Aktuelle Informationen zu z/VM

Dr. Manfred Gnirss
gnirss@de.ibm.com

Frank Heimes
frank.heimes@de.ibm.com

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
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*

* 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.

Acknowledgments – Platform Update Team

- Kevin Adamslan
- Alan Altmark
- Bill Bitner
- Miguel Delapaz
- Glenda Ford
- John Franciscovich
- Les Geer
- Susan Greenlee
- Dan Griffith
- Brian Hugenbruch
- Emily Hugenbruch
- Mark Lorenc
- Brian Wade

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

Topics

Teil 1 – Manfred Gnirss:

- Scalability
  - Large Memory Support
- HiperDispatch
- Miscellaneous Enhancements

Teil 2: Frank Heimes:

- IBM Wave

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

Scalability –
Large Memory Support

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

- Eliminate use of expanded storage for z/VM paging, allowing greater flexibility
  - Recommend converting all Expanded Storage to Central Storage
  - Expanded Storage will be used if configured (up to 128 GB)
  - SOD: z/VM 6.3 will be the last to support expanded storage

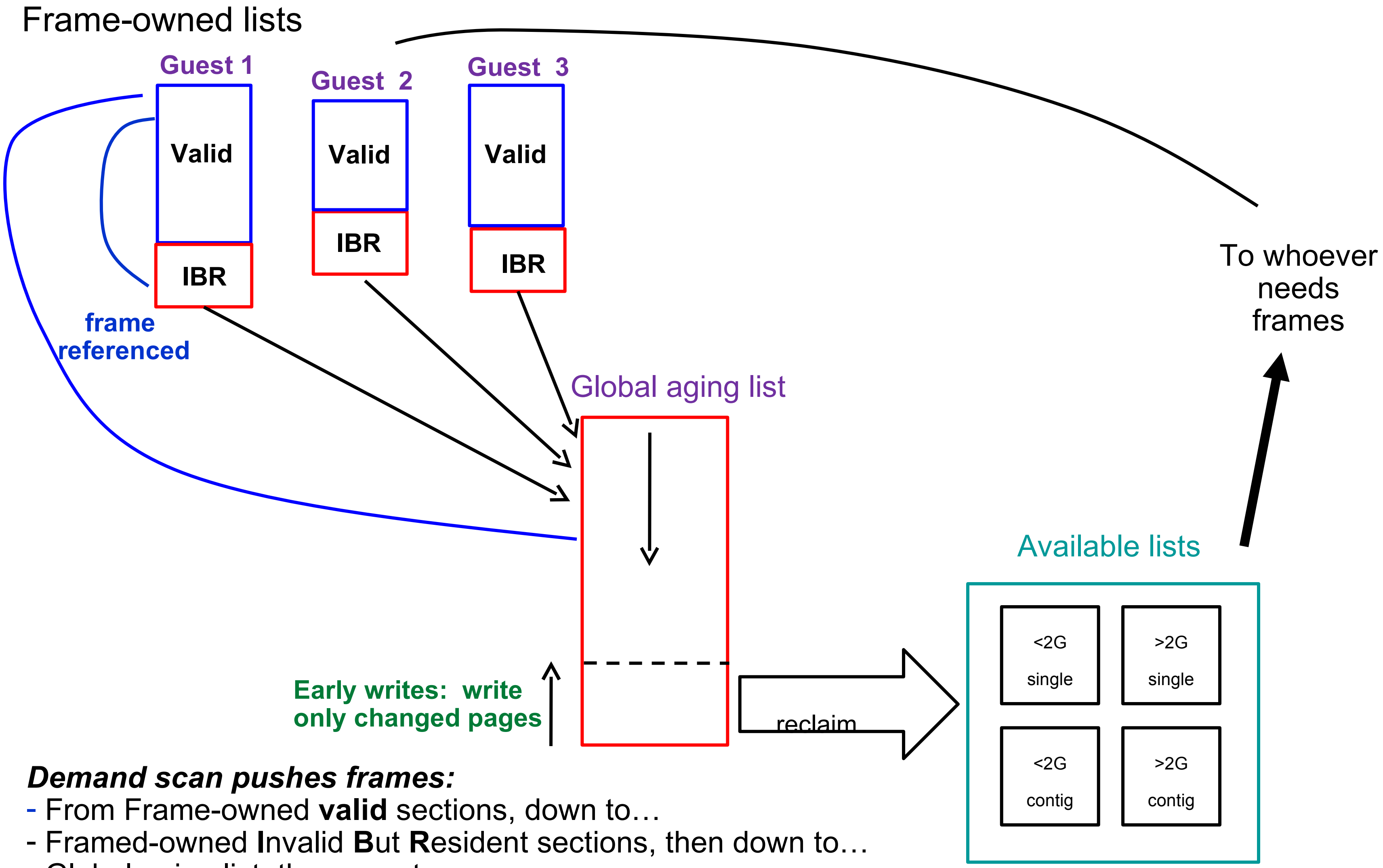
Large Memory Support: Reserved Storage

- 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: The Big State Diagram



Demand scan pushes frames:

- From Frame-owned **valid** sections, down to…
- Framed-owned **I**nvalid **B**ut **R**esident sections, then down to…
- Global aging list, then over to…
- Available lists, from which they...
- Are used to satisfy requests for frames

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: New and Changed Commands

- New commands to **SET** and **QUERY AGELIST** attributes
  - Size
  - Early Writes

- Enhanced **SET RESERVED** command
  - Reserve pages for NSS and DCSS
  - Define ***number of frames*** or ***storage size*** to be reserved
  - Define maximum amount of storage that can be reserved for system
  - **QUERY RESERVED** command enhanced to show information about above

- **STORAGE** config statement enhanced to set AGELIST and maximum reserved storage

- **INDICATE** commands
  - New "instantiated" pages count where appropriate

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

HiperDispatch

HiperDispatch

- Objective: Improve performance of guest workloads
  - z/VM 6.3 communicates with PR/SM to maintain awareness of its partition's topology to improve processor efficiency
    - Partition Entitlement and excess CPU availability
    - Better use of processor cache
    - Exploit cache-rich system design of System z10 and later machines

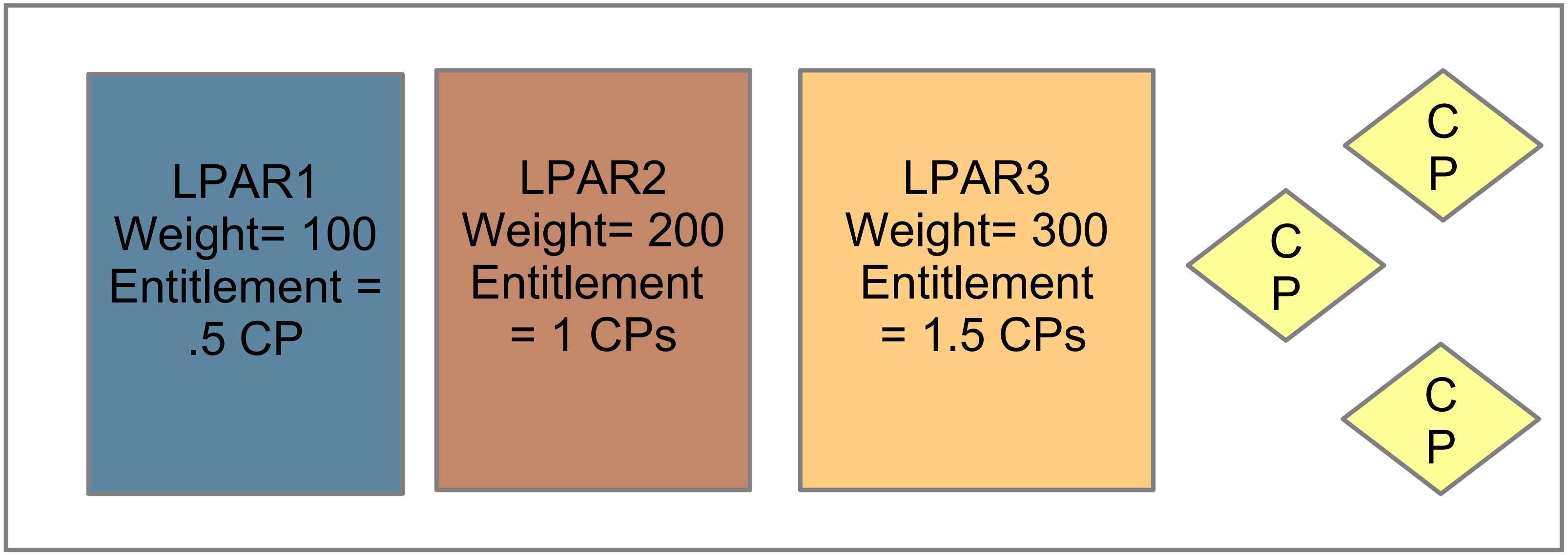
    z/VM polls for topology information/changes every 2 seconds

