What's New in the z/VM 6.3 Hypervisor *Session 17515*

John Franciscovich

IBM: z/VM Development

Endicott, NY

Presented by Bill Bitner
bitnerb@us.ibm.com

#SHAREorg

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

CELEBRATING 60 YEARS OF SHARE Influencing IT Since 1955

Trademarks

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

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

* 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.

Notice Regarding Specialty Engines (e.g., zIIPs, zAAPs and IFLs):

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 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 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.

Acknowledgements

- Bill Bitner
- Brian Wade
- Alan Altmark
- Emily Hugenbruch
- Mark Lorenc
- Kevin Adams
- Romney White

- … and anyone else who contributed to this presentation that I may have omitted

z/VM 6.3 Topics

- z/VM 6.3 Overview and Evolution
- Support for the IBM z13
  - Compatibility
  - Exploitation of z13 features
- 2014 Enhancements
  - Environment Information Interface
  - CPU Pooling
  - PCIe
- Highlights of z/VM 6.3 base release
  - Scalability
    - Large Memory Support
    - Enhanced Dump Support
  - HiperDispatch
- Additional Information
  - Virtual Networking
  - Technology Exploitation
  - Miscellaneous Enhancements

z/VM 6.3: Themes

- Reduce the number of z/VM systems you need to manage
  - Expand z/VM systems constrained by memory up to four times
    - Increase the number of Linux virtual servers in a single z/VM system
  - Exploit HiperDispatch to improve processor efficiency
    - Allow more work to be done per IFL
    - Support more virtual servers per IFL
  - Expand real memory available in a Single System Image Cluster up to 4 TB
- Improved memory management flexibility and efficiency
  - Benefits for z/VM systems of all memory sizes
  - More effective prioritization of virtual server use of real memory
  - Improved management of memory on systems with diverse virtual server processor and memory use patterns

z/VM 6.3: 2014 Enhancements

- **Environment Information Interface**
  - Available with APAR VM65419 / PTF UM34348

- **CPU Pooling**
  - Available with APAR VM65418 / PTF UM34348

- **PCIe / 10GbE RoCE Express Feature / zEDC Express Feature**
  - Available with:
    - IBM zEC12 or zBC12, driver 15, bundle 21
    - VM CP - APAR VM65417 / PTF UM34343
    - VM CMS - APAR VM65437 / PTF UM34401
    - VM TCP/IP - APAR PI20509 / PTF UI19055
    - VM DVF - APAR VM65572 / PTF UM34342
    - z/OS 2.1 - APAR OA43256 / PTF UA72717
    - z/OS 2.1 - APAR OA44482 / PTF UA73687
  - Fullfills 2013 Statement of Direction

z/VM 6.3: Recent Announcements

- Support for the IBM z13™
  - Compatibility
  - z/VM Enhancements to exploit z13 features
    - Simultaneous Multithreading (SMT)
    - Increased Processor Scalability
    - Multi-VSwitch Link Aggregation

Recent Announcements - z/VM 6.3 Exploitation of IBM z13

Expanding the Horizon of Virtualization

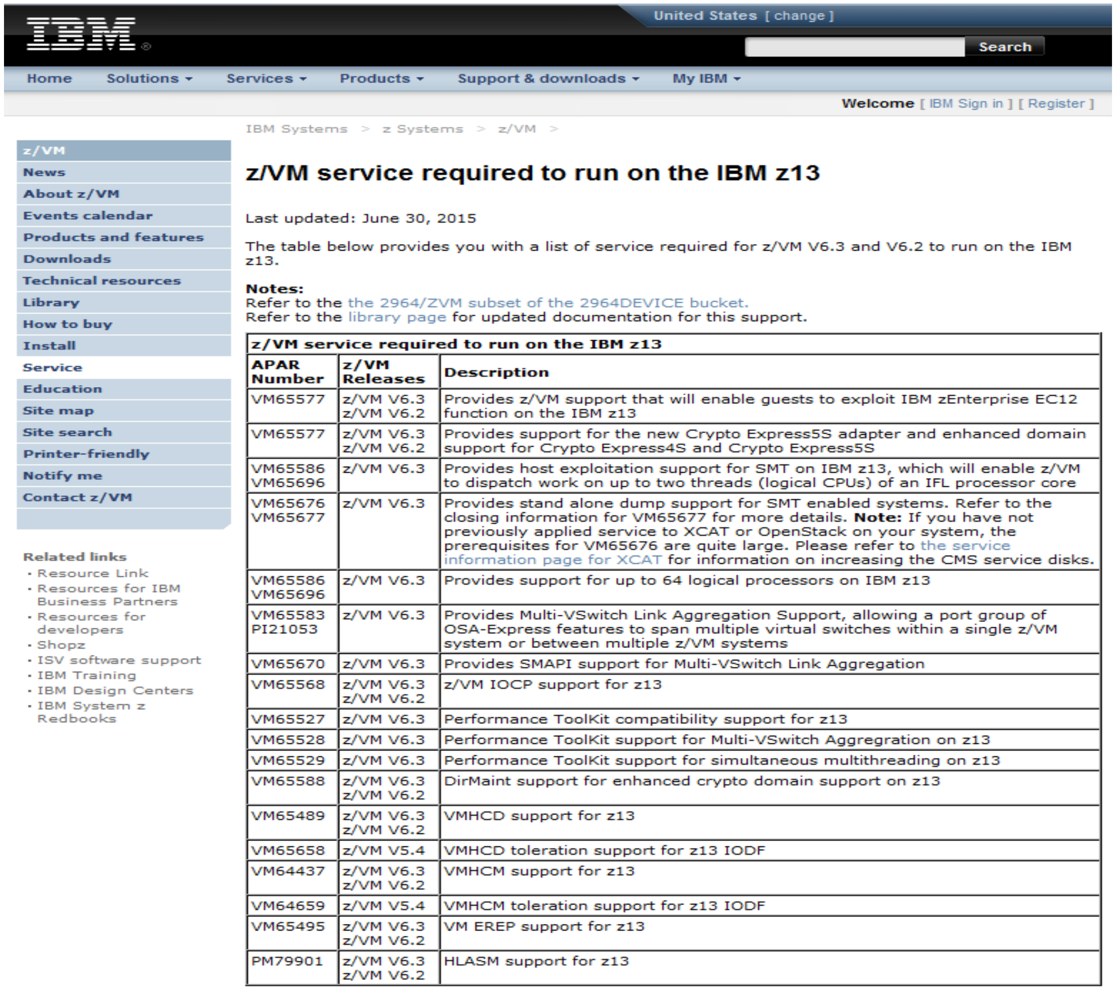
- Release for Announcement – The IBM z13™
  - January 14, 2015
  - Announcement Link

- z/VM Compatibility Support
  - PTFs available February 13, 2015
  - Also includes Crypto enhanced domain support
  - z/VM 6.2 and z/VM 6.3
  - No z/VM 5.4 support
  - Refer to bucket for full list

- Enhancements and Exploitation Support only on z/VM 6.3
  - IBM z13 Simultaneous Multithreading
  - Increased Processor Scalability
  - Multi-VSwitch Link Aggregation Support (Link Aggregation with Shared OSAs)

z/VM Service Required for the IBM z13

http://www.vm.ibm.com/service/vmreqz13.html



z/VM service required to run on the IBM z13

Last updated: June 30, 2015

The table below provides you with a list of service required for z/VM V6.3 and V6.2 to run on the IBM z13.

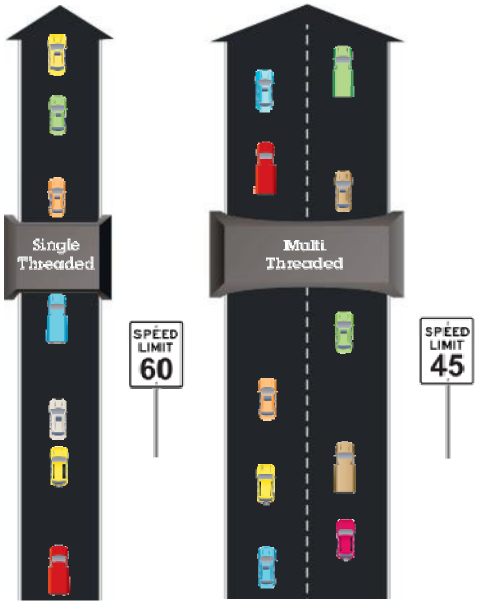
Notes:
Refer to the the 2964/ZVM subset of the 2964DEVICE bucket.
Refer to the library page for updated documentation for this support.

z/VM service required to run on the IBM z13

APAR Number	z/VM Releases	Description
VM65577	z/VM V6.3 z/VM V6.2	Provides z/VM support that will enable guests to exploit IBM zEnterprise EC12 function on the IBM z13
VM65577	z/VM V6.3 z/VM V6.2	Provides support for the new Crypto Express5S adapter and enhanced domain support for Crypto Express4S and Crypto Express5S
VM65586 VM65696	z/VM V6.3	Provides host exploitation support for SMT on IBM z13, which will enable z/VM to dispatch work on up to two threads (logical CPUs) of an IFL processor core
VM65676 VM65677	z/VM V6.3	Provides stand alone dump support for SMT enabled systems. Refer to the closing information for VM65677 for more details. **Note:** If you have not previously applied service to XCAT or OpenStack on your system, the prerequisites for VM65676 are quite large. Please refer to the service information page for XCAT for information on increasing the CMS service disks.
VM65586 VM65696	z/VM V6.3	Provides support for up to 64 logical processors on IBM z13
VM65583 PI21053	z/VM V6.3	Provides Multi-VSwitch Link Aggregation Support, allowing a port group of OSA-Express features to span multiple virtual switches within a single z/VM system or between multiple z/VM systems
VM65670	z/VM V6.3	Provides SMAPI support for Multi-VSwitch Link Aggregation
VM65568	z/VM V6.3 z/VM V6.2	z/VM IOCP support for z13
VM65527	z/VM V6.3	Performance ToolKit compatibility support for z13
VM65528	z/VM V6.3	Performance ToolKit support for Multi-VSwitch Aggregration on z13
VM65529	z/VM V6.3	Performance ToolKit support for simultaneous multithreading on z13
VM65588	z/VM V6.3 z/VM V6.2	DirMaint support for enhanced crypto domain support on z13
VM65489	z/VM V6.3 z/VM V6.2	VMHCD support for z13
VM65658	z/VM V5.4	VMHCD toleration support for z13 IODF
VM64437	z/VM V6.3 z/VM V6.2	VMHCM support for z13
VM64659	z/VM V5.4	VMHCM toleration support for z13 IODF
VM65495	z/VM V6.3 z/VM V6.2	VM EREP support for z13
PM79901	z/VM V6.3 z/VM V6.2	HLASM support for z13

