Goal-oriented dynamical performance management of KVM for IBM z Systems virtual server CPU resources

Yüksel Günal
ygu@us.ibm.com
IBM Systems
Poughkeepsie, NY, USA

April 6th, 2016

Trademarks

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

DB2*	ECKD	IBM*	LinuxONE	PR/SM	z13	z Systems
DB2 Connect	FICON*	Ibm.com	LinuxONE Emperor	Storwize*	zEnterprise*	z/VSE*
DS8000*	FlashSystem	IBM (logo)*	LinuxONE Rockhopper	XIV*	z/OS*	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 .
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.

Outline of this talk

- An introduction to KVM for IBM z Systems
- Objectives of Hypervisor Performance Manager on KVM for IBM z Systems (zHPM)
- zHPM Functional Capabilities
  - Concept of a workload resource group
  - Concept of a performance policy
  - Monitoring workloads
  - Basics of CPU management
  - CPU-critical support
  - How do cgroups CPU shares work?
  - RESTful APIs to interact with zHPM
- Usage Case 1
- Usage Case 2
- Demo
- References
- Q & A

z Systems and LinuxONE Virtualization Options

IBM z Systems now has three strategic virtualization platforms

- KVM for IBM z Systems
- IBM z/VM
- IBM Processor Resource/System Manager (PR/SM)

KVM

KVM for IBM z provides an open source choice for IBM z Systems and LinuxONE virtualization for Linux workloads. Best for clients that are not familiar with z/VM and are Linux centric admins.

z/VM

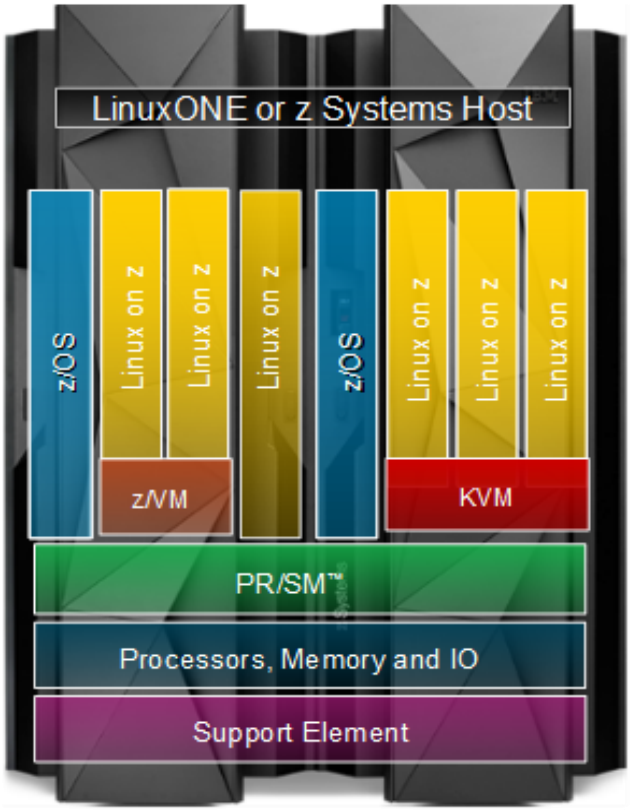
Proprietary Server Virtualization that is deeply integrated into System z. Complete hardware awareness. Supported on all IBM z Systems and LinuxONE servers. z/VM will continue to be enhanced to support Linux Workloads.

PR/SM

Divide one physical server into up to 85 logical partitions (LPAR) running a mix of multiple z/OS, z/VM, Linux, KVM for IBM z, Transaction Processing Facility (TPF) and z/VSE instances isolated and secured in parallel. Share resources across LPARs or dedicated to a particular LPAR. Running a mix of multiple z/OS, z/VM, Linux, TPF, KVM for IBM z and z/VSE instances isolated and secured in parallel.

