An Introduction to Xen Project Virtualization

Lars Kurth
Community Manager, Xen Project
Chairman, Xen Project Advisory Board
Director, Open Source, Citrix

lars_kurth

www.slideshare.net/xen_com_mgr/presentations

About Me

Was a contributor to various projects

Worked in parallel computing, tools, mobile and now virtualization

Community guy for the Xen Project

Working for Citrix

Accountable to the Xen Project Community

Chairman of Xen Project Advisory Board

Why Virtualize?

Consolidation

Servers/Equipment, Cooling, Floor space

Faster provisioning

Flexibility

Less dependency on specific Hardware

Co-existing OS environments

Increased uptime

Live migration, storage migration, fault tolerance, HA

Enhanced security

Strong Isolation

Architecture provides strong isolation between VMs

Grant tables

System Partitioning

Disaggregation: sandboxing parts of the system

Fine-grain control of VM capabilities

Enables multi-layered security approach

Secure I/O

Sandboxing disk, memory, etc. drivers

New classes of threat detection

Virtual Machine Introspection, alt2pm

Consolidation

Single SoC
Maintainability, BoM

Flexibility

Less dependency on specific Hardware
Co-existing OS environments

Additional Requirements

Security requirements (same as on previous slide)
Minimal IRQ latency
Safety Certification
Low or 0 scheduling overhead
Drivers for special I/O devices

Increasing alignment of Needs

SOFTLAYER
an IBM Company
ORACLE
CLOUD
Amazon Lightsail
CITRIX
SUSE
amazon
web services
Tencent 腾讯
rackspace
the open cloud company
HUAWEI
ORACLE
Alibaba Cloud
aliyun.com
XZD
XENZYNQ
DISTRIBUTION
NXP
epam
Xenbedded
Bitdefender
ais
GlobalLogic
Bromium
DornerWorks
BAE SYSTEMS
OpenXT
QUBES OS
A REASONABLY SECURE OPERATING SYSTEM
CRUCIBLE
XILINX
ais
galois
BAE SYSTEMS
Benefits
to the entire
Community

Xen, a type-1 Hypervisor with a twist

Introduction of key concepts

Console
VM0 (or Dom0)
Toolstack
Dom0 Kernel
*Back Driver
Native Driver
VM1 (DomU1)
Applications
Guest OS
*Front Driver
VM2
Applications
Guest OS
VMn
Applications
Guest OS
Config
Scheduler
MMU
Timers
Interrupts
Xen
I/O
Memory
CPUs
HW

Virtualization Modes (x86 & ARM)

User perspective (config file)

Marketing Term	Mode	With	Disk and Network	Emulated Motherboard, Legacy Boot	Privileged Instructions, Page Tables
HVM / Fully Virtualized	HVM		Qemu	Qemu	HW
HVM + PV drivers [1]	HVM	PV Drivers	PV	Qemu	HW
PVH	PV	pvh=1	PV	PV	HW
PV	PV		PV	PV	PV
ARM	N/A		PV	PV	HW

Windows

Linux, BSDs, …

[1] When running a Linux/BSD guest on HVM with PV drivers, we also use the term PVHVM

Why is PVH & PVH Dom 0 important?

ARM	N/A	PV	PV	HW
PVH	PV	PV	PV	HW
PV	PV	PV	PV	P

This is the most complex part of Xen today!

Simplicity: Less code & fewer Interfaces in Linux/FreeBSD

Security : smaller TCB and attack surface, fewer possible exploits
Clean-up : simplify Xen-Linux kernel, Xen-Any-OS interface

Better Performance & Lower Latency for some workloads

Dom0 must be a PV guest: run Dom0 using hardware extensions
32 bit: PV guest kernels were run in ring 1, userspace in ring 3 (HW isolation)
64 bit: no ring 1 & 2 ➔ kernel & user space must share ring 3 (TLB flushes)

Server Virtualization & Cloud Computing

Recent developments

Live Patching

A tale of improved collaboration within the Xen Project Community

Why did we develop Live Patching?

Affected AWS, Rackspace, IBM SoftLayer and many others

Deploying security patches may require reboots; Inconveniences users

How did we fix this?

