F0

What's New in the z/VM 6.3 Hypervisor

Session 16482

John Franciscovich

IBM: z/VM Development, Endicott, NY

#SHAREorg

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

SHARE in Seattle 2015

Copyright (c) 2014 by SHARE Inc. Except where otherwise noted, this work is licensed under http://creativecommons.org/licenses/by-nc-sa/3.0/

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2015 IBM Corporation

Expanding the Horizon of Virtualization

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

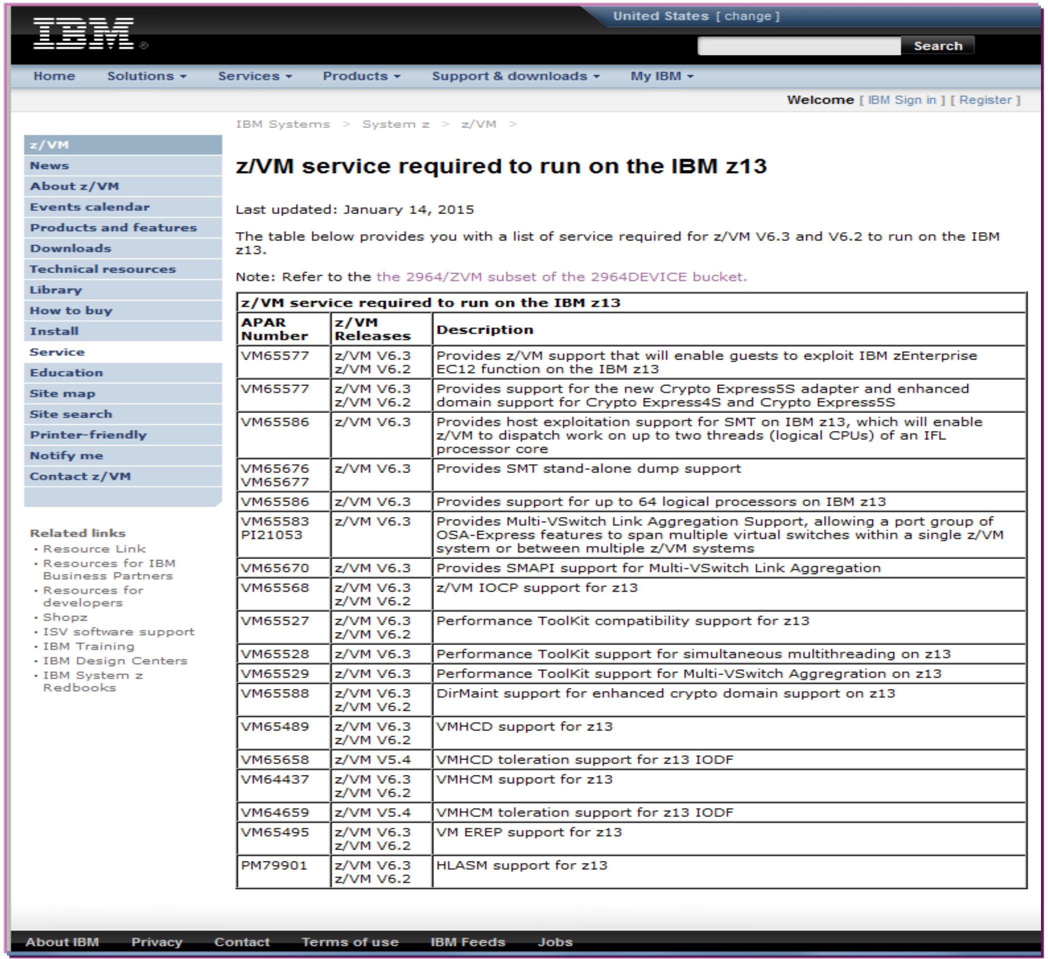
- z/VM Compatibility Support
  - PTFs available February 13, 2015
  - Also includes Crypto enhanced domain support
  - z/VM 6.2 and z/VM 6.3
  - No z/VM 5.4 support
  - Refer to bucket for full list

- 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)

© 2015 IBM Corporation

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: January 14, 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.

Note: Refer to the the 2964/ZVM subset of the 2964DEVICE bucket.

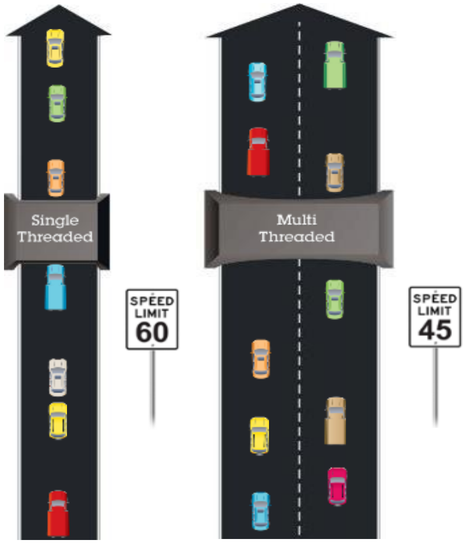
z/VM service required to run on the IBM z13

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	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 SMT stand-alone dump support
VM65586	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 z/VM V6.2	Performance ToolKit compatibility support for z13
VM65528	z/VM V6.3	Performance ToolKit support for simultaneous multithreading on z13
VM65529	z/VM V6.3	Performance ToolKit support for Multi-VSwitch Aggregration 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

© 2015 IBM Corporation

Simultaneous Multithreading (SMT) on z/VM

- Increase processing efficiency and overall throughput for many workloads
  - Amount of benefit for different workloads **will** vary
- Objective is to improve capacity, not performance.
- Allows z/VM to dispatch work on up to two threads of a z13 IFL
  - Up to 32 IFLs (cores) supported
  - z/VM support is only for IFLs
- Transparent to virtual machine
  - Guest does not need to be SMT aware
  - SMT is not virtualized to the guest
- SMT is disabled by default
  - Requires a System Configuration setting and re-IPL
  - When enabled, applies to the entire system
- VM65586 for z/VM 6.3 only
  - PTFs planned to be available March 13, 2015