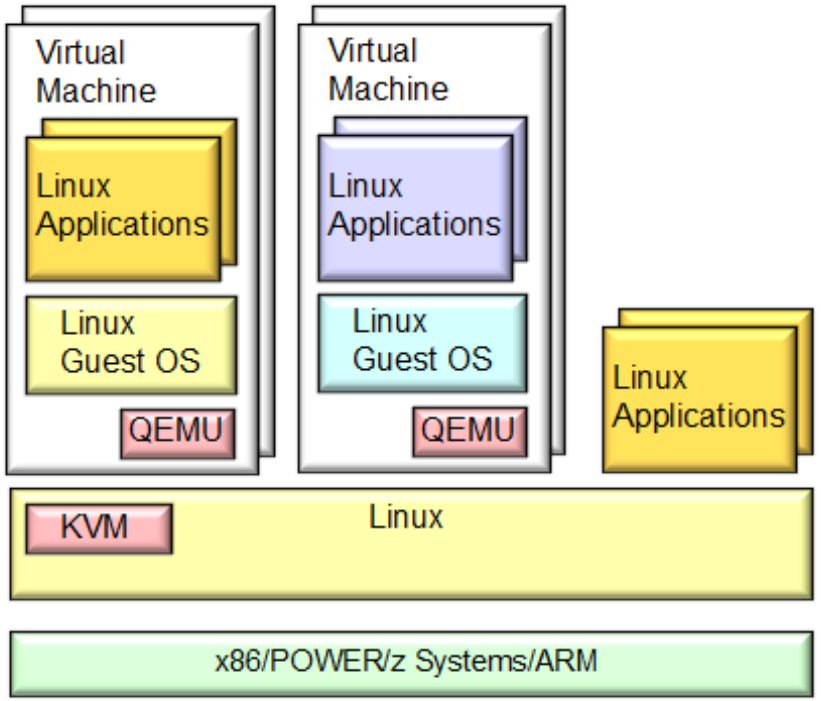
Standards based virtualization

- Standard KVM interfaces allow for quick startup for clients who are familiar with x86 Linux
- Standard management and operation controls leading to greater operational efficiencies
- KVM-based virtualization on z Systems and LinuxONE allows businesses to reduce costs by deploying fewer systems to run more workloads, sharing resources, and improving service levels to meet demand
- KVM open source solution for running virtual servers on z Systems and LinuxONE enables cloud deployments and big data solutions while reducing complexity and cost



Kernel Based Virtual Machine (KVM)

- **An open source hypervisor based on Linux**
  - Linux provides the base capabilities
  - KVM turns Linux into a hypervisor
  - QEMU provides I/O device virtualization and emulation
- **Provides flexibility in technology choices**
  - Open
  - Scalable
  - Economical



What problem is zHPM trying to solve?

- Protect important workloads while maintaining high utilization of KVM for IBM z Systems CPU resources.

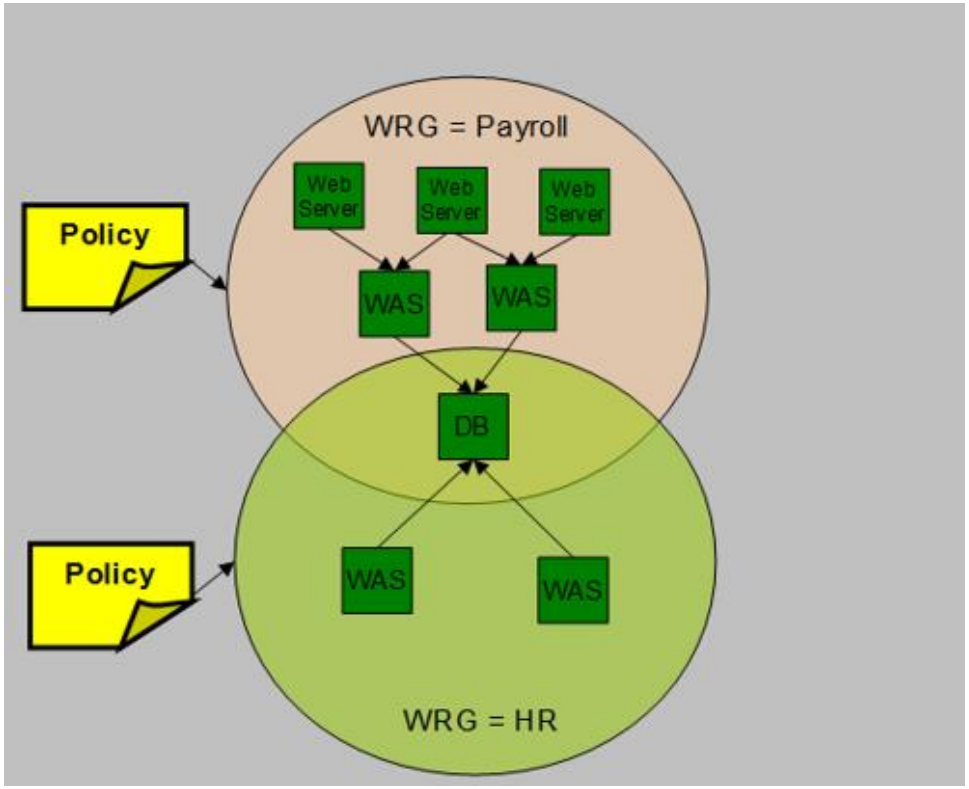
- Initial Release
  - Introduce the concept of a workload resource group
  - Introduce the concept of a goal-oriented performance policy that assigns business importance levels and performance objectives to virtual servers
  - Support performance policy based resource monitoring
  - Support goal-oriented CPU resources management
    - Dynamic control groups CPU shares management based on performance objectives and business importance levels

System z Hypervisor Performance Manager (zHPM)

