Managing your real hardware: Installation, Boot, Hardware changes

Olivier Crémel

Staff Engineer

VMWORLD 2006

Agenda

- Hardware choices
  - Real hardware vs. virtual hardware
  - Driven by Service Console or VMkernel
  - Hardware compatibility list
- Installation
  - BIOS settings
  - Device naming
- Running
  - Failure to boot
  - Alerts and warnings
  - Interrupt sharing

Real hardware vs. virtual hardware

- Virtual hardware is a convenient abstraction
  - Hides real hardware variation (NIC speed, chipset memory support)
  - Allows VMotion (outlives real hardware)
  - Simplifies software delivery (virtual appliances freed of hardware complexity)
- Real hardware has (almost) no impact on virtual hardware
  - Virtual storage is SCSI (don't go looking for your FC adapter in a virtual machine)
  - Virtual networking advertises Gb (whether underlying is 100Mb or 10Gb or even software)
  - Virtual graphics will not gain anything from an advanced graphics card
  - But processors will shine through

Driven by Service Console or VMkernel

- All supported networking and storage devices are driven by VMkernel
  - Even Service Console networking and storage
  - (With ESX Server 2.x, a NIC or storage controller could be dedicated to the Service Console or shared with it)
- All other devices are driven by the Service Console
  - Keyboard, mouse, video, sound
  - IDE CD-ROM drive, floppy controller
  - Serial and parallel interface
  - USB controllers and devices
  - IPMI
  - Any add-on devices (UPS, …)

Hardware Compatibility List

- Why do we have one ?
  - Testing, validation
  - Driver availability for VMkernel
  - Would ESX Server run on non-supported platforms ? Most likely.

- Restrictive checks
  - Prevents VMkernel from loading on non-supported platforms:
    - E.g. explicit check for AMD or Intel processors
  - May result in spurious warnings or alerts

Hardware choices

- Processors
  - Better not to mix and match (at least for now)
  - Mixed stepping is supported
  - BSP leads (APs are assumed to be at least as capable as BSP)
- Memory
  - Memory must be balanced across NUMA nodes
  - AMD Opteron-based systems are *de facto* NUMA
- Devices
  - Need a VMkernel driver
  - USB over IP (to not impair VMotion e.g. when USB dongles are used, otherwise a virtual machine becomes tied to real hardware)

Agenda

- Hardware choices
  - Real hardware vs. virtual hardware
  - Driven by Service Console or VMkernel
  - Hardware compatibility list
- Installation
  - BIOS settings
  - Device naming
- Running
  - Failure to boot
  - Alerts and warnings
  - Interrupt sharing

BIOS settings

- BIOS is important
  - Provides processor list, memory configuration, interrupt routing
  - MPS vs. ACPI, MPS is used by ESX Server but is no longer well-supported by OEMs
  - Make sure the latest BIOS is installed
- BIOS settings are important
  - Often used by OEMs to support older versions of system software :
    - Disabling NUMA on Opteron-based systems
    - Limiting CPUID leaves
    - Reordering BSP
    - MPS limitation (for single IOAPIC mode of operation)
  - Make sure the machine is not unwittingly shackled

Device naming

- With ESX Server 2.x
  - Created by VMkernel on the fly as drivers are loaded
  - Non-persistent
  - May change even without hardware changes
- With ESX Server 3.0
  - Created during installation
  - Mapping between name and PCI address
  - Can only change with hardware changes
    - Adding/removing a PCI card may cause bus renumbering

Device naming (cont.)

- Buses may be renumbered when buses are added/removed
  - Depends on how BIOS enumerates buses (holes)
  - Bridged cards contain buses (some quad-ported NIC)
- Names will be assigned based on old enumeration
  - Existing names may point to different PCI devices
  - Existing devices may become nameless and thus unusable
- Configuration needs to be regenerated
  - Be careful to follow documented procedure
  - Simply shutting down, adding/removing device, starting up may render the machine unusable

Agenda

- Hardware choices
  - Real hardware vs. virtual hardware
  - Driven by Service Console or VMkernel
  - Hardware compatibility list
- Installation
  - BIOS settings
  - Device naming
- Running
  - Failure to boot
  - Alerts and warnings
  - Interrupt sharing

