

An abstract geometric design featuring three blue circles of varying sizes, each composed of concentric rings of different shades of blue. Two thin blue lines intersect at a point, forming a V-shape. A large, light gray vertical rectangle is positioned on the right side of the page. The circles and lines are arranged in a way that suggests a sense of depth and movement.

NewBluePill 探索之路

架构说明及注释详解

So many open source projects. Why not Open
your Documents?

Superymk
2009-5-1

参与人员：

| | |
|----------|--|
| 作者 | 联络 |
| Superymk | Superymkfounder@hotmail.com |

文中所提到的代码源程序来自 <http://www.bluepillproject.org/>，作者 Invisible Things Lab，读者可下载 nbp-0.32-public.zip 文件，其中代码即为我所分析的主要代码。

读者可结合此书阅读源代码，再结合书中实验以及书后的三个实验，从而获得更深刻的领悟。

由于作者水平有限，书中出现错误在所难免，望批评指正。

发布记录

| 版本 | 日期 | 作者 | 说明 |
|----|----|----|----|
| | | | |
| | | | |
| | | | |

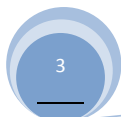
版权说明

本文档版权归原作者所有。

在免费、且无任何附加条件的前提下，可在网络媒体中自由传播。

如需部分或者全文引用，请事先征求作者意见。

如果本文对您有些许帮助，表达谢意的最好方式，是将您发现的问题和文档改进意见及时反馈给作者。当然，倘若有时间和精力，能为技术群体贡献自己的所学为最好的回馈。



目录

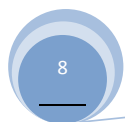
| | |
|---|----|
| 前言 | 8 |
| 鸣谢 | 9 |
| 关于作者..... | 10 |
| 本书简介..... | 11 |
| 一、 概述..... | 13 |
| Hypervisor 概述..... | 13 |
| 虚拟化的历史..... | 13 |
| 硬件虚拟化技术（HEV）..... | 13 |
| HEV 技术最佳实践..... | 14 |
| 已有的 HEV 技术平台介绍..... | 15 |
| SVM..... | 15 |
| 概述..... | 15 |
| 新的地址翻译机制..... | 16 |
| 相关结构和汇编指令..... | 17 |
| 基于 SVM 的 Hypervisor 开发逻辑..... | 18 |
| Intel-VTx..... | 18 |
| 概述..... | 18 |
| 新的地址翻译机制..... | 19 |
| 相关结构和汇编指令..... | 20 |
| 基于 VT 的 Hypervisor 开发逻辑..... | 22 |
| Intel-VTd(如果时间充裕则写)..... | 22 |
| NewBluePill 项目介绍..... | 22 |
| PART1 HEV 技术相关知识..... | 24 |
| 二、 深入 HEV 技术细节..... | 24 |
| HEV 下虚拟机启动过程..... | 24 |
| 启动过程模型..... | 24 |
| VT 技术下开启虚拟机的过程..... | 27 |
| SVM 技术下开启虚拟机的过程..... | 28 |
| HEV 下虚拟机关键结构体..... | 28 |
| VT 技术下的 VMCS 结构体..... | 29 |
| 虚拟机状态保存区（Guest-State Area）..... | 30 |
| 宿主机（Hypervisor）状态保存区（Host-State Area）..... | 32 |
| 虚拟机运行控制域（VM-Execution Control Fields）..... | 32 |
| VMEntry 行为控制域（VM-Entry Control Fields）..... | 36 |
| VMEXIT 行为控制域（VM-Exit Control Fields）..... | 37 |
| VMEXIT 相关信息域（VM Exit Information Fields）..... | 38 |
| SVM 技术下的 VMCB 结构体..... | 38 |
| HEV 下虚拟机关闭过程..... | 38 |
| VT 技术下关闭 Hypervisor 和虚拟机的过程..... | 39 |
| SVM 技术下关闭 Hypervisor 和虚拟机的过程..... | 39 |
| 总结..... | 39 |
| PART2 深入研究 NewBluePill..... | 43 |

| | | |
|----|------------------------------------|----|
| 三、 | 体验 NewBluePill | 43 |
| | 编译 NewBluePill | 43 |
| | 演示 NewBluePill | 45 |
| | 调试 NewBluePill | 48 |
| 四、 | NewBluePill 的启动和卸载 | 51 |
| | NewBluePill 驱动的启动过程 | 51 |
| | 构建私有页表 | 51 |
| | 初始化调试系统 | 52 |
| | 构建 Hypervisor 并将操作系统放入虚拟机 | 52 |
| | 进入 NewBluePill 的世界 | 53 |
| | 阶段 1 初始化 | 53 |
| | 阶段 2 初始化 | 62 |
| | NewBluePill 驱动的卸载过程 | 63 |
| | 拆除 Hypervisor 恢复原宿主机信息 | 64 |
| | SVM 技术下 NewBluePill 的卸载实现 | 65 |
| | VT 技术下 NewBluePill 的卸载实现 | 65 |
| | 关闭调试系统 | 65 |
| | 拆除私有页表 | 65 |
| 五、 | NewBluePill 内存系统 | 67 |
| | 1) 相关文件: | 67 |
| | 2) 技术背景: | 67 |
| | 3) 总体功能介绍: | 70 |
| | 4) 实现过程: | 70 |
| | MmInitManager() 方法 | 70 |
| | MmSavePage () 方法 | 71 |
| | MmSavePage () 方法 | 73 |
| | MmSavePage () 方法 | 73 |
| 六、 | NewBluePill 陷入事件管理系统 | 74 |
| | 相关文件 | 74 |
| | Trap 元素的生成、注册机制 | 74 |
| | Trap 元素的触发机制 | 75 |
| | 阶段 1 触发 | 76 |
| | 阶段 2 触发 | 76 |
| | SVM 技术下的阶段 2 触发 | 77 |
| | VT 技术下的阶段 2 触发 | 77 |
| | 各处理函数功能和实现 | 77 |
| | SVM 技术实现中各处理函数功能和流程 | 78 |
| | VT 技术实现中各处理函数功能和流程 | 78 |
| | VmxDspatchVmInstrDummy () 函数 | 78 |
| | VmxDspatchCpuid () 函数 | 78 |
| | VmxDspatchINVD () 函数 | 79 |
| | VmxDspatchMsrRead () 函数 | 79 |
| | VmxDspatchMsrWrite () 函数 | 80 |
| | VmxDspatchCrAccess () 函数 | 80 |

| | |
|--------------------------------|-----|
| VmxDspatchException() 函数 | 81 |
| VmxDspatchRdtsc() 函数 | 82 |
| VmxDspatchIoAccess() 函数 | 82 |
| 七、 NewBluePill 反探测系统 | 84 |
| 探测 NewBluePill | 84 |
| 通过指令执行耗时分析 | 84 |
| 通过观察 TLB 变化分析 | 85 |
| Blue Chicken 策略 | 85 |
| 相关文件 | 85 |
| 功能介绍和详细分析 | 85 |
| 时间欺骗——指令追踪策略 | 86 |
| 相关文件 | 86 |
| 功能介绍和详细分析 | 86 |
| 八、 NewBluePill 调试系统 | 89 |
| 相关文件 | 89 |
| 功能概述 | 89 |
| 实现细节 | 90 |
| NewBluePill 端调试系统部分 | 90 |
| DbgClient 端调试系统部分 | 91 |
| 总结 | 93 |
| PART3 实验部分 | 95 |
| 九、 动手写自己的第一个 HVM 程序 | 95 |
| 实验目的 | 95 |
| 实验概述 | 95 |
| 实验过程 | 95 |
| 十、 移植 NBP 到 32 位系统 | 96 |
| 十一、 开发基于 HEV 技术的注册码验证器 | 97 |
| 实验目的 | 97 |
| 实验概述 | 97 |
| 实验过程 | 98 |
| A. 其它有关 HVM 技术的项目 | 99 |
| B. 其它安全技术 | 100 |
| C. 相关软件和参考文档 | 102 |



前言



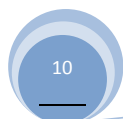
鸣谢

批注 [S1]: 完成本书后进行修改

特别鸣谢：戚正伟老师和徐昊所给的指导帮助和对本书的审核工作，如果没有他们的帮助，我无法在短时间内明白 NewBluePill 的核心思想，并如此深入的探索研究虚拟化技术。

鸣谢：上海交大虚拟机项目组在编写本书期间所给的鼓励和支持。

关于作者



本书简介

好的数据结构是程序的灵魂，一个定义优秀的数据结构能够使人立刻读懂其中的算法。所以我们更着重于介绍其中所涉及的数据结构，力求用最短的篇幅说明 NewBluePill 的世界。

~~本书假定读者仅理解最基本的术语，不具有嵌入式开发经验，当然如果有相关经验的话就可以更快的理解书中内容。~~

~~对于没有太多基础的读者，建议先看第三章一些背景知识，然后再看第二章和后续章节。~~
本书假定读者具有一定嵌入式软件开发经验，对于没有太多基础的读者，建议先阅读 *Windows Internals, 4th Edition* 和一些 x86 平台相关开发书籍。

写这本书的目的只是去引导读者阅读代码，因此并不具体到代码中涉及的每个角落。

本书涉及的表示方法：

| | |
|---|----------|
| <i>Windows Internals, 4th Edition</i> | 书名 |
| Hypervisor | 英文名词 |
| http://www.bluepillproject.org/ | 超级链接 |
| HvmSubvertCpu() | 书中代码及函数名 |

(This page is intended to be blank)

一、概述

在这一章中，我们先介绍一些贯穿全书的概念，比如 Hypervisor, VT-x, VT-d, SVM 等等。然后我们会简略介绍下 NewBluePill 项目背景及其所采用的硬件虚拟化技术。

这一章只是介绍这些技术大致的轮廓，详细内容会在后面各章节中逐一介绍。

Hypervisor 概述

虚拟化的历史

在讨论 Hypervisor 之前首先谈谈虚拟，虚拟 (virtualization) 指对计算机资源的抽象，一种常用的定义是“虚拟就是这样的一种技术，它隐藏掉了系统，应用和终端用户赖以交互的计算机资源的物理性的一面，最常做的方法就是把单一的物理资源转化为多个逻辑资源，当然也可以把多个物理资源转化为一个逻辑资源（这在存储设备和服务器上很常见）”

实际上，虚拟技术早在 20 世纪 60 年代就已出现，最早由 IBM 提出，并且应用于计算技术的许多领域，模拟的对象也多种多样，从整台主机到一个组件，其实打印机就可以看成是一直在使用虚拟化技术的，总是有一个打印机守护进程运行在系统中，在操作系统看来，它就是一个虚拟的打印机，任何打印任务都是与它交互，而只有这个进程才知道如何与真正的物理打印机正确通信，并进行正确的打印管理，保证每个 job 按序完成。

长久以来，用户常见的都是进程虚拟机，也就是作为已有操作系统的一个进程，完全通过软件的手段去模拟硬件，软件再翻译内存地址的方法实现物理机器的模拟，比如较老版本的 VMWare, VirtualPC 软件都属于这种。

在 2005 年和 2006 年，Intel 和 AMD 都开发出了支持硬件虚拟技术的 CPU，也就是在这时，x86 平台才真正有可能实现完全虚拟化¹。在 2007 年初的时候，Intel 还进一步的发布了 VT-d 技术规范，从而在硬件上支持 I/O 操作的虚拟化。随着硬件虚拟化技术越来越广泛的采用，开发者也开始虚拟技术来做一些其他的事情：当前 HVM 已经在虚拟机，安全，加密等领域上有所应用，例如 VMware Fusion, Parallels Desktop for Mac, Parallels Workstation 和 DNGuard HVM，随着虚拟化办公和应用的兴起，相信虚拟化技术也会在未来得到不断发展。

硬件虚拟化技术 (HEV)

有了虚拟技术的基本概念，下面我们谈谈硬件虚拟化技术。硬件虚拟化技术 (Hardware Enabled Virtualization, 本书中简称 HEV)，也就是在硬件层面上，更确切的说是在 CPU 里 (VT-d 技术是在主板上北桥芯片支持)，对虚拟技术提供直接支持，并通过这种设计提高虚拟效率、降低开发难度。在硬件虚拟化技术诞生前，编写虚拟机过程中，为了实现多个虚拟机上的真实物理地址隔离，需要编程实现把客户机的物理地址翻译为真实机器的物理地址。同时也需要给不同的客户机操作系统编写不同的虚拟设备驱动程序，使之能够共享同一真实硬件资源。硬件虚拟化技术则在硬件上实现了内存地址甚至于 I/O 设备的映射，因此大大简化了编写虚拟机的过程。而其硬件直接支持二次寻址和 I/O 映射的特性也提升了虚拟机在运行时的

¹ 完全虚拟化(Full Virtualization), 完整虚拟底层硬件，这就使得能运行在该底层硬件上的所有操作系统和它的应用程序，也都能运行在这个虚拟机上。

性能。¹

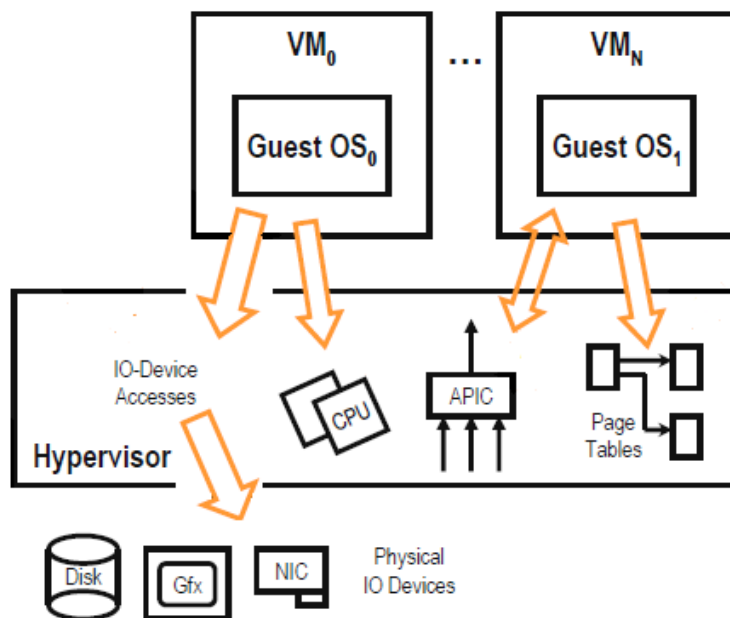
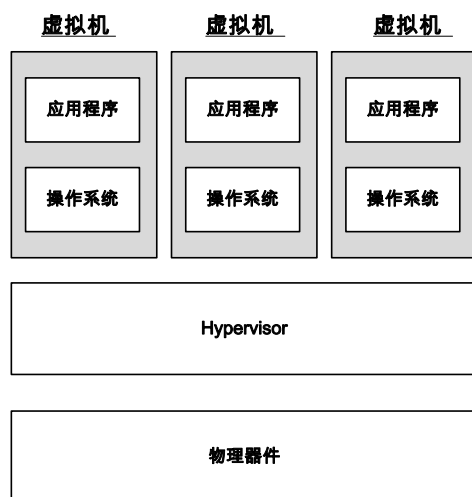


图 1.1 硬件虚拟化技术架构示意图

在硬件虚拟化技术中，一个重要的概念就是 HVM。HVM, Hypervisor Virtual Machine 的缩写（在本书中简称为 Hypervisor），是在使用硬件虚拟化技术时创建出来的特权层，该层提供给虚拟机开发者，用来实现虚拟硬件与真实硬件的通信和一些事件处理操作（如图 1.1），因此 Hypervisor 的权限级别要高于等于操作系统权限。

HEV 技术最佳实践



¹ 一些优化技术也在硬件中被采用，比如专门用于二次寻址的 TLB，详细信息可以参考 Intel 和 AMD 的手册

图 1.2 Hypervisor 使用架构图

前文中已经提过,Hypervisor 层的权限要高于等于操作系统的权限。操作系统的内核态已经处在了 Ring0 特权级上,因此 Hypervisor 层实际上要运行在一个新的特权级别上,我们称之为“Ring -1”特权级。同时需要新的指令,寄存器以及标志位去实现这个新增特权级的功能。

作为一种最佳实践方案,一般 Hypervisor 层的实现都是越简单越好。一方面,简单的实现能够尽量降低花在 Hypervisor 上的开销¹,毕竟大多数这些开销在原先的操作系统上是不存在的。另一方面,复杂的程序实现容易引入程序漏洞,Hypervisor 也是如此,且一旦 Hypervisor 中的漏洞被恶意使用,由于其所处特权级高于操作系统,将使隐藏在其中的病毒、恶意程序很难被查出。

批注 [S2]: 后面要描述 Hypervisor 的开销问题

Virtualization.pdf 10

已有的 HEV 技术平台介绍

现今两大主要硬件厂商 Intel 和 AMD 均以推出了支持硬件虚拟化技术的产品,两者大体功能和实现方法近似(意料之中,因为两家公司在过去你死我活的市场拼斗中,每次也都是实现功能和方法类似,只不过名字不同罢了)。下面我们简略介绍下这两家公司的支持 HEV 技术的平台,读者可以首先对这两种平台有概念。后面的章节中会有对这两个平台技术细节的更详细描述。

SVM

概述

AMD 芯片支持硬件虚拟化的技术被称作 AMD-V (在技术文档中也被称为 SVM,其全称是 AMD Secure Virtual Machine,在本书中我们仍称其为 SVM)。其主要是通过一组能够影响到 Hypervisor 和 Guest Machine (客户机,下文简称 Guest) 的中断实现的。AMD-V 技术设计目标如下: □

- 引入客户机模式 (Guest Mode) ²
- Hypervisor 和 Guest 之间的快速切换
- 中断 Guest 中特定的指令或事件(events)
- DMA 外部访存的保护
- 中断处理上的辅助并对虚拟中断 (virtual interrupt) 提供支持
- 新的嵌套页表用来实现地址翻译
- 一个新的 TLB (其实就是一个 Cache) 来减少虚拟化造成的性能下降。
- 对系统安全的支持

新的客户机模式 通过 VMRUN 指令即可进入这种新的处理器³模式,当进入客户机模式

¹ 关于 Hypervisor 的开销问题,后面的章节会有介绍

² x86 上原有处理器模式包括保护模式(Protected Mode),管理模式(SMM),实模式(Real Mode)

³ 注意:本书中强调处理器 (Processor) 和 CPU 的差别,文中若无特别说明,处理器指逻辑处理器 (Logical Processor)。在多核时代,一个 CPU 上可能会有多个核,而在操作系统视角中,一个核才是一个逻辑处理器,因此通过操作系统查看的逻辑处理器数量往往大于真实 CPU 的数量,并且逻辑处理器才是能够运行 Hypervisor 的基础。

后，为了辅助虚拟化过程，一些 x86 汇编指令的语义会发生变化。

外部访问保护 过去客户机（Guest）可以直接访问选定的 I/O 设备。现在硬件上已经实现这样的安全功能，能够阻止某个虚拟机拥有的某个设备访问其它虚拟机的内存。

中断上的支持 为了辅助中断的虚拟化，下列各项现在已经得到硬件支持，并且可以通过配置 VMCB 结构体¹的方法使用

- 1) 拦截物理中断分发（Intercepting physical interrupt delivery）发生在物理硬件上的中断能够让虚拟机发生一个中断，陷入 Hypervisor，从而使得 Hypervisor 可以首先处理这个中断。
- 2) 虚中断（Virtual Interrupts）Hypervisor 能够将为提供给客户机（Guest）一套虚拟的中断机制。它是这样实现的，Hypervisor 会给这个客户机复制出来一份 EFLAGS.IF 用做中断屏蔽位（Interrupt Mask Bit），同时复制 APIC²中的中断优先级寄存器提供给客户机，从而客户机就会去操纵这套假的中断机制，而不是直接去操纵物理中断。
- 3) 共享物理 APIC AMD 的 SVM 技术能够允许多个 Guest 共享同一物理 APIC，同时又能保护这个 APIC 以免某个客户机不慎或恶意的在未经其它客户机许可的情况下，将可接收中断优先级设置为高优先级，从而清空了所有其它 Guest 的中断。

被标记的 TLB（Tagged TLB） 为了降低 Guest 模式和 VMM 模式切换开销，，TLB 上新加了一个 ASID 标记（Address Space Identifier），这个标记可以区分 TLB 上的一块地址是 Hypervisor 范围内的地址还是 Guest 的地址，从而加速了地址翻译。

安全方面的支持 现在提供的安全方面的支持主要是利用和 TPM 模块（Trusted Platform Module）³的交互，基于与安全 Hash 值的比较。

批注 [S3]: 第 17 章 其它安全技术写些有关 TPM 技术的介绍

新的地址翻译机制

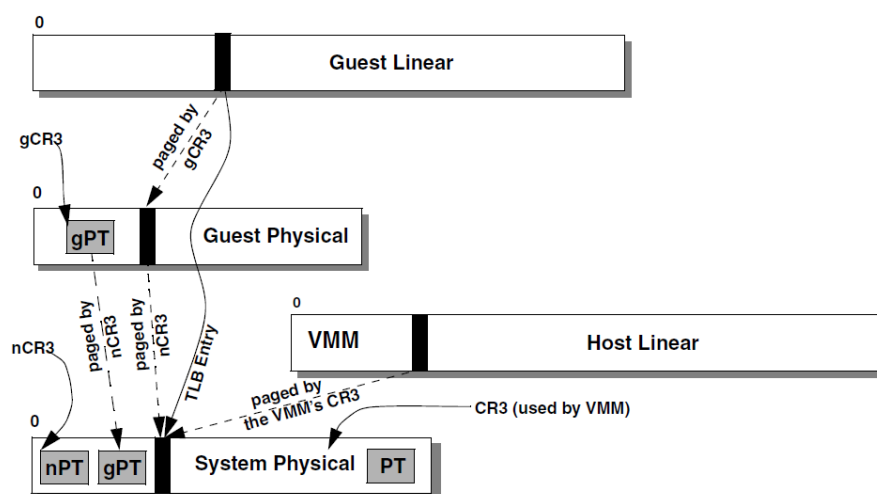
AMD 引入了新的地址翻译技术——嵌套页表翻译（Nested Page Table, NPT），用于支持两级地址翻译，这样就使得虚拟机管理器不用再自己软件维护一套影子页表⁴

¹ VMCB 结构体, Virtual Machine Control Block, 也称 VMCB 控制块, Intel 的相应结构体名称为 Virtual Machine Control Sector, VMCS。这个控制块用于通知物理 Processor 要拦截的事件，以及在进出 Hypervisor 上下文切换时保存 Hypervisor 和 Guest 的各项寄存器，后面的章节中会有对这个结构体的详细介绍

² APIC, Advanced Programmable Interrupt Controller, 高级可编程中断控制器，第三章有关于此主题内容。

³ TPM 技术会在“第 17 章 其它安全技术”一章中介绍。

⁴ 影子页表（Shadow Page Tables），常见于过去传统的虚拟机管理器中，由于存在从 Guest 线性地址到 Guest 物理地址，从 Guest 物理地址到真实物理地址两层地址翻译，所以在过去一般是要虚拟机软件自己维护两套页表去做这样的地址翻译。

图 1.3 嵌套页表翻译地址过程¹

嵌套页表翻译地址的过程如图 1.3 所示，这套机制的实现允许了从 Guest 线性地址到真实物理地址的翻译，也允许了在 Hypervisor 范围内的 Host 线性地址到真实物理地址的翻译。同时专门附加了一个 TLB 寄存器，用于缓存从 Guest 线性地址到真实物理地址的映射，从而提升了虚拟机的运行性能。

Note 关于嵌套页表技术的详细解释会在“第四章 深入 HEV 技术细节”一章中介绍，完整的描述请参考 *AMD64 Architecture Programmer's Manual, Volume 2: System Programming*

批注 [S4]: 记得所有书名用斜体

相关结构和汇编指令

在 SVM 中，VM 控制块被称为 VMCB (Virtual Machine Control Blocks)，其信息主要分为两块，第一块是控制信息存储部分，同时也包含是否允许拦截某特定异常的遮罩 (interception enable mask)，Guest 中不同的指令和事件都能以修改 VMCB 中相应控制位的方法拦截，SVM 支持的两类主要的拦截是异常拦截和指令拦截，第二部分则是 guest 的状态信息保存，这里会保存段寄存器以及大部分的虚拟内存的入口控制寄存器，不过浮点寄存器信息不会被保存。需要注意的是 VMCB 在不同的处理器间不共享，并且 VMCB 一定要保证是在 4K 页对齐的连续物理内存空间中。

SVM 中主要的指令有以下这些：

- **VMLoad** 从 VMCB 加载 guest 的状态，VMCB 与 guest 是有对应关系的。
- **VMMCALL** 通过该方法 guest 可以与 VMM 显式的交流，方法是利用生成 #VMEXIT 从 guest 层退到 VM 层。
- **VMRUN** 加载 VMCB，并开始执行 guest 层的指令，VMCB 的物理地址将通过 RAX 获得，这个 VMCB 对应于要执行的 guest
- **VMSAVE** 存储处理器状态的子集到 VMCB 中，这个 VMCB 的物理地址由 RAX 寄存

¹ 此图摘自 *AMD64 Architecture Programmer's Manual, Volume 2: System Programming*

器给出。

- STGI 用于设置全局中断标志（Global Interruption Flag）为1，这个指令属于Secure Virtual Machine。
- CLGI 用于设置全局中断标志（Global Interruption Flag）为0，同样这个指令属于Secure Virtual Machine。
- INVLPGA 使得TLB上一个ASID和一个虚拟页(Virtual Page)之间的映射关系无效，这个指令属于Secure Virtual Machine。
- SKINIT 安全的重新初始化CPU，使得CPU可以开始执行一段受信任的程序（trusted software）其方法是将该代码进行安全的哈希比较(secure hash comparison)。这也就是开发者可以开发一个更安全的VMM loader。这种安全手段可以在TPM的帮助下发挥更大作用
- 改进的MOV指令 现在的MOV指令可以直接读写CR8寄存器（任务优先级寄存器 Task Priority Register），因此可以用来提高SVM应用的性能。

基于 SVM 的 Hypervisor 开发逻辑

其实由上文的描述可以看出，开发基于SVM的Hypervisor最主要是编写一个循环，这个循环要包含VMRUN命令以便从VM层启动一个Guest虚拟机，也要包含一段程序用于处理当#VMEXIT发生后的异常情况，这其中可能要手动做一些必要的保存现场和恢复现场的工作，具体造成异常的起因等均可通过读取VMCB中的数据获得。不过SVM没有提供一个显示终止Hypervisor的指令，因此若有需要，则要用其它方法关闭Hypervisor。NewBluePill中对SVM的支持就是这样实现的，我们会在深入探究NewBluePill的章节中详细展示如何使用这些指令。

批注 [55]: 后面要介绍 NBP 中关于 SVM 技术的运用

Intel-VT_x

概述

Intel 芯片支持硬件虚拟化的技术被称为 Intel VT 技术（Intel® Virtualization Technology）。与 SVM 一样，其主要也是通过一组能够影响到 Hypervisor 和 Guest Machine 的中断实现的。

在 VT 技术中，与 SVM 类似的，设计架构上同样存在两种角色——虚拟机管理器（Virtual Machine Monitors, VMM）和客户机（Guest），两者分处在 VMX root 模式和 VMX non-root 两种模式下。VT 技术的设计目标是：

对于 VMM 层：（进入此层则代表进入了 VMX root 模式）

- 为每个虚拟机提供虚拟处理器，并且可以在恰当的时候把它放在真正的物理处理器上，从而使得这个虚拟处理器可以处理指令。
- VMM 层可以控制处理器资源，物理内存，管理中断和 I/O 操作

对于 Guest Machine：（进入此层则代表进入了 VMX non-root 模式）

- 每个虚拟机使用相同的接口来使用虚拟处理器，内存，存储设备等资源
- 每个虚拟机可以独立的不受干扰的运行，虚拟机间都是相互独立的
- 对于虚拟机来说，VMM 层像是完全透明的。

在 VT 技术下的 Hypervisor 生命周期如图 1.4 所示：

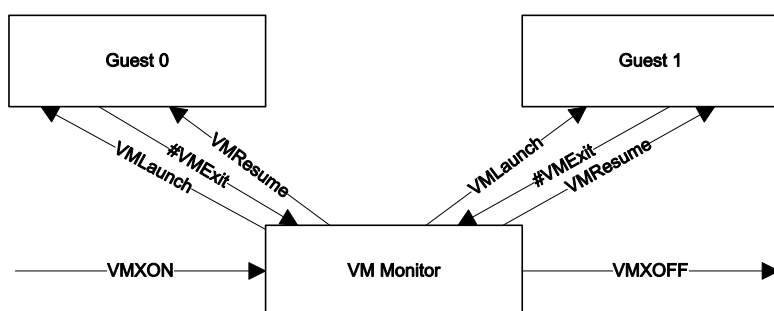


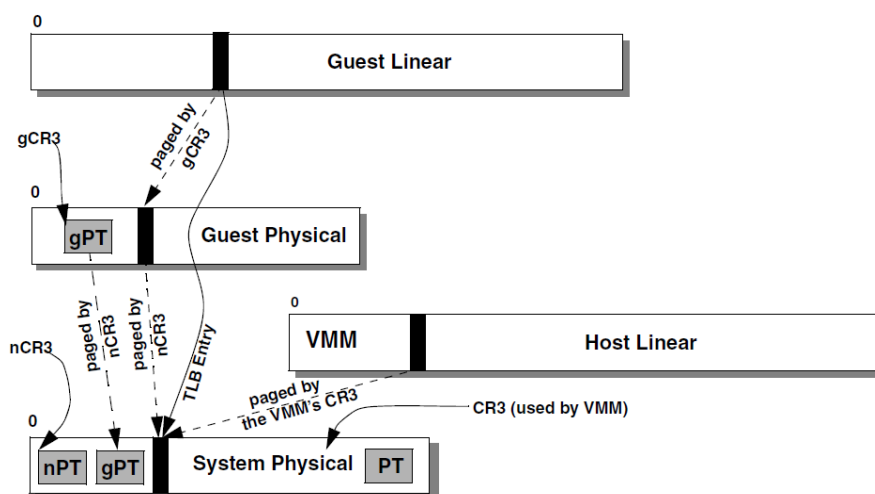
图 1.4 VT 技术中 Hypervisor 的生命周期

图示表明，软件通过执行 VMXON 指令进入 VMX Root 模式下，开启了虚拟机管理器的运行环境。然后通过使用 VMLaunch 指令使得目标系统正式运行在虚拟机中。当某条指令产生了 #VMEXIT 事件后，会陷入虚拟机管理器中，待其处理完这个事件，可以通过 VMXResume 指令将控制权移交回发生 #VMEXIT 事件的虚拟机。直到某个时刻，在 Hypervisor 中显示的调用了 VMXOFF 指令，Hypervisor 才会被关闭。