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

**Illustrative numbers only*

© 2015 IBM Corporation

Cores, Threads, and Logical Processors (CPUs)

- Core
  - When multi-threading is enabled (IFL only)
    - Contains multiple threads, each the equivalent of a logical processor (CPU)
    - Each core has same number of threads
  - When multi-threading is not installed or not enabled
    - Equivalent of a single logical processor

- Thread
  - When multi-threading is enabled
    - Synonymous with logical processor that is a member of a core
    - Each thread within a core has same processor type and polarization
    - Threads within a core share some hardware resources
      - Execution of one thread in a core can affect performance of other threads in the same core
  - When multi-threading is not installed or not enabled
    - Each core runs a single thread (logical processor)

- Logical Processor (CPU)
  - A thread in a core
  - All of the architected resources available to programs

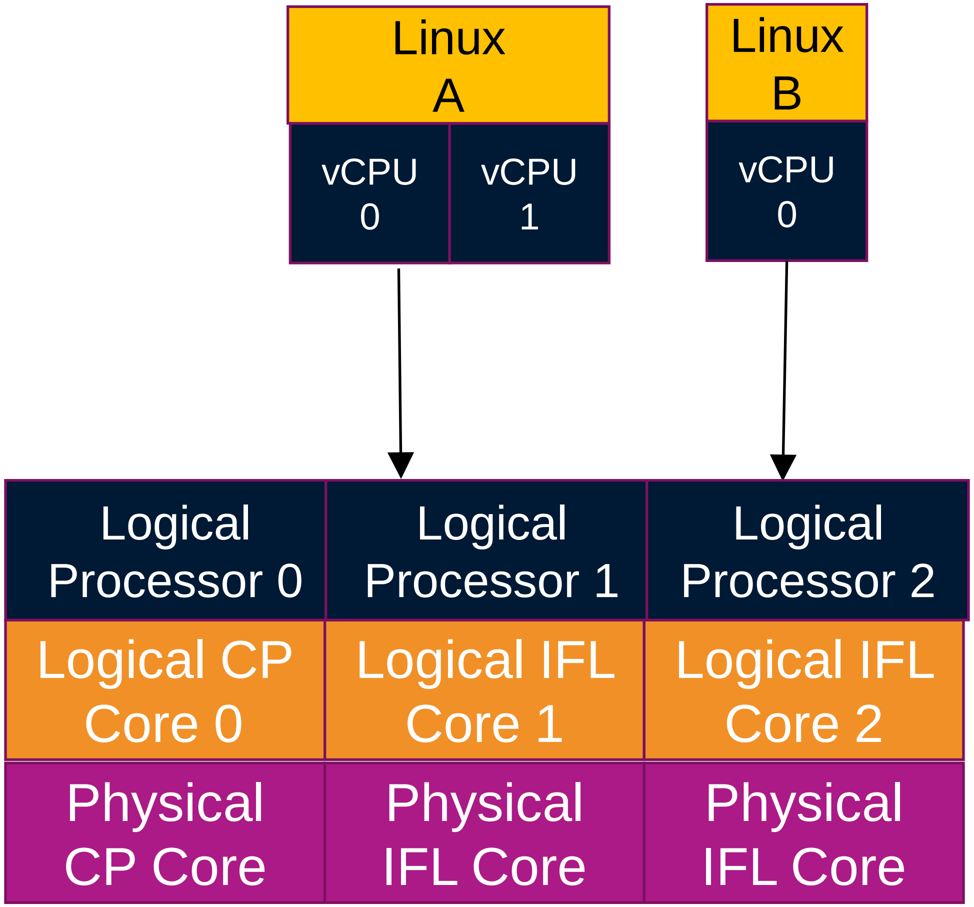
© 2015 IBM Corporation

Dispatching: SMT Disabled

- Physical IFLs and CPs are single-threaded

- z/VM creates a logical processor (CPU) for each associated logical IFL or CP

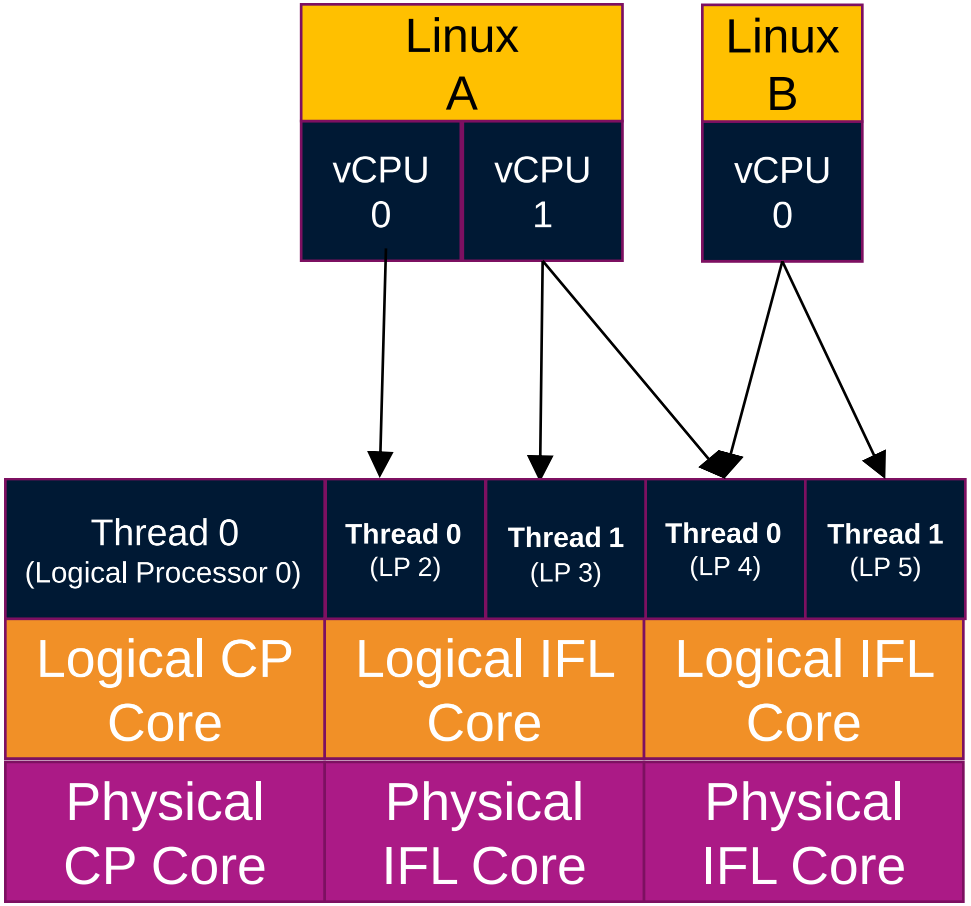
- The virtual processors of guests are dispatched on individual logical processors (CPUs)