- A multi-threaded Java program with some native libraries
- Interacts with libvirt (virtualization library) and Linux kernel
- Supports policy-based goal-oriented monitoring and management of CPU resources
- Shipped as part of KVM for IBM z Systems
  - Optionally enabled (zHPM does not run by default)
- Scope of management is single KVM for IBM z Systems instance
  - zHPM will have no knowledge outside of its zKVM instance
- Controlled through RESTful Web Services APIs and CLI
  - APIs
    - Point of integration with higher-level virtualization management solutions
    - Support for scripting
    - Fully documented external interface
  - CLIs provide support for local administration

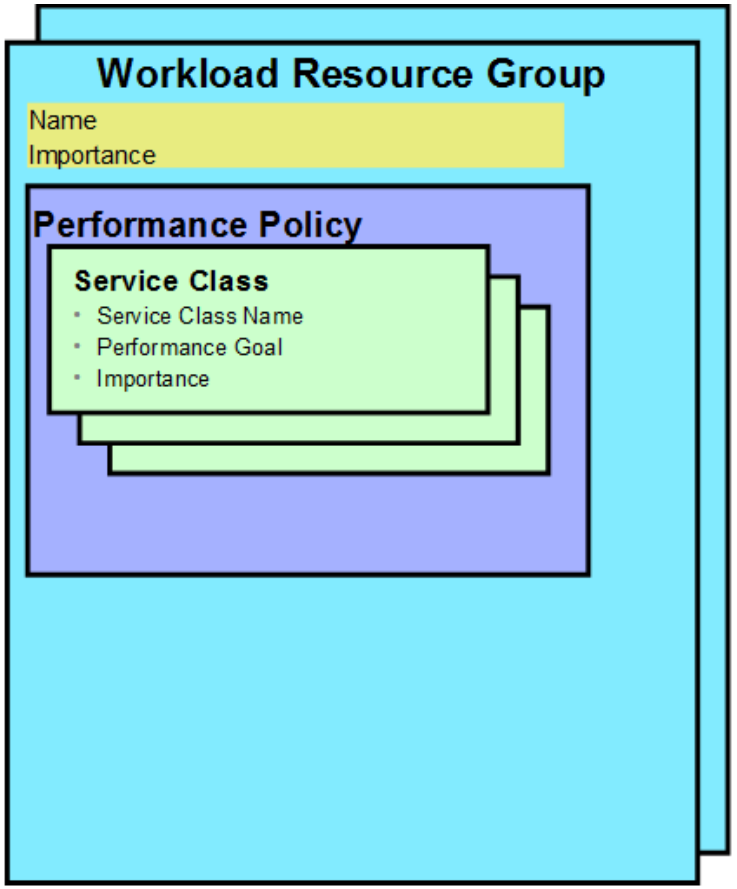
The concept of a Workload Resource Group

- A workload resource group (WRG) is a grouping mechanism and “management view” of virtual servers supporting a business application
- Provides the context within which associated platform resources are presented, monitored, reported, and managed
- In the figure we show a payroll workload and an HR workload example
- Management policies are associated to Platform Workload
  - Performance
  - Availability (Future)



The concept of a user-defined Performance Policy associated with a Workload Resource Group

- Performance policy defines performance objectives for virtual servers in a workload resource group
  - Conceptually similar to simplified z/OS WLM Policy
- Provides basis for monitoring and management of platform resources used by virtual servers in a Workload Resource Group
- Policy structure:
  - Policy contains a set of service classes
  - Virtual servers within the workload are assigned to a service class
  - A service class defines a performance goal and importance and a set of classification rules
  - Currently supports velocity goals



Monitoring and Reporting within the context of Workload Resource Groups (WRGs)

- Reporting capability that shows usage of hypervisor resources in a workload resource group context
- Performance goal vs actual performance reporting
  - Easily identify WRGs not achieving performance goals
- Drill down from overall WRG “health” view to contributions of individual virtual server
  - Quickly isolate virtual servers contributing to performance issues
    - From Workload Resource Group view monitor how each service class is performing
    - For each service class monitor the virtual servers that are associated with the service class

Basics of CPU management

- Objective function to optimize
  - Performance Index (PI) is calculated based on performance objective defined in a performance policy and actual performance
    - When PI is <= 1, performance goals are met. When PI > 1, performance goals are missed.
    - Performance is calculated as *actual velocity*, which is a function of CPU utilization and CPU delay. It is a number between 1 and 100. The higher the velocity, the better the performance.
    - Performance goals defined in terms of velocity goal in performance policies. 5 different velocity goals allowed:
      - Fastest
      - Fast
      - Moderate
      - Slow
      - Slowest
- Management knob to tune *is cgroups CPU shares*, a relative resource allocation control
- The higher importance workloads missing their performance goals are helped by moving CPU shares from virtual servers in lower importance workloads or same importance workloads with better performance (in this case pain is spread evenly). CPU shares moved are projected based on PI improvement projections.

How do cgroups CPU Shares work?