批注 [S6]: 仔细看完 EPT 技术后补充这一部分

新的地址翻译机制

Intel 同样引入了新的地址翻译技术——扩展页表翻译（Extended Page Table，EPT），用于支持两级地址翻译。

图 1.4 嵌套页表翻译地址过程¹

嵌套页表翻译地址的过程如图 1.3 所示，这套机制的实现允许了从 Guest 线性地址到真实物理地址的翻译，也允许了在 Hypervisor 范围内的 Host 线性地址到真实物理地址的翻译。同时专门附加了一个 TLB 寄存器，用于缓存从 Guest 线性地址到真实物理地址的映射，从而提升了虚拟机的运行性能。

¹ 此图摘自 AMD64 Architecture Programmer's Manual, Volume 2: System Programming

Note 关于扩展页表技术的详细解释会在“第四章 深入 HEV 技术细节”一章中介绍，完整的描述请参考 Intel 64 and IA-32 Architectures Software Developer's Manual Volume 3B, Chapter 24. Support for Address Translation

相关结构和汇编指令

在 VT 技术中，VM 控制块被称为 VMCS (Virtual Machine Control Structure)。VMCS 包括三个组成部分：

表 1.1 VMCS 区域的组成部分

| Byte 偏移量 | 内容 |
|----------|--|
| 0 | VMCS 版本标志 (Revision Identifier) |
| 4 | VMX 退出原因指示器 (VMX-abort indicator) ¹ |
| 8 | VMCS 数据区 |

批注 [S7]: 后文会详细介绍 VMX Abort

如表 1.1 所示，VMCS 区域的前四个字节用于 VMCS 版本标志，不同的 VMCS 格式对应的版本号也不同，而这个物理处理器可以加载的 VMCS 结构体的版本号会存储在 MSR 寄存器中，因此这样的设计也就给未来发展留下了空间。

在 VMCS 数据区中，主要有如下几个组成部分：

表 2.2 VMCS 数据区主要组成部分^①

| 名称 | 作用 |
|---|--|
| 虚拟机状态保存区 (Guest State Area) | 当发生了 #VMEXIT 事件时虚拟机当前状态保存于此，在重新进入虚拟机的时候再利用此处的数据恢复虚拟机的状态 |
| 宿主机状态保存区 (Host State Area) | 当发生了 #VMEXIT 事件时宿主机的状态利用此处数据恢复 |
| 虚拟机运行控制域 (VM Execution Control Fields) | 此处数据定义了虚拟机在什么情况下发生 #VMEXIT 事件，对 VMX non-root 模式有影响 |
| VMEXIT 行为控制域 (VM Exit Control Fields) | 此处数据定义了当 #VMEXIT 事件发生时要做的附加工作 (比如保存调试寄存器，加载全局性能控制寄存器等这些工作) |
| VMEntry 行为控制域 (VM Entry Control Fields) | 此处数据定义了当发生 #VMEntry 事件时 (通常是因为调用了 VMResume 汇编指令) 要做的附加工作。 |
| VMEXIT 相关信息域 (VM Exit Information Fields) | 此处数据在发生 #VMEXIT 事件时自动记录了发生原因和该事件的具体种类。这个域是只读的 |

VMX Abort 和 VMCS 数据区结构和用法会在后续章节中详细介绍。

VT 技术在设计时注明，没有任何标志位用于区分一个逻辑处理器 (Logical Processor)

当前正在执行 VMX root 模式下的指令还是执行的 VMX non-root 模式下的指令，这就确保了 Hypervisor 对虚拟机完全透明——因为虚拟机无从判断它当前是否运行在一个虚拟机下。最后要注意的是 VMCS 同样要求保证是在 4K 页对齐的连续物理内存空间中。

批注 [S8]: VMX Abort 和 VMCS 数据区结构和用法会在后续章节中详细介绍

批注 [S9]: 也许你可以考虑添加一部分内容，用于描述如何检测硬件虚拟机的存在。

¹ 如果在 VM Exit 的时候遇到问题，就会发生 VMX Abort，一旦发生，那么这个逻辑处理器会进入关闭状态 (Shutdown State)

VT 中主要的指令有以下这些:

维护VMCS结构体的指令

- **VMPTRLD** 参数为VMCS块的物理地址。该指令用来激活一块VMCS。修改该处理器的当前VMCS指针（Current-VMCS Pointer）指向传入的VMCS物理地址，并且激活该VMCS，如果要维护一块VMCS则必须先激活该VMCS。（否则不能用这些指令来维护）
- **VMPTRST** 用来存储当前VMCS指针（VMCS块物理地址）到指定位置。
- **VMCLEAR** 该指令用来使一块VMCS变为不活跃状态。该指令将标记为已启动状态（Launch State）的VMCS设置为不活跃状态（Inactive State/Clear State）并且更新该VMCS块所有区域信息并确保写入VMCS块内存中（这也就把对应虚拟机和Hypervisor的最新信息同时写入到VMCS块中），如果带操作的VMCS块就是当前VMCS指针所指向的VMCS块，那么该指针会被设置为无效地址
- **VMREAD** 通过指定的VMCS Encoding从当前VMCS块中读取一个参数。
- **VMWRITE** 通过指定的VMCS Encoding从当前VMCS块中写入一个参数。

与虚拟机管理器有关的指令

- **VMCALL** 这条指令用于Guest和Hypervisor进行通信。执行该汇编指令会产生一个#VMEXIT事件，从而使得可以陷入Hypervisor中。
- **VMLAUNCH** 这条指令用于启动当前VMCS指针所指的一个虚拟机，并且移交控制权给Guest。
- **VMRESUME** 这条指令用于从Hypervisor中恢复虚拟机的执行，并且移交控制权给Guest。
- **VMXOFF** 这条指令用于关闭Hypervisor。在下次执行VMXON开启Hypervisor前不得执行虚拟机相关汇编指令。
- **VMXON** 这条指令用于处理器进入VMX模式下，执行该指令后也就可以运行Hypervisor。传入的参数必须是4K页对齐的物理地址，这段内存用于支持后续VMX相关的操作。

VMLAUNCH 和 VMRESUME 指令的异同

VMLAUNCH 和 VMRESUME 命令都是将控制权移交到虚拟机，那么两者的区别呢？

两者运行的时机不同！

1. VMLAUNCH 指令会检查当前 VMCS 的启动状态是不是不活跃状态（相应标记位清空）。成功运行结果是该 VMCS 被标记为已启动状态。
2. VMRESUME 指令会检查当前 VMCS 的启动状态是不是已启动状态

所以，必须利用 VMLAUNCH 指令启动一个虚拟机。以后的某个时候，因为 VMEXIT 事件而陷入 Hypervisor 中，这个时候要恢复虚拟机的运行则要利用 VMRESUME 指令，正如图 1.4 所示。

管理VT相关的TLB的控制指令

- **INVEPT** 这条指令用于EPT地址翻译中，使TLB中缓存的地址映射失效
- **INVVPID** 这条指令用于在TLB中使某个VPID对应的地址映射失效

基于 VT 的 Hypervisor 开发逻辑

利用 VT 技术开发 Hypervisor 的过程不同于利用 SVM 技术的开发过程，最主要的差别是在 VT 技术中，Guest 和 Hypervisor 下面要运行的指令地址（RIP/EIP）是在 VMCS 中设置的，同时 Hypervisor 就是用于处理 VMEXIT 事件，因此就像现代操作系统为系统调用设置一个统一入口，并将入口地址存入 MSR 寄存器一样，在 VT 中，通常也将 Hypervisor 的这个入口 IP 设置为事件处理函数入口地址（Event Dispatcher Address）。在事件处理的最后，又通过一个 VMXRESUME 指令统一的返回到 Guest 的指令执行流程中。对于事件发生信息，同样可以通过读取 VMCB 中相应数据获得。NewBluePill 中也有对 VT 技术的支持，我们同样会在深入探究 NewBluePill 的章节中详细展示怎样使用这些指令。

批注 [S10]: 后面要介绍 NBP 中关于 VT 技术的运用

Intel-VTd(如果时间充裕则写)

NewBluePill 项目介绍

NewBluePill 项目 (<http://www.bluepillproject.org/>) 诞生于 2007 年，由 Invisible Things Lab 开发，现在对外公开的版本是 nbp-0.32-public 版本。该项目从 2007 年第三季度以来得到了 Phoenix 公司的大力支持。¹

该项目目的是开发这样一种恶意软件：

- 不用已有的方法的隐藏自己
- 即使它的隐藏方法众所周知，其它软件也不能探测到它
- 即使它的实现代码众所周知，其它软件也不能探测到它

可以看出，该目的与非对称密码的设计目的有异曲同工之妙。该项目充分发掘并利用利用 HEV 功能。当前在公开版本上已经实现的功能包括：

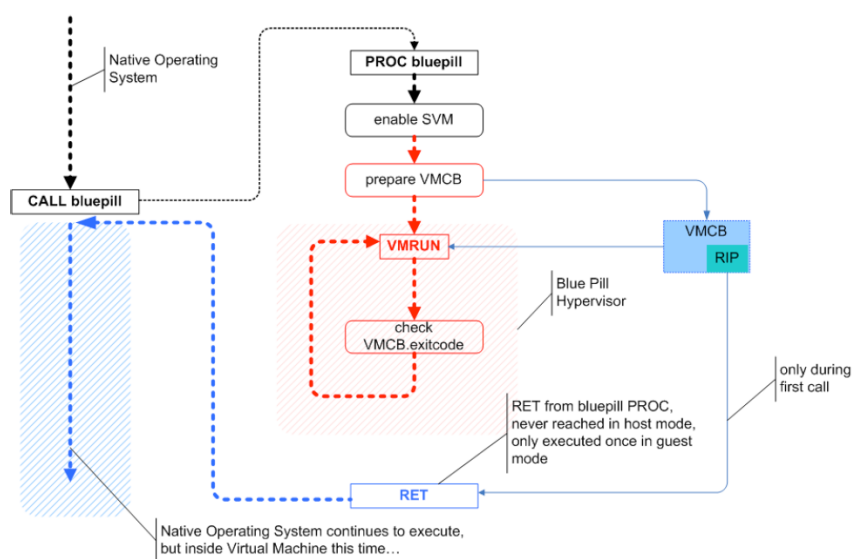
- 支持 SVM 和 VT-x 技术构建 Hypervisor
- 在操作系统运行时时刻动态加载和卸载，因此在操作系统完全不知情的情况下将操作系统放入了虚拟机中继续运行。
- 在 AMD 平台上支持嵌套 Hypervisor
- 一套自己的页表，用于实现内存隐藏，因此在操作系统中无法访问到 NewBluePill 的内存
- 反 Hypervisor 探测技术（Anti-Hypervisor Detector）：RDTSC 欺骗
- 反 Hypervisor 探测技术（Anti-Hypervisor Detector）：组织可信时间源的检测（Blue-Chicken 技术）

在其未公开的版本上，实现的功能包括：

- 在 Intel VT-x 平台上实现嵌套 Hypervisor

这些特性最终使得即使 NewBluePill 后于操作系统启动，操作系统却完全无法感知到 NewBluePill 的存在，这也就是 NewBluePill 的主要设计目标（如图 1.5 所示）。

¹ Phoenix 公司的虚拟化技术产品 HyperSpace，具体信息可以参考网上资料。与之类似的还有华硕公司（Asus Inc.）的 Instant On 技术

图 1.5 NewBluePill 的实现目标及思路¹

版权信息

本书引用 NewBluePill 代码版权归属于 Invisible Things Lab

/*

* Copyright holder: Invisible Things Lab

*

* This software is protected by domestic and International

* copyright laws. Any unauthorized use (including publishing and

* distribution) of this software requires a valid license * from the copyright holder.

*

* This software has been provided for the educational use

* only during the Black Hat training and conference. This

* software should not be used on production systems.

*

 $\ast/$

¹ 本图来源 *Subverting Vista™ Kernel For Fun And Profit*, Joanna Rutkowska

PART1 HEV 技术相关知识

二、深入 HEV 技术细节

在前一章中，我们简单介绍了 SVM 和 VT 技术，但是他们是如何被具体使用的呢？在本章中我们将详细介绍这些技术的细节：

- HEV 下虚拟机的启动过程
- VMEXIT 事件的陷入和处理
- 拆除 Hypervisor 和虚拟机
- EPT/NPT 翻译地址过程

直接阅读本章，可能会觉得理解其中内容却印象不深，推荐在阅读完全书后再次阅读本章——在理解了 NewBluePill 代码后，对本章内容自然会有更深的认识。

批注 [S11]: 本章写好后对这个列表更新

HEV 下虚拟机启动过程

“物有本末,事有终始,知所先后,则近道矣”——《大学》

想要了解 HEV 技术的本质，则要了解 HEV 要解决的问题和怎样解决这些问题。要熟悉这些，就要沿着虚拟机开启——运行——关闭的过程，看 HEV 技术是怎样融入其中的。所以我们首先就来看看在 HEV 技术的帮助下，虚拟机是怎样启动的。

启动过程模型

首先介绍下有了 HEV 技术后，启动虚拟机的方式。使用了硬件虚拟化技术的虚拟机可以有三种引导 Guest 操作系统的方式：

1. 存在特殊 OS/Host OS，后启动 Hypervisor 的虚拟机启动过程
 2. 存在特殊 OS/Host OS，先启动 Hypervisor 的虚拟机启动过程
 3. 不存在特殊 OS/Host OS，先启动 Hypervisor 的虚拟机启动过程
- 存在特殊 OS/Host OS，后启动 Hypervisor 的虚拟机启动过程。采用这种启动过程的虚拟机代表是 KVM，其启动过程如下：
 - a) 先启动宿主 Linux 操作系统
 - b) 在 Linux 中启动 KVM 设备，从而启动了 Hypervisor
 - c) 启动虚拟机，作为 Linux 进程运行

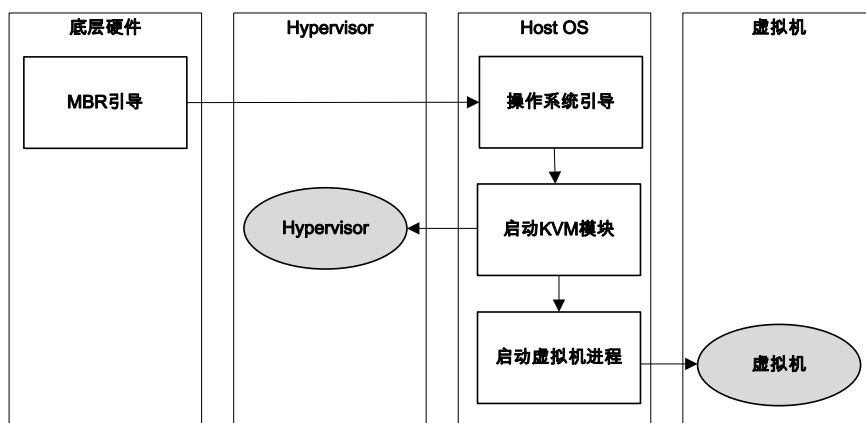


图 2.1 KVM 中虚拟机的启动过程

启动过程如图 2.1，可以看出，KVM 启动虚拟机的模式说明它不想脱离进程级虚拟机的本质，但是它要利用虚拟化技术进行加速。这样做的缺点在于需要一个 Host OS 充当载体。除 KVM 外，VMWare6.5 以上版本也是采用类似的架构，使用支持 HEV 技术的 CPU 进行加速。但是它们都需要再另外安装相应 Guest OS 上的驱动。

NewBluePill 也属于这样的启动模型，略有不同的是在成功启动 NewBluePill 后，它会把宿主 Windows 操作系统置于虚拟机中运行，详细过程可以参考本书后续章节。

- 存在特殊 OS/Host OS，先启动 Hypervisor 的虚拟机启动过程。采用这种启动过程的虚拟机代表是 Xen，其启动过程如下：
 - a) 先创建并启动 Hypervisor
 - b) 引导 Dom0
 - c) 由 Hypervisor 和 Dom0 一起协作创建虚拟机
 - d) 启动该虚拟机¹

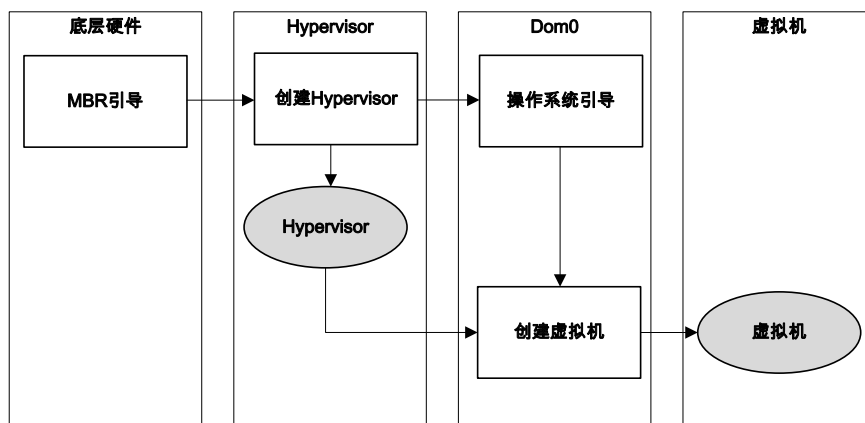


图 2.2 Xen 中虚拟机的启动过程

¹ Xen 中具体创建和启动虚拟机的过程会在“第 14 章 其它有关 HEV 项目”中介绍

启动过程如图 2.2，可以看出，Xen 中仍存在 Dom0 是因为它要适应过去未出现 HEV 技术时的架构，所以无论是 Dom0 还是 Hypervisor 的实现都比较笨重，并且安装和配置也比较麻烦，同样需要另外安装相应 Guest OS 上的驱动。但是不可忽视的是 Xen 的虚拟化效率最高。

- 不存在特殊 OS/Host OS，先启动 Hypervisor 的虚拟机启动过程。当前暂时没有采用这种启动过程的虚拟机软件（暂时称之为 UVM, Unknown Virtual Machine），其启动过程如下：

- a) 先创建并启动 Hypervisor
- b) 从 Hypervisor 中创建并启动虚拟机

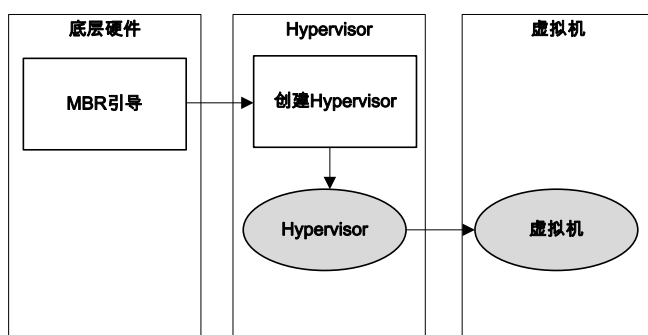


图 2.3 UVM 中虚拟机的启动过程

启动过程如图 2.3，这样的启动过程包括了如下组件：

表 2.1 UVM 模型下启动过程主要组件（传统 BIOS 启动）

| 组件 | 运行模式 | 作用 |
|-------------------|-----------------|---|
| 主引导扇区代码（MBR） | 16 位实保护模式 | 读取并加载活动分区启动扇区（Active Partition's Boot Sector） |
| 启动扇区（Boot Sector） | 16 位实保护模式 | 读取并运行磁盘上的 Hypervisor 创建程序 |
| Hypervisor 创建程序 | 16 位实保护模式，虚拟机模式 | 创建并初始化 Hypervisor，并创建至少一个虚拟机 |
| 虚拟机引导程序 | 虚拟机模式 | 任何已有的 BIOS 初始化程序或操作系统的 MBR 引导程序，目的是初始化虚拟机 |

批注 [S12]: 这种情况下，并不一定是启动扇区来读取 Hypervisor 运行的。例如在 EFI 情况下，是直接由 EFI 负责加载运行的。

这种虚拟机的设计目标在于：不需要在 Guest OS 中安装任何支持驱动。换句话说，Hypervisor 对于 Guest OS 完全透明，从而实现完全虚拟化（Full Virtualization）。这种方式的缺点是：Hypervisor 可能实现会很笨重，因而虚拟化效率不高，也会影响到系统安全，虚拟机的配置和管理可能也不易呈现给用户。

实验：阅读 Xen 和 KVM 的初始化部分代码

在 Xen 和 KVM 中，Hypervisor 的初始化代码都是用 C 编写的。阅读 Hypervisor 的初始化代码，对照图 2.1 和 2.2 体会其初始化的过程。

VT 技术下开启虚拟机的过程

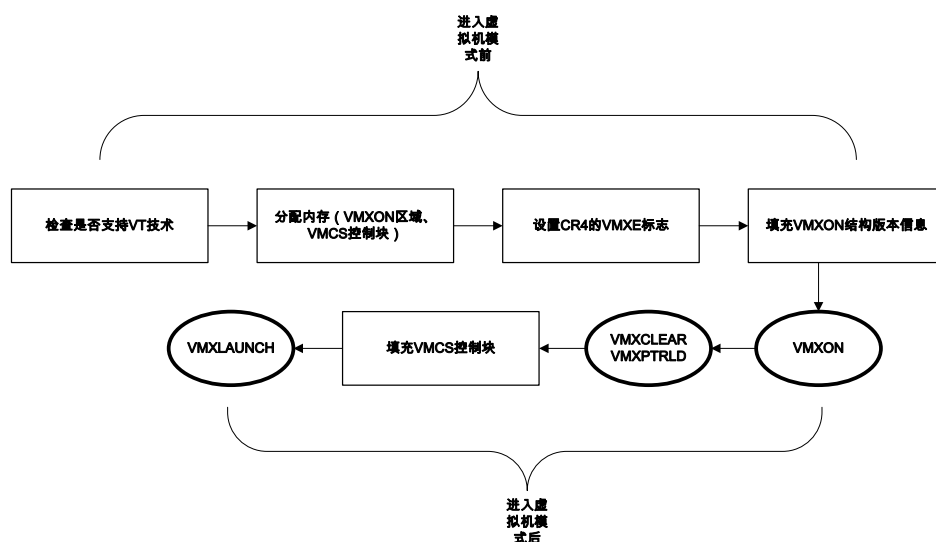


图 2.4 VT 技术下开启虚拟机的过程

在 VT 技术下进入虚拟机模式，开启虚拟机的全过程如图 2.4 所示。

首先检查当前 CPU 是否支持硬件虚拟化技术，这可以通过使用 `CPUID` 指令检查是否 `CPUID.1:ECX.VMX[bit 5]=1`（这句话表示操作数为 1 执行 `CPUID` 指令，检查返回的 `ECX` 寄存器 bit 5 位，也就是 `VMX` 位，查看结果是否为 1，下文均用这种表示法）。

开启虚拟机前，必须设置 `CR4.VMXE[bit 13]=1`，并且在内存中分配出来 `VMXON` 区域（`VMXON Region`）和 `VMCS` 控制块，后者也可以在进入虚拟机模式后再分配。需要注意的是，他们两者都必须分配在 4K 页对齐的内存区域上。

然后还需要初始化 `VMXON` 区域，需要把 `VMCS` 的版本号表示符（`VMCS Revision Identifier`）写入 `VMXON` 区域当中，该版本号可以通过访问 `IA32_VMX_BASIC` MSR 寄存器获得。除此以外不需要任何其它操作。只是要注意的是，在 `VMXON` 和 `VMXOFF` 指令之间的代码区中不要访问或者修改这个 `VMXON` 区域。

这之后通过执行 `VMXON` 指令即可以进入虚拟机模式。要注意的是，如果在执行 `VMXON` 指令时发现 `CR4.VMXE=0`，那么 `VMXON` 指令会发生无效操作数的异常（`#UD`¹）。最后，一旦进入虚拟机模式，`CR4` 和 `CR0` 寄存器中的一些与虚拟机模式相关的位将无法设置。

关于 `VMXON` 指令（`VMXON Instruction`）

`VMXON` 指令能否执行，同样也受 `IA32_FEATURE_CONTROL_MSR` 寄存器控制：

- Bit 0 是置锁位（Lock Bit） 如果该位为 0，那么 `VMXON` 指令不能执行；如果该位

¹ 关于异常的说明，请参考 *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 1* 中 Chapter 6.4 Interrupts and Exceptions 一节。

为 1，那么 WRMSR（写 MSR 寄存器指令）不能去写这个寄存器。该位在系统上电后便不能修改。BIOS 通过修改这个寄存器来设置是否支持虚拟化操作。在支持虚拟化操作的情况下，BIOS 还要设置 Bit 1 和 Bit 2

■ Bit 1 指示 VMXON 指令能否在 SMX¹操作环境中执行 如果这一位为 0，那么 VMXON 指令不能在 SMX 操作环境中执行

■ Bit 2 指示 VMXON 指令能否在 SMX²操作环境外执行 如果这一位为 0，那么 VMXON 指令不能在 SMX 操作系统外执行

如图 2.4 所指出的那样，执行 VMXON 指令后，我们需要执行 VMCLEAR 指令来初始化 VMCS 控制块。在第一章中我们介绍了 VMCS 区域的组成部分，在使用中，处理器会使用 VMCS 数据区来维护 VMCS 版本特定信息（implementation-specific information），VMCLEAR 指令会初始化这些信息，VT 技术推荐先执行 VMCLEAR 指令从而设置 VMCS 启动状态为空状态（clear），再执行 VMPTRLD 指令激活该 VMCS 块。

关于 VMCS 的启动状态

在第一章中我们在介绍了 VMLAUNCH 指令和 VMRESUME 指令的区别，里面同样涉及到 VMCS 的启动状态。

实际上，一个 VMCS 块的启动状态包含空状态（Clear）和已启动状态（Launched）两种状态，VMCLEAR 指令会把指定的 VMCS 块置为空状态，而 VMXRESUME 和 VMLAUNCH 两个指令会根据该状态的进行相应的操作。

要注意的是，通过 VMWRITE 指令是不能修改这个状态的。而且在转移一个 VMCS 块到另一个逻辑处理器上时，需要先使用 VMCLEAR 指令设置启动状态为空状态，再用 VMLAUNCH 启动该 VMCS 块所代表的虚拟机。因此尽量避免在不同逻辑处理器间转移执行 VMCS 块的操作。

利用 VMPTRLD 加载 VMCS 块之后，需要开发者按照需要配置 VMCS 块具体内容，VMCS 相关具体内容，会在本章稍后详加解释。

当这一切都设置好以后，利用 VMXLAUNCH 指令启动该 VMCS 块所代表虚拟机，至此虚拟机和 Hypervisor 均已初始化完毕，并能够成功开启虚拟机。

SVM 技术下开启虚拟机的过程

批注 [S13]: 阅读 AMD 手册补充这部分

HEV 下虚拟机关键结构体

VT 下的 VMCS 结构体和 SVM 下的 VMCB 结构体在开启虚拟机的过程中起着至关重要的作用，这一节我们就将深入探索 VMCS、VMCB 结构体的细节。

¹ SMX（安全扩展模式，Safer Mode Extensions）

² SMX（安全扩展模式，Safer Mode Extensions）

VT 技术下的 VMCS 结构体

在上一章中我们提到，VMCS 区域由 VMCS 版本标志（Revision Identifier）、VMX 退出原因指示器（VMX-abort indicator）、VMCS 数据区三部分构成。

在使用 VMCS 区域前，应当首先设置 VMCS 版本标志，因为文档中说明，这个域是由软件负责设置的，并且 VMPTRLD 指令在执行时会检查该 VMCS 块设置的版本标志与目标处理器能够接受的 VMCS 版本是否相符，不相符则会抛出异常。通常软件通过读取 IA32_VMX_BASIC MSR 寄存器来设置 VMCS 版本标志，因为这个 MSR 寄存器存有该处理器能够接受的 VMCS 版本标志。

VMX 退出原因指示器的产生是因为在 VM Exit 事件的时候如果遇到问题，系统就会发生 VMX Abort 事件，这会导致该逻辑处理器进入关闭状态。对于一个激活了的 VMCS，一个 VMX Abort 事件的发生并不会修改 VMCS 数据区，因此我们需要一种机制去判断是什么原因导致了 VMX Abort 事件的发生。通常，造成 VMX Abort 事件的原因包括：保存客户虚拟机 MSR 寄存器失败、当前 VMCS 区域损坏、加载 Hypervisor 的 MSR 寄存器失败等等。

VMCS 数据区构成了 VMCS 区域的主体，第一章中我们介绍了 VMCS 数据区的六个主要组成部分，下面我们将逐一介绍这六个主要部分又是怎样构成的。

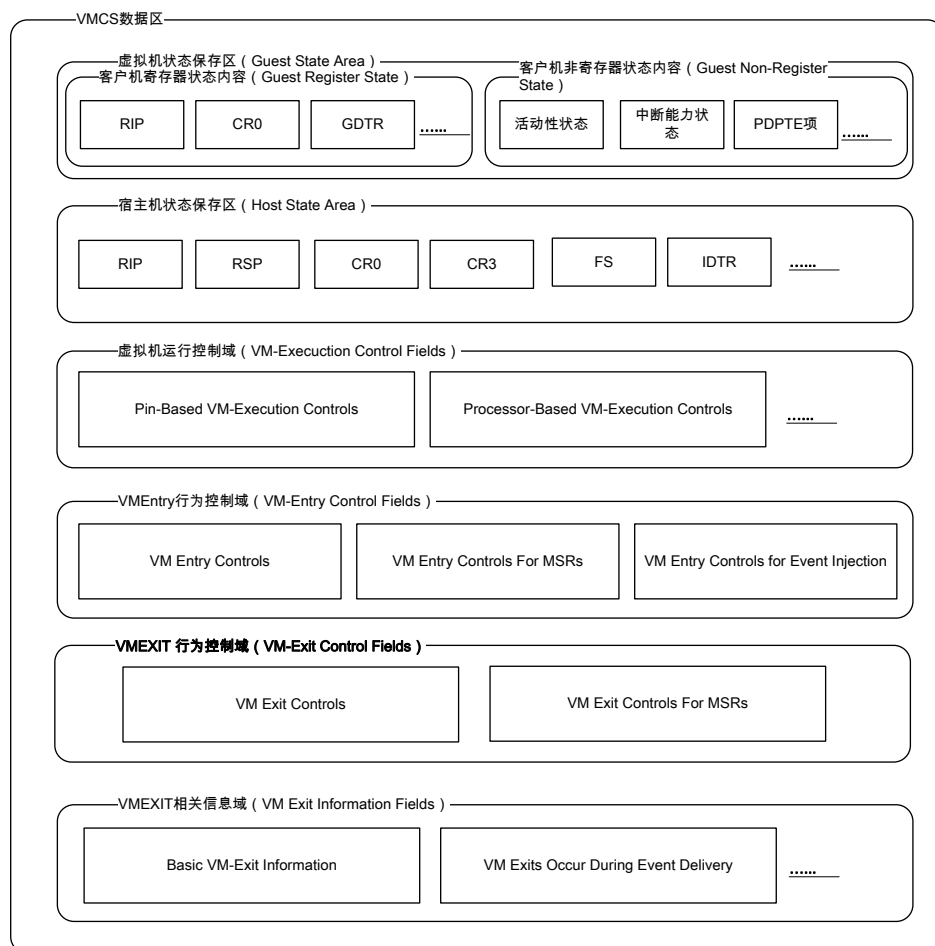


图 2.5 VMCS 数据区结构图

虚拟机状态保存区 (Guest-State Area)

