IBM z Systems

z/VM: Winner of the Greatest Hypervisor on Earth Competition

"43 Years in a Row"

Bill Bitner
z/VM Client Focus and Care

bitnerb@us.ibm.com

Trademarks

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

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

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

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

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

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

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

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

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

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

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

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

This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

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

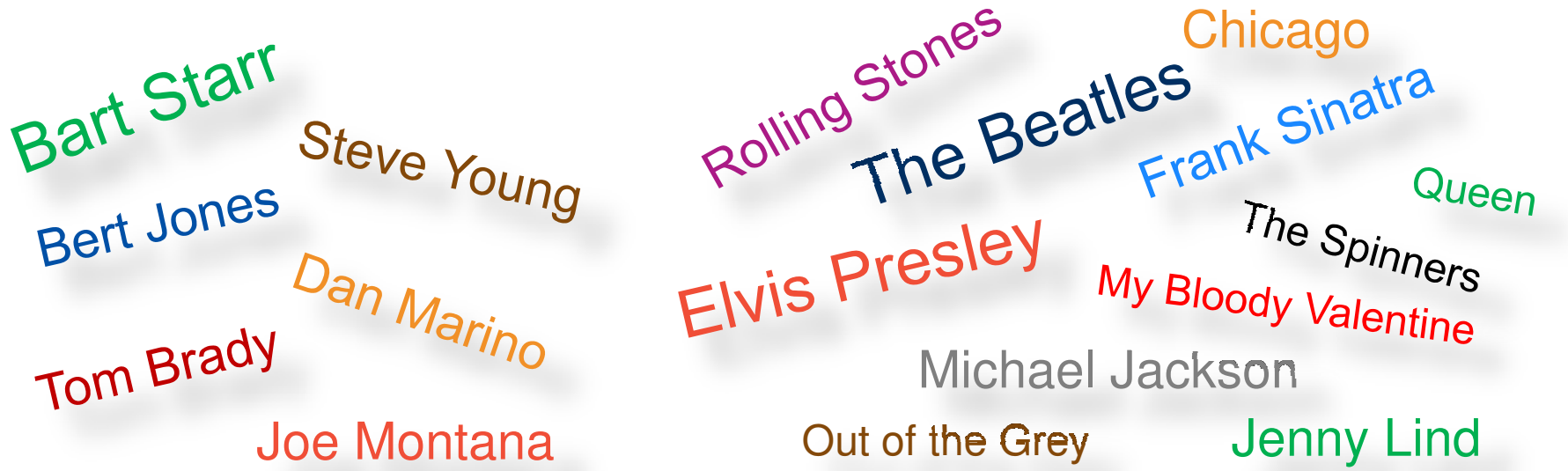
- The following people contributed charts, information, or thoughts used in this presentation:
  - Bob Albright
  - Susan Franciscovich
  - Reg Harbeck
  - Jeff Howard
  - Brian Hugenbruch
  - Emily Hugenbruch
  - Greg Kudamik
  - Reed Mullen
  - Rob Shisler
  - Paul D. Smith
  - Rick Troth
  - Helio Velloso de Almeida
  - Brian Wegener

What does it mean to be the Greatest?

Asked people to name and defend the "greatest":

- Quarterback (US Football)
  - Most Wins (Championships)
  - Makes things happen
  - Protects the ball / Consistent
  - Critical to team success
  - Body of work
  - Leverages teammates

- Singer/Band
  - Longest performing
  - Most songs/albums/gold
  - Impact to industry
  - Number of impersonators
  - Most famous
  - Meaningful
  - Talent
  - Cross Generational



z/VM Design Philosophy

Replication of the Architecture

- z/VM creates virtual machines with a high degree of architecture fidelity.
  - Obeys the rules of the z/Architecture Principle of Operations
  - Allows for a high level of trust that the virtualization provided by z/VM does not skew or contaminate or disrupt from functionality compared to running without z/VM.
  - Recursive virtualization

- Test programs used to validate System z Servers are also run against z/VM
- New processor features often implemented early in an internal z/VM version to aid in other software development
- This faithful replication of architecture gives ISVs a higher confidence that z/VM virtualization is a platform that can be supported.

z/VM – Part of a Bigger IBM Picture

- IBM has the entire stack
  - Hardware & Firmware
  - Hypervisors
  - Operating Systems
  - Middleware

- z/VM Inherits benefits of the platform

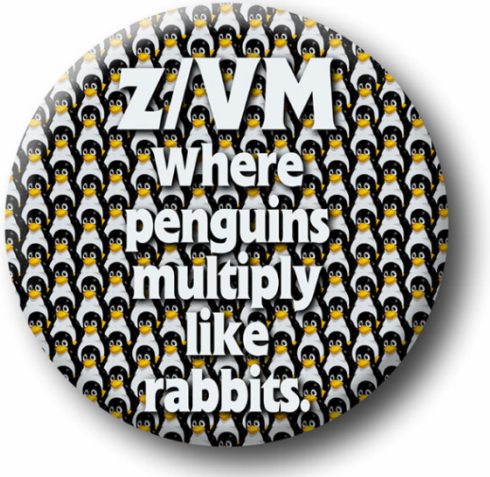
- Facilitates advances and interaction
  - Hardware assists
    - HPMA – Host Page Management Assist
  - Handshaking between Hypervisor and Operating Systems
    - Asynchronous Page Fault Processing
  - Multiple levels
    - QEBSM – QDIO Enhanced Buffer State Management
  - Testing advantages

The z/VM Community

- Long term connections, communication, and collaboration
  - VMSHARE electronic conference started in 1976
- z/VM customers and ISVs increase the level of help available
- z/VM Community tends to be friendlier, less flames, than other groups
- Long history of providing additional function and tools
  - Modifications to z/VM
  - Various tools and download packages
    - E.g. TRACK, SWAPGEN
- Long history of influencing and steering IBM
- z/VM Community – you're never alone

Adaptability to Varying Workloads

- The breadth and depth of z/VM systems is impressive



- z/VM Customers may span…
  - Memory >100 x's
  - System Processors 32 x's
  - Virtual Machine Size >800 x's
  - I/O Devices >500 x's

- z/VM supports them all and continues to adjust to changes in customer demographics

- Historically things have changed significantly, on the same code base
  - 1992: Over 20,000 OVVM CMS virtual machines
  - 2010: Over 500 Linux virtual machines

Protect the Customer

- Protect their investment
  - Compatibility of programs
  - Compatibility of data
  - Compatibility of behavior
  - Data to validate their investment
- Security
  - Certification
  - Integrity statement
- Reliability
  - Stability
  - Maintainability
  - Problem Determination
- Recognize needs of their business

