# 深度剖析
## ONTAP Select

NetApp
April 29, 2024

# 目录

# 深度剖析

## 存储

### 存储：一般概念和特征

了解适用于 ONTAP Select 环境的常规存储概念，然后再了解特定存储组件。

#### 存储配置的各个阶段

ONTAP Select 主机存储的主要配置阶段包括：

- 部署前的前提条件
  - 确保每个虚拟机管理程序主机均已配置完毕并做好 ONTAP Select 部署准备。
  - 此配置涉及物理驱动器， RAID 控制器和组， LUN 以及相关网络准备。
  - 此配置在 ONTAP Select 之外执行。
- 使用虚拟机管理程序管理员实用程序进行配置
  - 您可以使用虚拟机管理程序管理实用程序配置存储的某些方面（例如 VMware 环境中的 vSphere ）。
  - 此配置在 ONTAP Select 之外执行。
- 使用 ONTAP Select Deploy 管理实用程序进行配置
  - 您可以使用 Deploy 管理实用程序配置核心逻辑存储构造。
  - 这可以通过命令行界面命令显式执行，也可以在部署过程中由实用程序自动执行。
- 部署后配置
  - ONTAP Select 部署完成后，您可以使用 ONTAP 命令行界面或系统管理器配置集群。
  - 此配置在 ONTAP Select Deploy 之外执行。

#### 受管存储与非受管存储

由 ONTAP Select 访问和直接控制的存储是托管存储。同一虚拟机管理程序主机上的任何其他存储均为非受管存储。

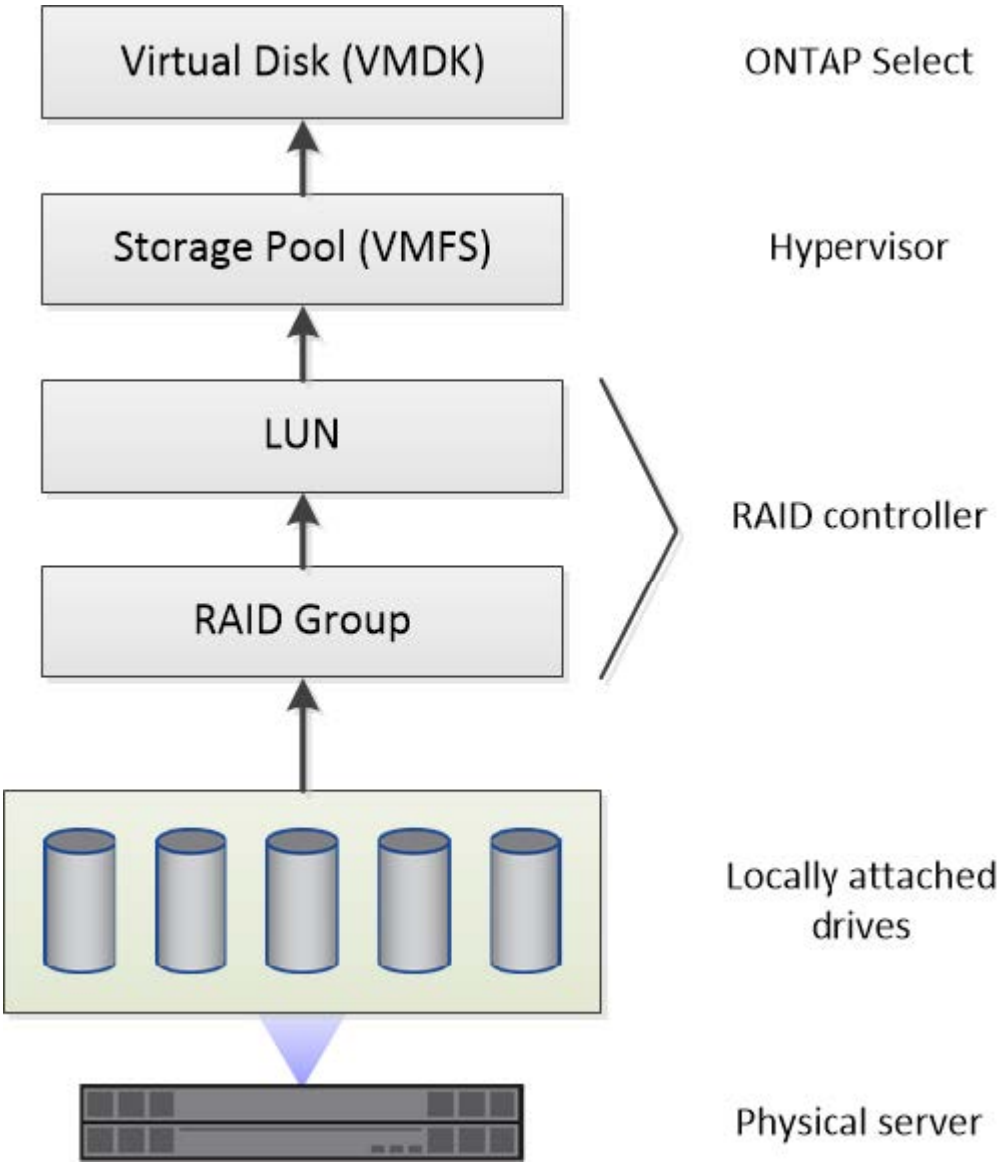
##### 同构物理存储

构成 ONTAP Select 受管存储的所有物理驱动器都必须是同构的。也就是说，在以下特征方面，所有硬件都必须相同：

- 类型（ SAS ， NL-SAS ， SATA ， SSD ）
- 速度（ RPM ）

#### 本地存储环境图示

每个虚拟机管理程序主机都包含可供 ONTAP Select 使用的本地磁盘和其他逻辑存储组件。这些存储组件采用物理磁盘的分层结构进行排列。




本地存储组件的特征

以下几个概念适用于 ONTAP Select 环境中使用的本地存储组件。在准备 ONTAP Select 部署之前，您应熟悉这些概念。这些概念按类别进行排列： RAID 组和 LUN ，存储池和虚拟磁盘。

将物理驱动器分组为 **RAID** 组和 **LUN**

一个或多个物理磁盘可以本地连接到主机服务器，并可供 ONTAP Select 使用。物理磁盘会分配给 RAID 组，然后这些 RAID 组会作为一个或多个 LUN 提供给虚拟机管理程序主机操作系统。每个 LUN 都会作为物理硬盘驱动器提供给虚拟机管理程序主机操作系统。

配置 ONTAP Select 主机时，应注意以下事项：

- 所有受管存储都必须通过一个 RAID 控制器进行访问
- 根据供应商的不同，每个 RAID 控制器支持每个 RAID 组的最大驱动器数

#### 一个或多个 RAID 组

每个 ONTAP Select 主机都必须有一个 RAID 控制器。您应为 ONTAP Select 创建一个 RAID 组。但是，在某些情况下，您可能会考虑创建多个 RAID 组。请参见 "最佳实践摘要"。

### 存储池注意事项

在准备部署 ONTAP Select 时，您应注意一些与存储池相关的问题。

> 在 VMware 环境中，存储池与 VMware 数据存储库是同义词。

#### 存储池和 LUN

每个 LUN 在虚拟机管理程序主机上都被视为本地磁盘，并且可以是一个存储池的一部分。每个存储池都使用虚拟机管理程序主机操作系统可以使用的文件系统进行格式化。

您必须确保在 ONTAP Select 部署过程中正确创建了存储池。您可以使用虚拟机管理程序管理工具创建存储池。例如，对于 VMware ，您可以使用 vSphere 客户端创建存储池。然后，存储池将传递到 ONTAP Select Deploy 管理实用程序。

### 管理虚拟磁盘

在准备部署 ONTAP Select 时，您应注意一些与虚拟磁盘相关的问题。

#### 虚拟磁盘和文件系统

ONTAP Select 虚拟机分配有多个虚拟磁盘驱动器。每个虚拟磁盘实际上都是存储池中的一个文件，由虚拟机管理程序维护。ONTAP Select 使用多种类型的磁盘，主要是系统磁盘和数据磁盘。

此外，您还应了解以下有关虚拟磁盘的信息：

- 要创建虚拟磁盘，存储池必须可用。
- 在创建虚拟机之前，无法创建虚拟磁盘。
- 您必须使用 ONTAP Select Deploy 管理实用程序创建所有虚拟磁盘（也就是说，管理员绝不能在 Deploy 实用程序之外创建虚拟磁盘）。

#### 配置虚拟磁盘

虚拟磁盘由 ONTAP Select 管理。使用 Deploy 管理实用程序创建集群时，系统会自动创建这些卷。

#### 外部存储环境图示

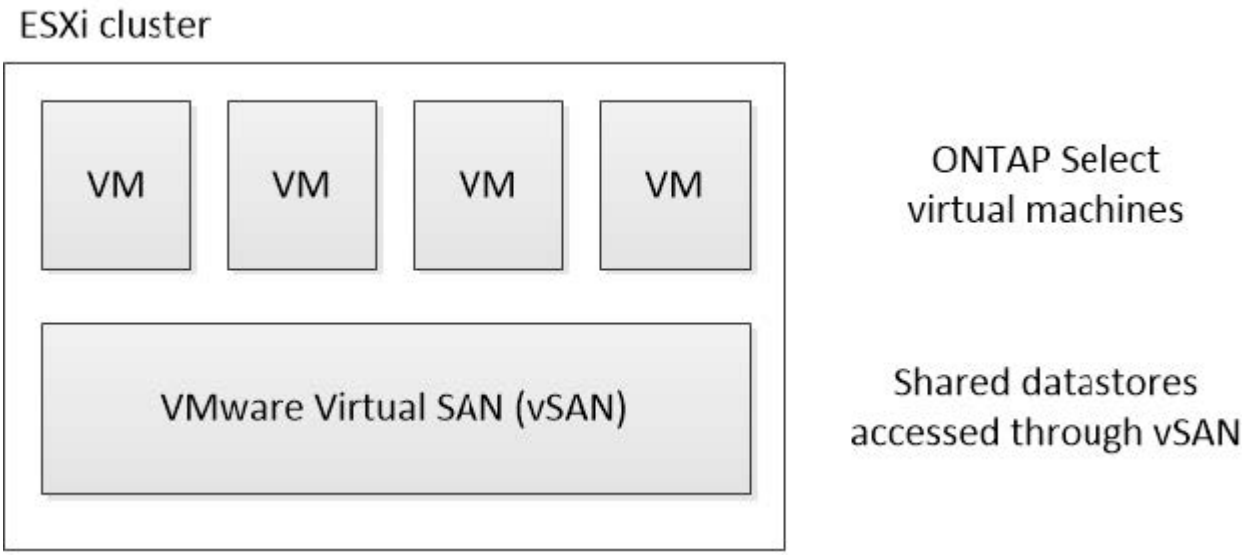
通过 ONTAP Select vNAS 解决方案 ， ONTAP Select 可以使用虚拟机管理程序主机外部存储上的数据存储库。可以使用 VMware vSAN 通过网络访问这些数据存储库，也可以直接在外部存储阵列上访问这些数据存储库。