这个区域保存了用于描述客户虚拟机状态的寄存器，当发生 #VMEXIT 事件的时候，虚拟机的状态会自动保存到这些域当中，而当发生 #VMENTRY 事件的时候，这些域中的值又被用来恢复虚拟机的运行状态。

这个区域可以保存的内容分为客户机寄存器状态内容 (Guest Register State) 和客户机非寄存器状态内容 (Guest Non-Register State) 两部分。客户机寄存器状态内容 (x86 上分别对应各自 x86 寄存器)：

- 1) 控制寄存器 CR0, CR3, CR4
- 2) 调试寄存器 DR7
- 3) RSP, RIP 和状态寄存器 RFLAGS
- 4) CS, SS, DS, ES, FS, GS, LDTR 和 TR 寄存器的下面各项

- 选择子 (Selector)
 - 基址 (Base Address)
 - 段长 (Segment limit)
 - 访问权限 (Access Rights)
- 5) GDTR 和 IDTR 信息
- 基址 (Base Address)
 - 段长 (Segment limit)
- 6) 一些 MSR 寄存器, 包括 IA32_DEBUGCTL, IA32_SYSENTER_CS, IA32_SYSENTER_ESP, IA32_SYSENTER_EIP, IA32_PERF_GLOBAL_CTRL, IA32_PAT, IA32_EFER

客户非寄存器状态内容包括:

- 1) 活动性状态 (Activity State) 这项表明了当前逻辑处理器的活动性, 也就是能否正常处理客户机程序指令, 包含 4 个状态:
 - a) 活跃的 (Active) 说明当前逻辑处理器可以执行客户机指令
 - b) 中断的 (HLT) 当前逻辑处理器不能执行客户机指令, 因为执行了一条 HLT 指令。
 - c) 关闭的 (Shutdown) 当前逻辑处理器不能执行客户机指令, 因为发生了严重的错误。
 - d) 等待 SIPI 中断 (Wait-for-SIPI) 当前逻辑处理器不能执行客户机指令, 因为在等待 Startup-IPI 中断¹
- 2) 中断能力状态 (Interruptibility State) x86 架构支持某些事件能被阻塞一段时间的特性, 包括被 STI 开中断指令阻塞, 被 MOV SS 阻塞, 被 SMI 中断阻塞和被 NMI 中断阻塞。这个域就包含了对应的描述信息。
- 3) 推迟调试的异常 (Pending Debug Exceptions) x86 支持延迟发送一些调试异常, 如某些情况下的单步调试。这个域就包含了对应的描述信息。
- 4) VMCS 连接指针 (VMCS Link Pointer) 该域保留, 用于未来扩展。
- 5) VMX 抢占计时器值 (VMX-Preemption Timer Value) 该域保存了虚拟机的 VMX 抢占计时器计数值。
- 6) PDPTE 项 (Page Directory Pointer Table Entries) 虚拟机状态保存区只有在虚拟机运行控制域中开启 EPT 模式才会用到此域, 其中包括 4 个 64 位数据: PDPTE0~PDPTE3。

关于 VMX 抢占计时器 (VMX-Preemption Timer)

VMX 抢占计时器是 VT 技术中这样的一种特性, 如果在 #VMEntry 事件发生后, 处理器硬件发现在虚拟机运行控制域中“启用 VMX 抢占计时 (Activate VMX-Preemption Timer)”被设置, 那么交由虚拟机执行指令时, 将启用一个 **VMX 抢占计时器**, 该计时器会倒数, 当倒数为 0 时会触发一个 #VMEXIT 事件陷入 Hypervisor 中。

如果“启用 VMX 抢占计时 (Activate VMX-Preemption Timer)”被设置, 同时 VMX 抢占计时器值为 0, 那么 #VMEXIT 事件会在 #VMENTRY 事件发生后执行任何指令前发生 (但是如果有推迟调试的异常 (Pending Debug Exceptions), 它们将先于 #VMEXIT 事件得到处理)。

VMX 抢占计时器的计时间隔 (Timer Interval) 是可以通过 IA32_VMX_MISC 寄存器设置

¹ SIPI (Startup Inter-Processor Interrupt), 用于多核平台初始化其它处理器。具体可参考 *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 3B* 手册

的 (0~31)，且为 2 的整数次幂于 TSC 寄存器值的变化。比如，如果我们设置该寄存器内容为 5，那么表示 TSC 寄存器中的值每增加 32，VMX 抢占计时器计数减 1。

但是要注意，由于节能模式的关系，所以当逻辑处理器在 C-states 的 C0, C1, C2 之外的其它状态时，VMX 抢占计时器不会产生 #VMEXIT 事件¹

对于系统管理中断 (SMIs, System Management Interrupts) 和系统管理模式 (SMM, System Management Mode)，VMX 抢占计时器的行为依赖于 Hypervisor 是否要处理这个模式：

- 默认情况下，VMX 抢占计时器在虚拟机处理系统管理中断时，也会继续倒计时时，只不过在倒数为 0 的情况下，VMX 抢占计时器要继续等到虚拟机脱离系统管理模式后再产生 #VMEXIT 事件。

- 如果 Hypervisor 的设定指出同样要监管系统管理中断和系统管理模式时，这个时候 VMX 抢占计时器的行为和在处理一般 VMEXIT 事件、VMEntry 事件时一样。

宿主机 (Hypervisor) 状态保存区 (Host-State Area)

这个区域记录了所有有关 Hypervisor 的状态信息，正如第一章中所提到的，这个区域保存的内容会在每次发生 #VMEXIT 事件时恢复到相应的寄存器中，以恢复 Hypervisor 的执行环境。

与虚拟机状态保存区不同，宿主机状态保存区只能存储有关寄存器的信息：

- 1) 控制寄存器 CR0, CR3, CR4
- 2) RSP, RIP
- 3) CS, SS, DS, ES, FS, GS 和 TR 寄存器的下面各项
 - 选择子 (Selector)
- 4) FS, GS, TR, GDTR 和 IDTR 信息
 - 基址 (Base Address)
- 5) 一些 MSR 寄存器，包括 IA32_SYSENTER_CS, IA32_SYSENTER_ESP, IA32_SYSENTER_EIP, IA32_PERF_GLOBAL_CTRL, IA32_PAT, IA32_EFER

虚拟机运行控制域 (VM-Execution Control Fields)

虚拟机运行控制域管理着虚拟机的运行，包含下列域：

- 1) 基于针脚的虚拟机执行控制 (Pin-Based VM-Execution Controls)

这个域用来管理中断等异步事件 (Asynchronous Event)²，如“启用抢占计时”就是在这个域中设置的。对于其中保留位的设置，软件要参考 IA32_VMX_PINBASED_CTLs 和 IA32_VMX_TRUE_PINBASED_CTLs 两个 MSR 寄存器的内容。

- 2) 基于处理器的虚拟机执行控制 (Processor-Based VM-Execution Controls)

该域包含两个 4 字节值，用于管理执行特定指令而产生的同步事件 (Synchronous Event)。这个控制域分为主要基于处理器的虚拟机执行控制和次要基于处理器的虚拟机执行控制两个，前者包括对 HLT、INVLPG、MWAIT、RDTSC、CR3 读取/存储、

¹ 有关节能模式和 C-states 的详细信息，可以参考网上相关资料

² 有些中断不受该域控制，触发 VMEXIT 事件

使用 I/O Bitmap 等等这些产生 VMEXIT 事件的控制；后者包括对开启 EPT 地址翻译、开启 VPID¹、开启虚拟 APIC（高级可编程中断控制器，关于此部分的介绍，可参考随后“关于可编程中断控制器”部分）等等的控制。对于其中保留位的设置，前者要参考 IA32_VMX_PROCBASED_CTLs 和 IA32_VMX_TRUE_PROCBASED_CTLs 两个 MSR 寄存器的内容，后者要参考 IA32_VMX_PROCBASED_CTLs2 MSR 寄存器。

3) 异常位图（Exception Bitmap）

该域仅有 32 位长，每一位代表当某种异常发生时，硬件会自动产生 VMEXIT 事件；如果某一位为 0，则表示这个异常会通过 IDT 表正常处理。这其中要注意的是缺页异常的处理略有不同，需要借助于 VMCS 中另外的两个域（Page Fault Error Code Mask 和 Page Fault Error Code Match）来处理。

4) I/O 位图地址（I/O Bitmap Addresses）

该域包含两个 64 位长的物理地址，指向两块 I/O 位图，只有在 Primary Processor-Based VM-Execution Controls.Use I/O Bitmaps[bit 25]=1 的情况下才会被使用。逻辑处理器在处理 I/O 指令时，会根据这些 I/O 位图而在访问相应地址时产生 VMEXIT 事件。VT 技术要求这些位图必须在 4K 页对齐的位置上。

5) 时间戳寄存器偏移值（Time-Stamp Counter Offset）

虚拟机运行控制域还包括一个时间戳寄存器偏移值域，这个域在 Primary Processor-Based VM-Execution Controls.RDTSC Exiting[bit 12]=0 且 Primary Processor-Based VM-Execution Controls.Use TSC Offsetting[bit 3]=1 时起作用。当客户机利用 RDTSC、RDTSCP 指令或者访问 IA32_TIME_STAMP_COUNTER MSR 寄存器的时候，得到的结果将是真实的值加上这个偏移量的和。

6) 虚拟机/Hypervisor 屏蔽和 CR0/CR4 访问隐藏设置（Guest/Host Masks and Read Shadows for CR0 and CR4）

这个域主要对 CR0 和 CR4 寄存器进行保护。虚拟机/Hypervisor 屏蔽中置位 1 的位说明 CR0 和 CR4 寄存器的相应位只能由 Hypervisor 修改，否则产生 VMEXIT 事件，置位 0 的部分说明 CR0 和 CR4 的相应位可以在虚拟机中修改。

7) CR3 访问目标控制（CR3-Targeting Controls）

包含最多 4 个 CR3 目标值²，当在虚拟机中执行 CR3 的赋值操作时，如果赋予的值是这些目标值中任意一个，那么硬件不会产生 VMEXIT 事件。

8) APIC 访问控制（Controls for APIC Accesses）

访问本地 APIC 的寄存器有三种方法：通过 xAPIC 模式访问、通过 x2APIC 模式访问、在 64 位模式下通过 mov CR8 指令来访问任务优先级寄存器（Task Priority Register, TPR）³。APIC 的访问控制实际上也是 VT 技术对 APIC 的虚拟化，在这个域中，通过配置“使用影子 TPR（Use TPR Shadow）”、“虚拟化 APIC 访问（Virtualize APIC Accesses）”、“虚拟

¹ 开启 VPID (Virtual-Processor Identifier)意味着缓存在翻译线性地址时会根据 VPID 进行翻译，其优点已在第一章中介绍。

² 未来可能在此处容纳更多的 CR3 目标值，因此在使用前参考 IA32_VMX_MISC MSR 寄存器来查看当前处理器详细信息。

³ 请参考 Intel 手册相关部分以获得有关三种 APIC 访问方法的详细资料。

化 x2APIC 模式 (Virtualize x2APIC Mode)” 对 APIC 进行虚拟访问。

9) MSR 位图地址 (MSR-Bitmap Address)

该域包含一个指向 MSR 位图区域的物理地址，当 Primary Processor-Based VM-Execution Controls.Use MSR Bitmaps[bit 28]=1 时会被使用。MSR 位图区域占据 4K 大小内存，分为 4 个连续区域：读低地址/高地址 MSR、写低地址/高地址 MSR。根据配置，当相应的 MSR 访问执行时，会发生 VMEXIT 事件。

10) 执行体 VMCS 指针 (Executive-VMCS Pointer)

该域 64 位长，用于 VT 技术对系统管理中断 (SMI) 和系统管理模式 (SMM) 进行监管 (也称 Dual-Monitor Treatment)

11) EPT 指针 (Extended Page Table Pointer)

该域包含了 EPT 页表的基地址 (也就是指向 EPML4 级页表的物理地址)，以及一些 EPT 页表的配置信息，当 Secondary Processor-Based VM-Execution Controls.Enable EPT[bit 1]=1 时启用。

12) 虚拟机标示符 (Virtual Processor Identifier, VPID)

虚拟机表示符定义为 16 位长，当 Secondary Processor-Based VM-Execution Controls.Enable EPT[bit 5]=1 时启用。

关于可编程中断控制器

■ x86 上的中断控制器

在大多数很古老的 x86 系统上都有一个 i8259A 可编程中断控制器 (Programmable Interrupt Controller, PIC)，或者新一些的，i82489 高级可编程中断控制器 (Advanced Programmable Interrupt Controller, APIC)。APIC 兼容 PIC，前者拥有 256 条中断线，而后者仅拥有 15 条中断线；前者支持多核技术，而后者并不支持。APIC 架构图如图 2.6 所示

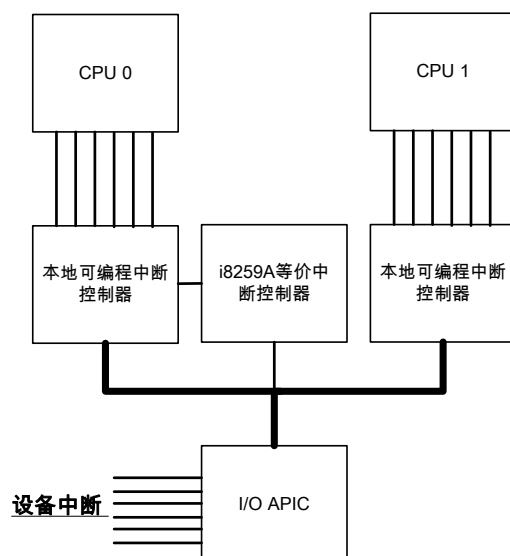


图 2.6 可编程中断控制器 (APIC) 架构图

APIC 包括下列组成部分:

- ◆ I/O APIC 用于从设备接收中断
- ◆ 本地可编程中断控制器 (Local APIC) 本地中断控制器利用一条私有总线, 从 I/O APIC 上接收中断, 然后向与其关联的 CPU 发送中断。
- ◆ i8259A 等价中断控制器 (i8259A Equivalent PIC) 负责把 APIC 的输入信号翻译成 PIC 等价的信号, 用于实现在 APIC 上对 PIC 的兼容

APIC 私有总线也要负责决定要把该中断信号发送到哪个 Local APIC 上 (Windows 有权决定是否要利用 I/O APIC 的这个算法), 这样做可以更好的利用处理器本地性。

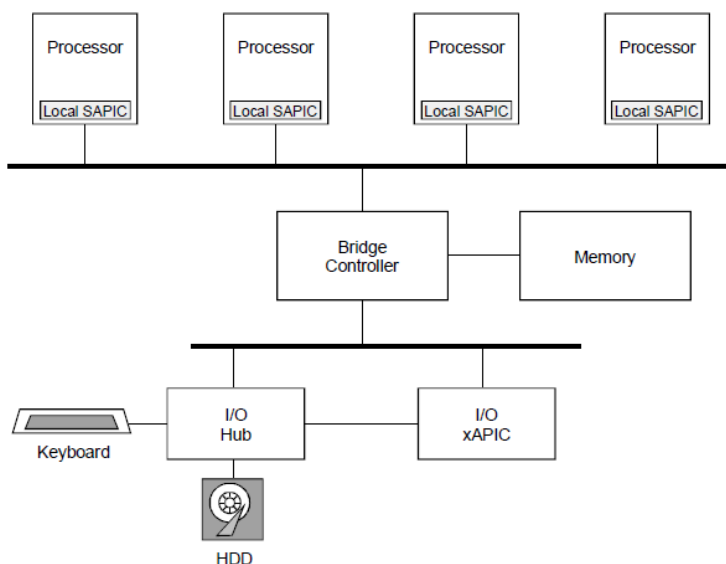
Local APIC 的另外一个功能是发送处理器间中断 (Inter-Processor Interrupts, IPI) 到其它的 Local APIC 上, 考虑下列情况: 每当一个时钟中断发生的时候, 在多核平台上, 显然只应该有一个处理器去负责更新系统时间, 否则系统时间将会在一次时钟中断下被多次更新。这个时候就需要 IPI 的帮忙了。其它情况比如更新 TLB Cache、在 DISPATCH_LEVEL 中断级别下调度一个线程、系统崩溃、系统关闭时均需要 IPI 中断的帮忙。

■ x64 上的中断控制器

因为 x64 平台要兼容于 x86 平台, 所以 x64 平台上的中断控制器与 x86 平台的相同。但是 x64 平台上的 Windows 要求一定要有 APIC 作为中断控制器。

■ IA64 上的中断控制器

IA64 平台使用 SAPIC (Streamlined Advanced Programmable Interrupt Controller) 作为其中断控制器。SAPIC 的 I/O APIC 不再使用一条私有总线传输中断, 而是利用系统总线, 进而通过北桥控制器来传输中断 (见图 2.7)。这样做的好处是能够更快的传输中断。

图 2.7 SAPIC 架构图¹

¹此图摘自 Intel Itanium Processor Family Interrupt Architecture Guide

另一个不同之处在于中断分发的路由算法上，由于 SAPIC 没有了私有 APIC 总线，所以中断路由算法就要写入固件（Firmware）当中。与 x86 中的情况类似，Windows 可以决定是否要使用这个算法。

实验：查看本机上的 APIC 配置

在多核机器上，可以在 windbg 中利用 !apic 命令查看本机 APIC 配置，“0: kd”表明当前运行在 CPU 0 上，m 表明当前中断被屏蔽。

```
0: kd> !apic
Apic @ fffe0000 ID:0 (40011) LogDesc:01000000 DestFmt:ffffff TPR FF
TimeCnt: 03f07410clk SpurVec:1f FaultVec:e3 error:0
Ipi Cmd: 02000000`000008e1 Vec:E1 FixedDel Lg:02000000 edg high
Timer..: 00000000`000300fd Vec:FD FixedDel Dest=Self edg high m
Linti0.: 00000000`0001001f Vec:1F FixedDel Dest=Self edg high m
Linti1.: 00000000`000004ff Vec:FF NMI Dest=Self edg high
TMR: 63, 83, B1
IRR: B1, D1
ISR: D1
```

Note 阅读 *Intel Itanium Processor Family Interrupt Architecture Guide* 文档可以获得关于 IA64 SAPIC 的更详细的指导信息

VMEntry 行为控制域（VM-Entry Control Fields）

VMEntry 行为控制域定义了 VMEntry 事件发生后硬件要立即做的事情，主要包括三部分：VMEntry 基本操作控制设置（VM Entry Controls）、VMEntry MSR 寄存器操作控制设置（VM Entry Controls For MSRs）和 VMEntry 注入事件控制设置（VM Entry Controls for Event Injection）。

VMEntry 基本操作控制设置包括：

- 1) 加载调试寄存器内容
 - DR7
 - IA32_DEBUGCTL MSR 寄存器
- 2) 虚拟机是否进入 x64 支持模式（x86 架构上永远为 0）
- 3) 进入系统管理模式（SMM）
- 4) 关闭 Dual-Monitor Treatment
- 5) 加载 IA32_PERF_GLOBAL_CTRL MSR 寄存器
- 6) 加载 IA32_PAT MSR 寄存器
- 7) 加载 IA32_EFER MSR 寄存器

对于其它保留位的设置，软件必须根据 IA32_VMX_ENTRY_CTLS MSR 寄存器和 IA32_VMX_TRUE_ENTRY_CTLS MSR 寄存器内容设置。

对于 VMEntry MSR 寄存器操作控制设置，细心的读者可能发现，前面我们在描述客户机状态域时提到过，硬件会在发生 VMEntry 事件后自动加载一些有关虚拟机状态 MSR 寄存器。这里 VMEntry MSR 寄存器操作控制设置允许开发人员恢复更多的 MSR 寄存器，主要通过下面两个域配置：

- 1) VMEntry MSR 寄存器加载数量 (VMEntry MSR-Load Count)
- 2) VMEntry MSR 寄存器加载地址 (VMEntry MSR-Load Address)

VMEntry 事件可以在所有虚拟机状态恢复完毕后，通过客户机 IDT 表触发一个中断。VMEntry 注入事件控制设置就是用来配置这个特性的，它主要有如下三个可配置部分：

- 1) VMEntry 中断信息域 (VM Entry Interruption Information Field)
- 2) VMEntry 异常错误码 (VM Entry Exception Error Code)
- 3) VMEntry 指令长度 (VMEntry Instruction Code)¹

VMEXIT 行为控制域 (VM-Exit Control Fields)

VMEXIT 行为控制域定义了 VMEXIT 事件发生后硬件要立即做的事情，主要包括两部分：VMEXIT 基本操作控制设置 (VM Exit Controls) 和 VMEXIT MSR 寄存器操作控制设置 (VM Exit Controls For MSRs)。

VMEXIT 基本操作控制设置包括：

- 1) 保存调试寄存器内容
 - DR7
 - IA32_DEBUGCTL MSR 寄存器
- 2) Hypervisor 地址空间大小 (x86 架构上永远为 0)
- 3) 加载 IA32_PERF_GLOBAL_CTRL MSR 寄存器
- 4) VMEXIT 保留外部中断原因信息 (Acknowledge Interrupt on Exit)
- 5) 保存 IA32_PAT MSR 寄存器
- 6) 加载 IA32_PAT MSR 寄存器
- 7) 保存 IA32_EFER MSR 寄存器
- 8) 加载 IA32_EFER MSR 寄存器
- 9) 保存 VMX 抢占计时器值

VMEXIT MSR 寄存器操作控制设置类似于上文中的 VMEntry MSR 寄存器操作控制设置，它允许开发人员在 VMEXIT 事件发生后保存更多有关虚拟机状态的 MSR 寄存器，并恢复更多有关 Hypervisor 状态的 MSR 寄存器，主要通过下面四个域实现：

- 1) VMEXIT MSR 寄存器保存数量 (VM-Exit MSR-Store Count)
- 2) VMEXIT MSR 寄存器保存地址 (VM-Exit MSR-Store Address)
- 3) VMEXIT MSR 寄存器加载数量 (VM-Exit MSR-Load Count)
- 4) VMEXIT MSR 寄存器加载地址 (VM-Exit MSR-Load Address)

¹ 对于软中断 (Software Interrupt)、软件异常 (Software Exception) 和特权软件异常 (Privileged Software Exception)，这个域用来决定填充到异常堆栈上的 RIP 地址指针值。

VMEXIT 相关信息域 (VM Exit Information Fields)

VMEXIT 相关信息域包含了最近#VMEXIT 事件相关的信息。这是一个只读的域，尝试利用 VMWRITE 指令向这个域中写入信息会失败。该域主要包括五个部分：VMEXIT 事件基本信息区 (Basic VM-Exit Information)、向量化事件 VMEXIT 信息区 (VM Exits Due to Vectored Events)、事件分发时 VMEXIT 信息区 (VM Exits Occur During Event Delivery)、指令执行时 VMEXIT 信息区 (VM Exits Due to Instruction Execution)、VM 指令错误信息域 (VM-Instruction Error Field)

VMEXIT 事件基本信息区包括：

- 1) 退出原因 (Exit Reason)
- 2) 退出条件 (Exit Qualification)
- 3) 客户机线性地址 (Guest-linear address)
- 4) 客户机物理地址 (Guest-physical address)

向量化事件 VMEXIT 信息区用于异常，外部中断，NMI 等造成的 VMEXIT 事件的信息呈现。包括：

- 1) VMEXIT 中断信息 (VMExit Interruption Information)
- 2) VMEXIT 中断异常号 (VMExit Interruption Error Code)

事件分发时 VMEXIT 信息区用于在虚拟机中传播事件时发生的 VMEXIT 事件的信息记录。包括：

- 1) IDT 向量表信息 (IDT-Vectoring Information)
- 2) IDT 向量表异常号 (IDT-Vectoring Error Code)

指令执行时 VMEXIT 信息区，该域记录在虚拟机中执行某些指令时发生的 VMEXIT 事件的信息。包括：

VMEXIT 指令长度 (VMEXIT Instruction Length)

VMEXIT 指令信息 (VMEXIT Instruction Information)

VM 指令错误信息域描述了执行虚拟机操作指令 (VMX Operation) 发生的最后一个 Non-Faulting 错误的信息。

SVM 技术下的 VMCB 结构体

VMEXIT 事件的陷入和处理

什么拦截的到什么拦截不到

有条件陷入、无条件陷入

SMM 和 Hypervisor

HEV 下虚拟机关闭过程

VT 技术下关闭 Hypervisor 和虚拟机的过程

对于 Hypervisor 的关闭，Intel 技术文档要求必须在 Hypervisor 下，通过执行 VMXOFF 指令来关闭虚拟机模式，通过检查 RFLAGS[CF]=0 和 RFLAGS[ZF]=0 来判断是否关闭成功，然后可以清除 CR4.VMXE 的值。

要注意的是，如果 Hypervisor 开启了 SMM 监控器（SMM Monitor），那么一定要先关闭这个监控器再拆除 Hypervisor，否则执行 VMXOFF 指令不成功。

另一个要注意的问题是，成功卸载 Hypervisor 后，由于此时已经没有了虚拟机模式，所以如果还需要继续运行操作系统的话，那么必须人工的还原操作系统运行环境，很多时候这指的是人工将 VMCS 保存的虚拟机环境填充为当前环境，一个显而易见的例子是：如果不人工设置 RIP/EIP，那么操作系统此时不可能继续执行指令。

SVM 技术下关闭 Hypervisor 和虚拟机的过程

EPT/NPT（Chapter 24）

Vmxon Region, VMCS, I/O 位图, Msr 位图

AMD Nested Page Table 详细技术细节 AMD 手册 P406

总结

SVM 和 VT 不同之处和使用时应注意的地方

通过前文的描述，看上去 SVM 技术和 VT 技术十分相似，但是实际上两者还是有一些不同之处。在开发过程中必须注意到这些不同之处，它们是正确并且高效实现 Hypervisor 的关键。

SVM 的开发逻辑中，VMRUN 和事件处理程序要处于同一循环中，这是因为 Hypervisor 的事件处理程序入口在 VMRUN 的下一条地址上，而在 VT 技术中，由于可以自由指定这个入口地址，因此可以在 VMCS 块中指定一个函数作为事件处理入口函数。

SVM 采用 ASID 作为 TLB 中 Guest 和 Hypervisor 地址的标记，而 VT 采用 VPIDs（Virtual-Processor Identifiers）作为 TLB 中不同虚拟机地址翻译的缓存标记，因此 VT 技术的缓存策略更精细所以更好些。

虽然 SVM 技术和 VT 技术都可以管理中断，管理资源，访问控制，但两者具体处理行为有一些差别，VT 技术将指令分为三种：无条件陷入的指令，有条件陷入的指令和不产生陷入的指令/事件。SVM 技术则分为了异常拦截和指令拦截，其中一些异常虽然会造成陷入，但是同时也会自动标记相应的异常寄存器（Exception Specific Registers）。应用的时候一定要根据手册上的描述给出相应的实现。

SVM 和 VT 技术在使用时一定要注意到，#VMEXIT 事件的产生来源于异常而不是行为，比如用户可以拦截 RDMSR 指令，但是发生的 sysenter 指令却不能拦截到，是因为在这种情况下，虽然 sysenter 有读取 MSR 寄存器的操作，但是因为没有提前在 VMCS/VMCB 中设定处理器遇到 sysenter 产生异常，所以处理器执行到 sysenter 指令当然也就不会产生 #VMEXIT 事件。