Empower the Customer

- "If we code it, they will use it"

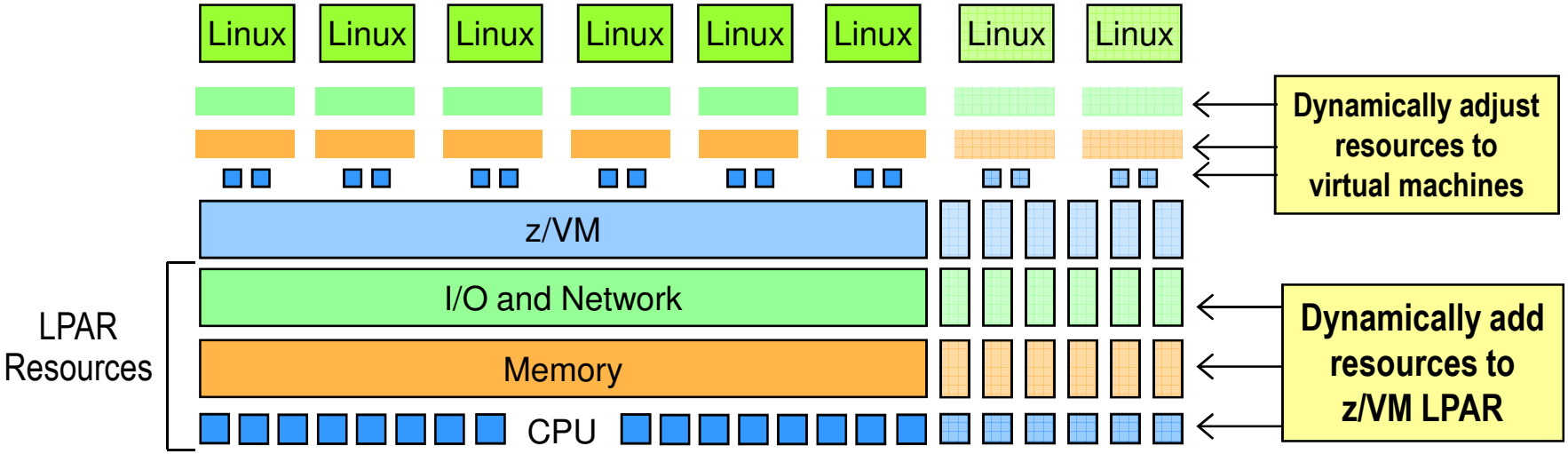
- Customization and extensions
  - CP exits
  - Utilities
    - REXX
    - Pipelines
    - OpenEdition
    - Sockets
    - Download pages
    - Community code

- Offers flexibility

Customer Value

Linux on z/VM: Flexible, efficient growth

- Clients can start small with Linux on z Systems and non-disruptively grow their environment as business dictates
- Users can dynamically add CPUs, memory, I/O adapters, devices, and network cards to a running z/VM LPAR
- z/VM virtualizes this capability for guest machines



Smart economics: non-disruptively scale the z/VM environment by adding hardware assets that can be shared with *every* virtual server

Extreme Consolidation and Scalability

- A benefit of virtualization is overcommitted resources. Done with the idea that not all of those resources will be used at the exact same time.
  - Consolidation of white space or unused physical resources
  - Discrete servers averaging 15% busy
  - Redundant servers for availability
  - Unaligned utilization peaks for various virtual servers

- z/VM scalability continues to be increased

Overcommitment of Resources - Processors

- How defined? Typically number of defined virtual processors to logical processors
- z/VM was designed to minimize overhead in dealing with virtual processors
- Really more about processor utilization, which can be measured
- Care does need to be taken to avoid over configuring virtual processors and incurring too much MP overhead.
  - We have data to help with that

Overcommitment of Resources - Memory

- How defined?
  - Real memory = total central + expanded storage (memory)
  - Virtual memory = total of logged on virtual machine defined size
    - Factor in other virtual spaces (e.g. Vdisk)
  - Warning: different people define differently
    - Example: WebSphere admins often equate virtual to JVM Heap size
    - Example: Some people use size of discrete server being consolidated instead of size of virtual machine

- Two rules of thumb:
  - If you don't want to think about it, run 1:1. But that can be wasteful.
  - If you go above 3:1, you should seek expert help

- Mishandling memory configurations is one of leading causes of performance problems when running Linux on z/VM
  - Common idea of just add more memory to a virtual machine can have negative effects
  - Excess memory capacity in Linux is often not used effectively

Factors Affecting Memory Overcommitment

- Guest provisioning: if oversized, easier to “squeeze down”
- Percentage of workload active at any one time
- Sensitivity to latency
  - Cognos “bursty” workload suffers from delays due to spikes in memory demand
  - WAS somewhat more tolerant of faults, provided heap not impacted
- Software levels
  - Newer WAS levels exhibit better idle behavior
- Software mix
  - Typically several types of virtual servers on the same z/VM host. OC ratio must be tuned to satisfy all.
- SLA stringency: “**all** transactions must complete in < 1 sec.” vs. “99.9% must complete in < 1 sec”.
- Capacity and bandwidth of paging I/O configuration

Memory Overcommitment – Features and Capabilities

- **Greater Efficiency Features**
  - Ballooning (CMM1, via cpuplugd or VMRM)
  - Page state handshaking (CMMA Lite)
  - Hierarchical paging (aging list)
  - Pageable page tables
  - Block paging
  - Guest swap to VDisk (allows smaller guests)
  - Shared memory (XIP filesystem on DCSS)

Memory Overcommitment – Features and Capabilities

- **Mitigate and manage negative impact**
  - Asynchronous host page faults
  - Real memory entitlement for individual guests (SET RESERVED)
  - Scheduler settings to adjust memory capacity calculations (avoids E-list abuse)
  - Paging I/O Throttling
    - Avoids memory depletion from queued I/O requests
    - Lets workload survive, run as fast as paging devices allow
- **Diagnosis, Tuning, and Capacity Planning aids**
  - Rich monitor data stream
  - Real time and post processing