Simultaneous Multithreading (SMT) on z/VM

- Objective is to improve capacity, not the speed of a single instruction stream
- z/VM can now dispatch work on up to two threads of a z13 IFL core
  - Up to 32 cores supported
- VM65586 for z/VM 6.3 **only**
  - PTF UM34552 became available March 13, 2015
- Requires z13 millicode bundle 11
- Transparent to virtual machines
  - Guests do not need to be SMT-aware
  - SMT is not virtualized to the guest
- z13 exploitation is for IFL cores only
- SMT is disabled by default
  - Requires a system configuration setting and re-IPL
  - When enabled, applies to the entire system
- Potential to increase the overall capacity of the system
  - Workload dependent

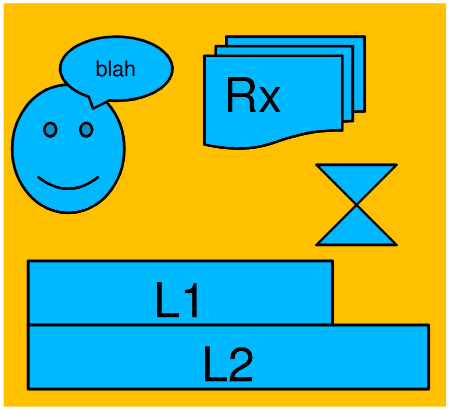


Which approach is designed for the higher volume of traffic? Which road is faster?

** Illustrative numbers only*

Cores and Threads

Single threaded cores



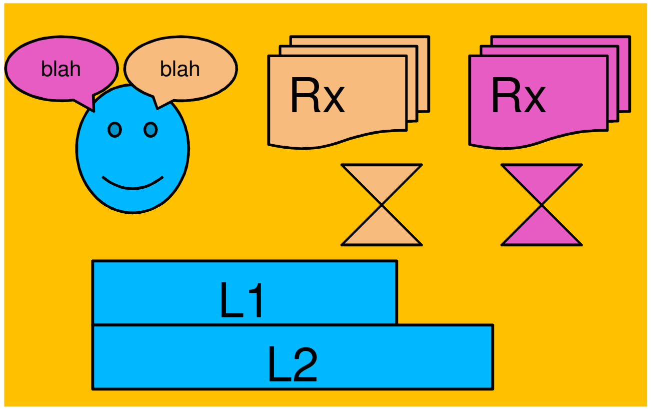
Core: L1, L2, address translator, …

Thread: PSW, registers, address translations, timers, … *execution context*

zEC12 had **one** thread per core.

What’s mine is mine, no sharing!

Multithreaded cores



Core: L1, L2, address translator, …

Thread: PSW, registers, address translations, timers, … *execution context*

z13 has **two** threads per core for IFLs and zIIPs. The rest have **one**.

The threads must share some core facilities!

How do I enable SMT on my z/VM system?

1. Install your IBM z13 mainframe
2. Install service for APAR **VM65586**
3. Set up an LPAR with at least some IFL engines
   - Could be a Linux-only LPAR with all IFLs
   - Could be a VM-mode LPAR with some IFLs
4. The system must be in *vertical polarization mode* (this is the default)
   Make sure you ***don't*** have an **SRM POLARIZATION HORIZONTAL** statement in your SYSTEM CONFIG.
5. The system must be using the *reshuffle dispatcher method* (this is the default)
   Make sure you ***don't*** have an **SRM DSPWDMethod REBALANCE** statement in your SYSTEM CONFIG.
6. Add the **MULTITHreading ENAble** statement to your SYSTEM CONFIG
7. Re-IPL your system!

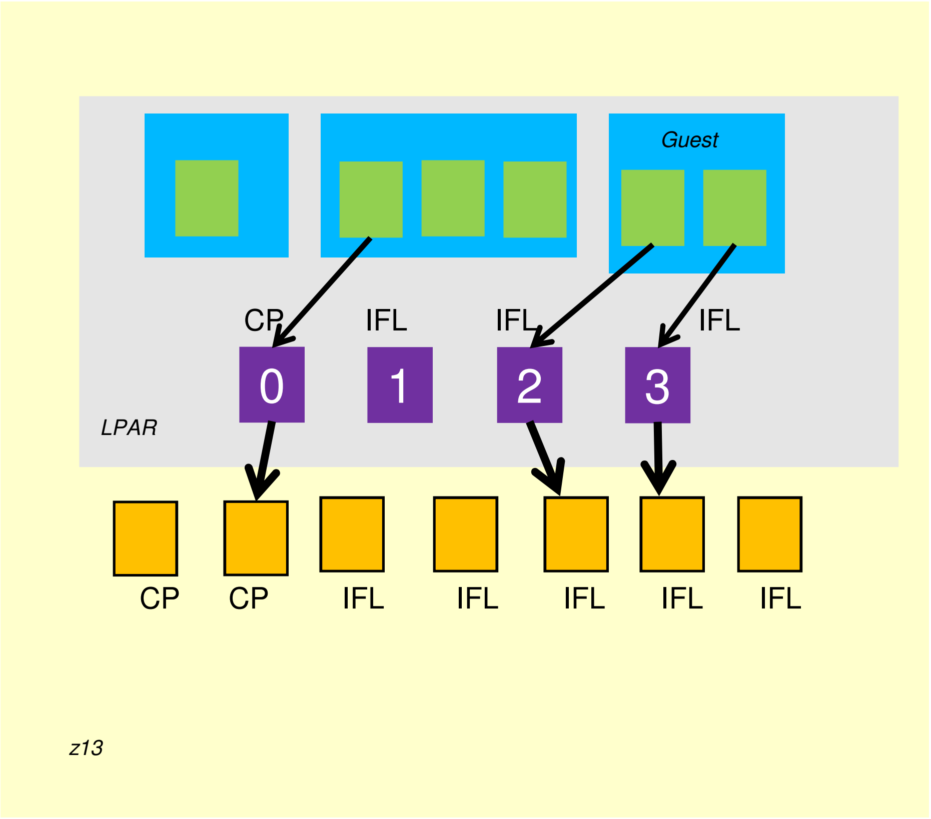
Enabling SMT – MULTIThreading statement

- **MULTIThreading** configuration statement allows you to specify either
  - maximum number of threads for all core types
  - different number of threads for each type
    - z/VM supports multithreading on only IFL cores

- **CPSYNTAX** has been updated to verify:
  - Are there multiple **MULTIThreading** statements?
  - Is the maximum activated thread value less than the number of threads specified for any type?
  - Is **MULTIThreading ENABLE** specified with any incompatible **SRM** statements?

I enabled SMT; what does that mean for guests?

SMT disabled



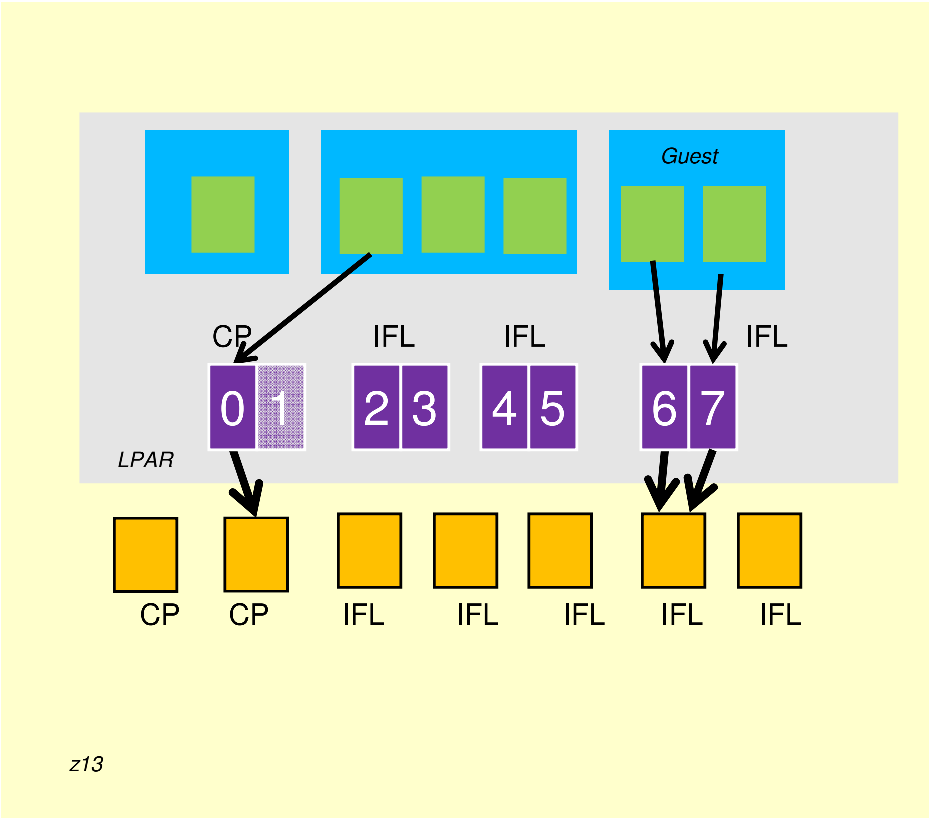
z/VM provides *virtual CPUs* for guests.
z/VM dispatches virtual CPUs on logical CPUs.

When the partition does not *opt in to SMT*, PR/SM provides *logical CPUs* for the partition.
PR/SM dispatches *one* logical CPU on a physical core at a time.

Each physical IFL core can run *two* streams of instructions at a time.
We say each one has two *threads.*
In this case for IFL cores, one thread goes unused.

I enabled SMT; what does that mean for guests?

SMT enabled



z/VM still provides virtual CPUs for guests.

z/VM still dispatches virtual CPUs on logical CPUs.