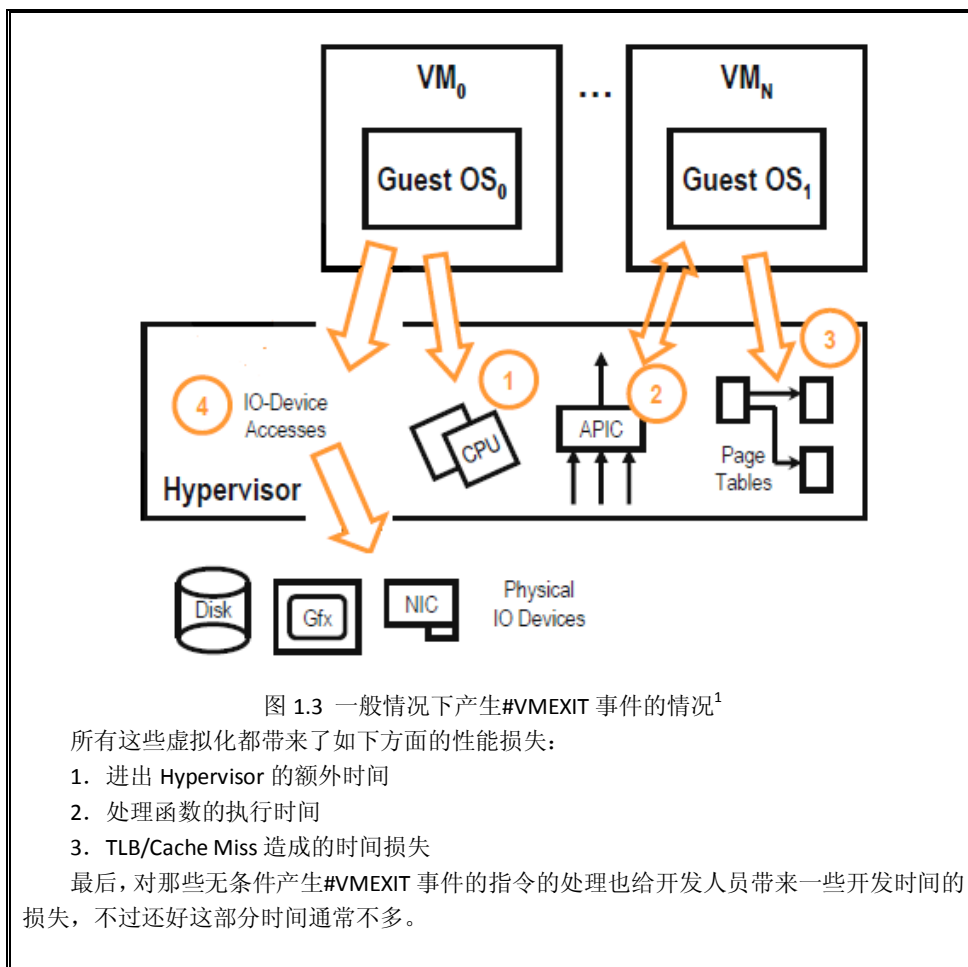
HEV 技术所带来的性能损耗

新技术在使得开发人员的世界变得更加美好的同时，也不可避免的带来性能上的冲击。

HEV 技术中，性能损耗最大的地方在于 Hypervisor 的引入及其所造成的需要进出 Hypervisor。一个最简单的例子，在普通的 x86 保护模式下，运行时刻执行到 CPUID 指令时，处理器会根据 EAX（RAX）寄存器的值直接读取 MSR 寄存器，并把结果写到 EAX~EDX（RAX~RDX）寄存器。但是在 Guest 模式下并且设置对 CPUID 指令进行拦截，那么每当 Guest OS 执行到 CPUID 指令时，处理器都会产生 #VMEXIT 事件，从而陷入 Hypervisor 中对该指令进行相应处理，这个过程中涉及到 Guest 模式寄存器的保存，Host 模式寄存器的恢复，填充 VMCS/VMCB 中相应的内容（都是一系列处理器自动完成的内存操作），然后 Hypervisor 中不可缺少的有 CPUID 指令陷入的处理，最后在 Hypervisor 处理完后，处理器要回到 Guest 模式，这又涉及到 Host 当前寄存器的保存，Guest 模式寄存器的恢复，以及 VMCS/VMCB 中相应内容的填充。显然，花费在这上面的指令周期数将是保护模式下 CPUID 一条汇编指令所消耗的指令周期数的成千上万倍以上。

而在更一般的情况下，下列四种产生 #VMEXIT 事件的情况都是需要处理的：

1. 访问特权级别的 CPU 的状态（Access to Privileged CPU State）
2. 中断虚拟化（Interrupt Virtualization）
3. 页表虚拟化（Page-Table Virtualization）
4. IO 设备虚拟化（IO-device virtualization）



Note 关于此章内容更详细的信息，请参考 *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 3B* 和 *AMD64 Architecture Programmer's Manual, Volume 2: System Programming, Chapter 15 Secure Virtual Machine*

¹ 此图摘自 *Intel® Virtualization Technology Processor Virtualization Extensions and Intel® Trusted execution Technology*, Gideon Gerzon

PART2 深入研究 NewBluePill

批注 [S14]: 在介绍了 2 家的技术后，加一些自己的总结，或者对比。不要光是介绍内容，要有自己的分析。

三、 体验 NewBluePill

首先介绍下我的平台，在整个项目中我用了两台计算机

PC1（调试机）：Intel Core 2 6300, 1G RAM, XP SP2 (x86) + windbg + WDK6001.18001

PC2(被调试机)：Intel Core 2 6300, 1G RAM, Windows Server 2008 Beta 1(X64), NewBluePill
只能运行于这台机器上。

编译 NewBluePill

了解了以上那么多，是不是很想亲自动手尝试下呢？不过先别急，还是先把工具准备好再说。

工具一共有下面几个：

1. Windbg
2. DebugView¹
3. InstDrv²
4. Windows Driver Kits (WDK 6001.18001)

总体来说编译 NewBluePill 的过程很简单。

步骤 1. 首先确保手上了 nbp-0.32-public.zip 这个代码³。然后解压缩到一个根目录，在这里我们假设是 D 盘。目录结构应该是这样的：

¹ DebugView，可以到 <http://download.sysinternals.com/Files/DebugView.zip> 下载

² InstDrv，可以到 <http://dl2.csdn.net/fd.php?i=23314208212665&s=0affa2ecb56fc0dcc14cff07345a388e> 下载

³ NewBluePill 项目源代码可以从 <http://www.bluepillproject.org/> 下载

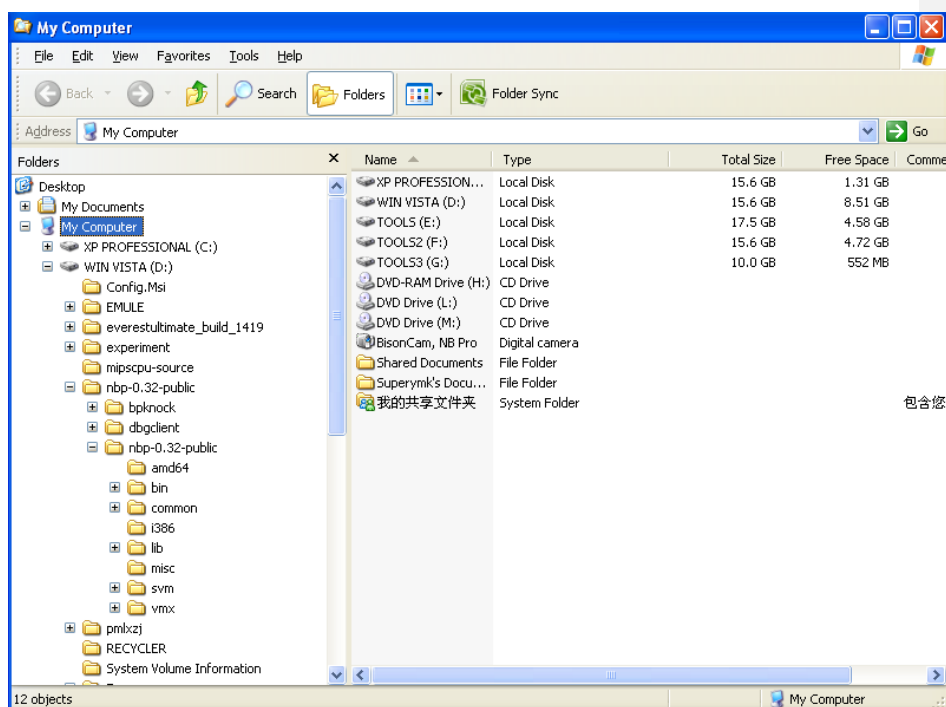


图 3.1 NewBluePill 项目目录结构

步骤 2. 然后打开 Launch Windows Vista and Windows Server 2008 x64 Checked Build Environment 编译环境:

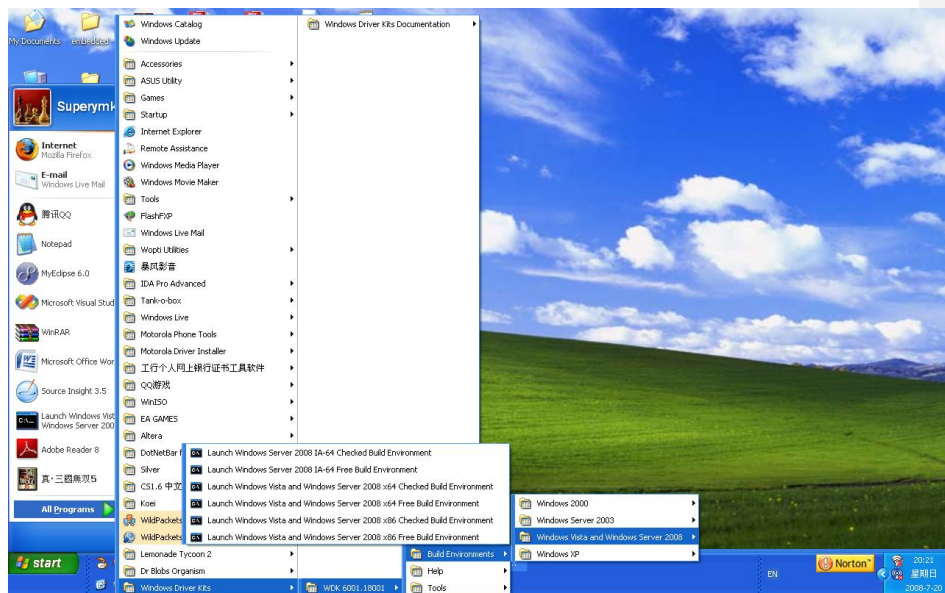


图 3.2 WinDDK 编译环境快捷方式的位置

步骤 3. 在该编译环境中执行 `nbp-0.32-public\NewBluePill-0.32-public\build_code.cmd`,

如果编译成功则会出现以下窗口：

```

1>Compiling - vmx\vmxdebug.c
2>Building Library - lib\amd64\svm.lib
1>Building Library - lib\amd64\vmx.lib
1>BUILD: Compiling and Linking d:\nbp-0.32-public\nbp-0.32-public\common directory
1>Assembling - amd64\msr.asm
1>Assembling - amd64\svm-asm.asm
1>Assembling - amd64\vmx-asm.asm
1>Assembling - amd64\common-asm.asm
1>Assembling - amd64\regs.asm
1>Assembling - amd64\cpuid.asm
1>Assembling - amd64\inststubs.asm
1>Compiling - common\newbp.c
1>Compiling - common\hvm.c
1>Compiling - common\portio.c
1>Compiling - common\comprint.c
1>Compiling - common\hypercalls.c
1>Compiling - common\traps.c
1>warnings in directory d:\nbp-0.32-public\nbp-0.32-public\common
1>d:\nbp-0.32-public\nbp-0.32-public\common\traps.c : warning C4819: The file contains a character that cannot be represented in the current code page (936). Save the file in Unicode format to prevent data loss
1>Compiling - common\interrupts.c
1>Compiling - common\common.c
1>Compiling - common\paging.c
1>Compiling - common\snprintf.c
1>Compiling - common\chicken.c
1>Compiling - common\dbgclient.c
1>Linking Executable - bin\amd64\newbp.sys
BUILD: Finish time: Sun Jul 20 20:22:39 2008
BUILD: Done

30 files compiled - 2 Warnings
2 libraries built
1 executable built

D:\nbp-0.32-public\nbp-0.32-public>ctags -R
'ctags' is not recognized as an internal or external command,
operable program or batch file.

D:\nbp-0.32-public\nbp-0.32-public>

```

图 3.3 显示编译成功信息的 WinDDK 控制台

如果看到这个提示，恭喜你，编译成功了！

演示 NewBluePill

运行 NewBluePill 就有一定要求了，首先要求必须运行在支持虚拟技术（HVM）的 CPU 上，并且推荐在 64 位或者支持虚拟 64 位技术的 CPU 上运行，原因是虽然 NewBluePill 程序中附带了支持 32 位 CPU 的代码，但是有几个函数在编译时（Vista x86 Checked Mode）会出现问题¹，而且有几个函数是未实现的，所以还是在 x64 上去跑吧。

下面是详细步骤：

步骤 1：重启被调试机，按 F8，然后选择 Disable Driver Signature Enforcement(切记一定要用这个模式启动，否则不能加载未签名的驱动程序)

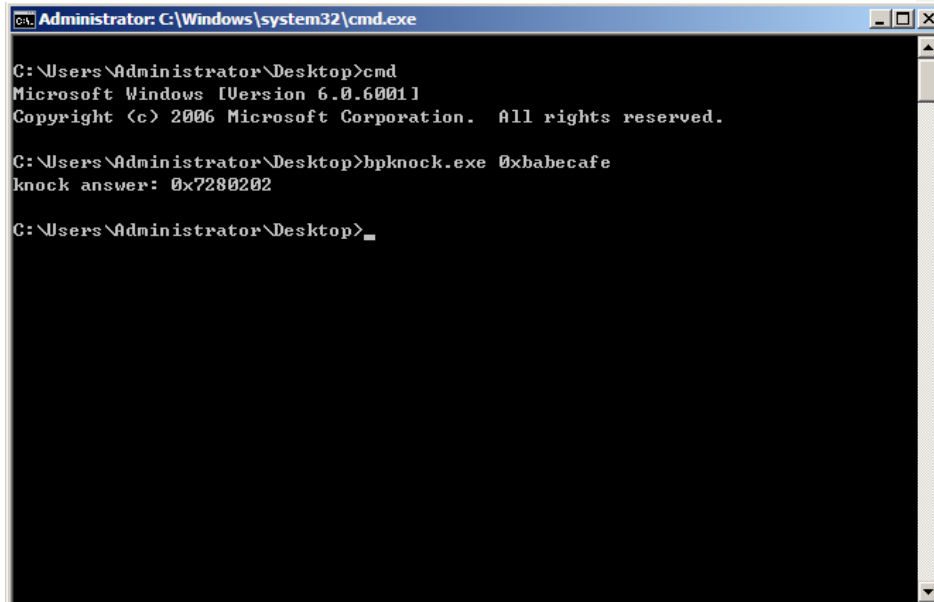
步骤 2：去 nbp-0.32-public 主目录及其子目录下找到下面几个编译生成的二进制文件：

¹ 这个问题会作为实验内容“第十二章 移植 NewBluePill 到 32 位系统”留给读者解决

批注 [S15]: 此处出现“第十二章 移植 NewBluePill 到 32 位系统”章号

bpknock.exe, dbgclient.sys, newbp.sys

步骤 3: 运行下 bpknock 0xbabecafe 看下没运行 NewBluePill 的输出结果。



```
Administrator: C:\Windows\system32\cmd.exe

C:\Users\Administrator\Desktop>cmd
Microsoft Windows [Version 6.0.6001]
Copyright (c) 2006 Microsoft Corporation. All rights reserved.

C:\Users\Administrator\Desktop>bpknock.exe 0xbabecafe
knock answer: 0x7280202

C:\Users\Administrator\Desktop>_
```

图 3.4 未加载 newbp 驱动的 bpknock 程序输出结果

步骤 4: 打开 DebugView, 在 DebugView 中的 Capture 菜单中选中下列项:

Capture Global Win32

Capture Kernel

Enable Verbose Kernel Output(这个一定要选中)

Pass-Through

Capture Events

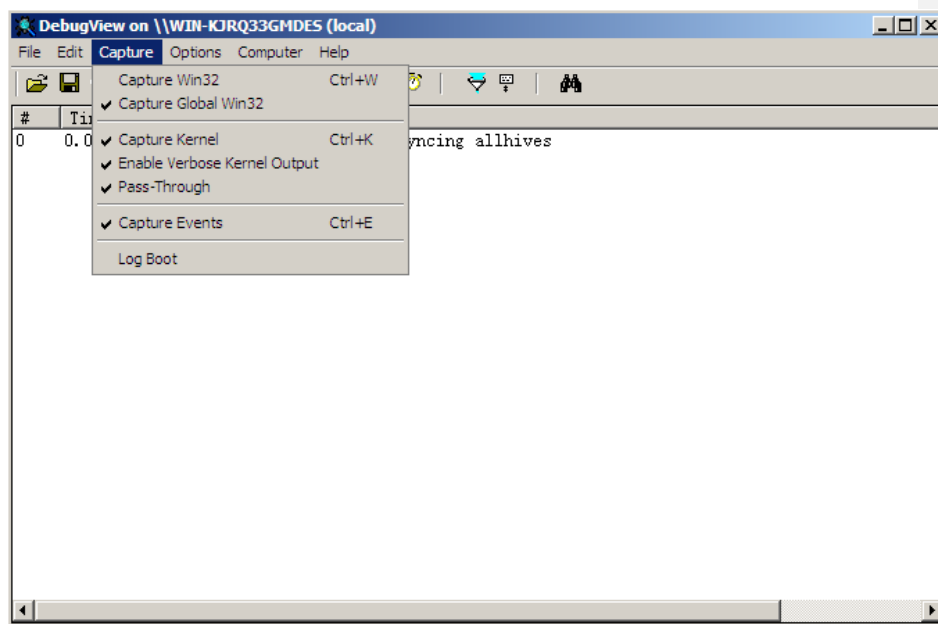
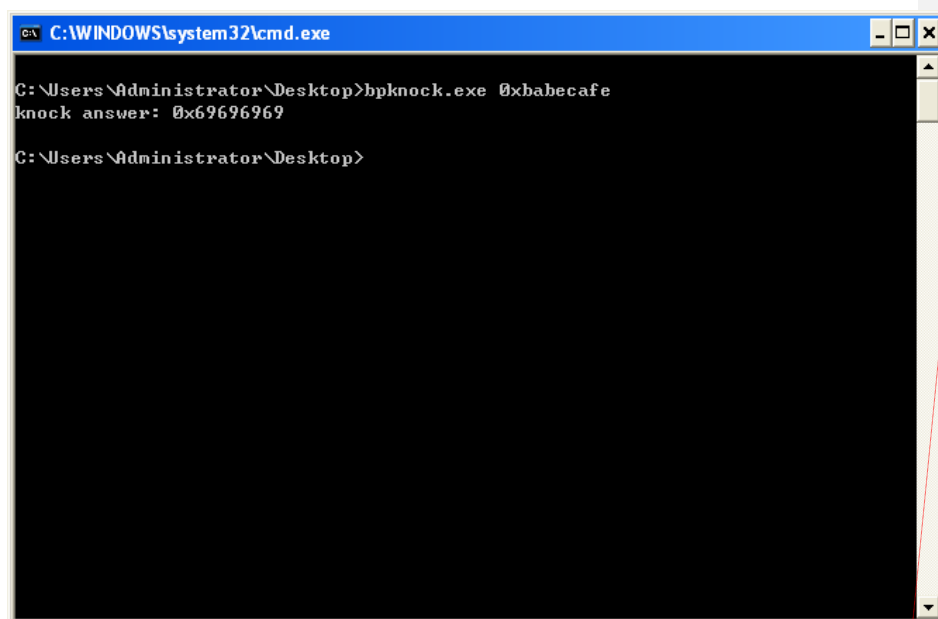


图 3.5 配置 DebugView

然后打开 InstDrv，先后加载并启动 dbgclient.sys 驱动和 newbp.sys 驱动¹

步骤 6：再运行下 bpknock 0xbabecafe 看下运行了 nbp 的输出结果。（如图 3.6 所示）



批注 [S16]: 以后再把这个换成 win2k8 下面的

图 3.6 加载 newbp 驱动的 bpknock 程序输出结果

¹ 这两个驱动分别在各自文件夹的 bin 子目录下

调试 NewBluePill

调试 NewBluePill 需要用到 WinDbg, 主要过程如下:

步骤 1. 为了调试过程中可以下断点 (切记做这一步只是为了以后能够调试, 并且使得 NewBluePill 驱动只能运行在操作系统的 debug 模式下), 修改 common 目录下的 newbp.c 文件, 在 DriverEntry 方法的一开始添加 CmDebugBreak() 方法调用¹ (如图 3.7 所示), 重新编译。修改后的代码如下:

```
00046: NTSTATUS DriverEntry (
00047:     PDRIVER_OBJECT DriverObject,
00048:     PUNICODE_STRING RegistryPath
00049: )
00050: {
00051:     NTSTATUS Status;
00052:     CmDebugBreak();
00053:     #ifdef USE_COM_PRINTS
00054:     PciInit ((PUCHAR) COM_PORT_ADDRESS);
00055:     #endif
00056:     ComInit ();
00057:     Status = MmInitManager ();
00058:     return Status;
}
```

图 3.7 添加 CmDebugBreak() 方法的位置

步骤 2: 参考 Debugging Windows Vista² 修改被调试机启动项和调试项 (这一步只需做这一次就可以)

步骤 3: 重启被调试机, 可以看到启动项中多了一个 DebugEntry [debugger enabled] 项, 选中它按 F8, 然后选择 Disable Driver Signature Enforcement 项启动

步骤 4: 调试机上设置 _NT_SYMBOL_PATH 环境变量, 指向 newbp.pdb 所在的目录, 用于链接符号表。

步骤 5: 调试机上启动 WinDbg, 单击 File 菜单选择 Kernel Debugging, 在弹出的对话框输入 Baud Rate 为 115200, Port 用 com1。³这是由于刚才在演示过程的第一步我们用的是默认配置, 如果调试端口发生相应改变, 这里也要改。

¹ CmDebugBreak() 函数实际上是一个 int 3 调用, 在非调试模式的 Windows 下, 这时这个中断的处理程序未注册, 因此执行 int 3 指令会死机。

² 文章来源: http://www.microsoft.com/whdc/driver/tips/debug_vista.msp

³ 除了利用串口线调试外, 也可以利用 1394 线进行调试, 这样做的好处是数据传输速度更快。具体方法可以参考网上相关资料。

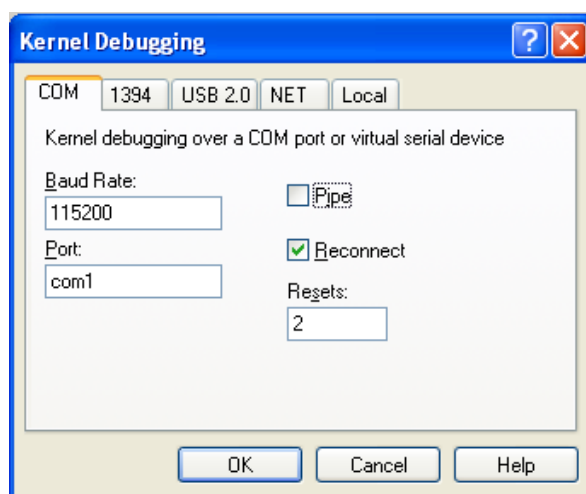


图 3.8 配置 Windbg

步骤 6: 被调试机上先后加载并启动 dbgclient.sys 和 newbp.sys 两个驱动, 运行 bpknock 程序, 开始调试。

如果出现 symbol 不能被加载的情况可以试试 WinDbg 中的 .reload 命令, 如果不行为可以试试用 .sympath 在 WinDbg 运行时设定 symbol 路径, 然后 .reload 重新加载符号表。

成功情况下的截图:

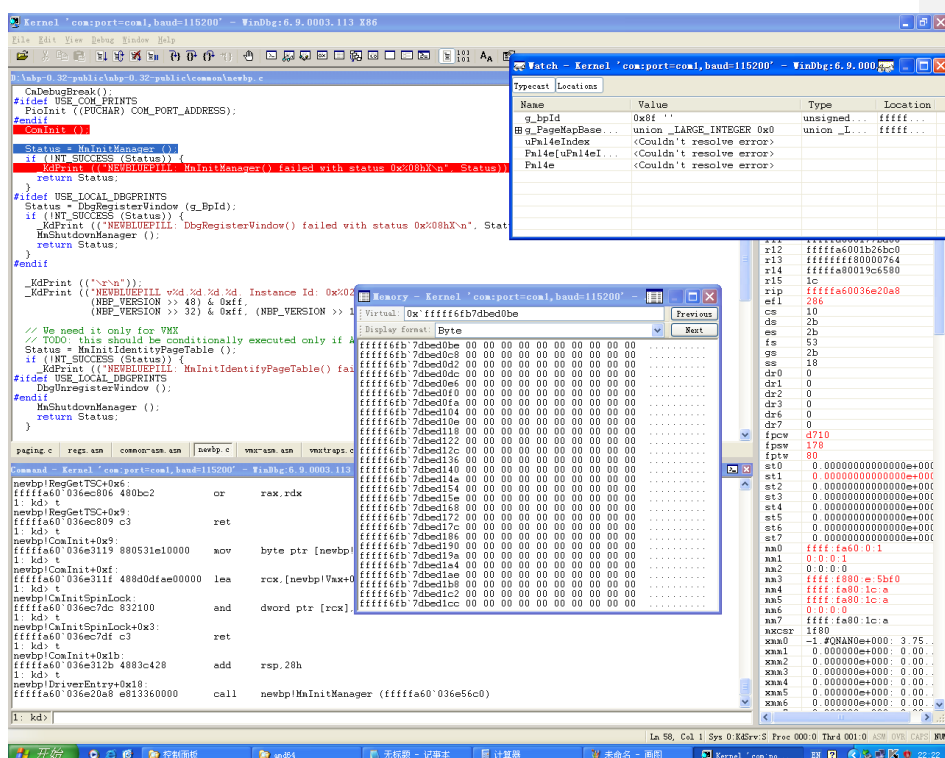


图 3.9 成功搭建调试平台

Ok，有了调试平台，我们就可以揭开 NewBluePill 的层层面纱了。

四、NewBluePill 的启动和卸载

在这一章中，我们将探究 NewBluePill 驱动的启动和关闭过程，从而将 NewBluePill 各组件串接起来（本章不涉及 dbgclient.sys 的启动过程，“第八章 NewBluePill 调试系统”会对其加以说明）。启动和卸载过程在 NewBluePill 中占据很大的比重，这一点可以从相关代码所占总代码量比重上看出：约 30% 的源文件均与启动和卸载过程有关。所以了解 NewBluePill 的启动和卸载，将对了解其功能实现有极大帮助。

在后续章节中，我们将逐一探索每个组件是如何完成其功能的。

NewBluePill 驱动的启动过程

NewBluePill 驱动入口在 common\newbp.c 文件中，入口函数为 DriverEntry 函数（Newbp.c 第 46 行）。这个函数流程图如图 4.1 所示：

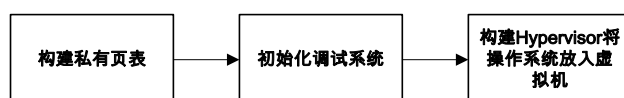


图 4.1 NewBluePill 启动流程图

下面我们将按照图 4.1 逐一介绍每部分运行过程。

构建私有页表

在 DriverEntry 函数中，从 58 行到 62 行，以及从 79 行到 114 行，都是在完成私有页表的构建。（对于内存相关部分，本章中我们只是列举出被调用的函数，每个函数的详细作用我们会在“第五章 NewBluePill 内存系统”中阐述）

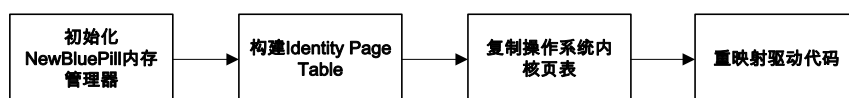


图 4.2 NewBluePill 初始化过程中构建私有页表的流程

- **初始化 NewBluePill 内存管理器** Newbp.c 中第 58 行到第 62 行，该代码调用函数 MmInitManager()，该函数会在内存中分配新空间作为 NewBluePill 自己的页表，并按照 x64 地址翻译机制构建出页表结构。
- **构建 Identity Page Table** Newbp.c 中第 79 行到第 87 行，不知道这个函数干什么用的
- **复制操作系统内核页表** Newbp.c 中第 89 行到第 97 行，该代码调用函数 MmMapGuestKernelPages()，该函数会根据当前 Windows 操作系统的内核页表内容，填充 NewBluePill 自己的页表。
- **重映射驱动代码** Newbp.c 中第 98 行到第 114 行，调用函数

批注 [S17]: 第八章 NewBluePill 其它系统
章号

批注 [S18]: “第五章 NewBluePill 内存系统”章号

批注 [S19]: 待调查后补充
TODO

MmMapGuestPages()。NewBluePill 作为驱动，必定要占据内核空间，此处调用该函数，就是要把自己占用的操作系统内核页面空间的页表信息复制到自己的页表中，从而为以后实现页表隐藏打下基础。

初始化调试系统

DriverEntry() 函数的 63 行到 75 行在初始化 NewBluePill 的调试系统。（对于调试系统部分，本章中我们只是列举出被调用的函数，每个函数的详细作用我们会在“第八章 NewBluePill 调试系统”中阐述）

NewBluePill 初始化本地调试窗口，是通过调用函数 DbgRegisterWindow() 实现的，这个函数主要作用是根据当前 NewBluePill 驱动实例的唯一 ID（驱动运行时读取处理器时间寄存器（Time Stamp Counter，TSC），并将低八位作为这个 ID），分配一段共享内存，并将打印信息全部保存在这段内存上，从而使得本机调试变得方便。（dbgclient.sys 会读取这段共享内存的内容并发送到调试机上，其实也可以调整下让它将这些内容保存在磁盘上）

NewBluePill 也直接支持利用串口发送调试信息到调试机上（不需 dbgclient.sys 的帮助）

构建 Hypervisor 并将操作系统放入虚拟机

构建 Hypervisor，并将操作系统放入虚拟机的工作，是在 DriverEntry 函数中的 116 行到 132 行完成的，主要调用的函数有两个：

- HvmInit() 函数
- HvmSwallowBluepill() 函数

HvmInit() 函数的作用是：确定当前系统架构是否支持 HEV 技术，并确定 NewBluePill 支持哪种 HEV 技术（Intel VT/AMD SVM）。最后根据获得的信息，将相应的处理函数组和平台信息捆绑在 Hvm 结构体上，该结构体可以通过在 windbg 下输入 dt Hvm 命令实现。

实验：查看 Hvm 结构体

在 NewBluePill 运行时，您可以在 windbg 下使用 dt Hvm 命令查看当前平台对应的 Hvm 结构体内容：

Lkd

批注 [S20]: 给出该实验的实验结果

实际上，检查是否支持 HEV 技术，和确定平台的函数（两者都由 Hvm->ArchIsHvmImplemented() 实现）在后面的 HvmSubvertCpu() 函数中（被 HvmSwallowBluepill() 调用，后面会讲到）再次出现，我们认为重复检查是不必要的。

HvmSwallowBluepill() 函数的作用是：该函数及其子函数给每个处理器（Processor）安装 NewBluePill 的 Hypervisor，这个函数也是 NewBluePill 主要逻辑的初始化入口。下面，我们就将进入这个入口背后的世界。

进入 NewBluePill 的世界

当我们进入了 `HvmSwallowBluePill()` 函数,也就踏入了 NewBluePill 的世界。为了便于理解,我们人为的把启动过程分为两个阶段,进入虚拟机模式前的初始化部分称为阶段 1 初始化 (Phase 1),进入虚拟机模式后的初始化部分称为阶段 2 初始化 (Phase 2)。