可以将 ONTAP Select 配置为使用虚拟机管理程序主机外部的以下类型的 VMware ESXi 网络数据存储库：

- VSAN （虚拟 SAN ）
- VMFS
- NFS

**vSAN** 数据存储库

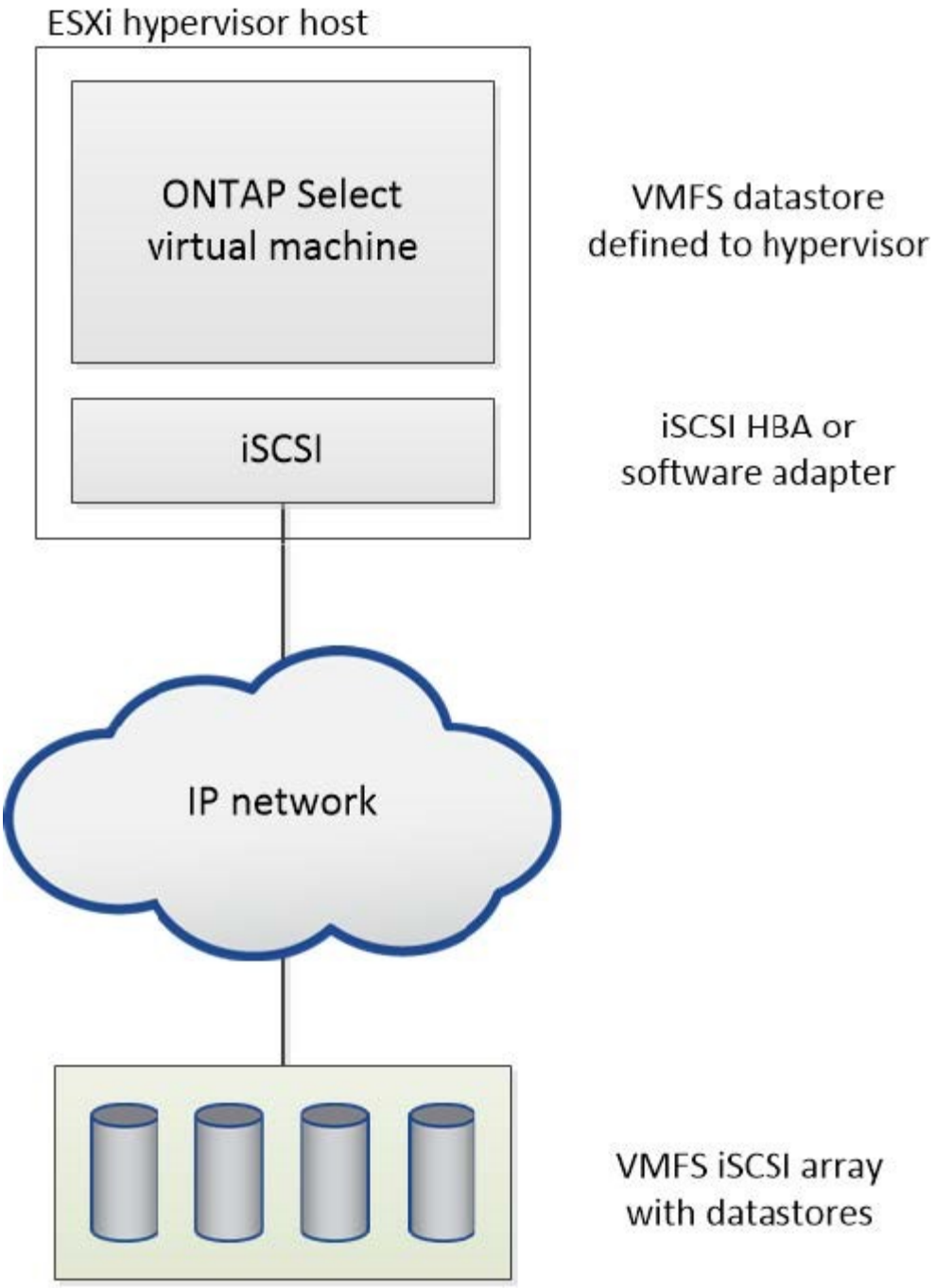
每个 ESXi 主机都可以有一个或多个本地 VMFS 数据存储库。通常，这些数据存储库只能由本地主机访问。但是， VMware vSAN 允许 ESXi 集群中的每个主机共享集群中的所有数据存储库，就像它们位于本地一样。下图说明了 vSAN 如何创建在 ESXi 集群中的主机之间共享的数据存储库池。




外部存储阵列上的 **VMFS** 数据存储库

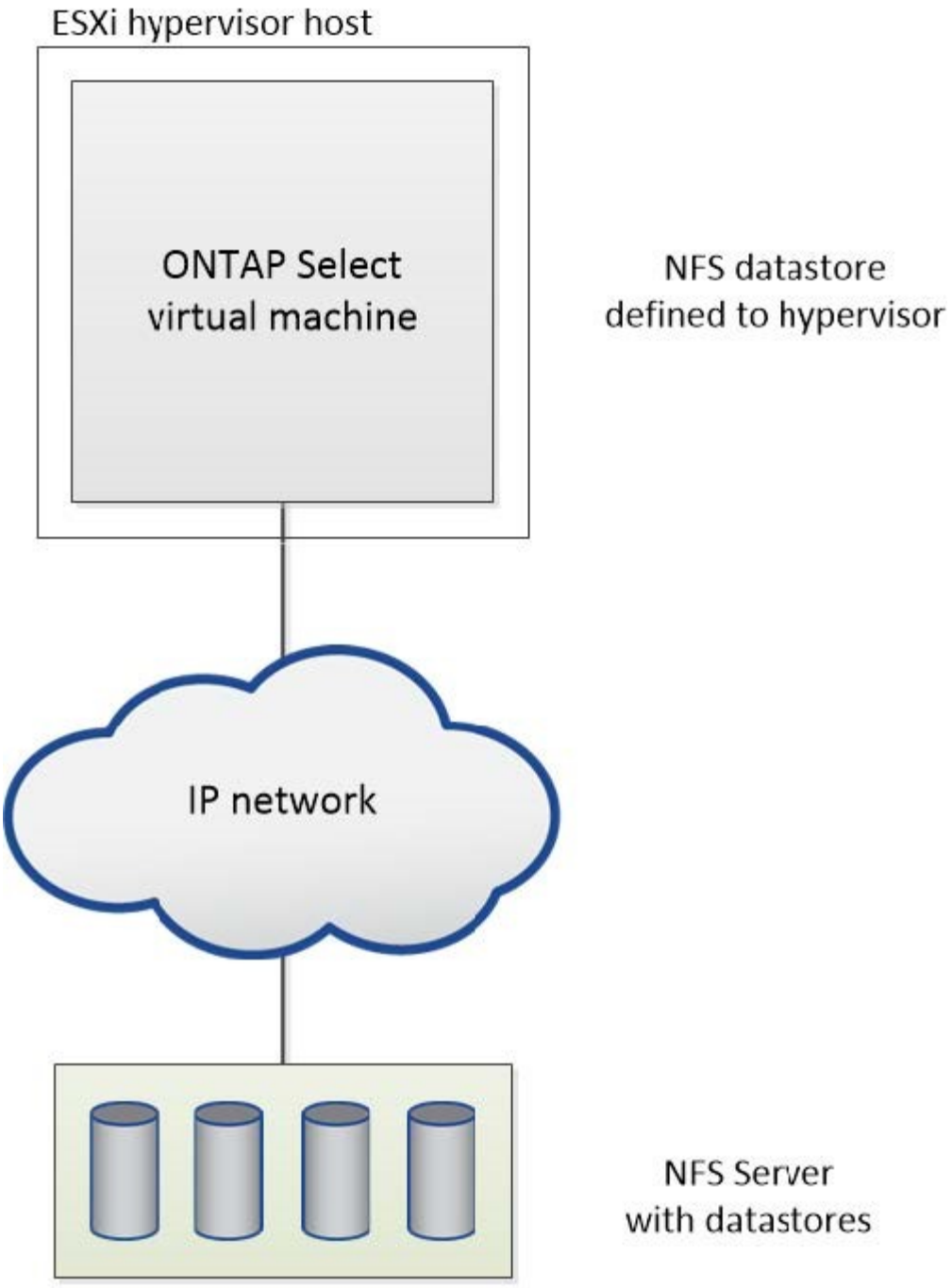
您可以创建驻留在外部存储阵列上的 VMFS 数据存储库。可以使用多种不同的网络协议之一访问存储。下图显示了使用 iSCSI 协议访问的外部存储阵列上的 VMFS 数据存储库。

> ONTAP Select支持VMware存储/SAN兼容性文档中所述的所有外部存储阵列、包括iSCSI、光纤通道和以太网光纤通道。




外部存储阵列上的**NFS**数据存储库

您可以创建驻留在外部存储阵列上的 NFS 数据存储库。存储可使用 NFS 网络协议进行访问。下图显示了通过 NFS 服务器设备访问的外部存储上的 NFS 数据存储库。




用于本地连接存储的硬件 **RAID** 服务

如果有可用的硬件 RAID 控制器， ONTAP Select 可以将 RAID 服务移至硬件控制器，以提高写入性能并防止物理驱动器出现故障。因此， ONTAP Select 集群中所有节点的 RAID 保护由本地连接的 RAID 控制器提供，而不是通过 ONTAP 软件 RAID 提供。

> ONTAP Select 数据聚合配置为使用 RAID 0 ，因为物理 RAID 控制器正在为底层驱动器提供 RAID 条带化。不支持其他 RAID 级别。

本地连接存储的 **RAID** 控制器配置

为 ONTAP Select 提供后备存储的所有本地连接磁盘都必须位于 RAID 控制器后面。大多数商用服务器都随附多个 RAID 控制器选项，价格各不相同，每个控制器选项的功能级别各不相同。其目的是支持尽可能多的这些选项，前提是它们满足控制器上的某些最低要求。

管理 ONTAP Select 磁盘的 RAID 控制器必须满足以下要求：

- 硬件 RAID 控制器必须具有电池备份单元（ BBU ）或闪存备份写入缓存（ FBWC ），并支持 12 Gbps 的吞吐量。
- RAID 控制器必须支持至少可承受一个或两个磁盘故障的模式（ RAID 5 和 RAID 6 ）。
- 驱动器缓存必须设置为已禁用。
- 必须将写入策略配置为回写模式，并在发生 BBU 或闪存故障时执行回退。
- 读取的 I/O 策略必须设置为缓存。

所有为 ONTAP Select 提供后备存储的本地连接磁盘都必须置于运行 RAID 5 或 RAID 6 的 RAID 组中。对于 SAS 驱动器和 SSD ，使用最多包含 24 个驱动器的 RAID 组可以使 ONTAP 获得将传入读取请求分散到更多磁盘的优势。这样可以显著提高性能。在 SAS/SSD 配置中，对单 LUN 配置和多 LUN 配置执行了性能测试。没有发现显著的差异，因此，为了简单起见， NetApp 建议创建最少数量的 LUN 来满足您的配置需求。

NL-SAS 和 SATA 驱动器需要一组不同的最佳实践。出于性能原因，最小磁盘数仍为 8 个，但 RAID 组大小不应超过 12 个驱动器。NetApp 还建议每个 RAID 组使用一个备用磁盘；但是，可以使用所有 RAID 组的全局备用磁盘。例如，您可以为每三个 RAID 组使用两个备用磁盘，每个 RAID 组包含 8 到 12 个驱动器。

> 旧版 ESX 的最大块区和数据存储库大小为 64 TB ，这可能会影响支持这些大容量驱动器提供的总原始容量所需的 LUN 数量。

**RAID**模式

许多 RAID 控制器最多支持三种操作模式，每种模式都表示写入请求所采用的数据路径存在显著差异。这三种模式如下：

- 直写。所有传入的 I/O 请求都会写入 RAID 控制器缓存，然后立即转储到磁盘，然后再向主机确认该请求。
- 写入。所有传入的 I/O 请求都会直接写入磁盘，从而绕过 RAID 控制器缓存。
- 回写。所有传入的 I/O 请求都会直接写入控制器缓存，并立即确认回主机。使用控制器异步将数据块转储到磁盘。

回写模式提供最短的数据路径，在数据块进入缓存后立即进行 I/O 确认。此模式可为混合读 / 写工作负载提供最低延迟和最高吞吐量。但是，如果不存在 BBU 或非易失性闪存技术，则在系统在此模式下运行时发生电源故障时，用户将面临丢失数据的风险。

ONTAP Select 要求具有电池备份或闪存单元；因此，我们可以确信，在发生此类故障时，缓存的块会转储到磁盘。因此， RAID 控制器必须配置为回写模式。

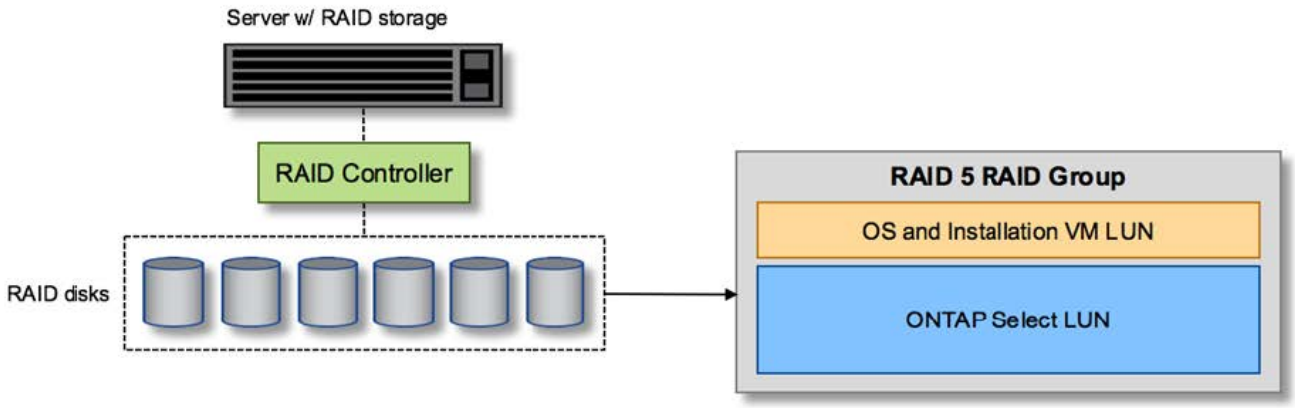
**ONTAP Select** 和操作系统之间共享的本地磁盘

最常见的服务器配置是，所有本地连接的磁盘轴都位于一个 RAID 控制器后面。您应至少配置两个 LUN ： 一个用于虚拟机管理程序，一个用于 ONTAP Select VM 。