© 2013, 2015 IBM Corporation

Dispatching: SMT Enabled

- Physical IFLs (or Cores) with SMT allow up to two threads to be used.
- Logical IFLs are presented to z/VM as in the past.
- z/VM creates a logical processor (CPU) associated with each thread for it to use.
- The virtual CPUs of guests can then be dispatched on different threads intelligently, based on topology information.
- In a mixed-engine environment, general purpose processors can not do threading, but a second logical processor address is consumed (LP 1 in example)



© 2013, 2015 IBM Corporation

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

- Add the **MULTITHreading ENAble** statement to your SYSTEM CONFIG
  - maximum number of threads for all core types
  - different number of threads for each type
    - z/VM only supports IFL cores for multithreading.

- 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.

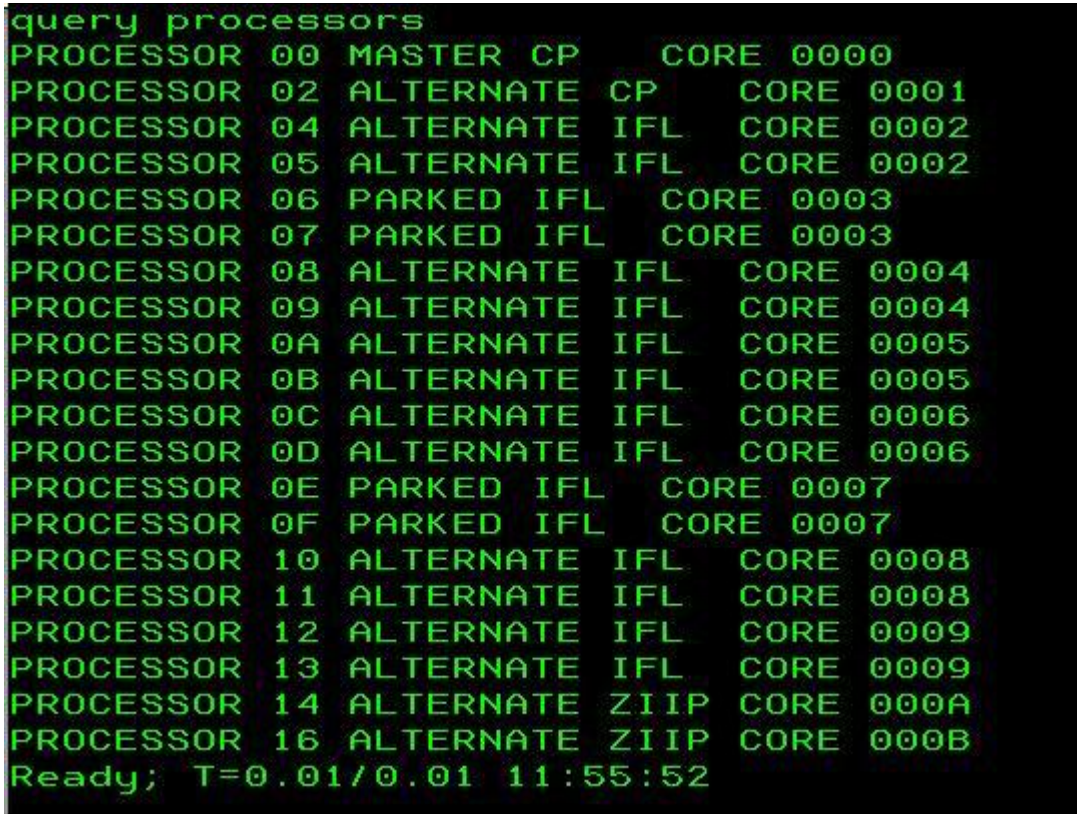
- The system must be using the *reshuffle dispatcher method* (this is the default)
  - Make sure you ***don't*** have an **SRM DSPWDMethod REBALANCE** statement in your SYSTEM CONFIG.

- Re-IPL your system!

© 2015 IBM Corporation

QUERY PROCessors with SMT

- Shows which core each thread/processor is on:



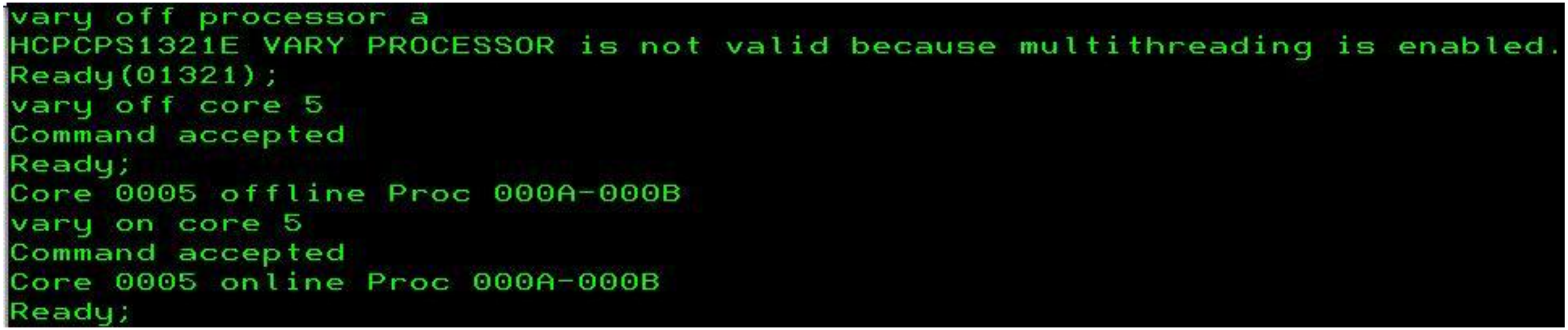
```
query processors
PROCESSOR 00 MASTER CP    CORE 0000
PROCESSOR 02 ALTERNATE CP    CORE 0001
PROCESSOR 04 ALTERNATE IFL   CORE 0002
PROCESSOR 05 ALTERNATE IFL   CORE 0002
PROCESSOR 06 PARKED IFL   CORE 0003
PROCESSOR 07 PARKED IFL   CORE 0003
PROCESSOR 08 ALTERNATE IFL   CORE 0004
PROCESSOR 09 ALTERNATE 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
```

