

Virtually Impossible

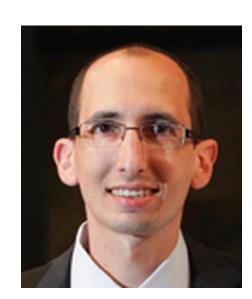
The Reality of Virtualization Security

Gal Diskin / Chief Research Officer / Cyvera LTD.

/WhoAml?

- Chief Research Officer @ Cvyera LTD
- Formerly Security Evaluation Architect of the Software & Services Group @ Intel®
- Before that Entrepreneur, Consultant, IDF
- Always a security "enthusiast" ©
 - Personal focus areas:
 - DBI, Fuzzing & Automated exploitation
 - Exploitation techniques & Mitigations
 - Vehicles & Traffic systems
 - Embedded systems





ThankZ & GreetZ

- My wife
 - For tolerating me doing security research
- Everyone at Cyvera, special thanks to:
 - Harel Baris for help with the presentation design
 - Gal Badishi and Ariel Cohen for reviewing
- All Intel security people
 - Especially my old team



What I will talk about today

Beyond why virtualization is virtually impossible to secure...

- Hardware assisted virtualization
- SW stacks and different virtualization approaches and related weaknesses
- The complexity in memory management and related weaknesses
- Computer platforms internals and related weaknesses
- Finally, I will present a small taxonomy of attacks against virtualization
- Special bonus potential VM escape ;-)

What is Virtualization?

- In the context of this talk replacing the CPU and computer platform with a virtual environment
- A bit of history:
 - Turing's universal computing machine
 - Popek and Goldberg virtualization requirements





Terminology

- A Virtual Machine Manager (VMM) is the software virtualizing privileged instructions and hardware
- A Virtual Machine (VM) is a software stack running under a VMM
- A Guest OS is the operating system of a VM
- A Host OS is the operating system controlling the VMM
- Root operation is when you execute inside a VMM



What is "secure" virtualization?

Security Goals:

Prevent modification of VMM and host OS by guests

Prevent guest OS from modifying another guest

Prevent guest from subverting hardware or firmware*

 Prevent guest from stealing data from other guest OS / host OS / VMM*

 Prevent DOS by guest OS* or getting unfair share of resources relative to other guests*

 Keep guest OS secure – don't harm normal OS defenses*

* Depending on the hypervisor design, might be a non-goal

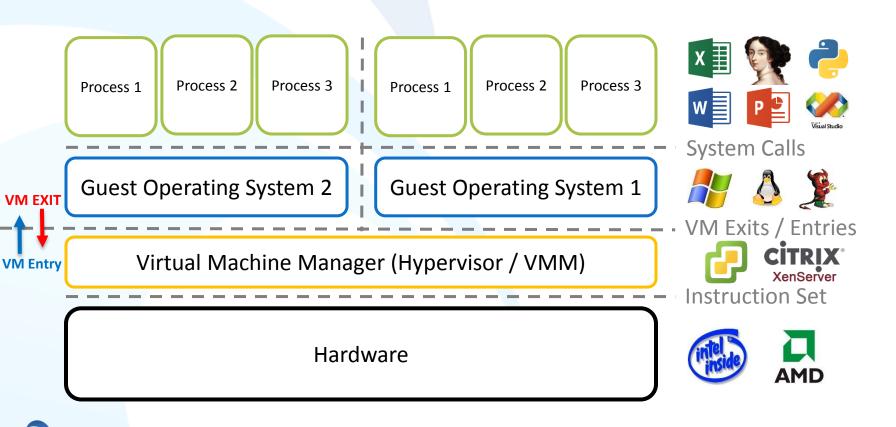


- CONFIDENTIAL -

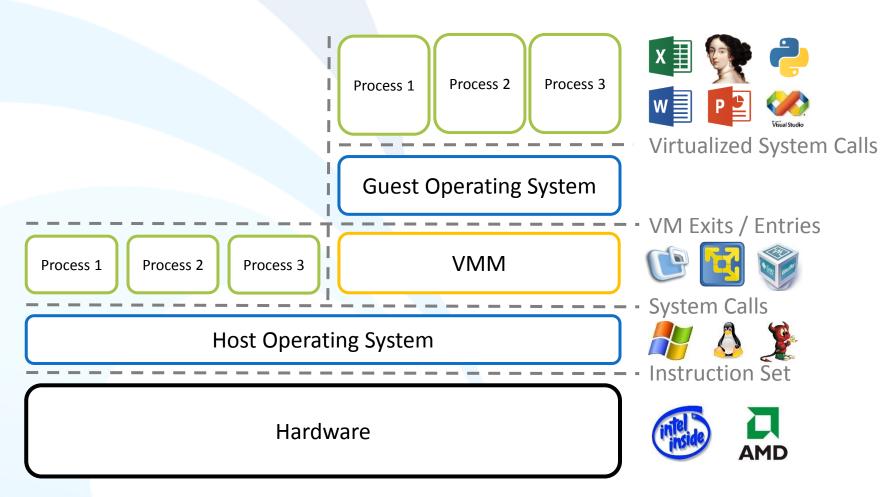
SOFTWARE STACKS

Piling different pieces of software

Software Stack Type 1 Hypervisor



Software Stack Type 2 Hypervisor



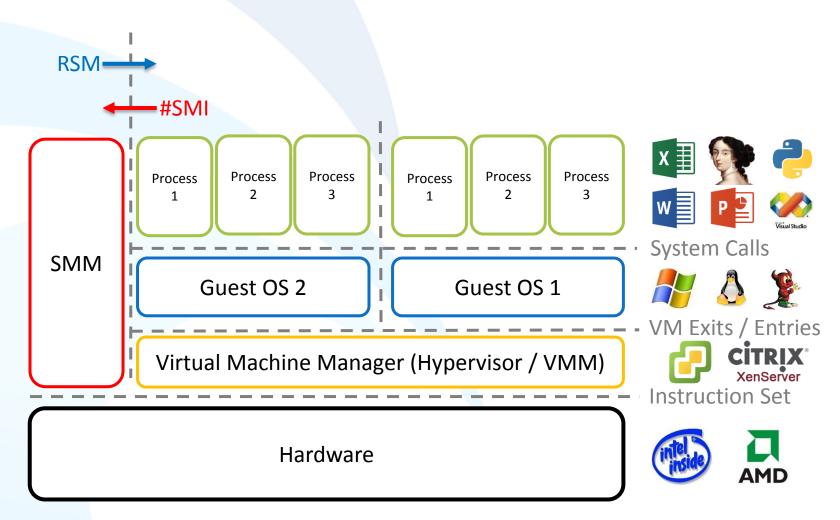


ISA emulation challenges

- A VMM needs to emulate every instruction or event it registers on
- A VMM <u>must</u> register to a certain set of instructions and x86 events known as the "fixed-1 exits"
 - e.g: CPUID, GETSEC, INVD, XSETBV and various VT ISA
- ISA emulation challenges
 - Specification
 - Corner cases
 - Deciding if the guest has the right privilege from root operation is hard
 - Confused deputy situation...