Processor Overcommitment – Features and Capabilities

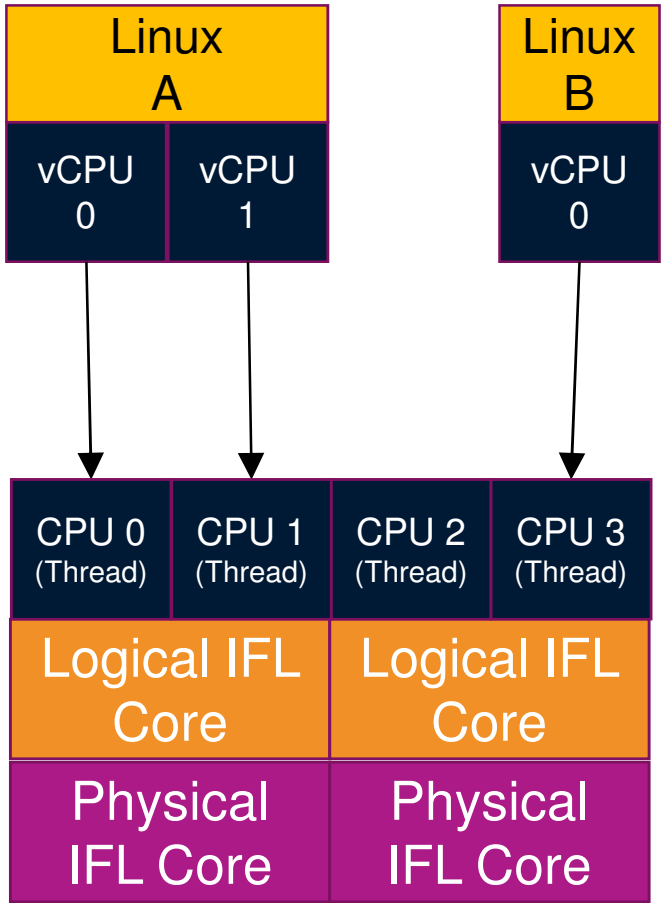
- **Greater Efficiency Features**
  - HiperDispatch
    - Processor affinity
    - Vertical CPU Mangement
  - Intelligent scheduling
    - Deadline & Consumption algorithms
  - Spin-locks yielding control to hypervisor
    - Guest and Host level
  - Start Interpretive Execution instruction
  - Assists in hardware and firmware
  - Test idle
  - SMT support
  - Independent dispatching of virtual processors
  - Linux CPUPLUGD

Processor Overcommitment – Features and Capabilities

- **Mitigate and manage the negative impact**
  - Share settings
    - Minimum
    - Limit
  - CPU Pooling
- **Diagnosis, Tuning, and Capacity Planning aids**
  - Rich monitor data stream
  - Real time and post processing

SMT in z/VM

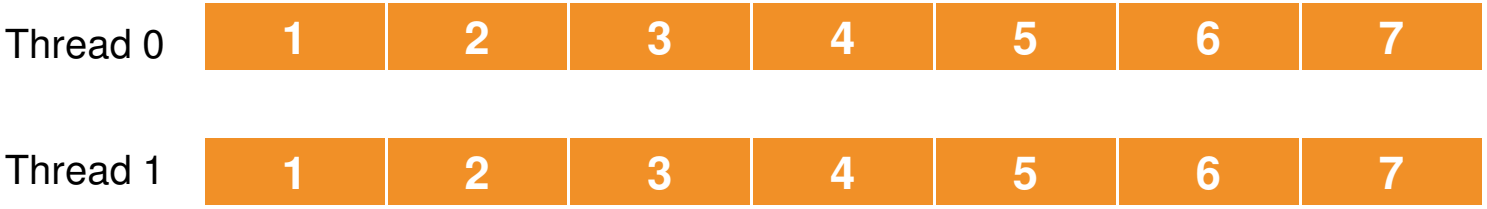
- **Physical IFL Cores (you purchase these) with SMT allow up to two threads to be used**
- **Logical IFL Cores are presented to z/VM as in the past (you define these in the logical partition profile on the HMC)**
- **z/VM creates a CPU or logical processor associated with each thread (reflected in commands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



