B51

Virtualization Basics

Brian K. Wade, Ph.D.

bkw@us.ibm.com

IBM System z Expo
September 17-21, 2007
San Antonio, TX

Trademarks

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

CICS*
DB2
DB2 Connect
DB2 Universal Database
e-business logo*
FICON
HiperSockets
IBM*
IBM logo
MQSeries*
Multiprise*
OS/390
RISC
S/390
S/390 Parallel Enterprise Server*
Virtual Image Facility
VM/ESA*
VSE/ESA
WebSphere
z/OS
z/VM
zSeries

* Registered trademarks of the IBM Corporation

The following are trademarks or registered trademarks of other companies.

Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation.
Tivoli is a trademark of Tivoli Systems Inc.
Linux is a registered trademark of Linus Torvalds.
Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
UNIX is a registered trademark of The Open Group in the United States and other countries.
Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.

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.

IBM considers a product "Year 2000 ready" if the product, when used in accordance with its associated documentation, is capable of correctly processing, providing and/or receiving date data within and between the 20th and 21st centuries, provided that all products (for example, hardware, software and firmware) used with the product properly exchange accurate date data with it. Any statements concerning the Year 2000 readiness of any IBM products contained in this presentation are Year 2000 Readiness Disclosures, subject to the Year 2000 Information and Readiness Disclosure Act of 1998.

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.

Credits

- People who contributed ideas and charts:
  - Alan Altmark
  - Bill Bitner
  - John Franciscovich
  - Reed Mullen
  - Brian Wade
  - Romney White
- Thanks to everyone who contributed!

Introduction

- We'll explain basic concepts of System z:
  - Terminology
  - Processors
  - Memory
  - I/O
  - Networking
- We'll see that z/VM *virtualizes* a System z computer:
  - Virtual processors
  - Virtual memory
  - ... and so on
- Where appropriate, we'll compare or contrast:
  - PR/SM or LPAR
  - z/OS
  - Linux

Terminology

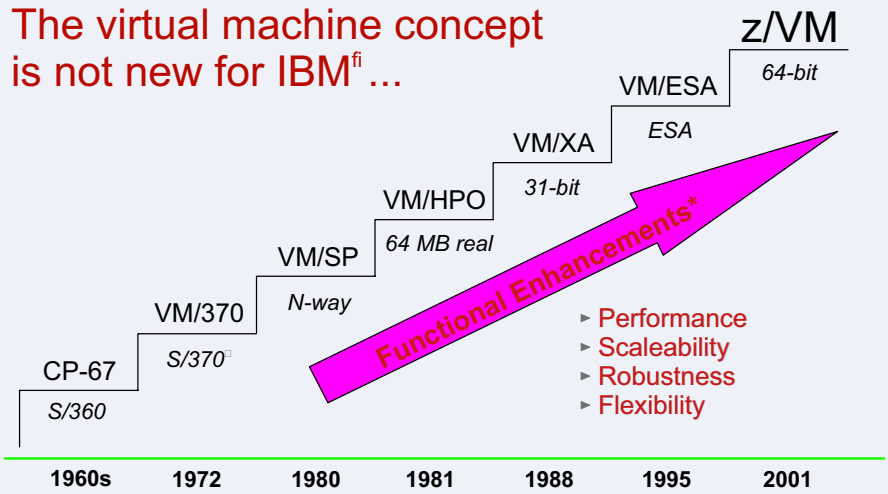
IBM System z Expo
September 17-21, 2007
San Antonio, TX

System z Architecture

- Every computer system has an *architecture*.
  - Formal definition of how the hardware operates
  - It's the hardware's functional specification
  - What the software can expect from the hardware
  - *What the hardware does*, not how it does it
- IBM's book z/Architecture Principles of Operation defines System z hardware
  - Instruction set
  - Processor features (registers, timers, interruption management)
  - Arrangement of memory
  - How I/O is to be done
- Different *models* implement the architecture in different ways.
  - How many processors there are
  - How the processors connect to the memory bus
  - How the cache is arranged
  - How much physical memory there is
  - How much I/O capability there is
- z900, z800, z990, z890, z9 EC, and z9 BC are all *models* implementing z/Architecture.

IBM Virtualization: Evolution

The virtual machine concept is not new for IBM fi ...



