Bay Bunch

Using z/VM in a SCSI Environment

zSeries & Linux Users' Group

09/24/2004

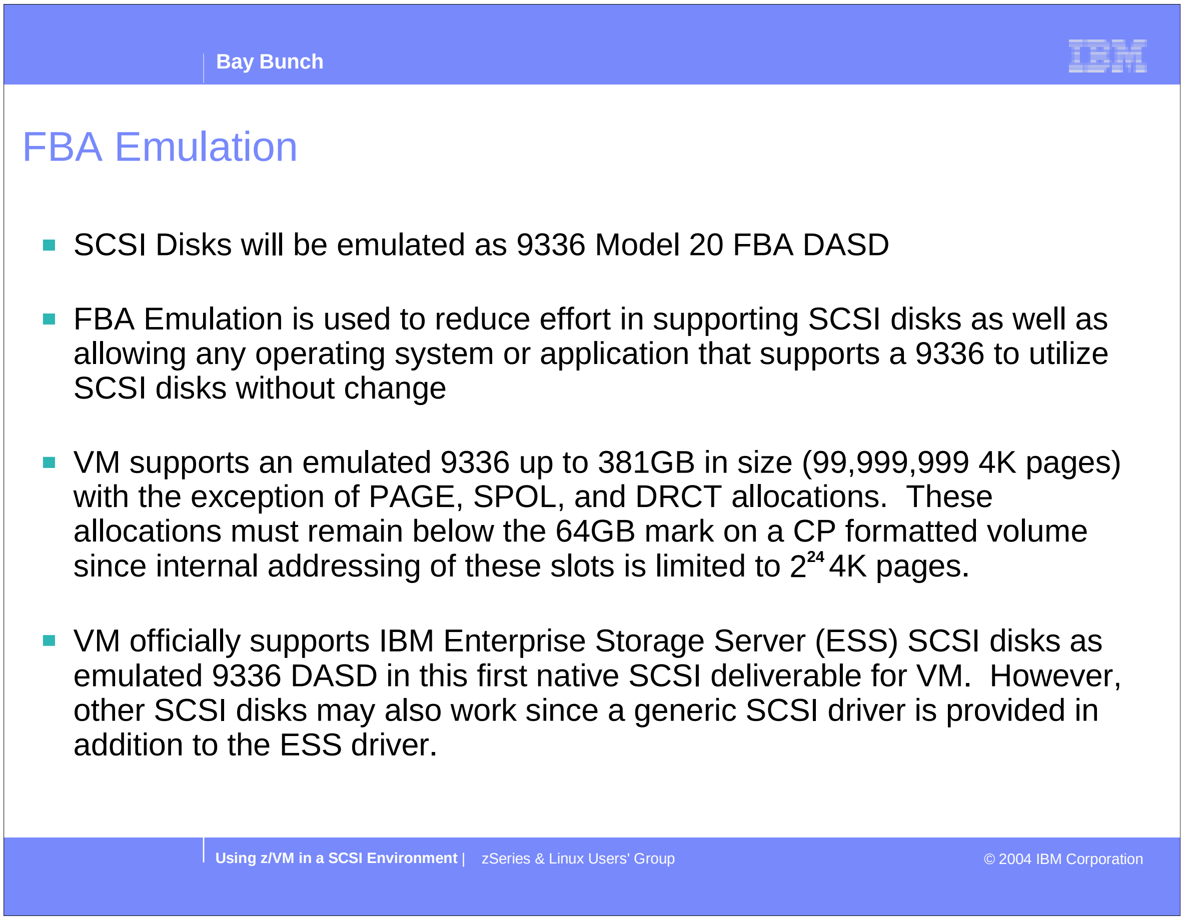
Steve Wilkins
wilkinss@us.ibm.com

Agenda

- Overview
- FBA Emulation
- C SCSI Driver Stack
- System Structure
- Configuration Elements
- Commands
- Configuration File Statement
- HCD
- Install
- Software/Firmware Requirements
- IPL
- Dump and Service
- Monitor Records
- Performance
- Trademarks

Overview

- VM will provide native support for SCSI disks for paging, spooling, and other system data
- Support will be provided by configuring SCSI disk LUNs to VM as 9336 FBA 512-byte block DASD
- VM guests (such as CMS, GCS, and VSE) may keep data on SCSI disk LUNs without requiring the guest software to have SCSI disk drivers
- IPL, Dump, and Service from/to SCSI disk LUNs will be provided to achieve a SCSI-only VM environment
- SCSI-only as well as mixed SCSI and ECKD environments are supported
- Available with z/VM 5.1.0 (9/2004)



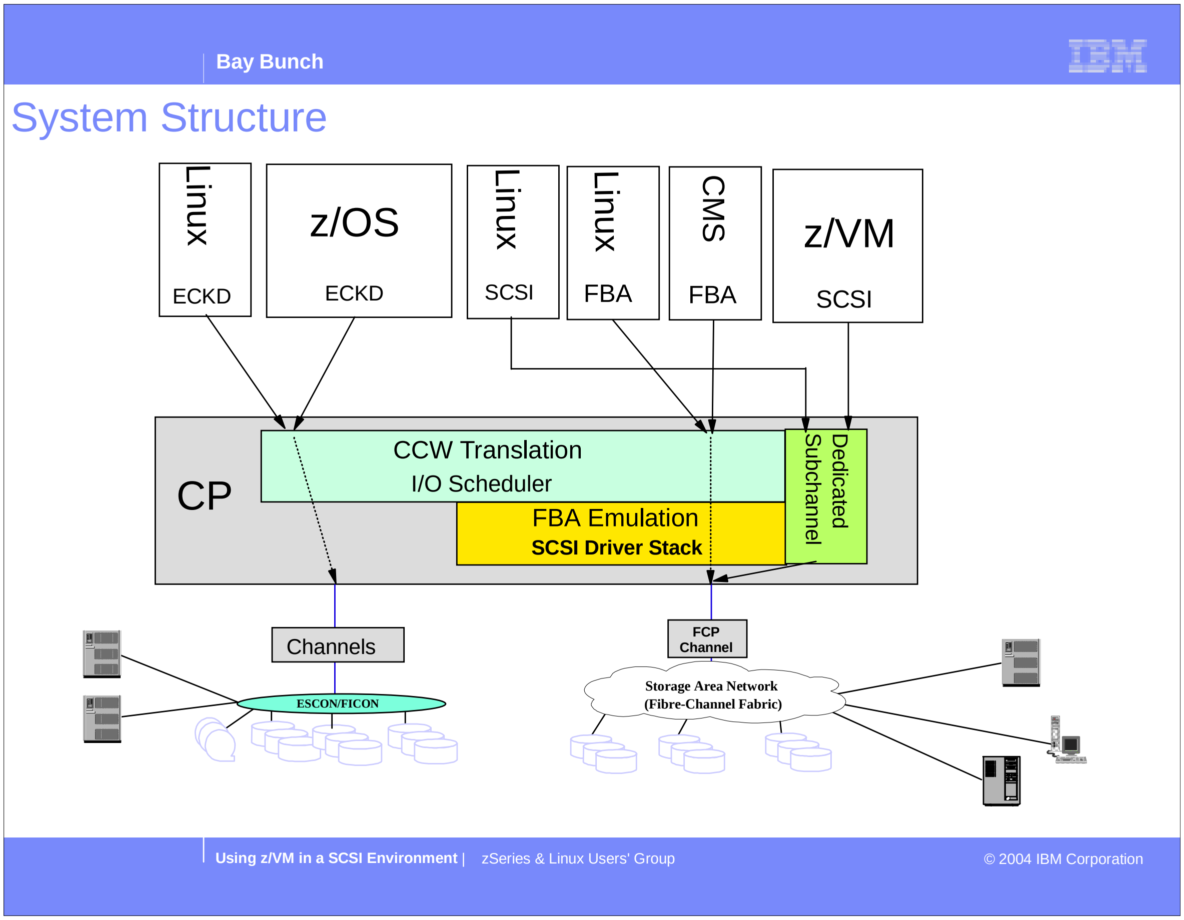
FBA Emulation

- SCSI Disks will be emulated as 9336 Model 20 FBA DASD
- FBA Emulation is used to reduce effort in supporting SCSI disks as well as allowing any operating system or application that supports a 9336 to utilize SCSI disks without change
- VM supports an emulated 9336 up to 381GB in size (99,999,999 4K pages) with the exception of PAGE, SPOL, and DRCT allocations. These allocations must remain below the 64GB mark on a CP formatted volume since internal addressing of these slots is limited to 2^{24} 4K pages.
- VM officially supports IBM Enterprise Storage Server (ESS) SCSI disks as emulated 9336 DASD in this first native SCSI deliverable for VM. However, other SCSI disks may also work since a generic SCSI driver is provided in addition to the ESS driver.