2015: Design with input from AWS, Alibaba, Citrix, Oracle and SUSE

Replace functions while running (old with new) in a payload

Stackable payloads can be applied and removed

2016: Xen 4.7 came with Live Patching for x86

2016: Xen 4.8 added extra x86 use-cases and ARM support

2017: XenServer 7.1 releases Live Patching in first commercial product

…

If you want to know more …

Specification & Status

xenbits.xen.org/docs/unstable/misc/livepatch.html
wiki.xenproject.org/wiki/LivePatch

Presentations, Videos, Demos

bit.do/live-patch-detailed-ppt
bit.do/live-patch-detailed-video

bit.do/live-patch-short-ppt
bit.do/live-patch-short-video

Pratap Sankar @ Flickr

Virtual Machine Introspection

A new way to protect against malware

Enablers: from xenaccess/xenprobes to LibVMI

Interesting research topic

Originally used for forensics (too intrusive for server virt)

VMI: enabling commercial applications

Hardware assisted VMI solves the intrusion problem

Collaboration between: Zentific, Citrix, BitDefender, Intel and others

Products

AIS Introvirt, BitDefender Hypervisor Introspection, Zentific Zazen

Traditional Cloud Security

Installed in-guest agents, e.g. anti-virus software, VM disk & memory scanner, network monitor, etc.

Can be disabled by rootkits and advanced persistent threats (APT)

A new model for Cloud Security?

Uses HW extensions to monitor memory (e.g. Intel EPT) ➔ Low Intrusion

Register rules with Xen to trap on and inspect suspicious activities (e.g. execution of memory on the dynamic heap)

Protection against attack techniques

All malware need an attack technique to gain a foothold
Attack techniques exploit specific software bugs/vulnerability

The number of available attack techniques is small
Buffer Overflows, Heap Sprays, Code Injection, API Hooking, …

Because VMI protects against attack techniques
It can protect against entirely new malware

Verified to block these advanced attacks in real-time
APT28, Energetic Bear, DarkHotel, Epic Turla, Regin, ZeuS, Dyreza, …
solely by relying on VMI

Protection against rootkits & APTs

Rootkits & APTs

Exploit 0-days in Operating Systems/System Software

Can disable agent based security solutions (mask their own existence)

VMI solutions operate from outside the VM

Thus, it cannot be disabled using traditional attack vectors

BUT:

VMI is not a replacement, for traditional security solutions

It is an extra tool that can be used to increase protection

If you want to know more …

Documentation

wiki.xenproject.org/wiki/Virtual_Machine_Introspection

Products

AIS Introvirt
XenServer
www.ainfosec.com

BitDefender HVI
XenServer
www.bitdefender.com

Protection & Remedial
Monitoring & Admin

Zentific Zazen (Apr 17)
Xen & XenServer & …
www.zentific.com

Protection & Remedial
Monitoring & Admin
Forensics & Data gathering
Malware analysis

Pratap Sankar @ Flickr

How secure is the Xen Project Hypervisor really?

All CVE's

2015+

Active initiatives to find bugs
XTF to help find bugs
Fuzzing of some components

Very few ARM issues

2016: 2/33
2015: 6/47
Does not use QEMU

Vulnerability data from cvdetails.com

CVE's by CVSS Severity

Average CSSV Scores

Xen: 4.7

Linux Kernel: 5.9

QEMU: 4.3

Known 0-Day Exploits

Xen: 0

Linux Kernel: 18

QEMU: 0

Low = 0.1-3.9; Medium = 4.0-6.9; High = 7.0-8.9; Critical = 9.0-10.0

Vulnerability Process Comparison

	Team	Process	Severity [1]	Type	CVEs	Days [2]	Who? [3]
					Responsible only		
Xen Hypervisor Includes Linux & QEMU vulnerabilities in supported Xen configurations	Yes	Yes	**All**	Responsible	Yes	14	**D, S, P**
OpenStack OSSA **OpenStack OSSN**	Yes Yes	Yes Yes	> Low <= Low	Responsible Full, post-fix	Yes	3-5	**D, S, P**
Linux Kernel via OSS security distros OSS security	Yes Yes	Partly [4] Yes No	> Low <= Low	Responsible Full, immediate	Yes	14-19	D [4]
QEMU via OSS security distros OSS security	Yes	Partly [4]	> Low <= Low	Responsible [5] Full, immediate	Yes	14-19	D [4]
Jailhouse	No	No					

1) Is the CVE severity used as cut-off for the process?
2) Days embargoed
3) D = Distros/Products, S = Public Service, P = Private Downstream
4) No own pre-disclosure list
5) Only handles x86 KVM bugs