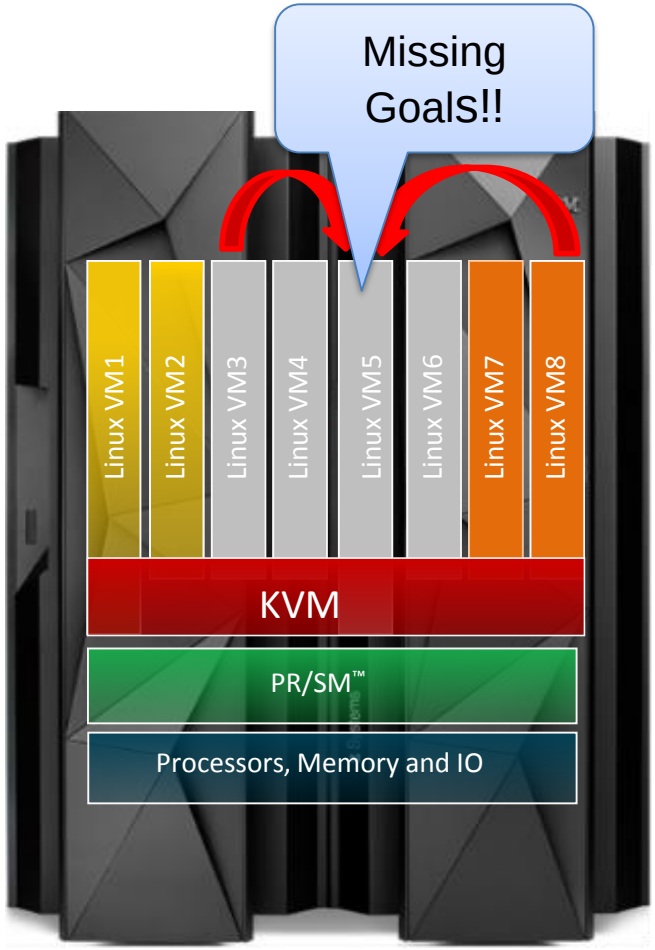
- Relative resource allocation knob
- Consider a KVM hypervisor with 4 CPUs and 2 virtual servers, VS_A and , VS_B
  - Assume VS_A has 1024 CPU shares and VS_B has 4096 CPU shares
    - VS_A will get $(4 * 1024/(1024 + 4096)) = 0.8\ CPUs$
    - VS_B will get $(4 * 4096/(1024 + 4096)) = 3.2\ CPUs$

Note that the calculations above assume the virtual servers have the virtual CPUs to support the CPU time that their CPU shares entitle them to. For instance, VS_B should have at least 4 virtual CPUs to be able to consume 3.2 CPUs.

zHPM's CPU management actions are zero-sum: a virtual server gets an amount of CPU shares equal to what is taken from one or more donor virtual servers.

