DPtech VSM 技术白皮书

目录

1 概述

随着网络规模的不断扩大，网络故障点越来越多，配置和维护的复杂度大幅增加。为了提高网络及安全设备的可靠性，简化管理和组网，提高网络易用性，本白皮书提出了一种将多台 L2~7 层物理设备虚拟成一台逻辑设备来管理和使用的技术，也就是虚拟交换矩阵（Virtual Switching Matrix ，即 VSM）技术。通过 VSM 技术，可大大简化组网的复杂性和提高网络的可靠性，同时网络也更容易配置和维护。本白皮书将介绍 VSM 实现原理及如何通过该技术实现交换机、路由器、防火墙、IPS、应用交付等 L2~7 层设备的虚拟化。

VSM 虚拟交换矩阵相对于传统的堆叠方式具有如下几个优势：

- 简化配置，提高带宽利用率。VSM 完全作为一台设备使用，无需使用 STP 等协议对链路进行阻塞，通过跨设备的链路聚合不仅能够提供链路冗余的功能，还能够支持链路负载分担，充分利用带宽。此外，VSM 组网下无需使用 VRRP 等冗余网关协议，大大简化网络配置。
- 降低运维成本。传统的堆叠需要对每台设备进行配置及版本升级，耗费大量维护成本。但是在 VSM 组网下，用户通过任意成员设备的任意端口均可以登录主（Master）设备页面，对 VSM 内所有成员设备进行统一管理，降低了维护难度和运维成本。
- 高可靠性。VSM 的高可靠性体现在多个方面。1、VSM 支持跨设备链路聚合及级联端口聚合，保障了 VSM 链路的可靠性和 VSM 设备间数据传输的可靠性；2、VSM 可以支持同一个 VSM 组内的多个主控冗余备份，保障了设备的可靠性； 3、表项和会话的同步备份，保证了会话和转发表项的严格一致，实现工作在主备模式下的主控和业务板卡的无缝切换。
- 广泛的产品线支持。VSM 可支持多种不同形态的产品，实现同类产品的多合一虚拟化，

从接入设备到核心设备均支持 VSM，支持的产品形态覆盖防火墙、IPS、上网行为管理及流控、应用交付、路由器、交换机等。

2 VSM 技术介绍

2.1 概念介绍

2.1.1 VSM 成员类别

VSM 中每台设备都称为成员设备。成员设备根据功能不同，分为两种不同的角色：

- Master：主成员设备，负责管理控制整个 VSM 系统。VSM 的所有配置信息都是由 Master 设备统一下发给所有的 Slave 设备 ;VSM 中运行的数据链路层及上层协议状态机的状态信息都统一由 Master 设备来维护管理，并将这些信息同步给 Slave 设备。
- Slave：备成员设备，由 Master 控制管理。作为 Master 设备的备份设备运行，同时 Slave 设备也可以进行数据业务转发。

当 Master 故障时，系统会自动从 Slave 中选举一个新的 Master 接替原 Master 工作。Master 和 Slave 均由角色选举产生。一个 VSM 中同时只能存在一台 Master，其它成员设备都是 Slave。

2.1.2 VSM 标识

即 VSM ID。VSM 中，每台设备都通过 VSM ID 来进行唯一标识，通过 VSM ID 来进行 VSM 成员角色的选举。

2.1.3 VSM 级联

VSM 系统中各成员设备通过 VSM 级联来形成 VSM 系统。用来做 VSM 级联的单板，称为级联板，包括支持级联的普通接口板和专用级联板两大类。对于框式设备，可通过万兆或 40G 接口板卡实现级联， 高端设备除支持接口板卡级联外，还支持专用级联板，实现多台设备的高性能无阻塞级联。对于盒式设备，支持万兆接口实现级联。当使用接口板做级联的时候允许单板上部分接口用于正常数据转发，部分接口用于级联。

级联板上用于 VSM 级联的端口称作为 VSM 级联口。设备之间通过 VSM 级联口形成 VSM 通道；使用多个物理口做级联时，级联口自动进行端口聚合。级联口分为上行级联口和下行级联口两类 ,设备之间互连时需将一台设备的上行级联口和相邻设备的下行级联口相连。