例如，假设一个 HP DL380 g8 具有六个内部驱动器和一个智能阵列 P420i RAID 控制器。所有内部驱动器均由此 RAID 控制器管理，系统上不存在任何其他存储。

下图显示了这种配置。在此示例中，系统上不存在其他存储；因此，虚拟机管理程序必须与 ONTAP Select 节点共享存储。

- 仅使用 RAID 管理磁盘轴的服务器 LUN 配置 *




通过从与 ONTAP Select 相同的 RAID 组配置操作系统 LUN ，虚拟机管理程序操作系统（以及也从该存储配置的任何客户端虚拟机）可以从 RAID 保护中受益。此配置可防止单驱动器故障导致整个系统停机。

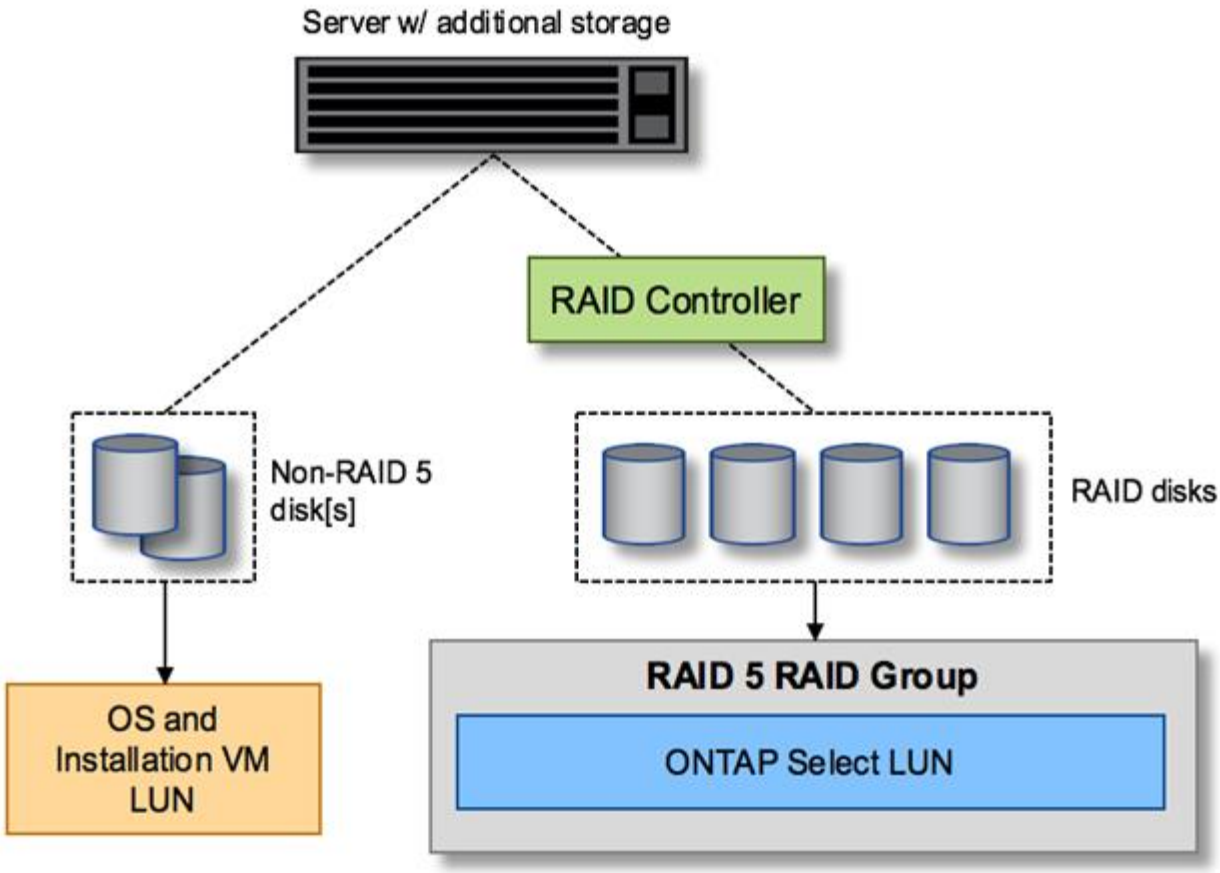
在 **ONTAP Select** 和操作系统之间拆分的本地磁盘

服务器供应商提供的另一种可能的配置包括为系统配置多个 RAID 或磁盘控制器。在此配置中，一组磁盘由一个磁盘控制器管理，该控制器可能提供 RAID 服务，也可能不提供 RAID 服务。第二组磁盘由硬件 RAID 控制器管理，该控制器能够提供 RAID 5/6 服务。

在这种配置模式下，可提供 RAID 5/6 服务的 RAID 控制器后面的一组磁盘轴应仅供 ONTAP Select VM 使用。根据所管理的总存储容量，您应将磁盘轴配置为一个或多个 RAID 组以及一个或多个 LUN 。然后，这些 LUN 将用于创建一个或多个数据存储库，其中所有数据存储库均受 RAID 控制器保护。

第一组磁盘是为虚拟机管理程序操作系统以及未使用 ONTAP 存储的任何客户端虚拟机预留的，如下图所示。

- 混合 RAID/ 非 RAID 系统上的服务器 LUN 配置 *




多个 **LUN**

在两种情况下，单 RAID 组 / 单 LUN 配置必须更改。使用 NL-SAS 或 SATA 驱动器时， RAID 组大小不得超过 12 个驱动器。此外，一个 LUN 可能会大于底层虚拟机管理程序存储限制，可以是单个文件系统块区最大大小，也可以是存储池总最大大小。然后，必须将底层物理存储拆分为多个 LUN ，才能成功创建文件系统。

**VMware vSphere** 虚拟机文件系统限制

在某些 ESX 版本上，数据存储库的最大大小为 64 TB 。

如果服务器连接的存储超过 64 TB ，则可能需要配置多个 LUN ，每个 LUN 都小于 64 TB 。创建多个 RAID 组以缩短 SATA/NL-SAS 驱动器的 RAID 重建时间也会导致配置多个 LUN 。

如果需要多个 LUN ，则需要考虑的一个主要问题是确保这些 LUN 的性能相似且一致。如果要在一个 ONTAP 聚合中使用所有 LUN ，则这一点尤其重要。或者，如果一个或多个 LUN 的一个子集具有截然不同的性能配置文件，我们强烈建议将这些 LUN 隔离在一个单独的 ONTAP 聚合中。

可以使用多个文件系统块区来创建一个数据存储库，该数据存储库的大小不超过数据存储库的最大大小。要限制需要 ONTAP Select 许可证的容量，请确保在集群安装期间指定容量上限。此功能允许 ONTAP Select 仅使用数据存储库中的一部分空间（因此需要许可证）。

或者，也可以先在一个 LUN 上创建一个数据存储库。如果需要更多空间，并需要更大的 ONTAP Select 容量许可证，则可以将该空间作为块区添加到同一个数据存储库中，但不能超过数据存储库的最大大小。达到最大大小后，可以创建新的数据存储库并将其添加到 ONTAP Select 中。这两种类型的容量扩展操作均受支持，并且可以使用 ONTAP Deploy storage-add 功能来实现。可以将每个 ONTAP Select 节点配置为最多支持 400 TB 的存

储。从多个数据存储库配置容量需要两步过程。

初始集群创建可用于创建占用初始数据存储库中部分或全部空间的 ONTAP Select 集群。第二步是使用其他数据存储库执行一个或多个容量添加操作，直到达到所需的总容量为止。有关此功能的详细信息，请参见一节 "增加存储容量"。

> VMFS 开销不为零（请参见 "VMware 知识库 1001618"），并且尝试使用数据存储库报告为可用的整个空间会导致集群创建操作期间出现虚假错误。

每个数据存储库中会保留 2% 的未使用缓冲区。此空间不需要容量许可证，因为 ONTAP Select 不会使用此空间。只要未指定容量上限， ONTAP Deploy 就会自动计算缓冲区的确切 GB 数。如果指定了容量上限，则会首先强制实施该大小。如果容量上限大小不超过缓冲区大小，则集群创建将失败，并显示一条错误消息，指出可用作容量上限的正确最大大小参数：

```
“InvalidPoolCapacitySize: Invalid capacity specified for storage pool
“ontap-select-storage-pool”, Specified value: 34334204 GB. Available
(after leaving 2% overhead space): 30948”
```

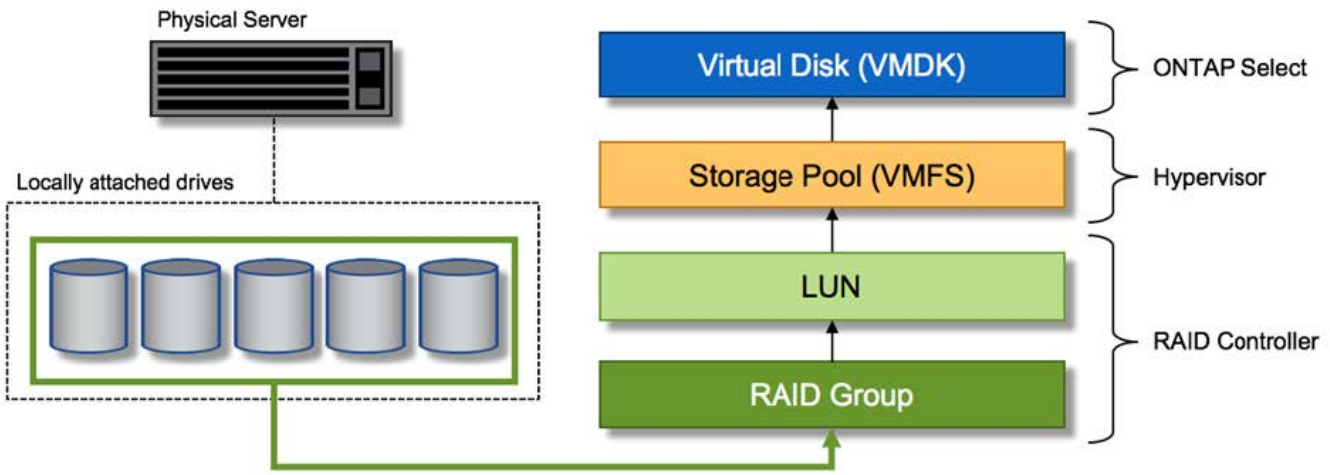
新安装和现有 ONTAP Deploy 或 ONTAP Select VM 的 Storage vMotion 操作均支持 VMFS 6 。

VMware 不支持从 VMFS 5 原位升级到 VMFS 6 。因此， Storage vMotion 是唯一允许任何 VM 从 VMFS 5 数据存储库过渡到 VMFS 6 数据存储库的机制。但是，除了从 VMFS 5 过渡到 VMFS 6 的特定目的之外，ONTAP Select 和 ONTAP Deploy 对 Storage vMotion 的支持也有所扩展，以涵盖其他情形。

**ONTAP Select 虚拟磁盘**

ONTAP Select 的核心是为 ONTAP 提供一组从一个或多个存储池配置的虚拟磁盘。ONTAP 会提供一组虚拟磁盘，这些虚拟磁盘会被视为物理磁盘，而存储堆栈的其余部分则由虚拟机管理程序进行抽象化。下图更详细地显示了这种关系，突出显示了物理 RAID 控制器，虚拟机管理程序和 ONTAP Select VM 之间的关系。

- RAID 组和 LUN 配置可通过服务器的 RAID 控制器软件进行。使用 VSAN 或外部阵列时不需要此配置。
- 存储池配置从虚拟机管理程序中进行。
- 虚拟磁盘由各个 VM 创建并拥有；在此示例中，虚拟磁盘由 ONTAP Select 创建并拥有。
- 虚拟磁盘到物理磁盘的映射 *