Managing Resources across Virtual Servers on a hypervisor

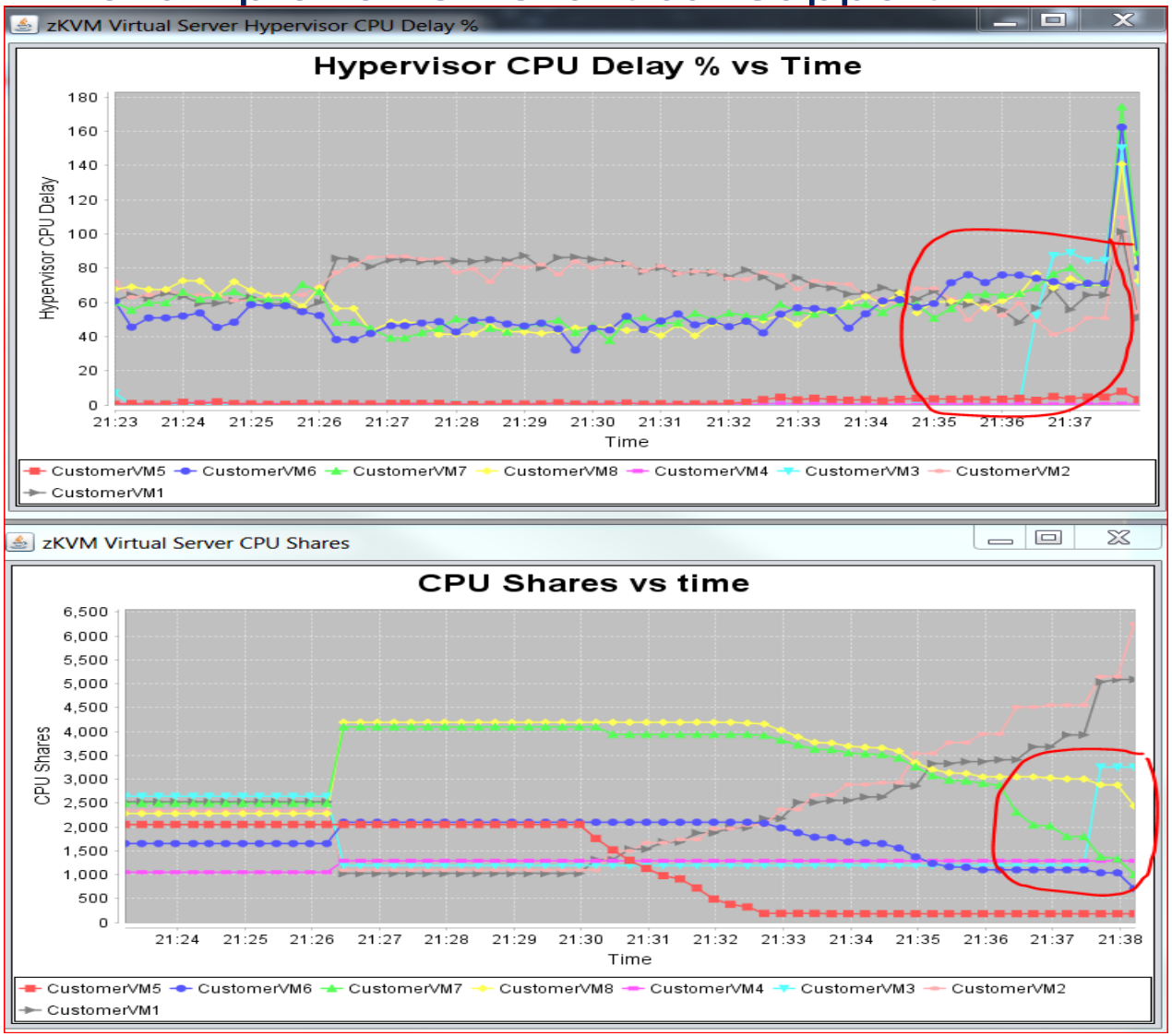
- Manage CPU resources across virtual servers to achieve performance goals
  - Detect that a virtual server is part of a WRG not achieving its goals
  - Determine that the virtual server performance can be improved with additional resources
  - Project impact on all affected virtual servers of reallocating resources
  - If good trade-off based on policy, redistribute processor resources
  - Current support for CPU management, potential to extend to other resources



Support for *CPU Critical* Workloads

- An attribute for service classes
- Virtual Servers associated with a service class designated as CPU-Critical will be monitored for sharp changes in workloads
  - If a CPU-critical virtual server suddenly experiences ***sharp changes*** **in hypervisor CPU delays**, its CPU shares will be adjusted by a large amount rather than rely on CPU management to do the job gradually
  - CPU-critical attribute does not override business importance

An example for CPU critical support



- Note in the upper graph on the left the sharp spike in CPU delay experienced by virtual server named “CustomerVM3” (the graph in light blue)
- Note in the lower plot how zHPM takes an immediate action to remedy the performance issue with this virtual server by increasing its CPU shares sharply

RESTful APIs to interact with zHPM - I

- Web services APIs for workload resource group operations
  - Add virtual server to a workload resource group
  - Create workload resource group
  - List workload resource groups
  - Get workload resource group properties
  - Get default workload resource group properties
  - Delete workload resource group
  - List performance policy for workload resource groups
  - List workload resource group detail for all workload resource groups
  - Remove a virtual server from a workload resource group
  - Update workload resource group to switch performance policy back to default
  - Update performance policy of default workload resource group

- Virtual server operations
  - List virtual servers known to zHPM
  - Get virtual server properties
  - Update virtual server properties
  - List virtual server detail for all virtual servers

RESTful APIs to interact with zHPM - II

- Metrics
  - Get raw metrics for the hypervisor
  - Get calculated metrics for the hypervisor
  - Get raw and calculated metrics for workload resource groups
  - Get velocity goal range for mappings
  - Get virtual server raw and calculated metrics for a workload resource group

- APIs for Virtual server CPU management
  - Get current CPU management setting
  - Enable/Disable CPU management
  - Report dynamic resource adjustments

