Virtualization Basics

Your name here
Software Engineer
z/VM Development Lab
Endicott, NY

built on IBM Virtualization Technology
Virtualization (As told from the view of z/VM)
• What is it
• A brief history
• How it's used today
What is a virtual machine?
Introduction to basic concepts
  • Processor, Memory, I/O, etc...
Overcommitment of resources
How do you build z/VM?
Dynamic resource management
Monitoring your system
What happens when things go wrong?
# Z Nomenclature

<table>
<thead>
<tr>
<th><strong>Intel, System p, etc.</strong></th>
<th><strong>IBM Z</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>Storage (central storage)</td>
</tr>
<tr>
<td>Disk Storage</td>
<td>DASD (Direct Access Storage Device)</td>
</tr>
<tr>
<td></td>
<td>FCP (Fibre Channel Protocol)</td>
</tr>
<tr>
<td>Computer Host System</td>
<td>CEC (Central Electronics Complex)</td>
</tr>
<tr>
<td></td>
<td>CPC (Central Processor Complex)</td>
</tr>
<tr>
<td></td>
<td>Server</td>
</tr>
<tr>
<td>Socket Slot</td>
<td>Book, drawer or node</td>
</tr>
<tr>
<td>Chip (e.g., multiple cores on a chip)</td>
<td>Core (e.g., multiple processors on a core)</td>
</tr>
<tr>
<td>CPU</td>
<td>Processor or Engine</td>
</tr>
<tr>
<td>GPU</td>
<td>CPU or PU</td>
</tr>
<tr>
<td>Processor</td>
<td>CP (Central Processor)</td>
</tr>
<tr>
<td>Core</td>
<td>IFL (Integrated Facility for Linux)</td>
</tr>
<tr>
<td>HT Core</td>
<td>zIIP (zSeries Integrated Information Processor)</td>
</tr>
<tr>
<td></td>
<td>ICF (Integrated Coupling Feature)</td>
</tr>
<tr>
<td></td>
<td>SAP (System Assist Processor)</td>
</tr>
<tr>
<td></td>
<td>IOP (I/O Processor)</td>
</tr>
<tr>
<td>Boot</td>
<td>IPL (initial program load)</td>
</tr>
</tbody>
</table>
Why Virtualize?
Virtualization is the creation of substitutes for real resources.

One physical resource can look like multiple virtual resources.

Virtual resources can have functions or features that are not available in their underlying physical resources.

Benefits:

– Consolidation of physical resources reduces hardware cost
– Optimization of workloads that run near each other
– IT flexibility and responsiveness – make changes without a hardware outage!
Are we there yet?

1964
CP-40 development begun, this will be the predecessor of today's z/VM.

1972
In 1972, VM/370 is launched, the first in the VM product line from IBM.

1984
In 1984, Amdahl introduces MDF (multiple domain feature) to partition the real device into several separate machines. IBM releases LPAR technology in 1988.

2001
VMware ESX Server is launched, marking the start of a new wave of virtualization. z/VM, the latest version of the VM operating system, debuted in 2000.

2013
The first release of Docker, a containerization technology, released to open source.
Businesses continue to prize flexibility, portability and openness as they develop new applications and delve into piles of data. The future is bright for virtualization.
What are we virtualizing? Z Architecture

- Every computer system has an architecture
  - It's the specs for how to build software on the system
  - Defined in z/Architecture Principles of Operation
  - z/VM must both follow and implement z/Architecture
Virtualization enables four fundamental capabilities:

- resource sharing
- resource aggregation
- emulation of function
- insulation

What does virtualization mean on IBM Z?

*Virtualization* is the ability for a computer system to share resources so that one physical machine can act as many virtual machines.

*z/VM* allows the sharing of the mainframe's physical resources such as:

- Disk
- Memory
- Network adapters
- CPUs

These resources are managed by a *hypervisor*. 
What does Virtualization look like on IBM Z?

OS

HYPERVISOR

HARDWARE

TYPE 1

native
(bare metal)

CMS

z/OS

CP

Linux

z/VM CP

z/OS

LPAR

LPAR

CEC
Real, Logical, & Virtual

It's important that everyone understands which view YOU are talking about!

What the accounting team sees (REAL)

What the system programmer sees (LOGICAL)

What the application developer sees (VIRTUAL)
IBM has the entire Z stack, with special assists that bring the hardware and software closer together.

