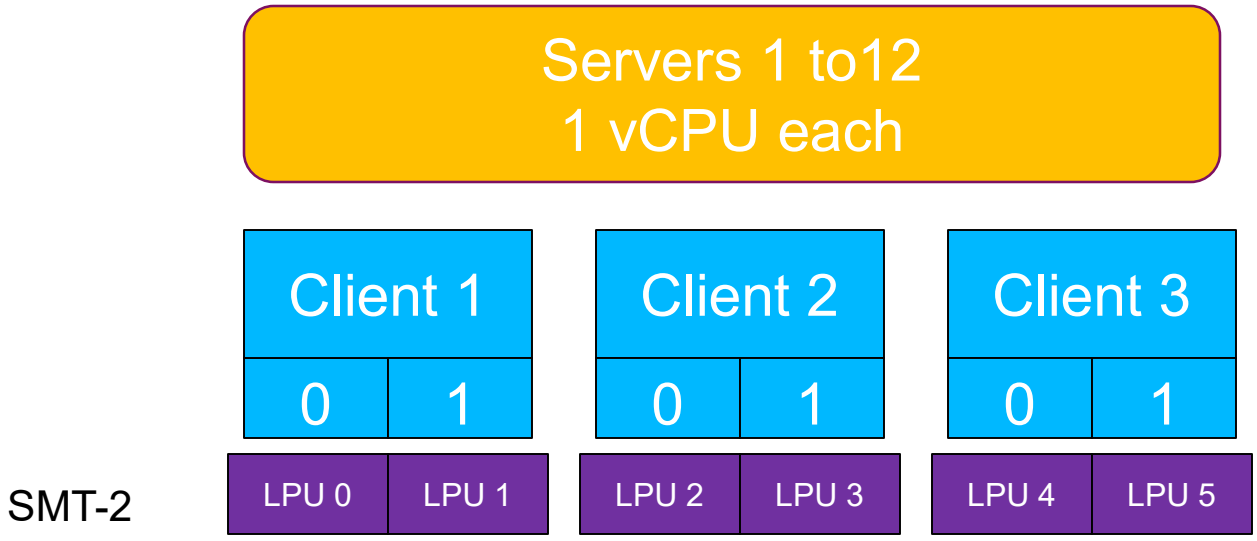


SMT Single Processor Serialization Application

Multithreading	Disabled	Enabled
SMT	SMT-0	SMT-2
Logical Cores	3	3
Logical Processors	3	6
Client virtual machines	3	3
Virtual processors per client	1	1
External Throughput Ratio	1.000	**0.649**
Internal Throughput Ratio	1.000	1.018
Response Time Ratio	1.000	**1.450**
Processor Utilization	95.8%	61.1%
Client utilization	70.5%	95.4%
SMT Core Busy %	95.8%	95.5%
SMT Avg Thread Density	na	1.28
Capacity Factor	na	104%
SMT Max. Capacity Factor	na	115%

Mitigation – Add virtual CPUs

- With only a total of 3 virtual IFLs from the client machines, all 6 logical processors can not be utilized at same time. And with two threads being slower than a dedicated core, we lost performance.
- One approach to mitigate, would be to add virtual IFLs to the client virtual machines.