Failure to boot

- Checks are being performed and may prevent VMkernel from loading
  - Missing interrupt information
  - Unbalanced NUMA memory
  - BIOS options to restrict visibility (e.g. CPUID limit)
- Device renaming may prevent boot disk from being found
- Fallback
  - Simple shell
  - dmesg will provide information

Failure to boot (examples)

- No pages allocated to Node 2 -- big mismatch between BIOS and SRAT memory maps, or MTRR error, or user removed all memory from a Node. Try checking memory or upgrading BIOS
- Unsupported BIOS setting, CPUID is limited
- Unsupported CPU, id0.name is …

Alerts and warnings

- While booting
  - Assumption about supported platform has been violated but not deemed serious enough to prevent VMkernel from loading:
    - mismatched processors (speed, number of cores, …)
    - incomplete BIOS information
  - May be spurious on a given configuration (PCI setup failure)
- While running
  - Machine Check Exceptions (memory errors)
  - Heartbeat loss (processor lock-up)

Interrupt sharing

- VMkernel gets all the interrupts
  - Interrupts are always targeted to a specific processor
    - Target processor can change (re-balancing)
    - Handled by resource scheduler (hardware balancing mechanism is not used)
    - Policy can be changed
  - Interrupts are dispatched to VMkernel device drivers
  - Interrupts corresponding to Service Console devices are forwarded to Service Console kernel which handles dispatching to Service Console device drivers
- Sharing can happen
  - Between VMkernel devices
  - Between VMkernel and Service Console

Interrupt sharing (cont.)

- Overview for VMkernel and Service Console

```
esx233 root # cat /proc/vmware/interrupts | more
Vector    PCPU  0    PCPU  1
0x21:          2          0     COS irq 1 (ISA edge), <VMK device>
0x29:          0          0     <COS irq 3 (ISA edge)>
0x31:          1          0     COS irq 4 (ISA edge), VMK serial
0x39:          0          0     <COS irq 6 (ISA edge)>
0x41:          0          0     <COS irq 8 (ISA edge)>
0x49:          1          0     COS irq 12 (ISA edge)
0x51:          0          0     <COS irq 13 (ISA edge)>
0x59:          0          0     <COS irq 14 (ISA edge)>
0x61:      11622      26938     <COS irq 5 (PCI level)>, VMK vmnic2
0x69:          0          0     COS irq 11 (PCI level)
0x71:      61517     132696     <COS irq 10 (PCI level)>, VMK vmnic0, VMK qla2300
0x79:     202308     299186     <COS irq 7 (PCI level)>, VMK vmnic1, VMK qla2300
0x81:       9053      17285     <COS irq 15 (PCI level)>, VMK cciss0

<> Means set up but not used currently
```

Interrupt sharing (cont.)

- Overview inside Service Console

```
esx233 root # cat /proc/interrupts | more
CPU0
 0:  111415681      vmnix-edge  timer
 1:          4      vmnix-edge  keyboard
 2:   14031287      vmnix-edge  VMnix interrupt
 4:          0      vmnix-edge  VMnix serial
11:          0     vmnix-level  usb-ohci
12:          1      vmnix-edge  PS/2 Mouse
```

Interrupt sharing (cont.)

- Sharing disrupts and lessens usefulness of re-balancing
  - Two VMkernel devices sharing an interrupt may have conflicting needs
  - Any interrupt used by the Service Console cannot be rebalanced and has to be tied to the BSP
- Sharing lowers performance
  - On any interrupt, all the VMkernel device drivers using that interrupt are called even though only one is likely to have work to do
  - On any interrupt used by the Service Console, a context switch to the Service Console will take place (not good if that interrupt is shared with a VMkernel NIC interrupting very frequently)

Preventing interrupt sharing

- Shuffling cards around
  - Interrupts are tied to physical slots
  - May not be effective if interrupts are constrained by the motherboard/chipset
- Disable unused onboard devices
  - USB interrupts are often shared but USB may not be used

Presentation Download

Please remember to complete your
session evaluation form
and return it to the room monitors
as you exit the session

The presentation for this session can be downloaded at
http://www.vmware.com/vmtn/vmworld/sessions/

Enter the following to download (case-sensitive):

Username: cbv_rep
Password: cbvfor9v9r

Some or all of the features in this document may be representative of feature areas under development. Feature commitments must not be included in contracts, purchase orders, or sales agreements of any kind. Technical feasibility and market demand will affect final delivery.

VMWORLD 2006