

服务器虚拟化基础设施性能优化研究

目前服务器虚拟化技术已非常普及，对集群性能优化的需求日趋增长。基础设施作为集群正常运作的基石，保障其各项功能、组件的高效协同运转至关重要。本文首先介绍了三种主流服务器虚拟化技术以及相应的CPU虚拟化实现模式，接着以VMware vSphere 5.5的全虚拟化体系结构为例，提出了集群基础设施性能优化的整体思路、优化项和核心影响因素，从物理CPU、物理存储、物理网络和服务BIOS四个方面详细阐述了优化的内容和方法。

陈乐然 王刚 陈威 徐小天/国网冀北电力有限公司电力科学研究院

0 引言

目前服务器虚拟化技术已被运用到大多数的企业级数据中心，通过构建宿主机集群，最大化地利用了现有的服务器硬件资源，降低了数据中心能耗，与企业IT发展绿色环保的目标相契合。在这种背景下，优化虚拟化环境的整体性能成为了当前学术界和工业界对云计算基础架构技术的研究热点之一。对服务器虚拟化性能的优化分为很多方面，包括基础设施性能优化、虚拟网络、虚拟机监视器（Virtual Machine Monitor, VMM）层、虚拟机、客户操作系统和虚拟基础设施等。其中保障基础设施的最佳性能是开展其他优化工作的基础和前提。在下文中我们将首先简要介绍服务器虚拟化的三种实现模式，并以VMware vSphere 5.5为例，从优化思路、优化项和核心影响因素三个角度详细说明基础设施的性能优化方法。



陈乐然/中级工程师

关键词/Keywords

服务器虚拟化 ·
基础设施性能优化 ·
VMware vSphere ·

1 主流服务器虚拟化技术

目前主流的服务器虚拟化技术分为

三类：全虚拟化、半虚拟化和硬件辅助虚拟化。全虚拟化的代表是VMware的vSphere；半虚拟化的代表包括Xen和微软Hyper-V；硬件辅助虚拟化多用于优化全虚拟化和半虚拟化性能，并不是独立的实现模式，近几年来，许多成熟的硬件辅助虚拟化技术相继问世，例如Intel的VT-x和AMD的AMD-V等。

1.1 全虚拟化

全虚拟化的实现方式是在客户操作系统（Operation System, OS）和物理硬件间增加VMM（也称为Hypervisor），将物理的CPU、存储、内存和I/O资源抽象化，通过VMM实时探测虚拟机发出的指令流，动态识别处于Ring0的特权或敏感指令进行二进制转译，模拟这些系统指令的行为并将其转发给物理硬件实现虚拟化。

全虚拟化的典型应用是VMware vSphere，其优点在于客户操作系统无需进行修改，具有极佳的兼容性，缺点是二进制转译对系统资源的消耗较重。得益于日渐成熟的硬件辅助虚拟化和一些降低转译开销的优化技术（例如转译缓存），全虚拟化的性能短板正在逐渐弱化。

1.2 半虚拟化

半虚拟化与全虚拟化的不同在于将虚拟化代码直接植入虚拟机操作系统，VMM无需捕获特权指令以及进行二进制转译，VMM通过智能编译器协助客户操作系统调用那些对虚拟化不友好的OS特权指令。

开源的Xen项目是半虚拟化的代表，它通过修改Linux内核实现对CPU、内存资源的虚拟，通过定制驱动程序实现对I/O的

虚拟。半虚拟化的实现模式虽然降低了系统的开销，但其性能很大程度上依赖于运行的负载。况且由于需要修改操作系统内核，可移植性方面也存在一定问题，尤其是对非开源的Windows系统兼容性较差。在实际工作中，也存在操作系统支持和维护困难等问题。

1.3 硬件辅助虚拟化

硬件辅助虚拟化最早出现在2006年，初衷是解决x86指令集中部分特权指令无法正常陷入和模拟的问题（详见下文中CPU虚拟化实现模式）。实现原理上，硬件辅助虚拟化技术为CPU增加了新的ROOT执行模式，并将VMM运行在这种模式下，以此将客户OS重新放置于Ring 0层，使关键指令可以在VMM上直接执行而无需进行二进制转译，同时无需修改操作系统内核，该项技术的出现对全虚拟化和半虚拟化的性能提升具有重大意义。