© 2015 IBM Corporation

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;
```

© 2015 IBM Corporation

Processor Time Reporting

- **Raw time** (the old way, but with new implications)
  - Time each virtual CPU is dispatched on a thread
    - Includes time thread is not doing work
  - The only time measurement 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 be if SMT were disabled
    - Adjusts CPU time for each thread

© 2015 IBM Corporation

Prorated Core Time (availability TBD)

- Prorated core time will divide the time a core is dispatched evenly among the threads dispatched in that interval
  - 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** option
  - Limited by the **SET SHARE LIMITHARD** command

  (currently raw time is used; raw time will continue to be used when SMT is disabled)

- **QUERY CPUPOOL** will show capacity in cores instead of CPUs

- Prorated core time will be reported in monitor records and the new Type F accounting record.

- Watch for APAR VM65680

© 2015 IBM Corporation

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!

© 2015 IBM Corporation

Increased CPU Scalability

- Various improvements to allow z/VM systems to be larger in terms of processors and more efficient, improving the n-way curve

- For z13
  - With SMT disabled, increases logical processors supported from 32 to 64
  - With SMT enabled, the limit is 32 IFLs (64 threads)

- For processors prior to z13
  - Limit remains at 32
  - May still benefit from improved n-way curves

- APAR VM65586 for z/VM 6.3 **only**
  - PTFs planned to be available March 13, 2015

© 2015 IBM Corporation

Areas Improved to Increase CPU Scalability

- Improvements were made to the following areas to improve efficiency and reduce contention
  - Scheduler lock
  - VSwitch data transfer buffers
  - Serialization and processing of VDisk I/Os
  - Memory Management

- Some areas needing improvement were known – others required thorough investigation and experimentation

- All tested workloads now show acceptable scaling up to a 64-way LPAR
  - Benefits are workload dependent

- Changes to processor parking algorithms

© 2015 IBM Corporation

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

Crypto Express5S

- Available on IBM z13
- Can be configured one of 3 ways
  - Shared or dedicated access
    1. IBM Common Cryptographic Architecture (CCA) coprocessor
    2. Accelerator
  - Dedicated access only
    3. IBM Enterprise Public Key Cryptographic Standards (PKCS) #11 (EP11) coprocessor
- Enhanced domain support
  - z13 supports up to 16 APs and up to 85 domains per AP
  - z/VM supports architected increases in
    - Maximum number of crypto features (APs) from 64 to 256
    - Maximum number of domains per AP from 16 to 256

© 2015 IBM Corporation

2014 Enhancements

© 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

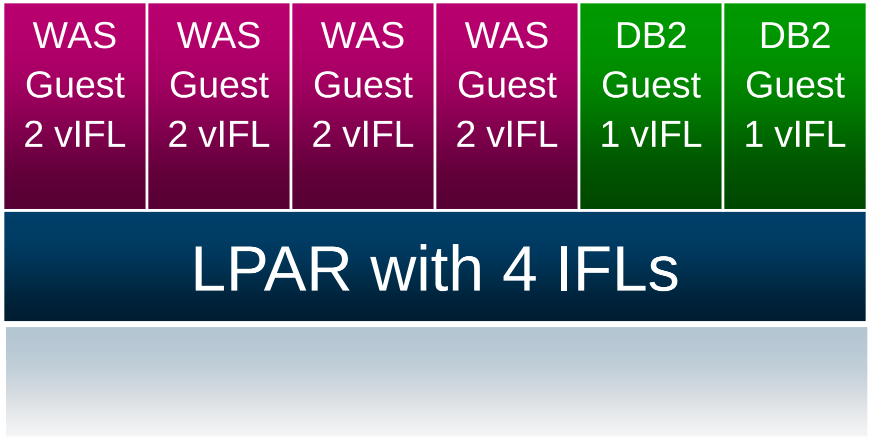
```
query cpupool all
CPU pool  Limit     Type      Members
LINUXP2  8.0 CPUs    IFL            0
CPPOOL10  12 %        CP            8
LINUXP3   30 %       IFL           20
LINUXP1  2.5 CPUs    IFL            6

