Contents

[A Comparison of Virtualization Technologies for HPC 2](#_Toc504832421)

[A GPGPU Transparent Virtualization Component for High Performance Computing Clouds 3](#_Toc504832422)

[A Survey on Virtualization Technologies 3](#_Toc504832423)

[An Evaluation of CUDA-enabled Virtualization Solutions 5](#_Toc504832424)

[High Performance Computing 7](#_Toc504832425)

# A Comparison of Virtualization Technologies for HPC

高性能虚拟化对比中的资料。

虚拟化是提高现有计算资源利用率的常用策略，特别是在数据中心内部。

然而，尽管这项技术能提高计算机资源的利用率，并且还能为其用户提供资源保证，但是目前，它在高性能计算应用中仍然受到很多限制。

虚拟化技术在计算机上的应用已经有30多年的历史了。

传统上，使用它就意味着为了获得虚拟机的便利而需要接受性能大幅降低的代价。如今，这种性能上的降低已经得好很好的缓解。由于更快的处理器的出现以及更高效的虚拟化解决方案的诞生，现在我们也能把一台普通的台式计算机作为虚拟机的宿主机来使用。

很快，大型计算集群将利用虚拟化的优势来提高集群的效用，同时可以减轻管理大量机器所产生的负担。事实上，虚拟机允许管理员更精确地控制其资源，同时保护主机节点免受用户软件的故障影响。这使得管理员可以从用户的角度提供几乎没有性能影响的“类似沙箱”的环境。

鉴于虚拟化的好处，我们相信基于虚拟机/虚拟服务器的计算集群将很快在高性能计算（HPC）社区中得到广泛采用。但是，到目前为止，还没有对各种虚拟化策略和实现进行全面的考察，特别是在HPC环境下。我们通过使用标准HPC基准检查集群性能的多个方面来填补这方面的文献空白。

Existing Virtualization Technologies

Full Virtualization：有时也称为硬件仿真。 在这种情况下，使用虚拟机管理程序运行未修改的操作系统，即时捕获并安全地转换/执行特权指令。由于捕获特权指令会导致性能显著下降，所以使用巧妙的策略来汇总多个指令并将它们一起转换。其他增强功能（例如二进制翻译）可以通过减少将来转换这些指令的需求来进一步提高性能。

Paravirtualization: 与完全虚拟化一样，半虚拟化也使用虚拟机管理程序，并使用术语虚拟机来表示虚拟化的操作系统。但是，与完全虚拟化不同，准虚拟化需要对虚拟化操作系统进行更改。由于使用特权指令是虚拟化系统性能损失的主要原因，它需要虚拟机与虚拟机管理程序进行协调，减少此类的性能丢失。其优点是半虚拟化的虚拟机通常比完全虚拟化的虚拟机要好。 然而，缺点是需要修改半虚拟化的虚拟机/操作系统，使其具有管理程序意识。这对没有足够可用源代码的操作系统会产生影响。

Operating System-level Virtualization：虚拟化最具侵入性的形式是操作系统级虚拟化。 与准虚拟化和完全虚拟化不同，操作系统级别的虚拟化不依赖于管理程序。相反，修改操作系统以在单个主机内安全隔离多个操作系统实例。 来宾操作系统实例通常被称为虚拟专用服务器（VPS）。 操作系统级虚拟化的优势主要在于性能。没有管理程序/指令陷阱是必要的。 这通常会导致近原生速度的系统性能。 主要缺点是所有的VPS实例共享一个内核。 因此，如果内核崩溃或受到威胁，则所有VPS实例都将受到威胁。 但是，拥有单个内核实例的好处是，由于多个内核的操作系统开销，消耗的资源更少。

Native Virtualization: 本地虚拟化利用硬件支持处理器本身内的虚拟化来帮助虚拟化工作。 它允许多个未修改的操作系统相互并行运行，前提是所有操作系统都能够直接在主处理器上运行。 也就是说，本地虚拟化不会模拟处理器。 这与完全虚拟化技术不同，在虚拟处理器上可以运行操作系统，通常性能较差。 在x86 64系列处理器中，Intel和AMD都通过Intel-VT和AMD-V虚拟化扩展来支持虚拟化。 支持虚拟化的x86 64处理器相对较新，但正在迅速普及。