虚拟磁盘配置

为了提供更加简化的用户体验， ONTAP Select 管理工具 ONTAP Deploy 会自动从关联的存储池配置虚拟磁盘并将其连接到 ONTAP Select VM 。此操作会在初始设置期间以及存储添加操作期间自动执行。如果 ONTAP Select 节点属于 HA 对，则虚拟磁盘会自动分配给本地和镜像存储池。

ONTAP Select 会将底层连接的存储拆分为大小相等的虚拟磁盘，每个虚拟磁盘不超过 16 TB 。如果 ONTAP Select 节点属于 HA 对，则在每个集群节点上至少创建两个虚拟磁盘，并将其分配给要在镜像聚合中使用的本地丛和镜像丛。

例如， ONTAP Select 可以为数据存储库或 LUN 分配 31 天的数据存储库或 LUN （部署虚拟机并配置系统和根磁盘后剩余的空间）。然后，创建四个 ~7.75TB 虚拟磁盘并将其分配给相应的 ONTAP 本地丛和镜像丛。

> 向 ONTAP Select VM 添加容量可能会导致 VMDK 的大小不同。有关详细信息，请参见一节 "增加存储容量"。与 FAS 系统不同，同一聚合中可以存在不同大小的 VMDK 。ONTAP Select 会在这些 VMDK 之间使用 RAID 0 条带，从而可以完全使用每个 VMDK 中的所有空间，而不管其大小如何。

虚拟化 **NVRAM**

NetApp FAS 系统通常配备物理 NVRAM PCI 卡，这是一种包含非易失性闪存的高性能卡。此卡使 ONTAP 能够立即确认传入的写入操作并返回到客户端，从而显著提升写入性能。此外，它还可以计划在称为转存的过程中将修改后的数据块移回速度较慢的存储介质。

商用系统通常不安装此类设备。因此，此 NVRAM 卡的功能已虚拟化并置于 ONTAP Select 系统启动磁盘上的分区中。因此，放置实例的系统虚拟磁盘极为重要。这也是该产品要求为本地连接的存储配置提供具有弹性缓存的物理 RAID 控制器的原因。

NVRAM 放置在自己的 VMDK 上。通过将 NVRAM 拆分为自己的 VMDK ， ONTAP Select VM 可以使用 vNVMe 驱动程序与其 NVRAM VMDK 进行通信。此外，还要求 ONTAP Select VM 使用与 ESX 6.5 及更高版本兼容的硬件版本 13 。

介绍的数据路径： **NVRAM** 和 **RAID** 控制器

最好通过在写入请求进入系统时浏览写入请求所占用的数据路径来突出显示虚拟化 NVRAM 系统分区与 RAID 控制器之间的交互。

传入到 ONTAP Select VM 的写入请求将定向到 VM 的 NVRAM 分区。在虚拟化层，此分区位于 ONTAP Select 系统磁盘中，即连接到 ONTAP Select VM 的 VMDK 。在物理层，这些请求会缓存在本地 RAID 控制器中，就像所有针对底层磁盘轴的块更改一样。此时，写入操作将确认回主机。

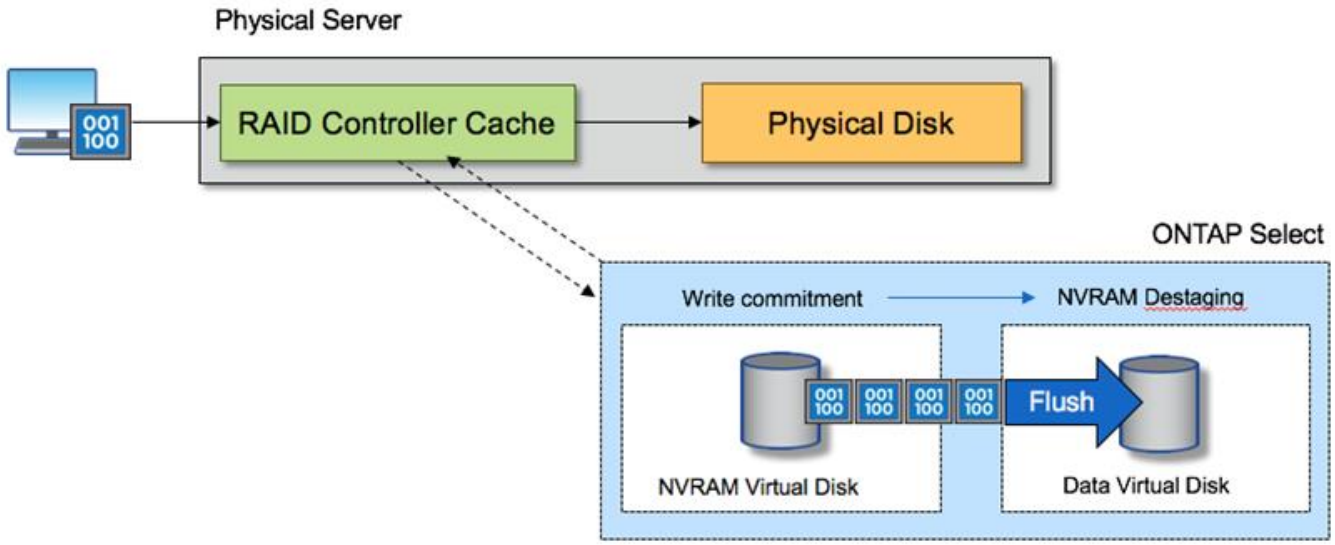
此时，该块在物理上驻留在 RAID 控制器缓存中，等待转储到磁盘。从逻辑上讲，该块驻留在 NVRAM 中，等待转存到相应的用户数据磁盘。

由于更改后的块会自动存储在 RAID 控制器的本地缓存中，因此传入到 NVRAM 分区的写入操作会自动缓存并定期转储到物理存储介质。这一点不应与定期将 NVRAM 内容刷新回 ONTAP 数据磁盘混淆。这两个事件是不相关的，发生时间和频率不同。

下图显示了传入写入所采用的 I/O 路径。其中重点介绍了物理层（由 RAID 控制器缓存和磁盘表示）与虚拟层（由虚拟机的 NVRAM 和数据虚拟磁盘表示）之间的区别。

> ⓘ 尽管 NVRAM VMDK 上更改的块会缓存在本地 RAID 控制器缓存中，但缓存无法识别 VM 构造或其虚拟磁盘。它会将所有更改过的块存储在系统上，其中 NVRAM 只是其中的一部分。如果虚拟机管理程序是从同一个后备磁盘轴配置的，则这包括绑定到该虚拟机管理程序的写入请求。

- 传入 ONTAP Select VM* 的写入




> ⓘ NVRAM 分区将在其自己的 VMDK 上分隔。该 VMDK 使用 ESX 6.5 或更高版本中提供的 vNVME 驱动程序进行连接。对于使用软件 RAID 的 ONTAP Select 安装来说，此更改最重要，因为这些安装不会从 RAID 控制器缓存中受益。

## 用于本地连接存储的软件 **RAID** 服务

软件 RAID 是在 ONTAP 软件堆栈中实施的 RAID 抽象层。它提供的功能与 FAS 等传统 ONTAP 平台中的 RAID 层相同。RAID 层执行驱动器奇偶校验计算，并针对 ONTAP Select 节点中的各个驱动器故障提供保护。

ONTAP Select 还提供了一个软件 RAID 选项，与硬件 RAID 配置无关。在某些环境中，硬件 RAID 控制器可能不可用或不受欢迎，例如在小型商用硬件上部署 ONTAP Select 时。软件 RAID 扩展了可用的部署选项，以包括此类环境。要在您的环境中启用软件 RAID ，请记住以下几点：

- 它随 Premium 或 Premium XL 许可证一起提供。
- 它仅支持 SSD 或 NVMe （需要高级 XL 许可证）驱动器用于 ONTAP 根磁盘和数据磁盘。
- ONTAP Select VM 启动分区需要一个单独的系统磁盘。
  - 选择一个单独的磁盘，即 SSD 或 NVMe 驱动器，以便为系统磁盘（在多节点设置中为 NVRAM ，启动 /CF 卡，核心转储和调解器）创建数据存储库。
- 注释 *
- 术语服务磁盘和系统磁盘可互换使用。
  - 服务磁盘是指在 ONTAP Select VM 中用于为集群，启动等各种项目提供服务的 VMDK 。
  - 服务磁盘实际位于一个物理磁盘上（统称为服务 / 系统物理磁盘），就像主机中显示的那样。该物理磁盘必须包含 DAS 数据存储库。ONTAP Deploy 会在集群部署期间为 ONTAP Select VM 创建这些服务磁盘。
- 无法在多个数据存储库之间或多个物理驱动器之间进一步分隔 ONTAP Select 系统磁盘。
- 硬件 RAID 未弃用。

本地连接存储的软件 **RAID** 配置

使用软件 RAID 时，最好不要使用硬件 RAID 控制器，但是，如果系统确实具有现有 RAID 控制器，则必须满足以下要求：

- 必须禁用硬件 RAID 控制器，以便可以将磁盘直接提供给系统（ JBOD ）。此更改通常可在 RAID 控制器 BIOS 中进行
- 或者，硬件 RAID 控制器应处于 SAS HBA 模式。例如，除了 RAID 之外，某些 BIOS 配置还允许使用 "AHCI" 模式，可以选择此模式来启用 JBOD 模式。这样可以启用直通，以便可以将物理驱动器视为主机上的物理驱动器。

根据控制器支持的最大驱动器数，可能需要额外的控制器。在 SAS HBA 模式下，确保 IO 控制器（ SAS HBA ）至少支持 6 Gb/ 秒的速度。但是， NetApp 建议使用 12 Gbps 的速度。

不支持其他硬件 RAID 控制器模式或配置。例如，某些控制器允许 RAID 0 支持，这种支持可能会人为地使磁盘实现直通，但其影响可能不受欢迎。支持的物理磁盘大小（仅限 SSD ）介于 200 GB 到 16 TB 之间。

> 管理员需要跟踪 ONTAP Select VM 正在使用哪些驱动器，并防止在主机上无意中使用这些驱动器。

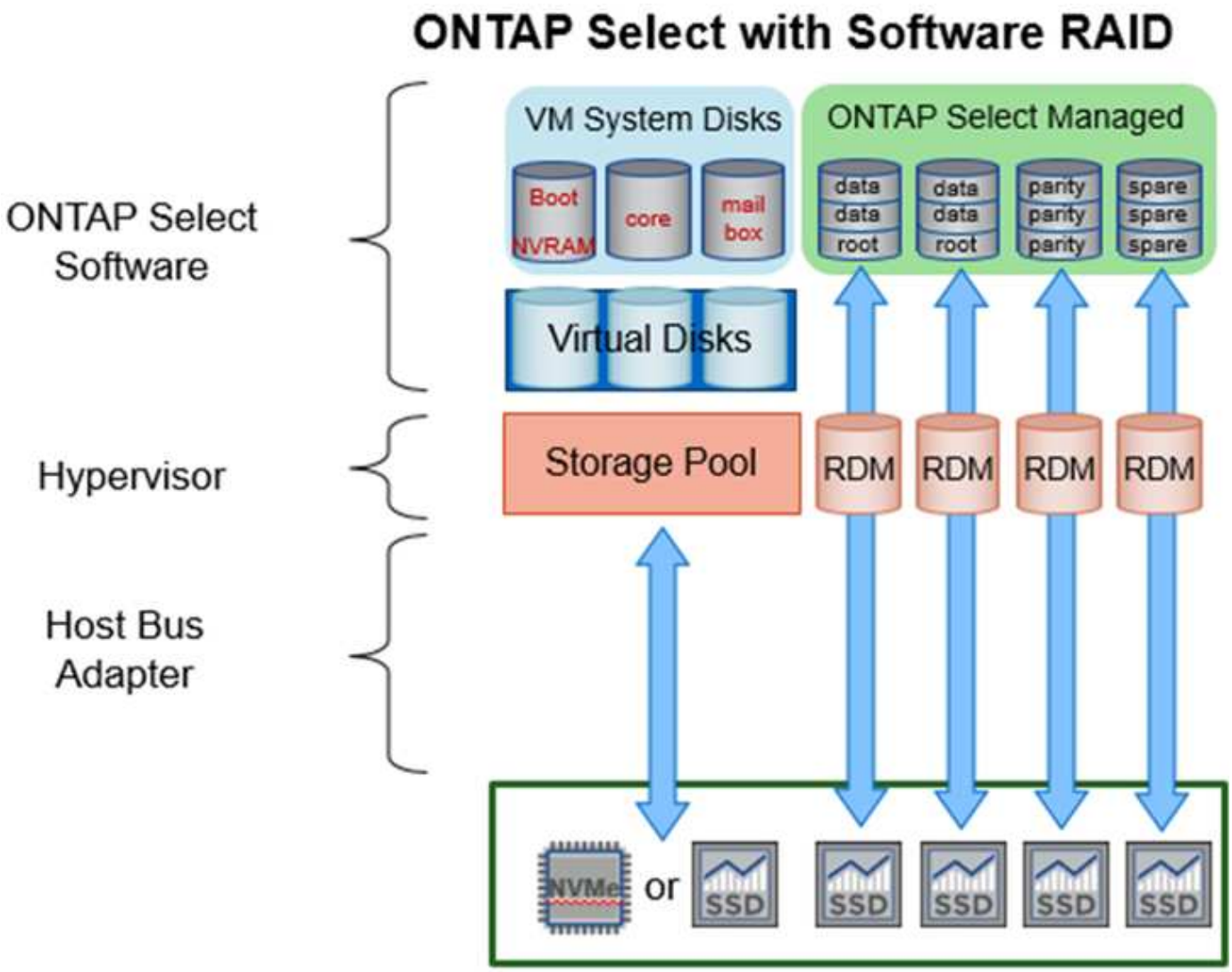
**ONTAP Select** 虚拟和物理磁盘

对于使用硬件 RAID 控制器的配置，物理磁盘冗余由 RAID 控制器提供。ONTAP Select 会显示一个或多个 VMDK ， ONTAP 管理员可以从中配置数据聚合。这些 VMDK 采用 RAID 0 格式进行条带化，因为使用 ONTAP 软件 RAID 会因硬件级别提供的故障恢复能力而变得冗余，效率低下且效率低下。此外，用于系统磁盘的 VMDK 与用于存储用户数据的 VMDK 位于同一个数据存储库中。