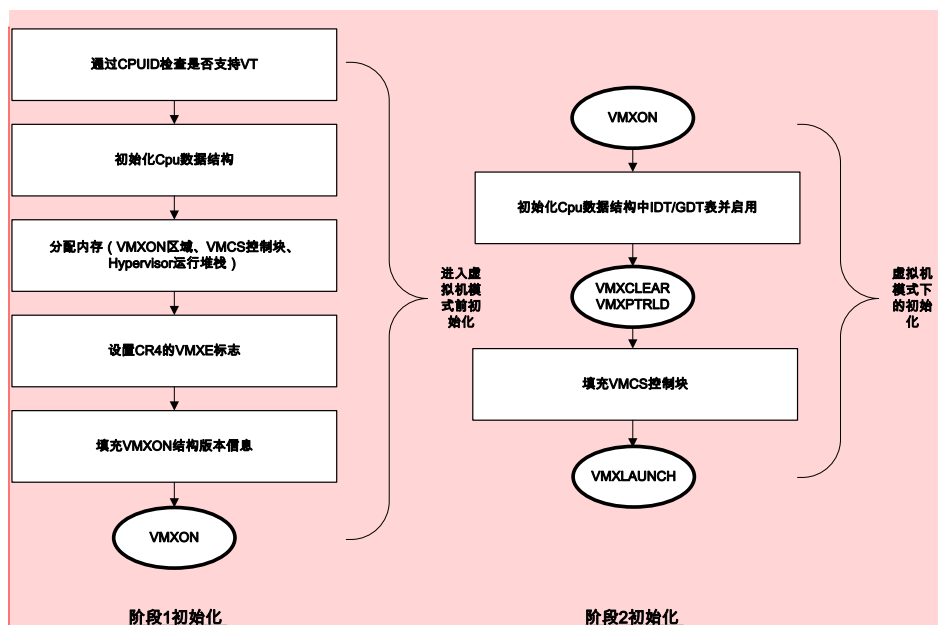


图 4.3 阶段 1 和阶段 2 初始化流程图

批注 [S21]: 此图应该分为 VT 和 SVM 两者
分别的初始化过程图

阶段 1 初始化

阶段 1 初始化主要任务是为每个处理器开启虚拟机模式。

`HvmSwallowBluePill()` 函数首先会在操作系统范围内申请一个互斥锁 `g_HvmMutex` (`common\Hvm.c` 第 556 行),当多个 NewBluePill 驱动同时启动/卸载时,该锁能够保证每个时刻只能有一个 NewBluePill 实例运行,以避免嵌套设置失败。随后, `HvmSwallowBluePill()` 函数遍历每个处理器 (`common\Hvm.c` 第 558 行~581 行)调用 `CmDeliverToProcessor()` 函数,进而通过回调手段调用 `HvmSubvertCpu()` 函数,后者用来指导对 VT/SVM 两种平台执行同样的侵染过程。

`HvmSwallowBluePill()` 函数中对每个处理器遍历的代码如图 4.4 所示,注意这个模式所采用的手法,遍历内部传入了回调函数,并且既要处理 `CmDeliverToProcessor()` 的运行结果,又要处理回调函数的运行结果。该模式在安装和卸载过程中均有出现。¹

¹ 在这段代码中一个小的 Bug 是 for 循环应该使用 `KeQueryActiveProcessors()` 函数而不是 `KeNumberProcessors` 变量来获得当前处理器个数,这个 bug 会影响到支持热插拔 CPU 的系统,后果是可能导致在这种系统上某些处理器未被侵染,或重复侵染而导致内存泄露、系统异常等其他问题。

```

00558:   for (cProcessorNumber = 0; cProcessorNumber < KeNumberProcessors; cProcessorNumber++) {
00559:
00560:       _KdPrint (("HvmSwallowBluepill(): Subverting processor #td\n", cProcessorNumber));
00561:
00562:       Status = CmDeliverToProcessor (cProcessorNumber, CmSubvert, NULL, &CallbackStatus);
00563:
00564:       if (!NT_SUCCESS (Status)) {
00565:           _KdPrint (("HvmSwallowBluepill(): CmDeliverToProcessor() failed with status 0x%08hX\n", Status));
00566:           KeReleaseMutex (&g_HvmMutex, FALSE);
00567:
00568:           HvmSpitOutBluepill ();
00569:
00570:           return Status;
00571:       }
00572:
00573:       if (!NT_SUCCESS (CallbackStatus)) {
00574:           _KdPrint (("HvmSwallowBluepill(): HvmSubvertCpu() failed with status 0x%08hX\n", CallbackStatus));
00575:           KeReleaseMutex (&g_HvmMutex, FALSE);
00576:
00577:           HvmSpitOutBluepill ();
00578:
00579:           return CallbackStatus;
00580:       }
00581:   } ? end for cProcessorNumber=0;cP... ?
00582:
00583:   KeReleaseMutex (&g_HvmMutex, FALSE);
00584:
00585:   if (KeNumberProcessors != g_uSubvertedCPUs) {
00586:       HvmSpitOutBluepill ();
00587:       return STATUS_UNSUCCESSFUL;
00588:   }

```

图 4.4 HvmSwallowBluePill() 函数中对每个处理器遍历的代码

CmDeliverToProcessor() 函数用于提高中断优先级 (DISPATCH_LEVEL IRQL)，并在指定的处理器上执行一个函数，从而可以不被打断¹的在该处理器上执行这个函数，同时在执行最后会恢复到原先执行该指令的处理器组设置上（也称亲核性，CPU Affinity，参见 Common.c 代码第 359 行）。在安装过程中，通过它来执行 HvmSubvertCpu() 函数，从而保证 HvmSubvertCpu() 函数的运行都作用于指定的处理器。

Important 在操作系统中断优先级高于等于 DISPATCH_LEVEL 的情况下，不能使用可分页的内存，因为访问一个已换出页的内存地址会发生一个内存页换入 (swap-in) 操作，而该操作是在较低的中断优先级（确切的说：APC_LEVEL）进行的。这种情况下，操作系统会蓝屏。

正确的做法是：如果确实要在 DISPATCH_LEVEL 或更高的中断优先级下分配内存，则要从操作系统的不分页内存池 (Nonpaged pool) 中分配内存，该段空间在操作系统存活期间内驻留在内存。

我们可以通过研究 MmAllocatePages() 函数 (common\Paging.c 第 275 行) 和其调用者出现的逻辑位置看到 NewBluePill 驱动是怎样注意这个问题的。

HvmSubvertCpu() 函数主要做了两件事情：

- 1) 创建并初始化 CPU 数据结构 (Hvm.c 中第 365 行到 418 行，453 行到 459 行，后者用于构建 NewBluePill 自己的 GDT、IDT 表，并在虚拟机模式中启用)。
- 2) 指导怎样侵染一个核。(Hvm.c 中第 420 行到 465 行，不包括 453 行到 459 行)

CPU 数据结构定义在 Hvm.h 文件的第 53 行，该数据结构在阶段 1 中初始化，阶段 2 中被使用。查找陷入处理函数、页表隐藏以及实现 Blue Chicken 反虚拟机探测技术都需要这个结构体的帮助，可以说，这个结构体是继 VMCS/VMCB 之后第二重要的结构体。下面先详细解释下这个数据结构

```

typedef struct _CPU
{

```

¹ 确切地说还是会被打断的，比如时钟中断就可以打断 NewBluePill 的运行，但是这些打断不会干扰到 NewBluePill 的运行。

```

PCPU      SelfPointer;      /*必须放在第一个，与处理事件逻辑有关*/

union
{
    SVM      Svm;
    VMX      Vmx;
};
/*标示所运行在哪种平台上 SVM/VT,里面含有构建该平台下 Hypervisor 和虚拟机的关键信息*/

```

```

ULONG      ProcessorNumber; /*该处理器在操作系统中的处理器序号*/
ULONG64     TotalTscOffset; /* */

```

批注 [S22]: 调查时间迷惑后补充

```

LARGE_INTEGER LpicBaseMsr; /*实际上未被用到*/
PHYSICAL_ADDRESS LpicPhysicalBase; /*实际上未被用到*/
PUCHAR        LpicVirtualBase; /*实际上未被用到*/

```

```

LIST_ENTRY GeneralTrapsList; /*该处理函数列表用于处理一般原因陷入的 VMEXIT 事件，比如对 cr0-cr4 寄存器的操作*/

```

```

LIST_ENTRY MsrTrapsList; /*该处理函数列表用于处理操作 MSR 寄存器而产生的 VMEXIT 事件，比如 rdmsr 指令和 wrmsr 指令*/

```

```

LIST_ENTRY IoTrapsList; /*该处理函数列表用于处理因 I/O 操作而产生的 VMEXIT 事件，实际上没有用到1*/

```

```

PVOID      SparePage; /*存储一个空页面的引用，用于后面的页表隐藏操作*/

```

```

PHYSICAL_ADDRESS SparePagePA; /*SparePage 原来的物理地址*/
PULONG64 SparePagePTE;

```

批注 [S23]: 调查后补充

```

PSEGMENT_DESCRIPTOR GdtArea; /*NewBluePill 自己的 GDT 表空间*/
PVOID      IdtArea; /*NewBluePill 自己的 IDT 表空间*/

```

批注 [S24]: 只是一个无用的 page，用于映射自己要访问的物理地址。如果要访问某个物理地址，只需填充 PTE 后使用 SparePage 访问即可。

```

PVOID      HostStack; /*NewBluePill Hypervisor 在该处理器上的运行堆栈空间，16 页大小，这段空间顶端放置了 CPU 结构体*/
BOOLEAN     Nested; /*是否是嵌套 NewBluePill*/

```

```

#ifdef INTERCEPT_RDTSCs

```

```

// variables for RDTSC tracing and cheating
ULONG64     Tsc; /*当因为 RDTSC 指令陷入时，要返回的虚假时间*/
ULONG64     LastTsc; /*上次因为 RDTSC 指令陷入返回的虚假

```

¹ 可能算一个 bug，当前确实有 I/O 造成的 VMEXIT 事件的处理函数，但是相关信息却并未放到 IoTrapsList 链表中。不过影响不大。MsrTrapsList 也有相似的问题，只有 NewBluePill 对 SVM 技术支持的实现中用到了这个链表。

```

时间*/
    ULONG64    EmulatedCycles;        /* 两次 RDTSC 指令陷入间,
NewBluePill 积累的时间欺骗量*/
    int        Tracing;                /* 两次 RDTSC 指令陷入间,
NewBluePill 还能记录欺骗时间的指令数 */
    int        NoOfRecordedInstructions; /*两次 RDTSC 指令陷入间,
NewBluePill 已经记录欺骗时间的指令数 */

#endif
#ifdef BLUE_CHICKEN

    int        ChickenQueueSize;       /*用于记录ChickenQueueTable 当前大小*/
    ULONG64    ChickenQueueTable[CHICKEN_QUEUE_SZ]; /*用于记录每次陷入 Hypervisor 时的时间寄存器值*/
    int        ChickenQueueHead, ChickenQueueTail; /* 指向 ChickenQueueTable 头尾元素位置*/

    UCHAR      OriginalTrampoline[0x600]; /*NewBluePill 卸载时,用于构造弹簧床代码的内存区域*/

#endif

    ULONG64    ComPrintLastTsc; /*上次 Com 口输出时间,调试系统会利用这个值*/

} CPU,
*PCPU;

```

实验：查看 Cpu 结构体

在 NewBluePill 调试状态下，您可以在 windbg 下，通过 bp 命令设置断点，使得被调试机系统停在 NewBluePill 包含 Cpu 结构的函数内，然后使用 dt Cpu 命令查看 Cpu 结构体内容。比如加载 NewBluePill 驱动后，当驱动停在 CmDebugBreak() 后，在 windbg 中输入 bp VmxDispatchCpuId，然后按 F5。当驱动停在 VmxDispatchCpuId() 后，输入 dt Cpu:

Lkd

批注 [S25]: 给出该实验的实验结果

有一点需要注意，那就是在 HvmSubvertCpu() 函数中处理 SparePage 的地方（hvm.c 第 414 行），在为 Cpu 结构体的 SparePagePTE 赋值的时候：

```
Cpu->SparePagePTE = (PULONG64)((((ULONG64)(Cpu->SparePage) >> 9) & 0x7fffffff8)+PT_BASE);
```

这么做是因为，

批注 [S26]: Cpu->SparePagePTE 的问题

关于页表项状态位

在 Hvm.c 的 417 行:

```
*Cpu->SparePagePTE |= (1 << 4);
```

这条语句是在设置其所指向的 PTE 是否禁用 Cache。关于硬件页表项各状态位的含义,可以参考 *Windows Internals, 4th Edition* 的 Chapter 7. Memory Management 中的 Address Translation 部分中的相关内容,或者 Intel/AMD 手册中的相关部分。

HvmSubvertCpu() 函数还指导了 NewBluePill 如何侵染一个核。过程按先后顺序分为三步:

- Hvm->ArchRegisterTraps(Cpu) 注册 VMEXIT 事件处理函数
- Hvm->ArchInitialize(Cpu, CmSlipIntoMatrix, GuestRsp)
开启虚拟机模式,成为 Hypervisor 并构建虚拟机以装入原来的操作系统
- Hvm->ArchVirtualize(Cpu) 开启虚拟机

可以看到,实际上 Hvm 结构体起到了硬件抽象的作用,对上层屏蔽了 SVM 和 VT 技术两者的差异,提高了代码的复用性。下面,我们将深入 NewBluePill 适应于每个技术的具体实现,查看阶段 1 是如何完成的。

SVM 技术下阶段 1 的初始化

在 NewBluePill 对 SVM 技术的支持中,HvmSubvertCpu() 函数中 Hvm 结构体的三个函数对应关系如下:

表 4.1 Hvm 结构体中三函数对应关系 (AMD 平台)

| Hvm 结构体中函数 | 映射函数 (svm\Svm.c 和 svm\Svmtraps.c) |
|--------------------------|-----------------------------------|
| Hvm->ArchRegisterTraps() | SvmRegisterTraps() |
| Hvm->ArchInitialize() | SvmInitialize() |
| Hvm->ArchVirtualize() | SvmVirtualize() |

首先是 SvmRegisterTraps() 函数

批注 [S27]: 对 SVM 技术的支持随后再写

VT 技术下阶段 1 的初始化

在 NewBluePill 对 VT 技术的支持中,HvmSubvertCpu() 函数中 Hvm 结构体的三个函数对应关系如下:

表 4.2 Hvm 结构体中三函数对应关系 (AMD 平台)

| Hvm 结构体中函数 | 映射函数 (vmx\Vmx.c 和 vmx\Vmxtraps.c) |
|--------------------------|-----------------------------------|
| Hvm->ArchRegisterTraps() | VmxRegisterTraps() |
| Hvm->ArchInitialize() | VmxInitialize() |

Hvm->ArchVirtualize()

VmxVirtualize()¹

VmxRegisterTraps 函数

首先是 VmxRegisterTraps() 函数，该函数通过调用 TrInitializeGeneralTrap() 函数将各事件与处理函数捆绑起来成为一个新的 Trap 元素，并通过调用 TrRegisterTrap() 函数将这个 Trap 存入 Cpu 结构的 GeneralTrapsList 链表中²。同时，由于 MSR 操作、I/O 操作、和其他通用操作间的不同，所以在 Trap 元素中又使用了另外的类型来存储三者 Trap 处理特定的信息。Trap 元素及其相关 NBP_TRAP_DATA_GENERAL、NBP_TRAP_DATA_MSR、NBP_TRAP_DATA_IO 结构体定义如下：

```
typedef struct _NBP_TRAP
{
    /*_NBP_TRAP 结构体存储了 Trap 类型, Trap 发生原因, Trap 处理函数等一系列的信息,
    用于在发生该类型 Trap 时搜索适当的处理函数*/
    LIST_ENTRY    le;          /*Windows 系统链表元素必备域, 且必须在头部*/

    TRAP_TYPE     TrapType;    /*记录该 Trap 的类型, 可以是 TRAP_DISABLED、
    TRAP_GENERAL、TRAP_MSR、TRAP_IO 之一*/
    TRAP_TYPE     SavedTrapType; /*用于在该 Trap 元素被 Disable 时, 备份
    TrapType, 以供在该元素被 Enable 时恢复*/

    union
    {
        NBP_TRAP_DATA_GENERAL    General;
        NBP_TRAP_DATA_MSR        Msr;
        NBP_TRAP_DATA_IO         Io;
    };                          /*用于该 Trap 元素记录附加信息, 详细内容参见这三个结构体注释*/

    NBP_TRAP_CALLBACK    TrapCallback;    /*记录若该 Trap 元素被选中,
    那么应该调用的事件处理函数*/
    BOOLEAN               bForwardTrapToGuest; /*记录是否应该把这个事件继续上传,
    在处理 MSR 造成的 VMEXIT 时会用到*/

} NBP_TRAP,
*PNBP_TRAP;
```

Trap 结构体定义 (Traps.h 中第 67 行到 87 行)

```
typedef struct _NBP_TRAP_DATA_GENERAL
{
```

¹ 该函数在阶段 2 初始化过程中被用到，因此不在本节中介绍。

² NewBluePill 对 VT 技术的支持实现中，所有的事件处理的 Trap 元素都放到了 Cpu->GeneralTrapsList 链表中，虽然与 VT 技术特性有关，但也应该算是一个 bug，不过不影响当前运行。

```

    ULONG    TrappedVmExit;          /*记录该 Trap 元素对于什么原因造成的
VMEXIT 事件进行处理, 该域在查找过程中用作键*/
    ULONG64   RipDelta;              /*这个值用于 NewBluePill Hypervisor
处理完 VMEXIT 事件返回虚拟机后, Guest_Rip 上要加的字节数以跳过触发 VMEXIT 事件的
指令, 如果为 0 则说明需要在陷入 Hypervisor 后, 从 VMCS 中获得该值*/
} NBP_TRAP_DATA_GENERAL,
*PNBP_TRAP_DATA_GENERAL;

```

NBP_TRAP_DATA_GENERAL 结构体定义 (Traps.h 中第 36 行到 41 行)

```

typedef struct _NBP_TRAP_DATA_MSR
{
    ULONG32    TrappedMsr;
    UCHAR      TrappedMsrAccess;
    UCHAR      GuestTrappedMsrAccess;
} NBP_TRAP_DATA_MSR,
*PNBP_TRAP_DATA_MSR;

```

NBP_TRAP_DATA_MSR 结构体定义 (Traps.h 中第 43 行到 49 行)

```

typedef struct _NBP_TRAP_DATA_IO
{
    ULONG TrappedPort;
} NBP_TRAP_DATA_IO,
*PNBP_TRAP_DATA_IO;

```

NBP_TRAP_DATA_IO 结构体定义 (Traps.h 中第 51 行到 55 行)

结合“第二章 深入 HEV 技术细节”中的内容我们可以看到, 在 `VmxRegisterTraps()` 函数中, NewBluePill Hypervisor 对所有的无条件产生 VMEXIT 事件的汇编指令都添加了相应的处理函数, 如表 4.3 所示。

批注 [S28]: 看完 AMD 后补充

批注 [S29]: “第二章 深入 HEV 技术细节”章号

表 4.3 VT 下 NewBluePill Hypervisor 对无条件陷入绑定的处理函数

| 造成无条件陷入的指令 | 对应硬件 VMEXIT 原因标记 ¹ | 处理函数 (vmx\Vmxttraps.c) |
|------------|----------------------------------|--------------------------|
| VMCALL | EXIT_REASON_VMCALL | VmxDispatchVmxInstrDummy |
| VMCLEAR | EXIT_REASON_VMCLEAR ² | VmxDispatchVmxInstrDummy |
| VMLAUNCH | EXIT_REASON_VMLAUNCH | VmxDispatchVmxInstrDummy |
| VMRESUME | EXIT_REASON_VMRESUME | VmxDispatchVmxInstrDummy |
| VMPTRLD | EXIT_REASON_VMPTRLD | VmxDispatchVmxInstrDummy |
| VMPTRST | EXIT_REASON_VM PTRST | VmxDispatchVmxInstrDummy |
| VMREAD | EXIT_REASON_VMREAD | VmxDispatchVmxInstrDummy |

¹ 所有这些标记定义 Vmx.h 中, 具体这些定义为什么这样取值, 请参考 *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 3B* 手册 Appendix I: VMX Basic Exit Reasons

² 在源代码中有一个 bug, 那就是在 Vmxttraps.c 的第 518 行和 519 行两个都是 EXIT_REASON_VMCALL, 其中一个应该是 EXIT_REASON_VMCLEAR。

| | | |
|---------|---------------------|--------------------------|
| VMWRITE | EXIT_REASON_VMWRITE | VmxDispatchVmxInstrDummy |
| VMXON | EXIT_REASON_VMXON | VmxDispatchVmxInstrDummy |
| VMXOFF | EXIT_REASON_VMXOFF | VmxDispatchVmxInstrDummy |
| CPUID | EXIT_REASON_CPUID | VmxDispatchCpuid |
| INVD | EXIT_REASON_INVD | VmxDispatchINVD |

此外，也对有条件陷入的指令添加了处理函数，表 4.4 列举了它们：

表 4.4 VT 下 NewBluePill Hypervisor 对有条件陷入绑定的处理函数

| 造成有条件陷入的指令 | 对应硬件 VMEXIT 原因标记 ¹ | 处理函数 (vmx\Vmxttraps.c) |
|---------------------|--|------------------------|
| RDMSR | EXIT_REASON_MSR_READ | VmxDispatchMsrRead |
| WRMSR | EXIT_REASON_MSR_WRITE | VmxDispatchMsrWrite |
| CR Ops | EXIT_REASON_CR_ACCESS | VmxDispatchCrAccess |
| 异常 | EXIT_REASON_EXCEPTION_NMI ² | VmxDispatchException |
| RDTSC | EXIT_REASON_RDTSC | VmxDispatchRdtsc |
| I/O 操作 ³ | EXIT_REASON_IO_INSTRUCTION | VmxDispatchIoAccess |

关于这些处理函数的功能，我们会在“第六章 NewBluePill 陷入事件管理系统”进行详细讨论。

批注 [S30]: “第六章 NewBluePill 陷入事件管理系统” 章号

VmxInitialize 函数

在注册完各陷入事件处理函数后，NewBluePill 会调用 VmxInitialize() 函数进行后续的阶段 1 初始化工作。

```
typedef struct _VMX
{
    PHYSICAL_ADDRESS    VmcsToContinuePA;    // MUST go first in the
    structure; refer to SvmVmrun() for details
    PVOID                _2mbVmcbMap;

    PHYSICAL_ADDRESS    OriginalVmcsPA;        /*VT 技术中用于控制虚拟机的
    VMCS 结构体物理地址*/
    PVOID                OriginalVmcs;        /*VMCS 结构体指针*/
    PHYSICAL_ADDRESS    OriginalVmxonRPA;    /*为了开启虚拟机环境所需的
    VMXON 区域的物理地址*/
}
```

批注 [S31]: 未注释完全，待完全写好启动过程后看看有没有什么新发现

¹ 所有这些标记定义 Vmx.h 中，具体这些定义为什么这样取值，请参考 *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 3B* 手册 Appendix I: VMX Basic Exit Reasons

² Intel 手册中说明，该值 (0) 既可以表示虚拟机由于异常而产生 VMEXIT 事件，也可以表示虚拟机由于 NMI 异常而产生 VMEXIT 事件。在 NewBluePill 中，根据 VMCS 块的配置，该值代表发生异常的陷入原因。

³ 注意到在 Vmxtrap.c 第 589 行，只有在头文件中预先定义 VMX_ENABLE_PS2_KBD_SNIFFER，NewBluePill Hypervisor 才会处理访问 I/O 造成的陷入。

```

    PVOID                OriginaVmxonR;        /*该指针指向 VMXON 区域*/

    PHYSICAL_ADDRESS      IOBitmapAPA;        /* 下面四项分别指向两段 I/O
Bitmap, 在 VMCS 的设置中使用*/
    PVOID                IOBitmapA;

    PHYSICAL_ADDRESS      IOBitmapBPA;
    PVOID                IOBitmapB;

    PHYSICAL_ADDRESS      MSRBitmapPA;        /*指向 MSR Bitmap, 在 VMCS 的
设置中被使用*/
    PVOID                MSRBitmap;

    ULONG64               GuestCR0;           /*用于缓存虚拟机中 CR0 控制寄存
器的新值, 主要用于 CR0 寄存器造成的陷入的处理函数中。在 VmxInitialize() 也被赋
值*/
    ULONG64               GuestCR3;           /*用于缓存虚拟机中 CR3 控制寄存
器的新值, 主要用于 CR3 寄存器造成的陷入的处理函数中。当虚拟机禁用分页时, 存储虚
拟机 CR3。在 VmxInitialize() 也被赋值*/
    ULONG64               GuestCR4;           /*用于缓存虚拟机中 CR4 控制寄存
器的新值, 主要用于 CR4 寄存器造成的陷入的处理函数中。在 VmxInitialize() 也被赋
值*/
    ULONG64               GuestEFER;          /*用于缓存虚拟机中 EFER MSR 寄
存器1的值, 在处理 MSR 和控制寄存器操作造成的陷入时均被用到。在 VmxInitialize()
也被赋值*/
    UCHAR                GuestStateBeforeInterrupt[0xc00];

} VMX,
*PVMX;

```

VMX 结构体定义 (vmx\Vmxc.h 中第 205 行到 231 行)

VmxInitialize() 函数主要职责是初始化 Cpu 结构体中的 Cpu->Vmx 项, 这之中最重要的莫过于分配 Vmxon 空间 (Vmxon Regions) (Vmxc.c 文件第 519 行)、VMCS 空间 (Vmxc.c 文件第 530 行)、IOBitmap (I/O 位图, Vmxc.c 文件第 544 行和 554 行) 和 MsrBitmap (Msr 位图, Vmxc.c 文件第 564 行)。具体这四个空间的作用可参考“第二章 深入 HEV 技术细节”中相关内容描述。

当这些结构体都被分配出来后, 阶段 1 的初始化工作也将要结束。在 VmxInitialize() 函数的第 573 行, 程序调用 VmxEnable() 函数。

VmxEnable() 函数中每个有关设置的步骤都很关键。首先设置 CR4 寄存器第 13 位为 1 (该位也称 CR4.VMXE, 用于执行 VMXON 开启虚拟机模式的前提) (Vmxc.c 文件第 46 行), 然后为了保险起见, 检查 CR4 和 IA32_FEATURE_CONTROL, 确保当前系统支持开启虚拟机模

¹ EFER MSR 寄存器 (Extended Feature Enable MSR), 这个寄存器包括了一些该处理器是否支持以及正在使用一些扩展特性的信息, 比如是否支持 x64 的信息就存储在其中。具体信息可参考 *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 3A* 和网上相关资料。

批注 [532]: “第二章 深入 HEV 技术细节” 章号

式。接下来在 `Vmxon` 区域中填充 VMCS 版本号 (`Vmx.c` 文件第 58 行到第 59 行)。最后通过调用 `VmxTurnOn()` 函数 (实际就是 `vmxon` 汇编指令) 开启虚拟机模式。

至此 VT 技术下阶段 1 的初始化工作全部完成。

阶段 2 初始化

进入虚拟机模式后, NewBluePill 开始阶段 2 的初始化工作, 由于 SVM 和 VT 技术存在一些不同, 所以我们会看到 NewBluePill 对两种技术在阶段 2 的初始化的实现也存在一定差异。

SVM 技术下阶段 2 的初始化

VT 技术下阶段 2 的初始化

NewBluePill 在执行完 `VMXON` 指令后返回到 `VmxInitialize()` 函数中继续运行, 该函数接下来的工作就是:

1. 初始化 VMCS 块 (`Vmx.c` 文件第 578 行到 584 行)
2. 填充虚拟机初始状态 (`Vmx.c` 文件第 588 行到 600 行)

通过阅读第二章 VMCS 结构体详细介绍, 我们可以发现 NewBluePill 遵照了 VMCS 的初始化方法, 首先在第 578 行为 VMCS 结构体的头部赋予版本号信息, 并且赋值内容参考了 `MSR_IA32_VMX_BASIC` MSR 寄存器, 紧接着, NewBluePill 调用了 `VmxSetupVMCS()` 函数进一步设置 VMCS 结构体。

`VmxSetupVMCS()` 函数主要功能是通过设置 VMCS 块来配置 NewBluePill Hypervisor。函数首先设置各个段选择子 (`Vmx.c` 文件第 282 行到 295 行), 随后配置 I/O 位图和 MSR 位图, 可以看到, 这两部分位图在此处只是在 VMCS 中挂载了地址, 而没有进行进一步的配置 (在定义了监听键盘事件的情况下, I/O 位图相应项被配置)。紧接着在第 316 行, NewBluePill 依次初始化了 VMCS 块中的时间戳、VMCS Link Pointer (置为空, 后面不会用到)、虚拟机调试控制寄存器等多项。

随后, 根据需求初始化两个重要的域: 基于针脚的虚拟机执行控制 (Pin-Based VM-Execution Controls) 和基于处理器的虚拟机执行控制 (Processor-Based VM-Execution Controls)。前者赋值为 0, 说明屏蔽了由外部中断、NMI 中断、虚 NMI 和 VMX 抢占计时器造成的 #VMEXIT 事件, 后者根据是否要使用 MSR 位图、监听键盘和监控 RDTSC 指令进行相应配置。

第 347 行到 349 行对是否监控页面异常进行配置, 这里的配置是所有的页面异常都不会产生 VMEXIT 事件。

关于页面异常 (Page Fault) 产生 VMEXIT 事件

在 Intel 的手册中提到, 页面异常在一定条件下可以产生 VMEXIT 事件。

当一个页面异常发生后, 逻辑处理器首先用页面错误号 (Page Fault Error Code, PFEC) 和 VMCS 结构体中的页面错误号掩码 (PFEC Mask) 进行与操作, 如果结果与页面错误号匹

配值（PFEC Match）相等，则检查异常位图（Exception Bitmap）中 Bit 14 是否为 1，如果是则发生 VMEXIT 事件；如果结果与页面错误匹配号不相等，则如果异常位图中 Bit 14 为 0 的话，同样也会发生 VMEXIT 事件。

因此，如果想要监控所有页面异常造成的 VMEXIT 事件，可以把异常位图 Bit 14 置 1，然后将页面错误号掩码和页面错误号匹配值置 0，反之如果不想监控页面异常，那么可以在异常位图 Bit 14 置 1 的情况下，将页面错误号掩码置 0，页面错误号匹配值置为 FFFFFFFFH。