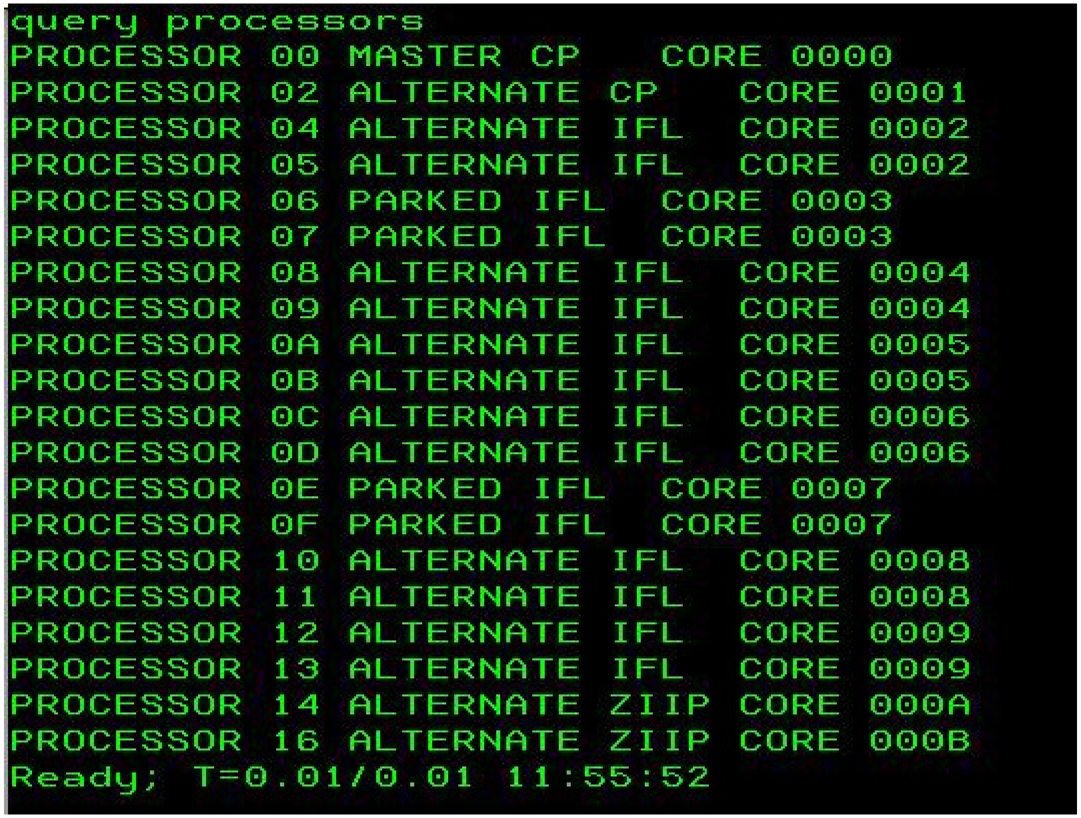
When the partition *opts in to SMT*, PR/SM provides logical CPUs for the partition
and groups them into *logical cores*.

PR/SM dispatches *one* logical core on a physical core at a time.

Each physical IFL core can run *two* streams of instructions at a time. We say each one has two *threads*. In this case for IFL cores, both threads are used.

QUERY PROCEssors with SMT

- Shows which core each logical CPU is on:

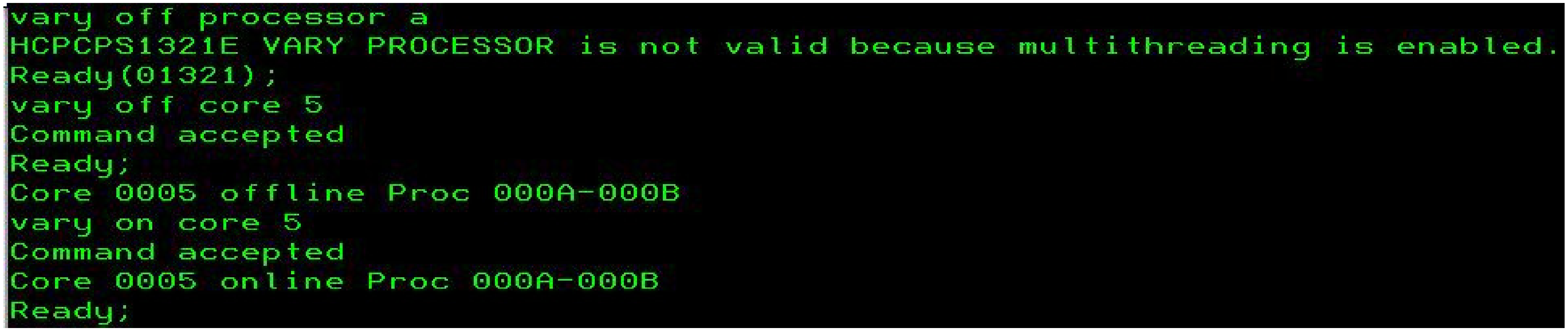


```
query processors
PROCESSOR 00 MASTER CP   CORE 0000
PROCESSOR 02 ALTERNATE CP   CORE 0001
PROCESSOR 04 ALTERNATE IFL  CORE 0002
PROCESSOR 05 ALTERNATE IFL  CORE 0002
PROCESSOR 06 PARKED IFL  CORE 0003
PROCESSOR 07 PARKED IFL  CORE 0003
PROCESSOR 08 ALTERNATE IFL  CORE 0004
PROCESSOR 09 ALTERNATE IFL  CORE 0004
PROCESSOR 0A ALTERNATE IFL  CORE 0005
PROCESSOR 0B ALTERNATE IFL  CORE 0005
PROCESSOR 0C ALTERNATE IFL  CORE 0006
PROCESSOR 0D ALTERNATE IFL  CORE 0006
PROCESSOR 0E PARKED IFL  CORE 0007
PROCESSOR 0F PARKED IFL  CORE 0007
PROCESSOR 10 ALTERNATE IFL  CORE 0008
PROCESSOR 11 ALTERNATE IFL  CORE 0008
PROCESSOR 12 ALTERNATE IFL  CORE 0009
PROCESSOR 13 ALTERNATE IFL  CORE 0009
PROCESSOR 14 ALTERNATE ZIIP CORE 000A
PROCESSOR 16 ALTERNATE ZIIP CORE 000B
Ready; T=0.01/0.01 11:55:52
```

Vary On and Off

- When SMT is enabled
  - Use **VARY CORE** to vary off or on an entire core
    - Multithread or single thread cores
    - Cannot vary a single thread of a core.
  - **VARY PROCESSOR** isn't allowed

- When SMT is not installed or not enabled
  - **VARY CORE** is the same as **VARY PROCESSOR**



```
vary off processor a
HCPCPS1321E VARY PROCESSOR is not valid because multithreading is enabled.
Ready(01321);
vary off core 5
Command accepted
Ready;
Core 0005 offline Proc 000A-000B
vary on core 5
Command accepted
Core 0005 online Proc 000A-000B
Ready;
```

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

Prorated Core Time (availability TBD)

- Prorated core time will divide the time a core is dispatched proportionally among the threads dispatched in that interval
  - Full time charged while a vCPU runs alongside an idle thread
  - Half time charged while a vCPU is dispatched beside another active thread
- Therefore:
  - CPU pool capacity consumed as if by cores
  - Suitable for core-based software licensing
- When SMT is enabled, prorated core time will be calculated for users who are
  - In a CPU pool limited by the **CAPACITY** or **LIMITHARD** option
  - Limited by the **SET SHARE LIMITHARD** command

    (currently raw time is used; raw time will continue to be used when SMT is disabled)
- Only CAPACITY-based CPU pools meet requirements for sub-capacity pricing
- **QUERY CPUPOOL** will report capacity in terms of cores' worth of processing power instead of CPUs'
- Prorated core time will be reported in monitor records and the new Type F accounting record.
- Watch for APAR VM65680

Live Guest Relocation Implications

- Guests can be relocated between SMT enabled and SMT disabled z/VM systems because SMT is transparent to guests
  - Capacity will be affected
    - Might require adjustment to the number of virtual CPUs
  - Because of differences in CPU time calculation they may see their CPU time advance at different rates.
    - But their time will never go backward!

Increased CPU Scalability

- Various improvements to help z/VM to run more efficiently when large numbers of processors are present, thereby improving the N-way curve

- APAR VM65586 for z/VM 6.3 **only**
  - PTF UM34552 available March 12, 2015

- For z13
  - With SMT disabled, increases logical processors supported from 32 to 64
  - With SMT enabled, the limit is 32 logical cores (yields at most 64 logical processors)

- For machines prior to z13
  - Limit remains at 32 logical processors
  - Might still benefit from improved N-way curves

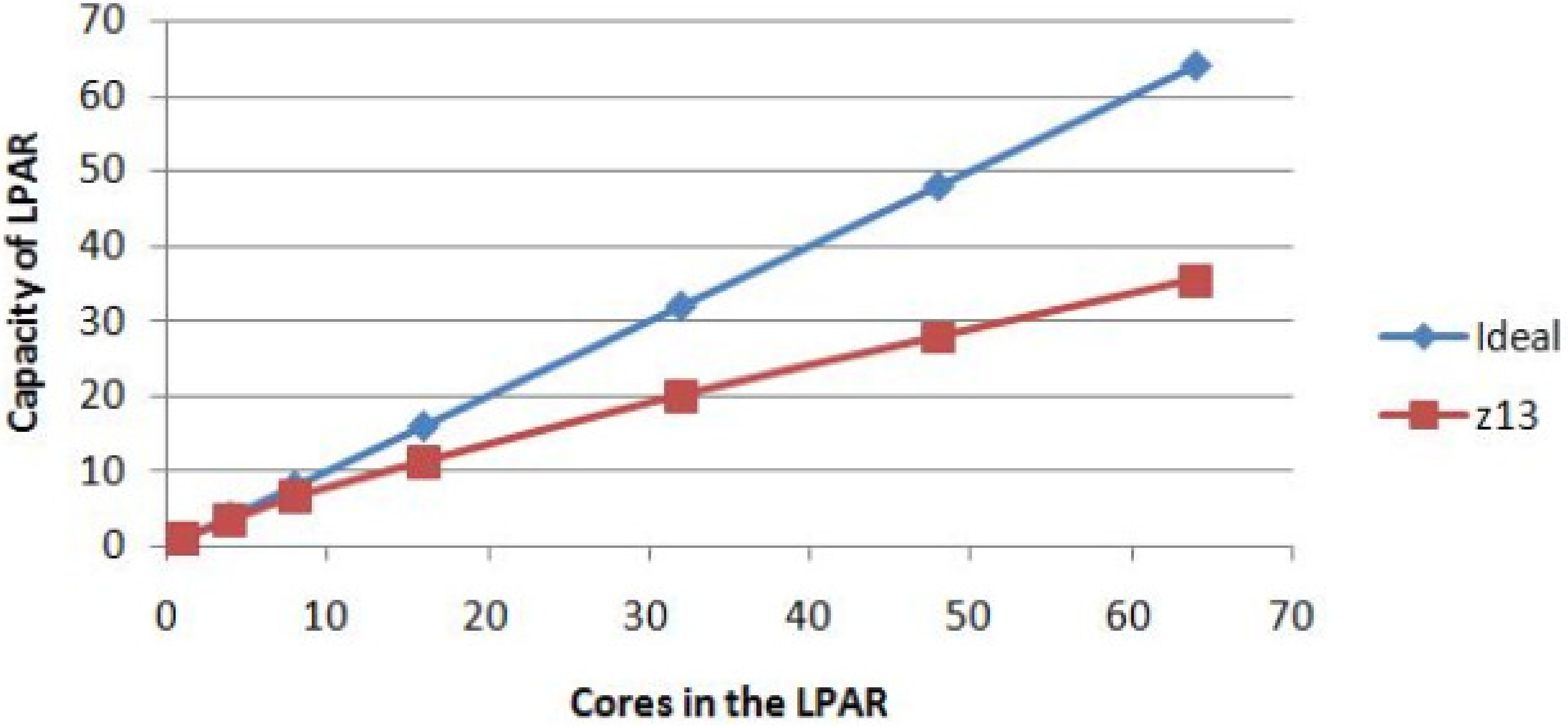
Areas Improved to Increase CPU Scalability