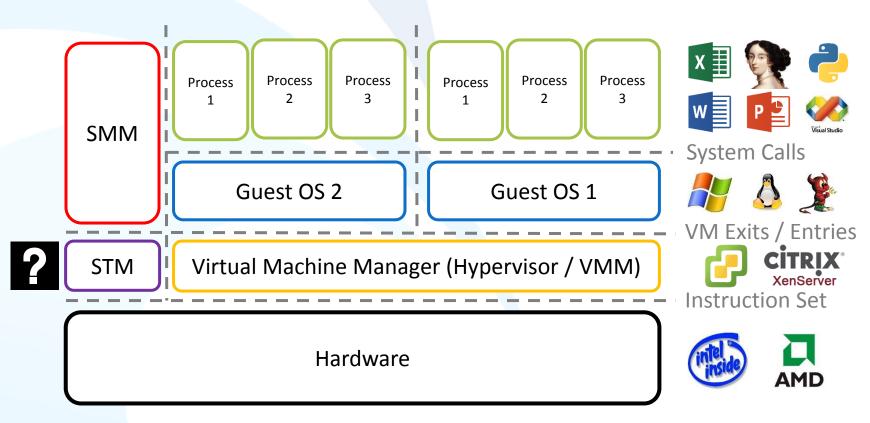


Software Stack SMM with VMM



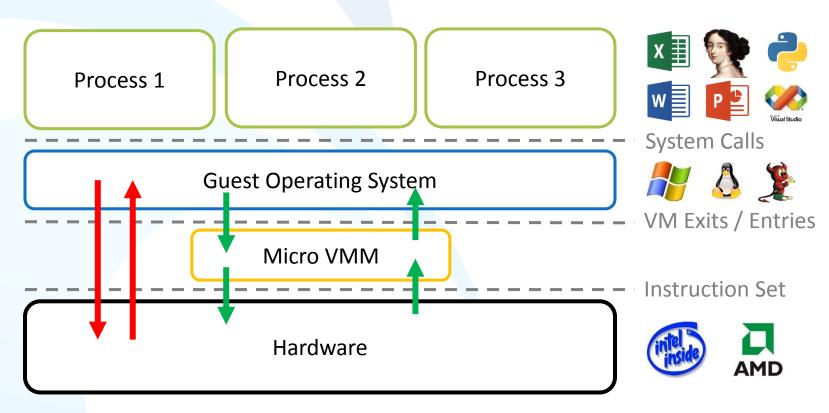


Software Stack SMM Transfer Monitor (STM)





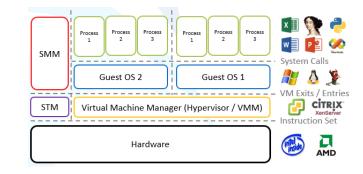
Micro-VMMs





- CONFIDENTIAL -

Section summary



- There are many ways to use hardware virtualization technology:
 - Type I, Type II, Micro-VMMs, ...
- Each approach has its own unique challenges:
 - Full HW virtualization: Secure a big implementation of SW emulation for all HW
 - Para-virtualization: Secure the guest OS interface with the host OS
 - All implementations: Emulate ISA correctly and securely
 - Micro-VMMs: Defend from HW subversion
- SMM is too privileged and where are the STMs?



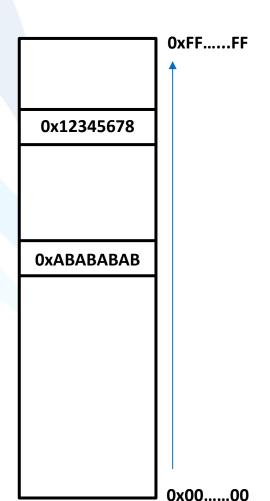
-CONFIDENTIAL-

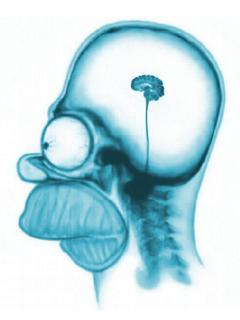
MEMORY

Where did I put that instruction?

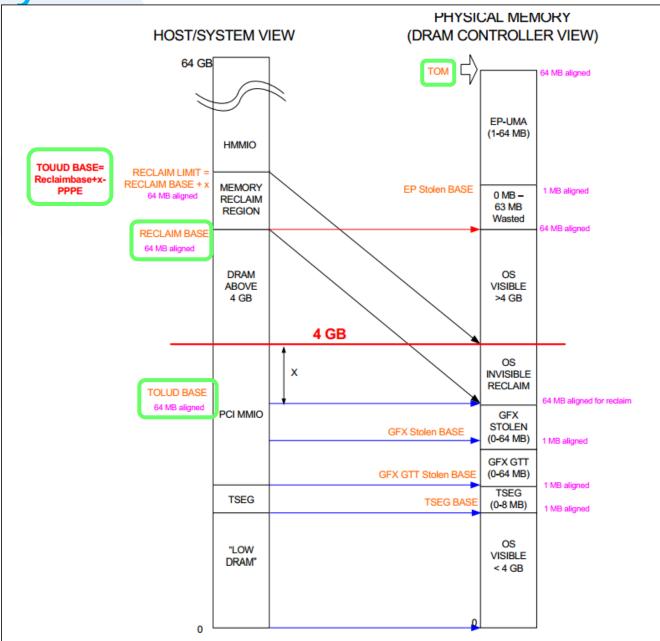
Memory is simple, right?