Additional Work Capacity

IFL (SMT disabled) – Instruction Execution Rate: 10

IFL (SMT enabled) – Instruction Execution Rate: 7

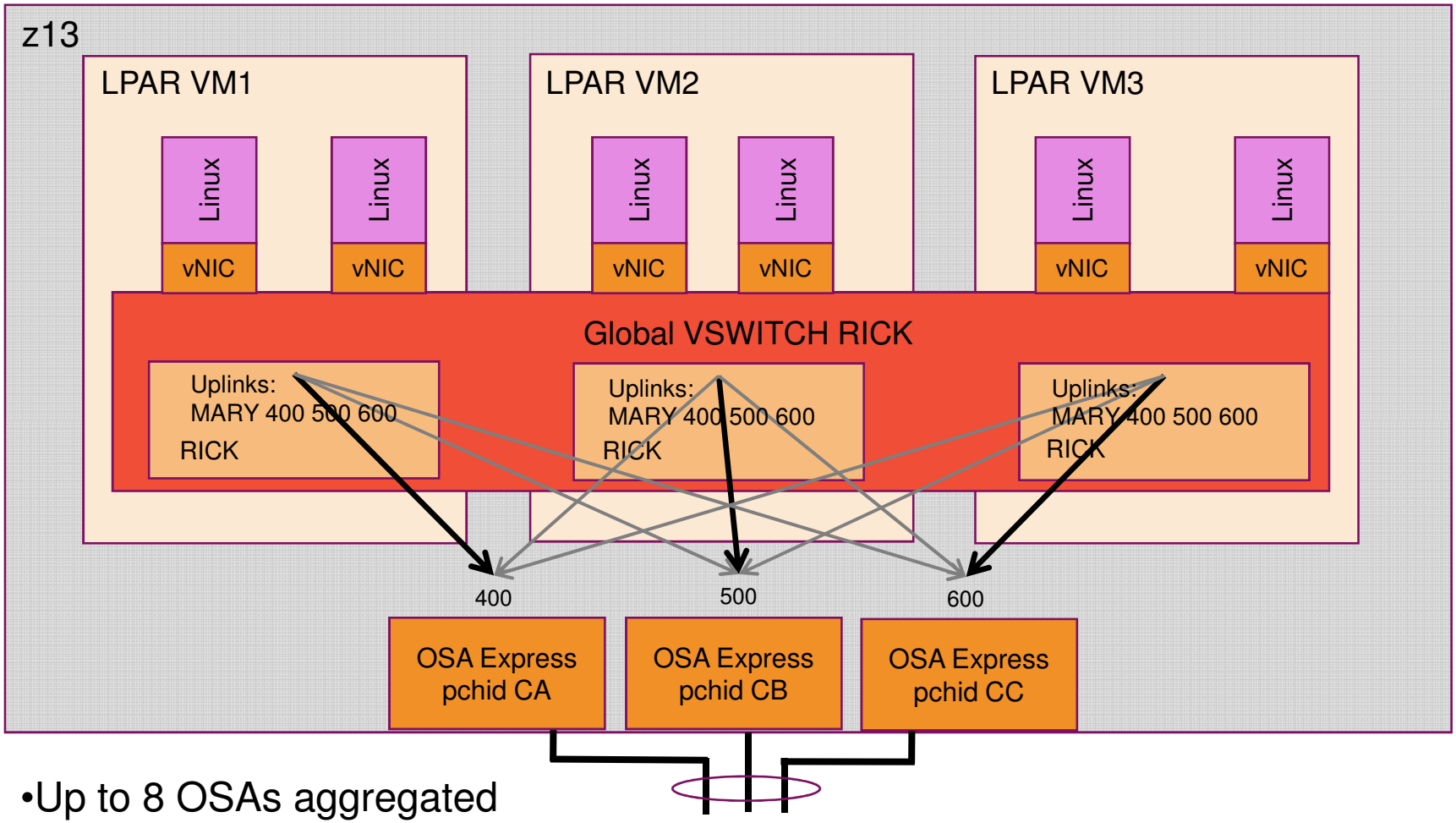


- Numbers are just for illustrative purposes
- Without SMT, 10 / second
- With SMT, 7 / second but two threads yields capacity of 14 / second

Processor Time Reporting

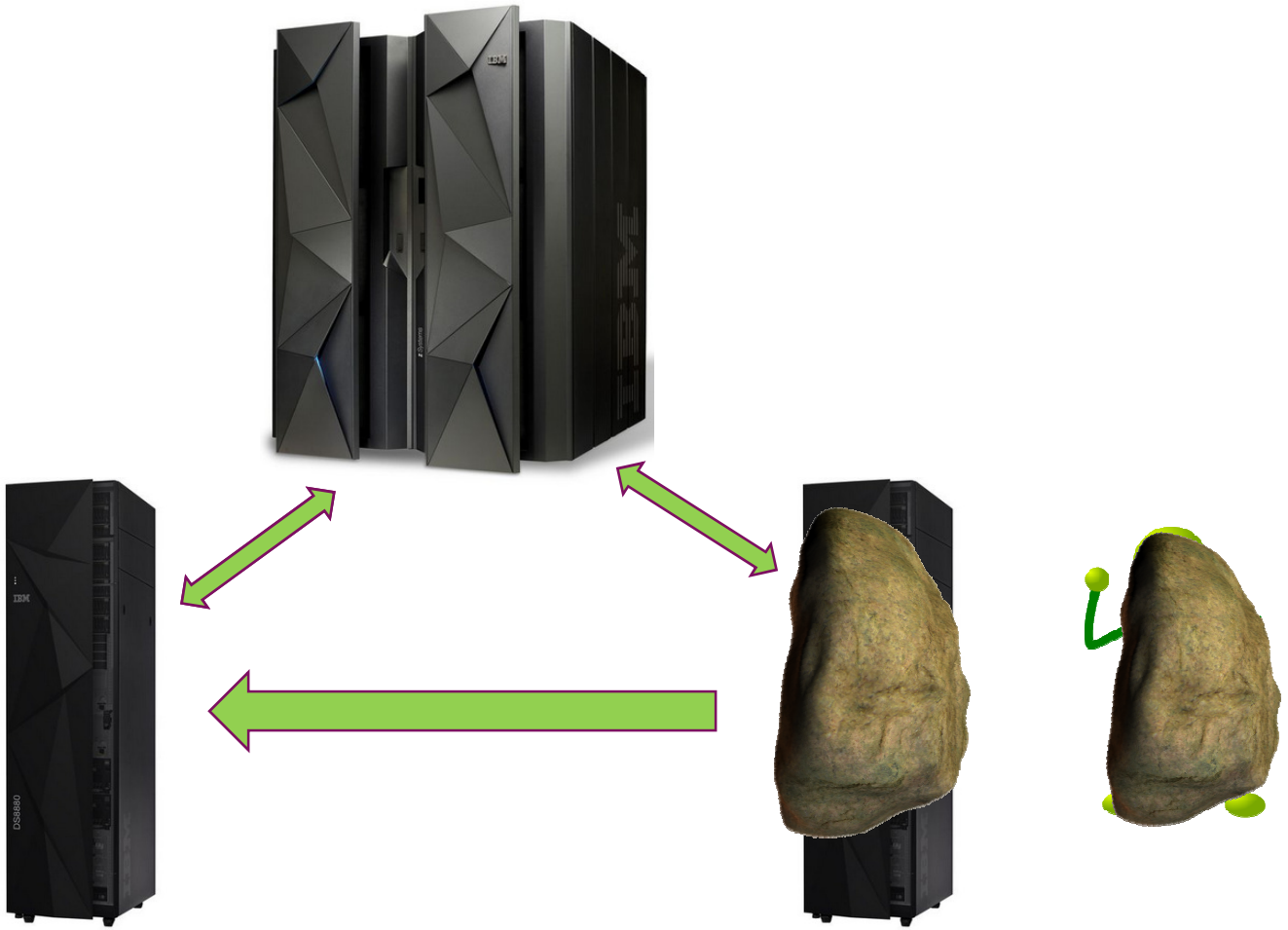
- **Raw time** (the old way, but with new implications)
  - Amount of time each virtual CPU is run on a thread
  - This is the only kind of time measurement available when SMT is disabled
  - Used to compute dispatcher time slice and scheduler priority
- **MT-1 equivalent time** (new)
  - Used when SMT is enabled
  - Approximates what the raw time would have been if the virtual CPU had run on the core all by itself
    - Adjusted downward (decreased) from raw time
  - Intended to be used for chargeback
- **Pro-rated core time** (new with VM65680)
  - Used when SMT is enabled
  - "Discounts" raw time proportionally when core is shared between active threads
    - Full time charged while a virtual CPU runs alongside an idle thread
    - Half time charged while vCPU is dispatched beside another active thread
  - Suitable for core-based software license metrics