- Improvements were made to the following areas to improve efficiency and reduce contention:
  - Scheduling & Dispatching
  - Virtual Networking
  - Virtual Disk in Storage
  - Memory Management
- Some areas needing improvement were known – others required thorough investigation and experimentation
- All tested workloads showed acceptable scaling up to…
  - … 64 logical processors when SMT is enabled
  - … 32 logical processors when SMT is not enabled
- Benefits are workload-dependent

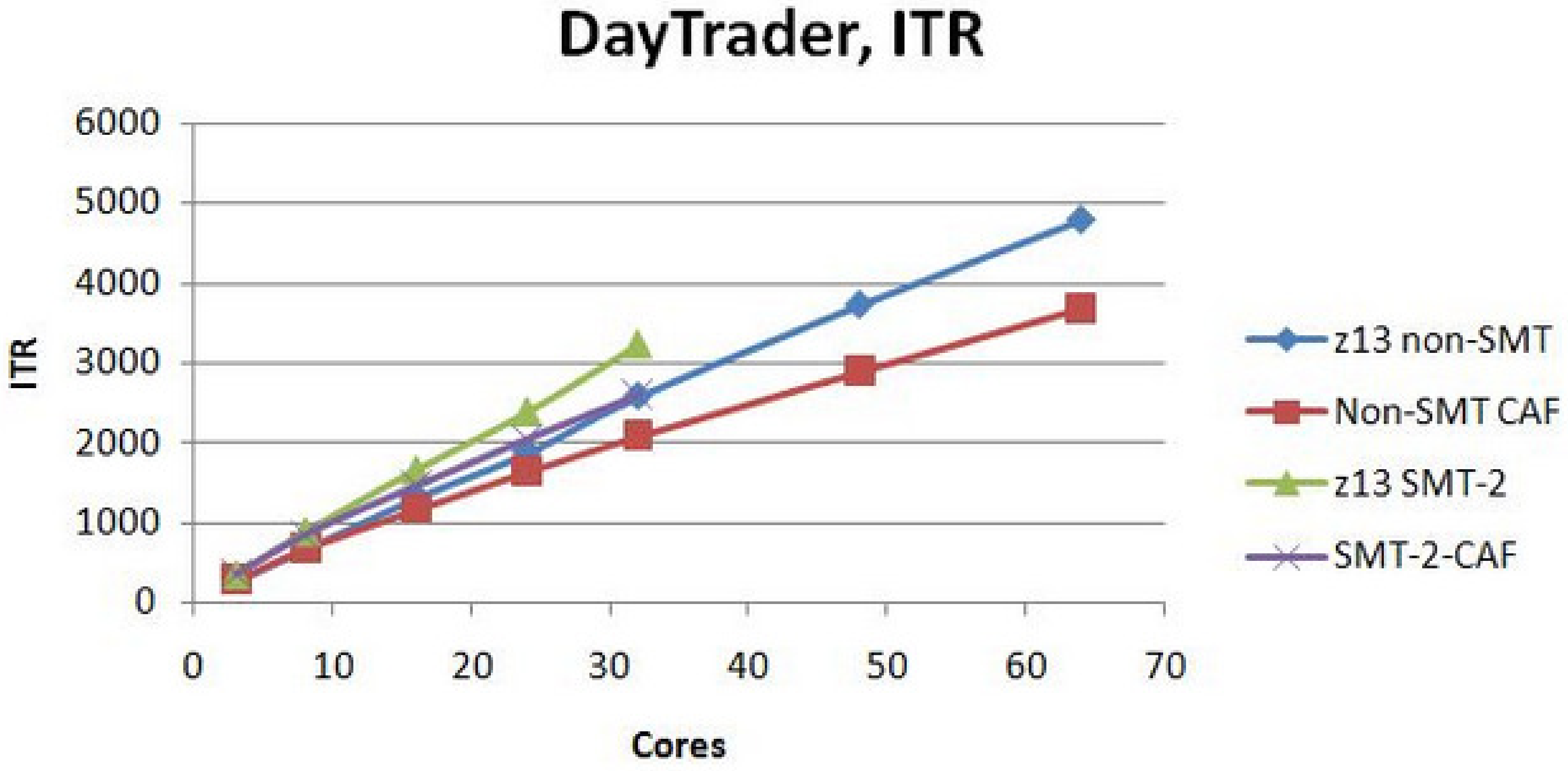
Areas Improved with Scalability Enhancements

- z/VM Scheduler Lock
  - Management of internal stacked work
  - Guests going into a wait state
- Locking for Memory Management
  - Most benefit during system initialization and when very constrained with memory
- Serialization and processing of VDisk I/Os
- Batching and processor-local queues for VSWITCH buffers

Hypothetical z13 Scaling Curves



- Ideal is linear scaling or if no impact from adding cores
- “z13” is based on the Capacity Adjustment Factor based on traditional z/OS scaling and other variables.



DayTrader (Java based Trading application) ITR scaling curves. 2964-NC9, dedicated LPAR, storage-rich. z/VM 6.3 with z13 exploitation SPE, non-SMT. z/VM 6.3 with z13 exploitation SPE, SMT-2.

Multi-VSwitch Link Aggregation

- Makes it possible to do Link Aggregation with VSwitches without the requirement for dedicated OSAs

- Allows a port group of OSA-Express features to span VSwitches within a single or multiple z/VM systems in same CEC
  - Cannot be shared with non-z/VM logical partitions or z/VM systems without support

- Only available on z13
  - Requires OSA enhancements introduced with the z13

- Allows better consolidation and availability while improving TCO

- APARs VM65583 and PI21053 for z/VM 6.3 **only**
  - PTFs planned to be available June 26, 2015

z Systems Crypto Enhancements

- z Systems *crypto architecture* itself grew
  - Maximum number of crypto features increased from 64 to 256
  - Maximum number of domains per AP increased from 16 to 256
  - The new architecture applies to Crypto Express4S and Crypto Express5S on z13

- z13 uses the new architecture to support more domains
  - Up to 16 APs and up to 85 domains per AP

- New Crypto Express5S card
  - Increased performance over Crypto Express4S
  - New functions: VISA Formatted Preserving Encryption, Elliptic Curve Cryptography

- Read more about it:
  - Ultimate Security with the IBM z13, an IBM Redbooks publication

z/VM Crypto Enhancements

- z/VM can handle the expanded z Systems architecture

- z/VM supports the new Crypto Express5S card
  - When configured as: (same situation as Express4S, by the way)

Config mode	Shared?	Dedicated?
IBM Common Cryptographic Architecture (CCA) coprocessor	yes	yes
IBM Enterprise PKCS #11 (EP11) coprocessor		yes
Accelerator	yes	yes

  - Only z/Architecture guests can use the new Crypto Express5S card (same as Express4S)
  - Authorized via **CRYPTO** statement in guest's CP directory entry (same as Express4S)

- New z/VM system configuration statement lets the administrator specify which domains should be used to fulfill shared crypto
  - Stops all the nondedicated domains from defaulting into shared mode
  - If no statement specified, only two domains get reserved for shared
  - **CRYPTO APVIRTUAL AP n1-n2 DOMAIN p1-p2**

2014 Enhancements

Environment Information Interface

- New programming interface allows guests to capture execution environment
  - Configuration and Capacity information
  - Various Levels:
    - Machine, logical partition, hypervisor, virtual machine, CPU pools

- New problem state instruction STore HYpervisor Information (STHYI)
  - Supported by z/VM 6.3
  - Tolerated by z/VM 6.2 ("function not supported")

- Used by IBM License Metric Tool (ILMT)
  - New ILMT 9.0.1 includes the ability to track CPU pools

CPU Pooling

- Define and limit the aggregate amount of CPU resources that a group of z/VM guests is allowed to consume
  - Allows capping of CPU utilization for a set of guests to better balance resource utilization
- Define one or more named pools in which a limit of CPU resources is set
  - No restrictions on number of pools or aggregate capacity (can overcommit)
- CPU pools coexist with individual share limits
  - More restrictive limit applies
- CPU pools in SSI clusters
  - Pool capacities are independent and enforced separately on each member
  - Live Guest Relocation
    - Destination member must have an identically named pool with same **TYPE** attribute
    - If limit is not required on destination, remove guest from pool before relocating
  - Recommend defining pools with identical names and types on all members of cluster

CPU Pooling: Use Cases

- Department resource requirements
  - Assign each department’s guests to CPU pool with contracted capacity

- Grow workloads without affecting existing requirements and limits
  - Add New Workload
  - Add Capacity
  - Combine LPARs
  - Handle fractional workload requirements

- Prevent resource over-consumption
  - Limit aggressive workloads

CPU Pooling: Defining and Managing

- Use the **DEFINE CPUPOOL** command to define named pools
  - **LIMITHARD** - % of system CPU resources
  - **CAPACITY** – number of CPUs
  - Define for a particular **TYPE** of CPU (**CP** or **IFL**)

