随着服务器虚拟化技术的成熟，数据中心部署的虚拟化服务器数量越来越多。虚拟机①的出现使数据中心服务器网络接入层出现了一个被称为VEB（Virtual Ethernet Bridge）的网络层。在服务器上采用纯软件方式实现的VEB就是通常所说的“VSwitch”。虽然VSwitch的实现方式简单，技术兼容性好，但也面临着诸多问题，例如VSwitch占用CPU资源导致虚拟机性能下降、虚拟机的流量监管、虚拟机的网络策略实施以及VSwitch管理可扩展性等问题。借助支持SR-IOV特性的网卡可以实现基于硬件的VEB。相比VSwitch，硬件EVB减少了CPU开销，但这种硬件方案与VSwitch类似，仍然不能有效解决虚拟机流量监管、网络策略实施及管理可扩展性问题。

为此，IEEE Data Center Bridging （DCB）任务组（DCB任务组是IEEE 802.1工作组的一个组成部分）正在制定一个新标准——802.1Qbg Edge Virtual Bridging（EVB），该标准将VEPA（Virtual Ethernet Port Aggregator）作为基本实现方案。VEPA的核心思想是，将虚拟机产生的网络流量全部交由与服务器相连的物理交换机进行处理，即使同一台服务器上的虚拟机间流量，也将在物理交换机上查表处理后，再回到目的虚拟机上。VEPA方式不仅借助物理交换机解决了虚拟机间流量转发，同时还实现了对虚拟机流量的监管，并且将虚拟机接入层网络纳入到传统服务器接入网络管理体系中。EVB标准同时还定义了VEB、VEPA和Director IO的混和实现方案，即多通道技术（Multichannel Technology）。

**一、 VEB（Virtual Ethernet Bridge）**

**1. 虚拟化服务器与VEB**

虚拟化服务器①是使用虚拟化软件（如VMWare ESX、Citrix XEN）在一台物理服务器上虚拟出一台或多台虚拟机（Virtual Machine，VM），安装在服务器上的虚拟化软件被称为VMM（Virtual Machine Monitor）。虚拟机运行在一个隔离环境中，是具有完整硬件功能的逻辑服务器，每个虚拟机具有自己的操作系统和应用程序。一台服务器上的多个虚拟机可以互不影响的同时运行，并复用物理机资源。虚拟化软件为虚拟机提供一套虚拟的硬件环境，包括虚拟的CPU、内存、存储设备、I/O设备（如网卡）以及虚拟交换机等，如图1。



图1虚拟化服务器模型

在虚拟化服务器中，虚拟以太网交换机是一个比较特殊的设备，具有重要的作用。虚拟机是通过虚拟交换机向外界提供服务器的。在虚拟化的服务器中，每个虚拟机都变成了一台独立的逻辑服务器，它们之间的通信通过网络进行。每个虚拟机被分配了一个虚拟网卡（不同的虚拟机网卡有不同MAC地址）。为实现虚拟机之间以及虚拟机与外部网络的通信，必须存在一个“虚拟以太网交换机”以实现报文转发功能。在IEEE的文档中，“虚拟以太网交换机”正式英文名称为“Virtual Ethernet Bridge②”，简称VEB。

VEB可以在VMM中采用纯软件方式实现，也可以借助支持SR-IOV特性的网卡通过全硬件方式实现。

**2. 软件VEB的技术特性**

软件VEB（或称VSwitch③）是目前较为成熟且产品化较好的技术方案。在一个虚拟化的服务器中，VMM为每个虚拟机创建一个虚拟网卡，对于在VMM中运行的VSwitch，每个虚拟机的虚拟网卡对应到VSwitch的一个逻辑端口上，服务器的物理网卡对应到VSwitch与外部物理交换机相连的上行逻辑端口上。

虚拟机的报文接收流程如下：VSwitch首先从物理网卡接收以太网报文，之后根据VMM下发的虚拟机MAC与VSwitch逻辑端口对应关系表（静态MAC表，VSwitch通常不做MAC地址学习）来转发报文。