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
```

© 2013, 2015 IBM Corporation

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

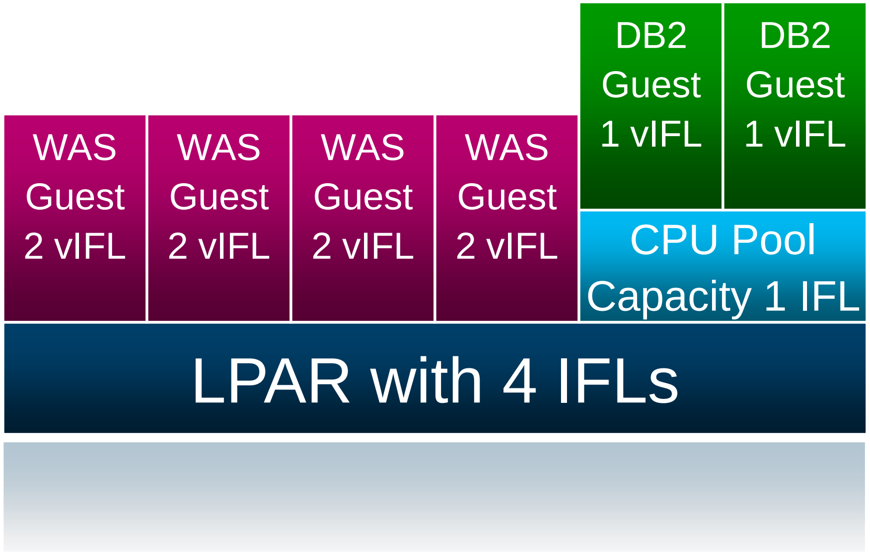
DB2 Guest 1 vIFL

LPAR with 4 IFLs

© 2013, 2015 IBM Corporation

Add New Workload: With CPU Pooling

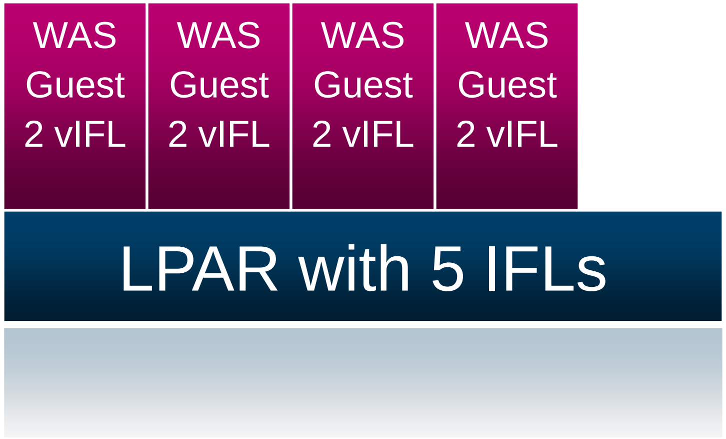
- 4 production guests for WAS
  - May consume up to 4 engines
- Create a 1-IFL pool
- Put the 2 DB2 production guests in the pool
  - DB2 is limited to 1 engine instead of 2



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

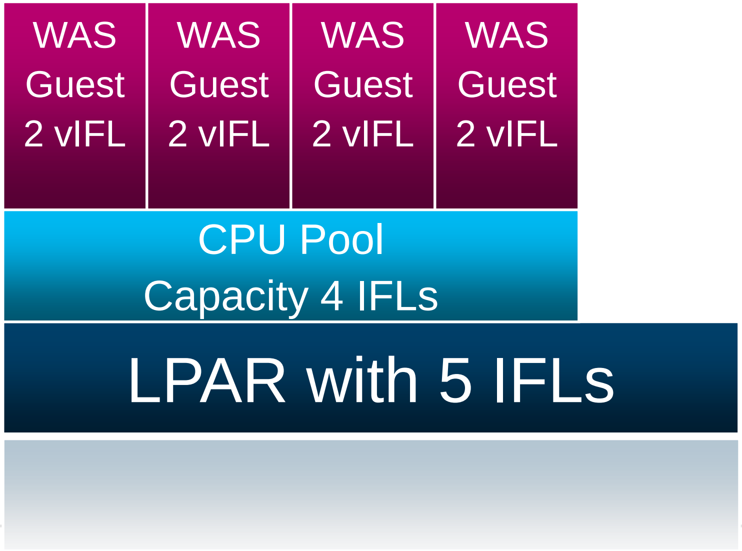
Add Capacity: Without CPU Pooling

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



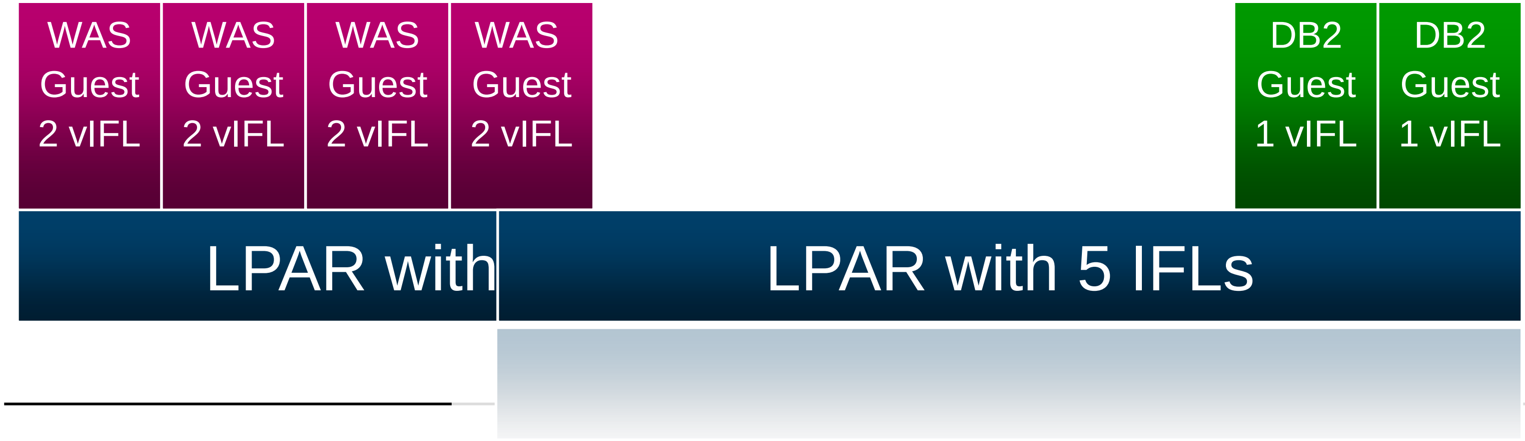
Add Capacity: With CPU Pooling

- 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



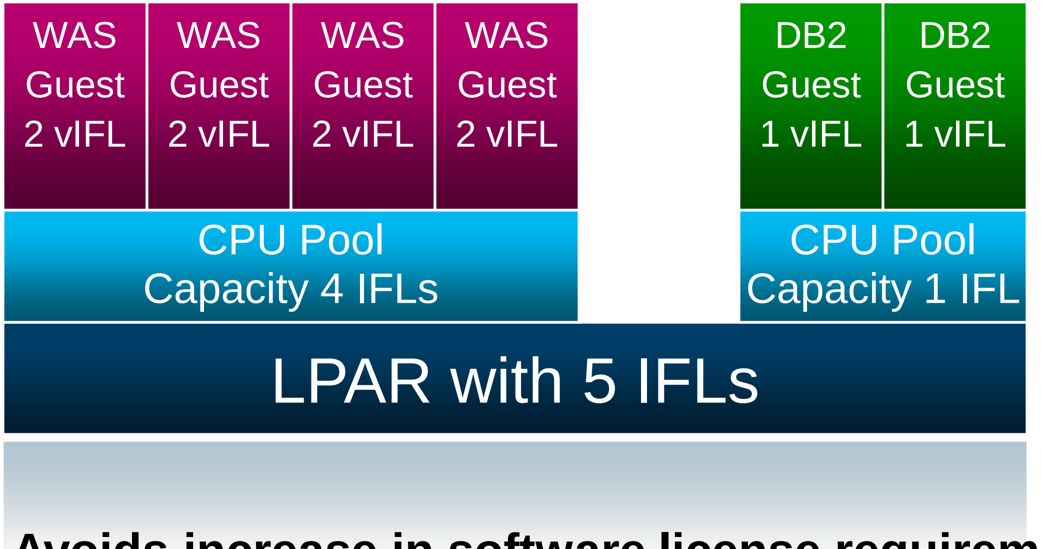
Combine LPARs: Without CPU Pooling

- 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

© 2013, 2015 IBM Corporation

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
```