针对数据中心核心交换以及骨干网核心路由等对性能和可靠性要求极高的应用场景 ,常用的通过级联口捆绑构成 VSM 组的方式可能存在级联带宽瓶颈问题，不能满足无阻塞交换要求。针对此类应用场景，VSM 支持无阻塞级联矩阵技术以提高级联带宽。VSM 无阻塞级联矩阵技术使用独立硬件设备实现 VSM 组成员之间的高速全连接互联，此设备称作级联矩阵，级联矩阵本身不出用户接口，全部接口都用来做级联。为保证级联的可靠性，支持 2 台级联矩阵和多台 VSM 成员设备的 N+2 全连接组网。

2.1.4 VSM 通道

级联口之间的连接构成了 VSM 通道 ,VSM 通道用于实现数据报文和 VSM 控制报文的传输，数据报文通过 VSM 通道在成员设备之间转发，可以看作是数据报文在同一设备上的不同接口板之间的转发一样。

2.2 VSM 的形成

2.2.1 VSM 配置

设备在加入 VSM 之前需要用户在 VSM 配置中开启 VSM 功能，并为每台设备分配一个唯一的 VSM ID，同时需要给每台 VSM 配置相应的上行和下行级联口。

2.2.2 物理连接

- VSM 连接介质

VSM 支持两种不同的连接介质：10G、40G 光口和 CX4、10G 电口。

- VSM 连接拓扑

VSM 连接的原则是本设备的上行级联口要连接对端设备的下行级联口。多台设备级联连接拓扑分成三种：

第一种是环形连接，环形连接的优点是即使有级联链路故障也会快速恢复，能够做到零丢包，并且不用配置任何协议防止环路。当环形拓扑有一条链路故障时，拓扑结构会切换到链形拓扑，此时对整个 VSM 系统的运行不会产生任何影响。

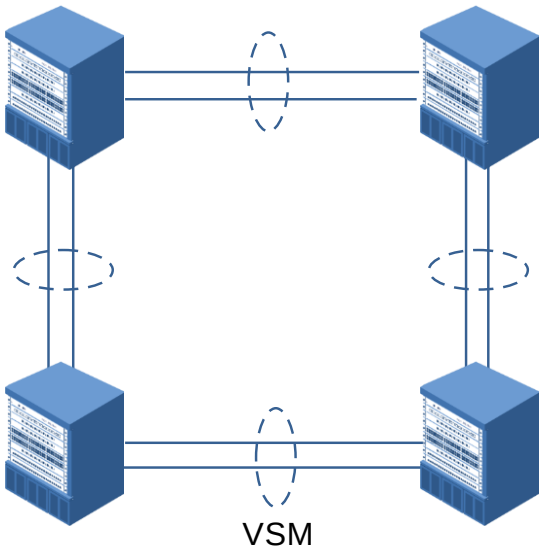

图 1 VSM 环形组网图

第二种是链式连接，链式连接是环形的一种简化，当有链路故障时 VSM 会分裂为两组 VSM。

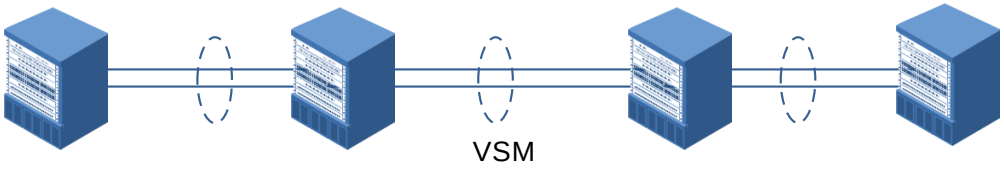

图 2 VSM 链式组网图

第三种是星形连接，如下图，图中的 VSM-M（VSM-MANAGER）是两个级联矩阵，VSM-O（VSM-OPERATOR）通过 VSM-M 进行连接，形成另一个 VSM 系统。其中交换框 VSM-M 专门负责级联并且负责管理， VSM-O 的级联口负责级联到级联矩阵 VSM-M，普通接口板接负责业务处理和流量转发。