- 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 of the z/VM LPAR
  - 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
    - “soft” affinity to dispatch virtual CPUs
    - No awareness of chip or book

HiperDispatch: Vertical Partitions

- *Vertical Polarization Mode*
  - Consolidates a partition's entitlement onto a subset of logical CPUs (attempts to minimize the number of logical processors, allowing LPAR to similarly manage 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%

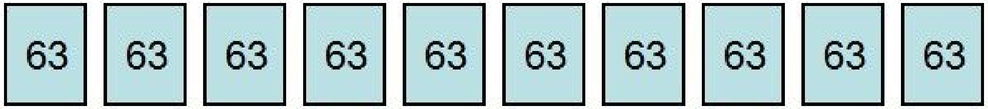
For FRED to run *beyond* **630%** busy, BARNEY has to leave some of its entitlement *unconsumed.*

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

HiperDispatch: Horizontal and Vertical Partitions

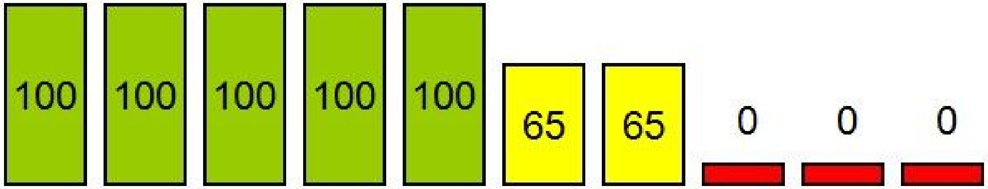
Two Ways To Get 630% Entitlement

Horizontally: 10 each @ 63%



- The logical processors are all created/treated equally.
- z/VM dispatches work evenly across the logical processors

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



- The logical processors are skewed to where some get greater share of the weight.
- z/VM dispatches work accordingly to the heavier weighted workload.

In vertical partitions:

- Entitlement is distributed unequally among LPUs.