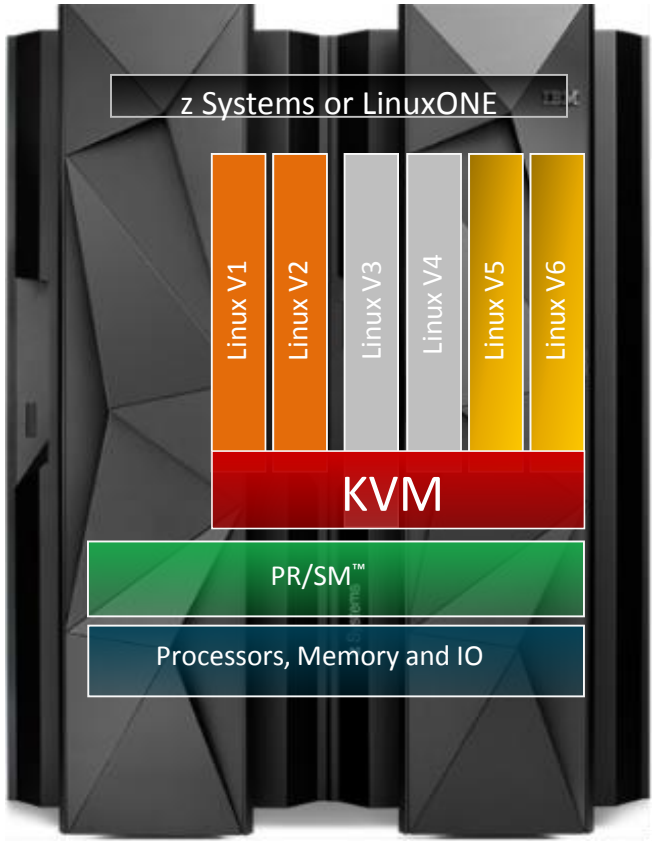
- APIs for zHPM diagnostics
  - Generate diagnostic dump
  - List current trace settings
  - Update current trace settings

Usage Case 1: hypothetical scenario

In this basic scenario, the hypothetical customer has a set of 6 virtual servers running development, test, and production workloads.

❑ There are six virtual servers (VS1-VS6) in a KVM for IBM z Systems™ hypervisor, and all are running WebSphere® Application Server (WAS).

- Two of these virtual servers, ***VS1 and VS2***, are running WAS for a development team.
- Two of these virtual servers, ***VS3 and VS4***, are running WAS for a test team.
- Two of these virtual servers, ***VS5 and VS6***, are running WAS for production.



Usage Case 1: identify workload resource groups

- ***WasDev*** workload resource group for development virtual servers:
  - is associated with virtual servers VS1 and VS2
  - has business importance level of low
  - has a user-defined service class called *WasDevServiceClass* with a business importance of medium and a performance goal of moderate.

- ***WasTest*** workload resource group for test virtual servers:
  - is associated with virtual servers VS3 and VS4
  - has business importance level of medium
  - has a user-defined service class called *WasTestServiceClass* with a business importance of medium and a performance goal of moderate.

- ***WasProd*** workload resource group for production virtual servers:
  - is associated with virtual servers VS5 and VS6
  - has business importance level of highest
  - has a user-defined service class called *WasProdServiceClass* with a business importance of medium and a performance goal of fast.

Usage Case 1: policies

- Sample Policy File for *WasProd* workload resource group, Wasprod.pp

```
{
   "performance-policy": {
      "perf-policy-info": {
         "name": "Production_Policy",
         "description": "WAS Production Performance Policy",
         "business-importance": "medium"
      },
      "service-classes": [{
         "name": "WasProdServiceClass",
         "description": "WAS Production Service class",
         "business-importance": "highest",
         "velocity-goal": "fast",
         "cpu-critical": true,
         "virtual-server-name-filters": [“.*”]
      }]
}
```

See the following link for other policy file samples:
http://www.ibm.com/support/knowledgecenter/SSNW54_1.1.1/com.ibm.kvm.v111.admin/usecase1.htm?lang=en

Usage Case 1: set up workloads and policies

- Run the following commands to set up the workloads and policies. (The policy files are expected to be in the directory where the commands are run.)
  - Set up the development workload resource group with its policy:
    # zhpm wrg-create --wrg-name WasDev --description 'WAS Development Sandbox' --perf-policy-file=./WasDev.pp
  - Set up the test workload resource group with its policy:
    # zhpm wrg-create --wrg-name WasTest --description 'WAS Test Sandbox' --perf policy-file=./WasTest.pp
  - Set up the production workload resource group with its policy:
    # zhpm wrg-create --wrg-name WasProd --description 'WAS Production Servers, be Careful!' --perf-policy-file=./WasProd.pp