虚拟机报文发送流程为：当报文的MAC地址在外部网络时，VSwitch直接将报文从物理网卡发向外部网络；当报文目的MAC地址是连接在相同VSwitch上的虚拟机时，则VSwitch通过虚拟机与逻辑网卡的对应关系表来转发报文。如图2所示。

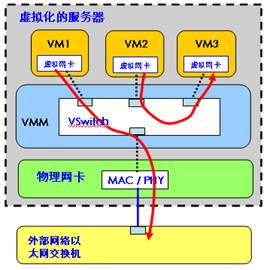


图2 VSwitch方案架构

VSwitch方案具有以下优点：

· 虚拟机间报文转发性能好。VSwitch实现虚拟机之间报文的二层软件转发， VSwitch对报文的转发能力只受限于CPU性能、内存总线带宽，因此虚拟机间报文的转发性能（带宽、延迟）非常好；

· 节省接入层物理交换机设备。例如，数据中心需要部署WEB服务器，且WEB服务器网关指向防火墙。这里可将一台服务器虚拟化成多个虚拟机，每个虚拟机作为一个WEB服务器，将VSwitch作为WEB服务器的网络接入层设备，将服务器物理网卡与防火墙端口互联即可完成组网，无需部署物理交换机设备；

· 与外部网络的兼容性好。VSwitch采用软件实现，对现有网络标准的兼容性好，所以VSwitch与外部网络设备不存在互联兼容性问题。

但VSwitch方案也存在一些缺点：

· 消耗CPU资源。虚拟机产生的网络流量越高，则基于软件实现的VSwitch就需要占用越多的CPU资源用于报文的转发处理，从而减弱了服务器支持更多虚拟机的能力。特别是在虚拟机到外部网络的流量很大时，CPU的开销会更大；

· 缺乏网络流量的可视性。VSwitch缺少内部流量监管能力，例如端口报文统计、端口流镜像、Netstream等特性。上述特性的缺失，一方面导致虚拟机之间的流量无法被网管系统所监管；另一方面也使得网络发生故障时，难于定位问题原因；

· 缺乏网络控制策略的实施能力。当前数据中心接入交换机都具有很多实现网络控制策略的特性，例如端口安全，QOS、ACL等。而VSwitch因顾及到CPU开销问题，通常不支持上诉特性。因此限制了数据中心的端到端网络控制策略（如端到端的QOS、整网安全部署策略等）的部署能力；

· 缺乏管理可扩展性。随着数据中心虚拟机数量的增加，VSwitch的数量随之增加，而传统的VSwitch必须被单独的配置管理，由此增加了网络的管理工作量。VMWare公司推出的“分布式交换机（DVW）”技术，可以将最多64个VSwitch作为一个统一的设备进行管理。但这种技术只有限的改善了管理扩展性问题，并未从根本上解决外部网络管理与VSwitch管理的统一性问题。

**3. 硬件VEB的技术特性**

硬件VEB的设计思想是将VEB功能从VMM移植到服务器物理网卡，通过网卡硬件改善VSwitch占用CPU资源而影响虚拟机性能的问题。这种硬件VEB方案必须采用支持SR-IOV（Single-Root I/O Virtualization）特性的PCIe网卡。

SR-IOV是对PCIe（PCI Express)标准的一种延伸，它可以使一个PCIe适配器（如网卡）通过一个共享的PCIe物理接口同时起到多个独立器件（多个独立网卡）的作用。如图3，物理服务器上的虚拟机可直接访问PCIe网卡上的寄存器和I/O队列, 由于绕过了VMM对虚拟机I/O请求的代理中转过程，可显著减少CPU开销，并且降低虚拟机访问外部网络的时延，这种工作模式也被称为“director IO”。