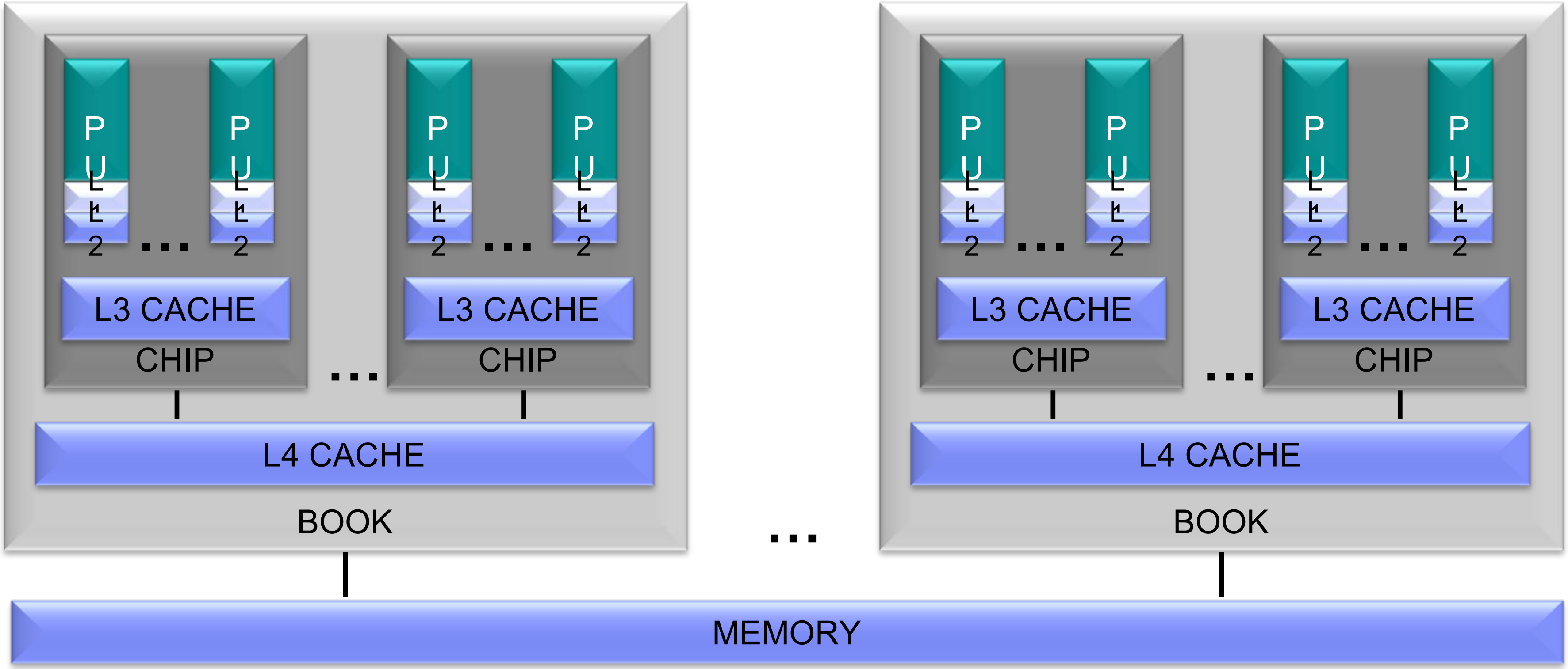
- Unentitled LPUs are useful only when other partitions are not using their entitlements.

- 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
- Goal is to re-dispatch work close (in terms of topology) to where it last ran
- Dispatcher is aware of the cache and memory topology. Dispatch virtual CPU near where its data may be in cache based on where the virtual CPU was last dispatched
- 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 – Vertical CPU Management

- Better use of cache can reduce the execution time of a set of related instructions

Especially:

- When CEC is constrained, the LPAR's entitlement is reduced to fewer IFLs
- z/VM and LPAR will cooperate
  - z/VM will concentrate the workload on a smaller number of logical processors
  - LPAR will redistribute the partition weight to give a greater portion to this smaller number of logical processors