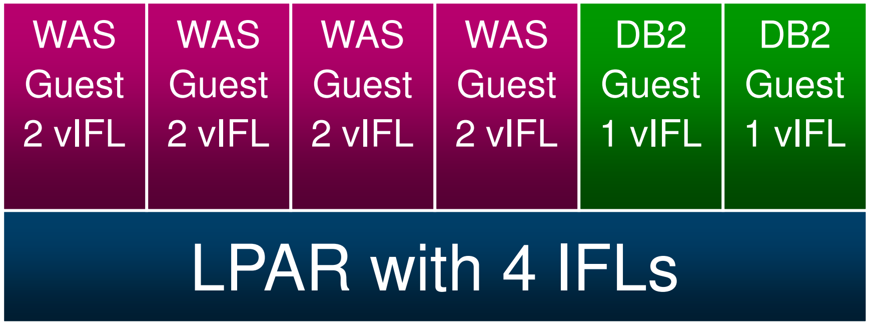
- Limits can be changed with the **SET CPUPOOL** command
- Assign and remove guests to/from a CPU pool with the **SCHEDULE** command
- Use **QUERY CPUPOOL** to see information about the pools that are defined on your system

```
query cpupool all
CPU pool  Limit     Type      Members
LINUXP2  8.0 CPUs    IFL            0
CPPOOL10  12 %        CP            8
LINUXP3   30 %       IFL           20
LINUXP1  2.5 CPUs    IFL            6

query cpupool linuxp1 members
CPU pool Limit      Type      Members
LINUXP1 2.5 CPUs     IFL            6
The following users are members of CPU pool LINUXP1:
D70LIN12 D79LIN03 D79ADM D79LIN10 D79LIN07
D79LIN04
```

Add New Workload: Without CPU Pooling

- 4 production guests for WAS
  - May consume up to 4 engines
- Add 2 production guests for DB2
  - May consume up to 2 engines



WAS Guest 2 vIFL

WAS Guest 2 vIFL

WAS Guest 2 vIFL

WAS Guest 2 vIFL

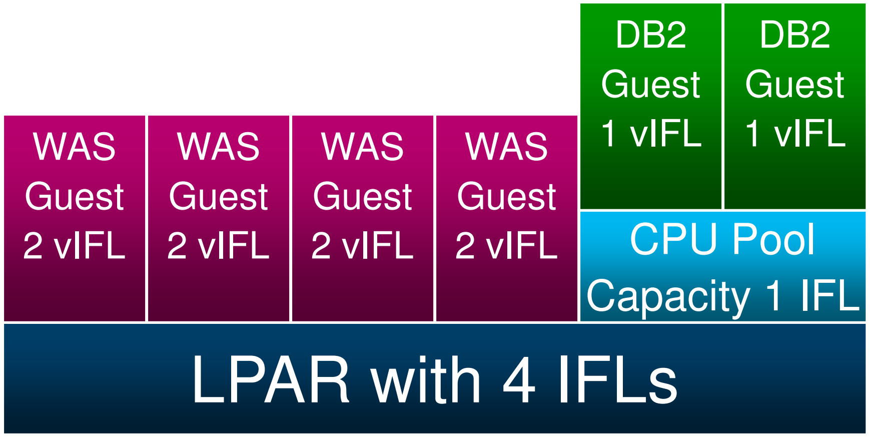
DB2 Guest 1 vIFL

DB2 Guest 1 vIFL

LPAR with 4 IFLs

Add New Workload: With CPU Pooling

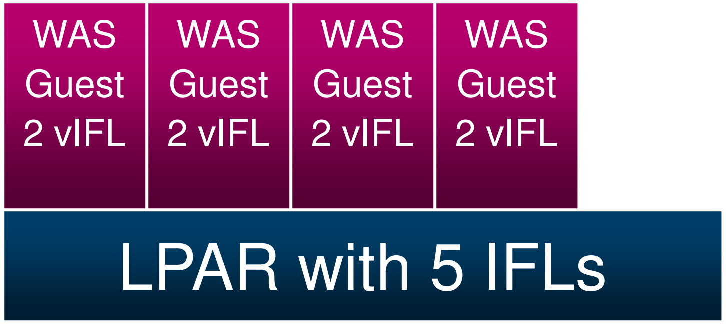
- 4 production guests for WAS
  - May consume up to 4 engines
- Create a 1-IFL pool
- Put the 2 DB2 production guests in the pool
  - DB2 is limited to 1 engine instead of 2



- Allows new workloads to be added cost effectively
- Encourages additional workload consolidation after initial success

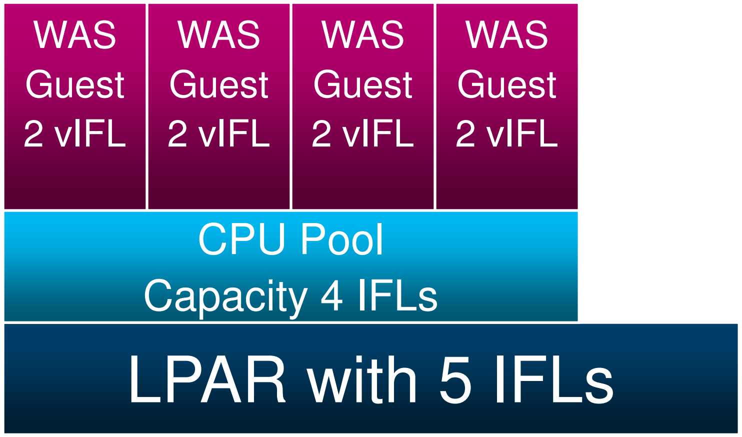
Add Capacity: Without CPU Pooling

- 4 production guests for WAS
  - May consume up to 4 engines
- Add another IFL to the LPAR
  - Limit for WAS increases to 5 engines



Add Capacity: With CPU Pooling

- **LPAR with 4 IFLs**
- **Set up CPU Pooling for 4 IFLs**
  - **Limits guests for WAS to 4 engines**
- **Add another IFL to the LPAR**
  - **WAS remains limited to 4 engines**
    - **Allows capacity to be added for new workload without increasing consumption of existing workloads**



WAS Guest 2 vIFL

WAS Guest 2 vIFL

WAS Guest 2 vIFL

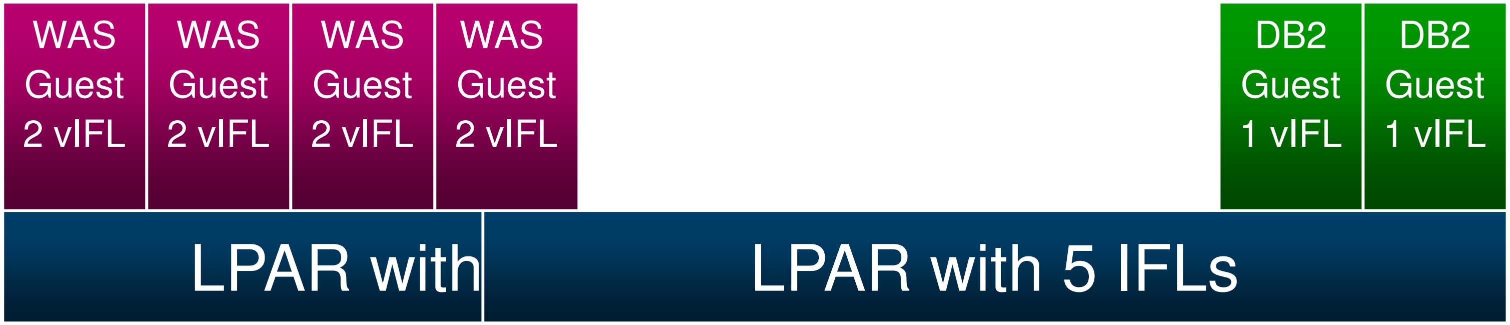
WAS Guest 2 vIFL

CPU Pool
Capacity 4 IFLs

LPAR with 5 IFLs

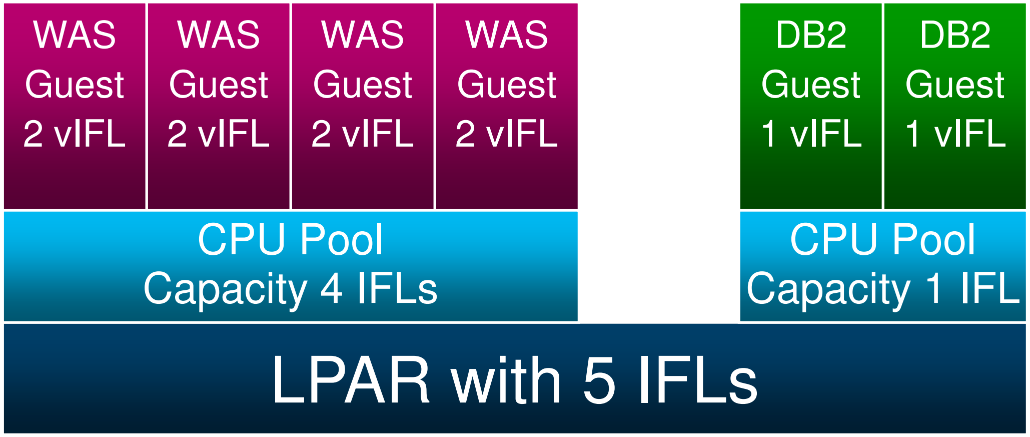
Combine LPARs: Without CPU Pooling

- LPAR with 4 IFLs and 4 production guests for WAS
  - May consume up to 4 engines
- LPAR with 1 IFL and 2 production guests for DB2
  - May consume up to 1 engine
- LPARs merge to one LPAR with 5 IFLs
  - Limit for WAS increases to 5 engines
  - Limit for DB2 increases to 2 engines



Combine LPARs: With CPU Pooling

- LPAR with 5 IFLs
- Create 2 Pools – one with 4 IFLs and one with 1 IFL
- Place the four WAS guests in the 4 IFL pool and the two DB2 guests in the 1 IFL pool
  - WAS remains limited to 4 engines
  - DB2 remains limited to 1 engine



- Avoids increase in software license requirements (and costs)
- Reduces z/VM system management and maintenance workload
- Consolidates resources (memory, paging, network) for greater efficiency

PCIe Support: Overview

- Basis for support for guest exploitation of
  - 10GbE RoCE Express Feature
  - zEDC Express Feature

- Allows guests with PCIe drivers to access PCI "functions" (devices)

- PCI functions can be dedicated to a guest
  - Guest must have PCI driver supporting specific function

Defining and Managing PCI Functions

- PCI functions are defined in the IOCP
  - May also be defined, modified, and deleted dynamically with new commands
    - **DEFINE PCIFUNCTION**
    - **MODIFY PCIFUNCTION**
    - **DELETE PCIFUNCTION**