# A GPGPU Transparent Virtualization Component for High Performance Computing Clouds

虚拟化技术是在网格上运行通用复杂的高性能科学软件的一个很有前途的工作，激发了一种新颖的计算模式，在这种模式中，虚拟化资源分布在真正的高性能硬件基础架构云中。在云计算中，硬件设备和软件应用分别通过硬件虚拟化软件即服务解决方案，使科学家有机会处理他们的具体研究需求[8]。特别是在并行计算应用领域，在云基础设施上实例化的虚拟集群在同一台真实机器上运行的虚拟机实例之间的消息传递性能差，而且不能够访问硬件特定的加速设备如GPU [11]， [20]。最近，科学计算经历了通用图形处理单元的加速数据并行计算任务。

一个最成功的基于GPU的加速系统是由nVIDIA提供的，并且依赖于支持高级语言工具的CUDA编程范例

目前，虚拟化不允许透明地使用加速器作为基于CUDA的GPU，因为虚拟/真实机器和访客/主机真实机器通信问题对基于弹性分配的资源的云计算基础设施的总体潜在性能造成严重的限制。

在本文中，我们提出的组件gVirtuS（GPU虚拟化服务）[1]是GPGPU透明虚拟化的结果，主要针对使用基于nVIDIA CUDA的加速板通过虚拟机实例来加速科学计算。

# A Survey on Virtualization Technologies

虚拟化是一种技术，它将计算资源结合或划分，以使用诸如硬件和软件划分或聚合，局部或全部机器模拟，仿真，分时等方法来呈现一个或多个操作环境。虚拟化技术在服务器整合，安全计算平台，支持多种操作系统，内核调试与开发，系统迁移等广泛领域找到了重要的应用，得到了广泛的使用。他们中的大多数向最终用户呈现类似的操作环境;然而，它们在其操作的抽象级别和底层架构上倾向于大不相同。本文调查了广泛的虚拟化技术，分析了它们的体系结构和实现，并提出了一个分类标准，根据它们的抽象级别进行分类。本文确定了以下抽象层次：指令集级，硬件抽象层（HAL）级，操作系统级，库级和应用级虚拟机。它研究每个类别的例子，并提供相对比较。它还提供了更广泛的虚拟化技术，并提供了可以扩展以适应这种分类下的未来虚拟化技术的见解。本文提出了一种非常轻量级的技术，我们称之为羽量级虚拟机（FVM），它可以用来在真实的环境中“试用”不受信任的程序，而不会对系统造成永久的损害。最后，它演示了FVM的有效性，将其应用于两个应用程序：安全移动代码执行和Windows程序的自动清理卸载。

虚拟机概念自20世纪60年代起就开始存在，当时它是由IBM首先开发的，用于提供对大型计算机的并发交互式访问。 每个虚拟机（VM）曾经是物理机器的一个实例，给用户一个直接访问物理机器的幻觉。 这是一个优雅而透明的方式，可以在昂贵的硬件上实现分时共享和资源共享。每个虚拟机都是底层系统的完全保护和隔离副本。 用户可以执行，开发和测试应用程序，而不用担心导致同一台计算机上的其他用户使用的系统崩溃。 因此，通过虚拟化来降低硬件购置成本，并通过让更多的用户同时进行工作来提高生产力。 随着硬件越来越便宜，多处理操作系统出现，虚拟机在20世纪70年代和80年代几乎灭绝。 随着二十世纪九十年代PC硬件和操作系统的多样化，虚拟化思想再次受到重视。 那么虚拟机的主要用途是在一台给定的机器上执行一系列的应用程序，最初的目标是针对不同的硬件和操作系统。 现在这个趋势还在持续。

只有在形式世界中，“虚拟性”才有别于“现实”，而具有相似的本质或效果。在计算机世界中，应用程序和世界其他地方的虚拟环境与真实环境的虚拟环境相同，尽管底层机制在形式上是不同的。由于各种原因，虚拟环境（或虚拟机）通常会呈现出与下面的物理机器（或资源）相比具有更多（或更少）能力的机器（或资源）的误导性图像。典型的计算机系统已经使用了许多这样的技术一个这样的例子是任何现代操作系统中的虚拟内存实现，它使进程使用内存通常远远超过其计算机必须提供的物理内存量。这个（虚拟内存）也使相同的物理内存可以在进程中被共享。类似地，多任务可以被看作是另一个例子，其中单个CPU以分时方式分区以向每个任务呈现某种虚拟CPU。在不同的设置中，可以将一组中速处理器组合在一起，以呈现具有非常高时钟速度的单个虚拟化处理器。当今世界有很多例子都是利用这种方法的。帮助建立这样的虚拟化对象的技术的伞可以说是实现有一个共同现象的任务，虚拟化。