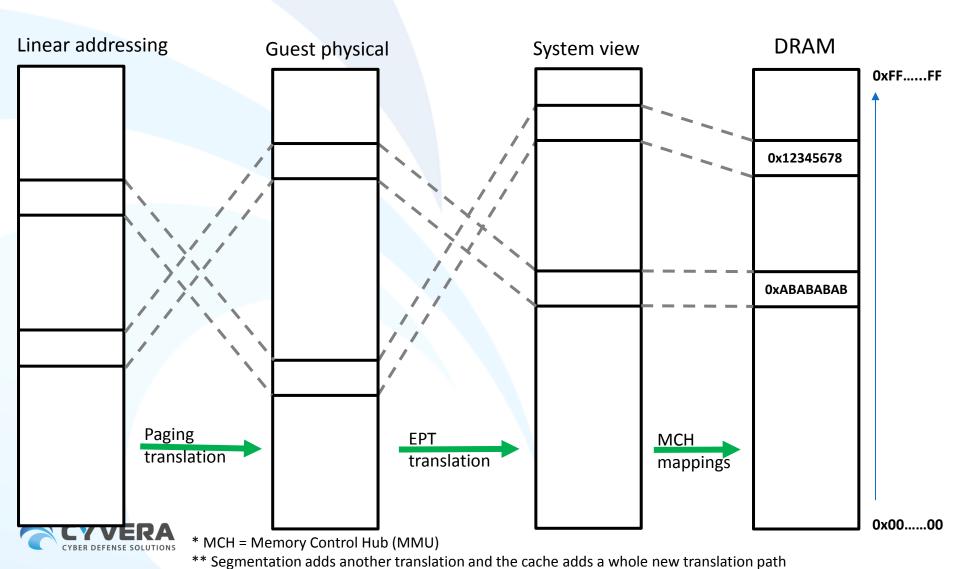


Memory - Intel manual

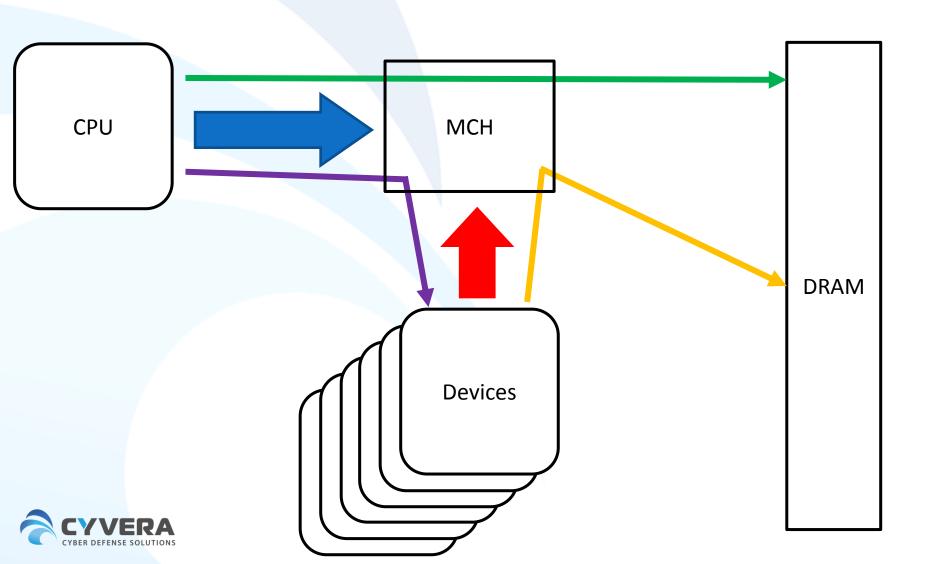




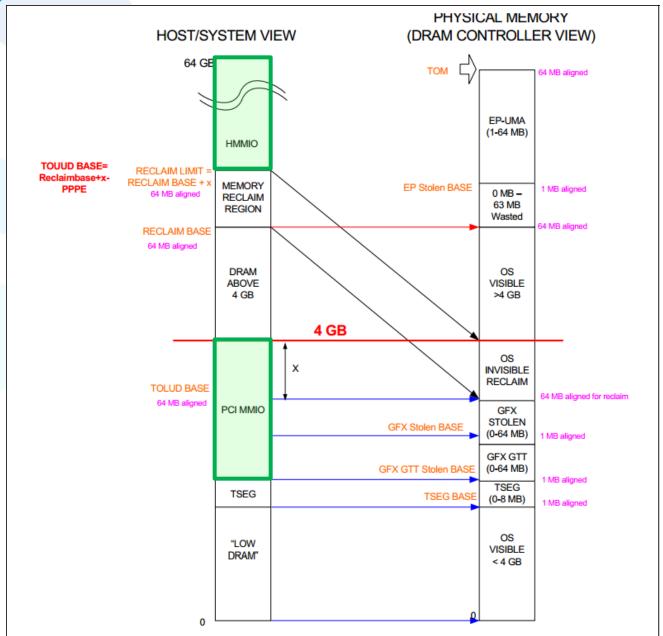
Memory – Address Translations



Memory – point of view



Memory - MMIO





Special address ranges

Figure 2-3. Main Memory Address Range

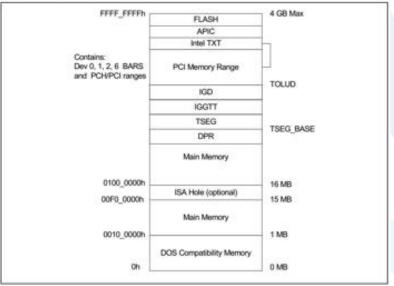


Figure 7-4. PCI Memory Address Range

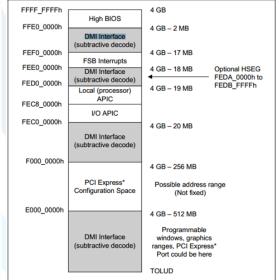


Figure 2-1. System Address Range

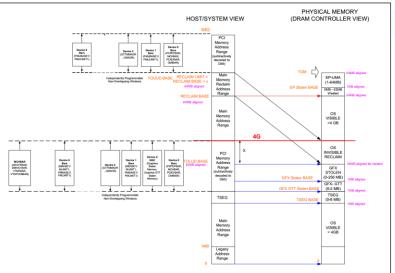
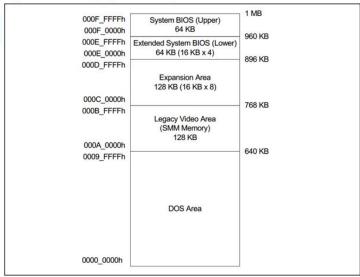


Figure 2-2. DOS Legacy Address Range



Cache!

Sorry, out of scope for today-

there is no end to it once you start discussing cache and security

Suffice to say that it adds another translation layer and that it is complex and performance

oriented

performance

Security



Section summary

- Memory is complex!
- Attackers with access to MMIO or physical memory addresses can compromise anything on the system
- Access to special address ranges is also dangerous
- EPT can help mitigate some of the problems
 - If you can configure it correctly, if it is available



- CONFIDENTIAL -

COMPUTER PLATFORMS

The insides

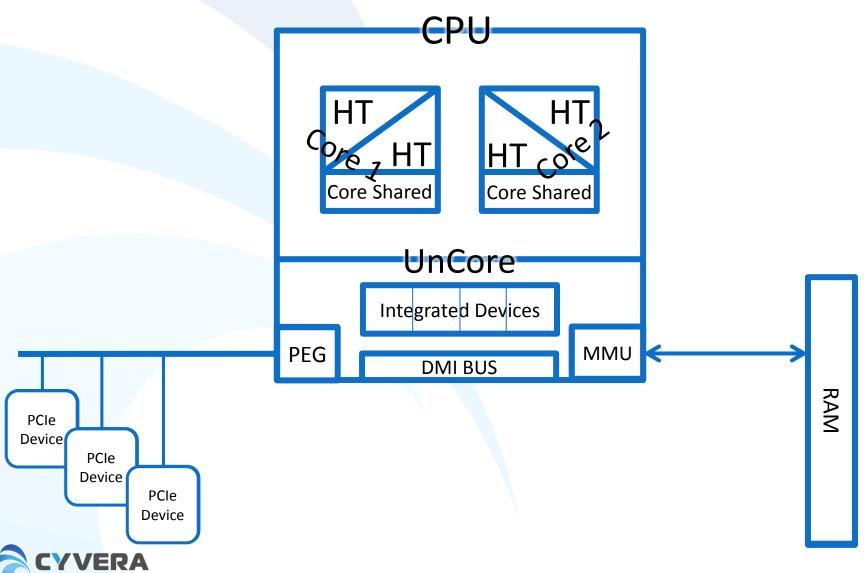
What is a computer?