随后在第 351 行到第 358 行，NewBluePill 配置为如果是外部中断造成的 VMEXIT 事件，那么这个中断原因会被保存在 VMEXIT 退出信息区中，同时根据当前平台决定所运行的虚拟机是虚拟 x86 平台还是虚拟 x64 平台。

从 374 行到 380 行，NewBluePill 继续初始化虚拟机，填充包括 IDT 表长，GDT 表长等信息。从 384 行到 397 行，NewBluePill 则继续初始化 Hypervisor，对 CR0、CR4 和 CR3 做处理。后面从 403 行到 418 行，NewBluePill 填充虚拟机的 CR0、CR3、CR4 和段选择子，设置好 Debug 寄存器后，在第 452 行到第 457 行，填充虚拟机将要运行的指令地址和堆栈地址，以及为 sysenter 指令入口填充地址。最后，NewBluePill 从第 486 行到第 494 行设置 Hypervisor 处理 sysenter 指令入口地址（其实就是虚拟机的 sysenter 处理地址，没有用到这个域），并设置好 VMEXIT 事件发生后进入 Hypervisor 的指令地址，也可以看作是 VMEXIT 事件的统一处理入口，这个入口指向 VmxVmexitHandler()¹ 函数，并且设置处理函数用堆栈，注意其地址是 Cpu 结构体地址，这是故意这样设置的，因为在处理过程中会利用这样的栈设置操作 Cpu 结构体²。至此 VMCS 初始化完毕。

VMCS 结构体设置完毕后，NewBluePill 的初始化工作也将要结束。HvmSubvert() 函数还会调用 VmxVirtualize() 函数，这个函数的唯一作用就是执行 VMXLaunch 指令³，正式启动虚拟机（Vmx.c 第 768 行），要注意的是这个函数不应该返回，因此如果返回则一定表明 STATUS_UNSUCCESSFUL。

至此，NewBluePill 的初始化过程全部完成。

NewBluePill 驱动的卸载过程

研究完启动过程后，接下来我们看下 NewBluePill 的卸载过程。卸载代码的入口是 DriverUnload() 函数（Newbp.c 第 20 行），与启动过程对应的，卸载过程流程如下：

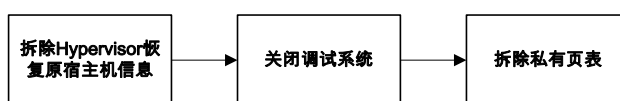


图 4.5 NewBluePill 卸载过程流程图

下面我们将按照图 4.5 逐一介绍每部分运行过程。

¹ 该函数及其功能会在“NewBluePill 陷入事件管理系统”一章中进行描述。

² 具体原因可参考网上关于函数调用规范的资料。

³ 在 x86 平台上还会将 Cpu 结构体放置在 VMCS 中 Host_Stack 指定的位置。

拆除 Hypervisor 恢复原宿主机信息

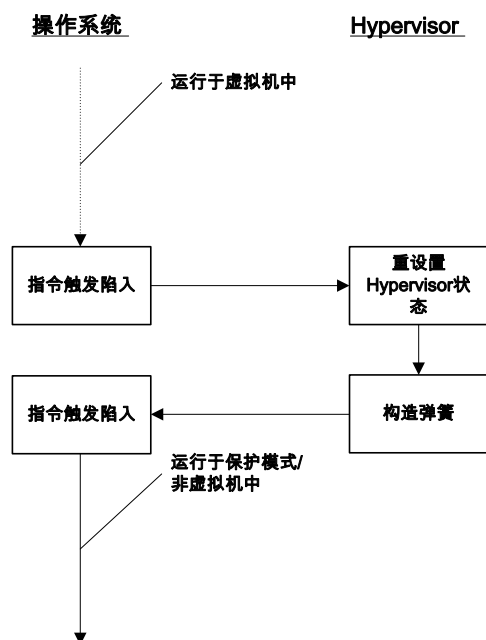


图 4.6 NewBluePill 卸载 Hypervisor 及恢复宿主机信息流程图

NewBluePill 卸载 Hypervisor 及恢复操作系统原先运行环境流程如图 4.6 所示。首先通过调用 `HvmSpitOutBluepill()` 函数（Newbp.c 第 31 行）来执行拆除 Hypervisor，并把操作系统放回原保护模式的工作，这项工作也是整个卸载过程中要做的主要工作。与启动过程不同的是，无论是 VT 技术还是 SVM 技术都没有提供直接的指令支持自动完成这一过程，同时又要求必须陷入到 Hypervisor 内部卸载，因此我们要注意 NewBluePill 为了实现卸载 Hypervisor 所采用的手法。¹

`HvmSpitOutBluepill()` 函数的作用是：该函数及其子函数通过 Hypercall 的手段为每个处理器卸载 NewBluePill Hypervisor 并进行后续恢复工作。可以看到，这里所采用的手法与 `HvmSwallowBluepill()` 函数类似，因此不再具体解释该函数具体实现。

在 `Hvm.c` 第 529 行，调用 `HvmLiberateCpu()` 函数，该函数又通过调用 `HcMakeHypercall()` 函数（`Hvm.c` 第 491 行）向 Hypervisor 送出申请卸载的消息（`NBP_HYPERCALL_UNLOAD`）。`HcMakeHypercall()` 函数会根据下层具体实现 HEV 技术的硬件（Intel/AMD）来执行相应的 Hypercall 实现在 Hypervisor 下的卸载。

下面我们分别介绍 NewBluePill 对 VT 技术和 SVM 技术实现的卸载过程：

批注 [S33]: 第二章卸载过程写好后改下这段的描述

批注 [S34]: 做实验看看 NewBluePill 是否能正确卸载。

¹ 实际上 NewBluePill 在卸载过程中是有 bug 的，在运行时也会发现卸载它会死机，不过这并不影响我们理解 NewBluePill 的卸载手法。

SVM 技术下 NewBluePill 的卸载实现

VT 技术下 NewBluePill 的卸载实现

在 VT 技术下，NewBluePill 的卸载是通过 VMCall 指令陷入的，并且在陷入过程中同样通过参数传入消息 NBP_HYPERCALL_UNLOAD，不过遗憾的是，默认情况下，NewBluePill 并没有在 VmxHandleInterception() 函数 (Vmx.c 第 128 行) 中对其进行处理，换句话说，默认情况下 NewBluePill Hypervisor 在 Intel 平台上是不能正常关闭的。¹

不过我们可以注意到，在 VmxHandleInterception() 函数中有这样一行 (Vmx.c 第 164 行)：

```
Hvm->ArchShutdown (Cpu, GuestRegs, TRUE);
```

虽然默认情况下没有开启 BLUE_CHICKEN 开关，但是 Blue Chicken 作为一种反虚拟机探测技术，在感知到有软件在尝试探测虚拟机时，会暂时拆除 Hypervisor 和虚拟机，所以实际上，利用这个函数是可以实现卸载 Hypervisor 功能的。

在 NewBluePill 对 VT 技术支持的实现中，这个函数指向 VmxShutdown() 函数，它负责具体卸载 Hypervisor 的步骤，同时为恢复操作系统在非虚拟机中执行设置弹簧床。该函数首先通过调用 VmxGenerateTrampolineToGuest() 函数 (Vmx.c 第 740 行) 设置弹簧床。

进入 VmxGenerateTrampolineToGuest() 函数可以发现，该函数通过传入的一块新内存作为弹簧床代码区，然后利用 CmGenerateMovReg() 函数将保存的虚拟机状态生成一系列机器码填入到该内存中，这些机器码对应的汇编指令功能是：将存储的虚拟机寄存器内容填充到 Hypervisor 对应寄存器中，从而 Hypervisor 和虚拟机完全一样。

VmxGenerateTrampolineToGuest() 函数执行完后，VmxShutdown() 函数将会执行 VmxDisable() 函数做最后拆除 Hypervisor 的工作，该函数执行 VMXOFF 指令并清空 CR4[VMXE]位，至此，虚拟机模式正式关闭。

批注 [S35]: 与实验：移植 NewBluePill 有关

关闭调试系统

DriverUnload() 函数的 38 行到 40 行在关闭 NewBluePill 的调试系统。(调试系统函数的详细作用我们会在“第八章 NewBluePill 调试系统”中阐述)

NewBluePill 关闭本地调试窗口，是通过调用函数 DbgUnregisterWindow() 实现的，这个函数主要作用是向窗口对应设备发送命令，通知其输出所有缓存信息然后撤销 NewBluePill 端为此设置的共享内存。

拆除私有页表

DriverUnload() 函数最后会关闭内存管理器，由于此时操作系统已经在使用 NewBluePill 自己维护的页表，所以 NewBluePill 分配的内存映射可从中找到。在释放页面时 MmShutdownManager() 会遍历一个内存信息链表，该链表原来用于存储 NewBluePill 私有内存的内存信息，通过该链表逐一释放自身所占每个页面。

¹ 在后面的移植 NewBluePill 的实验中，我们要解决这个 bug。

至此，NewBluePill 的卸载过程全部完成。

五、 NewBluePill 内存系统

1) 相关文件:

NewBluePill-0.32-public\common\Paging.c

NewBluePill-0.32-public\common\Paging.h

2) 技术背景:

分页机制是整个 nbp 的核心之一，称其核心是因为 nbp 所使用的很多其它部分都需要分页机制的支持，比如调试部分中调试信息的存放，还有整个 nbp 的代码段，以及 nbp 安装到每个 CPU 上后为了运行 nbp 所分配的堆栈区，这些内存（页）全部需要映射到 nbp 的分页机制上。

nbp 的分页机制是利用了 64 位系统特有的四级页表分页机制。AMD 和 Intel 对应的硬件实现基本相似。不过现在仍需定位 nbp 在哪里赋值给 EFER 和 CR4 寄存器，以及它赋的什么值，这影响到 nbp 如何启用相应的地址翻译机制。对于 AMD 的 CPU 来说，其各级页表寻址翻译总体过程如图 8.1 所示，每级 Virtual address 的地址结构要求可参考 AMD64 Architecture Programmer's Manual, Volume 2: System Programming 的第 131 页的四张图。Long Mode 的地址翻译机制还需要注意 CR3 的寄存器内容，在长模式下(Long Mode)，CR3 的寄存器内容包含如下部分（见图 8.2）:

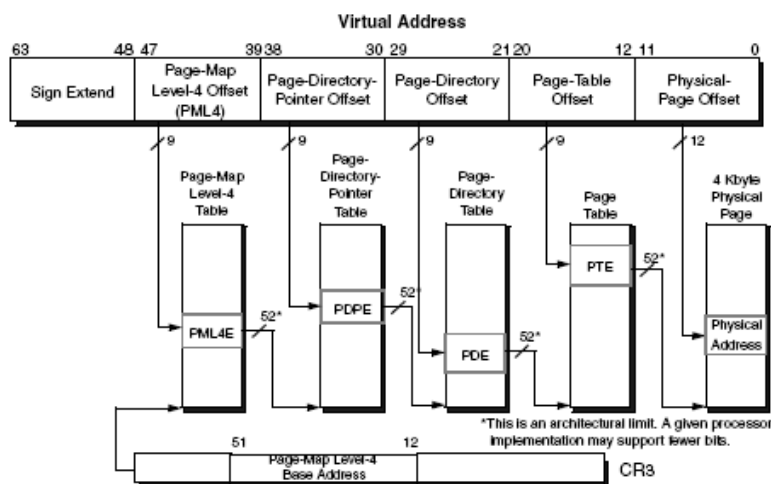


Figure 5-17. 4-Kbyte Page Translation—Long Mode

图 8.1 AMD 长模式(Long Mode)地址——4K 页翻译图

1. 页表基址域(Table Base Address Field), Bit 51-12, 这 40 位指向 PML4 页表基地址, PML4 页表必须是页对齐的。
2. 高页写穿位 (Page-Level Writethrough (PWT) Bit), Bit 3, 表明最高级别的页表使用写回(Writeback)还是写穿(Writethrough)策略
3. 高页可缓存位(Page-Level Cache Disable (PCD) Bit), Bit 4, 表明最高级别的页表是否可缓存。
4. 保留位(Reserved Bits), 保留位在写 CR3 寄存器时候应该清零。

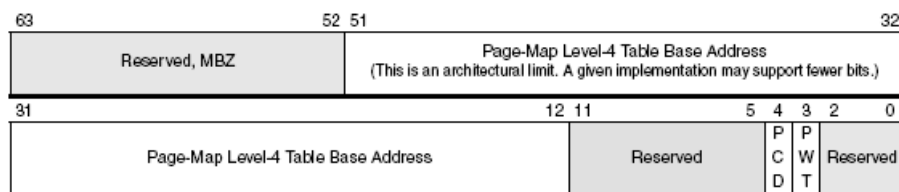


Figure 5-16. Control Register 3 (CR3)—Long Mode

图 8.2 AMD 长模式(Long Mode)中 CR3 寄存器内容

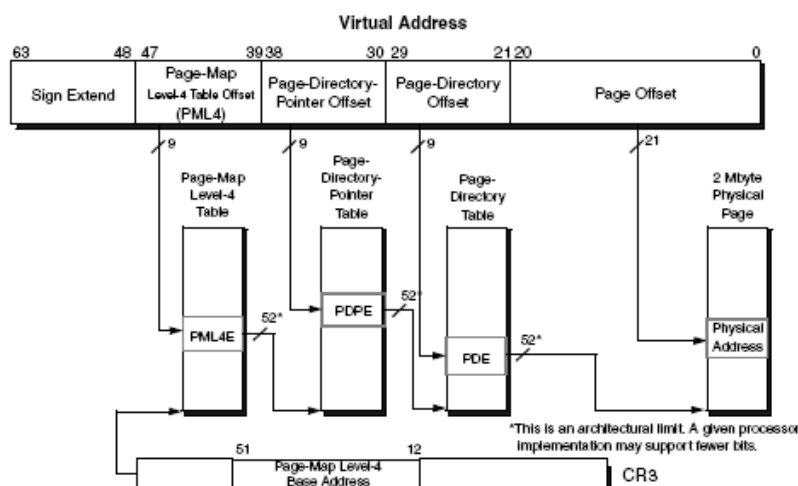


Figure 5-22. 2-Mbyte Page Translation—Long Mode

图 8.3 AMD 长模式(Long Mode)地址——2M 页翻译图

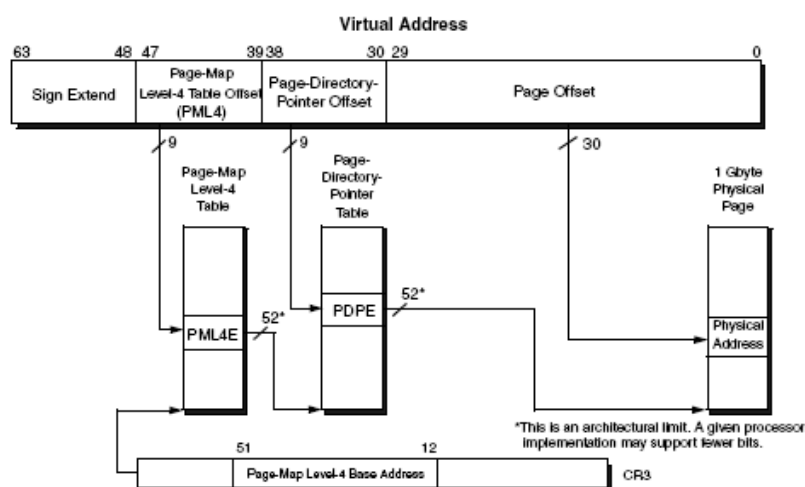


Figure 5-26. 1-Gbyte Page Translation—Long Mode

图 8.4 AMD 长模式(Long Mode)地址——1G 页翻译图

同时 AMD 也支持 2M 和 1G 页，通过分别挂载到 PD 和 PDP 即可实现，分别为 21 位寻址和 30 位寻址，比较简单因此不再叙述。（见图 8.3 和 8.4）

而对于 Intel 的 CPU 来说，这个翻译过程被称作 IA-32e Mode Linear Address Translation（IA-32e 模式地址翻译），各级页表的名称和翻译原理均十分相像，故在此仅列出 4K 页的映射过程(见图 8.5)，具体原理可参考 Intel 64 and IA-32 Architectures Software Developer's Manual Volume 3A 中第 3-42Vol3（第 126 页 Protect-Mode Memory Management）内容

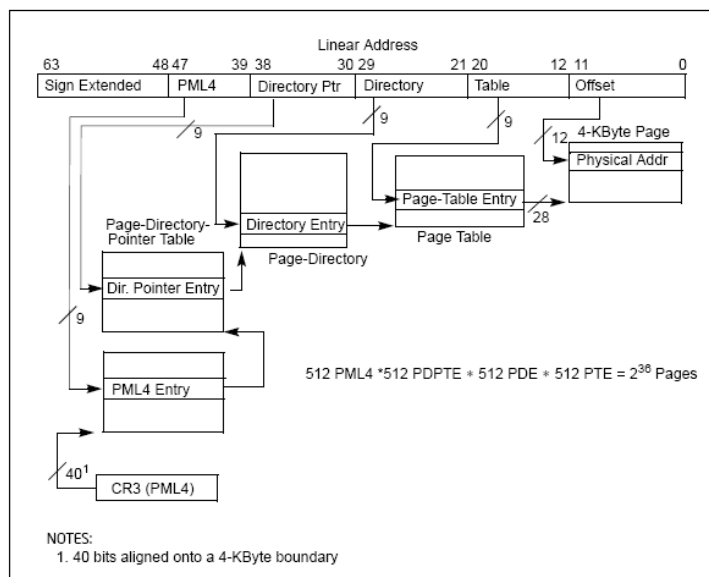


Figure 3-24. IA-32e Mode Paging Structures (4-KByte Pages)

图 8.5 Intel IA-32e 模式地址——4K 页翻译图

3) 总体功能介绍:

Nbp 的分页系统要完成的总体任务: 对 Windows 的内核地址分析, 并按照上面所提到的虚拟机地址翻译方法创建页表, 从而使得 nbp 可以在 VMM 层中仍旧可以使用 Windows API。

另一个目标是在 nbp 运行时分配自己的内存, 尽管分配的方法用到了 Windows API, 但是如何把新分配到的内存挂载到符合虚拟机地址翻译 (Address-Translation) 的页表上却是由 nbp 自己要负责的。

另外一点就是对于 Nbp 中自己产生的页表的 Host Address 和它在 Windows 中的 Virtual Address 并不相同, 这是由它要创建符合虚拟机地址翻译 (Address-Translation) 的页表造成的, Nbp 中自己产生的页的 Host Address 确是和 Windows 中的 Virtual Address 一样, 这个可能只是为了省事而已。

关于内存保护, nbp 的内存保护依赖于 Windows 操作系统。通过 Windows API 分配内存的方法使得各驱动间几乎无法访问同一块物理内存, 所以 nbp 无需再实现自己的内存保护方案。

实验: 查看 NewBluePill 的内存使用情况

我们可以通过一些工具来查看 NewBluePill 驱动的内存使用情况, 以证明确实隐藏了这些空间。比如 PoolMon (Windows Internals, 404 页)

批注 [S36]: 给出实验结果

4) 实现过程:

在谈了那么多背景知识后, 我们看一看 nbp 是去使用页表地址翻译的。nbp 利用了硬件提供的页表做地址翻译, 所以 nbp 在内存管理上主要做两件事情:

1. 根据硬件要求构建页表结构, 并将需要挂载的页挂载到这个页表结构上。
2. 在需要的时候改变 CR3 寄存器内容, 从而使 CPU 使用 nbp 的页表做地址翻译。
(这个时间点仍需确定) (CR3 更新内容后怎么找到 nbp 各段地址以继续运行?)

MmInitManager()方法

从 Newbp.c 的 DriverEntry 方法进入, 可以看到在通过 ComInit()方法设置当前运行的 bluepill 唯一标示 ID 后, 立即调用了 MmInitManager()方法来构建页表, 这个函数也是 nbp 内存管理部分的入口函数, 其作用是确定 PML4 页表的基地址 (第 621 行) 并初始化从 PML4 到 PT 每级一个页表, 当然此时也初始化好了他们之间的映射关系, 后者过程是通过 MmCreateMapping()方法和 MmUpdatePageTable()方法递归完成的。

可以观察到在 MmInitManager 中出现了两组地址翻译的操作, 一个是从 OS 的虚拟地址到内存物理地址的映射, 另一个是 VM 的虚拟地址 (或者说是 nbp 的虚拟地址) 到内存物理地址的映射。在 nbp 中, HostAddress 指 nbp 中 VM 寻址用的虚拟地址, PhysicalAddress 指物理内存地址, 而 GuestAddress 是指 OS 寻址用的虚拟地址。在后面

我们可以看到, nbp 为了实现方便, 把页的 GuestAddress 作为 HostAddress 直接挂在了 nbp 的页表上。

此外, nbp 为了方便管理这套页表 (主要是为了方便查找和释放某页), 利用 Windows 提供的双向链表将 nbp 自己所有的页表信息和页信息保存了起来并且顺序挂在这个双向链表上, 实现这个过程的方法是 MmSavePage(),

这里有一个细节就是每级页表所对应的 HostAddress, 依次定义如下: (定义在 \NewBluePill-0.32-public\commoncommon.h 的第 122 行)

```
#define PML4_BASE 0xFFFFF6FB7DBED000
#define PDP_BASE 0xFFFFF6FB7DA00000
#define PD_BASE 0xFFFFF6FB40000000
#define PT_BASE 0xFFFFF68000000000
```

此处的几个地址是固定的地址, 这几个地址的计算可参考参考书目一章中的 [4][5][6], 此处仅列出 Windows x64 address space layout (图 8.6, 512GB 4 level Page table map 行使我们所关心的, 这四个地址全在那段中)

当然更进一步的搜索发现在 xen 中也有完全相同的上面四个数字, 因此如果我们是开发自己的系统, 可以直接拿这几个 base address 来用, 常量不变。

MmSavePage ()方法

作用是将 nbp 分配出来的某个页表/页的信息保存起来并顺序挂在 g_PageTableList 这个双向链表上, 这个信息被存在 ALLOCATED_PAGE 结构体中, 主要记录了一个页表/页的物理地址, 在 OS 中的虚拟地址 (Guest Address, 某些时候使用以及释放该页时要借助于 OS 的帮忙), 在 nbp 内部的虚拟地址 (Host Address, 对于页来说这个地址就是其在 OS 中的虚拟地址), 分配类型 (Allocation Type), 页数量 (uNumberOfPages, nbp 只支持整页分配, 因此这项必为整数), 标志位 (Flags, 标记了关于该页的其它信息)

这里关于 PhysicalAddress 只取中间的 40 位, 而 HostAddress 要取高 52 位的一个原因是对于 PhysicalAddress 现在的内存远远不需要用 64 位来表示, 此处最细粒度是 4K 页, 因此低 12 位在此处是不用存的, 因为 CR3 中要求接受物理内存地址的中间 40 位,

| | |
|--------------------|---|
| 000007FFFFFFFFFFFF | User mode addresses - 8TB minus 64K |
| 000007FFFFFFFF0000 | |
| 000007FFFFFFFFFFFF | 64K no access region |
| | . |
| FFFFF0800000000000 | Start of system space |
| FFFFF6800000000000 | 512GB four level page table map |
| FFFFF7000000000000 | HyperSpace - working set lists and per process memory management structures mapped in this 512GB region |
| FFFFF7800000000000 | Shared system page |
| FFFFF780000001000 | The system working set information resides in this 512GB-4K region |
| | : |
| FFFFF8000000000000 | Mappings initialized by the loader |
| FFFFF9000000000000 | Session space |
| | This is a 512GB region |
| FFFFF9800000000000 | System cache resides here Kernel mode access only 1TB |
| FFFFFA800000000000 | Start of paged system area Kernel mode access only 128GB. |

图 8.6 Windows x64 结构地址空间图(x64 address space layout)

因此只需满足 CR3 的要求即可，当然如果为了以后 nbp 也能跑起来，CR3 高端部分开几位这里也就跟着开几位就可以了。

HostAddress 高 12 位必须保留，因为规范中规定这部分是 Sign extend 第 47 位得来的

关于分配类型(AllocationType)，在 nbp 中一共有三种内存分配类型：

- PAT_POOL 同 Windows 中的 a block of pool memory.
- PAT_CONTIGUOUS 同 Windows 中的 Contiguous Memory
- PAT_DONT_FREE 分配出来的多页内存，这种表示除了第一页外其它页的类型：此空间已被使用。（没什么别的含义）

关于标志 (Flags), 在 nbp 中对页表的标志有 5 个:

- AP_PAGETABLE: 表明该信息指一个页表的信息
- AP_PT: 该页表是 PT 级页表
- AP_PD: 该页表是 PD 级页表
- AP_PDP: 该页表是 PDP 级页表
- AP_PML4: 该页表是 PML4 级页表

◆ MmCreateMapping ()方法

这个方法主要作用是创建 PhysicalAddress 和 VirtualAddress (也就是 HostAddress) 之间的映射, 同时构造在这个翻译过程中所需要的每级页表。实际上构造完的结果可以使得用 VirtualAddress 去寻址 PhysicalAddress。这也是整个系统在分配页时所使用的方
法, 比如 MmAllocatePages(), MmAllocateContiguousPages(), MmMapGuestPages()等都调用该方法创建 VA 到 PA (虚拟地址到物理地址) 映射。

MmCreateMapping ()方法参数除了物理地址(Physical Address)和虚拟地址(Virtual Address)外, 还有一个 bLargePage 的布尔类型参数, 这个参数用于表明这个映射是否对应于一个大页, 在 nbp 中, 只有 2M 页和 4K 页两种, 当 bLargePage 为 true 时, 表明此时映射的是一个 2M 的大页, 一般情况下为 false, 表明映射 4K 页。

进入函数首先会调用 MmFindPageByPA()方法, 这个方法在此处的作用是获得每个 nbp 唯一的 PML4 页表信息, 其实最主要的是要获得其 OS 中的虚拟地址, 因为接下来会通过这个虚拟地址借助于 OS 的帮助构建后三级页表。

在这个函数中最主要的方法是 MmUpdatePageTable()方法, 它会根据虚拟地址用递归的方法创建完整的从 PML4 到 PT 各级页表, 具体实现方法会在 MmUpdatePageTable()方法中介绍。同时 PhysicalAddress 和 VirtualAddress 仍各取中间的 40 位和高 52 位, 此处这只是一种保证措施, 具体原因可参照上文。

MmSavePage ()方法

MmSavePage ()方法

六、 NewBluePill 陷入事件管理系统

在前面我们介绍 NewBluePill 的启动过程时，我们提到了在 NewBluePill 中存在一套陷入事件管理机制，同时列举出了一些事件/指令对应的处理函数。实际上，在 NewBluePill 中，陷入事件和处理函数的捆绑生成 Trap 元素、注册、激活是按照这样的流程进行的：

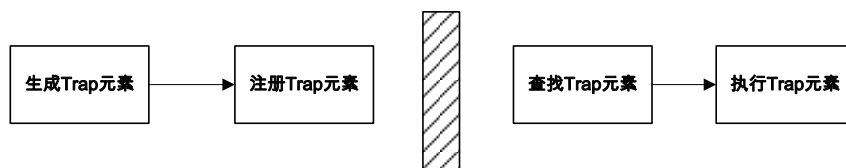


图 6.1 Trap 元素的生成、注册和执行

本章中，我们将深入研究 NewBluePill 陷入事件管理系统，并对这些处理函数的具体作用展开解释。

相关文件

NewBluePill-0.32-public\common\Traps.c
NewBluePill-0.32-public\common\Traps.h
NewBluePill-0.32-public\vmx\Vmxttraps.c
NewBluePill-0.32-public\vmx\Vmxttraps.h
NewBluePill-0.32-public\svm\Svmtraps.c
NewBluePill-0.32-public\svm\Svmtraps.h

Trap 元素的生成、注册机制

在描述 NewBluePill 的启动过程中我们已经提到，Trap 元素是 NewBluePill 用于捆绑 VMEXIT 事件和相应处理函数的基本单元，生成信息全部保存在 Trap 结构体中。

Trap 元素的生成是通过 TrInitializeGeneralTrap() 函数、TrInitializeMsrTrap() 函数和 TrInitializeIoTrap() 函数¹实现的。具体创建过程如下：

- 1) 为 Trap 元素分配内存
- 2) 填充 Trap 的类型和处理函数（TRAP_DISABLED²、TRAP_GENERAL、TRAP_MSR、TRAP_IO）
- 3) 根据 Trap 类型初始化其它内部域

具体数据结构及各项含义在 NewBluePill 的启动过程中已经提到，故不在此阐述。

¹ TrInitializeIoTrap() 函数在 NewBluePill 中并未具体实现。

² 通过 TrTrapDisable() 函数或者 TrTrapEnable() 函数控制这个类型。当一个 Trap 元素被 Disable 后，它所捆绑的处理函数不会被执行，但是该 Trap 元素仍然存在于列表内。

NewBluePill 陷入事件处理函数的注册是通过 `TrRegisterTrap()` 函数 (`Traps.c` 第 19 行) 实现的。这个函数的职责是根据传入的 `Trap` 元素的类型, 将 `Trap` 元素插入到相应链表中 (`Cpu` 结构体的 `GeneralTrapsList`、`MsrTrapsList` 或 `IoTrapsList`)。

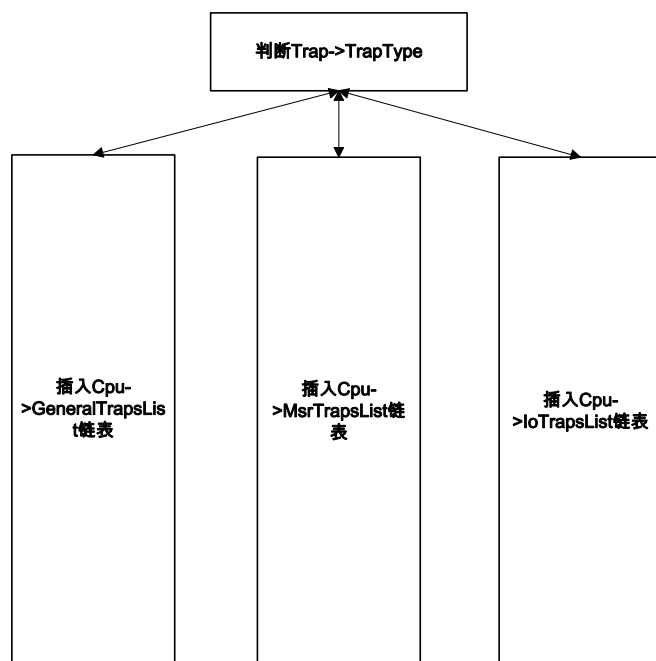


图 6.2 `TrRegisterTrap()` 函数的职责

由于使用的是 Windows 提供的链表机制, 所以在反注册一个 `Trap` 元素的时候只需提供 `Trap` 元素的挂钩, 即可利用 Windows API 删除整个元素, `TrDeregisterTrap()` 函数就是利用这样的机制反注册一个 `Trap` 元素的。