随着虚拟化概念在计算机科学广泛领域的应用越来越多，定义的规模越来越大。但是，本文只讨论这个问题，我们使用下面这个宽松的定义：“虚拟化是一种技术，它将计算资源合并或划分，以使用诸如硬件和软件划分或聚合，局部或整体机器模拟等方法呈现一个或多个操作环境，仿真，分时等等“。尽管虚拟化通常意味着分割和聚合，但为了本文的目的，我们只集中于分解问题（因为这些问题更为普遍）。因此，虚拟化层使用较低级别的资源来提供基础架构支持，以创建彼此独立且彼此隔离的多个虚拟机。有时，这样的虚拟化层也被称为虚拟机监视器（VMM）。虽然传统上VMM被用来指硬件之上和操作系统之下的虚拟化层，但是在很多情况下我们可以用它来表示一个通用层。在实际情况下，虚拟化如何有用可能有许多原因，其中包括以下几点：

服务器整合：将多台未充分利用的机器的工作量整合到更少的机器上，以节省基础设施的硬件，管理和管理

应用程序整合：传统应用程序可能需要更新的硬件和/或操作系统。 通过虚拟化较新的硬件并提供对其他硬件的访问，可以很好地满足这种传统应用程序的需求。

沙箱：虚拟机可用于为运行外部或信任度较低的应用程序提供安全，隔离的环境（沙箱）。 因此，虚拟化技术可以帮助构建安全的计算平台。

多种执行环境：可以使用虚拟化来创建多个执行环境（以各种可能的方式），并且可以通过保证指定的资源量来增加QoS。

虚拟硬件：它可以提供从未有过的硬件，例如 虚拟SCSI驱动器，虚拟以太网适配器，虚拟以太网交换机和集线器等。

# An Evaluation of CUDA-enabled Virtualization Solutions

摘要

虚拟化作为一种能够以低成本和低能耗的方式实现简单而有效的资源共享的技术，不仅在企业中，而且在高性能计算中也越来越受欢迎。 具有严格性能需求的应用程序通常使用图形处理器来加速其计算。 因此支持GPU加速的虚拟化解决方案正变得越来越重要。 本文通过虚拟机对三种框架（通过CUDA支持GPU加速的rCUDA，gVirtuS和Xen）进行详细评估。 我们描述了这三种解决方案的架构，并对它们的保真度，性能，复用和插入特性进行了比较和对比。

介绍

虚拟化可以广泛地定义为一种技术，该技术应用抽象层来划分运行在资源上的软件的计算资源。 这种抽象允许在同一物理计算资源上同时执行多个可能不同的隔离执行环境，从而实现不同用户的应用程序之间的有效资源共享并保持高系统利用率。 另外，从底层物理资源，主机操作系统等的抽象，有助于传统应用程序在虚拟机上运行，即使在计算系统升级时也是如此。

由于其支持灵活的操作系统种类和独立的执行环境，高效的细粒度资源共享，提高了生产力，可靠性和可用性，虚拟化已经在企业和高性能计算领域得到普及[1] [2]。 在过去的几年中，这项技术的快速发展已经导致计算趋势向按需计算，云计算和软件即服务模式发生了范式转变。 行业和机构更倾向于从亚马逊，微软和谷歌等云供应商那里租用计算设备，只为他们消费的计算资源付费，而不是在基础设施和维护上进行大量投资。 应用程序托管在云中的虚拟机上，并可能利用“按使用付费”业务模式来增加销售量并产生收入。

传统的虚拟化主要针对处理器，网络和存储虚拟化。 然而，随着GPU作为大规模并行多核设备的发展，可以满足医疗成像，生物信息学，计算金融和科学计算等各个领域不断增长的应用需求，GPU虚拟化正在变得越来越重要。 通过设备仿真，半虚拟化和设备专用于虚拟化环境中的来宾虚拟机的硬件设备（诸如GPU）的可用性使得不同的方法能够虚拟化GPU。 英特尔VT-d提供的硬件帮助消除了大量的虚拟化开销，因此提供了访问GPU的简单机制[3]。

然而，尽管有上述因素，GPU虚拟化，尤其是通用计算，仍然是一个正在被业界和学术界积极探索的新兴领域。 NVIDIA最近刚刚发布了第一款全面虚拟化GPU加速桌面虚拟化解决方案，采用NVIDIA VGX技术，支持全面虚拟化的Kepler GPU [4]。 这个解决方案正处于开发的beta阶段，尚未作为商业产品发布。