* Investments made in hardware, architecture, microcode, software

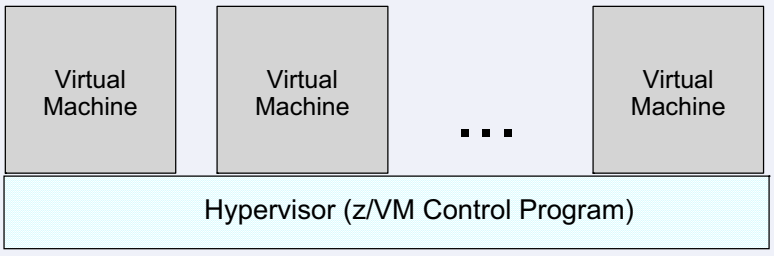
System z Parts Nomenclature

Intel, pSeries, etc.	zSeries
Memory	Storage (though we are moving toward "memory")
Disk, storage	DASD (Direct Access Storage Device)
Processor	Processor, CPU (central processing unit), engine, IFL (Integrated Facility for Linux), IOP (I/O processor), SAP (system assist processor), CP (central processor), PU (processing unit), zAAP (zSeries Application Assist Processor), zIIP (zSeries Integrated Information Processor)
Computer	CEC (central electronics complex)

Virtual Machines

IBM System z Expo
September 17-21, 2007
San Antonio, TX

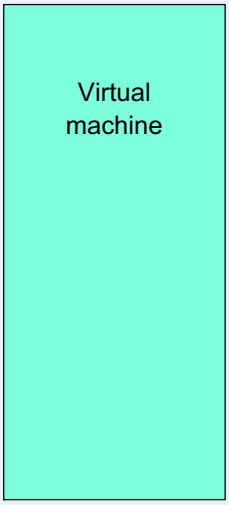
What: Virtual Machines



A **virtual machine** is an execution context that obeys the architecture.

The purpose of z/VM is to **virtualize** the real hardware:
- Faithfully replicate the z/Architecture Principles of Operation
- Permit any virtual configuration that could legitimately exist in real hardware
- Let many virtual machines operate simultaneously
- Allow overcommittment of the real hardware (processors, for example)
- Designed for many thousands of virtual machines per z/VM image (I have seen 40,000)
- Your limits will depend on the size of your physical System z computer

What: A Virtual Machine



z/Architecture

512 MB of memory

2 processors

Basic I/O devices:
- A console
- A card reader
- A card punch
- A printer

Some read-only disks

Some read-write disks

Some networking devices

We permit any configuration that a real System z machine could have.

In other words, we completely implement the z/Architecture Principles of Operation.

There is no "standard virtual machine configuration".

How: VM User Directory

Definitions of:

- memory
- architecture
- processors
- spool devices
- network device
- disk devices
- other attributes

```
USER LINUX01 MYPASS
CLASS G
STORAGE 512M
MAXSTORAGE 1024M
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 TWOBIT MR
```

Getting Started

- IML
  - Initial Machine (or Microcode) Load
  - Power on and configure processor complex
  - *LOGON* or *SET MACHINE*
  - Analogous to LPAR *image activation*
- IPL
  - Initial Program Load
  - Like *booting* a Linux system
  - z/VM lets us *IPL* a system in a virtual machine
  - Analogous to LPAR *load* function

How: CP Commands

- CP DEFINE
  - Adds to the virtual configuration somehow
  - CP DEFINE STORAGE
  - CP DEFINE PROC
  - CP DEFINE *{device} {device_specific_attributes}*
- CP ATTACH
  - Gives an entire real device to a virtual machine
- CP DETACH
  - Removes a device from the virtual configuration
- CP LINK
  - Lets one machine use another's disk device (read-write, read only)
- Changing the virtual configuration after logon is considered normal