- 381GB is (99,999,999 x 4096) / (2**30)
- 64GB is (2**24 x 4096) / (2**30)
- Page 16,777,215 is the last allowed page for Page, Spool, and Directory allocations.
- Allocations PERM, TDSK, and PARM may be placed above page 16,777,215.
- The ESS 2105 driver is specifically tuned for an ESS disk.
- The generic SCSI driver may work with other disk hardware if the hardware doesn't require a device dependent driver.
- Only ESS is officially supported.

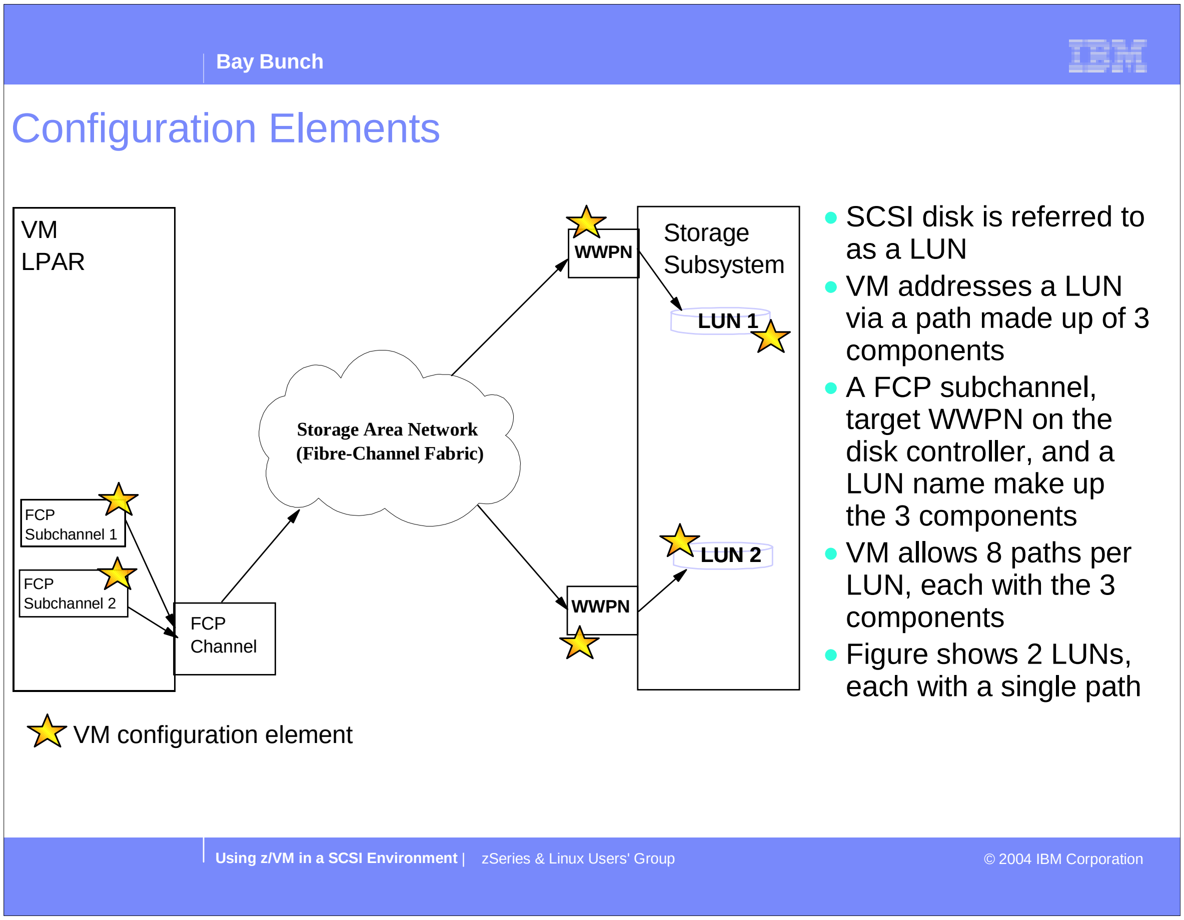
C SCSI Driver Stack

- A SCSI driver stack written in the C programming language has been coded to drive I/O to SCSI LUNs over FCP subchannels
- This is the first time for C in the CP Nucleus of VM. Therefore, C infrastructure (i.e., linkage and memory stack) has been added with this project. This will allow future CP projects to be written in C.
- This C SCSI driver stack exists in the real I/O layer of the CP component of VM (see next slide).
- An FBA channel program emulator intercepts real Start I/O (SSCH) requests from the real I/O dispatcher of VM and *transforms* these channel programs into API calls to the C SCSI driver stack
- Ending status from the C SCSI driver stack is in turn *transformed* into appropriate z/Architecture and FBA status to be returned to the originator of the FBA I/O request



System Structure

- Dedicated FCP subchannel support in z/VM 4.3.0 and up is used to run second-level guests that contain native support for SCSI disks
- This includes z/VM 5.1.0 and z/Linux



Configuration Elements

☆ VM configuration element

- SCSI disk is referred to as a LUN
- VM addresses a LUN via a path made up of 3 components
- A FCP subchannel, target WWPN on the disk controller, and a LUN name make up the 3 components
- VM allows 8 paths per LUN, each with the 3 components
- Figure shows 2 LUNs, each with a single path