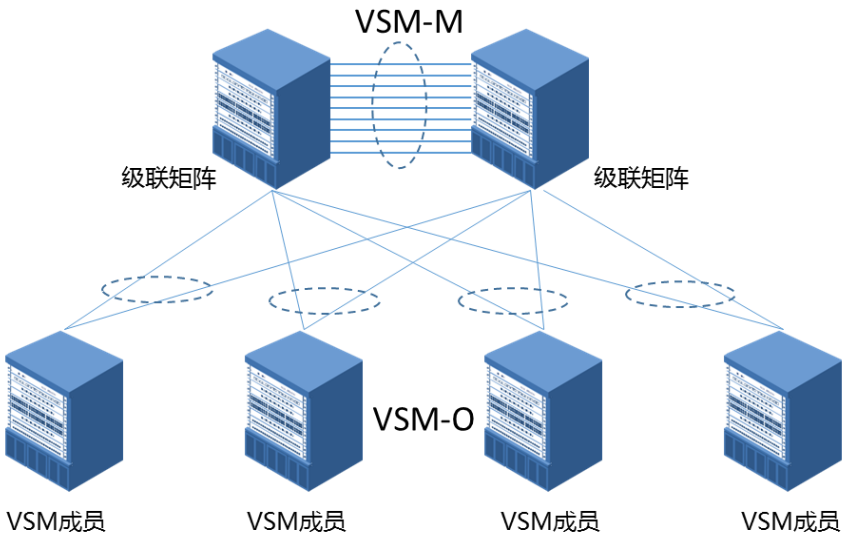

图 3 VSM 星形组网图

2.2.3 拓扑收集

每台成员设备都需要知道 VSM 组网的拓扑结构，该结构包含每个设备的 VSM ID，上

行结点设备的 ID，下行结点设备的 ID 以及其他基本信息。在设备加入 VSM 时会进行拓扑结构的收集，并将自己的信息发布到 VSM 组中。其他设备收集到非本设备发送的拓扑信息后，进行本地更新，然后将收集到的报文再转发出去，报文最终会转发到发送拓扑收集的源成员设备。当设备全部接收到自己从两种级联端口发送的拓扑收集报文时，说明本设备的拓扑结构收集完成。

拓扑收集完成后会进行设备 MAC 地址的同步，成员设备会检查 MAC 地址。当 Master 设备和 Slave 设备的 MAC 不一致时，Slave 设备将采用 Master 设备的 MAC。

2.2.4 成员主备选举

设备默认是以 Slave 的角色启动加入 VSM 的，拓扑收集完成后，成员设备之间需要选举出一个 Master 来对整个系统进行管理。

主备选举规则如下：

- 当只有一台成员设备时把该设备选举为 Master 设备；
- 当存在多台 Master 设备时需要进行选举，将 Master 设备中 VSM ID 小的设备选举为新的 Master 设备，其余的 Master 设备将重启并以 Slave 角色加入该 VSM；
- 当 VSM 中设备都为 Slave 设备时，VSM ID 最小的那台设备选举为 Master 设备，其他的都为 Slave 设备；

2.3 VSM 管理和维护

2.3.1 统一管理

VSM 将两台或多台设备虚拟成一台设备来统一管理，两台设备所有配置都保持一致。对于框式设备来说，备框的管理口 IP 地址和主框的保持一致。在 VSM 正常运行状态，只

有主框的主控板的管理口是可用的，主框的备用主控板和其他备框的管理口都不可用，当发生主备板或主备框切换时，新的主框的主控板的管理口变为可用，VSM 系统中其他设备的所有管理口变为不可用，这样用户使用管理口登录系统时，始终登录的都是主框，对系统中的所有设备进行统一管理。

对于盒式交换机或框式设备使用管理 VLAN 登录系统来说，主框和备框的管理 VLAN 的 IP 都是一致的，用户使用管理 VLAN 登录系统时始终登录的是主框，保证对系统中的所有设备的统一管理。