    ⟶ Update IOCP so you don't lose your dynamic definitions

- New or enhanced commands to manage PCI functions
  - **VARY PCIFUNCTION**
  - **ATTACH** (PCIFUNCTION operand)
  - **DETACH PCIFUNCTION**
  - **QUERY PCIFUNCTION**
    - *Sample query response:*

```
PCIF 00000003 ATTACHED TO USER01 00000001 DISABLED 10GbE RoCE
PCIF 00000004 FREE                        DISABLED 10GbE RoCE
PCIF 00000021 NOT CONFIGURED              STANDBY 10GbE RoCE
PCIF 00000026 NOT CONFIGURED              STANDBY 10GbE RoCE
PCIF 00000029 FREE                        DISABLED 10GbE RoCE
PCIF 00000032 ATTACHED TO USER02 00000032 ENABLED 10GbE RoCE
PCIF 00000033 FREE                        ERROR 10GbE RoCE
```

Enabling PCIe Support

- Make sure you have required hardware
  - IBM zEC12 or zBC12, driver 15, bundle 21
- System configuration flie
  - Enable new **PCI** feature on **FEATURES** statement
  - Define size of **IOAT** subpool (in megabytes) on **STORAGE** statement
    - Specify warning threshold percentage for usage

```
STORAGE IOAT 2 Megabytes WARN 80 Percent
```

  - Use **LOCKING** operand to define limits of available storage to be used by PCIe functions
    - Specify percentages to issue warning message and to fail lock request

```
STORAGE LOCKING WARN 50 Percent FAIL 80 Percent
```

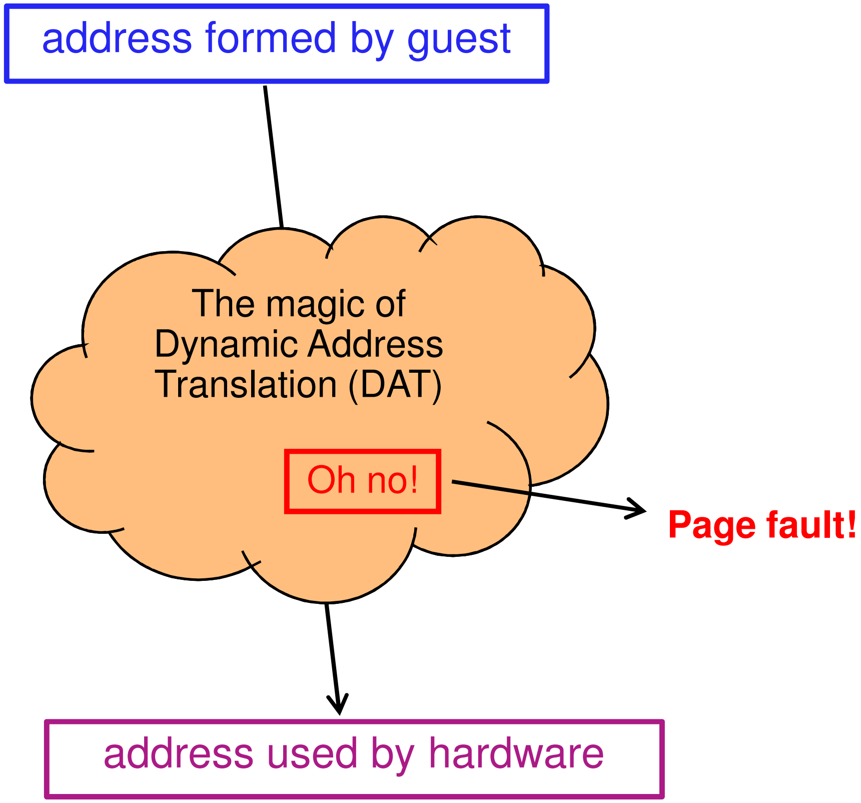
  - **QUERY FRAMES** shows **IOAT** and **LOCKING** settings and usage
- Review "Using PCIe Functions for z/VM Guests"
  - Chapter 16 (new) in CP Planning and Administration

z/VM 6.3 – Base Release

Large Memory Support

- Support for up to **1TB** of real memory (increased from 256GB)
  - Proportionately increases total virtual memory
  - Individual virtual machine limit of **1TB** is unchanged
- Improved efficiency of memory over-commitment
  - Better performance for large virtual machines
  - More virtual machines can be run on a single z/VM image (depending on workload)
- Paging DASD utilization and requirements have changed
  - No longer need to double the paging space on DASD
  - Paging algorithm changes increase the need for a properly configured paging subsystem
- Recommend converting all Expanded Storage to Central Storage
  - Expanded Storage will be used if configured

New Approach: Trial Invalidation



- Page table entry (PTE) contains an “invalid” bit

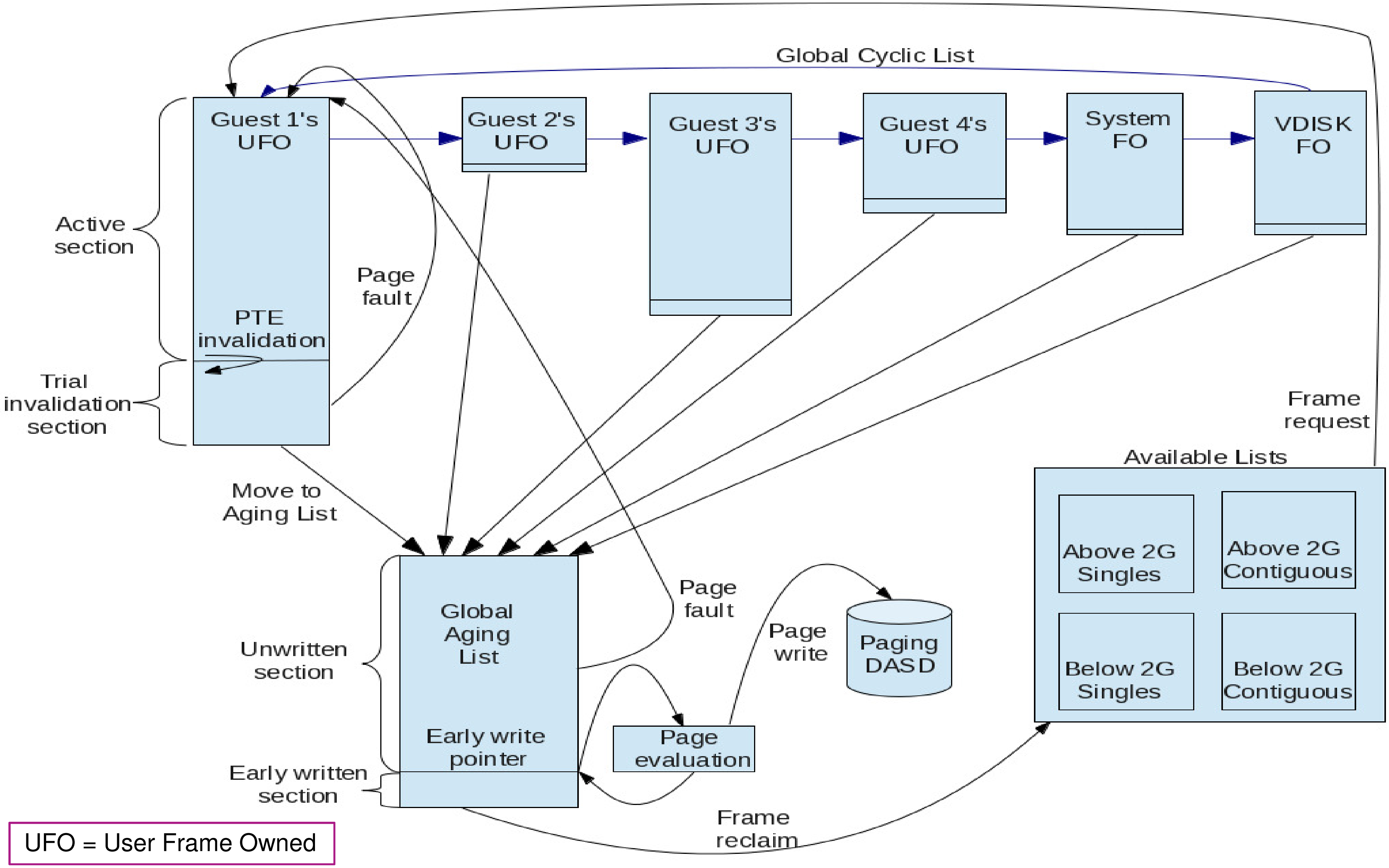
- What if we:
  - Keep the PTE intact but set the “invalid” bit
  - Leave the frame contents intact
  - Wait for the guest to touch the page

- A touch will cause a page fault, but…

- On a fault, there is nothing really to do except:
  - Clear the “invalid” bit
  - Move the frame to the front of the frame list to show that it was recently referenced

- We call this **trial invalidation**.

Memory Management Algorithm Visualization



UFO = User Frame Owned

Large Memory Support: Reserved Storage

- Reserved processing is improved
  - More effective at keeping specified amount of reserved storage in memory

- Pages can be now be reserved for NSS and DCSS as well as virtual machines
  - Set ***after*** **CP SAVESYS** or **SAVESEG** of NSS or DCSS
    - Segment does not need to be loaded in order to reserve it
    - Recommend reserving monitor segment (**MONDCSS**)

- Reserved settings do not survive IPL
  - Recommend automating during system startup

Large Memory Support: Reorder

- Reorder processing has been removed
  - Could cause "stalling" of large virtual machines
  - No longer required with new paging algorithms

- Reorder commands remain for compatibility but have no impact
  - **CP SET REORDER** command gives RC=6005, “not supported”.
  - **CP QUERY REORDER** command says it’s OFF.