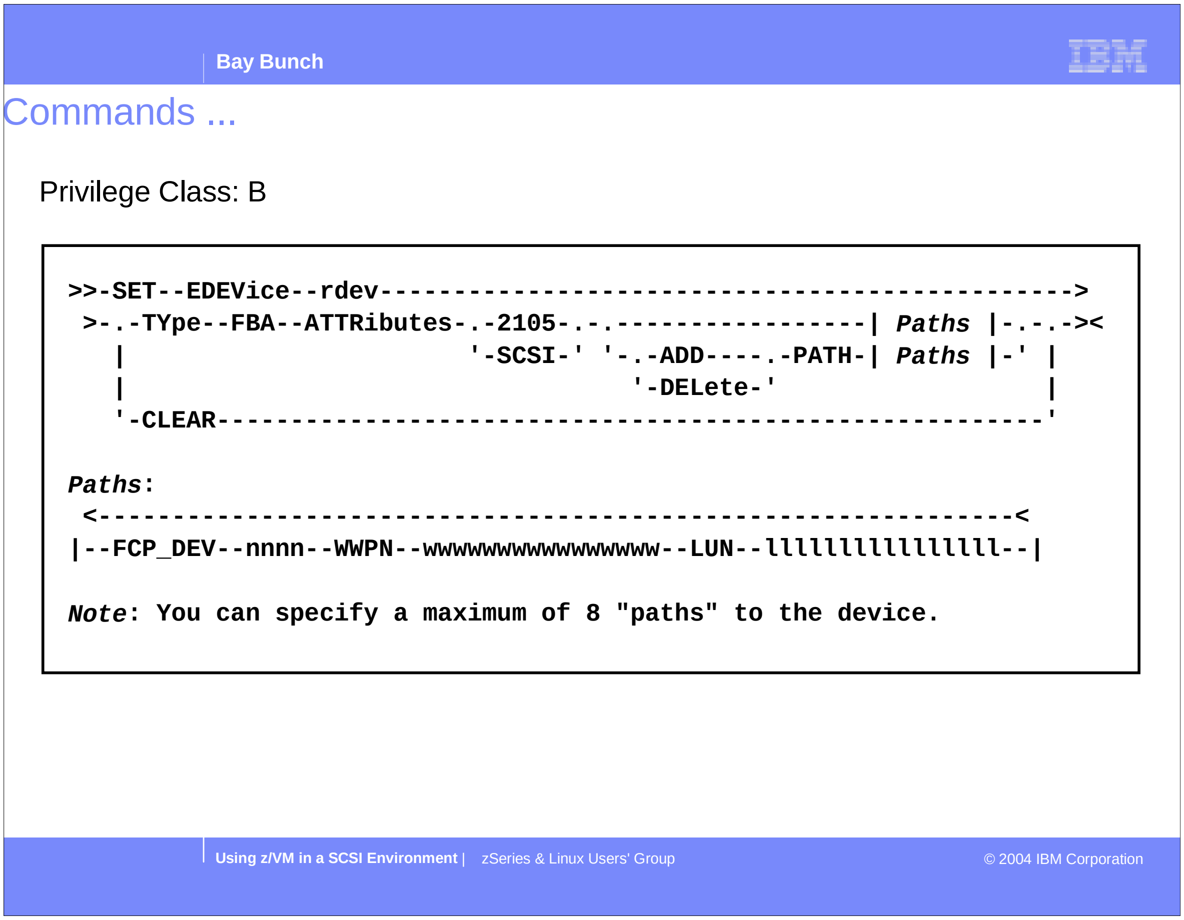
- WWPN is World Wide Port Name
- WWPN is worldwide unique

Commands

- New SET EDEVICE command provided to configure a SCSI disk LUN to the VM system as an emulated 9336 FBA DASD

- New QUERY EDEVICE command provided to obtain information related to SCSI disk LUNs defined to VM as emulated FBA DASD

- New DELETE EDEVICE command provided to remove a SCSI disk LUN from the VM system that has been previously defined as an emulated FBA DASD



Commands ...

Privilege Class: B

```
>>-SET--EDEVice--rdev------------------------------------------------>
 >-.-TYpe--FBA--ATTRibutes-.-2105-.-.------------------| Paths |-.-.-><
   |                       '-SCSI-' '-.-ADD-----.-PATH-| Paths |-' |
   |                                  '-DELete-'                   |
   '-CLEAR---------------------------------------------------------'

Paths:
 <---------------------------------------------------------------<
|--FCP_DEV--nnnn--WWPN--wwwwwwwwwwwwwwww--LUN--llllllllllllllll--|

Note: You can specify a maximum of 8 "paths" to the device.
```

- ATTR keyword tells VM which SCSI driver to use, either the ESS or Generic driver in the C SCSI driver stack
- FCP_DEV is the real device number associated with an FCP subchannel that has connectivity to the LUN
- All 16 digits of the WWPN and LUN must be specified. Otherwise, the VM command parser will add leading zeros causing the path to be invalid. For example, LUN 5150 must be specified as 5150000000000000.
- An EDEVICE must be varied off-line to ADD, DEL, or CLEAR path information.
- An EDEVICE takes up a slot in VM's real device space. That is, there can't exist another real device with the same real device number.
- Once an EDEVICE is defined, it is managed on VM like a real 9336 FBA. CP commands such as VARY, ATTACH, and QUERY execute as if the emulated disk were a real FBA. This applies also to user directory and system configuration file statements

Commands ...

Privilege Class: B

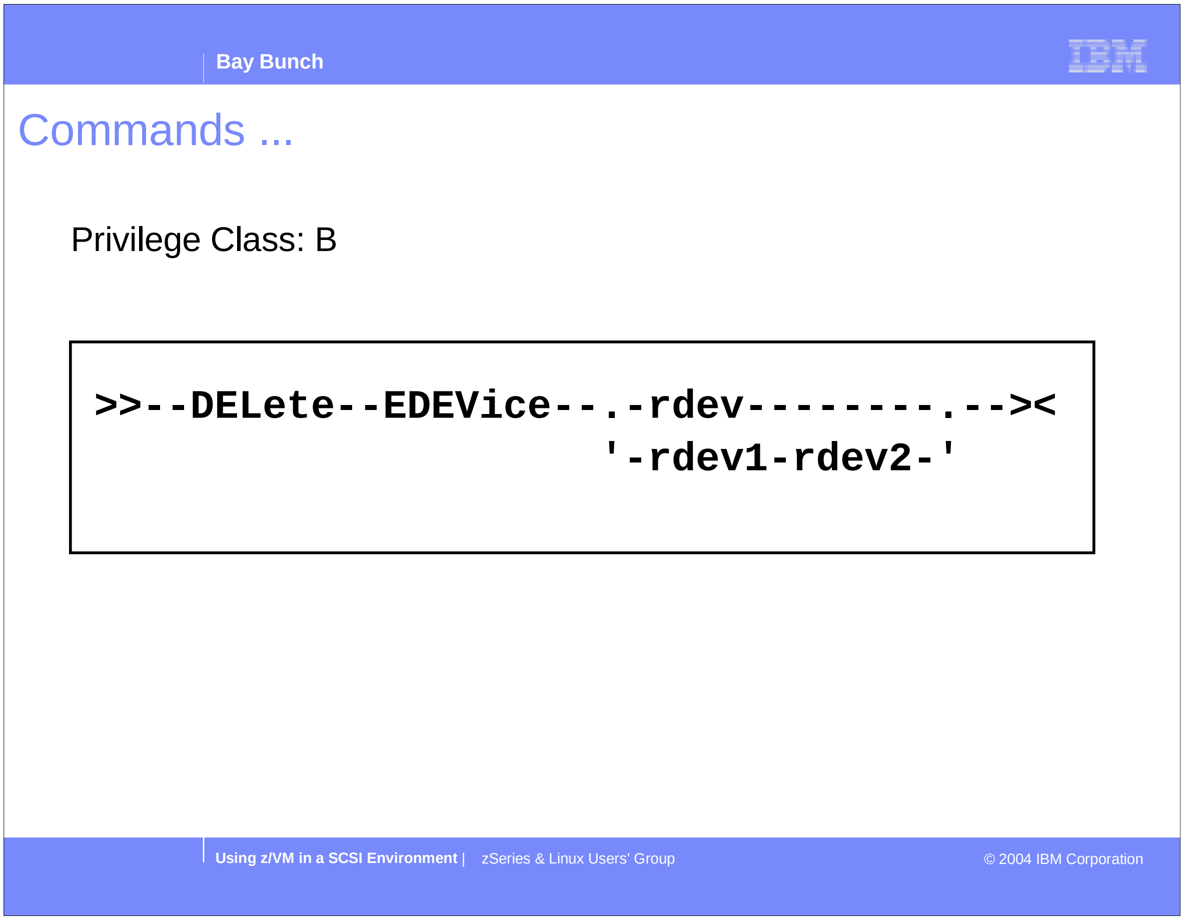
```
                     <---------<
>>-Query-EDEVice-.-.-rdev--------.-.---------.-.-><
                 | '-rdev1-rdev2-' '-DETAILS-' |
                 '-ALL-------------------------'
```

Commands ...

Query edevice responses

```
q edev 607
EDEV 0607 TYPE FBA ATTRIBUTES 2105
Ready;


q edev 607 details
EDEV 0607 TYPE FBA ATTRIBUTES 2105
   PATHS:
      FCP_DEV: 8100 WWPN: 5005076300C604DA LUN: 5137000000000000
      FCP_DEV: 8200 WWPN: 5005076300C604DA LUN: 5137000000000000
Ready;
```



Commands ...