Trap 元素的触发机制

由于 VT 技术和 SVM 技术的实现差异, 因此一个处理函数的触发过程在两个平台上也并不相同, 我们可以人为的把它分为两个阶段, 如图 6.3 所示。

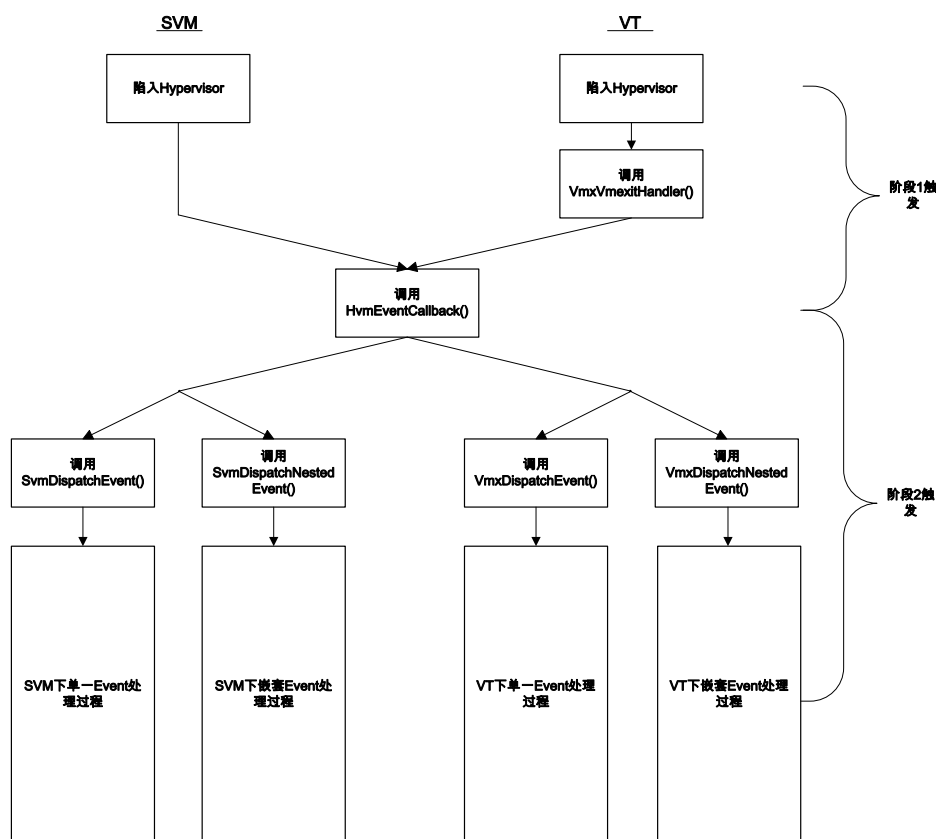


图 6.3 Trap 元素的触发机制图

阶段 1 触发

阶段 1 触发的过程比较简单，主要目标是进入 `HvmEventCallback()` 这个硬件抽象层的事件分发函数。NewBluePill 对 SVM 的实现中，在 `amd64\Svm-asm.asm` 文件第 109 行，`SvmVmrun()` 函数中，可以发现一旦发生 `#VMEXIT` 事件陷入 Hypervisor 后，会先保存各个通用寄存器以供事件处理函数在执行时可以获取虚拟机中断时的状态，然后立即执行 `HvmEventCallback()` 函数。而在 NewBluePill 对 VT 技术的实现中，VMCS 块中定义了一旦发生 `#VMEXIT` 事件陷入 Hypervisor，那么会执行 `VmxVmexitHandler()` 函数（`amd64\Vmx-asm.asm` 第 244 行），这个函数同样会保存各个通用寄存器后立即执行 `HvmEventCallback()` 函数。至此阶段 1 触发完成。

阶段 2 触发

阶段 2 触发过程比较复杂，因此我们按 VT 技术和 SVM 技术分开介绍。

SVM 技术下的阶段 2 触发

VT 技术下的阶段 2 触发

接下来，我们介绍下在 VT 技术中阶段 2 触发操作细节。对于 VT 技术，尽管在 `HvmEventCallback()` 函数中依旧可以调用 `Hvm->ArchIsNestedEvent()` 函数，但是由于 NewBluePill 没有针对 VT 技术实现嵌套事件分发功能，所以其对应的 `VmxIsNestedEvent()` 函数将永远返回 `FALSE`。且 `VmxDispatchNestedEvent()` 函数也不做任何事情，因此 NewBluePill 对 VT 技术阶段 2 触发的支持仅限于对单一事件的处理，`HvmEventCallback()` 函数通过调用 `Hvm->ArchDispatchEvent()` 函数（`Hvm.c` 文件第 244 行）实现这一点。

`Hvm->ArchDispatchEvent()` 函数对应函数是 `VmxDispatchEvent()` 函数，该函数会调用 `VmxHandleInterception()` 函数，并会通知它不要让虚拟机再继续处理这个事件。

`VmxHandleInterception()` 函数包含处理事件的主要逻辑，正如图 6.1 后半部分所示，它的工作包括获取陷入原因（Exit Code），指导查找 Trap 元素（`Vmx.c` 第 148 行到 153 行）、执行找到的 Trap 元素（`Vmx.c` 第 156 行到 158 行），并在打开 BLUE_CHICKEN 反 HVM 探测的情况下判断当前是否要暂时隐藏自己（`Vmx.c` 第 159 行到 167 行）。

查找 Trap 元素是通过 `TrFindRegisteredTrap()` 函数实现的（`Traps.c` 文件第 259 行）。这个函数会根据陷入原因而选择正确的 Trap 元素存储链表（`Traps.c` 文件第 275 行到 290 行），对于 Intel 平台，则一律指向 `GeneralTrapsList`。然后以 `Trap->General.TrappedVmExit` 的值为关键字逐一遍历该链表每个元素，并返回找到的第一个 Trap 元素。

执行 Trap 元素是通过 `TrExecuteGeneralTrapHandler()` 函数实现的（`Traps.c` 文件第 240 行）。该函数执行一个 Trap 元素内部捆绑的处理函数（`Traps.c` 文件第 251 行），同时最关键的是，在执行成功这个处理函数后，程序会调用 `Hvm->ArchAdjustRip()` 函数来调节虚拟机的 RIP（指令指针），使其在获得控制权后可以执行下一条指令¹，但是要注意的是，在执行 `Hvm->ArchAdjustRip()` 函数前，起到 RIP 偏移量作用的变量 `Trap->General.RipDelta` 必须已经设置好，它的设置工作通常是由各个处理函数内部实现。

最后执行的 Blue Chicken 策略，简单的说是在判断是否在短时间内发生了大量的 VMEXIT 陷入，如果发生，那么 NewBluePill 怀疑有人正在尝试探测自己，当前的实现中，NewBluePill 会通过 `Hvm->ArchShutdown()` 函数卸载自己。

当所有这些工作执行完后，`HvmEventCallback()` 函数返回，`VmxVmexitHandler()` 函数执行 `vmx_resume`（`Vmx-asm.asm` 文件第 258 行）返回到虚拟机中，虚拟机继续运行。整个 Trap 元素处理事件全部执行完成。

各处理函数功能和实现

熟悉了 NewBluePill 事件处理机制后，我们看看在 NewBluePill 中定义了哪些 VMEXIT 原因的处理函数及其具体执行流程，同样的，我们将会分开描述 VT 技术和 SVM 技术各处理函

¹ 在其它场合中，如果是因为 Page Fault 陷入，那么也可以用相同的手法在换页后使虚拟机重新执行发生页面异常的指令。

数的实现。

SVM 技术实现中各处理函数功能和流程

VT 技术实现中各处理函数功能和流程

前面我们曾经在表 4.3 和表 4.4 中列举过 VT 技术下陷入原因和处理函数的对应关系，接下来我们将详细阐述每个处理函数的流程。

VmxDispatchVmxInstrDummy() 函数

VmxDispatchVmxInstrDummy() 函数用于处理因 VMCALL、VMCLEAR、VMLAUNCH、VMRESUME、VMPTRLD、VMPTRST、VMREAD、VMWRITE、VMXON、VMXOFF 指令造成的陷入。这个函数不做任何事情，只是负责更新 Trap->General.RipDelta 域以便虚拟机执行后续指令（Vmxtrap.c 第 37 行）。

要注意的是，它还同时更新虚拟机的状态寄存器（RFLAGS）（Vmxtrap.c 第 39 行），该行作用是提供给虚拟机这些虚拟化指令执行失败（VMFAIL Invalid）的假象。

伪造虚拟化指令（VMX Operations）执行结果

可以通过伪装函数（pseudo-functions）来欺骗虚拟机虚拟化指令执行的结果。这些伪装函数包括 VMsucceed, VMfail, VMfailInvalid 和 VMfailValid。它们不是真正的函数，而是通过修改该状态寄存器（RFLAGS）的值来伪装虚拟化指令的执行结果。比如：

VMsucceed 的功能：CF<-0, PF<-0, AF<-0, ZF<-0, SF<-0, OF<-0

VMfail 的功能：如果 VMCS 指针有效，则是 VMfailValid，否则是 VMfailInvalid

VMfailValid 的功能：CF<-0, PF<-0, AF<-0, ZF<-1, SF<-0, OF<-0, VM-instruction error field 域设置为 ErrorNumber。

VMfailInvalid 的功能：CF<-1, PF<-0, AF<-0, ZF<-0, SF<-0, OF<-0[®]

更多信息请参考 *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 2B*, Chapter 5.2 Conventions。

VmxDispatchCpuid() 函数

VmxDispatchCpuid() 函数用于处理因 CPUID 指令造成的陷入。函数流程如图 6.4 所示：

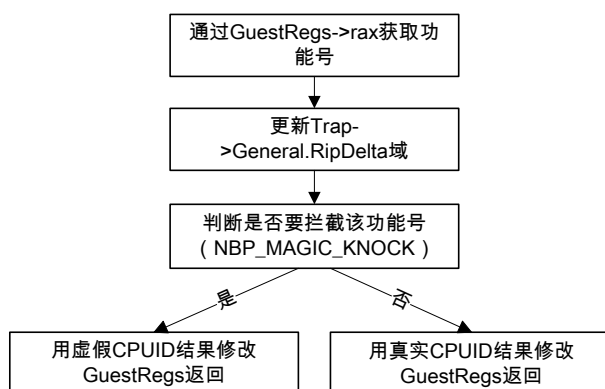


图 6.4 VmxDispatchCpuId() 函数流程图

首先获取虚拟机陷入时其 RAX 寄存器的值 (GuestRegs->rax)，该值将视作 CPUID 指令的功能号 (Vmxtrap.c 第 55 行)，然后更新 Trap->General.RipDelta 域以便虚拟机执行后续指令 (Vmxtrap.c 第 61 行)，随后判断如果该功能号是 NewBluePill 想要篡改的功能号，则篡改 GuestRegs 相应部分，从而在恢复虚拟机运行后返回给中断程序伪造的数据。

VmxDispatchINVD() 函数

VmxDispatchINVD() 函数用于处理因 INVD 指令造成的陷入。由于 INVD 指令只是负责让处理器缓存失效，因此这个函数与 VmxDispatchVmxInstrDummy() 函数一样不做任何事情，只是负责更新 Trap->General.RipDelta 域以便虚拟机执行后续指令 (Vmxtrap.c 第 499 行)。需要注意的是，由于 INVD 指令并不是虚拟化操作指令 (VMX Instructions)，因此在 VmxDispatchINVD() 函数的最后直接返回，而没有像 VmxDispatchVmxInstrDummy() 函数那样还有一步伪造虚拟化指令执行结果的步骤。

VmxDispatchMsrRead() 函数

VmxDispatchMsrRead() 函数用于处理因 RDMSR 指令造成的陷入。虚拟机中的 RDMSR 指令之所以能够陷入是因为在前面对 VMCS 结构体的设置中，默认情况下 NewBluePill 将 Primary Processor-Based VM Execution Controls.Use MSR bitmaps[bit 28]设置为 0，根据 Intel 文档的规定，这样设置就会造成虚拟机在执行 RDMSR 指令时发生 #VMEXIT 事件陷入到虚拟机中。VmxDispatchMsrRead() 函数的功能是：对于 Hypervisor 和虚拟机两者同一 MSR 寄存器号，其内容却不相同的，返回虚拟机 VMCS 结构体中对应 MSR 寄存器内容；否则返回 Hypervisor 的 MSR 寄存器内容（其实理论上也可以返回虚拟机 VMCS 结构体中对应 MSR 寄存器内容，但是考虑到两者被改变的 MSR 寄存器极少，而且读取虚拟机 VMCS 结构体中 MSR 寄存器内容的参数和普通读取 MSR 寄存器内容的参数是要做映射的，工作量会很大）。

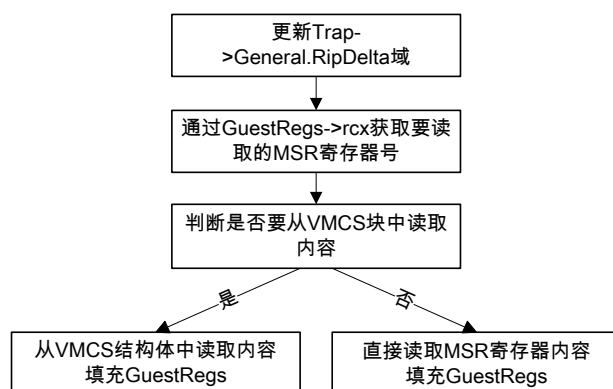


图 6.5 VmxDispatchMsrRead() 函数流程图

VmxDispatchMsrRead() 函数流程如图 6.5 所示。首先更新 Trap->General.RipDelta 域，然后获得虚拟机要读取的 MSR 寄存器号 (Vmxtraps.c 第 106 行)，随后根据该寄存器号选择是读取虚拟机 VMCS 结构体中对应 MSR 寄存器内容 (Vmxtraps.c 第 109 行到 124 行) 还是读取 Hypervisor 的 MSR 寄存器内容 (Vmxtraps.c 第 129 行到 130 行)。最后通过修改 GuestRegs 的 RAX 和 RDX 寄存器，将返回结果传回虚拟机。

这里存在一个问题，对于 GUEST_SYSENTER_CS、GUEST_SYSENTER_ESP、GUEST_SYSENTER_EIP 这三个 MSR 寄存器，实际上通过阅读前面章节中对启动过程的描述我们可以发现，他们的内容在 Hypervisor 和虚拟机中是一样的，所以此处 NewBluePill 作者选择仍从虚拟机 VMCS 结构体中读取，可能是为了代码运行时更保险一些。

VmxDispatchMsrWrite() 函数

VmxDispatchMsrWrite() 函数用于处理因 WRMSR 指令造成的陷入。与 RDMSR 指令一样，虚拟机中的 WRMSR 指令之所以能够陷入同样是因为前面对 VMCS 结构体的设置中 Primary Processor-Based VM Execution Controls.Use MSR bitmaps[bit 28]被设置为 0。

类似的，VmxDispatchMsrWrite() 函数首先更新 Trap->General.RipDelta 域，然后获得虚拟机要读取的 MSR 寄存器号 (Vmxtraps.c 第 157 行) 和要写入的值 (Vmxtraps.c 第 159 行到 160 行)，随后同样的，根据该寄存器号选择是写入虚拟机 VMCS 结构体中对应 MSR 寄存器域 (Vmxtraps.c 第 163 行到 177 行) 还是直接写入 MSR 寄存器 (Vmxtraps.c 第 183 行到 184 行)。

VmxDispatchCrAccess() 函数

VmxDispatchCrAccess() 函数用于处理因操作 CR0,CR3,CR4 控制寄存器造成的陷入，其主要流程如图 6.6 所示。函数首先更新 Trap->General.RipDelta 域 (Vmxtraps.c 第 221 行)，随后函数读取 VMCS 控制块中 VMEXIT 相关信息域 (VM Exit Information Fields) 中的退出条件域 (Exit Qualification)，并从中分离出被操作的控制寄存器号 cr (Vmxtraps.c 第 226 行) 和通用寄存器号 gp (Vmxtraps.c 第 225 行)。

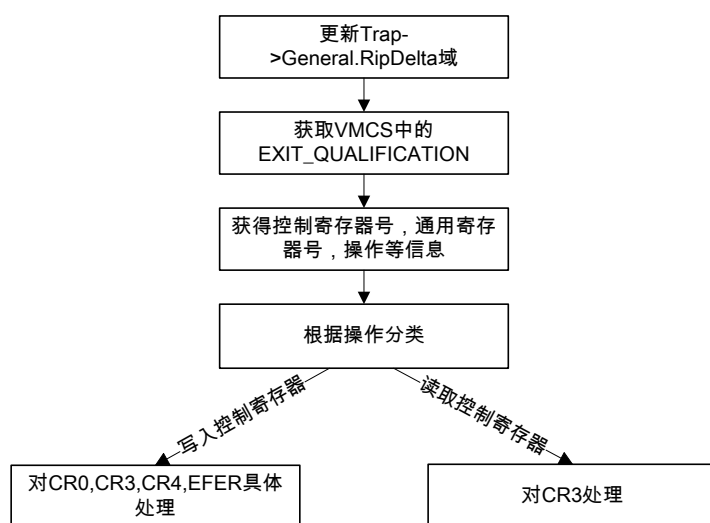


图 6.6 VmxDispatchCrAccess() 函数流程图

接下来 VmxDispatchCrAccess() 函数根据虚拟机对控制寄存器操作类型进行不同的处理过程。首先看看在写入控制寄存器操作中, VmxDispatchCrAccess() 函数是如何处理的。

在虚拟机准备写入 CR0 控制寄存器的情况下, 函数首先复制要写入的内容到 Cpu->Vmx.GuestCR0 中 (Vmxtraps.c 第 235 行), 有一点要注意的是, 该行代码使用 GuestRegs 结构体的方式决定了其中的元素位置是不能变化的。接下来函数判断 CR0 新值是否启用了分页机制 (Vmxtraps.c 第 236 行), 如果启动了分页机制, 函数会把 Cpu->Vmx.GuestCR3 域写入 VMCS 结构体中 Guest_CR3, 同时根据 EFER 寄存器的 EFER_LME (是否支持 x64) 来决定修改 EFER_LMA (是否开启 x64); 如果虚拟机禁用了分页机制, 那么函数会备份 VMCS 结构体中 Guest_CR3 内容, 并且修改 EFER_LMA, 不开启 x64 模式。最后修改 VMCS 结构体中的 CR0_READ_SHADOW 域, 同时调用 VmxUpdateGuestEfer() 决定是否要为虚拟机开启 x64 模式。¹

在虚拟机准备写入 CR0 控制寄存器的情况下, 函数首先复制要写入的内容到 Cpu->Vmx.GuestCR3 中 (Vmxtraps.c 第 267 行), 然后将新值写入到 VMCS 结构体的 Guest_CR3 返回。

对于 CR4 控制寄存器过程基本相同, 同样包括复制新值和更新 VMCS 结构体的 Guest_CR4 的操作。

而对于虚拟机将控制寄存器写入通用寄存器的操作, VmxDispatchCrAccess() 函数仅针对 CR3 做了处理。过程非常简单, 只是将 Cpu->Vmx.GuestCR3 域的内容复制到 GuestRegs 结构体相应变量 (Vmxtraps.c 第 306 行到 314 行)。

VmxDispatchException() 函数

VmxDispatchException() 函数用于处理因虚拟机发生异常而造成的陷入。这个函

¹ CR0 和 CR4 的处理过程最后都返回 False, 说明 NewBluePill 并不能正确处理写入这两个寄存器造成的陷入, 这一点要特别注意。

批注 [S37]: 为什么把 Cpu->Vmx.GuestCR3 域写入 VMCS 结构体中 Guest_CR3

批注 [S38]: 为什么把 Cpu->Vmx.GuestCR3 域写入 VMCS 结构体中 Guest_CR3

数是一个未完成的函数。首先读取出虚拟机的中断请求号（Vmxtaps.c 第 454 行），如果是调试异常¹则直接返回，然后为了保证计时的准确性对 VMCS 结构体中的 GUEST_INTERRUPTIBILITY_INFO 域赋值，使得不允许虚拟机中任何事件能被阻塞一段时间。随后该函数记录累计的执行周期数（Vmxtaps.c 第 473 行），并在追踪一定数量的虚拟机指令后清除虚拟机的 RFLAGS 的 TF 位，停止单步执行。最后的记录过程可与 SVM 中相应实现相同，但是 VT 和 SVM 一样这部分的实现并不完整。

VmxDispatchRdtsc() 函数

VmxDispatchRdtsc() 函数用于处理因 RDTSC 指令而造成的陷入。这个函数也是 NewBluePill 用于实现时间欺骗的关键函数，关于时间欺骗的具体原理我们会在“第七章 NewBluePill 反探测系统”叙述。该函数在更新 Trap->General.RipDelta 域（Vmxtaps.c 第 412 行）之后，根据时间欺骗策略填充作为结果的 Cpu->Tsc 域（Vmxtaps.c 第 418 行到 422 行），然后恢复各个中间变量（Vmxtaps.c 第 429 行到 432 行）并通过修改虚拟机的 RFLAGS 的 TF 位开启单步中断（Vmxtaps.c 第 436 行），从而可以在虚拟机恢复运行后，在停止追踪指令前，执行每条指令都会陷入到 Hypervisor 中并由 VmxDispatchException() 函数处理，最后修改 GuestRegs 结构体返回结果。

批注 [S39]: “第七章 NewBluePill 反探测系统”

VmxDispatchIoAccess() 函数

VmxDispatchIoAccess() 函数用于处理因 I/O 操作而造成的陷入。该函数主要功能是获取键盘和鼠标 I/O 操作的信息，如数据传送方向 In/Out、数据大小、数据值等。函数主要流程如图 4.7 所示：

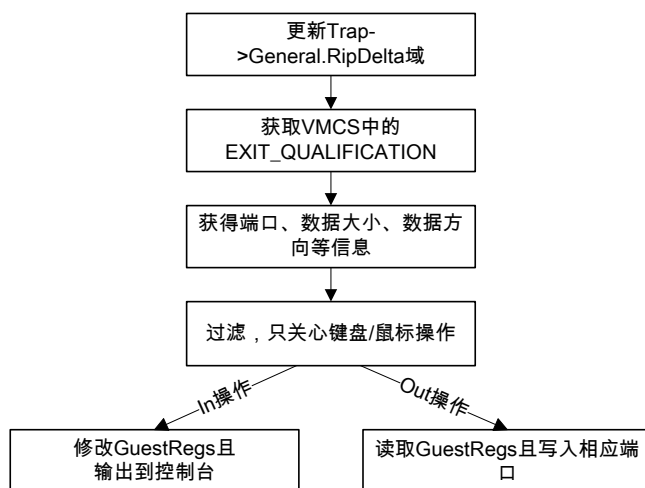


图 4.7 VmxDispatchIoAccess() 函数流程图

首先同样是更新 Trap->General.RipDelta 域（Vmxtaps.c 第 356 行），随后函数读取 VMCS

¹ 有关 Windows 下 x86/x64 异常和中断号对应关系，可以参考 Windows Internals, 4th Edition Chapter 3

控制块中 VMEXIT 相关信息域 (VM Exit Information Fields) 中的退出条件域 (Exit Qualification), 并由此获得端口、数据大小、数据方向等信息 (Vmxtraps.c 第 362 行到 368 行), 同时通过调用 `init_scancode()` 函数来初始化键盘扫描码。之后, 对于 In I/O 指令操作, 该函数会修改 `GuestRegs->rax` 域来传入虚拟机中, 同时, 如果操作对象是键盘, 那么该函数还会打印出所按的键。而对于 Out I/O 指令, 该函数负责读取 `GuestRegs->rax` 域, 并将内容输出到指定设备。

Note 关于退出条件域 (Exit Qualification) 对 I/O 操作记录信息的说明, 请阅读 *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 3B, Chapter 23.5* 相关部分

七、 NewBluePill 反探测系统

有了内存隐藏技术和针对多数陷入情况而实现的陷入处理，NewBluePill Hypervisor 看上去已经实现的很透明了，然而，NewBluePill 运行过程中的蛛丝马迹却还是能暴露 Hypervisor 自身，而面对这一切，NewBluePill Hypervisor 又是怎样应对的呢？本节我们就会介绍 NewBluePill 的探测原理和它对应实现的反探测技术。

探测 NewBluePill

根据 NewBluePill 的运行特点，探测 NewBluePill 的方法主要是如下几种：

- 设法创造陷入，通过测量时间判断是否 Hypervisor 在处理事件
- 执行一条可能陷入的指令，查看 TLB 上有多少项发生了变化从而确定是否存在 Hypervisor。
- 探测当前是否开启虚拟机模式

下面我们就将分别介绍每种探测方法和 NewBluePill 的对策：

通过指令执行耗时分析

通过在虚拟机中测量指令执行所消耗时间来判断当前是否存在 NewBluePill Hypervisor 是一个不错的方法，这种方法利用了引入 HEV 技术后带来的开销，如图 7.1 所示：

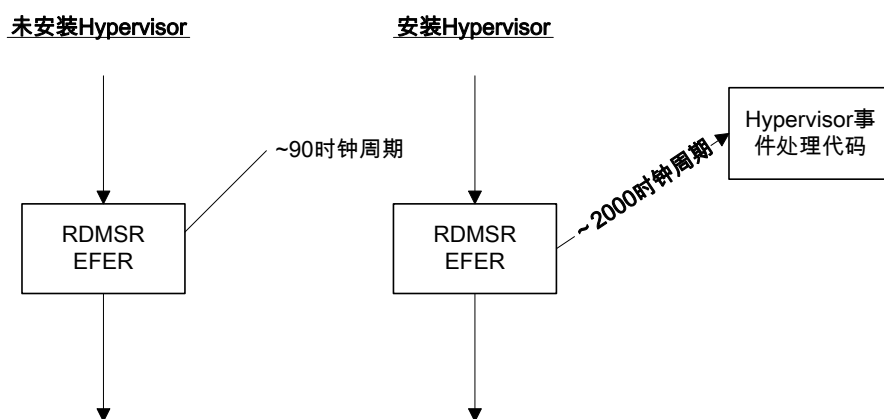


图 7.1 引入 NewBluePill Hypervisor 带来的时间开销

获得当前时间的方法有四种：

- 1) 读取 CPU 的时间戳寄存器 (Timer Stamp Counter)
- 2) 利用其它本地时钟 (如 HPET¹)
- 3) 真实计时器 (Real Time Clock)
- 4) 外部计时器 (External Clock)

¹ HPET (High Precision Event Timer, 高精度计时器)，未来将会替代真实计时器，每个 HPET 最多可以有 32 个计时器。有关 HPET 的具体资料请参考网上相关内容。

第一种方法利用 RDTSC 指令读取 CPU 的时间戳寄存器，这种方法获得的时间精度最高，可以精确到时钟周期，但是也是最容易欺骗的一种方法。

第二种方法同样有很高的精确度，但是 Hypervisor 通过对中断的监控也可以欺骗它的返回值。

真实计时器是一种 I/O 设备，计时精度为毫秒级，Hypervisor 仍能通过对 I/O 操作的监控欺骗返回值。

最后一种方法利用网络传输时间信息，且一般采用加密手段（如 NTP 协议），这就给 Hypervisor 的欺骗带来很大难度，只能针对某种特定实现欺骗。但是这种方法得到的时间精度很低，通常为 10ms 级，在一个典型的 1GHz 主频的处理器上，在这个时间间隔内已经执行了成百上千万条指令了。

NewBluePill 对于本地时间戳计时器的欺骗手段是利用指令追踪策略，如果这种探测升级为短时间内大量的探测的话，NewBluePill 的对策是 Blue Chicken 策略，本节稍后就会介绍这两个策略。

通过观察 TLB 变化分析

而通过观察 TLB 的变化去判断当前是否存在 NewBluePill Hypervisor 也是一种不错的手段。

众所周知，TLB 用于缓存最近访问的内存，无论是代码段或是数据段。在 NewBluePill Hypervisor 存在的情况下，一旦因为虚拟机执行某条指令而发生陷入，那么返回虚拟机中后 TLB 中内容一定是被污染的，只要有办法探测到这种污染就可以了，同样的，如果是短时间内大量的探测，NewBluePill 同样可以使用 Blue Chicken 策略自保。

然而，这种方法和第一种分析指令耗时的方法本质上确实在探测是否有 Hypervisor 存在而不是在探测是否存在 NewBluePill 等恶意 Hypervisor。考虑到当前和未来不断涌现的利用硬件虚拟化技术的软件，探测是否开启虚拟机模式及 Hypervisor 存在而确定是否存在类似于 NewBluePill 这样的恶意软件是行不通的。

Blue Chicken 策略

相关文件

NewBluePill-0.32-public\common\Chicken.c

NewBluePill-0.32-public\common\Chicken.h

功能介绍和详细分析

Blue Chicken 策略是这样的一种策略：当 NewBluePill 发现在短时间内出现了大量因为 VMEXIT 事件造成的陷入，NewBluePill 就怀疑有人在试图探测自己的存在，这时就会暂时卸载自己实现隐藏。

为了实现这一功能，Blue Chicken 采用了如图 7.2 的循环队列数据结构用于存储每次陷入发生时时间戳寄存器的值（TSC Value）。

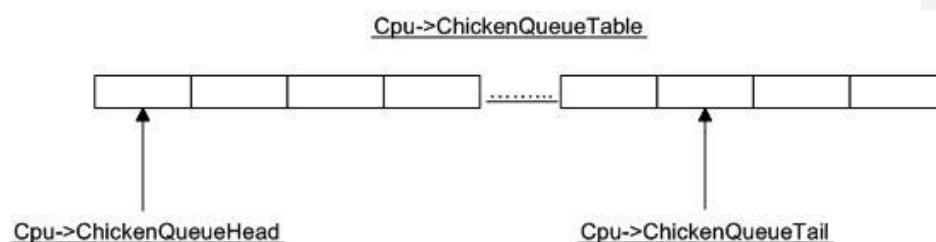


图 7.2 Blue Chicken 策略数据结构

进队操作通过 `ChQueueEnqueue()` 函数完成，进队元素永远插入 `Cpu->ChickenQueueTail` 所指的数组位置，当 `Cpu->ChickenQueueTail` 域已经指向了数组最后位置时，在队列未满的情况下指向数组头部，最后更新 `Cpu->ChickenQueueTail` 域。