Global z/VM Virtual Switch



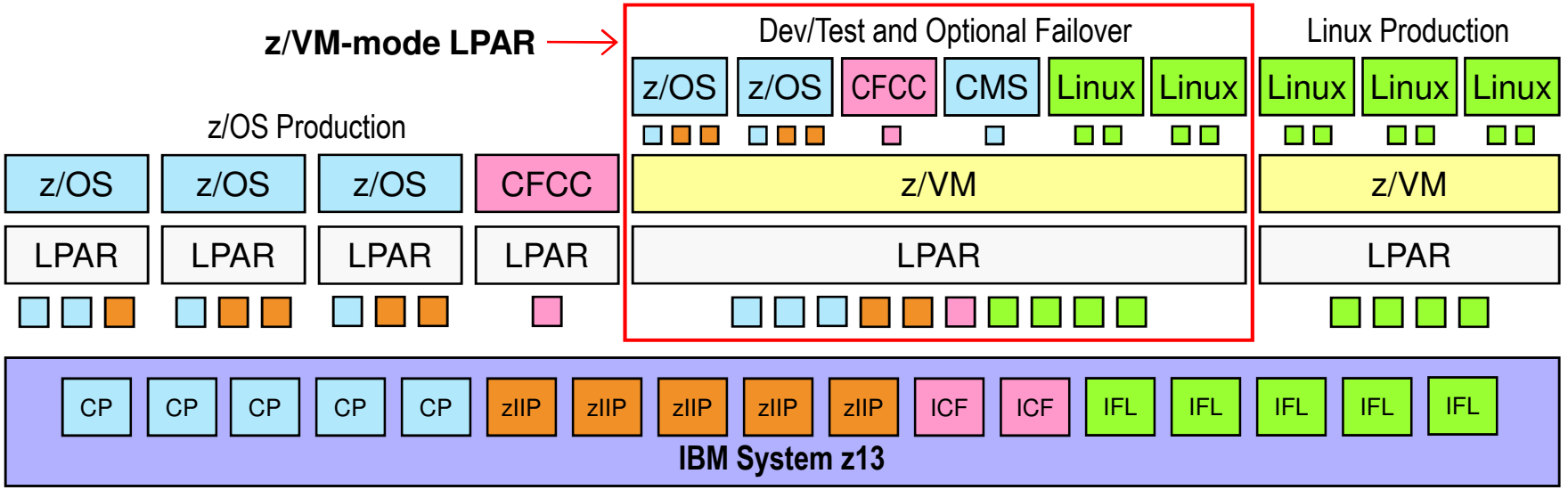
- Up to 8 OSAs aggregated
- Automatic failover
- Automatic balancing

GDPS – The Simple Explanation



z/VM-Mode LPAR Support for IBM zEnterprise Servers

- LPAR type (introduced with IBM System z10): *z/VM-mode*
  - Allows z/VM users to configure all CPU types in a z/VM logical partition
- Offers added flexibility for hosting mainframe workloads
  - Add *IFLs* to an existing standard-engine z/VM LPAR to host Linux workloads
  - Add *CPs* to an existing IFL z/VM LPAR to host z/OS, z/VSE, or traditional CMS workloads
  - Add z*IIPs* to host eligible z/OS specialty-engine processing
  - Test integrated Linux and z/OS solutions in the same LPAR



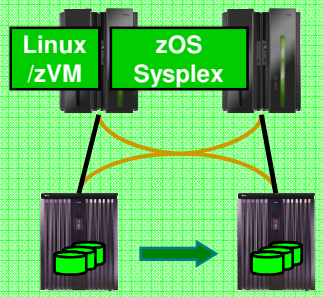
GDPS/PPRC for two sites: Metropolitan distance continuous availability (CA) and disaster recovery (DR) solution

Continuous Availability / Disaster Recovery within a Metropolitan Region

Two Data Centers

Systems remain active

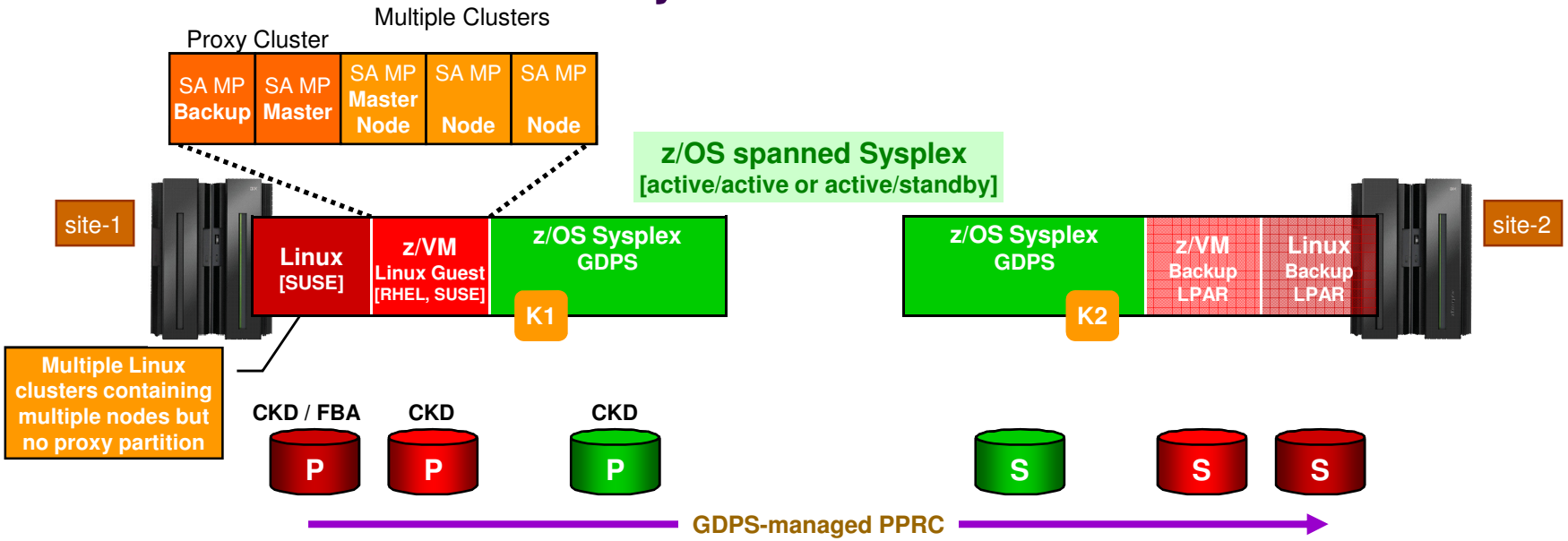
Multi-site workloads can withstand site and/or storage failures