Privilege Class: B

```
>>--DELete--EDEVice--.-rdev--------.--><
                     '-rdev1-rdev2-'
```

- An EDEVICE must be varied off-line to use the DELETE command.

Configuration File Statement

New SYSTEM CONFIG file statement performing same function as SET EDEVICE command

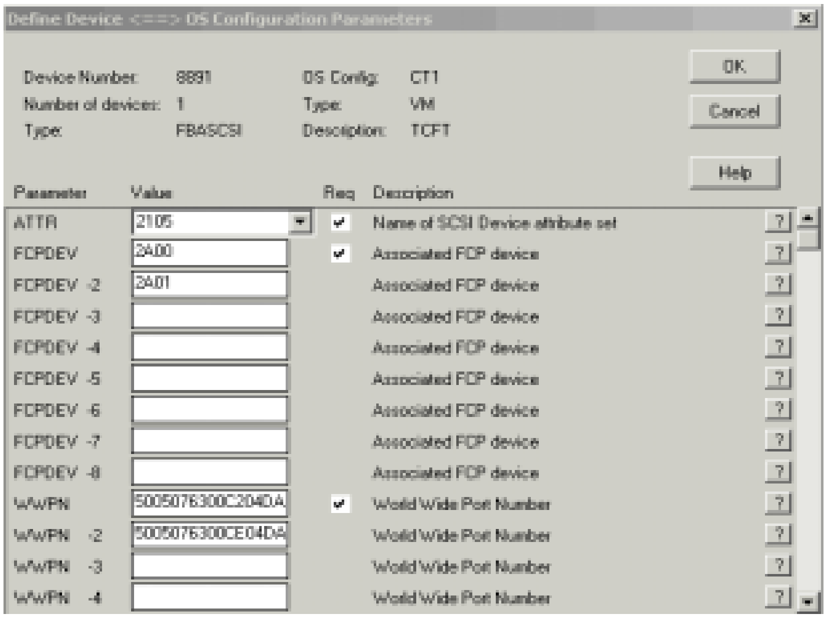
```
>>-EDEVice--rdev---TYpe--FBA--ATTRibutes--.-2105-.--| Paths |--><
                                          '-SCSI-'

Paths:
 <--------------------------------------------------------------<
|--FCP_DEV--nnnn--WWPN--wwwwwwwwwwwwwwww--LUN--llllllllllllllll--|

Note: You can specify a maximum of 8 "paths" to the device.
```

HCD

HCM GUI interface to VM's HCD support will allow an EDEV to be created with up to 8 paths.



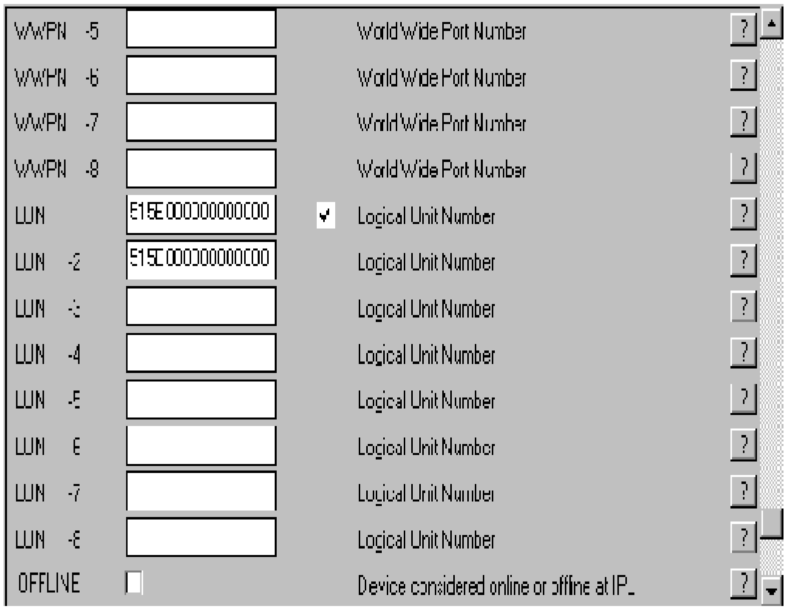
Device Number: 8891
Number of devices: 1
Type: FBASCSI
OS Config: CT1
Type: VM
Description: TCFT

Parameter	Value	Req	Description
ATTR	2105	☑	Name of SCSI Device attribute set
FCPDEV	2A00	☑	Associated FCP device
FCPDEV -2	2A01		Associated FCP device
FCPDEV -3			Associated FCP device
FCPDEV -4			Associated FCP device
FCPDEV -5			Associated FCP device
FCPDEV -6			Associated FCP device
FCPDEV -7			Associated FCP device
FCPDEV -8			Associated FCP device
WWPN	5005076300C204DA	☑	World Wide Port Number
WWPN -2	5005076300CE04DA		World Wide Port Number
WWPN -3			World Wide Port Number
WWPN -4			World Wide Port Number

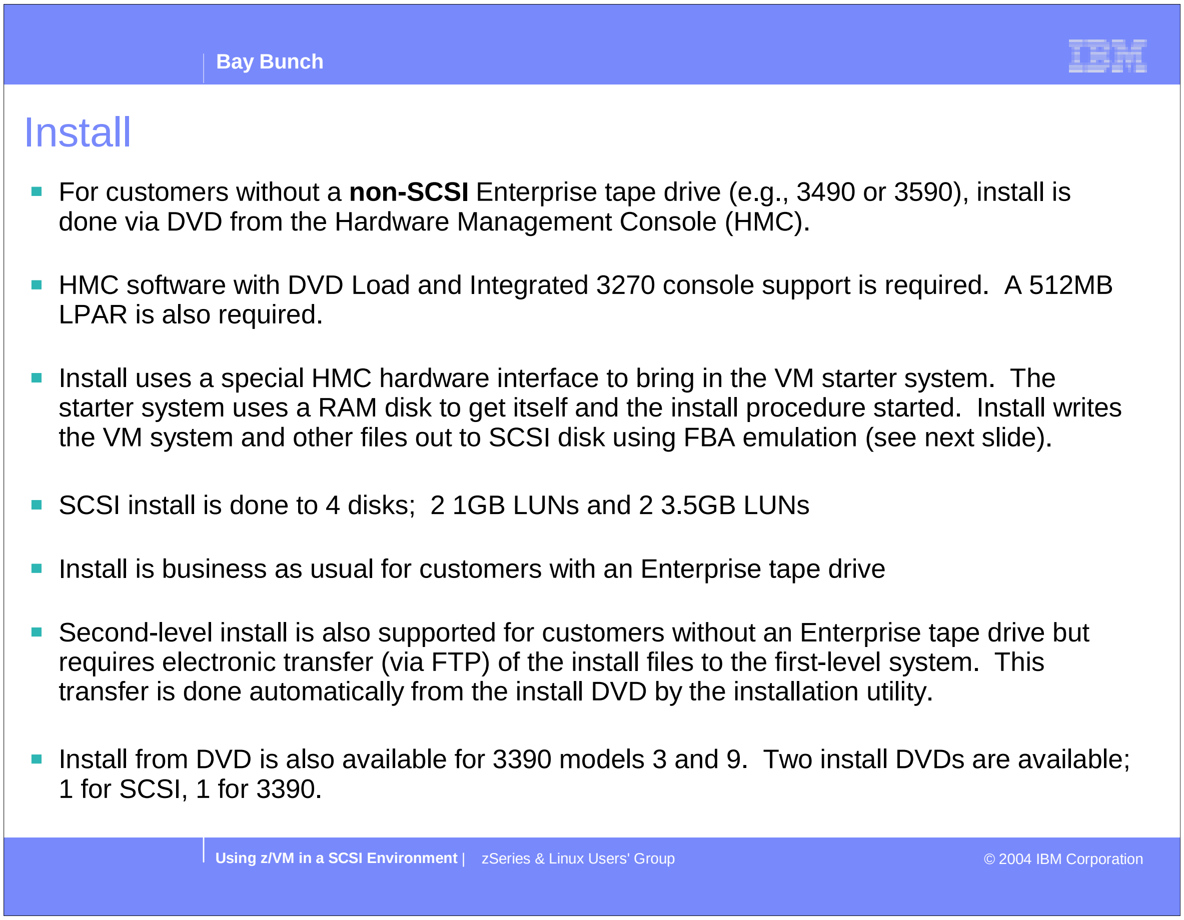
- VM HCD support can also be used to define emulated devices (EDEVs)
- HCM GUI takes input similar to SET EDEVICE command and EDEVICE configuration file statement