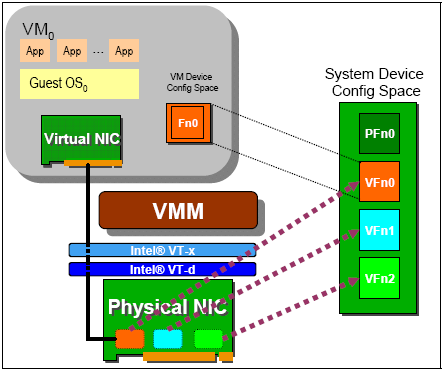


图3 SR-IOV实现PCIe虚拟化

采用硬件VEB方案时，VMM只实现SR-IOV网卡驱动程序（资源分配、中断处理等），不再参与虚拟机与外部网络，以及虚拟机之间的报文转发流程。如图4，对于从虚拟机发往外部网络的报文，由虚拟机操作系统的驱动程序直接操作网卡寄存器进行发送；对于外部网络发往虚拟机的报文，网卡根据目的MAC，将报文放入虚拟机对应的接收队列，虚拟机操作系统的驱动程序再通过DMA或中断方式进行接收处理；对于虚拟机之间的报文转发，网卡芯片通过查MAC表（静态配置，通常不支持MAC学习）进行转发处理。

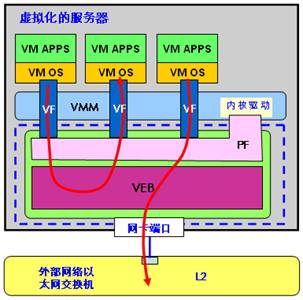


图4 基于SR-IOV的硬件VEB

基于SR-IOV网卡技术的硬件VEB方案的优点如下：

· 相对于VSwitch（软件VEB），减少了CPU占用率。借助SR-IOV技术将报文转发硬件化，不再需要类似VSwitch软件参与报文处理；

· 提高了虚拟机操作系统访问物理网卡的性能。尤其是当虚拟机与外部网络有大量报文流量时，由于实现了虚拟机对PCIe设备的直接访问和操作，显著降低了从虚拟机到物理网卡的报文处理延时；

· 支持SR-IOV的网卡，可虚拟化出256个虚拟网卡，提高了物理服务的网络虚拟化能力。

硬件VEB改进了VSwitch（软件VEB）的性能，但硬件方案仍然存在一些局限性：

· 缺乏网络可视性。支持SR-IOV的网卡不能解决网络流量监管问题。考虑到网卡成本，通常不会在网卡上添加流量监管处理芯片（如流量镜像、Netstream、端口报文统计信息等）。而且，由于硬件实现缺少灵活性，有可能硬件VEB的网络流量监管特性比VSwitch方案还少；

· 缺乏网络控制策略的实施能力。同样是考虑到网卡成本的问题，很难在网卡上实现基于硬件的网络控制策略（如QOS、ACL等）；

· 缺乏管理的可扩展性。每个物理网卡就是一个硬件VEB，所以随着虚拟化服务器数量的增加，网络中会存在很多硬件VEB需要分别进行管理，由此带来了管理可扩展性问题。相比VSwitch方案，由于一个VSwitch可包含多个物理网卡（如网卡捆绑上行），而硬件VEB则必须是每个物理网卡对应一个VEB，所以硬件VEB的管理可扩展性比VSwitch还差；

· 缺乏主流操作系统产品的支持。为支持SR-IOV技术，需要虚拟机的操作系统与VMM采用“准虚拟化 （Paravirtualized）”工作模式，这种模式要求虚拟机操作系统内核做必要的修改，以支持director IO。从目前了解到的情况看，主流操作系统厂商都不支持这种模式。“准虚拟化 + Director IO”方式主要用在Linux操作系统以及基于Linux内核实现的虚拟化技术上（XEN）。

**4. VEB技术特性总结**

VEB实现了虚拟机之间以及虚拟机与外部网络之间的报文转发。在VEB的两种实现方案中，软件VEB（VSwitch）的通用性好，虚拟机之间的转发性能好；硬件VEB基于网卡SR-IOV技术实现，对CPU资源占用小。但两种VEB方案始终没有解决的问题是：虚拟机间流量的监管、网络策略的部署，以及管理可扩展性。

