

MR-IOV: 下一代数据中心 I/O 虚拟化技术

喻 波, 胡怀湘

(华北计算技术研究所, 北京 100083)

摘 要: 服务器是构建数据中心的基础设施, 与计算机历史上存储与计算分离类似, PCI Express 等互连技术的出现使下一代数据中心 I/O 走出机箱, 构建外设网络成为可能。而如何对 I/O 资源进行高效地管理, 虚拟化技术的提出很好地解决了这个问题。通过讨论由 PCI-SIG 组织提出的 MR-IOV(多根 I/O 虚拟化)技术以及 MR-IOV 如何应用于数据中心的互连体系和 I/O 设备的共享, 勾画出 MR-IOV 在刀片服务器中的应用前景, 同时分析了 MR-IOV 存在的问题和当前研究现状, 为下一代数据中心交换结构的设计提供一定的参考。

关键词: 数据中心; 服务器; PCI Express; MR-IOV; I/O 共享; 交换结构

中图分类号: TP393

文献标识码: A

文章编号: 1673-629X(2013)10-0091-04

doi: 10.3969/j.issn.1673-629X.2013.10.023

MR-IOV: I/O Virtualization Technology for Next Generation Data Center

YU Bo, HU Huai-xiang

(North China Institute of Computing Technology, Beijing 100083, China)

Abstract: Server is the infrastructure to construct a data center. Similar to the breakaway between storage and computing in computer history, the emergence of interconnection technology such as PCI Express makes the possibility of I/O to go out of chassis to construct peripheral network in a data center. Meanwhile, how to manage the I/O recourses efficiently is a big problem faced now; the proposed virtualization technology has the ability to solve the problem. By discussing MR-IOV (Multi-Root I/O Virtualization) technology and how to deploy MR-IOV to a data center interconnection and I/O devices sharing, outline the application prospect in blade servers. Additionally, analyze existent problems and research status of MR-IOV, supplying some references for the design of switching fabric of the next generation data center.

Key words: data center; server; PCI Express; MR-IOV; I/O sharing; switching structure

0 引言

数据中心由众多服务器组成, 是各种应用与数据的资源池。由于数据中心的访问量巨大, 采用高速串行总线 PCI Express (Peripheral Component Interconnect Express, PCIe)^[1]可以缓解系统 I/O 的性能瓶颈。计算机发展史上, Ethernet、Fiber Channel 和 InfiniBand 等网络的发展使存储与计算分离, 形成存储网络。与之类似, PCI Express 互连技术的迅速发展使 I/O 设备走出机箱, 形成独立的外设网络成为可能。如何对外设资源进行高效的管理与利用, 虚拟化技术提供了很好的解决方案。

目前虚拟化技术已经广泛应用于计算机体系结构的各个层次, 不同层次的虚拟化带来不同的虚拟化概念, 如 CPU 虚拟化、内存虚拟化和 I/O 虚拟化等^[2]。系统虚拟化屏蔽了底层物理硬件, 通过 VMM (Virtual Machine Monitor) 可以高效地对物理资源进行集中管理和使用。

PCI-SIG 组织致力于 PCIe 和 I/O 虚拟化等标准规范的制定与维护, 分别于 2007 年和 2008 年发布 SR-IOV (Single Root I/O Virtualization)^[3]和 MR-IOV (Multi-Root I/O Virtualization)^[4]规范。PCIe 的 MR-IOV 架构解决了相应的 I/O 性能和安全问题, 为下一代数据中心的 I/O 虚拟化提供了可行的部署方案。文

收稿日期: 2012-12-17

修回日期: 2013-03-25

网络出版时间: 2013-07-24

基金项目: 国家“863”高技术发展计划项目 (2009AA01A405)

作者简介: 喻 波 (1988-), 男, 江西宜春人, 硕士研究生, 研究方向为网络存储、虚拟化技术; 胡怀湘, 高级工程师, CCF 会员, 研究方向为网络存储、抗恶劣环境计算机。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20130724.0953.019.html>

中通过比较传统数据中心和下一代数据中心的拓扑结构,分析 MR-IOV 技术的实现和应用前景,并提出了 MR-IOV 当前仍然存在的问题与大规模部署的难度。

1 PCIe I/O 虚拟化

数据中心的虚拟化有三个层面:服务器、网络和存储^[5]。虚拟化技术为物理资源的更高效利用、功耗的降低、不同硬件平台系统实例的移植等提供了可能,但 I/O 性能仍然是限制虚拟机在数据中心大规模部署的一个重要因素^[6]。