XTF: Testing API behavior

In essence a unikernel per test, with fewer safeguards in place to allow for easy testing of corner cases

Also used for Vulnerability Investigation and Testing

Picture by Lars Kurth

Summary on Security

Track Record

81% of Vulnerabilities Low and Medium
Average severity of vulnerabilities getting lower

Hardening Activities

Security Audits by Cloud and Product Vendors
Testing (fuzzing, XTF, code inspection, …)

Industry Leading Vulnerability Process

Includes QEMU and Kernel XSAs
Designed with input from Cloud Providers

Isolation

Limits impact of exploits

Picture by Lars Kurth

Xen Project in Security Applications

Technology enablers: *XSM*, vTPM & TXT (TEE), *Disaggregation & Driver Domains*

Qubes OS Architecture, Qubes OS 1.0, …

2009: Project Independence (Intel / Citrix)
2010: XenClient 1.0
2013: XenClient XT
2014: Became OpenXT (BAE Systems, Assured Information Security)
2015: Support for Cell Phones, Tablets and Embedded Devices

uXen (Bromium) – Windows only, thus never made it upstream

Crucible:Defense

Disaggregation Explained

Disaggregation Explained

Driver Domain Guest OS*: Linux, BSD, MiniOS, unikernel, …

XSM/FLASK Explained

VM

Fine-grained **policy**, controlling which hypervisor functionality is accessible to this (class of) VM

Effect: limit what an exploit in this VM could do

security | config | passthrough | inter-VM communication

hypervisor | domain(self) | domain(other) | memory (grant, mmu, shadow)

Attack Surface Reduction

Similar to **L**inux **S**ecurity **M**odules/SELinux
Same policy syntax as SELinux
Different types, roles, users and attributes
Same tools for policy compilation / verification (*checkpolicy*)

Edward Snowden
@Snowden

Follow

If you're serious about security, @QubesOS is the best OS available today. It's what I use, and free. Nobody does VM isolation better.

> **Qubes OS** @QubesOS
> Qubes OS 3.2 has been released!
>
> qubes-os.org/news/2016/09/2…

RETWEETS 2,294 LIKES 3,870

2:59 PM - 29 Sep 2016

151 2.3K 3.9K

Qubes VM Manager
Run Program...
Terminal Emulator
System Tools
DisposableVM
Domain: banking
personal
qubes
untrusted
vault
work
work-web
VM: firewallvm
VM: netvm
fedora-18-x64
personal: Add more shortcuts...
personal: Files
personal: Firefox
personal: Gnote
personal: LibreOffice
personal: Terminal
personal: Thunderbird
[Dom0] Qubes VM Manager
Name
State
CPU
MEM
dom0 4 % 2423 MB
netvm 0 % 200 MB
firewallvm 0 % 570 MB
work-web 0 % 840 MB
work 0 % 570 MB
qubes 0 % 620 MB
personal 0 % 570 MB
USB Service Domain
Xen
Banking Domain
Personal Domain
Firewall VM
enforces network policies
Network Domain
Dom0
Secure UI and sysadmin domain
User defined App VMs for individual apps or groups of apps

Qubes VM Manager
Run Program...
Terminal Emulator
System Tools
DisposableVM
Domain: banking
personal
qubes
untrusted
vault
work
work-web
VM: firewallvm
VM: netvm
fedora-18-x64
personal: Add more shortcuts...
personal: Files
personal: Firefox
personal: Gnote
personal: LibreOffice
personal: Terminal
personal: Thunderbird
[Dom0] Qubes VM Manager
Name
State
CPU
MEM
dom0 4 % 2423 MB
netvm 0 % 200 MB
firewallvm 0 % 570 MB
work-web 0 % 840 MB
work 0 % 570 MB
qubes 0 % 620 MB
personal 0 % 570 MB
USB Service Domain
Xen
Banking Domain
Personal Domain
Firewall VM
enforces network policies
Tor VM
pre-configured from whonix.org
Network Domain
Dom0
Secure UI and sysadmin domain
User defined App VMs for individual apps or groups of apps

If you want to know more …

Documentation

wiki.xenproject.org/wiki/Dom0_Disaggregation
wiki.xenproject.org/wiki/Xen_Security_Modules_:_XSM-FLASK

Products & Projects

Qubes OS

www.qubes-os.org

Secure OS

OpenXT

www.openxt.org

FOSS Platform for security research, security application and embedded appliance integration building on Xen & OpenEmbedded