1.4 CPU虚拟化实现模式

CPU虚拟化是服务器虚拟化的基础，其他核心技术还包括内存虚拟化、设备和I/O虚拟化。通过对比三类CPU虚拟化的实现模式，可以更好地理解各类虚拟技术的特性和异同。

在传统的x86体系架构下，规定了四种特权级别（Ring0、Ring1、Ring2、Ring3）用于实现对底层硬件的访问控制。通常用户级应用一般运行在Ring 3（最低权限），操作系统运行在Ring 0（最高权限）直接访问内存和硬件，执行特权指令。对x86架构进行虚拟化时，需要在操作系统下运行虚拟化层，用于创建和管理虚拟机，进行共享资源的分配。在宿主机层面，虚拟化层也是一个操作系统（例如VMware的ESXi），但是由于x86体系架构不允许在同一个特权层出现多个操作系统，这为x86架构的虚拟化实现带来了一定的困难。

对这一问题的解决方案是CPU的特权解除（Privilege Deprivileging）和“先陷入后模拟”技术。特权解除通过降低客户操作系统的运行特权级别，实现了VMM对虚拟机的全盘控制。客户操作系统被解除特权后一般会降至Ring1或者Ring3级，这个现象也被称为特权压缩（Ring Compression）。在这种情况下，一部分需要在

Ring0层级执行的指令（例如GDT、IDT、LDT和TSS）将无法直接运行，只能通过先陷入后模拟的方式，极大地增加了系统的开销。

另外，x86指令集存在的固有漏洞也阻碍了CPU虚拟化的实现，在x86指令集中存在十多条非特权级别的敏感指令（指非Ring0级，但需转交VMM进行处理的危险指令）不能正常地陷入和模拟，这些指令可绕过VMM的监控在较低特权级别上直接执行。为了应对上述问题，在全虚拟化的模式下解决办法是“动态二进制转译”（Dynamic Binary Translation），在半虚拟化模式下需要对客户操作系统的源码进行修改。

无论是二进制转译还是修改操作系统内核都属于补偿性的措施，彻底解决“指令陷入和模拟”问题还需要从CPU指令集入手，硬件辅助的虚拟化技术如下图所示。

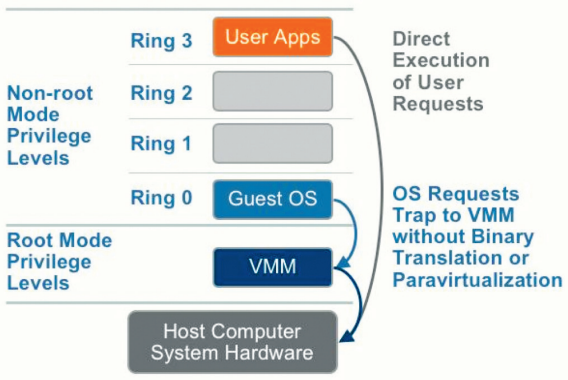


图 硬件辅助虚拟化实现模式

硬件辅助技术其实是对原有处理器架构的扩展补充，以Intel的VT-x为例，引入了两种专为虚拟化打造的操作模式：根操作模式（VMX Root Operation）和非根操作模式（VMX Non-Root Operation）。VMM运行在根模式下，客户操作系统运行在非根模式下，每个模式都设置了Ring0~Ring3四个特权级别。在这种架构下，原本工作在Ring0的VMM被转移到ROOT模式下运行，使得客户操作系统可以重回Ring0层级并顺利执行特权指令，削减了二进制转译带来的系统开销和修改内核代码引入的额外工作量。