- Monitor data is no longer recorded for Reorder

Large Memory Support: Planning DASD Paging Space

- Calculate the sum of:
  - Logged-on virtual machines' primary address spaces, plus…
  - Any data spaces they create, plus…
  - Any VDISKs they use, plus…
  - Total number of shared NSS or DCSS pages, … and then …
  - Multiply this sum by 1.01 to allow for PGMBKs and friends
- Add to that sum:
  - Total number of CP directory pages (reported by DIRECTXA), plus…
  - Min (10% of central, 4 GB) to allow for system-owned virtual pages
- Then multiply by some safety factor (1.25?) to allow for growth or uncertainty
- Remember that your system will take a PGT004 if you run out of paging space
  - Consider using something that alerts on page space, such as Operations Manager for z/VM

Enhanced Dump: Scalability

- Create dumps of real memory configurations up to 1 TB
  - Hard abend dump
  - SNAPDUMP
  - Stand-alone dump

- Performance improvement for hard abend dumps
  - Writes multiple pages of CP Frame Table per I/O
    - CP Frame Table accounts for significant portion of the dump
    - Previously wrote one page per I/O
  - Also improves time required for SNAPDUMPs and Stand-alone dumps

- Recommend allocating enough spool space for 3 dumps
  - See "Allocating Space for CP Hard Abend Dumps" in CP Planning and Administration manual

Enhanced Dump: Utilities

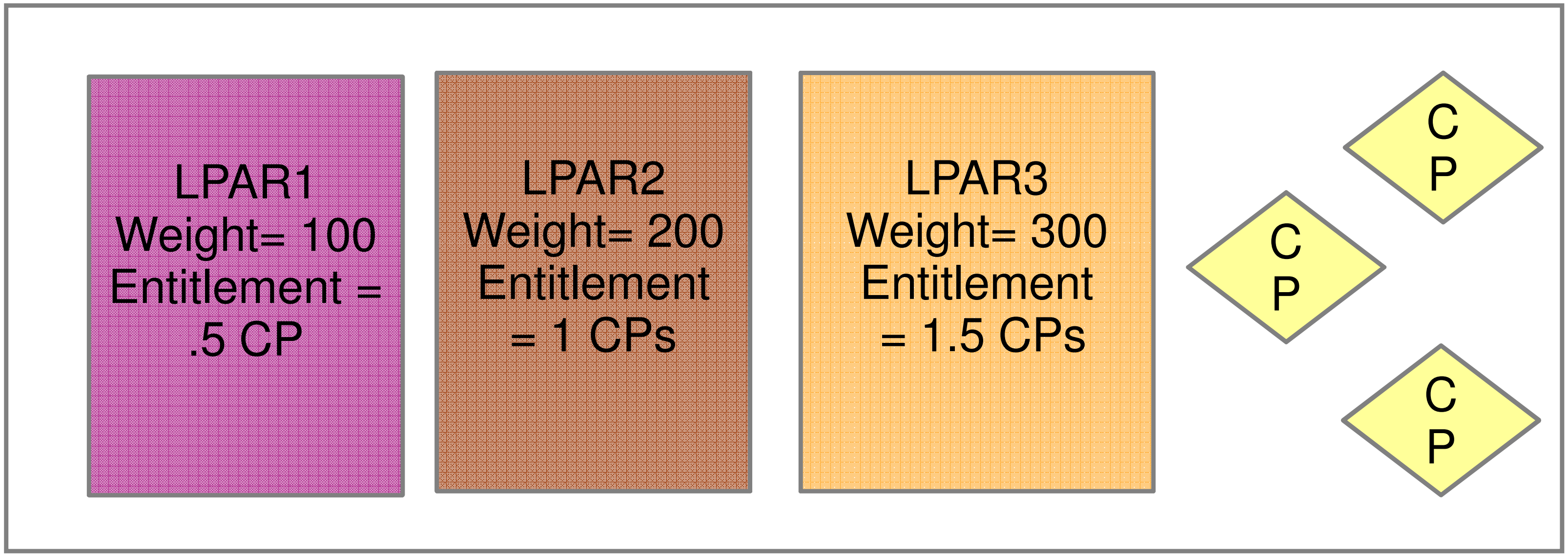
- New Stand-Alone Dump utility
  - Dump is written to disk – either ECKD or SCSI
    - Type of all dump disks must match IPL disk type
    - Dump disks for first level systems must be entire ECKD volumes or SCSI LUNs
    - Dump disks for second level systems may be minidisk "volumes"
  - Creates a CP hard abend format dump
    - Reduces space and time required for stand-alone dump
- **DUMPLD2** utility can now process stand-alone dumps written to disk
- VM Dump Tool supports increased memory size in dumps

HiperDispatch

- Objective: Improve performance of guest workloads
  - z/VM 6.3 communicates with PR/SM to maintain awareness of its partition's topology
    - Partition Entitlement and excess CPU availability
    - Exploit cache-rich system design of System z10 and later machines
  - z/VM polls for topology information/changes every 2 seconds
- Two components
  - Dispatching Affinity
  - Vertical CPU Management
- For most benefit, Global Performance Data (GPD) should be on for the partition
  - Default is ON

HiperDispatch: System z Partition Entitlement

- The allotment of CPU time for a partition
- Function of
  - Partition's weight
  - Weights for all other shared partitions
  - Total number of shared CPUs
- Dedicated partitions
  - Entitlement for each logical CPU = 100% of one real CPU



HiperDispatch: Horizontal Partitions

- Horizontal Polarization Mode
  - Distributes a partition's entitlement evenly across all of its logical CPUs
  - Minimal effort to dispatch logical CPUs on the same (or nearby) real CPUs ("soft" affinity)
    - Affects caches
    - Increases time required to execute a set of related instructions
  - z/VM releases prior to 6.3 always run in this mode

HiperDispatch: Vertical Partitions

- Vertical Polarization Mode
  - Consolidates a partition's entitlement onto a subset of logical CPUs
  - Places logical CPUs topologically near one another
  - Three types of logical CPUs
    - Vertical High (Vh)
    - Vertical Medium (Vm)
    - Vertical Low (Vl)
  - z/VM 6.3 runs in vertical mode by default
    - First level only
    - Mode can be switched between vertical and horizontal
    - Dedicated CPUs are not allowed in vertical mode

HiperDispatch: Partition Entitlement vs. Logical CPU Count

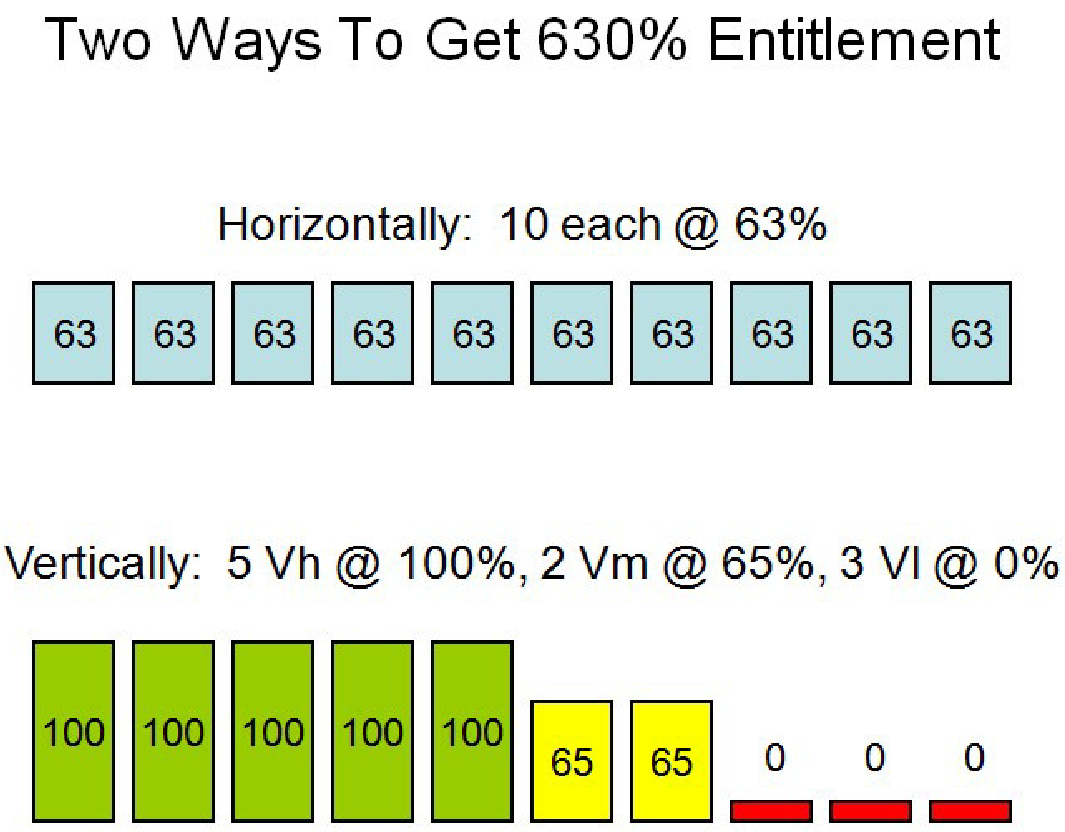
Suppose we have 10 IFLs shared by partitions FRED and BARNEY:

Partition	Weight	Weight Sum	Weight Fraction	Physical Capacity	Entitlement Calculation	Entitlement	Maximum Achievable Utilization
FRED, a logical **10-way**	**63**	100	63/100	1000%	1000% x (63/100)	**630%**	1000%
BARNEY, a logical **8-way**	**37**	100	37/100	1000%	1000% x (37/100)	**370%**	800%

For FRED to run *beyond* **630%** busy, BARNEY has to leave some of its entitlement *unconsumed.*

(CEC’s excess power XP) = (total power TP) - (consumed entitled power EP).

HiperDispatch: Horizontal and Vertical Partitions



Two Ways To Get 630% Entitlement

Horizontally: 10 each @ 63%

Vertically: 5 Vh @ 100%, 2 Vm @ 65%, 3 Vl @ 0%

In vertical partitions:

- Entitlement is distributed unequally among LPUs.