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

z/VM 6.3 – Base Release

© 2013, 2015 IBM Corporation

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

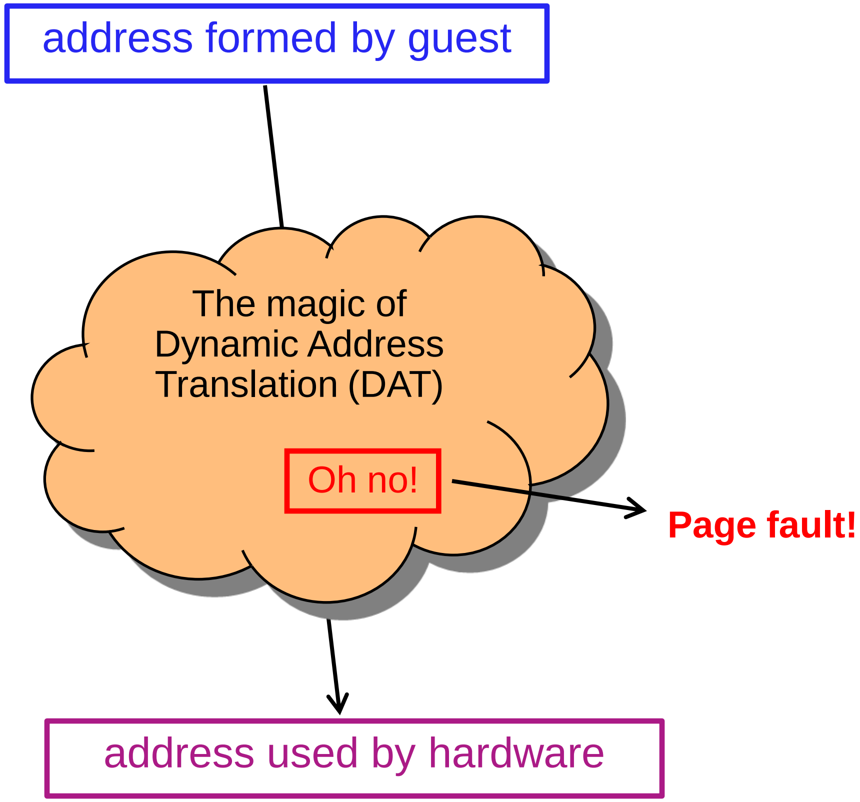
- Improved efficiency of memory over-commitment
  - Better performance for large virtual machines
  - More virtual machines can be run on a single z/VM image (depending on workload)