2 服务器虚拟化基础设施性能优化概述

目前，绝大部分的x86服务器都采用完全虚拟

化的实现模式，下文中我们将以全虚拟化的经典产品VMware vSphere 5.5为例，阐述服务器虚拟化基础设施性能优化的核心思路和技术要点。

优化的整体思路方面，笔者建议在环境建设的早期（规划阶段）就启动相关工作，因为一旦系统正式上线运行，对底层硬件设备的配置修改或升级改造将可能会造成业务中断、资源浪费，甚至是数据损毁等严重后果。正确的优化流程至少包括三个步骤：业务需求分析、产品功能调研、制定采购和实施方案。通过分析核心业务应用的运行需求，包括内存需求、CPU需求、数据总量、数据吞吐量和时间敏感性需求等，结合物理设备的功能特性（例如是否支持硬件辅助虚拟化，是否支持增强型VMotion兼容性模式、容错Fault Tolerant等），最终制定精细合理的设备采购和实施方案。

对服务器虚拟化基础设施的优化内容和核心影响因素如下表所示。

表 基础设施性能优化项及核心影响因素

优化项	核心影响因素I	核心影响因素II
物理CPU	CPU支持硬件辅助虚拟化	CPU虚拟化技术 例如Intel VT-x/AMD-V
		MMU虚拟化技术 例如Intel EPT/AMD RVI
		I/O虚拟化技术 例如VT-d/AMD-vi
	CPU支持高可用性功能组件	VMotion (DRS、DPM) Fault Tolerance容错
	存储设备的合理配置	/
物理存储	存储网络带宽瓶颈	存储网络可用带宽 链路过载问题 SAN存储 iSCSI、NFS存储 插槽、存储接口卡速率
		vFRC
		VAAI
	新技术应用	/
物理网络	网络带宽瓶颈	/
	网络适配器选型	/
	BIOS版本升级	/
服务器BIOS	服务器CPU选项	硬件辅助虚拟化功能 Turbo Boost 睿频加速技术 Hyper-Threading 超线程功能
		BIOS禁用不使用接口
		/
	物理接口管理	/
	分布式电源管理	/

3 基础设施性能优化

物理设施的性能是整个虚拟化环境高效运行的基石，对基础设施的性能优化起始于项目的规划阶段。在此阶段应至少完成以下三项工作。

一是理解现有的物理环境。包括对集群宿主机现有计算、存储和网络资源的梳理。对存储设备的类型、容量、冗余和RAID情况的统计。对网络拓扑的梳理和理解，重点关注可能存在的网络带宽瓶颈，例如存储设备接口卡、宿主机物理网卡、存储网络带宽及上行连接的交换机端口带宽等。

二是分析业务应用的运行需求。包括对虚拟硬件的需求（CPU、内存、存储空间等）、安全性需求、连续性需求、是否时间敏感和网络吞吐量等。

三是调研物理设施的兼容性需求。重点关注各类硬件设备对需要启用的虚拟化功能组件的兼容性，以及由于vSphere软件版本导致的虚拟硬件的兼容性问题。

对基础设施的性能优化简言之就是“合理规划”。在完全熟悉当前物理环境的前提下，结合业务应用对软硬件环境的实际需求以及设备的兼容性要求，在预算范围内采购最合适的物理设备，配合在软件层面的合理配置（包括服务器BIOS、ESXi软件和虚拟机、客户操作系统及虚拟硬件），实现整体环境的协同、高效且稳定运行。

3.1 物理CPU

目前主流的处理器都支持硬件辅助虚拟化功能（Hardware-Assisted Virtualization），具体可概括为三项技术特征，硬件辅助CPU虚拟化、MMU虚拟化和I/O MMU虚拟化。这些基于硬件的辅助技术极大地改善了全虚拟化、半虚拟化的性能，因此推荐所有的企业级集群都采用此项功能。

3.1.1 CPU虚拟化（Intel VT-x/AMD-V）

在上文中，我们详细介绍了CPU虚拟化的实现模式，这里不再赘述。需要注意的是，在处理特权指令时硬件辅助的CPU虚拟化技术并非完全禁用了传统的二进制转译功能，而是根据工作负载的特性，让VMM从硬件辅助虚拟化和二进制转译中自主选择。此外，对于运行在Intel处理器上的64位操作系统，处理器必须支持硬件辅助的CPU虚拟化功能，在进行服务器选型时一定要注