HCD

GUI continued



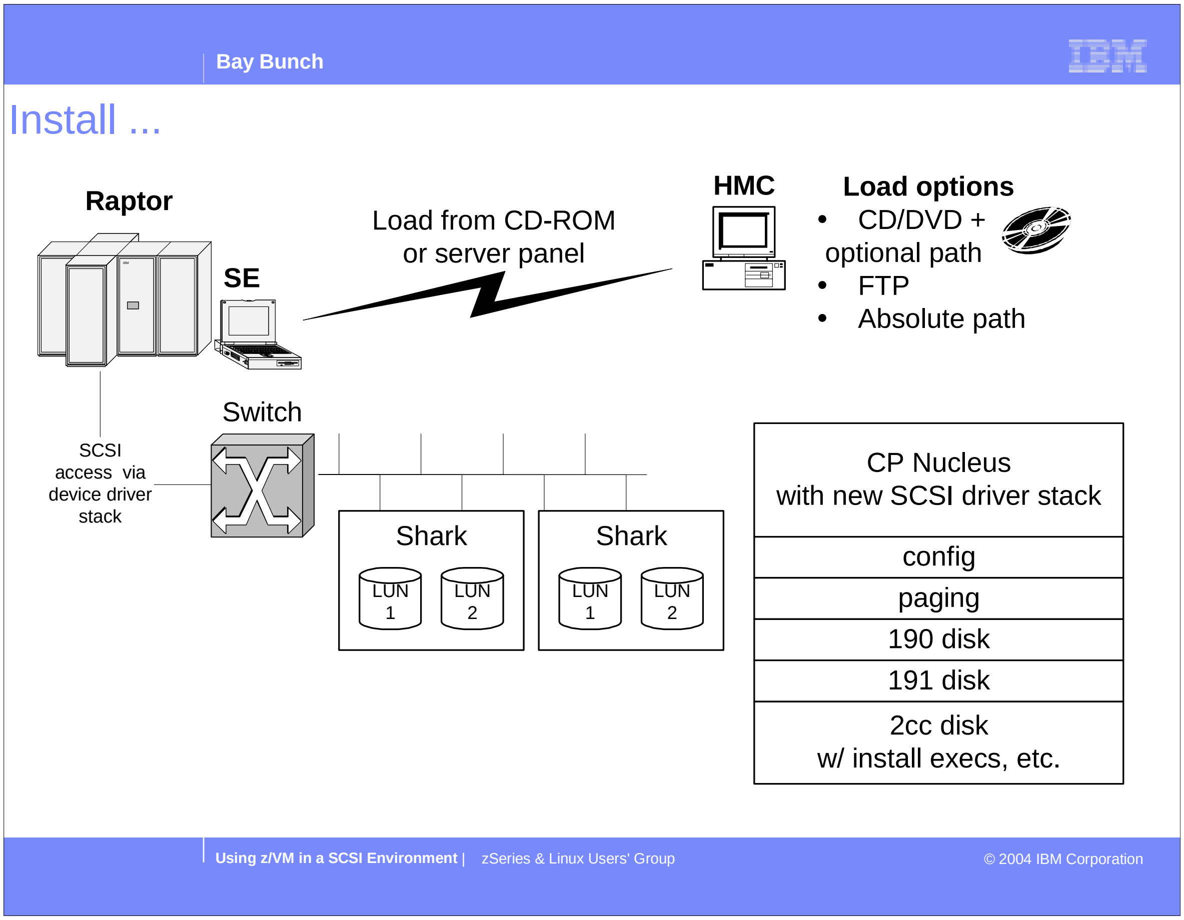
WWPN -5			World Wide Port Number
WWPN -6			World Wide Port Number
WWPN -7			World Wide Port Number
WWPN -8			World Wide Port Number
LUN	515E000000000000	☑	Logical Unit Number
LUN -2	515C000000000000		Logical Unit Number
LUN -3			Logical Unit Number
LUN -4			Logical Unit Number
LUN -5			Logical Unit Number
LUN -6			Logical Unit Number
LUN -7			Logical Unit Number
LUN -8			Logical Unit Number
OFFLINE	□		Device considered online or offline at IPL



Install

- For customers without a **non-SCSI** Enterprise tape drive (e.g., 3490 or 3590), install is done via DVD from the Hardware Management Console (HMC).
- HMC software with DVD Load and Integrated 3270 console support is required. A 512MB LPAR is also required.
- Install uses a special HMC hardware interface to bring in the VM starter system. The starter system uses a RAM disk to get itself and the install procedure started. Install writes the VM system and other files out to SCSI disk using FBA emulation (see next slide).
- SCSI install is done to 4 disks; 2 1GB LUNs and 2 3.5GB LUNs
- Install is business as usual for customers with an Enterprise tape drive
- Second-level install is also supported for customers without an Enterprise tape drive but requires electronic transfer (via FTP) of the install files to the first-level system. This transfer is done automatically from the install DVD by the installation utility.
- Install from DVD is also available for 3390 models 3 and 9. Two install DVDs are available; 1 for SCSI, 1 for 3390.

- The VM starter system on the DVD can also be used as an emergency tool utility
- z/VM Guide for Automated Installation and Service provides details on the installation procedure
- In addition to an FTP server, second-level installs also require a 64MB virtual machine with class B thru G privileges



Install ...

Raptor

SE

Load from CD-ROM
or server panel

HMC

Load options

- CD/DVD + optional path
- FTP
- Absolute path

Switch

SCSI
access via
device driver
stack

Shark

LUN 1

LUN 2

Shark

LUN 1

LUN 2

CP Nucleus
with new SCSI driver stack

config

paging

190 disk

191 disk

2cc disk
w/ install execs, etc.

- VM starter system is brought up via the 'Load from CD-ROM or Server' panel on the HMC. This panel is found by double-clicking the 'Single Object Operations' icon on the CPC RECOVERY window.
- VM starter system automatically comes up on userid MAINT. It is on MAINT that the install execs are run.