- A very complex device internally
- The logical software architecture can be complex
- Every modern computer system is also a complex high speed network of interconnecting hardware components using many communication protocols

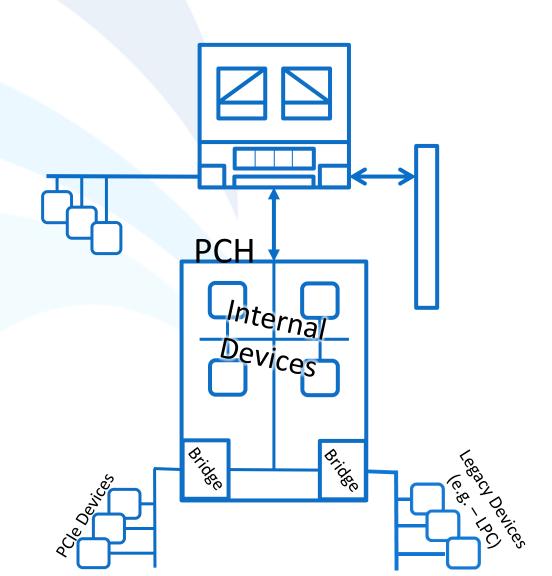




Computer Platform (1)



Computer platform (2)





Device virtualization

- XEN and KVM use a modified QEMU
 - No vulnerabilities there, right?