虽然虚拟化已经积极研究了十多年，但是GPU虚拟化只是在近几年才得到广泛的探索。 GPU虚拟化是指将物理GPU的特征展现给虚拟机（VM）的技术，同时保留专用虚拟GPU对每个VM的幻想。 如Dowty和Sugerman [5]所提出的，GPU虚拟化可以大致分为前端和后端技术。 在前端虚拟化中，图形驱动程序堆栈驻留在主机或管理程序中，VM通过API远程处理（其中图形调用通过远程过程调用转发到外部图形堆栈）或设备仿真（其中 在虚拟机中模拟虚拟GPU）。在后端虚拟化中，图形驱动程序堆栈位于VM中，允许直接访问物理GPU。 固定传递是一种流行的后端虚拟化技术，其中物理GPU专门用于虚拟机。

传统上，GPU虚拟化的目标是虚拟化图形渲染，但最近的重点已转移到GPGPU计算的虚拟化。业界和学术界已经提出了一些最新的CUDA虚拟化框架（CUDA是NVIDIA公司提出的一个计划模型，公开了用于GPGPU计算的NVIDIA GPU硬件）。其中一些是采用API远程处理或呼叫转移技术的前端虚拟化解决方案，例如vCUDA [6]，rCUDA [7]，[8]，gVirtuS [9]和GViM [10]。其他公司，如Citrix XenServer [11]，Xen [12]和VMware的vSphere [13]，使用固定传递机制，使用英特尔VT-d的硬件支持[14]。上述固定传递解决方案仅正式支持用于图形渲染的桌面虚拟化，而不是用于通用计算。在我们的研究中，我们尝试了Xen PCI直通机制，通过做一些修改来完成通用计算的GPU访问，这在后面的章节C中讨论。我们没有用VMware的vSphere服务器来实现。

前端虚拟化解决方案（vCUDA，GViM，rCUDA，gVirtuS）支持不同虚拟机之间的图形卡复用。另外，vCUDA还支持挂起和恢复等功能，使客户端会话可以跨计算机迁移，同时保留硬件加速功能[6]，而GViM则提出了在虚拟机之间协调共享GPU资源的技术[10]。使用后端虚拟化的Vmware ESX通过称为VMDirectPath 110 [15]的功能将物理GPU设备专用于VM来支持GPU计算。这提供了非常接近本地执行的性能，代价是缺乏对多路复用的支持。商业上，某些云提供商（如Amazon Elastic Compute Cloud（Amazon EC2）[16]）使用GPU虚拟化来提供包括GPGPU计算在内的GPU计算服务。这些目前利用固定直通机制的后端虚拟化。 Amazon Cluster GPU实例使用Xen完全虚拟化环境（Xen HVM）上的Xen pci-passthrough [12]。 Peerl托管的GPU私有云[17]使用传递，而Zillianss V-GPU [18]使用API​​远程处理。

# High Performance Computing

在过去的十年中，所谓的高性能计算的定义发生了巨大的变化。 1988年，“华尔街日报”刊登了一篇名为“杀手微软的攻击”的文章。 描述了由许多小型廉价处理器组成的计算系统如何很快使大型超级计算机过时。 那时候，个人电脑呢？ 耗资3000美元，可以每秒钟执行25万次操作。 耗资20000美元可以执行300万次运行，而一台价值300万美元的超级计算机每秒可以执行1亿次运行。 因此，为什么我们不能简单地连接400台个人电脑，以120万美元的价格实现超级计算机的性能呢？

这个愿景在某些方面已经成为现实，但并不像“凶手”的原始支持者那样，理论设想。相反，微处理器的性能已经在超级计算机性能上无情地获得了。这有两个原因。首先，有更多的技术？空间？以提高个人电脑领域的性能，而80年代后期的超级计算机正在推动性能。而且，一旦超级计算机公司突破了一些技术障碍，微处理器公司就可以在几年之后迅速采用超级计算机设计的成功要素。第二个也许是更重要的因素是，个人和商业电脑市场的兴起对性能的要求不断提高。诸如3D图形，图形用户界面，多媒体和游戏等计算机使用是这个市场的推动因素。有了这样一个巨大的市场，可用的研究资金投入到为国内市场开发廉价的高性能处理器。由于以前的超级计算机制造商正在被工作站公司收购（Silicon Graphics收购Cray公司，Hewlett-Packard公司在1996年收购了Convex公司），这种趋势正在向更快的小型计算机发展。

结果几乎每个有电脑接入的人都有一些“高性能”处理。 随着这些新型个人计算机的高速度增长，这些计算机将遇到超级计算机上常见的所有性能挑战。

虽然并不是所有的个人工作站用户都需要了解高性能计算的细节，但是对这些系统进行编程以获得最佳性能的人员将从对这些最新高性能系统的优缺点的理解中获益。