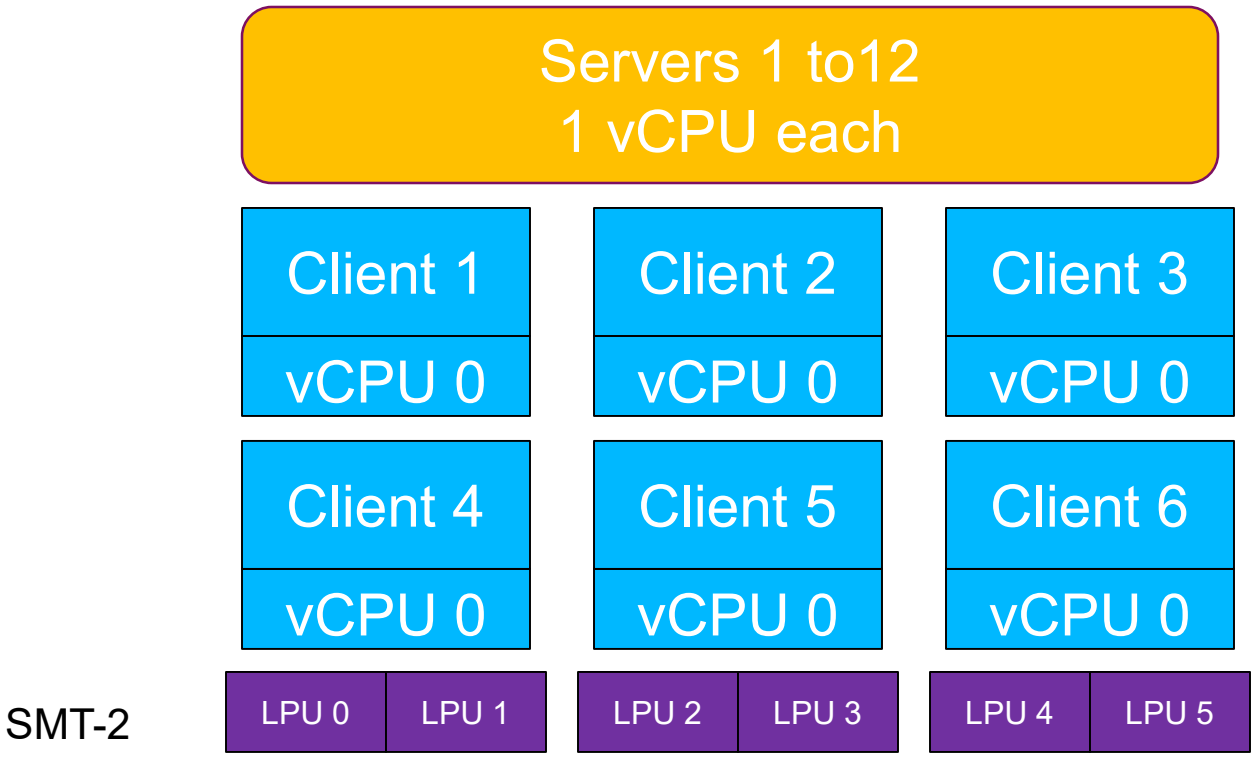


Mitigation – Add virtual CPUs

Multithreading	Disabled	Enabled	Enabled
SMT	SMT-0	SMT-2	SMT-2
Logical Cores	3	3	3
Logical Processors	3	6	6
Client virtual machines	3	3	3
Virtual processors per client	1	1	2
External Throughput Ratio	1.000	**0.649**	**1.065**
Internal Throughput Ratio	1.000	1.018	1.101
Response Time Ratio	1.000	**1.450**	**0.896**
Processor Utilization	95.8%	61.1%	92.6%
Client utilization	70.5%	95.4%	143.8%
SMT Core Busy %	95.8%	95.5%	95.1%
SMT Avg Thread Density	na	1.28	1.95
Capacity Factor	na	104%	149%
SMT Max. Capacity Factor	na	115%	151%

Mitigation – Replicate Clients

- Having 6 virtual IFLs for the clients could also be achieved for this workload by doubling the number of client virtual machines and returning them to a single virtual IFL



Mitigation – Double the Clients

Multithreading	Disabled	Enabled	Enabled	Enabled
SMT	SMT-0	SMT-2	SMT-2	SMT-2
Logical Cores	3	3	3	3
Logical Processors	3	6	6	6
Client virtual machines	3	3	3	6
Virtual processors per client	1	1	2	1
External Throughput Ratio	1.000	**0.649**	**1.065**	**1.303**
Internal Throughput Ratio	1.000	1.018	1.101	1.306
Response Time Ratio	1.000	**1.450**	**0.896**	**1.701**
Processor Utilization	95.8%	61.1%	92.6%	95.6%
Client utilization	70.5%	95.4%	143.8%	73.7%
SMT Core Busy %	95.8%	95.5%	95.1%	95.7%
SMT Avg Thread Density	na	1.28	1.95	1.99
Capacity Factor	na	104%	149%	na
SMT Max. Capacity Factor	na	115%	151%	na