| 1 | | CVE ID | CWE ID # of Exploits | Vulnerability Type(s) | Publish Date | Update Date | Score | Gained Access Level | Access | Complexity | Authentication | Conf. | Integ. | Avail. |
|--|--|--|---|--|--|---|---|---|--|--|--|---|--|---|
| 2 OFF-2013-2344 19 Overflow PPIV 2013-10-04 2013-10-03 6.0 None Local High Single system Complete Complete for overflow in the ECSI implementation in QENU, as used in Xen, when a SCSI controller has more than 256 at the devices, allows local users to gain privileges via a small transfer buffer in a NS command. 3 OFF-2013-2007 264 2013-06-71 2013-06-72 2013-010 30 None Local Medium Not required Complete Complete Complete or overflow in the 1000 device function in the 1000 device driver (New 1000-0) in QENU 1.1.0-rc2 and other versions, when the SEBP and LPE flags are disabled, allows remote attackers to easier service (guest OS crash) and possibly execute arbitrary guest code via a large packet. 5 OFF-2012-2551 20 +PfV 2012-11-23 2013-10-10 7.2 None Local Low Not required Complete Complete Soft-2012-2551 20 +PfV 2012-11-23 2013-10-10 7.2 None Local Low Not required Complete Complete Soft-2012-2551 20 +PfV 2012-11-23 2013-10-10 7.2 None Local Low Not required Complete Complete Soft-2012-2551 20 +PfV 2012-11-23 2013-10-10 7.2 None Local Low Not required Complete Complete Soft-2012-2551 20 +PfV 2012-11-23 2013-10-10 7.2 None Local Low Not required Partial Partial Soft-2012-2552 20 PfV 2012-2012-25 20 PfV 2012-25 20 PfV | | | | | | | | | Network | | | | | Partial |
| Suffer overflow in the SCSI implementation in QEMU, as used in Xen, when a SCSI controller has more than 256 attached devices, allows local users to gain privileges via a small transfer buffer in a UNIS command. 3 CVE-2011-2007 264 2013-05-21 2013-08-22 6.9 None Local Medium Not required Complete Complete He germu quest agent in Qemu 1.4.1 and earlier, as used by Xen, when started in daemon mode, uses weak permissions for certain files, which allows local users to read and write to these files. A CVE-2012-2013 112 DoS Devectode 2013-02-12 2013-10-10 Su None Remote Medium Not required Complete Complete Germitation in the e1000 receive function in the e1000 device driver (flow/e1000.c) in QEMU 1.3.0-rc2 and other versions, when the SEP and LPE flags are disabled, allows remote attackers to caus fearing (guest of Sc srah) and possibly vencute arbitrary guest code via a large packets. 5 CVE-2012-2515 20 +Priv 2012-11-23 2013-10-10 7.2 None Local Low Not required Complete Complete Remu, as used to Xen 4.0, 4.1, and possibly other products, when emulating certain devices with a virtual console backend, allows local Users to gain privileges via a crafted escape VT100 service Germination in Qemu 1.0 does not properly handle the failure of the mixtemp function, when in snapshot node, which allows local Users to gain privileges via a crafted escape VT100 service Germination in Qemu 1.0 does not properly handle the failure of the mixtemp function, when in snapshot node, which allows local users to overwrite or read arbitrary files via a symilink et negocified temporary file. 7 CVE-2011-2522 264 2012-06-21 2012-06-26 7.4 None Local Medium Not required Partial None he change_process_uid function in os-positic in Qemu 0.14,0 and earlier allows privileged quest users to occur arbitrary color to version in the virtio subsystem in genus-km 0.14,0 and earlier does not properly drop group privileges when the -runs option is used, which allows local guest users to accuse a denial of service (guest crash) or gain privileges via | se-a | fter-free vulnera | bility in the virtio-pci | implementation in Qem | u 1.4.0 through | 1.6.0 allows loc | al users to | cause a denial of | f service (da | emon crash) by | "hot-unplugging" a | virtio device. | | |
| 3 | 2 <u>C</u> \ | VE-2013-4344 | 119 | Overflow +Priv | 2013-10-04 | 2013-10-23 | 6.0 | None | Local | High | Single system | Complete | Complete | Complet |
| the qemu guest agent in Qemu 1.4.1 and earlier, as used by Xen, when started in deemon mode, a considerable of the quantity of the provided complete and the quantity of the q | | | SCSI implementation | in QEMU, as used in Xe | en, when a SCS | SI controller has | more than | 256 attached dev | vices, allows | local users to ga | ain privileges via a | small transfe | r buffer in a l | REPORT |
| Dos Exec Code 2013-02-12 2013-10-10 10 None Remote Medium Not required Complete Complete Overflow in the e1000_recious function in the e1000 device driver (hw/e1000-c) in QEMU 1.3.0-r2 and other versions, when the SBP and LPE flags are disabled, allows remote attackers to cause for service (quest to Scrash) and possibly execute arbitrary guest code vise a large packet. 5 VF-2012-2515 20 + Priv 2012-11-23 2013-10-10 7,2 None Local Low Not required Complete Complete Privacy as used in Xen 4.0, 4.1 and possibly other products, when emulating certain devices with a virtual console backend, allows local OS guest users to gain privileges via a crafted except V1100 s. register to everywrite of a "device models" address space. 6 OFE-2012-2652 20 2012-06-20 2013-01-10 4.4. None Local Medium Not required Partial Partial Privacy Privac | 3 <u>C</u> \ | VE-2013-2007 | 264 | | 2013-05-21 | 2013-08-22 | 6.9 | None | Local | Medium | Not required | Complete | Complete | Complet |
| Overflow The 1000 receive function in 1000 receive function | he q | emu guest agent | t in Qemu 1.4.1 and e | arlier, as used by Xen, | when started in | n daemon mode, | uses weak | permissions for | certain files, | which allows los | cal users to read ar | nd write to the | se files. | |
| of service (guest OS crash) and possibly execute arbitrary guest code via a large packet. S VE-2012-3515 20 | 4 🕚 | VE-2012-6075 | 119 | | 2013-02-12 | 2013-10-10 | 9.3 | None | Remote | Medium | Not required | Complete | Complete | Complet |
| termu, as used in Xen 4.0, 4.1 and possibly other products, when emulating certain devices with a virtual console backend, allows local OS guest users to gain privileges via a crafted escape VT100 singlers the overwrite of a "device model's address space." 6 CVE-2012-2652 2012-08-07 2013-04-18 4.4 None Local Medium Not required Partial Partia | | | _ | | | | 3.0-rc2 and | other versions, | when the SE | P and LPE flags | are disabled, allow | s remote atta | ckers to caus | se a denia |
| 19gers the overwrite of a "device model"s address space." 2012-08-07 2013-04-18 4.4 None Local Medium Not required Partial Pa | 5 <u>C</u> \ | VE-2012-3515 | 20 | +Priv | 2012-11-23 | 2013-10-10 | 7.2 | None | Local | Low | Not required | Complete | Complete | Comple |
| the bdry_open function in Qemu 1.0 does not properly handle the failure of the mkstemp function, when in snapshot node, which allows local users to overwrite or read arbitrary files via a symlink at snapscified temporary file. 7 CVE-2011-2527 264 2012-06-21 2012-06-26 7.1 None Local Low Not required Partial None he change_process_uid function in os-posix.c in Qemu 0.14.0 and earlier does not properly drop group privileges when the -runas option is used, which allows local guest users to access restricted fost. 8 CVE-2011-2212 119 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete Utfer overflow in the virtio subsystem in qemu-kvm 0.14.0 and earlier allows privileged guest users to cause a denial of service (guest crash) or gain privileges via a crafted indirect descriptor relate virtiqueue in and out requests." 9 CVE-2011-1751 20 DoS Exec Code 2012-06-21 2013-02-13 7.4 None Local Medium Single system Complete Complete Network he pciej_write function in hw/acpi_piix4.c in the PIIX4 Power Management emulation in gemu-kvm does not check if a device is hotpluggable before unplugging the PCI-ISA bridge, which allows privileges to cause a denial of service (guest crash) and possibly execute arbitrary code by sending a crafted value to the 0xaeo6 (PCI_ED_BASE) I/O port, which leads to a use-after-free related to "active complete Comple | | | | | emulating certa | in devices with a | virtual con | sole backend, all | lows local OS | guest users to | gain privileges via | a crafted esc | ape VT100 se | equence t |