Processors

IBM System z Expo
September 17-21, 2007
San Antonio, TX

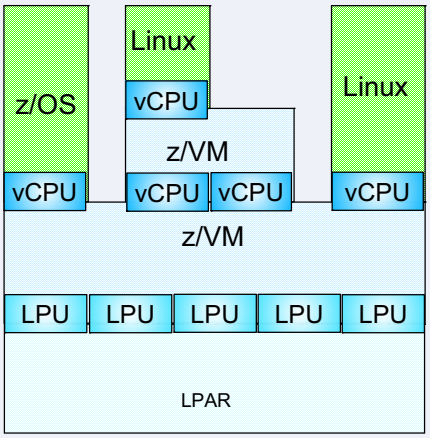
What: Processors

- Configuration
  - Virtual 1- to 64-way
    - Defined in user directory, or
    - Defined by CP command
  - We call these dispatchable units *virtual processors*
  - A real processor can be dedicated to a virtual machine
- Controls and Limits
  - Scheduler selects virtual processors according to apparent CPU need
  - "Share" setting - prioritizes real CPU consumption
    - Absolute or relative
    - Target minimum and maximum values
    - Maximum values ("limit shares") either hard or soft
  - Share for virtual machine is divided among its virtual processors

Phrases Associated with Virtual Machines

- In VM:
  - *Guest:* a system operating in a virtual machine, also known as *user* or *user ID* or *userid*
  - *Running under VM:* running a system as a guest of VM
  - *Running on (top of) VM:* same as *running under VM*
  - *Running second level:* running VM as a guest of VM
- In relationship to LPAR (partitioning):
  - *Logical partition:* LPAR equivalent of a virtual machine
  - *Logical processor:* LPAR equivalent of a virtual processor
  - *Running native:* running VM on the bare hardware, that is, without LPAR present
  - *Running in basic mode:* same as *running native*

What: Logical and Virtual Processors



How: Start Interpretive Execution (SIE)

- SIE = "Start Interpretive Execution", an instruction
- z/VM (like the LPAR hypervisor) uses SIE to "run" virtual processors
- Our processors contain special hardware (registers, etc.) to make SIE fast
- SIE has access to:
  - A control block describing the virtual processor state (registers, etc.)
  - The Dynamic Address Translation (DAT) tables for the virtual machine
- z/VM gets control back from SIE for various reasons:
  - Page faults
  - I/O channel program translation
  - Privileged instructions
  - CPU timer expiration (dispatch slice end)
  - Other, including CP asking to get control for special cases
- CP can also shoulder-tap SIE from another processor to end SIE

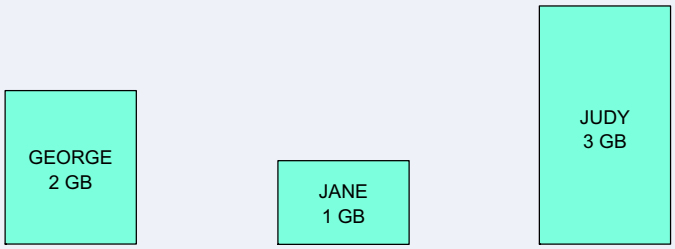
How: Scheduling and Dispatching

- VM
  - *Scheduler* determines priorities based on share setting and other factors
  - *Dispatcher* runs a virtual processor on a logical processor
  - Virtual processor runs for (up to) a *minor time slice*
  - Virtual processor keeps competing for (up to) an *elapsed time slice*
- LPAR hypervisor
  - Uses *weight* settings for partitions, similar to share
  - Dispatches logical processors on real engines
- Linux
  - *Scheduler* handles prioritization and dispatching of processes
  - Process runs for a time slice or *quantum*

Memory

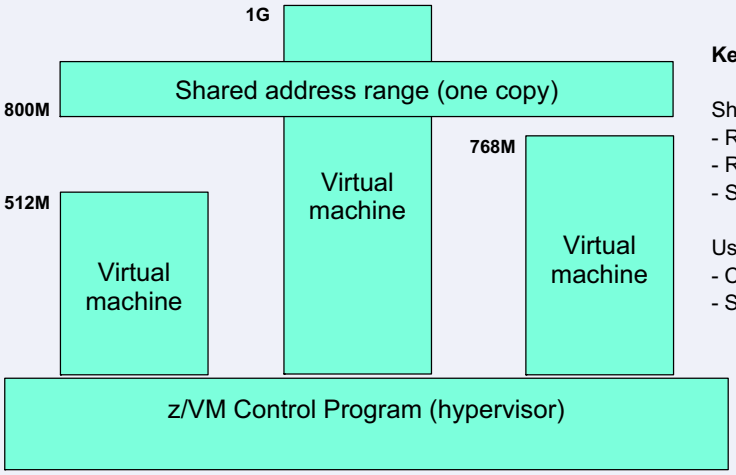
IBM System z Expo
September 17-21, 2007
San Antonio, TX