- Handshaking capabilities between the Hypervisor and Operating Systems
- Asynchronous Page Fault Processing
  - Guest virtual CPUs can execute other tasks while page(s) are brought back into memory
- QEBSM (QDIO Enhanced Buffer State Management)
  - QDIO data transfers are mediated by firmware and not z/VM

World-class clustering and mobility, allowing virtual machines to move to new hardware without an outage.

Why z/VM?

**Scale** – not just large machines, but fully utilized large machines

Resource **management** facilities for large and diverse workloads.

IBM has an awesome **community** of helpful folks. Ask questions at conferences, via listservs, and over Slack channels.
Virtual Machines
Just like the real thing, we have

- CPUs
- Memory
- Control Unit
- Disk
- Tape
- Printer
- Consoles

Operating system
Just lots of them!
Virtual Machines
- 85 LPARS per CEC
- 99,999 Guests per z/VM LPAR
- 7 levels of guests

Processors
- 170 Cores per CEC
- 80 Logical Cores per z/VM LPAR
- 64 Virtual Cores per z/VM guest

Memory
- 32 TB per CEC
- 2 TB per z/VM LPAR
- 1TB per z/VM guest

I/O
- 6 CSS, 4 Subchannel sets per CSS
- 65535 devices per z/VM LPAR
- 24576 devices per z/VM guest
Intro to Processors and z/VM

z/VM supports a certain number of processors (logical or virtual), on which a task or transaction can execute. A processor is the source of "power" for a virtual machine.

There are different engine types for different scenarios.

CP – general purpose

IFL – used to run LINUX on VM

z/VM and LPAR work together for efficiency. Unsure how much power you need? z/VM can work with LPAR to determine what processors you need and park the rest!

Running CPUs 90% busy doesn't scare us. We want you to get what you pay for!
Intro to Memory and z/VM

While each LPAR will carve up the real memory available to the box, virtual machines are welcome to as much virtual memory as they like.*

Fully utilizing your memory is not a bad thing - z/VM expects to have to move some memory to paging media.

You can reserve real memory for virtual machines, if you never want them to get paged out.

Virtual machines can share storage ranges.

*Terms and conditions apply – current VM maximum of 1TB, YMMV
Intro to I/O and z/VM

z/VM can accommodate an extremely large variety of I/O devices, allowing for incredible flexibility

- These can be storage devices, displays, printers, or just about anything else.

- In many cases, these devices are virtualized using system resources rather than actually being physically present.

- Even if z/VM doesn't support an I/O device that a user is trying to attach, it will generally allow that device to be attached anyway. The user will configure the device as needed.
Device Management Concepts

Different ways of getting from my house to the mall

- **Dedicated or attached**
  - My parents give me my own car to go to the mall
- **Virtualized**
  - I have to share the car keys with my whole family, so I can go to the mall sometimes
- **Simulated**
  - I shop online so I don't have to go anywhere physically
- **Emulated**
  - I take the bus to the mall
Device Management Concepts

Linux 1
Minidisk 1
Minidisk Cache (High-speed, in-memory disk cache)

Linux 2
Minidisk 2

Linux 3
Minidisk 3

TDISK 1
Minidisk Cache (High-speed, in-memory disk cache)

TDISK 2
Minidisk Cache (High-speed, in-memory disk cache)

TDISK 3
Minidisk Cache (High-speed, in-memory disk cache)

DS8000® Storage Server

LUN

SCSI Disks attached to z/VM appear to guests and rest of VM as emulated FBA.
FBA = Fixed Block Architecture
Examples of devices that z/VM supports:

<table>
<thead>
<tr>
<th>Device type</th>
<th>Model numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct-Access Storage Devices (DASD)</td>
<td>3380, 3390</td>
</tr>
<tr>
<td>Tapes</td>
<td>3480, 3490, 3590</td>
</tr>
<tr>
<td>Printers</td>
<td>3268, 3287</td>
</tr>
<tr>
<td>Unit Record Devices</td>
<td>3505 reader, 3525 punch</td>
</tr>
<tr>
<td>Displays</td>
<td>3270</td>
</tr>
<tr>
<td>Communication Controllers</td>
<td>3745</td>
</tr>
<tr>
<td>OSA Devices</td>
<td>OSA-2, OSA-Express, OSA-Express2</td>
</tr>
</tbody>
</table>
FICON vs FCP vs PCIe

What's right for you?