传统的数据中心拓扑结构如图 1 所示。服务器中集成了网卡 (Network Interface Card, NIC) 和 FC 适配卡 (Host Bus Adapter, HBA),通过以太网或 FC 交换机与以太网和存储网络连接。传统数据中心广泛应用的虚拟化技术是服务器层面的虚拟化,即每一个服务器主机 (Host) 可以虚拟出多个系统映像 (System Image, SI) 或系统实例 (System Instance, SI),SI 作为不同的应用服务器提供服务。

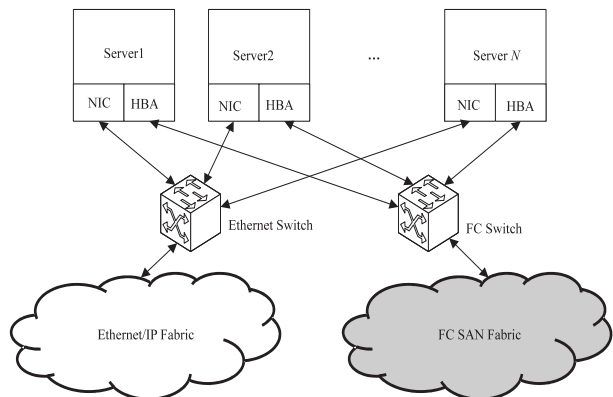


图 1 传统数据中心拓扑结构

传统的虚拟化环境中,多个 SI 共享一个单独的 I/O 设备,VM 管理软件需要处理不同的客户机操作系统 (Guest OS) 之间以及其分配的 I/O 设备之间数据的交换与传递,这是相当耗费资源的,尤其是虚拟机总线域的隔离和不同虚拟机 (Virtual Machine, VM) 间通信时信息流的安全^[7]。换句话说,一个 PCIe 设备在一个指定的时间内,只能与一个虚拟机 (VM₁) 绑定,而其他虚拟机 (VM₂) 访问 VM₁ 绑定的 PCIe 设备时,需要先向 VM₁ 发送请求,由 VM₁ 从 PCIe 设备获得数据后,再传送给 VM₂。使用这种方法将极大地增加虚拟机访问 PCIe 设备的延时,同时干扰其他虚拟机的正常运行。

PCI-SIG 组织提出的 SR-IOV 技术在此背景下诞生。SR-IOV 技术是单个主机的多个 VM 共享一个 PCIe 设备,将一个物理 PCIe 设备 (Physical Function, PF) 模拟成多个虚拟设备 (Virtual Function, VF),其中每一个虚拟设备与一个虚拟机绑定,从而便于不同的虚拟机访问同一个 PCIe 设备。

MR-IOV 技术增强了 SR-IOV 的功能,解决了多个处理器系统对一个 PCIe 总线域共享的问题,其本质是将一个物理 PCIe 总线域分解为多个虚拟的 PCIe 总线域,多个处理器系统可以与多个 PCIe 总线域对应,实现不同 PCIe 总线域的隔离^[8]。MR-IOV 的提出将全面改变现有数据中心的架构,I/O 与计算分离形成网络将是未来数据中心发展的方向。

PCI-SIG 组织提出的 MR-IOV 规范呈现出一个 I/O 虚拟化的图景,如图 2 所示。将 NIC 卡和 HBA 卡从服务器主机分离出来,通过 MR-IOV Switch 的连接形成 PCIe Fabric,每个服务器主机可以运行多个 SI,而通过对 PCIe Switch 的配置,可以虚拟出多个 NIC 和 HBA 卡,每个 SI 可以访问一个 PCIe 设备,从而实现 I/O 的共享。MR-IOV 的结构简化了数据中心的架构,可以明显降低功耗和成本,而且便于集中管理 I/O 资源,精简服务器计算单元。

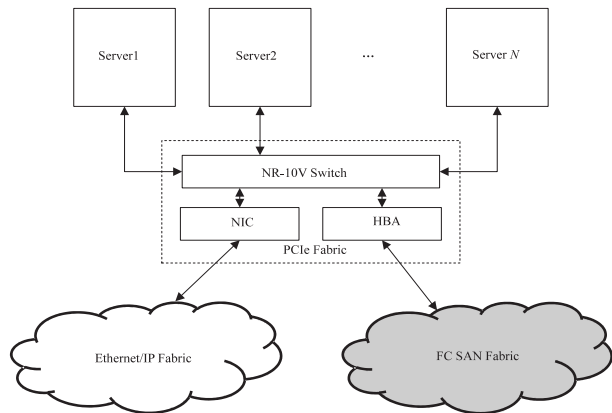


图 2 使用 MR-IOV 的下一代数据中心拓扑结构

2 MR-IOV Switch

MR-IOV Switch 在 MR-IOV 规范中称为 MRA (Multi-Root Aware) Switch,是构建下一代数据中心交换结构的核心部件。MR-IOV Switch 是一种支持 I/O 虚拟化的 PCIe Switch,其内部结构如图 3 所示。