What: Virtual Memory



- Configuration
  - Defined in CP directory entry or via CP command
  - Can define storage with gaps (useful for testing)
  - Can attach expanded storage to virtual machine
- Controls and Limits
  - Scheduler selects virtual machines according to apparent need for storage and for paging capacity
  - Virtual machines that do not fit criteria are placed in the *eligible list*
  - Can *RESERVE* an amount of real storage for a guest's pages
  - Can *LOCK* certain specific guest pages into real storage

What: Shared Memory

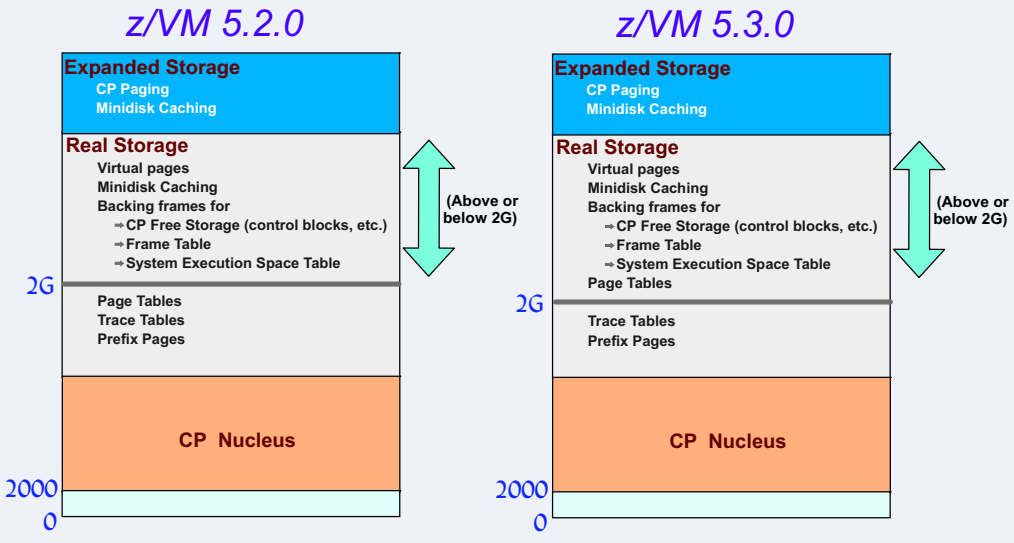


Key Points:

Sharing:
- Read-only
- Read-write
- Security knobs

Uses:
- Common kernel
- Shared programs

More: Layout of Real Storage



How: Memory Management

- VM
  - Demand paging between central and expanded
  - Block paging with DASD (disk)
  - Steal from central based on LRU with reference bits
  - Steal from expanded based on LRU with time stamps
  - Paging activity is traditionally considered normal
- LPAR
  - Dedicated storage, no paging
- Linux
  - Paging to "swap disks" (aka "swap extents")
  - Traditionally considered bad