Fiber Connection (FICON)
- Fiber-optic channel technology
- Well established on IBM Z
- Command driven

Fiber Channel Protocol (FCP)
- Small Computer System Interface (SCSI) protocol
- Utilizes Fiber channel connection

Peripheral Component Interconnect Express (PCIe)
- PCI standard used to connect high-speed devices
- New kid on the block
Intro to Networking and z/VM

Options include:

- Virtual Switch
  - MAC or IP oriented (Layer 2 or Layer 3)
  - Link Aggregation
  - Global VSwitch can span multiple LPARs
- HiperSockets
  - HiperSockets VSwitch Bridge
- OSA – Physical adapter
  - Ports can be shared or dedicated
  - Vswitch can be connected to one to give external connectivity
Intro to SPOOL in z/VM

- z/VM simulates unit-record devices for virtual machines
  - Simulated reader, punch, printer, console for every virtual machine

- Simulated UR devices are still widely used on z/VM; upcycled to
  - Exchange application data
  - Share data files among multiple virtual machines
  - Collect console output
  - Store data collected by various system functions

- "Intermediary" between information in memory and information that you might devote long term storage to

- Data in form of "files" is stored on disks that are owned by the z/VM system
Intro to Security and z/VM

Crypto accelerators offload the hard math of cryptography, making security faster and easier.
Intro to Crypto and z/VM

CP Assist for Cryptographic Function (CPACF)
- Included in every CP and IFL processor
- Features:
  - Support for symmetric and hashing algorithms
  - Pseudo-random number generator (TRNG on some models)
- Guest use
- Host use
  - z/VM TLS/SSL server
  - Encrypted paging
  - RACF Password Encryption

Crypto Express cards
- Features
  - Asymmetric and hashing algorithm offload
  - Host master key storage (HSM)
  - Hardware RNG
  - PKCS #11 cryptographic support
- Guest use
  - Directly attached (APDED)
  - Shared (APVIRT)
- Host use
  - z/VM TLS/SSL server
Overcommitment

Doing more with less
Resource Overcommitment

What is it?

- When the virtual resources defined to an OS (Or multiple OS's) exceed the physical resources of the machine
- Fundamental to the idea of virtualization
- Possible because a virtual resource will not need 100% of the physical resource 100% of the time
- Enables data center consolidation
- What things can be overcommitted? CPU, Memory, I/O, Networking, Crypto
Two logical processors defined to LPAR

z/VM running a z/OS and a Linux guest, each guests with two virtual CPUs

This configuration (four virtual CPUs backed by two logical CPUS) is overcommitted 2:1

How's it work?

- When vCPUs need to run they must be dispatched onto the logical processor
- No two vCPUs can be dispatched on the same logical processor at the same time
- Hypervisor must manage vCPU access to the LPU
Tools of CPU management

- Entitlement – The amount of real CPU time a logical partition is entitled based on LPAR configuration.

- Processor Topology – The amount of real CPU a logical processor is entitled
  - Vertical High – entitled to 100% of a real CPU
  - Vertical Medium – entitled to 50-100% of a real CPU
  - Vertical Low – entitled to 0% of a real CPU

- Consumption/Utilization - How much of a real CPU is being used (for LPAR or for entire CEC)

- Share Setting – How much entitlement is due after comparison to other users

- Processor Affinity – A vCPU that must run on a logical processor of the same type
Determining "Who's on first?!"

Gather Partition Information:
- Processor Topology
- LPAR weights
- Processor Utilization

Gather Hypervisor Information:
- Share Settings
  - Absolute/Relative
- Hard/Soft
- Resource Pool Information
- Processor Affinity

Schedule vCPU based on relative priority:
- Use partition and Hypervisor information to choose the best logical processor to run on.