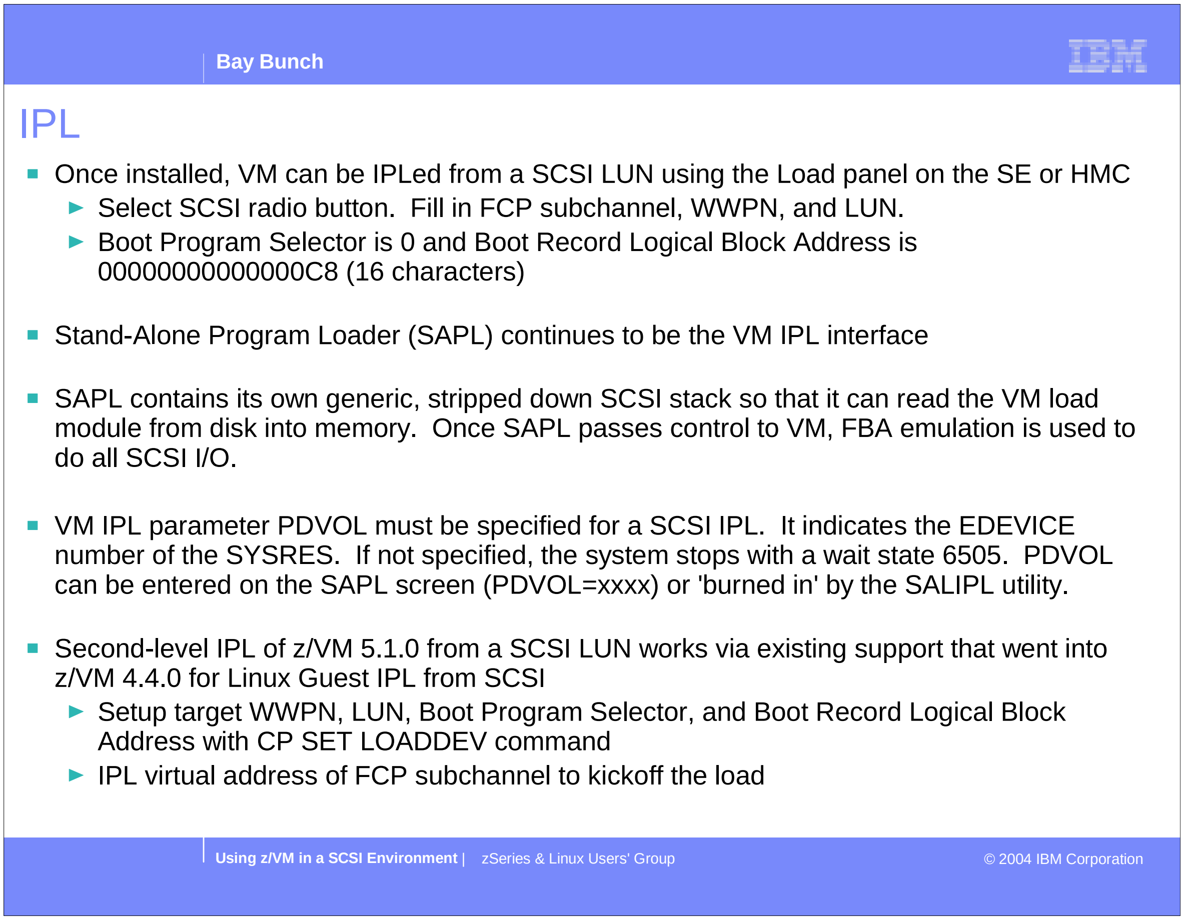
HMC SE Requirements

- The Hardware Management Console (HMC) must be communicating with the Support Element (SE). The HMC can only communicate with versions of the SE that are equal to or lower than the HMC. For example, HMC version 1.8.0 can communicate with a SE at version 1.7.3, or 1.8.0, but it can not communicate with a SE at version 1.8.2.

- The following minimum SE levels are required to install from DVD:
  - zSeries 800 - Support Element (SE) version 1.7.3, Engineering Change (EC) J11213, change level 146 or higher must be active.
  - zSeries 890 - Support Element (SE) version 1.8.2. No Licensed Internal Code changes are required.
  - zSeries 900 - Support Element (SE) version 1.7.3, Engineering Change (EC) J11213, change level 146 or higher must be active.
  - zSeries 990 - Support Element (SE) version 1.8.0, Engineering Change (EC) J12560, change level 054 or higher must be active. Or, Support Element (SE) version 1.8.2. No Licensed Internal Code changes are required.

FCP Firmware Requirements

- The following zSeries driver levels and FCP channel code levels are the minimum required to run z/VM 5.1.0 with SCSI disks:
  - zSeries 800 - Driver D3G with FCP Code 0.28, MCL J11233 #014
  - zSeries 890 - Driver D55 with FCP Code 3.04, MCL J13471 #003
  - zSeries 900 - Driver D3G with FCP Code 0.28, MCL J11233 #014
  - zSeries 990 - Driver D52 with FCP Code 2.07, MCL J12951 #004 for *GA2 level z990*
  - zSeries 990 - Driver D55 with FCP Code 3.04, MCL J13471 #003 for *GA3 level z990*



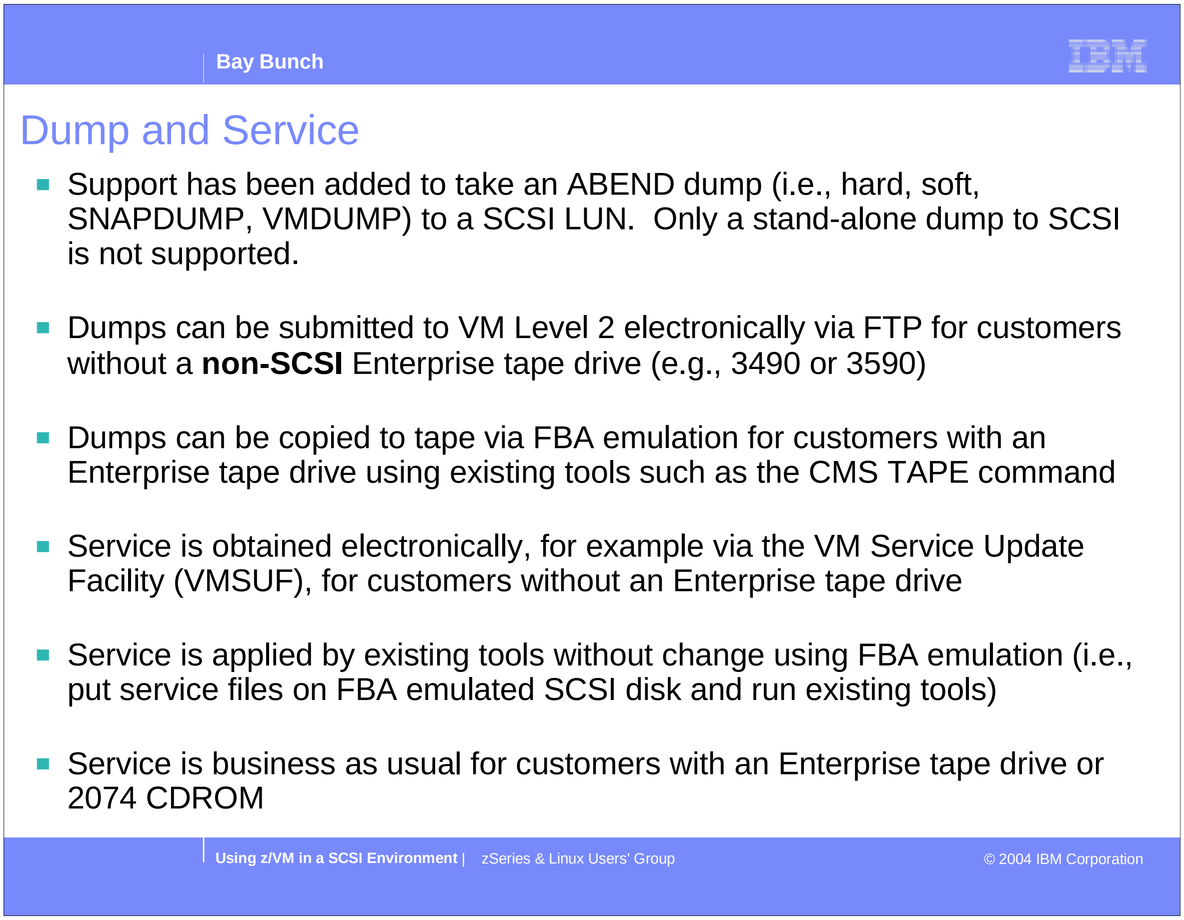
IPL

- Once installed, VM can be IPLed from a SCSI LUN using the Load panel on the SE or HMC
  - Select SCSI radio button. Fill in FCP subchannel, WWPN, and LUN.
  - Boot Program Selector is 0 and Boot Record Logical Block Address is 00000000000000C8 (16 characters)
- Stand-Alone Program Loader (SAPL) continues to be the VM IPL interface
- SAPL contains its own generic, stripped down SCSI stack so that it can read the VM load module from disk into memory. Once SAPL passes control to VM, FBA emulation is used to do all SCSI I/O.
- VM IPL parameter PDVOL must be specified for a SCSI IPL. It indicates the EDEVICE number of the SYSRES. If not specified, the system stops with a wait state 6505. PDVOL can be entered on the SAPL screen (PDVOL=xxxx) or 'burned in' by the SALIPL utility.
- Second-level IPL of z/VM 5.1.0 from a SCSI LUN works via existing support that went into z/VM 4.4.0 for Linux Guest IPL from SCSI
  - Setup target WWPN, LUN, Boot Program Selector, and Boot Record Logical Block Address with CP SET LOADDEV command
  - IPL virtual address of FCP subchannel to kickoff the load