Performance Measurements

- **SMT2 Ideal Application**
- Maximum Storage Configuration
- Maximum Logical Processor Configuration
- **Linux-only mode with Single Processor serialization Application**
  - **Mitigation 1: Increasing virtual processors**
  - **Mitigation 2: Increasing servers in workload**
- Linux-only mode with Master Processor Serialization Application
- z/VM-mode with Master Processor Serialization Application
- CPU Pooling Workload
- **Live Guest Relocation (LGR) Workload**
- For a more details about performance results see: http://www.vm.ibm.com/perf/reports/zvm/html/1q5smt.html

Performance Measurements: Live Guest Relocation

25 Linux guests relocated while running three workloads

- PING – to simulate network traffic
- BLAST– to simulate I/O
- PFAULT- to simulate referencing storage

Relocation was done synchronously using the SYNC option of VMRELOCATE command

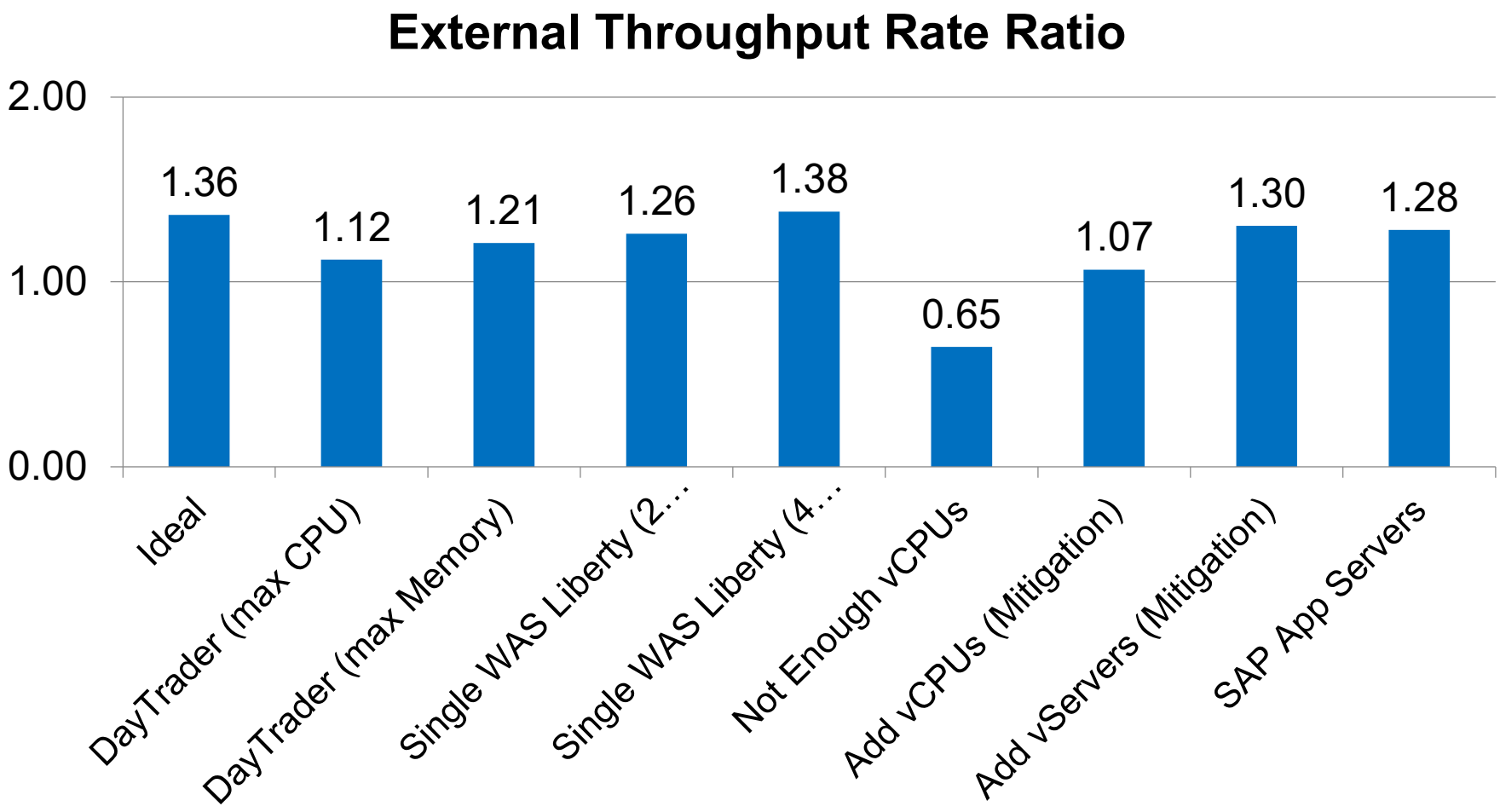
Results:

- Relocation time increased by 10%
- Quiesce time increased by 26%
- PFAULT (71%) and BLAST (34%) completions increased
- Total number of pages relocated during quiesce increased by 51%

Conclusion:

With SMT2, the BLAST and PFAULT workloads were changing pages more frequently, thus causing more pages to be moved during quiesce time.

Variation in Impact of SMT



Base measurement is with SMT disabled. All measurements on z13.

Performance Measurements: Conclusion

- Results in measured workloads **varied widely**.
- Best results were observed for applications having highly parallel activity and no single point of serialization.
- No improvements were observed for applications having a single point of serialization.
- To overcome bottlenecks, workload adjustment should be done where possible.
- While very rare, workloads that have a heavy dependency on the z/VM master processor are not good candidates for SMT-2.
- The multithreading metrics (provided by the SMTMET tool) provide information about how well the cores perform when SMT is enabled. There is **no direct relationship with workload performance** (ETR, transaction response time)
- **Measuring workload throughput and response time is the best way to know whether SMT is providing value to the workload.**

Summary

Summary

- SMT provides potential throughput improvements
  - Especially if workloads experiencing CPU Wait or cache misses

- Before moving to SMT
  - Gather throughput and response time data
  - Gather monitor and CPU MF data
  - Evaluate and adjust virtual machines (i.e., add virtual CPUs) that are approaching virtual CPU utilization limits (e.g., 85%) at peak times

- After moving to SMT
  - Compare throughput and response time data
  - Validate virtual CPU configurations
  - Evaluate the new SMT metrics
  - Assess both core resources and thread (logical processor) resources

Backup: Performance Toolkit Changes

Monitor Changes

New Monitor	Record Name
Domain 5 Record 20	MT CPUMF counters

Change Monitor	Records Name
Domain 0 Record 2	Processor data (per processor)
Domain 0 Record 15	Logical CPU utilization (global)
Domain 0 Record 16	CPU utilization in a logical partition)
Domain 0 Record 17	Physical CPU utilization data for LPAR management
Domain 0 Record 19	System data (global)
Domain 0 Record 23	Formal spin lock data (global)
Domain 1 Record 4	System configuration data
Domain 1 Record 5	Processor configuration data (per processor)
Domain 1 Record 16	Scheduler settings
Domain 1 Record 18	CPU capability change
Domain 2 Record 4	Add user to dispatch list
Domain 2 Record 5	Drop user from dispatch list
Domain 2 Record 7	Set SRM changes
Domain 2 Record 13	Add VMDBK to limit list
Domain 2 Record 14	Drop VMDBK from limit list
Domain 4 Record 2	User logoff data
Domain 4 Record 3	User activity data
Domain 4 Record 9	User activity data at transaction end
Domain 5 Record 1	Vary on processor
Domain 5 Record 2	Vary off processor
Domain 5 Record 11	Instruction counts per processor
Domain 5 Record 13	CPU-measurement facility counters
Domain 5 Record 16	Park/unpark decision
Domain 5 Record 17	Real CPU data
Domain 5 Record 19	CPU pool utilization

Perfkit Screen SYSCONF (FCX180) – SMT Disable

```
FCX180  Run 2015/02/15 08:52:14      SYSCONF
                                     System Configuration, Initial and Changed
From 2015/02/14 16:04:29                                                    SYSTEMID
To   2015/02/14 16:14:59                                                  CPU 2964-704
For    630 Secs 00:10:30             "This is a performance report for SYSTEM XYZ"      z/VM   V.6.3.0
```

```
Multithreading          Disabled, No MULTITHREADING statement

 Server Time Protocol (STP) facility configuration
 XRC_TEST enabled            No        XRC_OPTIONAL enabled
 STP H/W feature installed   No        STP H/W feature enabled
 STP Timestamping enabled    No        STP Timezone usage enabled  No
 STP is active               No        STP is suspended            No
 STP susp. message issued    No
 STP TOD clock offset        +00:00:00.0000000000

 Initial Status on 2015/02/14 at 16:04, Processor 2964-704
                          Total  Conf  Stby  Resvd  Ded  Shrd
 Real Proc: Cap 492.0000    103     4     0     99
 Sec. Proc: Cap 492.0000     99    99     0      4
 Log. IFL : CAF       41      8     4     4      0    4     0

 <------- Processor --------->  Core/
 Num Serial-Nr Type Status    Thread
   0   012F17  IFL  Master      00/0
   1   012F17  IFL  Alternate   01/0
   2   012F17  IFL  Alternate   02/0
   3   012F17  IFL  Alternate   03/0
 Processor Configuration Mode: LINUX
```