Dispatch vCPU onto Logical Processor:
- Run CPU under SIE
- Let vCPU run for AT LEAST a Minor Time Slice.
- Compete with other vCPUs for unused or Elapsed Time Slice.
Overcommitting Memory

• 1TB defined to LPAR
• z/VM running 50 Linux guests, each with 50GB of memory = 2500 GB ~ 2.5 TB
• This configuration is memory overcommitted ~2.5:1

How's it work?
• We can only keep so much in memory
  • The 1TB must have z/VM's own structures and memory as well as guest memory
• So we choose some memory to page out to disk
  • Both guest and host memory are up for grabs
  • Try to make intelligent choices about what isn't needed, working hand in hand with the dispatcher
• We bring back in paged out memory when it is needed
What is a VSWITCH?

- Similar to an ordinary network switch, but with the ability to interface with both physical and virtual devices simultaneously
- The system recognizes the result as a single LAN segment
- VSWITCHes can be used to overcommit an OSA, as all of the physical and virtual switches attached are recognized as one single LAN and subnet
How do you Build z/VM?
What do you need?

- LPAR definition - HMC/SE, what PR/SM controls
- IOCDS - I/O definition – or use DPM to give you both IOCDS & PR/SM
- System config – What's defined to the z/VM system
- User Directory – What's defined to the z/VM users
System Config: What makes up my system?

What is defined in the system configuration file?

• Operator_Consoles statement to find a console
• vSWITCH statements to define our virtual network
• CP_Owned statements to define the residence volume
• And so much more!!!

What is defined in a directory entry?

- Login credentials
- Authorizations
- Access to system resources
- Command privileges
- Virtual machine characteristics

Tools to help you

- DIRECTXA – creates and compiles your user directory
- Directory manager – Does this work for you!

USER LINUX01 MYPASS 512M 1024M G
MACHINE ESA 2
IPL 190 PARM AUTOCR
CONSOLE 01F 3270 A
SPOOL 00C 2540 READER *
SPOOL 00D 2540 PUNCH A
SPOOL 00E 1403 A
SPECIAL 500 QDIO 3 SYSTEM MYLAN
LINK MAINT 190 190 RR
LINK MAINT 19D 19D RR
LINK MAINT 19E 19E RR
MDISK 191 3390 012 001 ONEBIT MW
MDISK 200 3390 050 100 TSOBIT MR
Dynamic Resource Management
Dynamic Processor Management

- Many ways to manage CPU resource dynamically
- No IPL required
- Can dynamically grow CPU footprints for peak service windows
- Hiper-dispatch can consolidate workload onto a fewer number of cores if full resource not needed

VARY PROC/CORE

SMT Enabled – CORE
SMT Disabled – PROC

Allows a user to bring an offline core defined to LPAR into a z/VM configuration, or takes an online core out of a z/VM configuration

DYNAMIC SMT

While SMT enabled, allows you to dynamically switch MT threading levels.

HIPER-DISPATCH (Park/Unpark)

Based on current and projected consumption, will logically VARY off threads/cores that aren't needed.

Adjustable via:
CP SET SRM UNPARKING
Dynamic Memory Management

- Many ways to manage memory resources dynamically
- No IPL required*
- Memory can be taken from one LPAR and given to another
- Virtual memory can be added to or removed from guests
- Linux itself can bring memory on and offline

Define Storage

Used to change the memory on a guest.

Memory may be defined as online or in STANDBY state to the guest.

Set Storage

Use to add more memory to your z/VM system.

Memory to be added must be available in the STANDBY pool.

On the z14 and later machines, use this command to remove memory. Only memory in the RECONFIGURABLE pool may be removed.

Removed memory will go to the STANDBY pool.

Chmem

Linux commands are provided to bring more memory online, or take memory offline to the Linux OS. Memory is added to and removed from the STANDBY pool for the guest.
What needs a planned outage?

- Hardware upgrades
- Enabling/Disabling MT
- Updating maximum real memory size
- Applying service
- Adding/Removing certain types of devices
Single Systems Image (SSI) Feature

Simplified Management of Multi-z/VM environment

1. Manage a single user directory

2. Cluster management from any member
   - Apply maintenance to all members in the cluster from one location!
   - Issue commands from one member to operate another!

3. Built-in cross-member capabilities

4. Resource coordination and protection of network and disks

5. Relocate a running Linux guest from one system to another

FREE in z/VM 7.1
The INDICATE and MONITOR commands provide a system performance measurement facility for z/VM. These commands provide a method of:

1. **Early Detection**
   An unfavorable trend can be detected early so that corrective action can be taken before it develops into a serious performance problem. Early detection is done by tracking key performance indicators over time and comparing them to established limits of acceptable performance. These indicators are often various measures of response time.