| Any Control of the property file. 7 OVE-2011-2527 264 2012-06-21 2012-06-26 2.1 None Local Low Not required Partial None he change_process_uid function in os-posix.c in Qemu 0.14.0 and earlier does not properly drop group privileges when the -runas option is used, which allows local guest users to access retricted fost. 8 OVE-2011-2212 119 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Network Network Network (Indicated in the virtio subsystem in qemu-kvm 0.14.0 and earlier allows privileged guest users to cause a denial of service (guest crash) or gain privileges via a crafted indirect descriptor relate intriqueue in and out requests." 9 OVE-2011-1751 20 DoS Exec Code 2012-06-21 2013-02-13 7.4 None Local Medium Single system Complete Complete Network Network (Indicated in the PIIX4 Power Management emulation in qemu-kvm does not check if a device is hotpluggable before unplugging the PCI-ISA bridge, which allows privileges to cause a denial of service (guest crash) and possibly execute arbitrary code by sending a crafted value to the 0xae08 (PCI_EJ_BASE) I/O port, which leads to a use-after-free related to "activate" (Indicated in the Virtio-blk driver (hw/virtio-blk.c) in qemu-kvm 0.14.0 allow local guest users to cause a denial of service (guest crash) and possibly gain privileges via a (1 CVE-2011-1750 119 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Network Networ | 6 <u>C</u> | VE-2012-2652 | | | 2012-08-07 | 2013-04-18 | 4.4 | None | Local | Medium | Not required | Partial | Partial | Partia |
| he change_process_uid function in os-posix.c in Qemu 0.14.0 and earlier does not properly drop group privileges when the -runas option is used, which allows local guest users to access restricted fost. 8 CVE-2011-2212 119 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete uffer overflow in the virtio subsystem in qemu-kym 0.14.0 and earlier allows privileged guest users to cause a denial of service (guest crash) or gain privileges via a crafted indirect descriptor relate virtiqueue in and out requests." 9 CVE-2011-1751 20 DoS Exec Code 2012-06-21 2013-02-13 7.4 None Local Medium Single system Complete Complete Network 10 CVE-2011-1750 119 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete Overflow in the virtio-blk driver (hw/virtio-blk.c) in qemu-kym does not check if a device is hotpluggable before unplugging the PCI-ISA bridge, which allows privagers to cause a denial of service (guest crash) and possibly execute arbitrary code by sending a crafted value to the 0xae08 (PCI_E)_BASE) I/O port, which leads to a use-after-free related to "activates" and the virtio-blk driver (hw/virtio-blk.c) in qemu-kym 0.14.0 allow local guest users to cause a denial of service (guest crash) and possibly gain privileges via a (1 CVE-2011-1750 119 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete Utiple heap-based buffer overflows in the virtio-blk driver (hw/virtio-blk.c) in qemu-kym 0.14.0 allow local guest users to cause a denial of service (guest crash) and possibly gain privileges via a (1 CVE-2011-0011 287 Bypas 2012-06-21 2012-06-21 4.3 None Local High Not required Partial Partial Network 11 CVE-2011-0011 287 Bypas 2012-06-21 2012-06-21 7.2 None Local Low Not required Complete Complete Overflow in the usb_host_handle_control function in the USB passthrough handling implementation in usb-linux.c in QEMU before 0.11.1 allows guest OS users to cause a denial of service (guest rash) or pace | | | | ot properly handle the | failure of the m | kstemp function | , when in sr | napshot node, wh | nich allows lo | ocal users to ove | rwrite or read arbi | trary files via | a symlink at | tack on a |
| B CVE-2011-2212 119 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Network Wetwork | 7 <u>C</u> | VE-2011-2527 | 264 | | 2012-06-21 | 2012-06-26 | 2.1 | None | Local | Low | Not required | Partial | None | None |
| Network Network Network Network Network One of the price of the p | | hange_process_u | uid function in os-posi | x.c in Qemu 0.14.0 and | l earlier does n | ot properly drop | group privi | leges when the - | runas option | is used, which a | allows local guest u | sers to access | s restricted fi | les on the |
| Virtqueue in and out requests." 9 CVF-2011-1751 20 DoS Exec Code 2012-06-21 2013-02-13 7.4 None Local Medium Single system Complete Complete Network 10 CVF-2011-1750 119 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete Network 10 CVF-2011-1750 119 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete Network 11 CVF-2011-1750 129 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete Network 12 CVF-2011-1750 139 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete Network 13 CVF-2011-1750 149 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete Network 14 CVF-2011-0011 287 Bypass 2012-06-21 2012-06-21 4.3 None Local High Not required Partial Partial Network 15 CVF-2011-0011 287 Bypass 2012-06-21 2012-06-21 4.3 None Local High Not required Partial Partial Network 16 CVF-2011-0011 287 Bypass 2012-06-21 2012-06-21 7.2 None Local Low Not required Complete Overflow in the usb_host_handle_control function in the USB passthrough handling implementation in usb-linux.c in QEMU before 0.11.1 allows guest OS users to cause a denial of service (guest crash) and possibly service (guest crash) and possibly service (guest crash) and possibly gain privileges via a (1 to VF-2011-0011) 287 Bypass 2012-06-21 2012-06-21 4.3 None Local High Not required Partial Partial Network 17 CVF-2010-0297 119 DoS Exec Code 2010-02-12 2010-08-21 7.2 None Local Low Not required Complete Complete Overflow in the usb_host_handle_control function in the USB passthrough handling implementation in usb-linux.c in QEMU before 0.11.1 allows guest OS users to cause a denial of service (guest crash) and possibly execute arbitrary code on the host OS via a crafted USB packet. | 8 <u>C\</u> | VE-2011-2212 | 119 | DoS Overflow +Priv | 2012-06-21 | 2012-06-26 | 7.4 | None | | Medium | Single system | Complete | Complete | Comple |
| Network None Network Networ | | | | | arlier allows pri | vileged guest use | ers to cause | e a denial of serv | rice (guest cr | rash) or gain pri | vileges via a crafte | d indirect des | criptor relate | d to |
| sers to cause a denial of service (guest crash) and possibly execute arbitrary code by sending a crafted value to the 0xae08 (PCI_EJ_BASE) I/O port, which leads to a use-after-free related to "active mers." 10 CVE-2011-1750 119 DoS Overflow +Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete Network 11 CVE-2011-1750 129 DoS Overflow + Priv 2012-06-21 2012-06-26 7.4 None Local Medium Single system Complete Complete Network 12 CVE-2011-1011 287 Bypass 2012-06-21 2012-06-21 4.3 None Local High Not required Partial Partial Network 13 CVE-2011-0011 287 Bypass 2012-06-21 2012-06-21 4.3 None Local High Not required Partial Partial Network 14 CVE-2011-0011 287 Bypas 2012-06-21 2012-06-21 7.2 None Local Low Not required Complete Complete Complete Network Network 15 CVE-2010-0297 119 DoS Exec Code 2010-02-12 2010-08-21 7.2 None Local Low Not required Complete Complete Overflow 16 CVE-2010-0297 129 DoS Exec Code 2010-02-12 2010-08-21 7.2 None Local Low Not required Complete Complete Overflow 17 Complete Complete Overflow in the usb_host_handle_control function in the USB passthrough handling implementation in usb-linux.c in QEMU before 0.11.1 allows guest OS users to cause a denial of service (guest range) or possibly execute arbitrary code on the host OS via a crafted USB packet. | | | | emu-kvm 0.14.0 and ea | | | | | | | | | | |
| Network Ultiple heap-based buffer overflows in the virtio-blk driver (hw/virtio-blk.c) in qemu-kvm 0.14.0 allow local guest users to cause a denial of service (guest crash) and possibly gain privileges via a (1 cyt-2011-0011 287 Bypass 2012-06-21 2012-06-21 4.3 None Local High Not required Partial Partial Network emu-kvm before 0.11.0 disables VNC authentication when the password is cleared, which allows remote attackers to bypass authentication and establish VNC sessions. 1.2 Cyt-2010-0297 119 DoS Exec Code 2010-02-12 2010-08-21 7.2 None Local Low Not required Complete Overflow uffer overflow in the usb_host_handle_control function in the USB passthrough handling implementation in usb-linux.c in QEMU before 0.11.1 allows guest OS users to cause a denial of service (guest crash) and possibly gain privileges via a (1 cyt-2010-021) and | /irtqu | ueue in and out r | requests." | | 2012-06-21 | 2013-02-13 | 7.4 | None | | Medium | Single system | Complete | Complete | Comple |