VSM 登录主设备的管理页面对设备进行统一管理。在管理页面中可以看到所有成员设备的配置情况，在管理页面上进行配置，配置会下发到所有成员设备。

当设备人为重启或者故障重启，配置不会丢失，Slave 设备在启动过程中会向 Master 设备请求批量同步配置信息，然后 Slave 以新的配置完成初始化，保证 Slave 起来以后无缝加入 VSM。VSM 在运行中，在管理页面修改的所有配置都将同步到所有设备，保证当主设备故障后，配置数据不丢失。

2.3.2 新成员设备加入

当有新设备加入 VSM 中时，会产生级联端口的 up 事件，当级联端口 up 时，会发送拓扑改变通知，VSM 系统会重新进行拓扑收集，当收集完成时进行选举。如果新加入的成员设备的 VSM 角色是 Slave，则以 Slave 的身份加入 VSM 系统中；如果新加入的成员设备的 VSM 角色是 Master，那么重新选举出新的 Master 设备，其余的 Master 成员设备重启并以 Slave 的角色重新加入 VSM。

2.3.3 已有成员设备离开

VSM 维护过程中，当有设备离开 VSM 中时，会产生级联端口的 down 事件。当级联端口 down 时，该成员设备会发出拓扑改变通知，然后所有成员设备都会重新进行拓扑收集，如果离开的成员设备是 Slave，那么它离开后，其他成员角色不会发生变化继续正常运行；如果离开的成员设备是 Master，那么它离开后，将会在其他成员设备间选举出新的 Master 设备用于管理整个 VSM 系统。

2.3.4 VSM 分裂检测

前文描述的成员设备离开，如果离开的原因是因为设备故障重启，那么它再以 Slave 的身份加入 VSM 即可，没有任何影响；但是如果它离开的原因是因为级联链路故障，那么它所在的 VSM 系统中也会选举出新的 Master，那么在这个组网中将出现两组 VSM 系统，并且它们的 MAC 地址是一致的，这就导致 VSM 分裂。为了解决这个问题，VSM 系统支持 VSM 分裂检测，如果发现有相同的 VSM 系统，那么分裂检测机制会使产生冲突的系统进入静默状态，此时冲突的系统不会进行报文转发和学习，也不会处理任何业务。这样，在应用组网中就不会存在冲突的 VSM 系统。

2.3.5 VSM 在线升级

VSM 支持在线升级功能，分为两步走：第一，先将备机进行“孤岛隔离”，对 Slave 设备进行版本升级和配置工作，在 “孤岛隔离”的情况下，Slave 不会进行业务处理和流量转发，只有 Master 进行业务处理；第二，Slave 升级完成按新版本重新启动后，再将 Master 进行 “孤岛隔离”，此时 Slave 切换为 Master，接管所有流量业务处理，然后被“隔离”的设备升级版本再启动加入 VSM 系统。在升级的过程中能够保证业务正常。

具体实施方法是：准备升级的 Slave 设备先开启“孤岛隔离”状态，这时各业务口均被禁用，停止 Slave 设备上所有业务和数据转发，只有管理口可用于配置管理。接着配置备机的 VSM ID、级联口，更新配置信息，升级软件版本（由于级联口也被禁用，因此不会对系统运行产生影响）。然后点击“一键升级”，使 Slave 设备重启。Slave 设备完全重启后还处在孤岛状态，首先通过一个管理口的带外通道通知 Master 设备升级已完成，然后自己切换为 Master 设备，并且去除孤岛状态，此时 Slave 设备以 Master 的身份正常运行在网络中。

Master 设备在接收到 Slave 设备发送的更新完版本通知后，立即将自己切换到“孤岛隔离”状态，各业务口均被禁用，停止 Master 设备上所有业务和数据转发，只有管理口可用于配置管理。然后通过管理口的带外通道向新的 Master 设备索要新的软件版本，新的 Master 设备向旧的 Master 设备同步完软件版本后，以新的软件版本重新启动。重启过程中，新的 Master 设备向自己同步配置信息，并且检查主控板与业务板软件版本与新的 Master 设备是否一致，若不一致则从新的 Master 同步版本。

重启后 Slave 设备将变为 Master 设备，而原 Master 设备变为 Slave 设备。这时即完成了 VSM 在线升级的整个过程。

如果 VSM 系统中存在超过两台设备，则应先将一台 Slave 设备进行“孤岛隔离”、在线升级，然后先后通知 Master 设备和其他 Slave 设备进行版本升级。