2. **Basis for Performance Problem Determination**
   After a problem has been identified, the monitor data serves as the basis for determining the likely cause of the problem. This is done by comparing the current data that reflects the problem to past data collected when performance was adequate.

**CP monitor data** is the key source of information normally used to monitor the performance of a z/VM system.

The **INDICATE** and **MONITOR** commands provide a system performance measurement facility for z/VM. These commands provide a method of:

- Obtaining system resource usage data while z/VM is operating so that steps can be taken then or later to improve performance
- Collecting measurement data using a z/VM monitor for later analysis by system analysts.

**Performance Toolkit for VM** assists operators and system programmers with system console operation in full screen mode, and with performance monitoring on z/VM systems.
There are ways to connect into the system to monitor what's going on at a deeper level, or to issue commands to the hardware. These system services are for privileged users only and usually require special permissions. Examples:

- VMevent – feed of system events
- Monitor – performance and monitoring data
- Accounting – events that affect chargeback
- Security – anything that changes permission or access
Data sources:
- Guests
- CP itself
Houston...we have a problem
<table>
<thead>
<tr>
<th>What happened?</th>
<th>Soft ABEND</th>
<th>Hard ABEND</th>
<th>Hung User</th>
<th>Hung System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Something went wrong in the program, but CP is ok.</td>
<td>- Something went wrong with the control program.</td>
<td>- A specific user is stuck in CP and cannot continue.</td>
<td>- The entire system is stuck in CP and cannot continue.</td>
</tr>
<tr>
<td></td>
<td>- System stays up, Virtual Machine may terminate.</td>
<td>- The entire system terminates.</td>
<td>- The rest of the system continues operation.</td>
<td></td>
</tr>
</tbody>
</table>
Determining Why

CP will dump system storage to aid debug for a hard or soft abend.

**SNAPDUMP** or **RESTART** dumps help gather information for system or user hangs.

**IMPORTANT**: configuring your system with adequate dump space will reduce headaches should you need to use it!

Dumping z/VM is not instantaneous and can take on the order of minutes depending on the size of your system. The longer you wait, the more information you’ll have.
If you stay ready, you don't have to get ready
Installing z/VM
Installing z/VM

- z/VM install guide has worksheets to help you gather the information you need, fill these out first!
- Many shops install z/VM second level to try out the process and test new levels

Optional Features

If you've purchased DirMaint, Perfkit, RACF or any additional paid features, have their Program Directories available when you install.

There are special codes necessary to enable these.

Have your device information ready before you start installation: Disk, Networking for the z/VM TCPIP stack, CTCs if you're using SSI.

Pictures help! As much as you can, make layouts of your device configurations, networking layouts, and SSIs.
Servicing z/VM
Servicing z/VM

• Every 2 years, in the 3rd quarter, z/VM releases a new level of the product.
  • We'll always have one "stable" and one "new function" release in service
  • Releases will be out of service 6 months after the N+2 release comes out (e.g., z/VM 6.4 would end service 6 months after z/VM 7.2 was available)
• z/VM 6.4 is our current stable release
• z/VM 7.1 is our current new function release
• In between releases, new function and fixes are shipped via APARs.

PMR
Problem Management Report
what you open when you find a problem,
e.g., 65118,442,000

APAR
Authorized Program Analysis Report
for new function and problem fixes between releases,
e.g., VM66209

PTF
Program Temporary Fix
actual code for the APAR,
e.g., UM35371

RSU
Recommended Service Upgrade
group of PTFs we think all customers should apply,
e.g., z/VM 7.1 RSU 1901
Where to go when you're stuck

Community Forums
- IBMVM Listserv (z/VM community) http://listserv.uark.edu/cgi-bin/wa?A0=IBMVM
- LINUX 390 Listserv (Linux on Z) http://www.cavmen.org/list390.html
- Additional Listserv can be found here https://www.vm.ibm.com/techinfo/listserv.html

Technical Publications

Redbooks

Ask an IBMer
- Attend Conferences
- Engage User Groups
- Contact the z/VM experts! https://www.vm.ibm.com/forms/
Conclusion

Flexibility

Hardware Integration

Community

Testing and Debugging

Overcommitment