**二、 802.1Qbg EVB标准**

**1. EVB标准的设计思想**

正是认识到软件和硬件VEB的局限性，IEEE 802.1工作组正着手制定一个新标准802.1Qbg Edge Virtual Bridging（EVB），**其核心思想是，将虚拟机产生的网络流量全部交由与服务器相连的物理交换机进行处理，即使同一台服务器的虚拟机间流量，也发往外部物理交换机进行转发处理**。如图5，由VM1发往VM2或VM3的报文，首先被发往外部交换机，查表后，报文沿原路返回服务器。802.1Qbg标准的作者——HP 的Paul Congdon，将这种工作模式形象的描述为“发卡弯④（hairpin turn）”转发。

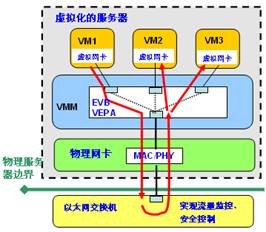


图5 EVB的基本架构

EVB改变了传统的VEB对报文的转发方式，使得大多数报文在外部网络交换机被处理。与VEB相似的，EVB可以通过软件方式实现（类似于VMM中的VSwitch），也可以在支持SR-IOV技术的网卡上实现。由于将所有流量都引向外部交换机，因此与虚拟机相关的流量监管、控制策略和管理可扩展性问题得以很好的解决。但是，由于流量被从虚拟机上引入到外部网络，EVB技术也带来了更多网络带宽开销的问题。例如，从一个虚拟机到另一个虚拟机的报文，占用的网络带宽是传统方式的两倍，其中一半带宽用于从源虚拟机向外网交换机传输，另一半带宽用于从外部交换机向目的虚拟机传输。EVB的出现并不是完全替换VEB方案，但是EVB对于流量监管能力、安全策略部署能力要求较高的场景（如数据中心）而言，是一种优选的技术方案。

以太网交换机在处理报文转发时，对于从一个端口上收到的报文，不会再将该报文从该端口发回。因此，当使能EVB特性的服务器接入到一个外网交换机上时，该交换机相应端口必须支持上述“发卡弯”转发方式。当前大多数交换机的硬件芯片都能支持这种“发卡弯”转发特性，只要修改驱动程序即可实现，不必为支持“发卡弯”方式而增加新的硬件芯片。

另一个由EVB技术引起的变化是服务器对从外部网络接收到组播或广播报文的处理方式。由于EVB从物理网卡上收到的报文可能是来自外部交换机的“发卡弯”报文，也就是说报文源MAC是虚拟化服务器上的虚拟机MAC，这种报文必须进行过滤处理，以避免发送该报文的虚拟机再次从网络上收到自己发出的组播或广播报文。因此，当前的操作系统或网卡驱动都需要做相应的修改。

EVB标准有如下的技术特点：

· 借助“发卡弯”转发机制将外网交换机上的众多网络控制策略和流量监管特性引入到虚拟机网络接入层，简化了网卡的设计，减少了虚拟网络转发对CPU的开销；

· 使用外部交换机上的控制策略特性（ACL、QOS、端口安全等）实现整网端到端的策略统一部署；

· 使用外部交换机增强虚拟机的流量监管能力，如各种端口流量统计、Netstream、端口镜像等。

前文仅描述了EVB的设计思路以及实现EVB方案带来的好处。实际上EVB定义了两种报文转发方案：VEPA（Virtual Ethernet Port Aggregator）和多通道（Multichannel Technology）。VEPA是EVB标准定义的基本实现方案，该方案不需要对虚拟机发出的以太网报文做改动即可实现“发卡弯”转发。多通道技术则定义了通过标签机制实现VEB、Director IO和VEPA的混和部署方案，借助多通道技术，管理员可以根据网络安全、性能以及可管理等方面的需求，来选择虚拟机与外部网络的接入方案（VEB、Director IO或VEPA）。

**2. EVB的基本实现方案——VEPA**

IEEE 802.1工作组在VEPA技术基础上实现IEEE 802.1Qbg EVB标准，是因为VEPA技术对当前网卡、交换机、现有以太网报文格式和标准影响最小。如图5。