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
```

HiperDispatch: New and Changed Commands

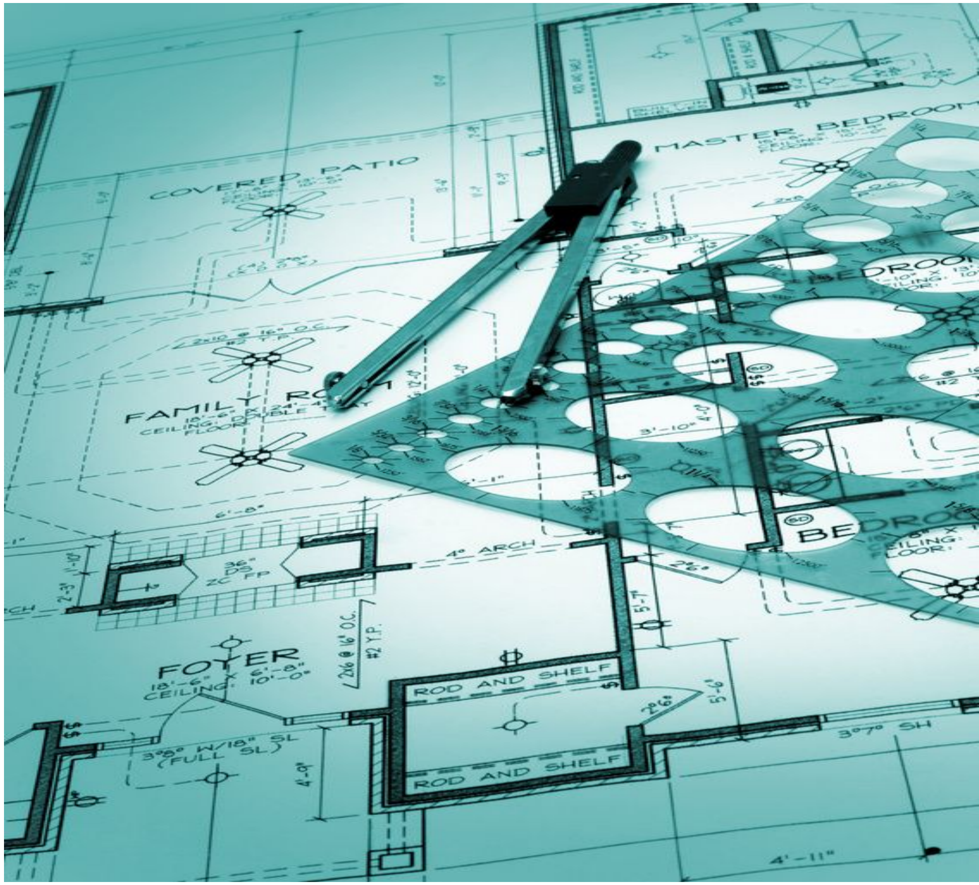
- **INDICATE LOAD**
  - AVGPROC now represents average value of the portion of a real CPU that each logical CPU has consumed

- **SET SRM** command can be used to change default settings for attributes related to HiperDispatch
  - Review monitor data and/or performance reports before changing

z/VM 6.3 Enhancements: Available June 27, 2014

Enhancing the Foundation for Virtualization

- Release for Announcement – zBX and zEnterprise System Enhancements
  - February 24, 2014
- Software Enhancements
  - CPU Pooling
  - Environment Information Interface
- Hardware Support
  - 10GbE RoCE Express Feature
  - zEDC Express Feature
- Available June 27, 2014



z/VM 6.3 Enhancements: CPU Pooling

- Fine grain CPU limiting for a group of virtual machines

- 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 identical pools on all members of cluster

- Support Details
  - z/VM 6.3 with APAR VM65418 - June 27, 2014

z/VM 6.3 Enhancements: CPU Pooling

- 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 **SET CPUPOOL** command
- Assign and remove guests to/from a CPU pool with the **SCHEDULE** command

z/VM 6.3 Enhancements: Environment Information Interface

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

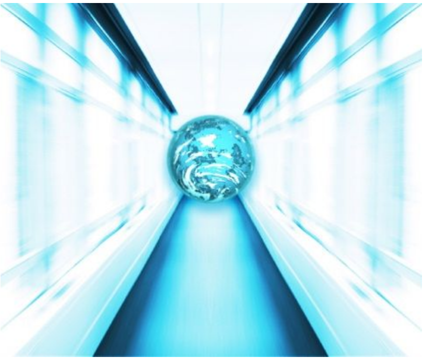
- New problem statement instruction STore HYpervisor Information (STHYI)
  - Supported by z/VM 6.3
  - Tolerated by z/VM 6.2 ("function not supported")

- Includes support for CPU Pooling enhancement

- Foundation for future software licensing tools

- Support details:
  - z/VM 6.3 with APAR VM65419 – June 27, 2014

10GbE RoCE Express Feature



- Support for RDMA over Converged Ethernet for guests
- Based on new hypervisor PCIe support
- Designed to support z/OS's Shared Memory Communications-Remote Direct Memory Access (SMC-R) in z/OS V2.1
- Helps reduce CPU resource consumption
- Support details:
  - IBM zEC12 or zBC12 with appropriate millicode (driver 15)
  - z/VM 6.3 with APAR VM65417 – June 27, 2014