2.4 控制平面工作原理

控制平面主要是通过控制成员设备来统一处理协议数据，达到统一管理的目的。协议报文包括路由协议报文、二层协议报文、DHCP、DNS 等等，协议报文统一由 VSM 的 Master 处理。

当 VSM 组的 Master 成员收到协议报文时，由 Master 成员处理；当 Slave 成员收到协议报文时，它不对协议报文进行处理，报文会通过 VSM 级联通道转发到 Master 成员处理。Master 处理完成后，并通过 VSM 级联通道将相关信息备份到 Slave 设备。

总而言之，控制平面统一由 Master 处理，解决由成员设备分开处理发生的路径不一致和状态不一致的问题。

2.5 数据平面工作原理

2.5.1 跨框链路聚合

跨框链路聚合即在不同的 VSM 成员设备上的端口配置为聚合组，配置时会将聚合配置信息向 Master 和 Slave 分别下发全局聚合信息。跨框聚合同样具有链路备份，负载分担的功能，VSM 也支持配置静态跨框聚合和动态跨框聚合。

2.5.2 L2~3 层设备转发原理

二层转发 :由于 VSM 作为一台设备来运行 ,其广播域包括所有 VSM 成员设备 ,当 VSM 成员设备收到广播报文、未知单播或未知组播时，不仅会在本设备内广播，而且报文将通过级联口广播到其他 VSM 成员设备，所有成员设备将对二层报文进行学习。VSM 成员设备收到已知报文时，查找二层表项，将报文送到正确出口，如果出口不是本设备，报文通过 VSM 级联通道送到其他设备。

三层转发 : VSM 成员设备收到的三层报文查找表项有路由但是没有 ARP 时，需要将报文上送 CPU 进行 ARP 学习，如果本成员设备是 Master，跟单台设备处理一致，如果本成员设备是 Slave，通过 VSM 级联通道转发给 Master 成员，Master 处理完成后，将表项下发给 Slave。VSM 成员设备收到的三层报文时查找三层表项，能够查到出口的转发方式和