GDPS/PPRC
active/active, active/standby configs

- Provides Parallel Sysplex and server management
- Simplifies and streamlines data replication management
- Manages remote copy environment using HyperSwap function and keeps data available for operating systems and applications (extends Parallel Sysplex CA function to disk data)
- Facilitates faster recovery time for planned and unplanned outages
- Ensures successful recovery via automated processes
- Enhances data consistency across all secondary volumes for both System z and distributed systems
- Leverages Distributed Cluster Management (DCM) to interface with distributed environments to provide an enterprise-level disaster recovery solution
- Combines with GDPS/Global Mirror or GDPS/XRC to provide a three-site solution for higher availability and disaster recovery

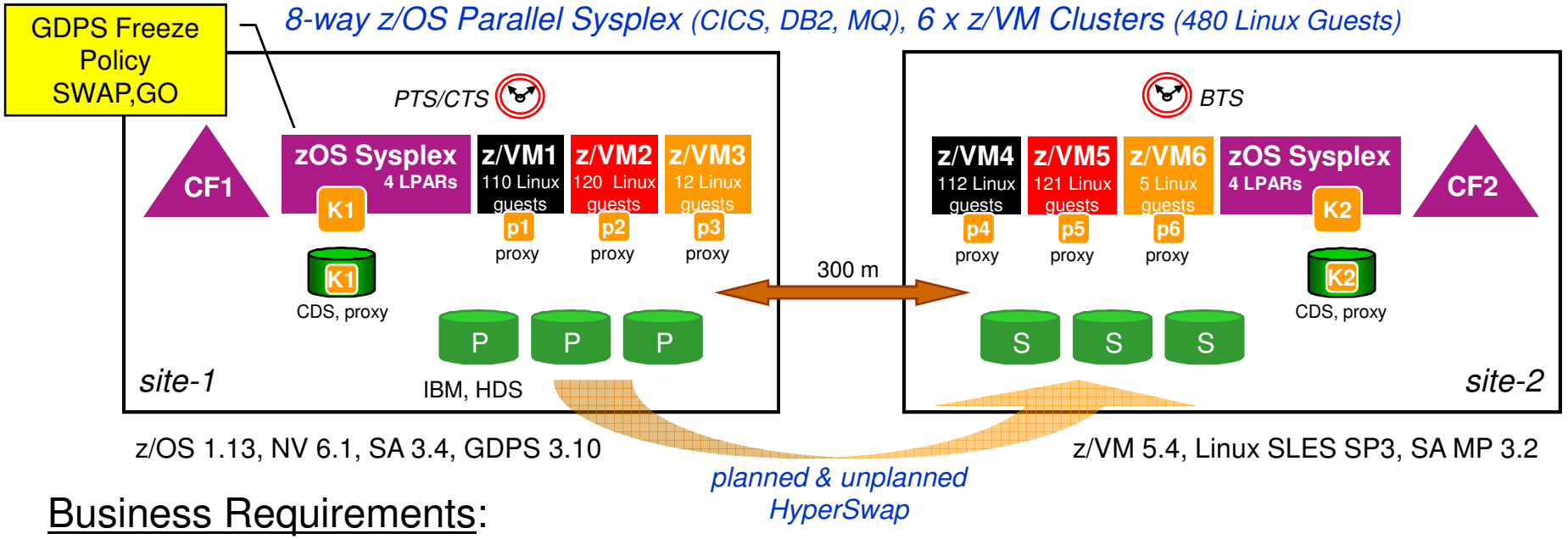
GDPS/PPRC xDR: Linux guest & native Linux on System z – Continuous Data Availability



- **Multiplatform Resiliency for IBM z Systems**
- Coordinated HyperSwap – z/OS, z/VM with its guests, and native Linux
- Graceful shutdown and startup (re-IPL in place) of Linux clusters or nodes
- z/VM SSI Live Guest relocation
- Graceful shutdown of z/VM
- Coordinated takeover – recovery from a Linux node or cluster failure
- Multiple SA MP Linux cluster are supported as are multiple z/VM systems & Linux LPARs

Coordinated recovery for planned and unplanned events

GDPS/PPRC xDR – MSW



Business Requirements:

- No data loss (RPO 0 sec)
- Continuous data availability for z/OS and Linux hosted by z/VM
- Coordinated disaster recovery for heterogeneous System z applications (RTO < 1 hour)

z/OS PPRC Pairs	z/OS LSS	z/VM PPRC Pairs	z/VM LSS	Planned HS RESYNC UIT	Planned HS SUSPEND UIT	Unplanned HyperSwap UIT
7,354	34	3,725	26	25 sec	18 sec	21 sec

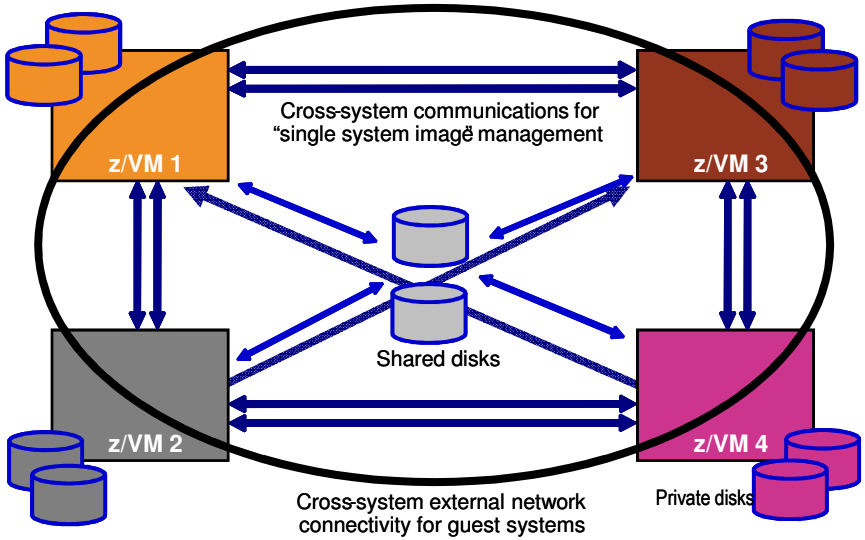
UIT = User Impact Time (seconds)
RPO = Recovery Point Objective
RTO = Recovery Time Objective