Disabled as Config file does not contain a ‘multithreaded enabled’ statement.

Total of 4 cores and each core has a thread 0 associated with it.

Perfkit Screen SYSCONF (FCX180) – SMT Enabled

```
FCX180  Run 2015/02/15 08:52:10    SYSCONF
                                   System Configuration, Initial and Changed
From 2015/02/14 16:31:32
To   2015/02/14 16:42:02                                                  CPU 2964-704
For    630 Secs 00:10:30           "This is a performance report for SYSTEM XYZ"      z/VM   V.6.3.0
```

```
 Multithreading          Enabled

 Initial Status on 2015/02/14 at 16:31, Processor 2964-704
                          Total  Conf  Stby  Resvd  Ded  Shrd
 Real Proc: Cap 492.0000    103     4     0     99
 Sec. Proc: Cap 492.0000     99    99     0      4
 Log. IFL : CAF       41      8     4     0      0    4     0

 <------- Processor --------->  Core/
 Num Serial-Nr Type Status    Thread
   0   012F17  IFL  Master      00/0
   1   012F17  IFL  Alternate   00/1
   2   012F17  IFL  Alternate   01/0
   3   012F17  IFL  Alternate   01/1
   4   012F17  IFL  Alternate   02/0
   5   012F17  IFL  Alternate   02/1
   6   012F17  IFL  Alternate   03/0
   7   012F17  IFL  Alternate   03/1
 Processor Configuration Mode: LINUX
```

The z/VM system is enabled for SMT.

Total of 4 cores and each core has both a thread 0 and a thread 1 associated with it.

Perfkit Screen SYSSET (FCX154) – SMT Enabled

```
FCX154  Run 2015/02/15 08:52:10          SYSSET
                                         System Scheduler Settings, Initial and Changed
From 2015/02/14 16:31:32                                                        SYSTEMID
To   2015/02/14 16:42:02                                                    CPU 2964-704
For    630 Secs 00:10:30                 "This is a performance report for SYSTEM XYZ" z/VM   V.6.3.0
```

```
Initial Scheduler Settings: 2015/02/14 at 16:31:32
LIMITHARD algorithm    Consumption
DSPWD method           Reshuffle
Polarization           Vertical
Global Perf. Data      ON
EXCESSUSE: CP   ......  CPUPAD: CP     6400%
           ZAAP ......          ZAAP      0%
           IFL  ......          IFL       0%
           ICF  ......          ICF       0%
           ZIIP ......          ZIIP      0%
Multithreading          Enabled
              <--------- Threads ---------->
              H/W Requested System Activated
Max Threads             Max                2
CP core         1       Max      1         1
IFL core        2       Max      2         2
ICF core        2       Max      1         1
ZIIP core       2       Max      1         1

Changed Scheduler Settings
Date  Time      Changed
..... ........  No changes processed
```

For SMT to be enabled:
1. z/VM Dispatch Workload Algorithm must be at default of Reshuffle.
2. HiperDispatch polarization must be vertical.