二层转发一致。

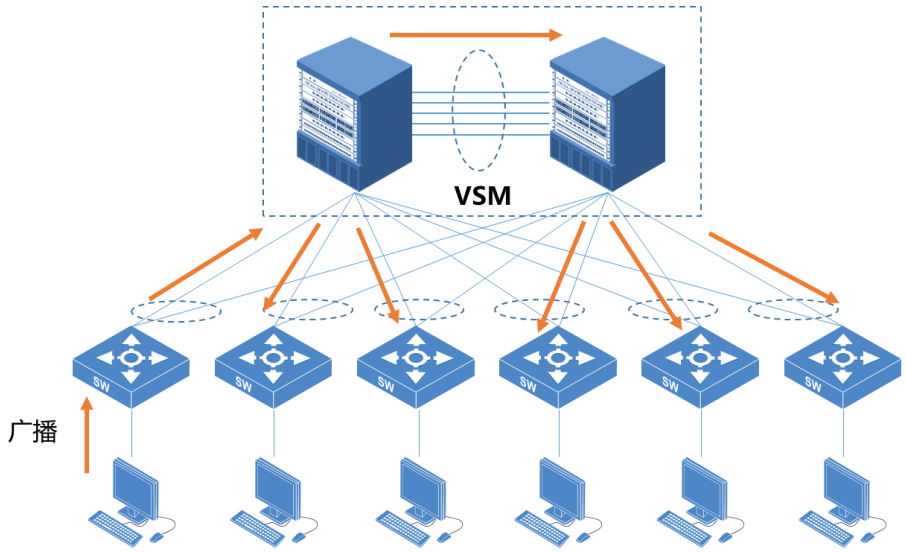

图 4 VSM 广播转发组网图

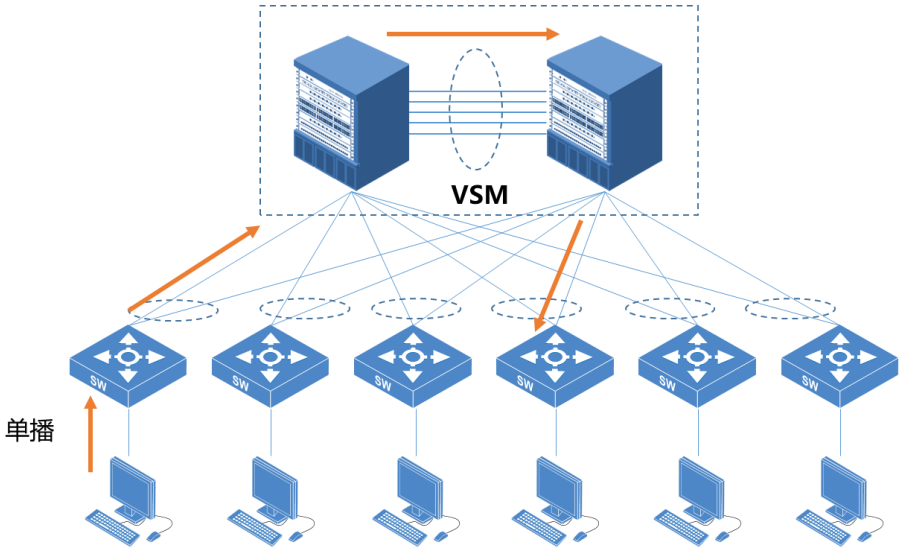

图 5　VSM 单播转发组网图

2.5.3　L4~7 层设备转发原理

盒式 L4~7 设备 VSM 虚拟化时，所有成员设备的 CPU 等物理资源工作在全局模式下。首先将收到的报文做全局流分类，随后通过交换芯片和级联口将报文调度到某个成员设备进

行 L4~7 层业务处理。提供多种云调度算法解决在不同组网下的双向流量同源同宿问题，例如在不做 NAT 情况，可依据源 IP+目的 IP 地址对匹配双向流量；在做 NAT 的情况下，通过域内域外不同算法的方式来实现同源同宿。会话在所有成员之间进行备份，当有成员设备故障时，业务流量能够快速切换到其他成员设备处理。

框式设备 VSM 虚拟化时，L4~7 层业务流量通过“流定义”技术实现在 VSM 组内的调度。首先按端口、IP 地址、VLAN 等多种参数定义数据流，然后将流量逐跳匹配到预定义的物理或虚拟业务模块上，同时 WEB 配置界面预设“在线部署”、“透明部署”等多种“流定义”模式，实现数据流定义和调度的图形化操作。VSM 组内的业务板可以配置成两种模式。一种是主备模式，不同成员设备的同类型业务板可以配置为主备业务板，报文默认流定义到主业务板，只有当主业务板故障后，流量才会切换到备业务板。另一种是云板卡模式，不同成员设备的同类型业务板可以配置为虚拟云，报文根据云调度算法上送到某个业务板完成业务处理。和盒式的 L4~7 层转发一样，VSM 会保证报文的同源同宿，即保证同一会话的报文去同一块业务板处理。

2.5.4 优先本框转发原理

报文进入 VSM 成员设备后查找转发表项，出口如果是聚合口，报文将按照负载分担的算法来找物理端口，那么报文就有可能总是要通过级联通道转发出去。为了减少设备间级联通道的带宽使用，VSM 可以配置为优先本框转发模式，即报文从哪个设备进入优先从哪个设备的物理口转发出去，仅当聚合中本设备的物理端口全部 down 或者故障时，才会从 VSM 组内其他成员设备转发出去。