I/O Resources

IBM System z Expo
September 17-21, 2007
San Antonio, TX

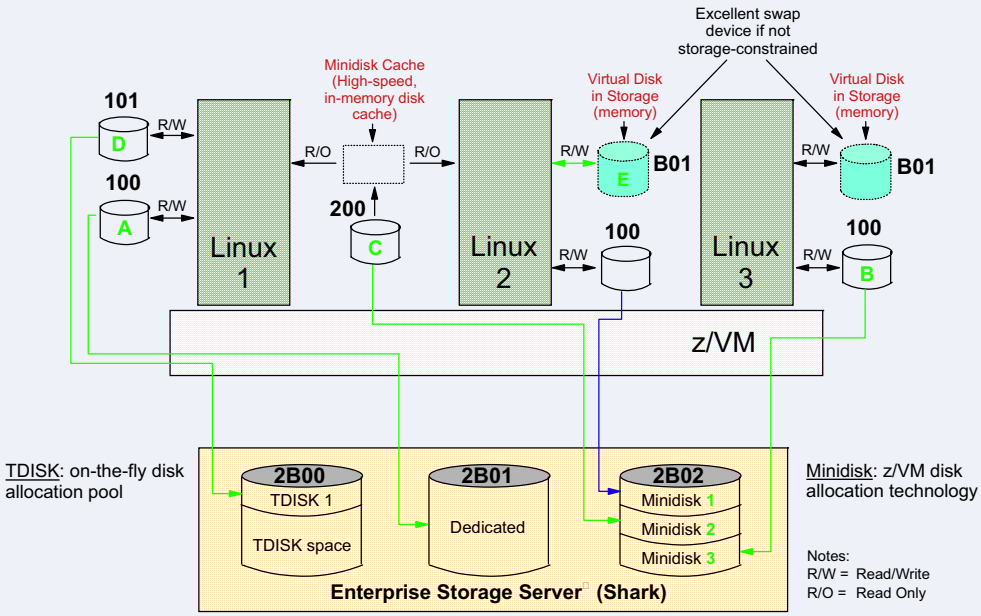
What: Device Management Concepts

- *Dedicated* or *attached*
  - The guest has exclusive use of the entire real device
- *Virtualized*
  - A slice (in time or in space) of a real device
  - E.g., DASD or crypto
- *Simulated*
  - No real hardware (all smoke and mirrors)
  - Virtual disks, virtual NICs, virtual channel-to-channel adapters (CTC)
- *Emulated*
  - Provide a device of one type using real hardware of a different type
  - E.g., CP can emulate FBA using SCSI LUNs

What: Device Management Concepts

- Vocabulary
  - *RDEV:* real device, or control block representing it
  - *VDEV:* virtual device, or control block representing it
  - *Subchannel:* Hardware-managed control block representing device in I/O operations
- Controls and Limits
  - Indirect control through share setting
  - Real devices can be "throttled"
  - Channel priority can be set for a virtual machine
  - MDC fair share limits (can be overridden)

What: Virtualization of Disks



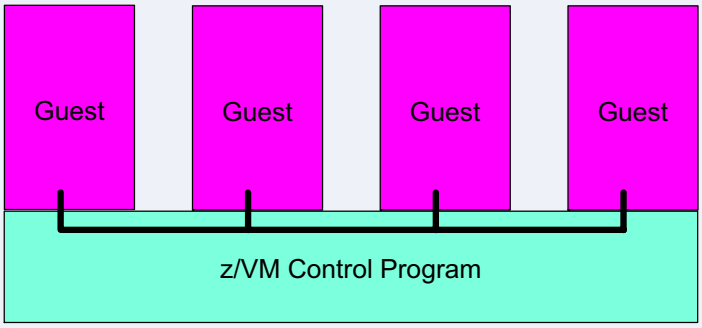
What: Data-in-Memory

- Minidisk Cache (MDC)
  - Caches reads of non-dedicated disks (minidisks)
  - Write-through cache
  - Great performance
  - Lots of tuning knobs
- Virtual Disk in Storage (VDISK)
  - Like a RAM disk, but pageable
  - Volatile (of course)
  - Appears to be an FBA disk
  - Can be shared among virtual machines
  - Plenty of knobs here too

Networking

IBM System z Expo
September 17-21, 2007
San Antonio, TX

What: Virtual Networking

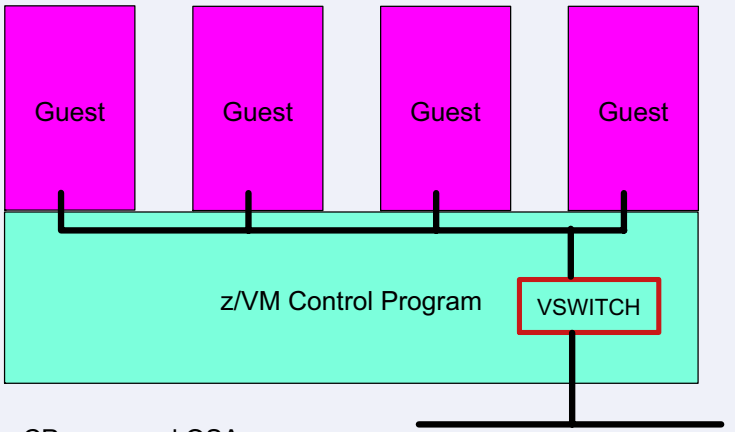


- Create a simulated LAN segment, called a *guest LAN,* inside z/VM
- Give the guest a simulated network adapter, called a *guest LAN adapter*
- *Couple* the fake adapter to the fake LAN
- At this point the guest is on a network