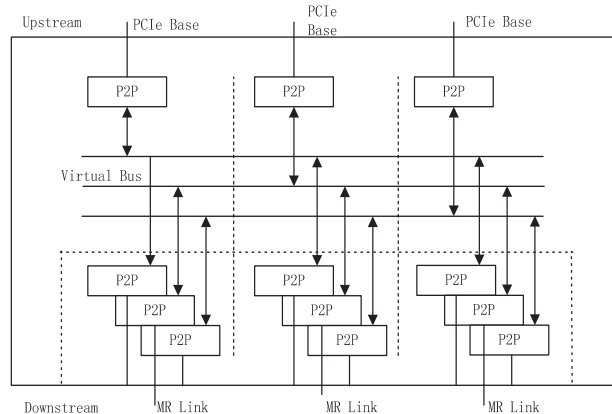


图 3 MR-IOV Switch 的内部结构

MR-IOV Switch 由多个上游端口和下游端口组

成,上游端口可以与多个 RP(Root Port)连接,这个 RP 可以是 MR-IOV RP 也可以是传统的 RP。下游端口可以与多个 EP(End Point)连接,也可以连接 SR-IOV 设备和传统的 PCIe 设备。

使用 MR-IOV Switch 可以组成多个虚拟 PCIe 总线域 VHs(Virtual Hierarchies),这些 VH 是通过软件 MR PCIM(Multi-Root PCI Manager)管理和维护的。如图 3 所示,MR-IOV Switch 由 3 组 P2P(PCI-to-PCI)桥组成,每一组 P2P 可以组成 1 个 PCIe 总线域,这 3 个 PCIe 总线域的地址空间独立,虚拟机可以对外部设备隔离访问。

现代计算架构越来越重视冗余和可靠性,MR-IOV Switch 结构通过 I/O 虚拟化能够提供 I/O 共享的功能,同时能够提供两个级别的冗余和故障切换,即 1+1(一主一备)模式和 $N+1$ (多主一备)模式。主机间通过中间结果暂存器等交换各自的状态信息,信息在虚拟总线(Virtual Bus)上传递,实现主机的热插拔和故障切换。

3 MR-IOV 应用前景

在下一代数据中心的建设中,MR-IOV 技术可能最先在刀片服务器中得到实现^[9]。刀片服务器是一种在标准高度的机箱内插装多个卡式的服务器单元,每一块“刀片”实际上是一块系统主板,通过“板载”硬盘启动自己的操作系统。

传统的刀片服务器采用的 I/O 互连方案如图 4 所示。6 个刀片分别有一个双通道 10 GB NIC 网卡和一个双通道 8 GB FC 适配器连接到冗余的交换模块。而采用 MR-IOV 方案的刀片服务器的优势是服务器单元可以不需要集成 I/O 设备,通过 MR-IOV Switch 实现服务器单元共享 I/O 设备的目的。如图 5 所示,使用 MR-IOV Switch 芯片来实现服务器链路的扩展与 I/O 的共享。