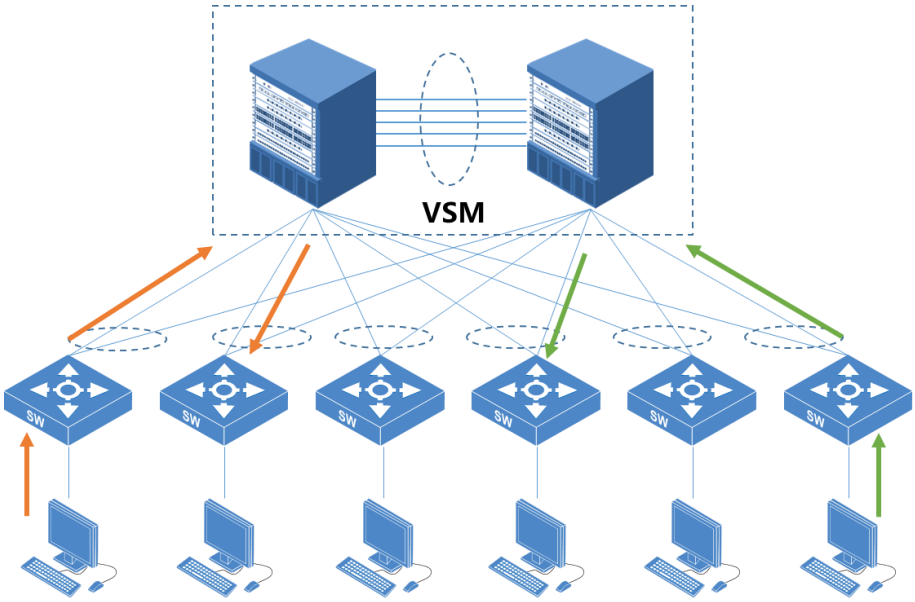

图 6 VSM 优先本框转发组网图

3 VSM 状态备份

3.1 会话备份

VSM 环境中，各个成员设备的会话要保持一致，如果有成员设备故障，原本由它处理的业务可以安全切换到其他成员设备上去。在成员设备会话新建和变动时，将会话实时备份给其他设备，确保当流量切换到其他设备时该设备上已拥有该会话，实现业务流量的无中断切换。另外，其它与现场流量相关的信息，例如健康检查的状态信息、动态路由表等也需要同步到其它设备。

3.2 策略备份

在流量转发路径上对流量的去向起控制作用的策略表项，如流定义、包过滤和 ACL 等，需要在各成员设备上保持有相同的配置。各成员设备上除了有属于本机的策略配置外，还保存着其它各成员设备上的配置，这样，当进行主备切换后，新 Master 上的策略配置与原

Master 的配置一致，VSM 组的流量策略不受主备切换的影响。

3.3 各种协议状态备份

VSM 组中，由 Master 管理各成员设备上的所有槽位、接口和资源，各数据链路层及上层协议状态机也由 Master 维护。当 Master 和 Slave 间进行切换时，为了使新的 Master 能无缝接收各协议运行状态，保持协议状态机运行的连续性，Master 会将协议状态机实时的同步到各 Slave 上，确保一旦发生切换时，新的 Master 立即接管槽位、接口和在运行的协议计算。