Thinking Outside the Box

- System z does have real networking hardware
  - *Open Systems Adapter (OSA):* a genuine LAN adapter
  - *HiperSocket:* a LAN adapter that talks only from one partition to another
- Who's going to own the real hardware?
  - The z/VM Control Program: the *Virtual Switch*
    - CP itself owns a real Open Systems Adapter
    - CP does the data switching functions
      - Layer 2 switch or layer 3 switch
    - The guests are effectively on the external network
  - A guest: a *virtual machine router*
    - A specific guest owns a real networking device (OSA or HiperSocket)
    - Said guest is also coupled to one or more guest LANs
    - This guest performs IP packet routing functions
    - IP subnet boundaries
    - Linux guest or VM TCP/IP stack

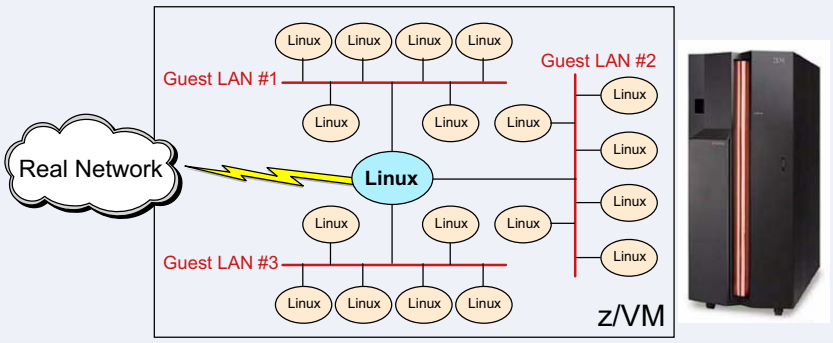
What: Virtual Switch



- CP owns real OSA
- Guests are on the external LAN

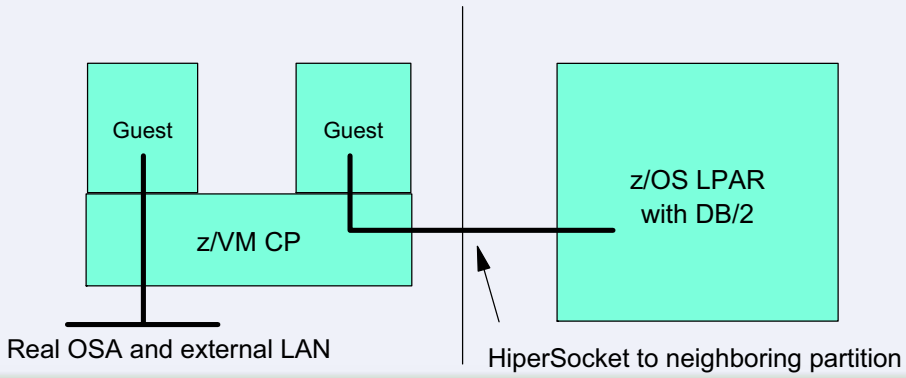
What: Virtual Machine Router

- One or more guest LANs
  - Each guest LAN is its own IP subnet
- Anointed guest is IP router
  - Is coupled to all the guest LANs
  - Also owns a piece of real network gear (OSA or HiperSocket)



There's Always Real Hardware

- Dedicate or attach a real OSA to the guest
- Dedicate or attach a real HiperSocket to the guest