- HMC/SE Load Address field for first-level IPL is the FCP subchannel providing access to the SCSI LUN
- HMC/SE Load Parameter for first-level or second-level IPL is the console address for SAPL to use. Use 'SYSG' on a first-level IPL to designate the integrated 3270 console.
- Second-level IPL uses same Boot Program Selector and Boot Record Logical Block Address as a first-level IPL

SALIPL

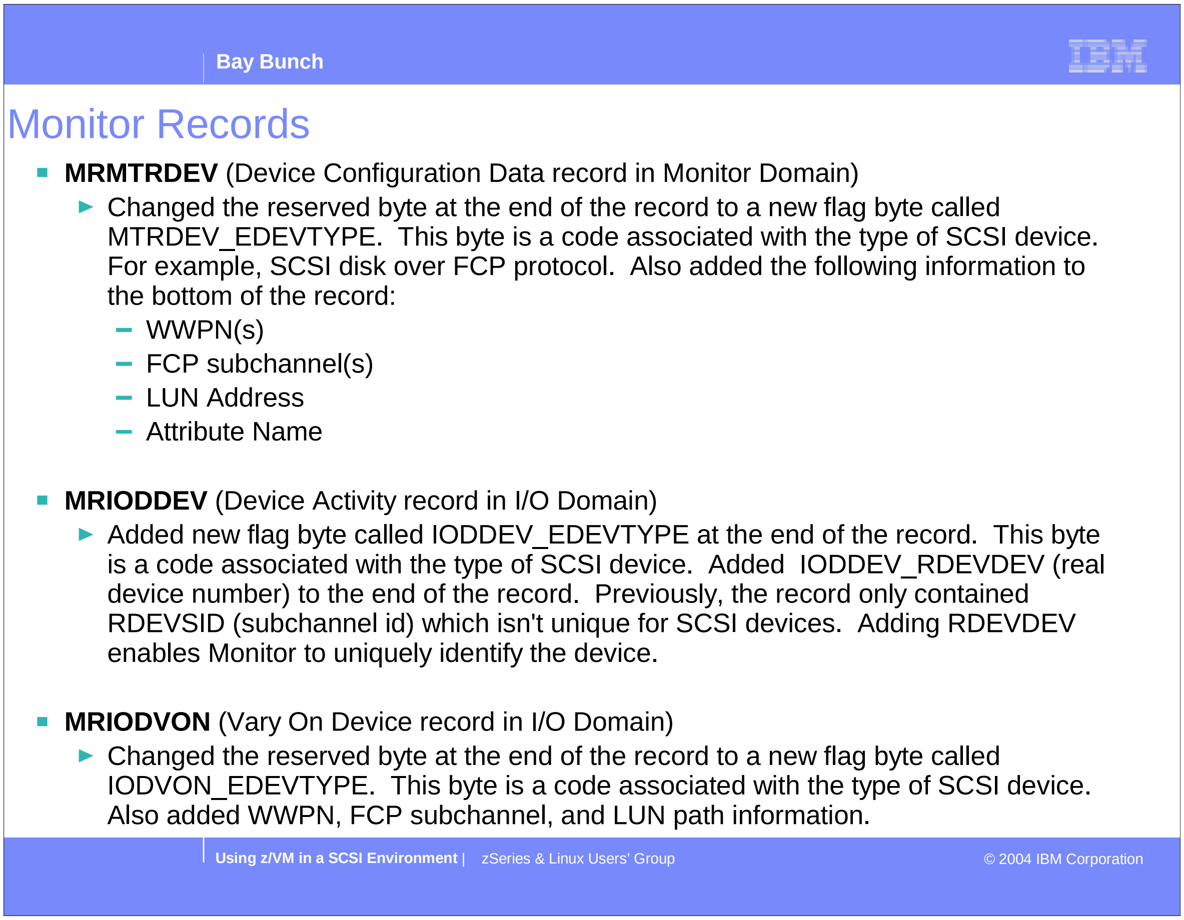
- The SALIPL utility continues to be used to setup SAPL for IPL.
- SALIPL now writes to blocks 5-207 if the device is a FBA (SCSI or not). SALIPL used to only write to blocks 5-31. This affects the size of the RECOMP area for a CMS minidisk containing SAPL. It also affects the placement of allocations such as PAGE and SPOL when SAPL is put on a CP formatted volume.
- SALIPL must run in a virtual machine to setup a SCSI disk for IPL. SALIPL can be run against either:
  - A virtual device, such as a fullpack minidisk, on an emulated FBA DASD
  - A virtual FCP subchannel where new SALIPL parameters WWPN and LUN designate the target SCSI disk
- The SYSRES device and device containing the PARM area must be the same for a SCSI IPL. This isn't the case for an ECKD IPL.
- A second-level SCSI IPL can either be done by issuing the IPL command against a virtual FCP subchannel with access to the LUN or against a virtual device on an emulated FBA DASD. First-level IPLs are SCSI-only since FBA emulation doesn't exist on the hardware.



Dump and Service

- Support has been added to take an ABEND dump (i.e., hard, soft, SNAPDUMP, VMDUMP) to a SCSI LUN. Only a stand-alone dump to SCSI is not supported.
- Dumps can be submitted to VM Level 2 electronically via FTP for customers without a **non-SCSI** Enterprise tape drive (e.g., 3490 or 3590)
- Dumps can be copied to tape via FBA emulation for customers with an Enterprise tape drive using existing tools such as the CMS TAPE command
- Service is obtained electronically, for example via the VM Service Update Facility (VMSUF), for customers without an Enterprise tape drive
- Service is applied by existing tools without change using FBA emulation (i.e., put service files on FBA emulated SCSI disk and run existing tools)
- Service is business as usual for customers with an Enterprise tape drive or 2074 CDROM