  - z/OS 1.12, z/OS 1.13, z/OS 2.1 with APAR OA43256
  - Fulfills 2013 Statement of Direction

zEDC Express Feature

- Guest support for zEDC Express Feature
- High performance, low latency, low CPU consumption compression
- Possible disk utilization reduction
- Support details:
  - IBM zEC12 or zBC12 with appropriate millicode (driver 15)
  - z/VM 6.3 with APAR VM65417 – June 27, 2014
  - z/OS 1.12, z/OS 1.13, z/OS 2.1 with APAR OA43256
  - z/OS 1.12, z/OS 1.13, z/OS 2.1 with APAR OA44482
  - Fulfills 2013 Statement of Direction

z/VM 6.3 Enhancements: PCIe Support

- 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

- **DEFINE PCIFUNCTION** defines real PCI function to I/O configuration
  - PCI Function ID (PFID) used to reference a function

- Basis for support for guest exploitation of 10GbE RoCE Express feature
  - Supports z/OS's Shared Memory Communications-Remote Direct Memory Access (SMC-R) in z/OS V2.1

- Support details:
  - IBM zEC12 or zBC12 with appropriate millicode (driver 15)
  - z/VM 6.3 with APAR VM65417 – June 27, 2014
  - z/OS 1.12, z/OS 1.13, z/OS 2.1 with APAR OA43256
  - Fulfills 2013 Statement of Direction

Fragen?

Dr. Manfred Gnirss

Senior IT Specialist

IBM Client Center

IBM Germany Lab

Schönaicher Strasse 220
71032 Böblingen, Germany

Phone +49 (0)7031-16-4093
gnirss@de.ibm.com

More Information

z/VM 6.3 resources

http://www.vm.ibm.com/zvm630/

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/

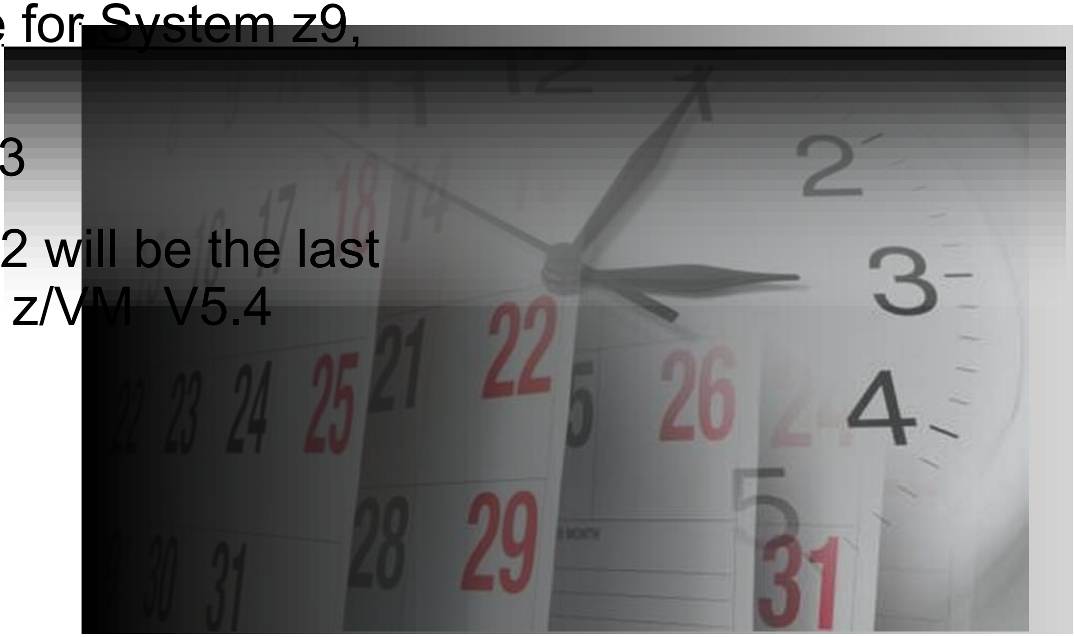
Live Virtual Classes for z/VM and Linux

http://www.vm.ibm.com/education/lvc/

z/VM Version 5 Release 4

- The last release of z/VM to support IBM System z9® and older processors
  - **No longer available as of March 12, 2012**
  - Also supports the IBM zEnterprise® EC12 (zEC12) and IBM zEnterprise BC12 (zBC12)

- End of Service was been extended to **December 31, 2014** or end of IBM service for System z9, whichever is ***later***
  - Statement of Direction 2013
    - The zEC12 and zBC12 will be the last processors to support z/VM V5.4



z/VM Version 6

Security Certification Plans

UPDATE

- Common Criteria (ISO/IEC 15408)
  - z/VM V6.1 has been certified: BSI-DSZ-CC-0752