Beyond Virtualization

IBM System z Expo
September 17-21, 2007
San Antonio, TX

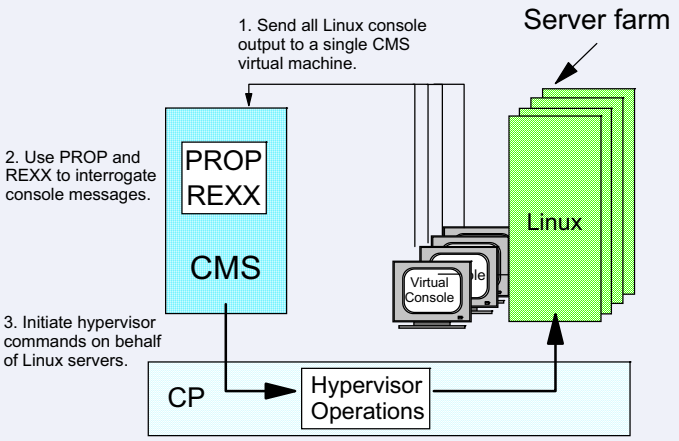
What: Other Control Program Interfaces

- Commands
  - Query or change virtual machine configuration
  - Debug and tracing
  - Commands fall into different *privilege classes*
  - Some commands affect the entire system
- Inter-virtual-machine communication
  - Connectionless (VMCF) or connection-oriented (IUCV) protocols
  - These pre-date TCP/IP, guest LANs, etc.
- System services
  - Guest has an enduring connection to CP via IUCV
  - Guest and CP exchange information on the IUCV connection
  - Guest is in a long-term relationship with CP (performance, accounting, security)
- Diagnose instructions
  - These are really programming APIs (semantically, procedure calls)
  - Large number of "entry points" provides many callable functions for guests

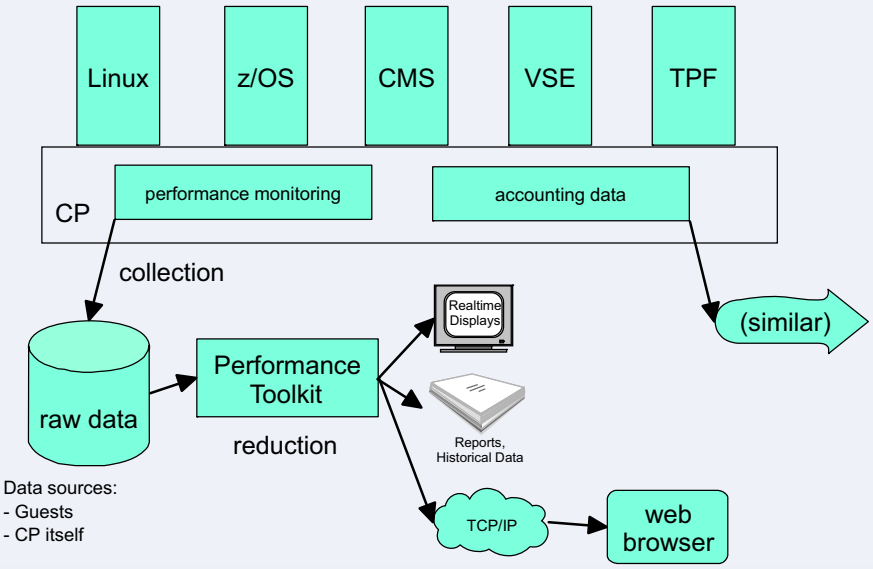
What: Debugging a Virtual Machine

- Tracing a virtual machine's execution
  - CP TRACE command: can trace
    - Instructions (specific opcodes, specific addresses)
    - Storage references or alterations
    - Register alterations
    - Use of various address spaces
    - ... on and on and on ...
  - Single-step through execution, or run and collect trace to a spool file
  - Trace points can trigger other commands
- Display or store into guest memory
  - Helpful especially in conjunction with tracing
  - Can select address space to display or alter
  - Options for translation: ASCII, EBCDIC, System z opcodes (disassemble)
  - Locate strings in memory

What: Programmable Operator



What: Performance and Accounting Data



References

- VM web site: www.vm.ibm.com
- Publications there:
  - www.vm.ibm.com/pubs/
  - Follow the links to the latest z/VM library
  - Of particular interest:
    - z/VM CP Command and Utility Reference
    - z/VM CP Planning and Administration
    - z/VM CP Programming Services
    - z/VM Performance
  - IBM Systems Journal, vol. 30, no. 1, 1991
    - Good article on SIE

End of Presentation

Question and Answer Time

IBM System z Expo
September 17-21, 2007
San Antonio, TX