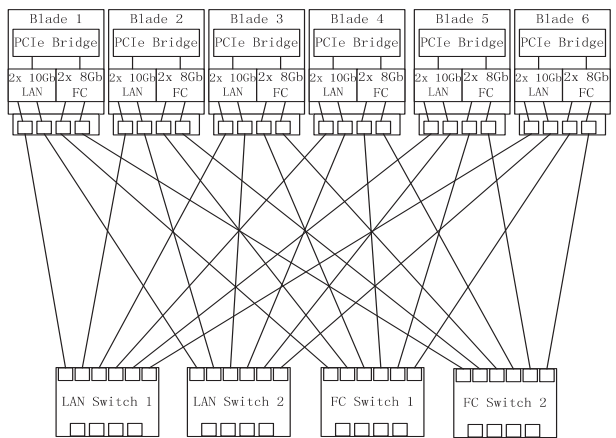


图 4 传统刀片服务器架构

从图 4 和图 5 两种刀片服务器的拓扑结构对比可

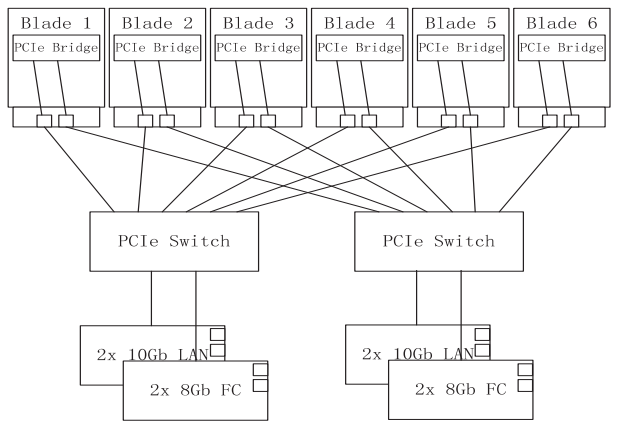


图 5 采用 MR-IOV 的刀片服务器架构

以看出,同样功能情况下,MR-IOV 方案的刀片结构得到很大程度的简化,通过 MR-IOV Switch 可以实现 I/O 设备的共享。两种结构使用的交换器(Switch)和适配器(Adapter)数量比较如表 1 所示。

表 1 相同功能的 MR-IOV 方案 and 传统方案所用适配器与交换器数量的比较

适配器和交换器数量	传统方案	MR-IOV 方案
10 GB Eth. PCIe 网卡	6	2
8 GB FC PCIe 适配器	6	2
10 GB Eth. 交换机	2	-
8 GB FC 交换机	2	-
PCIe 交换器	-	2

上面的例子中,相同功能的 MR-IOV 方案使用的 I/O 设备数量是传统方案的 1/3。由于 MR-IOV 的设备共享特性,刀片越多,设备数量呈线性递减,这大大降低了硬件成本,节约了主板空间,同时降低了服务器的功耗,而且 I/O 设备易于灵活配置,便于服务器功能的扩展。

MR-IOV 规范为硅片生产商提供了硬件设计的思路和方法。但是,MR-IOV Switch 作为 I/O 虚拟化的核心部件,其设计与实现的难度相当大。PLX 等公司在 2009 年发布了支持多根的 MR Switch 芯片,但截至目前仍没有 Switch 芯片支持多根 I/O 虚拟化,基于 PCIe Switch 的多根 I/O 虚拟化方案还在论证和实验阶段^[10-13]。

MR-IOV 的架构简化了数据中心的结构,却将复杂性留给了 MR-IOV Switch,而且 I/O 设备需要提供对 MR-IOV 的支持,这也将是限制 MR-IOV 大规模部署的一个重要因素。现阶段的替代方案是利用 MR Switch 构建服务器系统,通过 PCIe Switch 的非透明桥 NTB(Non-Transparent Bridging)等特性实现系统互连,从而达到 I/O 共享的目的^[14]。

MR Switch 只是实现了 MR-IOV 规范的部分功能,MR 系统方案当前也存在一些问题,如何配置

Switch,使之兼容各种处理器(x86、MIPS、ARM等)和操作系统(Windows、Linux等),这是使用现阶段的Switch构建服务器系统的关键。PLX、IDT、LSI等公司生产的MR Switch芯片支持的上游端口数量并不多,这不利于大规模部署数据中心数量庞大的服务器系统。如何利用现有的Switch来设计数据中心互连体系结构,实现数据中心的平稳升级,这也是当前需要研究的问题。

4 结束语

PCI Express的提出统一了局部I/O总线,I/O虚拟化成为数据中心基础架构的重要研究课题。MR-IOV的提出使I/O设备与机箱分离,形成外设网络,使对I/O资源的集中高效管理与利用成为可能。文中通过分析PCIe I/O虚拟化技术(MR-IOV)及其在数据中心的应用,勾勒出下一代数据中心I/O互连的拓扑结构及部署情况,分析了MR-IOV实现的难度以及现阶段MR系统方案的问题,对工程技术人员研究新型数据中心架构有一定的参考意义。

参考文献:

- [1] Wilen A H, Schade J P, Thornburg R. Introduction to PCI express - A hardware and software developer's guide[M]. [s. l.]: Intel Press, 2003.
- [2] 英特尔开源软件技术中心, 复旦大学并行处理研究所. 系统虚拟化-原理与实现[M]. 北京: 清华大学出版社, 2009.
- [3] PCI-SIG. Single root I/O virtualization and sharing specification[S]. 2007.
- [4] PCI-SIG. Multi-Root I/O virtualization and sharing specification[S]. 2008.
- [5] Chen Jyh-shing. Virtualization practices: Providing a complete virtual solution in a box[EB/OL]. 2012[2012-11-15]. <http://www.snua.org/education/tutorials/2012/fall#virtualization>.
- [6] Rixner S. Network virtualization: Breaking the performance barrier[J]. ACM Queue, 2008, 6(1): 36-43.
- [7] Homölle B, Schröder B, Brütt S. Multi root I/O virtualization (MR-IOV)[C]//Proceedings of the 1. GI/ITG KuVS Fachgespräch Virtualisierung. Paderborn, Germany: [s. n.], 2008: 11-18.
- [8] 王 齐. PCI Express 体系结构导读[M]. 北京: 机械工业出版社, 2010: 358-362.
- [9] PLX Technology, Inc. ExpApp 57 - 8648 in bladed systems[EB/OL]. 2007[2012-11-15]. http://www.plxtech.com/files/pdf/apps/ExpApp57_8648_Servers.pdf.
- [10] Krishnan V. Towards an integrated IO and clustering solution using PCI express[C]//Proceedings of the 9th IEEE International Conference on Cluster Computing. Austin, Texas, USA: [s. n.], 2007: 259-266.
- [11] IDT Corporation. Using PCI express as the primary system interconnect in multiroot compute, storage, communications and embedded systems (White Paper)[EB/OL]. 2008[2012-11-15]. <http://www.idt.com/document/idt-pcie-multi-root-white-paper>.
- [12] Bert L. Accelerating storage performance in virtualized servers using SR-IOV and MR-IOV[J]. SNS Europe, 2011(4): 8-9.
- [13] Waldspurger C, Rosenblum M. I/O virtualization[J]. Communications of the ACM, 2012, 55(1): 66-72.
- [14] Wong H. PCI express multi-root switch reconfiguration during system operation[D]. Massachusetts: Massachusetts Institute of Technology, 2011.

(上接第 90 页)

- 2005: 100-110.
- [4] 朱仲英. 传感网与物联网的进展与趋势[J]. 微型电脑应用, 2010, 26(1): 1-3.
- [5] 朱沛胜, 段世惠. 泛在网络发展现状分析[J]. 电信网技术, 2009(7): 18-22.
- [6] 崔 莉, 鞠海玲, 苗 勇, 等. 无线传感器网络研究进展[J]. 计算机研究与发展, 2005, 42(1): 163-174.
- [7] Heinzelman W R, Chandrakasan A, Balakrishnan H. Energy-efficient communication protocol for wireless microsensor networks[C]//Proceedings of the 33rd Annual Hawaii International Conference on System Sciences. [s. l.]: IEEE, 2000: 3005-3014.
- [8] Manjeshwar A, Agrawal D P. TEEN: a routing protocol for enhanced efficiency in wireless sensor networks[C]//Proceedings of the 15th Parallel and Distributed Processing Symposium. San Francisco: IEEE Computer Society, 2001: 2009-2015.
- [9] Lindsey S, Raghavendra C S. PEGASIS: power efficient gathering in sensor information systems[C]//Proc of the IEEE Aerospace Conf. New York: IEEE Press, 2002: 1125-1130.
- [10] Peng D, Zhang Q. An energy efficient cluster-routing protocol for wireless sensor networks[C]//Proc of ICCDA. [s. l.]: IEEE, 2010: 530-533.
- [11] Mollanejad A, Khanli L M, Zeynali M, et al. EHRP: novel energy-aware hierarchical routing protocol in wireless sensor network[C]//Proc of ICUMT. [s. l.]: IEEE, 2010: 970-975.
- [12] Nawaz R, Hussain S A, Abid S A, et al. Beaconless multihop routing protocol for wireless sensor networks[C]//Proc of ICCSN. [s. l.]: IEEE, 2011: 721-725.

MR-IOV：下一代数据中心I/O虚拟化技术

作者：[喻波](#)，[胡怀湘](#)，[YU Bo](#)，[HU Huai-xiang](#)

作者单位：[华北计算技术研究所, 北京, 100083](#)

刊名：[计算机技术与发展](#)



英文刊名：[Computer Technology and Development](#)

年，卷(期)：2013(10)

本文链接：http://d.wanfangdata.com.cn/Periodical_wjfz201310023.aspx