4 VSM 组网应用

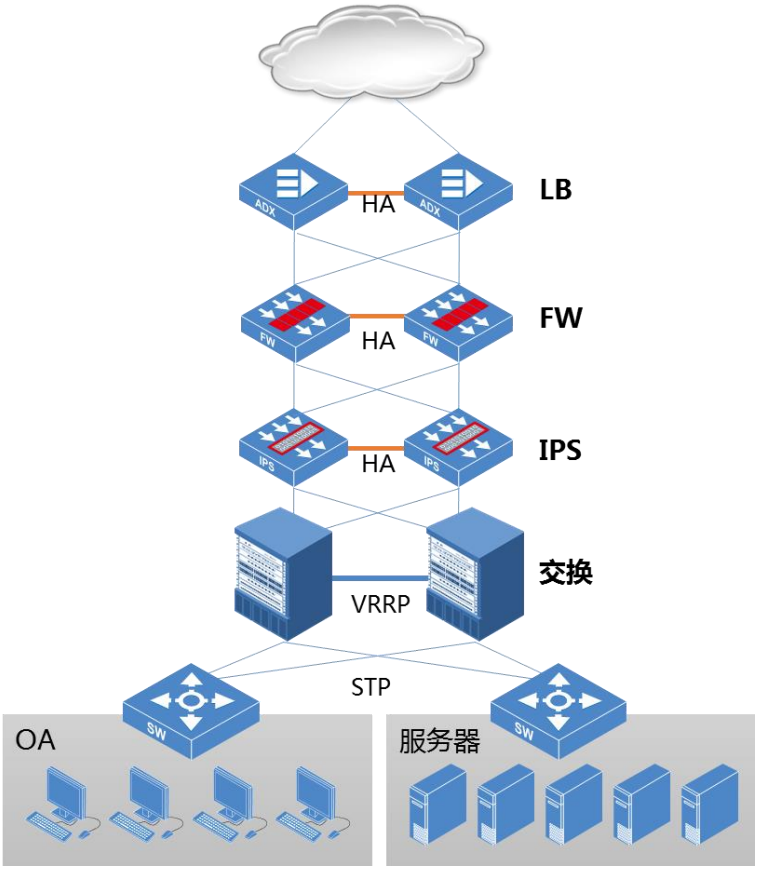

图 7 传统串行组网示意图

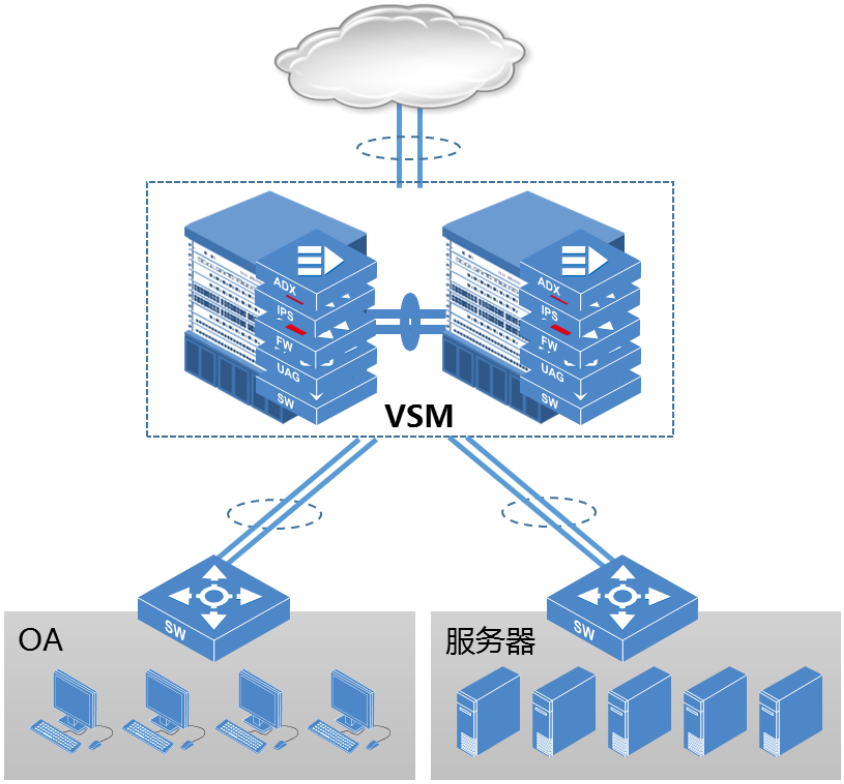

图 8 VSM 组网示意图

如图 7 和图 8 所示，DPX 板卡混插 VSM 组网和传统网络、安全和应用交付组网方式相比有以下几个优势：

- 简化组网

  传统组网实现双机热备功能需要使用 VRRP 网关冗余协议或者通过路由选路方式实现，VSM 组网能够对数据包进行负载分担处理，因此无需 VRRP 等技术就能够实现双机协同工作并且简化组网。

- 弹性扩展

  VSM 组网能够通过增加同类业务板卡弹性扩展性能，增加不同业务板卡扩展系统业务功能。

- 提高防护能力

  例如网络中独立的两台 IPS 由于流量的选路问题，可能会导致入侵检测设备不能对一整条完整的流量进行检测，使得带有攻击的网络流量进入内部网络。VSM 组网环境下可

以对完整的数据流量进行检测防护，阻断攻击流量。

- 简化设备管理

传统部署方案中，各个设备独立管理，维护成本高。VSM 组网方式下，所有机框和业务板卡提供统一管理 IP 和管理界面，所有配置和管理工作一站完成。