6/2014

Single System Image (SSI) Feature

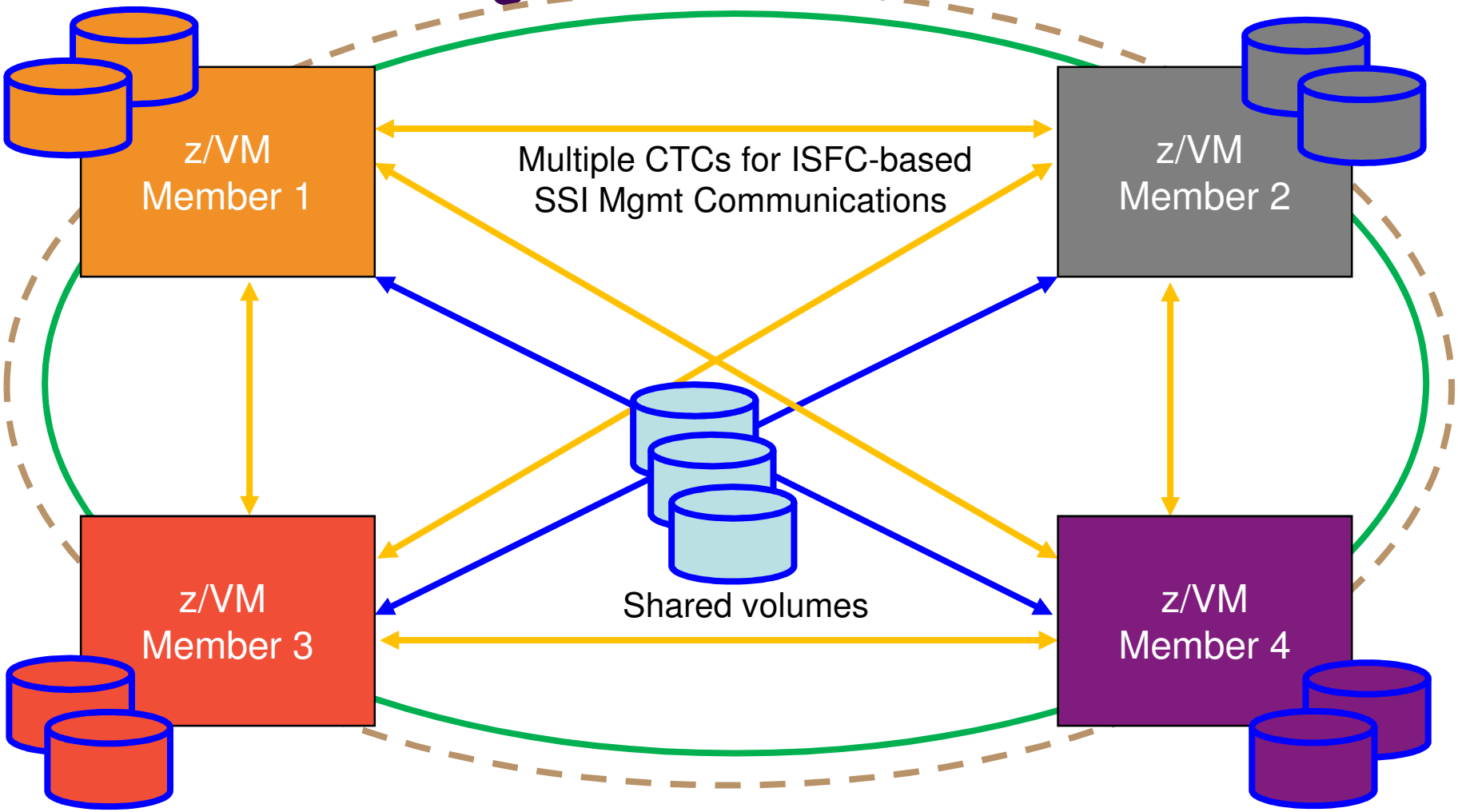
Clustered Hypervisor with Live Guest Relocation

- Optional priced feature, available starting with z/VM 6.2
- Connect up to four z/VM systems as members of a Single System Image cluster
- Cluster members can be run on the same or different System z servers
- Simplifies management of a multi-z/VM environment
  - Single user directory
  - Cluster management from any member
    - Apply maintenance to all members in the cluster from one location
    - Issue commands from one member to operate on another
  - Built-in cross-member capabilities
  - Resource coordination and protection of network and disks



- Allows Live Guest Relocation of running Linux guests

SSI Cluster Configuration



SSI Cluster Management: Greater Reliability

- Cross-checking of configuration details as members join cluster and as resources are used:
  - SSI membership definition and identity
  - Consistent definition of shared spool volumes
  - Compatible virtual network configurations (MAC address ranges, VSwitch definitions)
- Cluster-wide policing of resource access:
  - Volume ownership marking to prevent dual use
  - Coordinated minidisk link checking
  - Autonomic minidisk cache management
  - Single logon enforcement
- DirMaint
  - Main DirMaint virtual machine which can run on any of the members
  - Main DirMaint coordinates with satellite virtual machines on other members
  - A member that is down will be brought “up to speed” when re-started.

SSI Cluster Management: Addressing Problems

- Communications failure “locks down” future resource allocations until resolved
  - Existing running workloads continue to run
  - Prevents new accesses to resources
  - Cluster could temporarily be split and workloads continue to run

- Added the new “REPAIR” option to IPL for severe problem resolution
  - Meant for use with a single member cluster to repair
  - Allows correcting various problems that aren’t addressable in standard cluster.

A word or two on skills

Which one might have a slightly larger instruction manual?

z/VM 6.4 Preview

Preview IBM z/VM 6.4

- Preview announcement 216-009, dated February 16, 2016
  - http://www.vm.ibm.com/zvm640/index.html
- Planned availability date Fourth Quarter 2016
- A release born from customer feedback
- Key components:
  - Enhanced technology for improved scaling and total cost of ownership
  - Increased system programmer and management capabilities
- New Architecture Level Set (ALS) of z196 and higher

Improved Scalability and TCO

- z/VM Paging enhancements
  - Use of HyperPAV when available to increase bandwidth for paging
  - Increases number of paging I/Os that can be in-flight at once
  - Exploitation for Paging, Spooling, z/VM user directory, and minidisk pools that are mapped to z/VM data spaces.

- Guest large page support
  - Enhanced DAT facility for guest use
  - 1 MB pages
  - Decreases memory needed for DAT structures by guest with Enhanced DAT support
  - z/VM maps to 4KB pages at the host level.

- Guest Transactional Execution support
  - Potential efficiency and scaling improvements for guests and guest software that exploits
  - Alternative for serializing a set of operations.