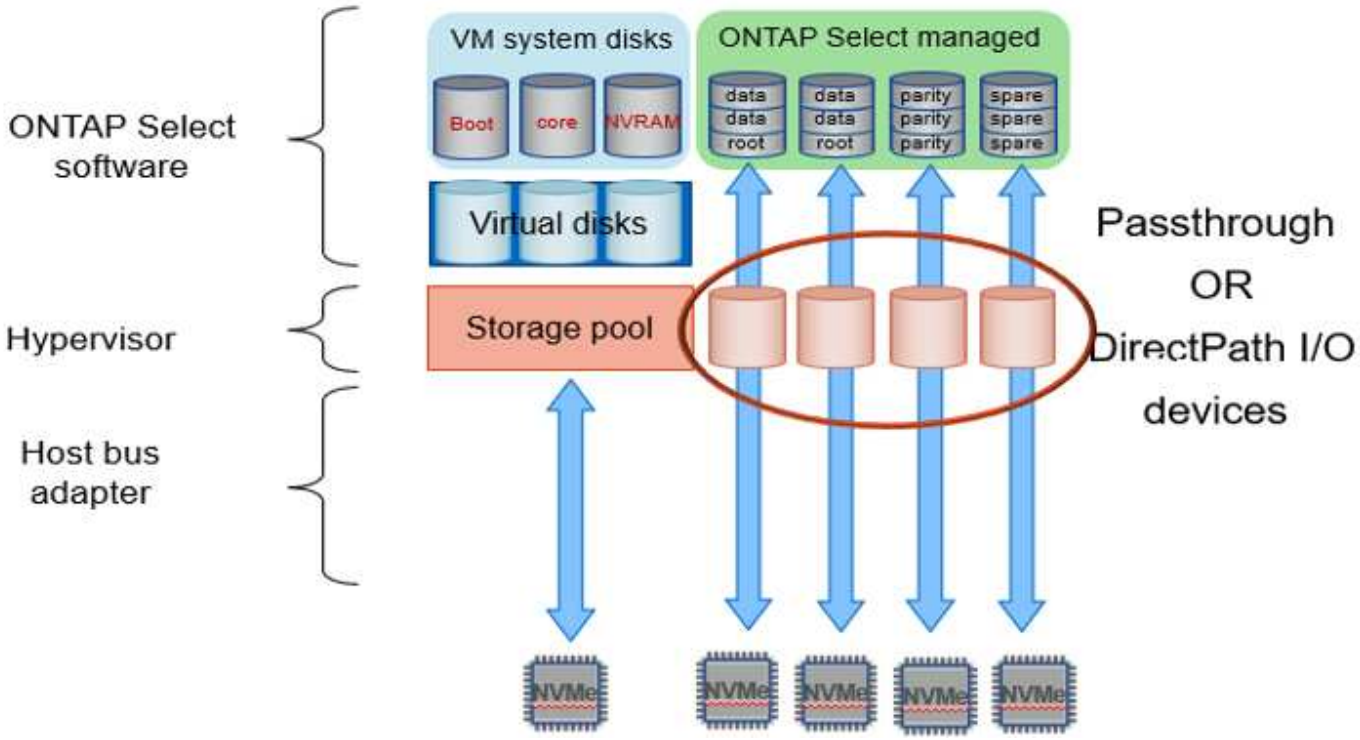
使用软件 RAID 时， ONTAP Deploy 会为 ONTAP Select 提供一组虚拟磁盘（ VMDK ）和物理磁盘原始设备映射（ RDM ），用于 SSD ，并为 NVMes 提供直通或 DirectPath IO 设备。

下图更详细地显示了这种关系，突出显示了用于 ONTAP Select VM 内部的虚拟化磁盘与用于存储用户数据的物理磁盘之间的区别。

- ONTAP Select 软件 RAID ： 使用虚拟化磁盘和 RDM *




系统磁盘（ VMDK ）位于同一个数据存储库中，并且位于同一个物理磁盘上。虚拟 NVRAM 磁盘需要一个快速且持久的介质。因此，仅支持 NVMe 和 SSD 类型的数据存储库。



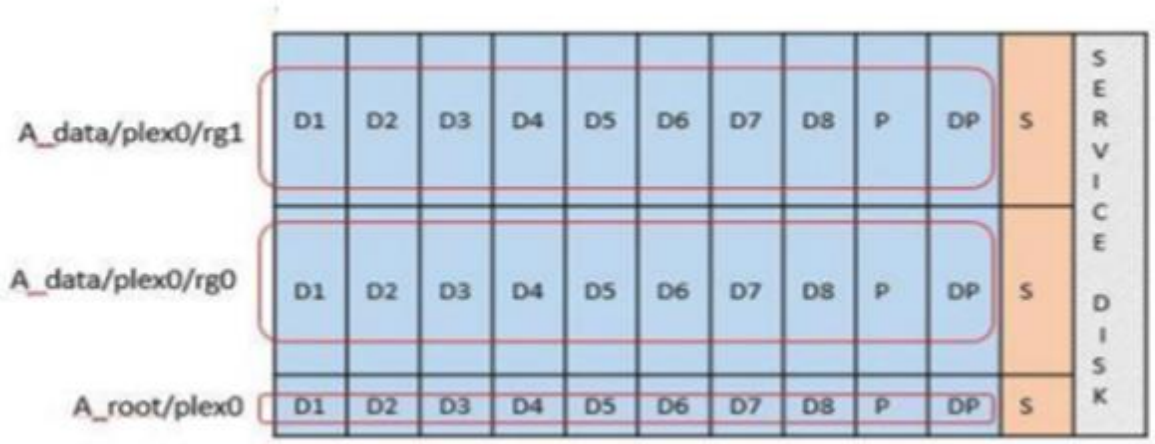

系统磁盘（VMDK）位于同一个数据存储库中，并且位于同一个物理磁盘上。虚拟 NVRAM 磁盘需要一个快速且持久的介质。因此，仅支持 NVMe 和 SSD 类型的数据存储库。在使用 NVMe 驱动器存储数据时，出于性能原因，系统磁盘也应是 NVMe 设备。在全 NVMe 配置中，最好使用 Intel Optane 卡作为系统磁盘。

> 在当前版本中，无法在多个数据存储库或多个物理驱动器之间进一步分隔 ONTAP Select 系统磁盘。

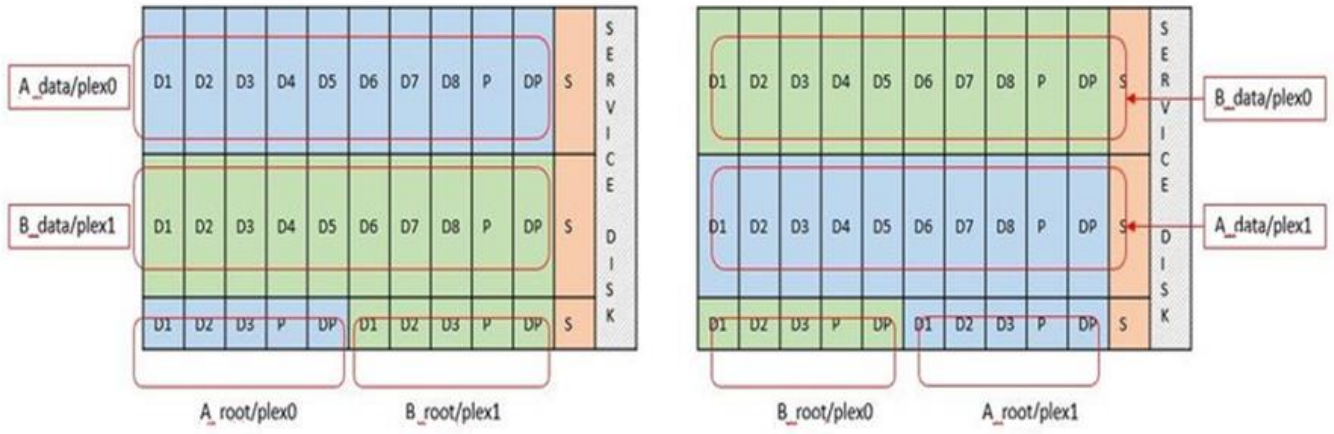
每个数据磁盘分为三部分：一个小根分区（条带）和两个大小相等的分区，用于创建 ONTAP Select VM 中可以看到的两个数据磁盘。对于单节点集群和 HA 对中的节点，分区使用根数据数据（RD2）模式，如下图所示。

`P` 表示奇偶校验驱动器。 `DP` 表示双奇偶校验驱动器和 `S` 表示备用驱动器。

- 用于单节点集群的 RDD 磁盘分区 *




- 多节点集群（HA 对）的 RDD 磁盘分区 *




ONTAP 软件 RAID 支持以下 RAID 类型： RAID 4 ， RAID-DP 和 RAID-TEC 。这些 RAID 构造与 FAS 和 AFF 平台使用的 RAID 构造相同。对于根配置， ONTAP Select 仅支持 RAID 4 和 RAID-DP 。对数据聚合使用 RAID-TEC 时，整体保护为 RAID-DP 。ONTAP Select HA 使用无共享架构将每个节点的配置复制到另一节点。这意味着每个节点都必须存储其根分区及其对等方根分区的副本。由于数据磁盘具有一个根分区，因此数据磁盘的最小数量将因 ONTAP Select 节点是否属于 HA 对而异。

对于单节点集群，所有数据分区都用于存储本地（活动）数据。对于属于 HA 对的节点，一个数据分区用于存储该节点的本地（活动）数据，另一个数据分区用于镜像来自 HA 对等方的活动数据。

直通（ **DirectPath IO** ）设备与原始设备映射（ **RDM** ）

VMware ESX 当前不支持将 NVMe 磁盘作为原始设备映射。要使 ONTAP Select 直接控制 NVMe 磁盘，必须在 ESX 中将 NVMe 驱动器配置为直通设备。请注意，将 NVMe 设备配置为直通设备需要服务器 BIOS 的支持，这是一个中断过程，需要重新启动 ESX 主机。此外，每个 ESX 主机的最大直通设备数为 16 。但是， ONTAP Deploy 将此限制为 14 。每个 ONTAP Select 节点最多 14 个 NVMe 设备这一限制意味着，全 NVMe 配置将提供极高的 IOPS 密度（ IOPS/TB ），但会影响总容量。或者，如果需要具有更大存储容量的高性能配置，建议的配置为：较大的 ONTAP Select VM 大小，系统磁盘的 Intel Optane 卡以及用于数据存储的 SSD 驱动器的标称数量。