  - Evaluated to EAL 4+ for the Operating System Protection Profile (OSPP) with:
    - Virtualization extension (-VIRT)
    - Labeled Security extension (-LS)
- Federal Information Protection Standard (FIPS) 140-2
  - z/VM V6.1 System SSL is FIPS 140-2 Validated(TM)
  - Enablement requirements for certificate database and servers
    - http://csrc.nist.gov/groups/STM/cmvp/documents/140-1/1401val2012.htm#1735
- z/VM V6.2 and z/VM 6.3 are designed to conform to both Common Criteria and FIPS 140-2 evaluation requirements

A Certification Mark of NIST, which does not imply product endorsement by NIST, the U.S. or Canadian Governments.

SOD 23.7.2013: Security Evaluation of z/VM 6.3

IBM intends to evaluate z/VM V6.3 with the RACF Security Server feature, including labeled security, for conformance to the Operating System Protection Profile (OSPP) of the Common Criteria standard for IT security, ISO/IEC 15408, at Evaluation Assurance Level 4 (EAL4+).

- We continue the practice of taking every other release through certification.
- Evaluation is with inclusion of RACF Security Server optional feature.
- See http://www.vm.ibm.com/security/ for current z/VM Security information.
- Update: z/VM 6.3 is formally listed as "In Certification" (with Certification ID BSI-DSZ-CC-0903) on the BSI website: https://www.bsi.bund.de/EN/Topics/Certification/incertification.html

SOD 23.7.2013: FIPS Certification of z/VM 6.3

IBM intends to pursue an evaluation of the Federal Information Processing Standard (FIPS) 140-2 using National Institute of Standards and Technology's (NIST) Cryptographic Module Validation Program (CMVP) for the System SSL implementation utilized by z/VM V6.3.

- Federal Information Protection Standard (FIPS) 140-2
  - Target z/VM 6.3 System SSL is FIPS 140-2 Validated*
  - Enablement requirements for certificate database and servers
    - http://csrc.nist.gov/groups/STM/cmvp/documents/140-1/1401val2012.htm#1735

- See http://www.vm.ibm.com/security/ for current z/VM Security information.

A Certification Mark of NIST, which does not imply product endorsement by NIST, the U.S. or Canadian Governments.