BAE SYSTEMS ais

Crucible:Defense

starlab.io

Xen Project based virtualization platform for technology protection, cyber-hardening, and system integrity for aerospace & defense systems

Pratap Sankar @ Flickr

Xen Project in Embedded

Vendors Active in the Community

Dornerworks
dornerworks.com/xen

Consulting
Xen Embedded Distros

Xen for Xilinx Zynq
Xen for NXP i.MX 8

ARLX Hypervisor
DO-178 (EAL6+), IEC 62304, ISO 26262
MILS EAL
FACE, VICTORY, ARINC 653

Starlab
starlab.io

Crucible and Crucible:Defense
Xen embedded hypervisor
In progress: DO-178, MILS EAL

Uses a minimal Dom0 using MiniOS, disaggregation and XSM/FLASK

AIS
ainfosec.com

BAE Systems
baesystems.com

Galois
galois.com

Maintain FreeRTOS Xen Port
Developed and maintain HalVM

Precedents of military grade certification for Xen based systems

www.slideshare.net/xen_com_mgr/art-certification & www.youtube.com/watch?v=UyW5ul_1ct0
www.linux.com/news/xen-project/2017/2/how-shrink-attack-surfaces-hypervisor

Pratap Sankar @ Flickr

Additional Requirements

Security requirements ✓
Safety certification ✓
Low or 0 scheduling overhead
Minimal IRQ latency
Drivers for special I/O devices

Schedulers & Interrupt Latency

Partitioning the System

Xen supports **several different** schedulers with different properties.

Partitioning the System

Xen supports **several different** schedulers with different properties.

Dedicated to 1 VCPU (pinning)
➔ no scheduler overheads

Xen Schedulers

Scheduler	Use-cases	Xen 4.8	Future plans
Credit	General Purpose	**Supported** **Default**	Supported Optional
Credit 2	General Purpose Optimized for lower latency, higher VM density	**Supported**	**Default**
RTDS	Soft & Firm Real-time **Multicore** Embedded, Automotive, Graphics & Gaming in the Cloud, Low Latency Workloads	Experimental Better XL support <1µs granularity	Supported Hardening Optimization
ARINC 653	Hard Real-time **Single core** Avionics, Drones, Medical	**Supported** Compile time	

Legend:
Likely in 4.9 or 4.10
Possible in 4.10

Example: Xilinx Xen Zynq

dornerworks.com/wp-content/uploads/2017/01/Xen-Zynq-Distribution-XZD-Users-Manual.pdf

IRQs: Physical follows virtual

IRQ injection

Always on the CPU running the vCPU

IRQs: Physical follows virtual

IF

vIRQ target changes or vCPU is moved

THEN

vIRQ is moved immediately

IRQs: Physical follows virtual

Xilinx ZynqMP board
(four Cortex A53 cores, GICv2)

WARM_MAX (excluding the first 3 interrupts): <2000ns

marc.info/?l=xen-devel&m=148778423725945
marc.info/?l=xen-devel&m=148839743820338

IRQs always shadow the vIRQ

➔ minimizes latency

ARM IRQs: no maintenance interrupts

IRQ received by DomU

DomU performs EOI

The guest kernel issues an "EOI" at the end of the interrupt service routine, to notify the HW that the IRQ handling is finished.

No maintenance IRQ

Additional context switch to handle EOI.

Use EOI support in HW to directly EOI the physical IRQ

PV Drivers and Protocols for various use-cases

Existing

net, block, console
keyboard, mouse, USB
framebuffer, *GPU sharing**

New

9pfs
PVcalls
multitouch, sound, display, DRM

Developing New Ones

Easy to write (GPL and BSD samples)
Kernel and User Space

*) A number of different approaches by different vendors in different market segments are being deployed, which are PV-like, but not strictly a PV protocol

Xen Project in Automotive

Vehicles are becoming the ultimate mobile device

Vendors that we know use Xen

GlobalLogic

Product: Nautilus

bit.do/gl-nautilus

First product in production expected in Q1 2018

Supports:

HW: Renesas R-Car Gen2 & Gen3, TI Jacinto6, Intel Apollo Lake, Qualcomm 410C, Sinlinx A33

Guests: Linux up to 4.9 • Android M, N, N-Car • QNX, ThreadX, FreeRTOS

PV Drivers for: GPU, Audio, HW accelerated Video codecs, DRM, …

Contributions:

27 smaller features from 2013 to 2016

EPAM

Demo

youtube.com/watch?v=jMmz1odBZb8

Interesting Features:

Container based telematics applications running in a Xen VM that can be downloaded from a cloud service

Ongoing Contributions:

ABIs for PV Sound, PV Display & PV DRM

Leading development of co-processor sharing framework

LG Electronics

Demo

bit.do/lg-xen-demo-2016

Bosch Car GmbH

Contributions

10 smaller features in 2016

Perseus

Founded by Xen maintainer

bit.do/perseus-2017

Pratap Sankar @ Flickr

GlobalLogic's Nautilus

RVC = Rear View Camera
SVS = Service Vehicle Soon
HMI = Human Machine Interface
Tell Tales = various status indicators

EPAM

AWS
Telematics Simulation Agent ver 1.0
Telematics Simulation Agent ver 2.0
Monitoring Dashboard
Driver Behavior Based Insurance Backend
Dom0 - Control
Dom0 Services
Minimal rootfs
Linux Kernel w/o HW Drivers
DomD – HW Drivers & Cluster
Cluster Simulation App
Wayland/Weston
Wayland BE (Events/Display)
OpenGL ES
ALSA w PV_ALSAS_BE
Linux Kernel with GPU and other HW Drivers
DomU Fusion
Telematics simulation Agent (Acceleration, Braking, Corning, GPS)
Containers
Container mgmt tool
Minimal rootfs with systems library
Linux Kernel w/o HW Drivers
DomU – Linux IVI
IVI Simulation App
MW Frameworks
PV DISPLAY
PV EVENTS
PV SOUND
Linux Kernel with GPU and w/o other HW Drivers
TrustZone
Trusted Apps
Hypervisor
OP-TEE OS
R-Car H3 Platform
TZ monitor

A diverse, vibrant and growing community

Hypervisor Git Commits

Stats are impacted by release model (code freezes) and our transition to 2 releases per year

2015: Hypervisor Stack Top Players

Top:

Citrix	48%
Suse	17%
Oracle	6%
Intel	6%
Red Hat	4%
Linaro	3%
FreeBSD	2%
Star Lab	1%
Other	13%

Others:

Fujitsu

Invisible Things Lab

BitDefender

Huawei

Zentific

Verizon

Cavium

GlobalLogic

NSA

…

2016: Hypervisor Stack Top Players

Top:

Citrix	39%
Suse	22%
Oracle	7%
ARM	5%
Red Hat	4%
Linaro	3%
Intel	3%
Star Lab	2%
BSD	2%
Fujitsu	2%
Bitdefender	2%
Zentific	1%
NSA	1%
Zentific	1%
Qualcomm	1%
Huawei	1%
Other	6%

First-time contributors in 2016:

ARM

Aporeto

Bosch Car Multimedia Gmbh

Netflix

Qualcomm

Xilinx

Why should I use Xen?

Picture by Lars Kurth

Extremely Flexible and Versatile

Proven in different markets
Easy to port to new environments
Easy to develop new PV drivers
Highly customizable

Security and Resilience

Isolation, Partitioning, Security Features
Track record in handling security issues

Safety

Examples of Military Grade Certification

Vibrant and Diverse Community

Covering Server, Cloud, Security, Embedded, Automotive

Picture by Lars Kurth

More Resources

Developer Portal: bit.do/xen-devs
Xen on ARM whitepaper: bit.do/xenarm-white
Xen on ARM wiki: bit.do/xenarm-wiki

Port Xen to a new SOC: bit.do/xenarm-porting
Add Xen support Xen to your OS: bit.do/xenarm-os

Device Passthrough presentation: bit.do/xenarm-pt
OE meta-virtualization Xen recipe: bit.do/xenmeta
OpenXT (Xen + OpenEmbedded): openxt.org
Xenbedded presentation: bit.do/xenbedded

Monthly ARM Community Call: bit.do/xenarm-call

Engage!

Lists and IRC on freenode:
xen-devel@lists.xenproject.org
xen-users@lists.xenproject.org
#xenarm or #xen-devel

Internships in 2017:
Google Summer of Code (Students)
Outreachy (Women and other groups)
bit.do/xen-internships

Xen Project Developer and Design Summit:
July 11-13, Budapest, Hungary
CfP: open until April 14
bit.do/xensummit17

Questions

www.slideshare.net/xen_com_mgr/presentations

Picture by Lars Kurth