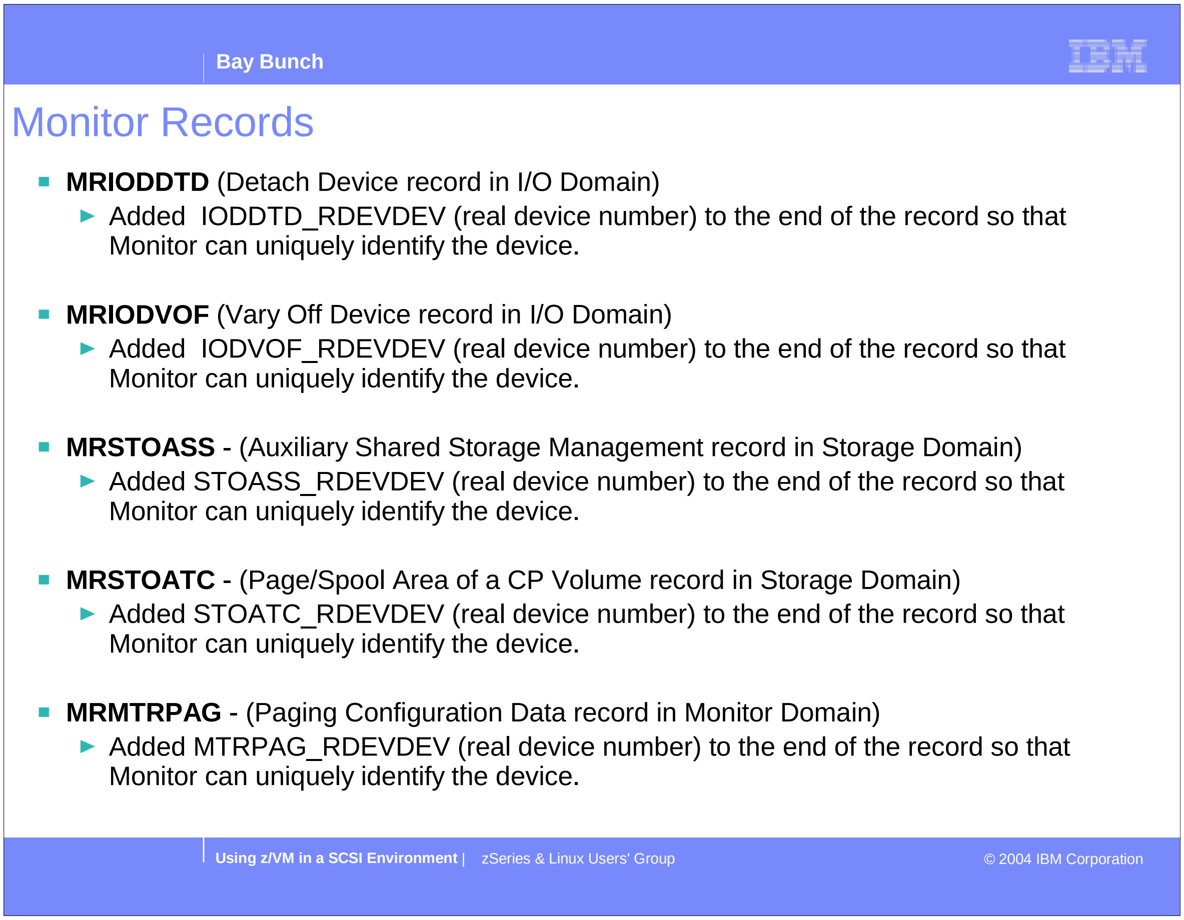
- z/VM Service Guide provides details on applying service



Monitor Records

- **MRMTRDEV** (Device Configuration Data record in Monitor Domain)
  - Changed the reserved byte at the end of the record to a new flag byte called MTRDEV_EDEVTYPE. This byte is a code associated with the type of SCSI device. For example, SCSI disk over FCP protocol. Also added the following information to the bottom of the record:
    - WWPN(s)
    - FCP subchannel(s)
    - LUN Address
    - Attribute Name
- **MRIODDEV** (Device Activity record in I/O Domain)
  - Added new flag byte called IODDEV_EDEVTYPE at the end of the record. This byte is a code associated with the type of SCSI device. Added IODDEV_RDEVDEV (real device number) to the end of the record. Previously, the record only contained RDEVSID (subchannel id) which isn't unique for SCSI devices. Adding RDEVDEV enables Monitor to uniquely identify the device.
- **MRIODVON** (Vary On Device record in I/O Domain)
  - Changed the reserved byte at the end of the record to a new flag byte called IODVON_EDEVTYPE. This byte is a code associated with the type of SCSI device. Also added WWPN, FCP subchannel, and LUN path information.

- _EDEVTYPE in all records has the same byte code definition.
- RDEVSID for an emulated device is inherited from one of the FCP subchannels defined on the SET EDEVICE command or EDEVICE configuration file statement.
- WWPN, FCP subchannel, and LUN path information in record MRIODVON is similar to the information in MRMTRDEV.



Monitor Records

- **MRIODDTD** (Detach Device record in I/O Domain)
  - Added IODDTD_RDEVDEV (real device number) to the end of the record so that Monitor can uniquely identify the device.
- **MRIODVOF** (Vary Off Device record in I/O Domain)
  - Added IODVOF_RDEVDEV (real device number) to the end of the record so that Monitor can uniquely identify the device.
- **MRSTOASS** - (Auxiliary Shared Storage Management record in Storage Domain)
  - Added STOASS_RDEVDEV (real device number) to the end of the record so that Monitor can uniquely identify the device.
- **MRSTOATC** - (Page/Spool Area of a CP Volume record in Storage Domain)
  - Added STOATC_RDEVDEV (real device number) to the end of the record so that Monitor can uniquely identify the device.
- **MRMTRPAG** - (Paging Configuration Data record in Monitor Domain)
  - Added MTRPAG_RDEVDEV (real device number) to the end of the record so that Monitor can uniquely identify the device.

- _RDEVDEV is the 'real' device number defined on the SET EDEVICE command or EDEVICE configuration file statement.
- It was only added to monitor records that did not already contain _RDEVDEV.

Monitor Records

- **MRSEKSEK** - (Seek Data record in Seek Domain)
  - Added SEKSEK_RDEVDEV (real device number) to the end of the record so that Monitor can uniquely identify the device.

- **MRIODSZI** (New I/O Domain Monitor Record 24)
  - This is the SCSI Device Activity Record and will be cut in addition to the existing MRIODDEV record. The first 20 bytes of the record will contain the standard header for monitor records. The rest of the record will contain the following:
    - RDEVDEV
    - Bytes in a block (512)
    - -> amount of time device is active
    - -> kbytes/sec transfer rate
    - -> #transfers to/from disk
    - -> #blocks read from disk
    - -> #blocks written to disk
    - -> #seek operations
    - -> I/O queue depth
    - Statistics on per path basis (same as -> arrows)

Performance

Negatives

- Significant path length increase, so one needs to insure proper planning for enough processor cycles to handle the I/O rate
- Monitoring capabilities are not as strong as traditional ECKD in isolating problems at the channel, control unit, and device levels.
- FCP paths are not expected to be significantly faster than FICON (from a hardware perspective)
- Further increases the VM "overhead" numbers for Linux guests

Positives

- Allows exploitation of minidisk cache
- Allows minidisks on SCSI to be shared
- Improves performance measurement capabilities compared to FCP subchannels dedicated to a guest
- FCP paths faster than ESCON
- Greater flexibility in size of minidisks based on SCSI devices tending to be larger in general
- Allows potential use of other tuning options not available for dedicated FCP subchannels:
  - SET THROTTLE
  - SET IOPRIORITY

Performance Recommendations

- Continue to use dedicated FCP subchannels for Linux guests except when:
  - sharing minidisks can add disk and administrative savings
  - significant common data exists in read-mostly usage and minidisk cache can benefit
  - other VM minidisk management capabilities add significant value

- Continue to use traditional ECKD DASD for paging and spooling except when:
  - no ECKD DASD is available
  - little paging and spooling activity exists
  - sufficient processor resources are available to handle increased path length

- Continue to use traditional ECKD DASD for CMS minidisks except when:
  - no ECKD DASD is available
  - I/O to CPU ratios are low (minidisk cache helps lower this ratio)
  - sufficient processor resources are available to handle increased path length

- Consider the new z/VM native SCSI support when:
  - it would facilitate moving off of ESCON channels to FCP
  - large minidisks are desired

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
DFSMS/MVS*
DFSMS/VM*
e business(logo)*
Enterprise Storage Server*
ESCON*
FICON
GDDM*
HiperSockets
IBM*
IBM(logo)*
Language Environment*
MQSeries*
Multiprise*
MVS
NetRexx
OpenEdition*
OpenExtensions
OS/390*
Parallel Sysplex*
PR/SM
QMF
RACF*
RAMAC*
RISC
S/370
S/390*
S/390 Parallel Enterprise Server
VisualAge*
VisualGen*
VM/ESA*
VTAM*
VSE/ESA
WebSphere*
z/Architecture
z/OS*
zSeries*
z/VM*

* 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.
Penguin (Tux) compliments of Larry Ewing

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.