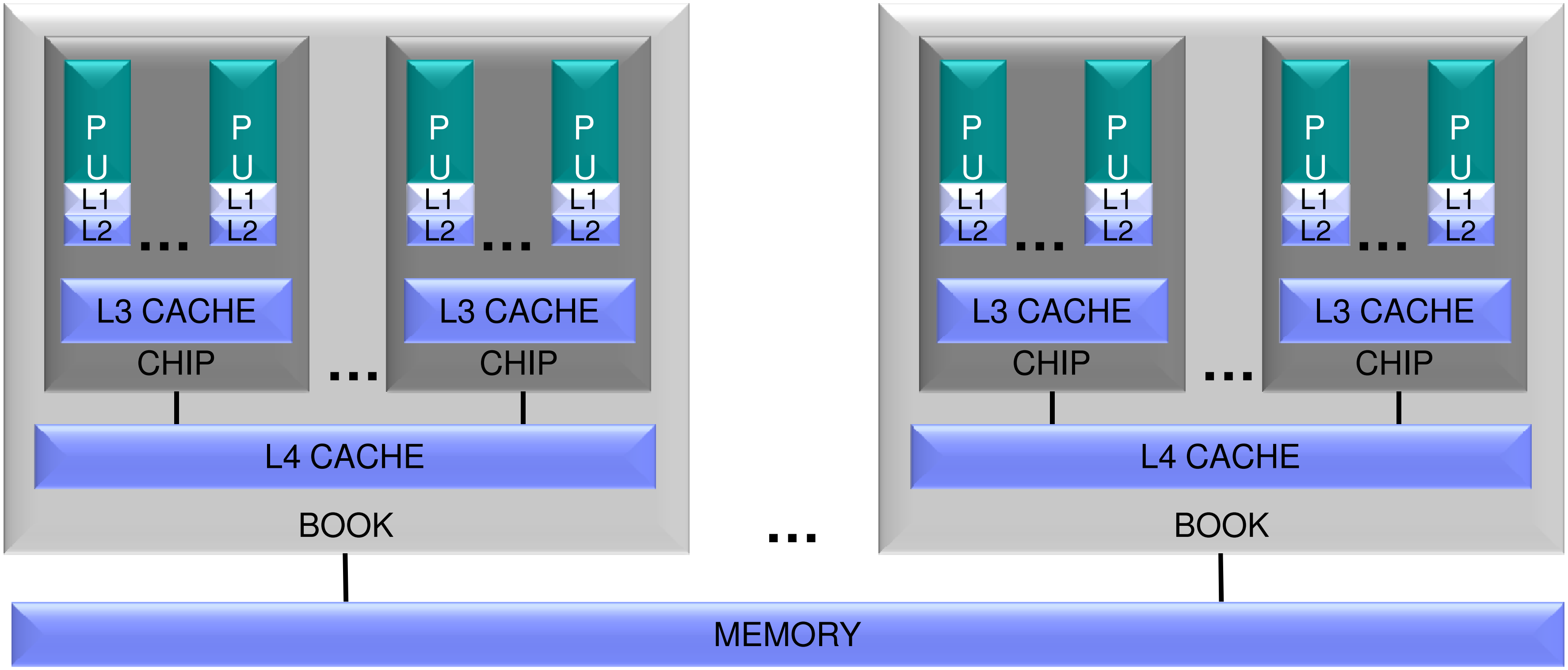
- Unentitled LPUs are useful only when other partitions are not using their entitlements.

- PR/SM tries very hard not to move Vh LPUs.

- PR/SM tries very hard to put the Vh LPUs close to one another.

- Partition consumes its XPF on its Vm and Vl LPUs.

HiperDispatch: Dispatching Affinity



- Processor cache structures have become increasingly complex and critical to performance
- z/VM 6.3 groups together the virtual CPUs of n-way guests
  - Dispatches guests on logical CPUs and in turn real CPUs that share cache
  - Goal is to re-dispatch guest CPUs on same logical CPUs to maximize cache benefits
  - Better use of cache can reduce the execution time of a set of related instructions

HiperDispatch: Parked Logical CPUs

- z/VM automatically parks and unparks logical CPUs
  - Based on usage and topology information
  - Only in vertical mode

- Parked CPUs remain in wait state
  - Still varied on

- Parking/Unparking is faster than **VARY OFF/ON**

HiperDispatch: Checking Parked CPUs and Topology

- **QUERY PROCESSORS** shows PARKED CPUs

```
PROCESSOR nn MASTER type
PROCESSOR nn ALTERNATE type
PROCESSOR nn PARKED type
PROCESSOR nn STANDBY type
```

- **QUERY PROCESSORS TOPOLOGY** shows the partition topology

```
q proc topology
13:14:59 TOPOLOGY
13:14:59   NESTING LEVEL: 02  ID: 01
13:14:59     NESTING LEVEL: 01  ID: 01
13:14:59       PROCESSOR 00  PARKED     CP     VH  0000
13:14:59       PROCESSOR 01  PARKED     CP     VH  0001
13:14:59       PROCESSOR 12  PARKED     CP     VH  0018
13:14:59     NESTING LEVEL: 01  ID: 02
13:14:59       PROCESSOR 0E  MASTER     CP     VH  0014
13:14:59       PROCESSOR 0F  ALTERNATE  CP     VH  0015
13:14:59       PROCESSOR 10  PARKED     CP     VH  0016
13:14:59       PROCESSOR 11  PARKED     CP     VH  0017
                         .
                         .
                         .
13:14:59   NESTING LEVEL: 02  ID: 02
13:14:59     NESTING LEVEL: 01  ID: 02
13:14:59       PROCESSOR 14  PARKED     CP     VM  0020
13:14:59     NESTING LEVEL: 01  ID: 04
13:14:59       PROCESSOR 15  PARKED     CP     VM  0021
13:14:59       PROCESSOR 16  PARKED     CP     VL  0022
13:14:59       PROCESSOR 17  PARKED     CP     VL  0023
```

Additional Information

- z/VM 6.3 resources
  - http://www.vm.ibm.com/zvm630/
  - http://www.vm.ibm.com/zvm630/apars.html
  - http://www.vm.ibm.com/events/

- z/VM 6.3 Performance Report
  - http://www.vm.ibm.com/perf/reports/zvm/html/index.html

- z/VM Library
  - http://www.vm.ibm.com/library/

- Licensing
  - IBM License Metric Tool 9.0.1
    - https://ibm.biz/cpupoolilmt
  - z/VM Software
    - http://www-03.ibm.com/systems/z/resources/swprice/zipla/zvm.html
  - Linux on System z Middleware
    - http://www-03.ibm.com/systems/z/resources/swprice/subcap/linux.html

- Live Virtual Classes for z/VM and Linux
  - http://www.vm.ibm.com/education/lvc/

Thanks!

John Franciscovich
IBM
z/VM Design and Development
Endicott, NY
francisj@us.ibm.com

Session 17515

Appendix A:
z/VM 6.3 Base: Virtual Networking

Virtual Networking: Live Guest Relocation Enhancements

- Live Guest Relocation supports port-based virtual switches
  - – New eligibility checks allow safe relocation of a guest with a port-based VSwitch interface
  - – Prevents relocation of an interface that will be unable to establish proper network connectivity
  - – Adjusts the destination virtual switch configuration, when possible, by inheriting virtual switch authorization from the origin

Virtual Networking: VSwitch Recovery and Stall Prevention

- Initiate controlled port change or failover to a configured OSA backup port
  - Minimal network disruption

- SET VSWITCH UPLINK **SWITCHOVER** command
  - Switch to first available configured backup device
  - Switch to specified backup device
    - Specified RDEV and port number must already be configured as a backup device
  - If backup network connection cannot be established, original connection is reestablished
  - Not valid for a link aggregation or GROUP configured uplink port

Virtual Networking: VSwitch Support for VEPA Mode

- Virtual Edge Port Aggregator (VEPA)
  - IEEE 802.1Qbg standard
  - Provides capability to send all virtual machine traffic to the network switch
  - Moves all frame switching from CP to external switch
  - Relaxes "no reflection" rule
  - Supported on OSA-Express3 and later on zEC12 and later
- Enables switch to monitor and/or control data flow
- z/VM 6.3 support
  - New **VEPA OFF/ON** operand on SET VSWITCH command

Appendix B:
z/VM 6.3 Base: Technology Exploitation

Crypto Express4S

- Available on zEC12 and zBC12

- Supported for z/Architecture guests
  - Authorized in directory (CRYPTO statement)

- Shared or Dedicated access when configured as
  - IBM Common Cryptographic Architecture (CCA) coprocessor
  - Accelerator

- Dedicated access only when configured as
  - IBM Enterprise Public Key Cryptographic Standards (PKCS) #11 (EP11) coprocessor

FCP Data Router (QEBSM)

- Allows guest exploitation of the Data Router facility
  - Provides direct memory access (DMA) between an FCP adapter's SCSI interface and real memory
  - Guest must enable the Multiple Buffer Streaming Facility when establishing its QDIO queues

- **QUERY VIRTUAL FCP** command indicates whether
  - Device is eligible to use Data Router facility
    - `DATA ROUTER ELIGIBLE`
  - Guest requested use of Data Router facility when transferring data
    - `DATA ROUTER ACTIVE`

- Monitor record updated:
  - Domain 1 Record 19 – MRMTRQDC – QDIO Device Configuration Record

FICON DS8000 and MSS Support

- FICON DS8000 Series New Functions
  - Storage Controller Health message
    - New attention message from HW providing more details for conditions in past reflected as Equipment Check.
    - Intended to reduce the number of false HyperSwap events.
  - Peer-to-Peer Remote Copy (PPRC) Summary Unit Check
    - Replaces a series of state change interrupts for individual DASD volumes with a single interrupt per LSS
    - Intended to avoid timeouts in GDPS environments that resulted from the time to process a large number of state change interrupts
- Multiple Subchannel Set (MSS) support for mirrored DASD
  - Support to use MSS facility to allow use of an alternate subchannel set for Peer-to-Peer Remote Copy (PPRC) secondary volumes
  - New **QUERY MSS** command
  - New MSS support cannot be mixed with older z/VM releases in an SSI cluster

Satisfies SODs from October 12, 2011

Appendix C:
z/VM 6.3 Base: Miscellaneous Enhancements

IPL Changes for NSS in a Linux Dump

- Allows contents of NSS to be included in dumps created by stand-alone dump tools such as Linux Disk Dump utility
  - New **NSSDATA** operand on IPL command

- **NSSDATA** can only be used if the NSS:
  - is fully contained within the first extent of guest memory
  - does not contain SW, SN or SC pages
  - is not a VMGROUP NSS

- See http://www.vm.ibm.com/perf/tips/vmdump.html for information on differences between VMDUMP and Linux Disk Dump utility

Specify RDEV for System Volumes

- Prevents wrong volume from being attached when there are multiple volumes with the same volid

- Optionally specify RDEV along with volid in system configuration file
  - **CP_OWNED** statement
  - **USER_VOLUME_RDEV** statement (new)

- If specified, disk volume must match both in order to be brought online

- No volume with specified volid is brought online when
  - Volume at RDEV address has a different volid than specified
  - There is no volume at specified RDEV address

Cross System Extensions (CSE) Withdrawn in z/VM 6.3

- Function has been replaced by z/VM Single System Image (VMSSI) feature
  - **XSPOOL** … commands no longer accepted
  - **XSPOOL_** … configuration statements not processed (tolerated)
- CSE cross-system link function is still supported
  - **XLINK** … commands
  - **XLINK_** … configuration statements
- CSE XLINK and SSI shared minidisk cannot be used in same cluster
- Satisfies Statement of Direction (October 12, 2011)

OVERRIDE Utility and UCR Function Withdrawn

- "Very OLD" method for redefining privilege classes for
  - CP Commands
  - Diagnose codes
  - other CP functions

- To redefine privilege classes, use
  - **MODIFY COMMAND** command and configuration statement
  - **MODIFY PRIV_CLASSES** command and configuration statement

- Satisfies Statement of Direction (October 12, 2011)