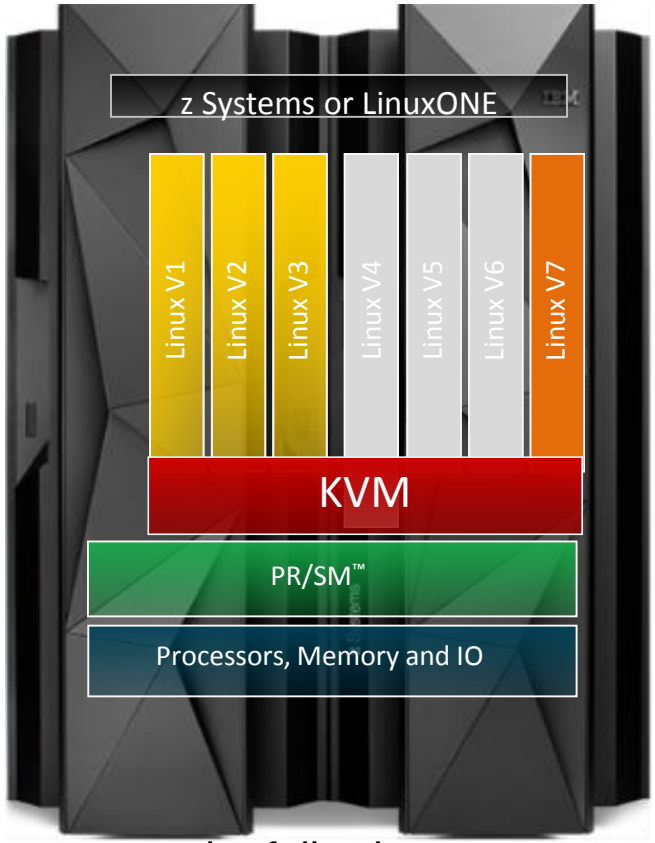
Usage Case 1: associate virtual servers with workloads

- Associate the virtual servers with the corresponding workload resource groups as follows:
  - Associate virtual servers running development work with *WasDev* workload resource group:
    # zhpm vs-wrg-add --wrg-name WasDev --vs-name VS1
    # zhpm vs-wrg-add --wrg-name WasDev --vs-name VS2
  - Associate virtual servers running test work with *WasTest* workload resource group:
    # zhpm vs-wrg-add --wrg-name WasTest --vs-name VS3
    # zhpm vs-wrg-add --wrg-name WasTest --vs-name VS4
  - Associate virtual servers running development work with *WasProd* workload resource group:
    # zhpm vs-wrg-add --wrg-name WasProd --vs-name VS5
    # zhpm vs-wrg-add --wrg-name WasProd --vs-name VS6

- Enable CPU management. Then zHPM will optimize CPU resource utilization based on business importance and performance goals.
  # zhpm config --cpu-mgmt on

Usage Case 2

- This scenario assumes that the hypothetical customer is running two sets of three-tiered Web applications and some CPU intensive batch workloads. The customer has a set of virtual servers running development, test, and production workloads.



  - There are seven virtual servers (VS1-VS7) in a KVM for IBM z Systems™ hypervisor.
  - ***VS1, VS2, and VS3*** are supporting three-tier Web transactions by running IHS, WAS, and DB2®, respectively. These transactions are considered to be *very important* to the business.
  - ***VS4, VS5, and VS6*** are supporting three-tier Web transactions by running IHS, WAS, and DB2, respectively. These transactions are considered to be *less important* to the business as compared to the transactions running in VS1, VS2, and VS3.
  - ***VS7*** runs nightly backup processes.

For policy file samples, commands to set up the environment, see the following: *http://www.ibm.com/support/knowledgecenter/SSNW54_1.1.1/com.ibm.kvm.v111.admin/usecase2.htm?lang=en*

Summary

- Bring workload-aware virtualization management to KVM for IBM z Systems
  - Concepts derived from z/OS goal oriented management approach
    - Concept of a workload resource group
    - Policy-based goal-oriented monitoring and management of CPU resources
    - Goals are defined in terms of “velocity”
- Enable exploitation with higher level virtualization management solutions
  - REST APIs allow programmatic interaction with higher level virtualization management tools.
  - CLI allows local administration
- Evolve capabilities over time
  - New resource management capabilities like memory management