意这一点。

3.1.2 MMU虚拟化 (Intel EPT/AMD RVI)

内存管理单元 (Memory Management Unit, MMU) 是一个硬件设备, 用于实现虚拟地址 (Virtual Address) 到物理地址 (Physical Address) 的映射, 以及内存访问权限检查。映射的实现基于页表 (Page Table) 和旁路转换缓冲 (Translation Lookaside Buffer, TLB)。在进行内存虚拟化时, 相比于传统物理环境, 物理地址即机器地址 (Machine Address), 但是对虚拟操作系统而言, 所“看到”的物理地址并非实际的机器地址, 也就是说客户OS对内存的访问需要经过两次地址转换。因此, VMM需要先对MMU进行虚拟化处理, 使客户OS可以顺利地访问主机内存。

在未使用硬件辅助虚拟化的系统中, 一般采用影子页表 (Shadow Page Tables) 完成从虚拟地址到物理地址、再到机器地址的映射。此时普通的内存调用将不会产生额外的开销, 但却需要一定的资源维持影子页表的同步更新。硬件辅助MMU虚拟化解决了影子页表的内在开销问题, 由硬件支持实现地址的两级映射, ESXi无需再消耗资源维护影子页表。

3.1.3 I/O虚拟化 (VT-d/AMD-vi)

除了CPU和内存虚拟化, 最后一个需要虚拟化的组件是设备I/O, 其本质是对虚拟设备和物理设备间I/O请求路径的管理。相较于对物理硬件的直接访问, 基于软件的I/O虚拟化具有更丰富的特性和更简化的管理方式。例如在网络方面, 虚拟网卡和虚拟交换机可以创建虚拟网络而无需消耗物理网络的带宽, 网卡组合 (NIC teaming) 使得多个物理网卡变成逻辑上的一块网卡, 对虚拟机而言物理网卡的故障转移是透明的。高效的I/O虚拟化关键在于既要保留虚拟化的优势同时也要着力降低对CPU资源的消耗。

对于硬件辅助的I/O虚拟化, 各大处理器厂商推出了硬件辅助的I/O MMU虚拟化功能 (Intel VT-d/AMD-vi), 该I/O内存管理单元实现了对DMA和I/O设备中断请求的重定向, 使得虚拟机可以直接访问物理网卡、存储控制器 (HBA卡) 和GPU等硬件I/O设备。

3.2 物理存储

存储性能优化是一个牵涉面很广的问题, 涉及工作负载、存储硬件、RAID类型、缓存大小和条带大小等。在规划时, 可以着重考虑以下三个核心问题: 一是后端存储设备的合理配置; 二是存储网络中的带宽瓶颈; 三是在ESXi中启用VAAI、vFRC等优化技术。

底层存储设备的配置极大地影响着环境的整体性能, 在部署前应参考存储厂商提供的最佳实践方案, 在此基础上对其他性能因素进行优化, 达到整体的最佳优化效果。

带宽瓶颈方面, 需要关注的因素包括:

1) 存储设备可用带宽。设备的可用网络带宽对时间敏感应用的影响极大, 此外集群的部分功能 (例如Storage vMotion) 对存储网络的带宽也有着很强的依赖性。

2) 链路过载问题。在VMware中虚拟网络和物理网络的交互通过上行链路实现, 所谓链路过载, 是指单条物理链路上承载了太多的虚拟数据流量, 例如当大数据流量的应用系统和iSCSI流量共用同一条传输链路时, 该链路将出现过载, 承载的应用系统和iSCSI的性能将会受到严重影响。链路过载的解决办法是分配充足的物理网卡, 为特殊的应用系统 (时间敏感、大数据流量)、iSCSI和NFS单独划分一块或多块物理网卡, 形成相互的隔离。

3) 对于端到端的光纤存储网络, 独立于以太网, 无需考虑链路过载的问题。但需要注意各端传输速率的一致性, 包括主机侧的HBA卡、光纤交换机和存储阵列上的光纤接口卡。

4) 对于基于以太网的iSCSI和NFS存储, 容易产生链路过载和丢包问题, 应尽量排除可能存在的网络带宽瓶颈。此外, 上述两种存储的网络协议解析是在ESXi宿主机内部进行的, 相比于其他类型的存储设备可能会消耗更多的CPU资源。