Maximum number of threads activated on this z/VM. Activated column = minimum(H/W, Requested, System)

Perfkit Screen PRCLOG (FCX304) – SMT Disabled

```
FCX304  Run 2015/02/15 08:52:14          PRCLOG                                                    Page   56
                                          Processor Activity, by Time
 From 2015/02/14 16:04:29                                                                          SYSTEMID
 To   2015/02/14 16:14:59                                                          CPU 2964-704    SN 12F17
 For    630 Secs 00:10:30                "This is a performance report for SYSTEM XYZ"             z/VM   V.6.3.0 SLU 0000
```

							<--- Percent Busy ---->				<-- Rates per Sec. --->			<----- Paging ------->			<Co>	< Di>	
Interval End Time	CPU	Type	PPD	Ent.	DVID	Pct Park Time	Total	User	Syst	Emul	Inst Siml	DIAG	SIGP	SSCH	<2GB /s	PGIN /s	Fast Path %	Page Read /s	<mm> Msgs /s	< ag> X'9C' /s	Core/ Thread
>>Mean>>	0	IFL	VhD	100	0000	0	95.7	95.5	.2	88.2	38153	551.3	22.8	37.1	.0	.0	....	.0	.2	.0	00/0
>>Mean>>	1	IFL	VhD	100	0001	0	95.7	95.5	.2	88.2	37536	492.2	10.3	2.7	.0	.0	....	.0	.0	.0	01/0
>>Mean>>	2	IFL	VhD	100	0002	0	95.6	95.4	.2	88.0	38178	509.8	74.0	2.9	.0	.0	....	.1	.0	.0	02/0
>>Mean>>	3	IFL	VhD	100	0003	0	95.5	95.3	.2	87.8	38532	508.4	8.8	4.8	.0	.0	....	.1	.1	.0	03/0
>>Total>	4	IFL	VhD	400	MIX	0	382.5	381.6	.9	352.1	152k	2062	115.9	47.5	.0	.0	....	.2	.3	.0	MIX

Report remains similar to the past, especially with SMT disabled. You will again see the Core/Thread nomenclature.

Core/
Thread
00/0
01/0
02/0
03/0
MIX

Perfkit Screen PRCLOG (FCX304) – SMT Enabled

```
04  Run 2015/02/15 08:52:10          PRCLOG                                                    Page   56
                                      Processor Activity, by Time
m 2015/02/14 16:31:32                                                                        SYSTEMID
  2015/02/14 16:42:02                                                                  CPU 2964-704   SN 12F17
    630 Secs 00:10:30                 "This is a performance report for SYSTEM XYZ"        z/VM   V.6.3.0 SLU 0000
```


terval d Time	C P U	Type	PPD	Ent.	DVID	Pct Park Time	Total	User	Syst	Emul	Inst Siml	DIAG	SIGP	SSCH	<2GB /s	PGIN /s	Fast Path %	Page Read /s	<mm> Msgs /s	< ag> X'9C' /s	Core/ Thread
Mean>>	0	IFL	VhD	100	0000	0	84.7	84.5	.2	77.0	30035	416.7	1124	34.6	.0	.0	....	.2	.2	.0	00/0
Mean>>	1	IFL	VhD	100	0000	0	84.3	84.1	.2	76.8	29845	447.8	1054	2.0	.0	.0	....	.0	.0	.0	00/1
Mean>>	2	IFL	VhD	100	0001	0	84.5	84.4	.2	76.8	31053	439.6	1098	1.4	.0	.0	....	.0	.0	.0	01/0
Mean>>	3	IFL	VhD	100	0001	0	84.6	84.4	.2	77.0	30648	491.9	1028	1.2	.0	.0	....	.0	.0	.0	01/1
Mean>>	4	IFL	VhD	100	0002	0	84.5	84.3	.2	77.0	29912	535.7	1106	1.7	.0	.0	....	.0	.0	.0	02/0
Mean>>	5	IFL	VhD	100	0002	0	84.9	84.7	.2	77.5	29667	526.1	1029	1.3	.0	.0	....	.0	.0	.0	02/1
Mean>>	6	IFL	VhD	100	0003	0	84.8	84.6	.2	77.3	29368	450.1	1062	2.1	.0	.0	....	.1	.0	.0	03/0
Mean>>	7	IFL	VhD	100	0003	0	84.7	84.5	.2	77.3	29026	566.8	1027	2.0	.0	.0	....	.0	.0	.0	03/1
Total>	8	IFL	VhD	800	MIX	0	677.0	675.5	1.5	616.6	240k	3875	8527	46.2	.0	.0	....	.2	.3	.0	MIX

With SMT enabled, you see each thread is shown as a “Logical CPU” on this report. The utilizations are of the thread, no longer the “core”.