| Equest to the virtio_blk_handle_write function or (2) read request to the virtio_blk_handle_read function that is not properly aligned. 1 CVE-2011-0011 287 Bypass 2012-06-21 2012-06-21 4.3 None Local High Not required Partial Partial Network 1 CVE-2011-0011 287 Bypass 2012-06-21 2012-06-21 4.3 None Local High Not required Partial Partial Network 2 CVE-2010-0297 119 DoS Exec Code 2010-02-12 2010-08-21 7.2 None Local Low Not required Complete Overflow 2 CVE-2010-0297 19 DoS Exec Code 2010-02-12 2010-08-21 7.2 None Local Low Not required Complete Complete Overflow 2 CVE-2010-0297 19 DoS Exec Code 2010-02-12 2010-08-21 7.2 None Local Low Not required Complete Complete Overflow in the usb_host_handle_control function in the USB passthrough handling implementation in usb-linux.c in QEMU before 0.11.1 allows guest OS users to cause a denial of service (guest hand) or possibly execute arbitrary code on the host OS via a crafted USB packet. | rirtqu 9 <u>C\</u> ne po sers | ueue in and out r VE-2011-1751 ciej_write functio to cause a denia | requests." 20 on in hw/acpi_piix4.c i | DoS Exec Code | agement emula | tion in qemu-kvr | n does not | check if a device | Network is hotplugg | able before unpl | ugging the PCI-ISA | bridge, which | h allows privi | Comple |
| Network amu-kvm before 0.11.0 disables VNC authentication when the password is cleared, which allows remote attackers to bypass authentication and establish VNC sessions. 2 CVE-2010-0297 119 DOS Exec Code 2010-02-12 2010-08-21 7.2 None Local Low Not required Complete Complete Overflow Overflow Uffer overflow in the usb_host_handle_control function in the USB passthrough handling implementation in usb-linux.c in QEMU before 0.11.1 allows guest OS users to cause a denial of service (guest hang) or possibly execute arbitrary code on the host OS via a crafted USB packet. | ne posers | ueue in and out r VE-2011-1751 ciej_write functio to cause a denia s." | requests." 20 on in hw/acpi_piix4.c i | DoS Exec Code n the PIIX4 Power Manash) and possibly execu | agement emula ute arbitrary co | ition in qemu-kvr de by sending a | n does not crafted valu | check if a device ue to the 0xae08 | Network is hotplugg; (PCI_EJ_BA | able before unpl .SE) I/O port, wh | ugging the PCI-ISA nich leads to a use- | bridge, which | h allows privi ated to "activ | Comple leged gu e qemu |
| 2 CVE-2010-0297 119 DoS Exec Code 2010-02-12 2010-08-21 7.2 None Local Low Not required Complete Complete Overflow uffer overflow in the usb_host_handle_control function in the USB passthrough handling implementation in usb-linux.c in QEMU before 0.11.1 allows guest OS users to cause a denial of service (guest hang) or possibly execute arbitrary code on the host OS via a crafted USB packet. | ne posers mers | ueue in and out r VE-2011-1751 ciej_write functic to cause a denia s." VE-2011-1750 ele heap-based bi | requests." 20 on in hw/acpi_piix4.c is of service (guest crist) 119 uffer overflows in the | DoS Exec Code n the PIIX4 Power Manash) and possibly executors DoS Overflow +Privious virtio-blk driver (hw/virtio-blk driver (hw/ | agement emula ute arbitrary co 2012-06-21 rtio-blk.c) in qe | ation in gemu-kvr de by sending a 2012-06-26 mu-kvm 0.14.0 a | 7.4 | check if a device ue to the 0xae08 None guest users to ca | Network is hotpluggs (PCI_EJ_BA Local Network use a denial | able before unpl SE) I/O port, wh | ugging the PCI-ISA nich leads to a use- Single system | bridge, which after-free rela | h allows privi ated to "activ Complete | Comple leged gu e qemu Comple |
| Overflow Uffer overflow in the usb_host_handle_control function in the USB passthrough handling implementation in usb-linux.c in QEMU before 0.11.1 allows guest OS users to cause a denial of service (guest hang) or possibly execute arbitrary code on the host OS via a crafted USB packet. | ne posers ners 0 C | ueue in and out r VE-2011-1751 ciej_write functic to cause a denie s." VE-2011-1750 ele heap-based bi st to the virtio_b | requests." 20 on in hw/acpi_piix4.c i al of service (guest cri 119 uffer overflows in the alk_handle_write funct | DoS Exec Code n the PIIX4 Power Man- ash) and possibly execu- DoS Overflow +Priv virtio-blk driver (hw/vi- ion or (2) read request | agement emula ute arbitrary co 2012-06-21 rtio-blk.c) in qe to the virtio_bl | ation in qemu-kvr de by sending a 2012-06-26 mu-kvm 0.14.0 a k_handle_read fi | 7.4 allow local gunction that | check if a device ue to the 0xae08 None guest users to ca is not properly a | Network is hotpluggs (PCI_EJ_BA Local Network use a denial aligned. Local | able before unpl SE) I/O port, wh Medium of service (gues | ugging the PCI-ISA nich leads to a use- Single system st crash) and possil | c bridge, which after-free rela Complete Day gain privik | h allows privi ated to "activ Complete eges via a (1 | Comple leged gu e qemu Comple |
| hang) or possibly execute arbitrary code on the host OS via a crafted USB packet. | ne personers O C ultiple que: | ueue in and out r VE-2011-1751 ciej_write functic to cause a denia s." VE-2011-1750 le heap-based bi st to the virtio_b VE-2011-0011 | on in hw/acpi_piix4.c i al of service (guest cri 119 uffer overflows in the llk_handle_write funct 287 | DoS Exec Code In the PIIX4 Power Maniash) and possibly execution DoS Overflow +Privition or (2) read request Bypass | agement emula ute arbitrary co 2012-06-21 rtio-blk.c) in qe to the virtio_bl 2012-06-21 | tion in qemu-kvr de by sending a 2012-06-26 mu-kvm 0.14.0 c k_handle_read for 2012-06-21 | 7.4 allow local gunction that | check if a device ue to the 0xae08 None guest users to ca is not properly a | Network is hotpluggs (PCI_EJ_BA Local Network use a denial aligned. Local Network | able before unpl SE) I/O port, wh Medium of service (gues | ugging the PCI-ISA nich leads to a use- Single system st crash) and possil Not required | c bridge, which after-free rela Complete Day gain privik | h allows privi ated to "activ Complete eges via a (1 | Comple leged gu e qemu Comple) write |
| 3 CVE-2009-3616 399 Exec Code 2009-10-23 2009-12-19 8.5 None Remote Medium Single system Complete Complete | irtquiers O C ultipl que: | veue in and out of VE-2011-1751 ciej_write function to cause a denie s." VE-2011-1750 le heap-based bist to the virtio_b VE-2011-0011 -kvm before 0.1: | on in hw/acpi_piix4.c is all of service (guest critis) 119 uffer overflows in the lik_handle_write funct 287 1.0 disables VNC auth | DoS Exec Code In the PIIX4 Power Maniash) and possibly execution DoS Overflow +Privion or (2) read request Bypass Intentication when the past DoS Exec Code | agement emula ute arbitrary co 2012-06-21 rtio-blk.c) in qe to the virtio_bl 2012-06-21 ssword is cleare | otion in qemu-kvr de by sending a 2012-06-26 mu-kvm 0.14.0 a k_handle_read for 2012-06-21 ed, which allows | 7.4 allow local gunction that | check if a device ue to the 0xae08 None guest users to call is not properly a None sckers to bypass | Network a is hotpluggi (PCI_EJ_BA Local Network use a denial aligned. Local Network authenticatic | able before unpl SE) I/O port, wh Medium of service (gue: High on and establish | ugging the PCI-ISA nich leads to a use- Single system st crash) and possit Not required VNC sessions. | bridge, which after-free rela Complete Dly gain privile Partial | h allows privi ated to "activ Complete eges via a (1 | Comple leged gu e qemu Comple) write |
| | rirtqu 9 © one posers mers 0 © one posers mers 1 © one posers 2 © one posers uffer | weue in and out of VE-2011-1751 ciej_write function to cause a denie of very service of the very service of v | on in hw/acpi_piix4.c i al of service (guest cri 119 uffer overflows in the alk_handle_write funct 287 1.0 disables VNC auth 119 usb_host_handle_con | DoS Exec Code In the PIIX4 Power Maniash) and possibly execu- DoS Overflow +Privion or (2) read request Bypass Interior when the pair DoS Exec Code Overflow Ittorion in the USB | agement emula ute arbitrary co 2012-06-21 rtio-blk.c) in qe to the virtio_bl 2012-06-21 ssword is cleare 2010-02-12 | ation in qemu-kvr de by sending a 2012-06-26 mu-kvm 0.14.0 a k_handle_read for 2012-06-21 ad, which allows 2010-08-21 andling impleme | 7.4 allow local question that 4.3 remote atta | check if a device ue to the 0xae08 None guest users to ca is not properly a None ackers to bypass None | Network a is hotplugg; (PCI_EJ_BA Local Network uuse a denial aligned. Local Network authenticatie Local | able before unpl ISE) I/O port, wh Medium of service (gues High on and establish Low | ugging the PCI-ISA nich leads to a use- Single system st crash) and possil Not required VNC sessions. Not required | complete Complete Partial Complete | n allows privi ated to "activ Complete ages via a (1 Partial | Compleged green e qemu Complete y write Parti |