DEMO: see zHPM in action

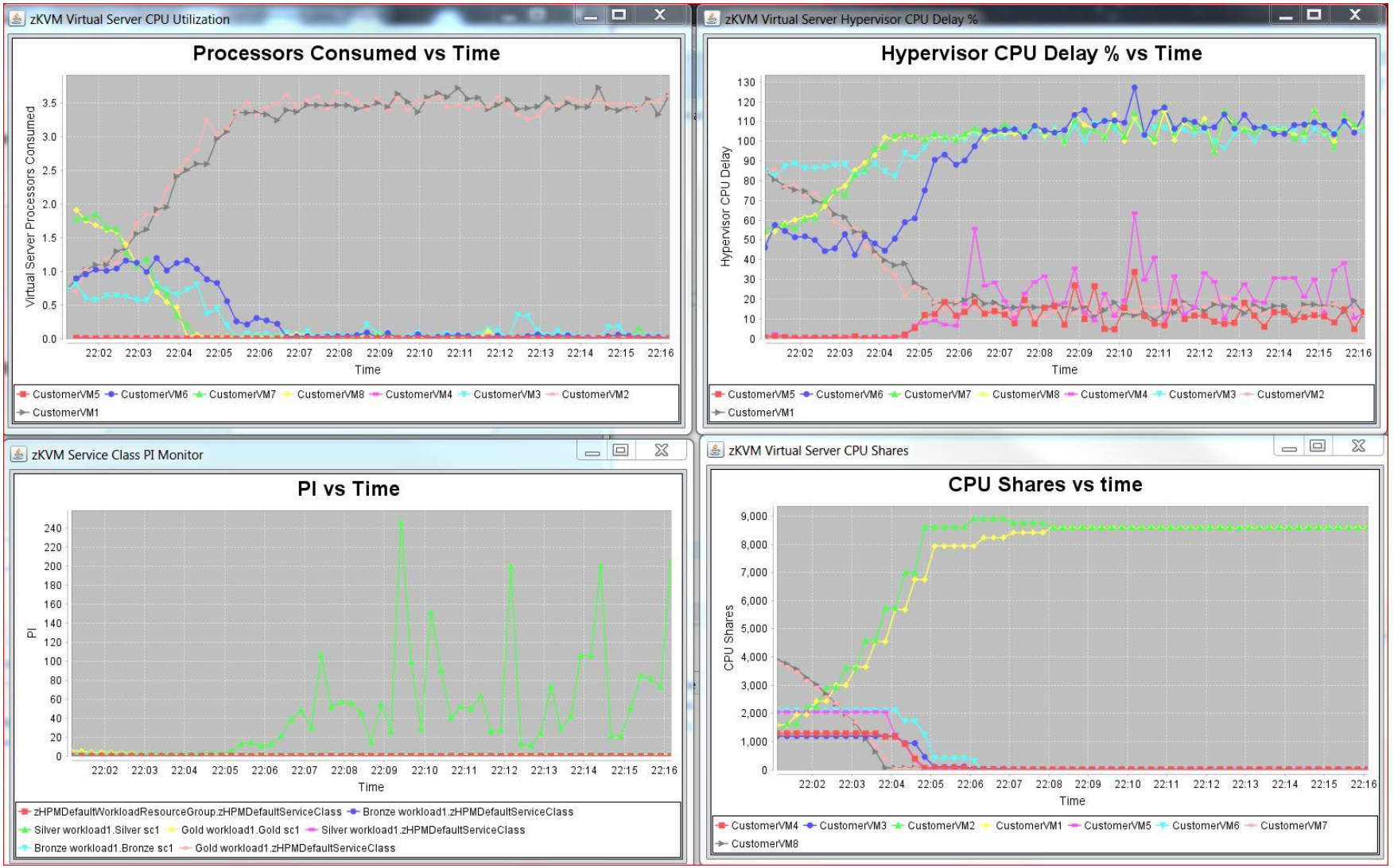
Demo set-up

- Hypervisor has 8 CPUs
- 8 Virtual Servers in the hypervisor, all running CPU-intensive work like Monte-Carlo simulations
- 3 Workloads
  - ***Gold workload*** has the *highest* business importance and has a service class with *fastest* performance objective
  - ***Silver workload*** has *medium* level business importance and has a service class with *moderate* performance objective
  - ***Bronze workload*** is the least important workload and has a service class with *slowest* performance objective
- Virtual Servers “CustomerVM1” and “CustomerVM2” are part of the Gold workload
- Virtual Servers “CustomerVM7”, “CustomerVM8” are in Bronze workload
- The rest of the virtual servers are in Silver workload

Expected Result

As work increases on important virtual servers and performance objectives are being missed, zHPM takes CPU actions to help important work. Least important workloads get hit, first.

Expected run captured in real time monitors



Now, start the demo run

- *Turn CPU management on*

For More Information on KVM for IBM z Systems

- Portal: KVM for IBM z Systems
  - http://www-03.ibm.com/systems/z/solutions/virtualization/kvm/

- Product documentation at http://www-01.ibm.com/support/knowledgecenter/linuxonibm/liaaf/lnz_r_kvm.html
  - KVM for IBM z Systems: Planning and Installation Guide SC27-8236-00
  - KVM for IBM z Systems: Administration Guide SC27-8237-00
  - Linux on z Systems: Virtual Server Management SC34-2752
  - Linux on z Systems: Virtual Server Quick Start SC34-2753
  - Linux on z Systems: Device Drivers, Features, and Commands for Linux as a KVM Guest SC34-2754
  - Linux on z Systems: Installing SUSE Linux Enterprise Server 12 as a KVM Guest SC34-2755

- Redbook: Getting Started with KVM for IBM z Systems
  - http://www.redbooks.ibm.com/redpieces/abstracts/sg248332.html?Open

- **Hypervisor Performance Manager on IBM Knowledge Center**
  - http://www.ibm.com/support/knowledgecenter/SSNW54_1.1.1/com.ibm.kvm.v111.admin/part2.htm?lang=en

- IBM z Systems Development Blog
  - https://www.ibm.com/developerworks/community/blogs/e0c474f8-3aad-4f01-8bca-f2c12b576ac9/entry/Dynamic_Resource_Management_with_KVM_for_IBM_z_Systems_and_LinuxOne?lang=en

Q & A

THANK YOU