5) 设备插槽和存储接口卡传输速率兼容性问题。服务器上的插槽分为很多种, 支持不同的总线结构, 例如PCI、PCI-X、PCI Express (PCIe) 及PCIe 2.0 (aka PCIe Gen 2) 等, 需要保证存储接口卡接入相匹配的插槽。例如一块单端口的16Gb光纤HBA卡至少需要配备PCI Express (PCIe) G1

X8或PCIe G2 X4插槽，而双端口的16Gb光纤HBA卡至少需要配备PCIe G2 X8的插槽等。

新技术应用方面，vSphere 5.5集成了两款强大的存储性能优化技术：vFRC和VAAI。

vFRC：vSphere Flash读缓存（vSphere Flash Read Cache）是一种基于软件的高速缓存，通过将数据存储在本地读缓存中，显著提高虚拟机性能表现并降低延迟，大幅缩短负载的响应时间。实现时需要几个先决条件：vSphere Flash Infrastructure Layer、支持VMware vFRC软件的硬件、vSphere版本不低于5.5。其中vSphere Flash Infrastructure Layer由使用各类标准接口的本地Flash存储设备组成，包括使用SATA和串行SCSI（SAS）接口的硬盘式SSD产品，以及一些使用PCIe接口的卡片扩展式设备。

VAAI：vStorage阵列集成应用接口（vStorage APIs for Array Integration）。类似于处理器厂商推出的Intel-VT和AMD-V技术，VMware通过与存储厂商的合作预先将代码植入存储设备中，将一部分原本在ESXi中进行的复杂运算过程转到存储阵列中离线完成，从而减轻主机端的运算压力，提升工作效率。主要体现在以下三个方面：

1) 离线拷贝（Hardware – accelerated cloning或fullcopy或Copy Offload）。实现虚拟机的快速拷贝，适用于Storage vMotion、基于模板的虚拟机部署。

2) 离线并发写（Block Zeroing）。虚拟磁盘在初次写入时，ESXi会发送大量指令对磁盘进行清零，造成巨大的资源消耗。Block Zeroing将该工作转移至阵列中进行，无需主机端进行任何操作。

3) 硬件辅助锁定（Scalable Lock Management）。当多个虚拟机文件系统（Virtual Machine File System, VMFS）访问同一共享逻辑卷时，vSphere为防止冲突将启用保护机制锁定整个逻辑卷。通过硬件辅助的锁定功能，可以离线地对个别磁盘数据块进行锁定，而不影响主机对其他数据块的访问，以此提高存储空间的使用效率。

综上所述，物理存储的性能优化首先需要参考最佳实践做好存储设备的配置工作，其次是合理规划存储网络拓扑并分配足够的网卡资源，避

免潜在的带宽瓶颈和链路过载问题，最后是按需采购支持vFRC和VAAI功能的服务器和存储设备。

3.3 物理网络

物理网络包含上文中提到的存储网络，两者的优化思路非常相近。首先是对环境的熟悉和理解，其次是排除潜在的带宽瓶颈。需要考虑的核心问题包括网络传输设备、链路、接口卡带宽的一致性，插槽和接口卡的传输速率匹配等。

此外为了获得最佳的网络性能，推荐使用支持以下功能的网络适配器：

- 1) 数据包检验（Checksum Offload）。
- 2) TCP分段分载（TCP Segmentation Offload, TSO）。
- 3) 处理64位DMA地址。
- 4) 巨型帧处理。
- 5) LRO功能（Large receive offload）。

3.4 服务器BIOS

服务器BIOS的优化属于软件层面。软件层的性能优化其实涉及很多方面，例如宿主机的BIOS配置、ESXi软件和虚拟机的配置、客户操作系统参数和虚拟硬件配置等。

出于对ESXi软件和虚拟机的保护，建议在服务器首次开机时进行BIOS优化。优化项包括：

- 1) 将BIOS升级至当前最新版本。
- 2) 开启处理器的硬件辅助虚拟化功能（VT-X、AMD-V、EPT、RVI等）。
- 3) 确保启用全部CPU插槽及其上的核心。
- 4) 若支持，启用“Turbo Boost”英特尔睿频加速技术。
- 5) 若支持，开启Hyper-Threading超线程功能。
- 6) 对于支持非统一内存访问（NUMA）功能的系统，关闭节点交错（Node Interleaving）。
- 7) 禁用不使用的物理接口，例如串口、USB和网络接口等。
- 8) 将电源管理选项设置为“OS Controlled Mode”，帮助ESXi启用CPU的节能功能，实现对主机的能耗管理。