VT-d (IOMMU)

- Used for virtualizing chipset components
- DMA remapping
 - Paging for devices
 - Nested translations
- Interrupt remapping

Second-level paging structure PASID Table SL-PDPE SL-PDE SL-PTE First-level paging SL-PML4E SL-PTE structure pointer SL-PTE SL-PDE PDPE SL-PTE SL-PDE SL-PDE SL-PTE SL-PDPE SL-PML4E SL-PDE SL-PTE Final Page

Figure 3-12. Nested Translation with 4-KByte pages

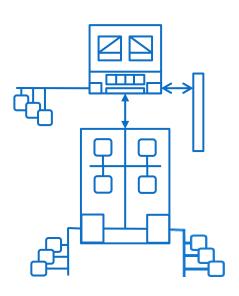
- Allows directing interrupts coming from hardware There is a good paper that explains the need for it by Rafal Wojtczuk and Joanna Rutkowska:
 - Following the White Rabbit: Software attacks against Intel® VT-d technology
- What about older systems where you don't have VT-d?



Section summary

- Computer hardware is complex!
- Emulating necessary components is hard:
 - Multiple CVEs already found in ACPI and APIC virtualization as well as QEMU
- VT-d helps virtualizing DMA and hardware interrupts
 - If used correctly





- CONFIDENTIAL -

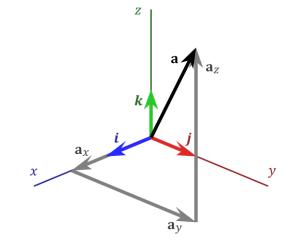
ATTACK VECTORS TAXONOMY

Potential and practical ones

- CONFIDENTIAL -

Basic Vectors

- ISA Implementation
 - Emulating x86 isn't easy...



- Performance monitoring
 - Classic side channels
 - Real Time Instruction Tracing new feature coming up
- Old systems
 - New defenses were introduced with the latest HW
- New features
 - Approach to new features (CPU/PCH)
 - Whitelist or Blacklist?



Address Space Attacks



- IO Address Space
 - How many IO ports are there in x86?
 - What happens if we configure port overlaps?
- MMIO Overlaps
 - As discussed during the presentation
- Special memory ranges access and overlaps
 - What happens when a guest can access special ranges?
- MSR address space
 - MSRs define system configuration and behavior



Privileged Software



- Corrupted ACMs
 - ACMs run in a very high privilege, if you can compromise one...
- CPU/PCH Firmware(s)
 - Compromise of the CPU or PCH firmware naturally allows an attacker to control any VM
- BIOS & SMM
 - The BIOS is a common component of the platform and controls both configuration and SMM code



Other Interesting Vectors



- Intentional misconfiguration
 - It is possible to misconfigure PCIe config space, MSRs or MMIO constants in order to create unexpected situations for the VMM
- Server platforms (are fun!)
 - Platforms and CPUs for the server market have special features, all of those usually run in very high privilege
- Errata
 - What if there is an Errata in the CPU/PCH behavior we rely on to emulate something - sucks, right? ☺



Bonus: Interesting Errata

- The below Errata appears in the June 2013 revision for 2nd generation core CPUs
- Sounds like an exploitable issue <u>IF</u> you can prevent reload of CR3 with 32bit value

BK124. The Upper 32 Bits of CR3 May be Incorrectly Used With 32-Bit Paging

Problem: When 32-bit paging is in use, the processor should use a page directory located at the

32-bit physical address specified in bits 31:12 of CR3; the upper 32 bits of CR3 should be ignored. Due to this erratum, the processor will use a page directory located at the

64-bit physical address specified in bits 63:12 of CR3.

Implication: The processor may use an unexpected page directory or, if EPT (Extended Page Tables)

is in use, cause an unexpected EPT violation. This erratum applies only if software enters 64-bit mode, loads CR3 with a 64-bit value, and then returns to 32-bit paging without changing CR3. Intel has not observed this erratum with any commercially

available software.

Workaround: Software that has executed in 64-bit mode should reload CR3 with a 32-bit value

before returning to 32-bit paging.

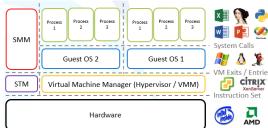
Status: For the steppings affected, see the Summary Tables of Changes.



Summary

- Computer platforms are complex
- There are several approaches to virtualizing HW, each with its own inherent weaknesses
 - Full hardware virtualization: slower and uses SW emulation, therefore prone to SW vulnerabilities
 - Direct hardware access: prone to malicious HW manipulations (micro-VMMs)
- Better defenses are available only with new and sometimes also high end HW





- CONFIDENTIAL-

THE END

Contact me at: http://www.cyvera.com/contact

-CONFIDENTIAL-

Useful tools and info for people interested in virtualization research

- LOLA by Jeff Forristal (free)
 - Because a Python interface to the HW rocks!
- 8 series PCH manuals
- 2nd generation core Errata
- Intel software developer manuals
 - Volume 3 contains most information about VT
 - Other volumes are also useful to understand what is emulated
- Patience!
 - Hardware debugging, reading long technical manuals



How different virtualization SW works

- VMware Player
 - Emulates 440BX motherboard (15 years old)
 - Monitors PCIe configuration, at least IO ports, to some degree but because of Win95 and Win3.1 compatibility, not security!



Disclaimer

 All product logos and names used in this presentation are the property of their respective owners. I make no claim for ownership on those. I am merely using them as examples of such products