- 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

© 2013, 2015 IBM Corporation

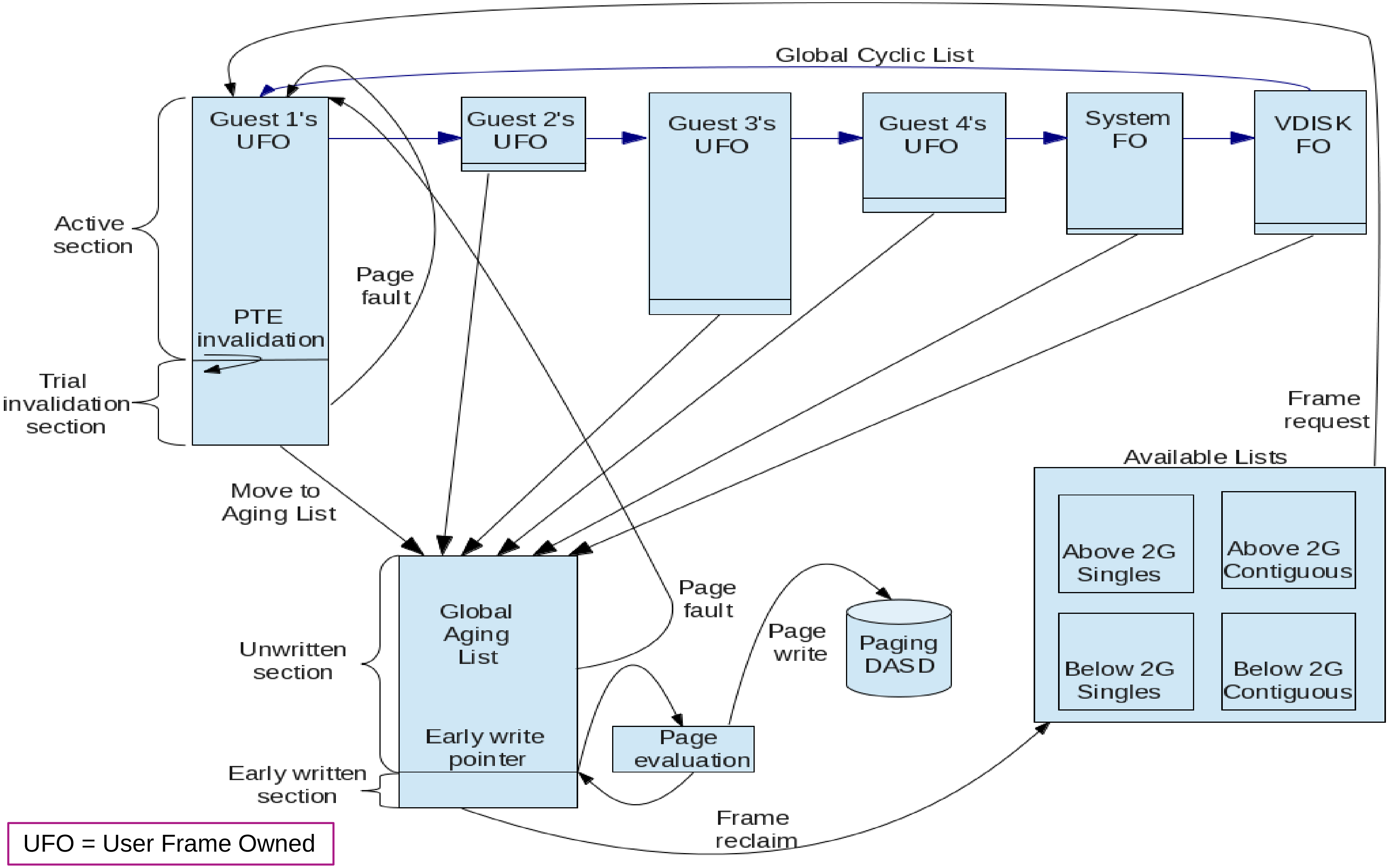
New Approach: Trial Invalidation



- 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**.