Improved Scalability and TCO

- Memory scalability improvements
  - Enhanced algorithms to further improve the efficiency of memory management
  - Provide a foundation for future enhancements in scaling and efficiceny

- FlashSystems support for FCP-attached SCSI disks.
  - Removes requirement of a San Volume Controller (SVC) to use FlashSystems for z/VM system volumes and EDEVs

System Programmer & Management Capability

- QUERY SHUTDOWN command
  - Allows better understanding of state of the system
  - Allows for increased programmatical management of the system

- CP environment variables
  - New framework to allow information to be set and queried for automatic processing
  - Example: Indicate system is being started for Production or DR Test or Actual DR

- New management queries for SCSI environment.
  - Allows SCSI detailed information to be gathered for emulated devices (EDEVs)

System Programmer & Management Capability

- CMS Pipelines enhancements
  - Pipelines is a powerful programming construct available in the CMS environment
  - Objective is to make available, with the product, many of the advances made to Pipelines since it was last updated in the product
  - Allows use of various tools and programming without the need to download additional code

- DirMaint to RACF Connector
  - Modernizes the Connector with a collection of functional enhancements
  - Brings processing in line with modern z/VM practices
  - Allows better passing of directory information to RACF
  - Facilitates proper security policy in environment managed by IBM Wave for z/VM or OpenStack

System Programmer & Management Capability

- Upgrade In Place migration enhancements
  - Upgrade In Place migration was introduced in z/VM 6.3
  - Enhanced to allow migration to z/VM 6.4 from
    - z/VM 6.2 or z/VM 6.3 (but not both at same time in cluster)
    - Supports migration for clustered or non-clustered systems

z/VM 6.4 Supported Hardware

- Following z Systems servers:
  - z13
  - z13s
  - LinuxONE Emperor
  - LinuxONE Rockhopper
  - IBM zEnterprise EC12
  - IBM zEnterprise BC12
  - IBM zEnterprise 196
  - IBM zEnterprise 114

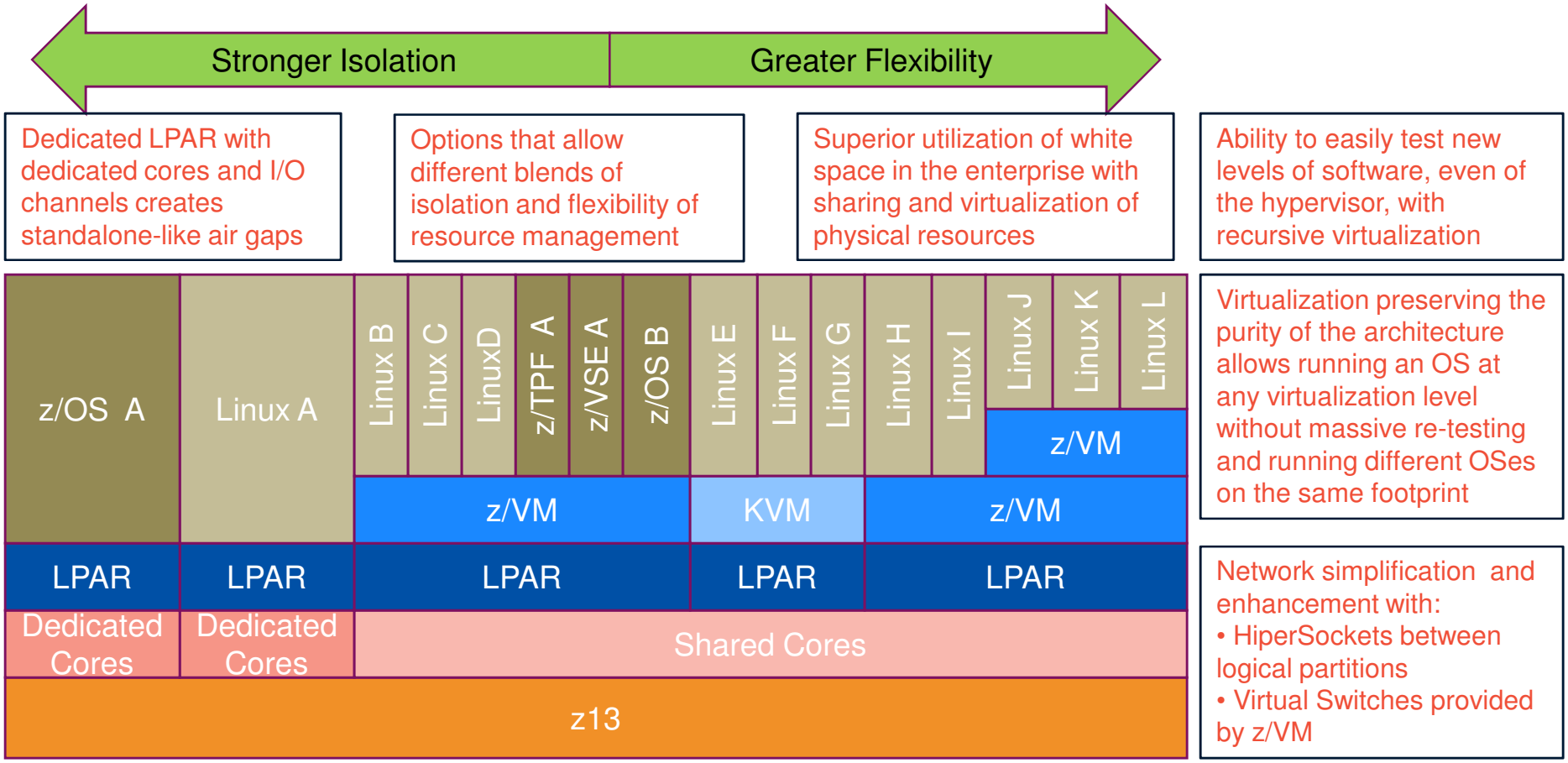
- Electronic and DVD install
  - No tapes

What's the Point Again?

Challenges of IT?

- The bigger the server I buy, the better the economies of scale. How can I buy the biggest machine and use it effectively?
- There is that one workload that needs to be strongly isolated for regulatory reasons. How can I support that and still leverage economies of scale?
- Globalization, workload variations, month-end processing, and the need for rapid deployment of new solutions creates huge variations in resource requirements and peaks for different workloads. Is there a solution that makes up for unbendable workloads with greater flexibility in resource management?
- Migrating to new releases or service used to be a ‘nice to have’. Now with security patches and other demands, it’s a requirement. How can I make it easier to keep existing software running, apply patches, and test it all?
- My administrators are so busy upgrading hardware and migrating that they don’t have time to support new business projects. Isn’t there a way to do hardware updates more quickly?
- All my applications seem to use different operating systems for different purposes. Can I save expenses by collocating them all on one platform?

IBM z Systems: The Solution to the Challenges of IT



Summary

z/VM has a history and a future of Leading