3.5 EVO：SDDC超融合服务器虚拟化

根据上文的介绍，可以看出对基础设施的性能优化需要进行大量的前期调研和精细的方案规

划，这里笔者推荐一个更为快捷的解决方案，那就是VMware在2015年的VMworld大会上推出的超融合服务器虚拟化产品EVO:SDDC。

EVO:SDDC是一个服务器虚拟化的“一揽子”解决方案，它以软件的形式集成在VMware合作伙伴的服务器设备中，其自动化管理引擎EVO:SDDC manager可以大幅简化虚拟资源池的部署与管理成本。其他核心组件包括：服务器虚拟化平台vSphere、网络虚拟化平台NSX、存储虚拟化平台VSAN以及虚拟化管理平台vCenter Server。

VMware认为，EVO:SDDC有能力在2 h内搭建一个具有8 000台虚拟机的私有云环境，足见其功能的强大。同时也从侧面说明，为适应企业高速膨胀的IT需求，云环境的快速部署和精简管理将是未来的发展方向。

(上接第577页)

图2所示，主要包括三个控制模块：存储平台、应用平台和管理平台，它们的构建都是在运用云计算技术和云基础设施。基于云计算技术的电能信息采集物联网通过云基础设施（包括配电室、用户侧的传感器网络、大型集群设备等）搜集和整理数据和信息，然后传递到应用层的云计算平台进行处理，其中云应用平台是最主要的模块，在这一平台进行数据交换操作，通过控制该平台计算数据、处理数据，使得物联网技术发挥作用；云存储平台所要存储的是物联网中的数据和信息，它实现所有用户共享虚拟存储池；云管理平台负责物联网整个平台的监控，保证平台内所有数据交换的顺利运行，保证数据的安全性、可靠性。

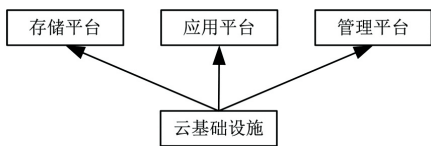


图2 云计算与物联网结合的基本框架

5 结束语

物联网技术是一项新兴的技术，可以有效促

4 结束语

服务器虚拟化的性能优化其实还包括许多方面，例如ESXi和虚拟机、客户操作系统和虚拟硬件的性能优化等，鉴于篇幅的限制不能一一详述。在对基础设施进行性能优化时，对各项影响因素的分析和考虑需要在项目的规划阶段开展，同时要求规划人员对应用系统的需求、现有物理环境和各类虚拟化功能有着深刻的理解和认识。

作者简介

陈乐然（1987-），男，硕士研究生，中级工程师，主要研究方向为信息安全。

（收稿日期：2015-12-10）EA
（编辑：崔文静）

进智能电网的发展，但是物联网技术的发展需要各项高端技术的融合，云计算与物联网技术的结合可以实现数据的准确采集、快速处理和可靠存储，达到高效计算和存储电能信息的目的，有效提高了电力企业的运营效率。

参考文献

[1] 白洋. 面向大数据的电力设备状态监测信息聚合研究[D]. 昆明：昆明理工大学，2014.
[2] 王淑祥，马素霞. 基于Hadoop的海量电能质量监测数据分析平台研究[J]. 中国科技信息，2013(13): 79-80.
[3] 陈全，邓倩妮. 云计算及其关键技术[J]. 计算机应用，2009，9(9): 2562-2567.
[4] 刘砚. 基于云计算的物联网系统架构研究[J]. 科技信息，2012(1): 209-210.

作者简介

吴宇（1980-），男，本科，电气工程师，主要研究方向为电力营销、用电信息采集和电力器具计量。

（收稿日期：2015-12-10）EA
（编辑：崔文静）