> 要充分利用 NVMe 性能，请考虑较大的 ONTAP Select VM 大小。

直通设备和 RDM 之间还有其他区别。RDM 可以映射到正在运行的虚拟机。直通设备需要重新启动 VM 。这意味着，任何 NVMe 驱动器更换或容量扩展（驱动器添加）操作步骤 都需要重新启动 ONTAP Select VM 。驱动器更换和容量扩展（驱动器添加）操作由 ONTAP Deploy 中的工作流决定。ONTAP Deploy 可管理单节点集群的 ONTAP Select 重新启动以及 HA 对的故障转移 / 故障恢复。但是，请务必注意使用 SSD 数据驱动器（无需 ONTAP Select 重新启动 / 故障转移）与使用 NVMe 数据驱动器（需要 ONTAP Select 重新启动 / 故障转移）之间的区别。

物理和虚拟磁盘配置

为了提供更加简化的用户体验， ONTAP Deploy 会自动从指定的数据存储库（物理系统磁盘）配置系统（虚拟）磁盘，并将其连接到 ONTAP Select VM 。此操作会在初始设置期间自动执行，以便 ONTAP Select VM 可以启动。RDM 将进行分区，并自动构建根聚合。如果 ONTAP Select 节点属于 HA 对，则数据分区会自动分配给本地存储池和镜像存储池。此分配会在集群创建操作和存储添加操作期间自动进行。

由于 ONTAP Select VM 上的数据磁盘与底层物理磁盘相关联，因此使用更多物理磁盘创建配置会对性能产生影响。

> 根聚合的 RAID 组类型取决于可用磁盘的数量。ONTAP Deploy 会选择适当的 RAID 组类型。如果为节点分配了足够的磁盘，则会使用 RAID-DP ，否则会创建 RAID-4 根聚合。

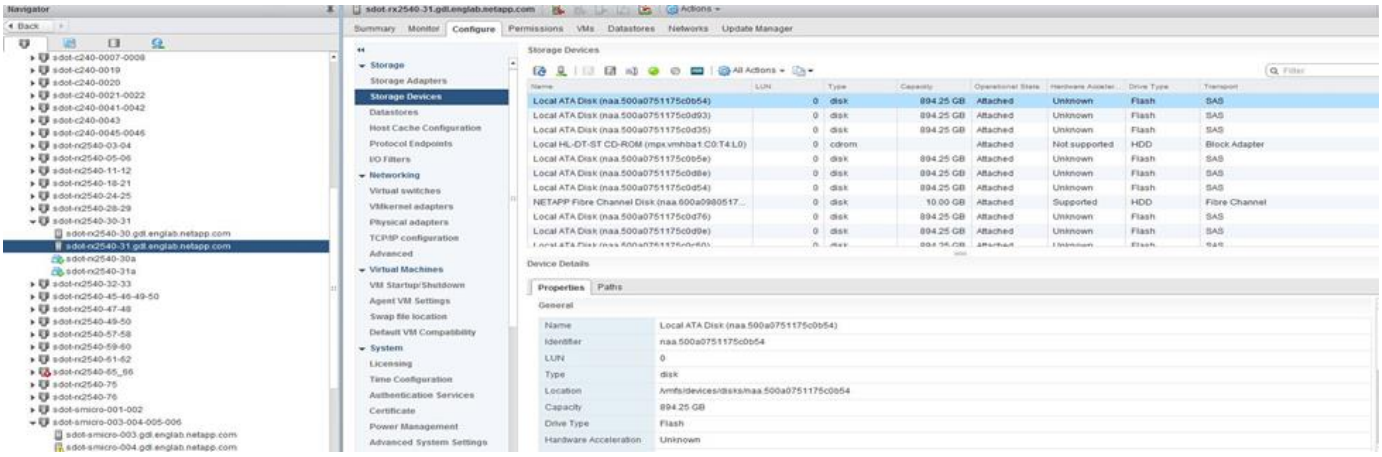
在使用软件 RAID 向 ONTAP Select VM 添加容量时，管理员必须考虑物理驱动器大小和所需的驱动器数量。有关详细信息，请参见一节 "增加存储容量"。

与 FAS 和 AFF 系统类似，只能向现有 RAID 组添加容量相等或更大的驱动器。容量较大的驱动器的大小合适。如果要创建新的 RAID 组，则新的 RAID 组大小应与现有 RAID 组大小匹配，以确保整体聚合性能不会下降。

将**ONTAP Select**磁盘与对应的**ESX**磁盘进行匹配

ONTAP Select 磁盘通常标记为 NET x.y 。您可以使用以下 ONTAP 命令获取磁盘 UUID：

```
<system name>::> disk show NET-1.1
Disk: NET-1.1
Model: Micron_5100_MTFD
Serial Number: 1723175C0B5E
UID:
*500A0751:175C0B5E*:00000000:00000000:00000000:00000000:00000000:00000000:
00000000:00000000
BPS: 512
Physical Size: 894.3GB
Position: shared
Checksum Compatibility: advanced_zoned
Aggregate: -
Plex: -This UID can be matched with the device UID displayed in the
‘storage devices’ tab for the ESX host
```




在 ESXi Shell 中，您可以输入以下命令，使给定物理磁盘（通过 na.unique-id 标识）的 LED 闪烁。

```
esxcli storage core device set -d <naa_id> -l=locator -L=<seconds>
```

使用软件 **RAID** 时出现多个驱动器故障

系统可能会遇到多个驱动器同时处于故障状态的情况。系统的行为取决于聚合 RAID 保护和故障驱动器的数量。

RAID4 聚合可以承受一个磁盘故障， RAID-DP 聚合可以承受两个磁盘故障，而 RAID-TEC 聚合可以承受三个磁盘故障。

如果故障磁盘数小于 RAID 类型支持的最大故障数，并且备用磁盘可用，则重建过程将自动开始。如果备用磁盘不可用，则聚合将在降级状态下提供数据，直到添加备用磁盘为止。

如果故障磁盘数超过 RAID 类型支持的最大故障数，则本地丛将标记为故障，并且聚合状态为降级。数据由 HA 配对节点上的第二个丛提供。这意味着，节点 1 的任何 I/O 请求都会通过集群互连端口 e0e （ iSCSI ）发送到物理上位于节点 2 上的磁盘。如果第二个丛也发生故障，则聚合将标记为发生故障，并且数据不可用。

必须删除并重新创建故障丛，才能恢复正确的数据镜像。请注意，如果多磁盘故障导致数据聚合降级，则根聚合也会降级。ONTAP Select 使用根 - 数据 - 数据（ RDD ）分区方案将每个物理驱动器拆分为一个根分区和两个数据分区。因此，丢失一个或多个磁盘可能会影响多个聚合，包括本地根聚合或远程根聚合的副本，以及本地数据聚合和远程数据聚合的副本。

```
C3111E67::> storage aggregate plex delete -aggregate aggr1 -plex plex1
Warning: Deleting plex "plex1" of mirrored aggregate "aggr1" in a non-
shared HA configuration will disable its synchronous mirror protection and
disable
         negotiated takeover of node "sti-rx2540-335a" when aggregate
"aggr1" is online.
Do you want to continue? {y|n}: y
[Job 78] Job succeeded: DONE

C3111E67::> storage aggregate mirror -aggregate aggr1
Info: Disks would be added to aggregate "aggr1" on node "sti-rx2540-335a"
in the following manner:
      Second Plex
        RAID Group rg0, 5 disks (advanced_zoned checksum, raid_dp)
                                                                Usable
Physical
          Position   Disk                      Type           Size
Size
          ---------- ------------------------- ---------- --------
--------
          shared     NET-3.2                   SSD               -
-
          shared     NET-3.3                   SSD               -
-
          shared     NET-3.4                   SSD         208.4GB
208.4GB
          shared     NET-3.5                   SSD         208.4GB
208.4GB
          shared     NET-3.12                  SSD         208.4GB
```

```
208.4GB

      Aggregate capacity available for volume use would be 526.1GB.
      625.2GB would be used from capacity license.
Do you want to continue? {y|n}: y

C3111E67::> storage aggregate show-status -aggregate aggr1
Owner Node: sti-rx2540-335a
 Aggregate: aggr1 (online, raid_dp, mirrored) (advanced_zoned checksums)
  Plex: /aggr1/plex0 (online, normal, active, pool0)
   RAID Group /aggr1/plex0/rg0 (normal, advanced_zoned checksums)
                                                                  Usable
Physical
     Position Disk                        Pool Type     RPM     Size
Size Status
     -------- -------------------------- ---- ----- ------ --------
-------- ----------
     shared   NET-1.1                      0   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-1.2                      0   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-1.3                      0   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-1.10                     0   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-1.11                     0   SSD        -  205.1GB
447.1GB (normal)
  Plex: /aggr1/plex3 (online, normal, active, pool1)
   RAID Group /aggr1/plex3/rg0 (normal, advanced_zoned checksums)
                                                                  Usable
Physical
     Position Disk                        Pool Type     RPM     Size
Size Status
     -------- -------------------------- ---- ----- ------ --------
-------- ----------
     shared   NET-3.2                      1   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-3.3                      1   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-3.4                      1   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-3.5                      1   SSD        -  205.1GB
447.1GB (normal)
     shared   NET-3.12                     1   SSD        -  205.1GB
447.1GB (normal)
10 entries were displayed..
```

> 要测试或模拟一个或多个驱动器故障、请使用 `storage disk fail -disk NET-x.y -immediate` 命令：如果系统中有备用磁盘，聚合将开始重建。您可以使用命令检查重建状态 `storage aggregate show`。您可以使用 ONTAP Deploy 删除模拟故障驱动器。请注意、ONTAP已将驱动器标记为 `Broken`。驱动器实际上未损坏，可以使用 ONTAP Deploy 重新添加。要擦除损坏的标签，请在 ONTAP Select 命令行界面中输入以下命令：

```
set advanced
disk unfail -disk NET-x.y -spare true
disk show -broken
```

最后一个命令的输出应为空。

#### 虚拟化 **NVRAM**

NetApp FAS 系统通常配备物理 NVRAM PCI 卡。此卡是一种高性能卡，包含非易失性闪存，可显著提升写入性能。为此，它授予 ONTAP 立即确认传入的写入客户端的能力。此外，它还可以计划在一个称为转存的过程中将修改后的数据块移回速度较慢的存储介质。

商用系统通常不安装此类设备。因此， NVRAM 卡的功能已虚拟化并置于 ONTAP Select 系统启动磁盘上的分区中。因此，放置实例的系统虚拟磁盘极为重要。

## **vSAN** 和外部阵列配置

虚拟NAS (vNAS)部署支持虚拟SAN (VSAN)上的ONTAP Select集群、某些HCI产品以及外部阵列类型的数据存储库。这些配置的底层基础架构可提供数据存储库故障恢复能力。

最低要求是，底层配置受 VMware 支持，并应列在相应的 VMware HCL 上。

#### **vNAS** 架构

所有不使用 DAS 的设置都使用 vNAS 命名。对于多节点 ONTAP Select 集群，这包括一个架构，同一 HA 对中的两个 ONTAP Select 节点共享一个数据存储库（包括 vSAN 数据存储库）。节点也可以安装在与同一共享外部阵列不同的数据存储库上。这样可以提高阵列端存储效率，从而减少整个 ONTAP Select HA 对的整体占用空间。ONTAP Select vNAS 解决方案的架构与使用本地 RAID 控制器的 DAS 上的 ONTAP Select 非常相似。也就是说，每个 ONTAP Select 节点仍有一份其 HA 配对节点数据的副本。ONTAP 存储效率策略的范围为节点范围。因此，最好使用阵列端存储效率，因为它们可能会应用于两个 ONTAP Select 节点的数据集。

HA 对中的每个 ONTAP Select 节点也可能使用单独的外部阵列。在将 ONTAP Select MetroCluster SDS 与外部存储结合使用时，这是一个常见的选择。

在为每个 ONTAP Select 节点使用单独的外部阵列时，两个阵列必须提供与 ONTAP Select VM 类似的性能特征，这一点非常重要。

##### **vNAS** 架构与具有硬件 **RAID** 控制器的本地 **DAS** 的对比

vNAS 架构在逻辑上与具有 DAS 和 RAID 控制器的服务器的架构最相似。在这两种情况下， ONTAP Select 都会占用数据存储库空间。该数据存储库空间会划分到 VMDK 中，这些 VMDK 构成传统的 ONTAP 数据聚合。ONTAP Deploy 可确保在集群 -create 和 storage-add 操作期间， VMDK 大小正确并分配给正确的丛（对于 HA 对）。

使用 RAID 控制器时， vNAS 与 DAS 之间存在两个主要区别。最直接的区别是， vNAS 不需要 RAID 控制器。vNAS 假定底层外部阵列可提供具有 RAID 控制器设置的 DAS 所能提供的数据持久性和故障恢复能力。第二个更微妙的区别在于 NVRAM 性能。