出队操作通过 `ChQueueDequeue()` 函数完成，`Cpu->ChickenQueueHead` 指向的元素最先出队，然后更新 `Cpu->ChickenQueueHead` 所指的位置。

`ChickenShouldUninstall()` 函数是 Blue Chicken 策略的核心函数。该函数首先获得队列头尾两元素内部所存储的值（`Chicken.c` 函数第 100 行），然后判断如果在一定时间限度内发生了 `CHICKEN_QUEUE_SZ` 多次陷入的话（`Chicken.c` 函数第 101 行），那么返回 `TRUE` 通知 NewBluePill Hypervisor 卸载自身（`Vmx.c` 文件第 161 行，`Svm.c` 文件第 609 行）。

时间欺骗——指令追踪策略

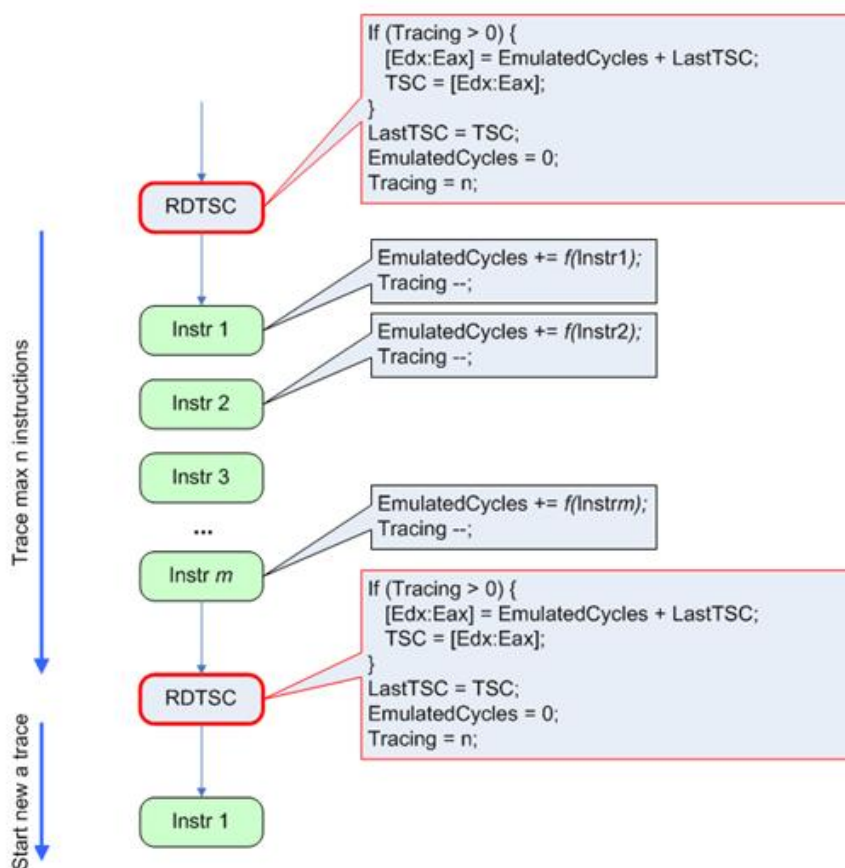
相关文件

NewBluePill-0.32-public\vmx\Vmxtraps.c

NewBluePill-0.32-public\svm\Svmtraps.c

功能介绍和详细分析

NewBluePill 的时间欺骗策略是通过指令追踪（Instruction Tracing）方法完成的，示意图如下所示：

图 7.3 NewBluePill 指令追踪策略示意图¹

NewBluePill 假设，探测器在检测时必然会反复执行一段指令并测量所消耗的时间。所以为了能够更精确的欺骗时间，NewBluePill 选择测量这段程序中的每条指令的真实执行时间。为了达到这一目的，一个关键要素就是在虚拟机访问时间戳寄存器后²，NewBluePill Hypervisor 必须开启单步中断，这样才能监控虚拟机后面执行的若干条指令，获得这些指令的执行信息。

接触过其它嵌入式平台的读者可能会问，为什么在这里我们不能使用通过手册中的信息生成的一张填充有每种指令执行周期的表格呢？

NewBluePill 之所以不这么做是因为，当前绝大多数的处理器都有流水线和缓存，这些因素造成了在运行时我们无法精确的知道每条指令在执行时具体消耗多少指令周期。此外，在比较特殊的情况下，如果某条指令正好处于页面的边缘，那么换页引入的开销我们也是无法通过预先生成的表格考虑到的，因此，运行时刻指令周期数的测量精度要高于预先静态生成表格这种方法的精度。

这种策略具体于每种平台的实现在代码中体现为 INTERCEPT_RDTSCs 开关之间的部分，所涉及的具体处理函数为：

¹ 该图摘自 IsGameOver, <http://www.bluepillproject.org/stuff/IsGameOver.ppt>

² 从第六章相关处理函数中我们可以看到，这包括了 RDTSC、RDTSCP 指令和对 TSC MSR 寄存器的直接操作。

SVM 平台:

- SvmDispatchDB () 函数
- SvmDispatchRdtsc () 函数
- SvmDispatchRdtscp () 函数
- SvmDispatchMsrTscRead () 函数

VT 平台:

- VmxDispatchRdtsc () 函数
- VmxDispatchException () 函数

这些函数具体过程在本书第六章中已有具体描述, 故不在此赘述。

Note 更多关于 NewBluePill 反探测技术的内容请参考 <http://www.bluepillproject.org/> 相关内容以及 *Subverting Vista™ Kernel for Fun and Profit*, Joanna Rutkowska

八、 NewBluePill 调试系统

Ok, 经过前面这些的讲解, 想必各位读者已经对 NewBluePill 各部分机理已经有了很多了解, 那么作为一个运行在如此底层并且具备一定代码量的系统, NewBluePill 的打印信息是怎样送出的呢 (每次分析蓝屏信息和 dump 文件肯定不是最好的方法, 通过分析打印信息能够节省很多调试时间)? 为什么不利用 WinDDK 的 API 而非要自己实现输出呢? 本章将对这些问题作出解释。

批注 [540]: 这个部分还需要更深入的探索

相关文件

```
Dbgclient\Dbgclient.c
NewBluePill-0.32-public\dbgclient\Dbgclient.c
Dbgclient\Dbgclient.h
NewBluePill-0.32-public\dbgclient\Dbgclient.h
Dbgclient\Dbgclient_ioctl.h
NewBluePill-0.32-public\dbgclient\Dbgclient_ioctl.h
NewBluePill-0.32-public\common\Portio.c
NewBluePill-0.32-public\common\Portio.h
NewBluePill-0.32-public\common\Comprint.c
NewBluePill-0.32-public\common\Comprint.h
```

功能概述

在 NewBluePill 中, 打印信息的输出主要有两种方式, 通过端口输出和通过共享内存窗口本地输出。后一种方式的示意图如图 8.1 所示:

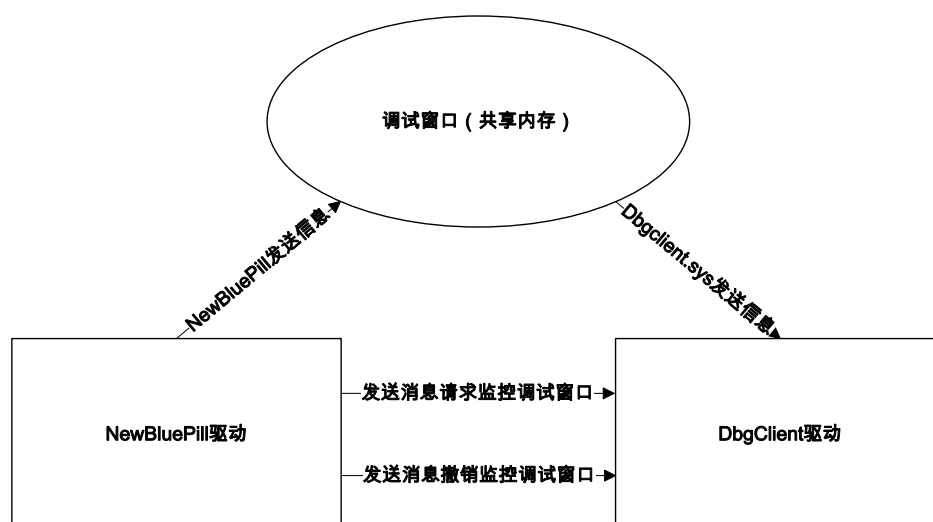


图 8.1 NewBluePill 本地输出信息方式示意图

这样做的原因在于，如果在 NewBluePill 内部完全通过 WinDDK API 中的 DbgPrint() 函数输出信息的话，那么在。。。。。。情况下会因为。。。。而死机¹。因此，为了获得这部分的调试信息，NewBluePill 创建一段内存，并把所有调试信息都写到这段内存中，随后当系统的 IRQ 级别降低时，Dbgclient.sys 会读取这些内容并在虚拟机中安全的调用 DbgPrint() 函数，将信息打印出来。

批注 [541]: 调查之

实现细节

下面我们看下 NewBluePill 调试系统具体是怎样实现的，NewBluePill 的调试系统与 HEV 技术本身无关，因此本章所叙述内容完全适用于 VT 平台和 SVM 平台。

我们将按照信息生产者 NewBluePill 和信息消费者 DbgClient 分别予以介绍。

NewBluePill 端调试系统部分

初始化过程

在“第四章 NewBluePill 的启动和卸载”中的启动过程时，我们提到过 NewBluePill 调试系统的初始化，本节我们将以此为入口探索 NewBluePill 端调试系统是如何初始化以及被使用的。

在 Newbp.c 文件第 53 行到 56 行，DriverEntry() 函数中，这几行代码在对 NewBluePill 调试系统做一部分初始化工作：如果准备使用串口输出调试信息，则利用 PpioInit() 函数指定输出信息用串口，然后无论是串口输出还是本地输出，都会通过调用 ComInit() 函数（CommonInit 的缩写）为该调试系统绑定 NewBluePill 全局唯一 ID 号 g_BpId，并初始化控制写入调试信息的自旋锁 g_ComSpinLock。

随后，程序运行到 Newbp.c 文件第 64 行时，该行在 NewBluePill 启用本地输出时启用，通过调用 DbgRegisterWindow() 函数，NewBluePill 调试系统申请了 5 页大小的内存作为调试窗口内存区，同时填充 DebugWindow 结构体¹，用于保存对该调试窗口的描述信息。

```
typedef struct _DEBUG_WINDOW
{
    UCHAR    bBpId;           /*保存 NewBluePill 全局唯一号（实例间唯一）*/
    PVOID    pWindowVA;       /*保存 NewBluePill 调试窗口虚拟地址*/
    ULONG    uWindowSize;     /*保存 NewBluePill 调试窗口大小 */
} DEBUG_WINDOW,
*PDEBUG_WINDOW;
```

最后，DbgRegisterWindow() 函数会调用 DbgSendCommand() 函数将 DebugWindow 结构体实例发送到 DbgClient 驱动（common\Dbgclient.c 文件第 72 行，该驱动将自身注册为“\\Device\\ntldbgclient”对象），DbgClient 驱动在接收到该消息后就会把该实例

¹ 该结构体实例既用于 NewBluePill 端又会通过设备通信方式传输到 DbgClient 端，因此 DebugWindow 结构体的定义会在两端出现且定义一致。

添加到监视列表上。DbgSendCommand() 函数的作用就是利用 Windows 设备消息发送机制向 DbgClient 注册设备发送信息。至此，NewBluePill 端调试系统初始化完成。

卸载过程

NewBluePill 端调试系统的卸载过程相对简单一些，它并不需要显式地做例如释放自旋锁资源等等的过程，因为在 NewBluePill 生命周期结束后，Windows 会管理这些。在卸载过程中唯一要做的事情就是在启用本地调试信息输出的情况下调用 DbgUnregisterWindow() 函数（Newbp.c 文件第 39 行）释放调试窗口所占内存，并通知 DbgClient 将对 NewBluePill 调试窗口的监视从监视列表中去除掉。DbgUnregisterWindow() 函数仍会调用 DbgSendCommand() 函数（common\Dbgclient.c 文件第 83 行），不过这次发送的消息换成了请求移除一个监视调试窗口的消息。

使用过程

在 NewBluePill 中，打印消息全部使用 _KdPrint 宏，该宏定义在 Common\Common.h 的第 92 行到第 96 行。在启用调试信息输出的情况（ENABLE_DEBUG_PRINTS，默认开启）下，NewBluePill 会通过调用 ComPrint() 函数（定义在 common\Comprint.c 文件第 106 行）进行统一的打印信息输出管理工作。

ComPrint() 函数同时管理对本地调试窗口的信息输出和通过串口的信息输出，函数首先获得旋锁，保证了某一时刻只能对一个打印信息请求进行处理。在打开 COMPRINT_OVERFLOW_PROTECTION 开关的情况下，该函数可以对溢出情况进行处理，即禁止在 COMPRINT_QUEUE_TH 时钟周期内输出超过 COMPRINT_QUEUE_SZ 行字符串，为了达到这个目的，这里同样维护了一个类似于 Blue Chicken 策略的循环队列，用于填充最近打印若干行字符串发生的时刻（同样是保存时间戳寄存器内容），组装字符串所用到的 snprintf() 函数（common\Comprint.c 文件第 164 行）是一个第三方开源库提供的，在此不进行分析。此后函数里通过调用 _ComPrint() 函数（common\Comprint.c 文件第 175 行）将要输出的字符串送到指定设备，这个函数会处理输出到端口或者本地调试窗口的具体细节。最后 ComPrint() 函数释放旋锁，整个处理过程结束。

DbgClient 端调试系统部分

从 NewBluePill 端调试系统部分的初始化过程我们可以看出，DbgClient 必须先于 NewBluePill 启动，这样才可以在 Windows 中注册相应设备从而能够接收到 NewBluePill 发送的消息。下面我们就介绍下 DbgClient 的启动、卸载和使用过程。

初始化过程

DbgClient 作为一个独立的驱动而存在，因此它有自己的 DriverEntry() 入口函数（定义在 dbgclient\Dbgclient.c 文件第 283 行）。该函数的工作流程图如图 8.2 所示：

批注 [S42]: DbgClient 的作用？为什么作者要这么设计

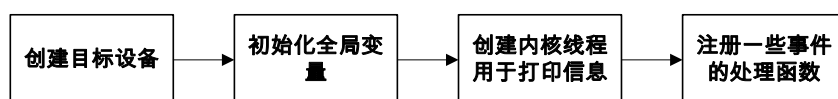


图 8.2 DbgClient 入口函数流程图

进入 `DriverEntry()` 函数，首先为目标设备声明了“`\\Device\\itldbgclient`”¹和“`\\DosDevices\\itldbgclient`”²两个设备名（`dbgclient\\Dbgclient.c` 文件第 293 行到 294 行），并分配了 `DEBUG_WINDOW_IN_PAGES` 页数的内存（`dbgclient\\Dbgclient.c` 文件第 297 行），在将来使用 `DbgPrint()` 函数打印信息时，会首先复制要打印的内容到该段内存中，起到了缓存和快速处理²的作用。随后 `DriverEntry()` 函数调用系统 API，将 `DbgClient` 驱动自身创建为设备，并以这两个设备名注册这同一个设备（`dbgclient\\Dbgclient.c` 文件第 303 行到 314 行）。

随后 `DriverEntry()` 函数初始化 `g_DebugWindowsList`，`g_DebugWindowsListMutex`，`g_ShutdownEvent` 这几个全局变量（`dbgclient\\Dbgclient.c` 文件第 316 行到 318 行），这三个全局变量用途如下：

- `g_DebugWindowsList` `DbgClient` 被设计用来同时监控多个调试窗口的输出信息，因此该列表用于存储每个调试窗口对应 `DebugWindow` 结构体实例。
- `g_DebugWindowsListMutex` 该互斥锁用于互斥执行 `PrintData()` 函数，该函数负责调用 `DbgPrint()` 函数打印信息。
- `g_ShutdownEvent` 这个对象被用于保证打印线程在 `DbgClient` 驱动被卸载前不会停止。使用了 Windows 内部对象的消息通知技术（Event Notification）。

初始化全局变量工作完成后，`DriverEntry()` 函数会创建内核线程，用于打印各个调试窗口的输出信息（`dbgclient\\Dbgclient.c` 文件第 320 行到 349 行），其中会测试下所创建的线程拥有的访问权限（`dbgclient\\Dbgclient.c` 文件第 334 行到 347 行）。

此时打印信息线程已成功开启并运行，`DriverEntry()` 函数最后做的工作就是注册事件处理函数 `DriverDispatcher()`（`dbgclient\\Dbgclient.c` 文件第 352 行到 354 行），该函数在 `NewBluePill` 开启、关闭、向 `DbgClient` 设备发送命令时会被执行。

此时 `DbgClient` 设备顺利启动。

卸载过程

`DbgClient` 通过调用 `DriverUnload()` 函数执行卸载工作，其卸载过程如图 8.3 所示：

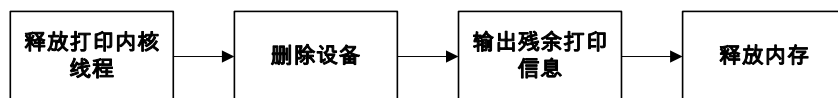


图 8.3 DbgClient 卸载过程流程图

该函数首先向打印线程监听的 `g_ShutdownEvent` 事件发送消息，这会使得线程退出循环结束（`dbgclient\\Dbgclient.c` 文件第 61 行到 69 行），然后判断 `g_pScanWindowsThread` 是否指向在初始化过程中所创建的打印内核线程，如果是则释放对它的引用（`dbgclient\\Dbgclient.c`

¹ 该设备名会把设备注册到系统全局对象空间上（`\\GLOBAL??\\`）。

² 一般情况下，I/O 操作慢于内存操作，用这种机制可以提升 `NewBluePill` 生产消息的速度。

文件第 250 行到 252 行), 此时该线程引用为 0, Windows 也就可以自动终止该线程并回收其占用的空间。随后, DriverUnload() 函数删除自身所代表的设备 (dbgclient\Dbgclient.c 文件第 256 行到 259 行) 并调用 PrintData() 函数输出可能留在调试窗口中的残余的打印信息, 随后遍历 g_DebugWindowsList 链表释放其中每个元素所占用的空间, 最后函数释放 g_pDebugString 所占用的内存空间 (dbgclient\Dbgclient.c 文件第 275 行到 276 行)。至此, DbgClient 的卸载过程全部完成。

使用过程

通过对 NewBluePill 端调试系统启动、关闭和使用过程的介绍, 我们可以知道它主要利用了 Windows 的对象通信机制。作为调试信息的接收者, DbgClient 在初始化过程的最后定义了在新 BluePill 端调试系统使用它的时候, 均要移交给 DriverDispatcher() 函数做具体处理工作, 下面我们就看看 DbgClient 是如何被使用的。

DriverDispatcher() 函数仅起到过滤事件的作用, 它会调用 DeviceControl() 函数 (dbgclient\Dbgclient.c 文件第 224 行) 仅针对 IRP_MJ_DEVICE_CONTROL 造成的触发进行处理¹。DeviceControl() 函数根据 NewBluePill 端调试系统发送的消息命令类型进行不同的处理: 如果是 IOCTL_REGISTER_WINDOW 类型, 也就是 NewBluePill 端调试系统请求 DbgClient 监控一个调试窗口, 那么 DeviceControl() 函数会首先遍历监控链表 g_DebugWindowsList, 查看是否已存在对这个调试窗口的监视 (dbgclient\Dbgclient.c 文件第 112 行到 123 行), 如果不存在则将其封装到一个 DEBUG_WINDOW_ENTRY 类型结构体中并将该实例添加到监控链表最后位置上 (dbgclient\Dbgclient.c 文件第 125 行到 145 行)。

如果是 IOCTL_UNREGISTER_WINDOW 类型, 也就是 NewBluePill 端调试系统请求 DbgClient 撤销监控一个调试窗口, 那么 DeviceControl() 函数会首先调用 PrintData() 函数输出所有调试窗口的残余信息 (dbgclient\Dbgclient.c 文件第 162 行), 随后遍历监控链表 g_DebugWindowsList, 移除目标调试窗口信息 (dbgclient\Dbgclient.c 文件第 166 行到 185 行)。

而 DbgClient 所创建的线程则在 DbgClient 生命周期内一直运行, 在 DbgClient 被卸载前, 该线程执行的 ScanWindowsThread() 函数由于一直等待不到 g_ShutdownEvent 事件而超时, 因此陷入死循环中 (dbgclient\Dbgclient.c 文件第 61 行到 69 行), 从而得以不断执行 PrintData() 函数。

PrintData() 函数是负责利用 WinDDK 中 DbgPrint() 函数输出信息的核心函数, 该函数首先利用互斥锁 g_DebugWindowsListMutex 确保自己的执行不会被打印线程干扰, 然后遍历监控链表 g_DebugWindowsList 中每一个调试窗口, 并复制其中内容到本地临时内存中 (dbgclient\Dbgclient.c 文件第 31 行), 随后遍历要打印的内容, 将回车符替换为 NULL, 再利用 DbgPrint() 函数逐行打印这些内容 (dbgclient\Dbgclient.c 文件第 35 行到 41 行)。当打印任务完成后释放互斥锁并成功返回。

总结

正如本章开始中所提到的那样, 如果在 NewBluePill 内部完全通过 WinDDK API 中的 DbgPrint() 函数输出信息的话, 那么在..... 情况下会因为..... 而死机, 读者可以通过修改_KdPrint 的宏定义来测试这一点, 并用 windbg 观察此时的输出和函数调用堆栈,

批注 [S43]: 调查之

¹ 有关 Windows IRP (Input/Output Request Packets) 的具体信息, 请参考网上相关内容。

相信一定有意意外收获。

这也就提醒了我们，在底层上实现自己的系统，调试系统的支持必不可少，它就像脚手架一样帮助我们实现这个系统，当已有的支持不能很好的帮助我们实现这一目标时，我们就必须自己实现——好的调试系统能够减少大量的开发时间和测试时间。

关于 Bpknock 触发器

由于 bpknock 很简单，所以本书中不再另开章节对其进行专门讲述。

Bpknock 程序是 NewBluePill 的演示程序（bpknock\Bpknock.c 文件），其作用是通过调用 cpuid 这条汇编指令触发 #VMExit 事件，使得 NewBluePill 陷入 Hypervisor 处理异常，最后将自己定义的返回结果赋值给相应寄存器，再回到 OS 中读出该修改过的返回结果，从而证明 NewBluePill 确实生效。核心函数分析如下：

```
ULONG32 __declspec(naked) NBPCall (ULONG32 knock) { // 使用
__declspec(naked)来自己管理函数调用堆栈
    __asm {
        push    ebp
        mov     ebp, esp
        push    ebx        ;保护 ebx,ecx,edx 寄存器
        push    ecx
        push    edx
        cpuid          ;要用 cpuid 触发异常陷入 VMM
        pop     edx
        pop     ecx
        pop     ebx
        mov     esp, ebp
        pop     ebp        ;恢复局部栈
        ret
    }
}
```

PART3 实验部分

九、动手写自己的第一个 HVM 程序

实验目的

1. 实践 Intel 平台 VT 技术相关指令
2. 掌握 VM Root 层 On the fly 方式的安装方法

实验概述

按照几乎所有新技术新语言的学习习惯，我们同样从 HelloWorld 开始做自己的第一个基于 HVM 的程序，当然为了便于调试，我们采用了跟 nbp 一样的 On the fly 安装。

我们的 HelloWorld HVM 程序通过如下方法实现：拦截 cpuid 调用，根据从 Console 过来的传入值生成相应传出值，但是如果传入值为 100，那么我们修改 eax,ebx,edx 三个寄存器的内容为 Helloworld，为了证明我们的 Helloworld 是有效的，可以做如下对比实验：卸载掉我们的 HelloWorld 驱动，然后通过 console 传入 100，察看输出，如果不再是 Helloworld 而是其它东西，那么我们的实验成功。

由于仅是为了说明我们可以利用 HVM 成功写出程序，所以我们在 Nbp 基础上删除了内存管理和调试模块，输出也仅利用 DbgPrint 而不再需要 sprintf。同时我们也删除了它对于 SVM 和 VMX 两种技术的平台无关性。我们的 hello world 将仅支持 Intel 平台。这样一来，我们的整个逻辑将仅包含本书五六两章提到的 nbp 最基本的程序逻辑，这将大大简化代码尺寸并便于我们理解和以后修改我们的 HVM Helloworld 程序。

批注 [S44]: Reference to 5.6

实验过程

十、 移植 NBP 到 32 位系统

十一、 开发基于 HEV 技术的注册码验证器

实验目的

1. 实践 VT 技术相关指令
2. 实践 VT 技术中 VMX 抢占计时器技术
3. 实践 NewBluePill 的内存隐藏技术

实验概述

在前面的实验中，我们已经对虚拟化技术有了初步的了解。这个实验将展示虚拟化技术在实际生活中的运用。

如今，共享软件和大多数的商业软件都在使用注册码技术来保护自己的版权，然而，由于验证程序处于 Ring-3 特权态（少数处于 Ring-0 特权态），因此很容易通过动态分析的手段改变跳转/改变语义或者推算出算法，从而破解掉软件。

这个问题的根源在于两个：

- 1) 注册码验证系统的运行空间操作系统可见
- 2) 注册码的验证一般只有一次¹

引入虚拟化技术后，由于我们拥有了比操作系统更高的权限，再加上我们在 NewBluePill 中所看到的内存隐藏技术，所以我们可以尝试去解决这个问题。考虑下面的模型：

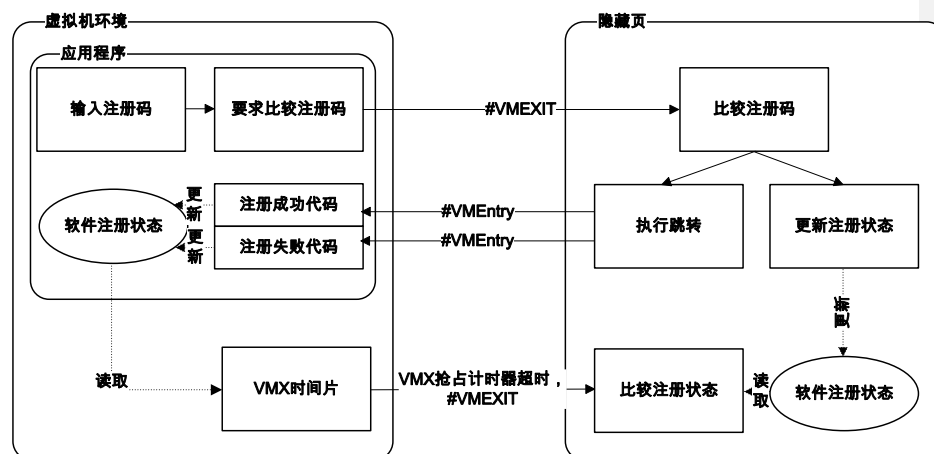


图 11.1 注册码验证器流程图

由图 11.1 所示，将比较注册码的关键部分内存隐藏，从而无法通过动态分析看到这部分内存。此外，通过 VMX 抢占计时器每过一定时间判断一次软件真实注册状态，防止破解者通过修改应用程序的注册失败代码语义而修改保存在应用程序中的软件注册状态²。当在

¹ 其实这个问题也与操作系统可见所有用户程序空间有关，在普通情况下，软件多次验证注册码基本不能带来任何好处。

² 在应用程序部分保存软件注册状态可以起到优化的作用，当其状态为未注册时，可以不必再去通过 VMX

Hypervisor 中发现应用程序的软件注册状态不等于自己保持的注册状态时，终止应用程序的运行。

实验过程

抢占计时器去陷入检查。

A. 其它有关 HVM 技术的项目

Xen (version>3)

Xen 创建虚拟机过程:

- a) 由 Dom0 创建对应客户机的 VT 域, 并载入客户机的虚拟 BIOS。然后, 完成这些工作以后, 由 VMM 创建虚拟机可用的内存映射,
- b) 在 Dom0 设备模块中完成相应的配置, 如设备绑定, IRQ 映射, PCI 配置等。然后创建事件通道。
- c) 把配置信息传送给客户机的虚拟 BIOS。

<http://cache.baidu.com/c?m=9d78d513d98916f34fede53a4b029026475bda257a95c7140cc98e15d2735b360f3ea4ac2755475294d27c1050f3540cbab12b7235012aaa8cc9fc09cabae47472de7029771a81&p=8a33f915d9c041c30be2932a1e4a&user=baidu>

HVM

B. 其它安全技术

TPM 技术

C. 相关软件和参考文档

相关软件:

- AMD SimNow™ Simulator: <http://developer.amd.com/cpu/simnow/Pages/default.aspx>
- VMware Fusion
- Parallels Desktop for Mac
- Parallels Workstation
- DNGuard HVM

参考文档:

[1] Virtualization <http://www.answers.com/virtualization>

[2] An Introduction to Hardware-Assisted Virtual Machine (HVM) Rootkits
<http://www.megasecurity.org/papers/hvmrootkits.pdf>

[3] Pacifica – Next Generation Architecture for Efficient Virtual Machines
http://developer.amd.com/assets/WinHEC2005_Pacifica_Virtualization.pdf

[4] 64 비트 윈도우 커널 분석 AMD64 (AMD64 架构的 64 位 Windows 内核分析)
<http://greemate.tistory.com/attachment/ck010000000001.pdf>

[5] 64 비트 윈도우 커널 분석 IA64 (IA64 架构的 64 位 Windows 内核分析)
<http://greemate.tistory.com/attachment/dk010000000001.pdf>

[6] Windows Internals, Fourth Edition

↩ Full Virtualization, http://en.wikipedia.org/wiki/Full_virtualization

↩ *AMD64 Architecture Programmer's Manual, Volume 2: System Programming*, Rev 3.14 P367

↩ *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 3B*, Chapter 20.3

Organizations of VMCS Data

↩ *Intel 64 and IA-32 Architectures Software Developer's Manual Volume 2B*, Chapter 5.2
Conventions