VEPA的实现是基于现在的IEEE标准，不必为报文增加新的二层标签，只要对VMM软件和交换机的软件升级就可支持VEPA的“发卡弯”转发。

与VEB方案类似，VEPA方案可以采用纯软件方式实现，也能够通过支持SR-IOV的网卡实现硬件VEPA。其实，只要是VEB能安装和部署的地方，就都能用VEPA来实现，但VEB与VEPA各有所长，并不存在替代关系。VEPA的优点在于，完全基于IEEE标准，没有专用的报文格式。而且容易实现，通常只需要对网卡驱动、VMM桥模块和外部交换机的软件做很小的改动，从而实现低成本方案目标。

**3. 对VEPA的增强——通道技术（Multichannel Technology）**

多通道技术是通过给虚拟机报文增加IEEE标准报文标签，以增强VEPA功能的一种方案，由HP公司提出，最终被IEEE 802.1工作组接纳为EVB标准的一种可选方案。

多通道技术方案将交换机端口或网卡划分为多个逻辑通道，并且各通道间逻辑隔离。每个逻辑通道可根据用户需要定义成VEB、VEPA或director IO的任何一种。每个逻辑通道作为一个独立的到外部网络的通道进行处理。多通道技术借用了802.1ad S-TAG（QinQ）标准，通过一个附加的S-TAG和VLAN-ID来区分网卡或交换机端口上划分的不同逻辑通道，如图6。多通道技术可组合出多种方案：

· 多个VEB或VEPA共享同一个物理网卡。管理员可能需要特定虚拟机使用VEB，以获得较好的交换性能；也可能需要其他的应用使用VEPA，以获得更好的网络控制策略可实施性和流量可视性，并要求上述使用VEB或VEPA的虚拟机同时部署在一个物理服务器上。对于这些情况，管理员通过多通道技术即可解决VEB与VEPA共享一个外部网络（网卡）的需求。

· 直接将一个虚拟机映射到物理网卡上（Director IO），而其它的虚拟机仍然通过VEB或VEPA共享物理网卡。传统的VMM软件都能支持直接将虚拟机映射到物理网卡上，但这个虚拟机将独占这个物理网卡，由此造成网络资源的浪费。多通道技术使外部物理交换机通过报文的S-TAG识别网络流量来自哪个VEAP/VEB，或来自哪个直接映射的网卡。反之亦然。

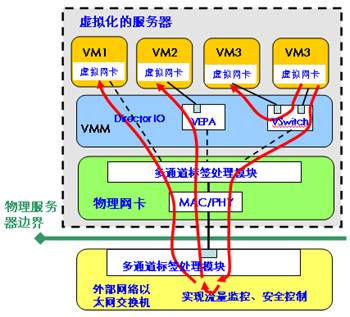


图6 多通道技术的架构

多通道技术需要网卡和外部交换机支持S-TAG和QinQ操作。所以在某些情况下，可能要求网卡或交换机做硬件升级，而VEPA方案对设备硬件没有要求，几乎在所有的VMM和外部物理交换机都能实现。多通道技术为管理员提供了选择虚拟机与外部网络接入方式的手段。对于一台物理服务器上的多个虚拟机，管理员可根据网络安全、性能及可管理性的需求，采用VEB、Director IO或VEPA中的任何一种方式为虚拟机提供网络接入。

**4. 关于802.1Qbh Bridge Port Extension标准**

端口扩展设备是一种功能有限的物理交换机，通常作为一个上行物理交换机的线卡使用。端口扩展技术需要为以太网报文增加TAG，而端口扩展设备借助报文TAG中的信息，将端口扩展设备上的物理端口映射成上行物理交换机上的一个虚拟端口，并且使用TAG中的信息来实现报文转发和策略控制。

当前市场上已有端口扩展设备，如Cisco的Nexus 2K就是Nexus 5K的端口扩展器。VN-TAG是Cisco为实现端口扩展而定义的一种私有以太网报文标签格式，这种报文格式不是建立在IEEE已定义的各种标准之上。VN-TAG为报文定义了虚拟机源和目的端口，并且标明了报文的广播域。借助支持VN-TAG技术的VSwitch和网卡，也能够实现类似EVB多通道的方案，但是VN-TAG技术有一些缺点：