**vNAS NVRAM**

ONTAP Select NVRAM 是 VMDK 。换言之， ONTAP Select 在块寻址设备（ VMDK ）上模拟字节寻址空间（传统 NVRAM ）。但是， NVRAM 的性能对于 ONTAP Select 节点的整体性能绝对重要。

对于使用硬件 RAID 控制器的 DAS 设置，硬件 RAID 控制器缓存充当事实上的 NVRAM 缓存，因为对 NVRAM VMDK 的所有写入操作首先托管在 RAID 控制器缓存中。

对于 vNAS 架构， ONTAP Deploy 会使用名为单实例数据日志记录（ SIDI ）的启动参数自动配置 ONTAP Select 节点。如果存在此启动参数，则 ONTAP Select 将绕过 NVRAM 并将数据有效负载直接写入数据聚合。NVRAM 仅用于记录写入操作更改的块的地址。此功能的优势在于，它可以避免双重写入：一个写入 NVRAM ，另一个写入在 NVRAM 转存时。此功能仅适用于 vNAS ，因为本地写入 RAID 控制器缓存的额外延迟可忽略不计。

SIDI 功能与所有 ONTAP Select 存储效率功能不兼容。可以使用以下命令在聚合级别禁用 SIDI 功能：

```
storage aggregate modify -aggregate aggr-name -single-instance-data
-logging off
```

请注意，如果关闭了 SIDI 功能，则写入性能会受到影响。禁用此聚合中所有卷上的所有存储效率策略后，可以重新启用 SIDI 功能：

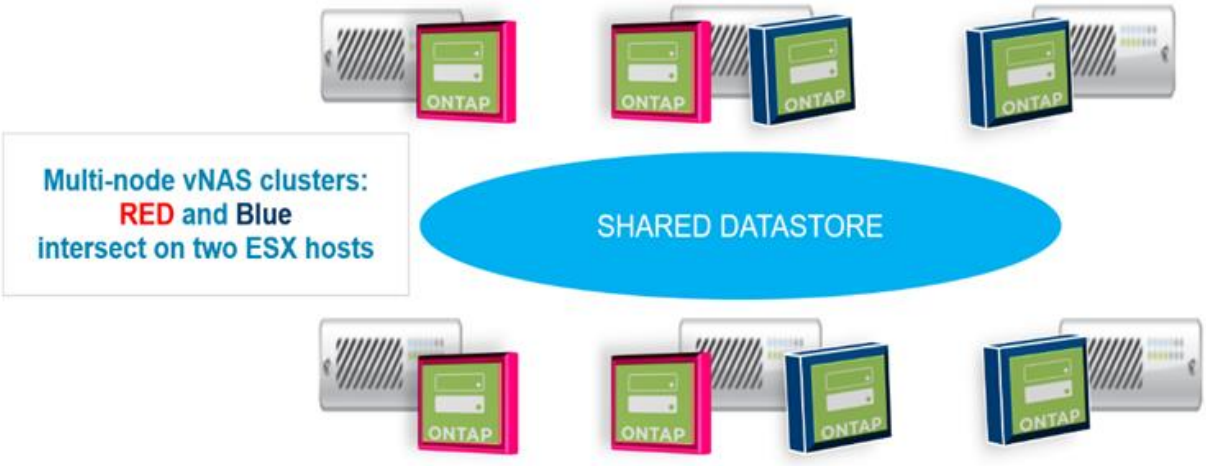
```
volume efficiency stop -all true -vserver * -volume * (all volumes in the
affected aggregate)
```

使用**vNAs**时、请在**ONTAP Select**节点上配置

ONTAP Select 支持在共享存储上使用多节点 ONTAP Select 集群。ONTAP Deploy 支持在同一 ESX 主机上配置多个 ONTAP Select 节点，前提是这些节点不属于同一集群。请注意，此配置仅适用于 vNAS 环境（共享数据存储库）。使用 DAS 存储时，不支持每个主机使用多个 ONTAP Select 实例，因为这些实例争用同一硬件 RAID 控制器。

ONTAP Deploy 可确保首次部署多节点 vNAS 集群时不会将同一集群中的多个 ONTAP Select 实例放置在同一主机上。下图显示了正确部署两个在两个主机上交叉的四节点集群的示例。

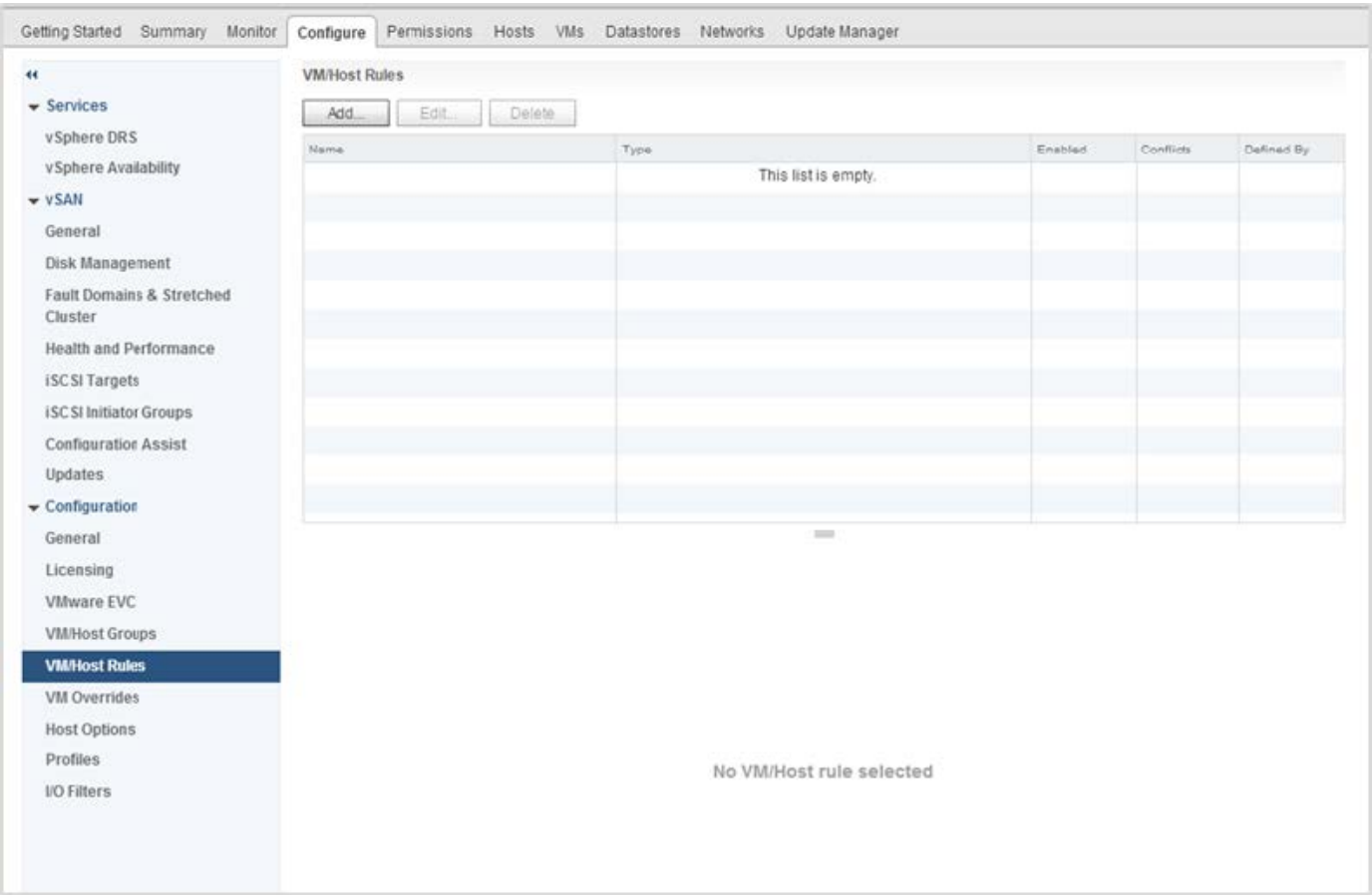
- 首次部署多节点 vNAS 集群 *

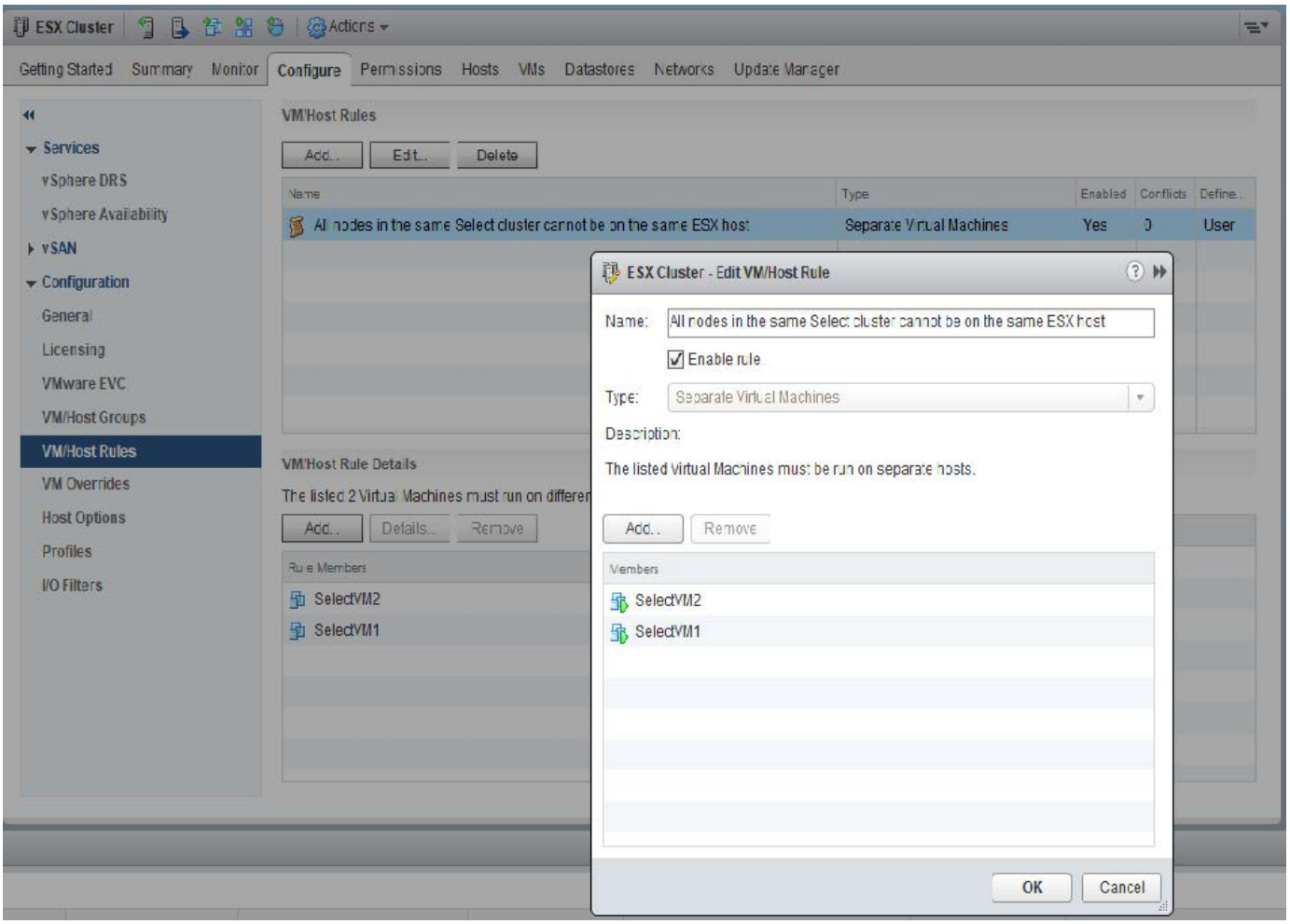



部署后，可以在主机之间迁移 ONTAP Select 节点。这可能会导致配置不是最佳配置，并且不受支持，因为同一集群中的两个或更多 ONTAP Select 节点共享同一个底层主机。NetApp 建议手动创建 VM 反关联性规则，以便 VMware 自动在同一集群的节点之间保持物理隔离，而不仅仅是同一 HA 对中的节点。

> 反关联性规则要求在 ESX 集群上启用 DRS 。

有关如何为 ONTAP Select VM 创建反关联性规则的示例，请参见以下示例。如果 ONTAP Select 集群包含多个 HA 对，则该集群中的所有节点都必须包含在此规则中。