© 2013, 2015 IBM Corporation

Memory Management Algorithm Visualization



UFO = User Frame Owned

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

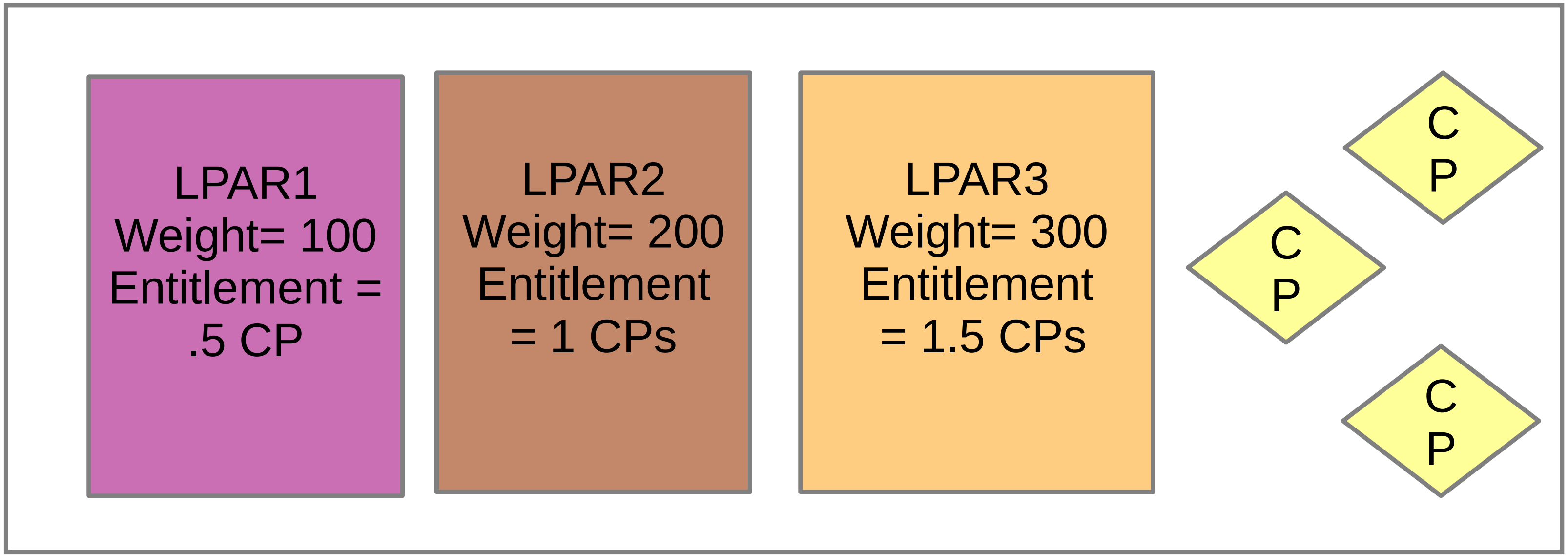
HiperDispatch

- 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

© 2013, 2015 IBM Corporation

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



© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

HiperDispatch: Partition Entitlement vs. Logical CPU Count

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

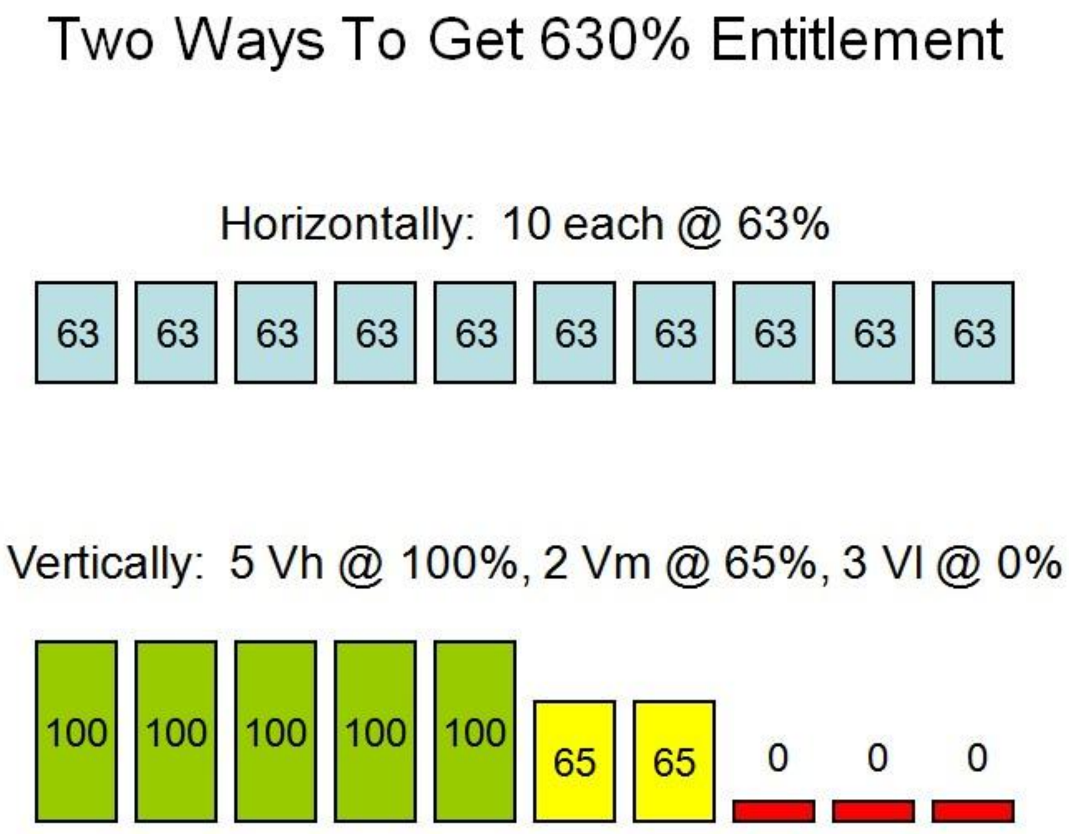
Partition	Weight	Weight Sum	Weight Fraction	Physical Capacity	Entitlement Calculation	Entitlement	Maximum Achievable Utilization
FRED, a logical **10-way**	**63**	100	63/100	1000%	1000% x (63/100)	**630%**	1000%
BARNEY, a logical **8-way**	**37**	100	37/100	1000%	1000% x (37/100)	**370%**	800%

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).

© 2013, 2015 IBM Corporation

HiperDispatch: Horizontal and Vertical Partitions



In vertical partitions:

- Entitlement is distributed unequally among LPUs.

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

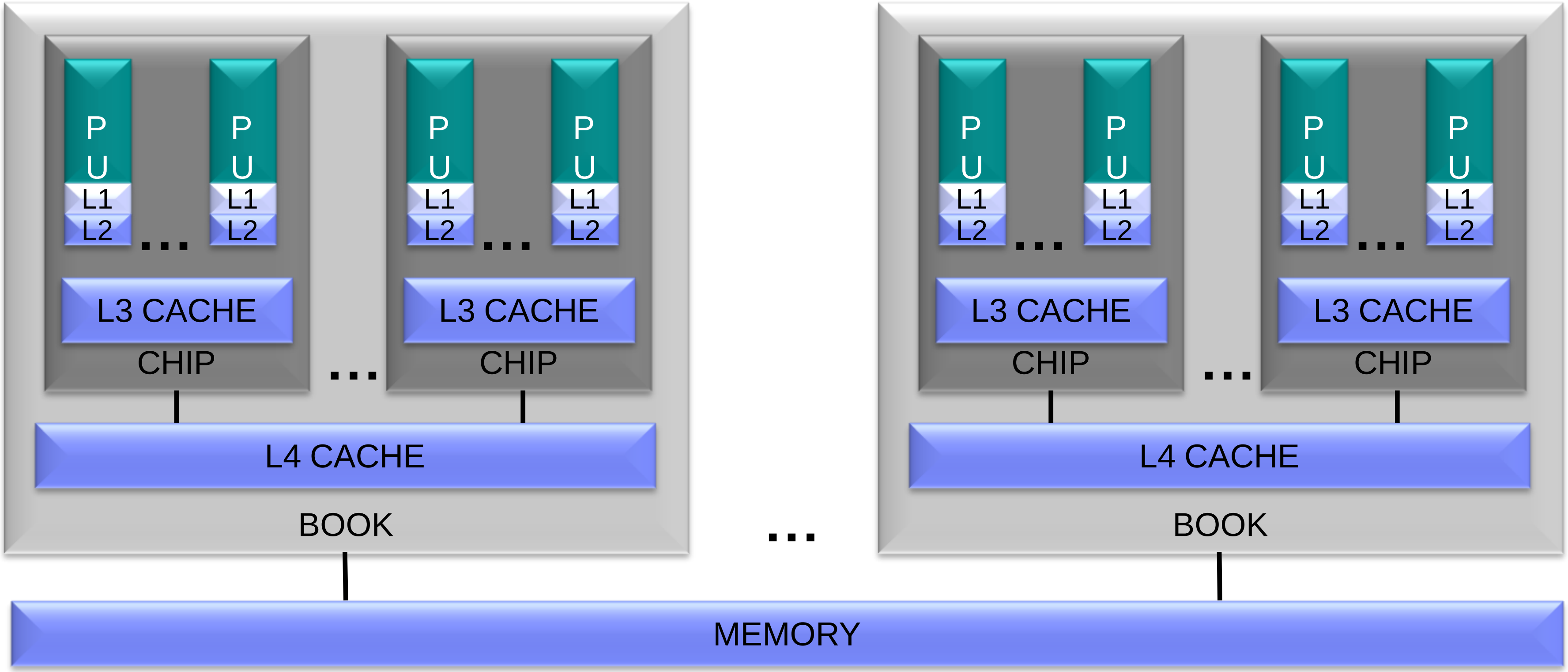
- PR/SM tries very hard not to move Vh LPUs.

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

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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**

© 2013, 2015 IBM Corporation

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
```

© 2013, 2015 IBM Corporation

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/

© 2013, 2015 IBM Corporation

Thanks!

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

Session 16482

Appendix A:
z/VM 6.3 Base: Virtual Networking

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

Appendix B:
z/VM 6.3 Base: Technology Exploitation

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

Appendix C:
z/VM 6.3 Base: Miscellaneous Enhancements

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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

© 2013, 2015 IBM Corporation

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)

© 2013, 2015 IBM Corporation

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)

© 2013, 2015 IBM Corporation