· VN-TAG是一种新提出的标签格式，没有沿用现有的标准（如IEEE 802.1Q、 IEEE 802.1ad、IEEE 802.1X tags ）。

· 必须要改变交换机和网卡的硬件，而不能只是简单的对现有的网络设备软件进行升级。也就是说，VN-TAG的使用需要部署支持VN-TAG的新网络产品（网卡、交换机、软件）。

最初IEEE 802.1工作组曾考虑将“端口扩展”作为EVB标准的一部分，但是最终决定将端口扩展发展成一个独立的标准，即802.1 Bridge Port Extension。Cisco曾向IEEE 802.1Q工作组建议，将其私有的VN-TAG技术作为实现EVB的一种可选方案，但工作组最终没有接纳这个提案。Cisco最近修改了VN-TAG技术草案，修改后的草案称为M-TAG，该方案的主要目标仍是为了实现端口扩展设备与上行交换机之间的通信标准化。

**5. 802.1Qbg EVB的标准化进程**

在本文写作的时候，IEEE 802.1Qbg的授权请求（PAR）已经通过，正式的标准化过程正在进行中。IEEE 802.1选择VEPA技术草案作为EVB标准的基础，因其使用现有的技术标准，并对现有的网络产品和设备产生的影响最小。多通道技术作为一种可选项，也在EVB标准中定义。多通道技术提出了一种标准化TAG机制，以实现VEPA、VEB及Director IO的灵活部署。同样的情况，IEEE 802.1gbh的授权请求（PAR）也通过了，正式的标准化过程正在进行中。IEEE 802.1工作组已接受了Cisco提出的M-TAG技术草案，将其作为端口扩展标准的基础。

**三、 结束语**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **VEB** | | **802.1Qbg EVB** |
| **VSwitch** | **SR-IOV网卡** |
| 虚拟机间报文转发性能 | 5 | 4 | 3 |
| 虚拟机与外部网络报文转发性能 | 2 | 5 | 4 |
| 对服务器CPU的开销 | 1 | 5 | 3 |
| 虚拟机间流量的可视性 | 1 | 1 | 5 |
| 虚拟机流量的网络控制策略部署能力 | 1 | 1 | 5 |
| 管理可扩展性 | 4 | 2 | 5 |
| 在现有网络环境上实现的难易程度 | 4 | 3 | 4 |

\*注：满分为5，分值越接近5，表示该项技术参数越优良。

表1 VEB方案与EVB方案的综合对比

“没有最好的方案，只有最适合的”。VEB的优点是虚拟机之间的报文转发性能高，而且软件EVB（VSwitch）的兼容性好，易于实现。而EVB的优点在于虚拟机流量的监管能力、网络策略部署能力以及管理可扩展性。EVB与VEB各有所长，并不存在绝对替代关系，也正因为这个原因，EVB标准又定义了“多通道技术”。建议用户在设计虚拟服务器接入层网络时，根据实际需求选择合适的技术方案。

另外，市场上现存的一些与虚拟服务器接入层网络相关的产品，采用的方案并不是802.1工作组承认的标准技术。建议用户在评估设备的过程中，从保护现有设备投资角度出发，充分评估特定厂商产品的技术特性，确认其在将来能与当前正在标准化进程中的EVB标准充分兼容。

Floating veb

网卡嵌入交换功能。VF之间的流可以在网卡上进行交换，交换基于MAC及VLAN。同时提供基于VLAN，VF和接口的流镜像。

    XXV710 (700系列)：

    在82599的基础上，增加了除MAC及VLAN以外的匹配规则，可以讲VXLAN等Tunnel报文导入指定的VF及队列上。此外，还增加了floating VEB的功能，顾名思义，**floating VEB就是VF之间的交换，与物理脱离，并可以与普通的VEB共存，从而达到了流在VEB之间隔离，并不受端口状态影响的目的。**