由于以下原因之一，可能会在同一 ESX 主机上找到同一 ONTAP Select 集群中的两个或更多 ONTAP Select 节点：

- 由于 VMware vSphere 许可证限制或未启用 DRS ， DRS 不存在。
- 绕过 DRS 反关联性规则，因为 VMware HA 操作或管理员启动的虚拟机迁移优先。

请注意， ONTAP Deploy 不会主动监控 ONTAP Select VM 位置。但是，集群刷新操作会在 ONTAP Deploy 日志中反映此不受支持的配置：


## 增加存储容量

ONTAP Deploy 可用于为 ONTAP Select 集群中的每个节点添加和许可额外的存储。

ONTAP Deploy 中的存储添加功能是增加所管理存储的唯一方法，不支持直接修改 ONTAP Select VM 。下图显示了启动存储添加向导的 "+" 图标。


以下注意事项对于容量扩展操作的成功非常重要。要添加容量，需要使用现有许可证来涵盖总空间量（现有空间加上新空间）。导致节点超出其许可容量的存储添加操作将失败。应首先安装具有足够容量的新许可证。

如果向现有 ONTAP Select 聚合添加了额外容量，则新存储池（数据存储库）的性能配置文件应与现有存储池（数据存储库）的性能配置文件类似。请注意，不能将非 SSD 存储添加到安装了类似于 AFF 的特性（已启用闪存）的 ONTAP Select 节点。也不支持混合使用 DAS 和外部存储。

如果将本地连接的存储添加到系统中以提供额外的本地（ DAS ）存储池，则必须构建额外的 RAID 组和 LUN （或 LUN ）。与 FAS 系统一样，如果要向同一聚合添加新空间，应注意确保新 RAID 组的性能与原始 RAID 组的性能相似。如果要创建新聚合，则如果清楚了解新聚合的性能影响，则新 RAID 组布局可能会有所不同。

如果数据存储库的总大小不超过 ESX 支持的最大数据存储库大小，则可以将新空间作为块区添加到同一个数据存储库中。可以动态地向已安装 ONTAP Select 的数据存储库添加数据存储库扩展，而不会影响 ONTAP Select 节点的操作。

如果 ONTAP Select 节点属于 HA 对，则应考虑其他一些问题。

在 HA 对中，每个节点都包含其配对节点的数据的镜像副本。向节点 1 添加空间要求向其配对节点 2 添加相同的空间量，以便将节点 1 中的所有数据复制到节点 2 。换言之，在节点 1 的容量添加操作中添加到节点 2 的空间在节点 2 上不可见或不可访问。将空间添加到节点 2 ，以便在发生 HA 事件期间，节点 1 的数据得到完全保护。

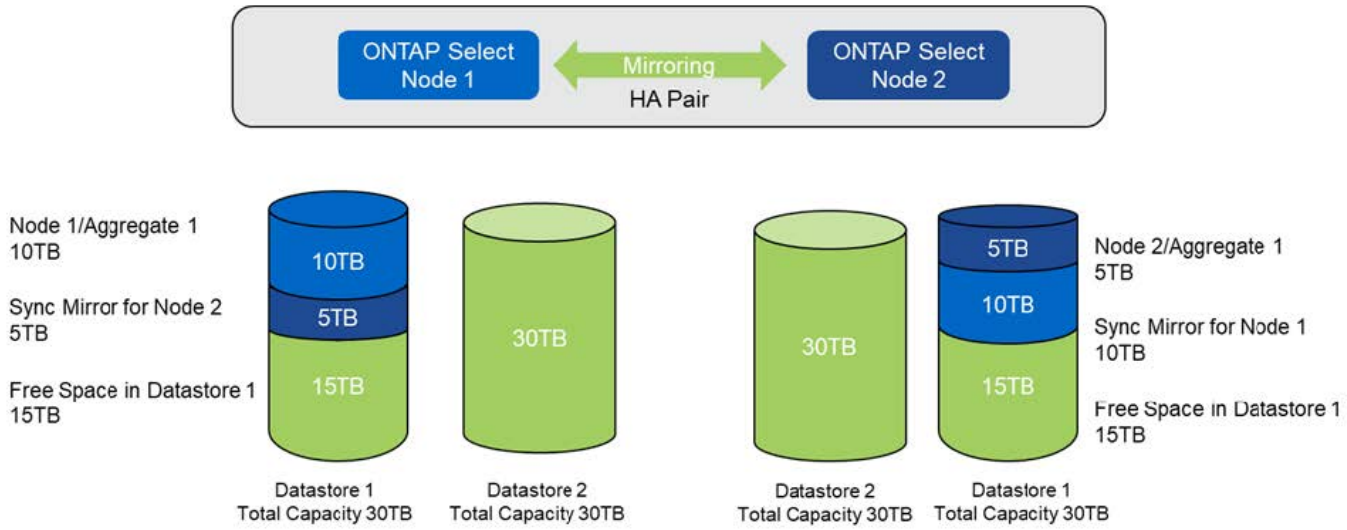
有关性能，还有一个额外的注意事项。节点 1 上的数据会同步复制到节点 2 。因此，节点 1 上新空间（数据存储库）的性能必须与节点 2 上新空间（数据存储库）的性能相匹配。换言之，在两个节点上添加空间，但使用不同的驱动器技术或不同的 RAID 组大小，可能会导致性能问题。这是因为 RAID SyncMirror 操作用于在配对节点上维护数据副本。

要增加 HA 对中两个节点上用户可访问的容量，必须执行两个存储添加操作，每个节点一个。每次存储添加操作都需要在两个节点上增加空间。每个节点上所需的总空间等于节点 1 上所需的空间加上节点 2 上所需的空间。

初始设置包含两个节点，每个节点具有两个数据存储库，每个数据存储库具有 30 TB 的空间。ONTAP Deploy 会创建一个双节点集群，其中每个节点都会占用数据存储库 1 中的 10 TB 空间。ONTAP Deploy 会为每个节点配置 5 TB 的活动空间。

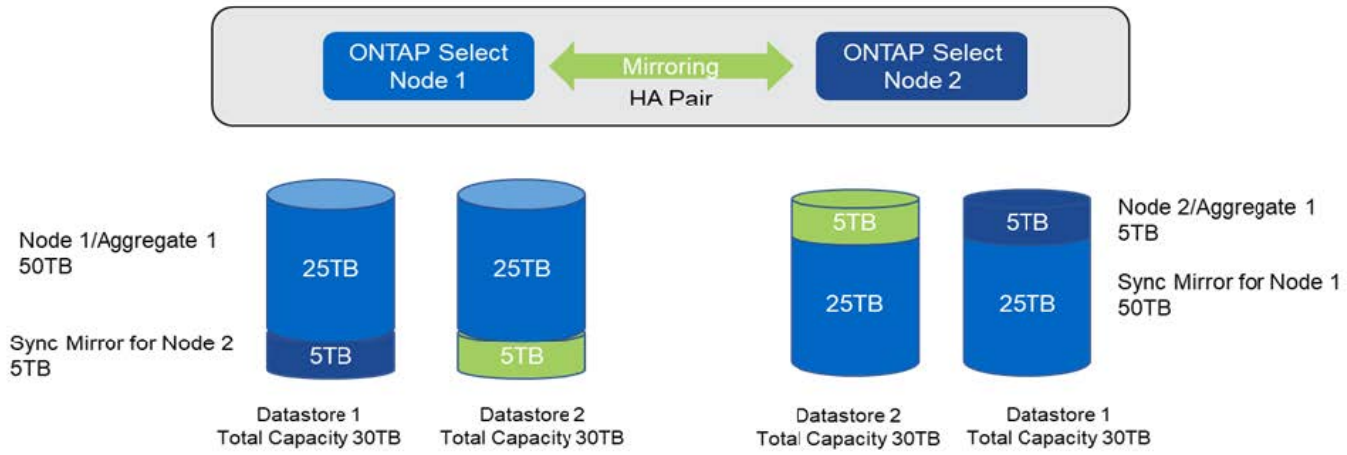
下图显示了节点 1 的单个存储添加操作的结果。ONTAP Select 仍会在每个节点上使用相等的存储容量（ 15 TB ）。但是，节点 1 的活动存储（ 10 TB ）比节点 2 （ 5 TB ）更多。两个节点均受到完全保护，因为每个节点都托管另一节点的数据副本。数据存储库 1 中还有额外的可用空间，数据存储库 2 仍完全可用。

- 容量分布：在一次存储添加操作之后分配和可用空间 *




节点 1 上的两个额外的存储添加操作会占用数据存储库 1 的其余部分和数据存储库 2 的一部分（使用容量上限）。第一个 storage-add 操作会占用数据存储库 1 中剩余的 15 TB 可用空间。下图显示了第二个 storage-add 操作的结果。此时，节点 1 管理着 50 TB 的活动数据，而节点 2 管理着原始 5 TB 的活动数据。

- 容量分布：对节点 1* 执行两次额外的存储添加操作后的分配和可用空间




容量添加操作期间使用的最大 VMDK 大小为 16 TB 。集群创建操作期间使用的最大 VMDK 大小仍为 8 TB 。ONTAP Deploy 会根据您的配置（单节点或多节点集群）以及要添加的容量创建大小正确的 VMDK 。但是，在集群创建操作期间，每个 VMDK 的最大大小不应超过 8 TB ，在存储添加操作期间，最大大小不应超过 16 TB 。

使用软件**RAID**增加**ONTAP Select**的容量

同样，也可以使用 storage-add 向导来增加使用软件 RAID 的 ONTAP Select 节点所管理的容量。此向导仅会显示可用的 DAS SDD 驱动器，这些驱动器可以作为 RDM 映射到 ONTAP Select VM 。

虽然可以将容量许可证增加一 TB ，但在使用软件 RAID 时，无法以物理方式将容量增加一 TB 。与向 FAS 或 AFF 阵列添加磁盘类似，某些因素决定了可在一次操作中添加的最小存储量。

请注意，在 HA 对中，向节点 1 添加存储要求节点的 HA 对（节点 2 ）上也具有相同数量的驱动器。在节点 1 上执行一次 storage-add 操作会同时使用本地驱动器和远程磁盘。也就是说，使用远程驱动器确保节点 1 上的新存储在节点 2 上进行复制和保护。要在节点 2 上添加本地可用的存储，必须在两个节点上分别执行存储添加操

作并使用相同数量的单独驱动器。

ONTAP Select 会将任何新驱动器分区为与现有驱动器相同的根，数据和数据分区。分区操作会在创建新聚合期间或扩展现有聚合期间执行。每个磁盘上的根分区条带大小设置为与现有磁盘上的现有根分区大小匹配。因此，两个相等的数据分区大小中的每一个都可以计算为磁盘总容量减去根分区大小除以 2 。根分区条带大小是可变的，它是在初始集群设置期间按如下所示计算的。所需的总根空间（单节点集群为 68 GB ， HA 对为 136 GB ）除以任何备用驱动器和奇偶校验驱动器的初始磁盘数。要添加到系统的所有驱动器上的根分区条带大小保持不变。

如果要创建新聚合，则所需的最小驱动器数会因 RAID 类型以及 ONTAP Select 节点是否属于 HA 对而异。

如果要向现有聚合添加存储，则需要考虑一些额外的注意事项。可以将驱动器添加到现有 RAID 组，前提是 RAID 组尚未达到最大限制。此处也适用向现有 RAID 组添加磁盘轴的传统 FAS 和 AFF 最佳实践，因此，在新磁盘轴上创建热点可能是一个潜在问题。此外，只能将数据分区大小相等或更大的驱动器添加到现有 RAID 组中。如上所述，数据分区大小与驱动器原始大小不同。如果要添加的数据分区大于现有分区，则新驱动器的大小将会合适。换言之，每个新驱动器的一部分容量仍会处于未使用状态。

也可以使用新驱动器在现有聚合中创建新的 RAID 组。在这种情况下， RAID 组大小应与现有 RAID 组大小匹配。

## 存储效率支持

## ONTAP Select 提供的存储效率选项与 FAS 和 AFF 阵列上的存储效率选项类似。

使用全闪存VSAN或通用闪存阵列部署ONTAP Select 虚拟NAS (vNAS)时、应遵循使用非SSD直连存储(DAS)的ONTAP Select 的最佳实践。

只要您的DAS存储具有SSD驱动器和高级许可证、就会在新安装中自动启用类似于AFF的特性。

如果具有类似于 AFF 的特性，则在安装期间会自动启用以下实时 SE 功能：

- 实时零模式检测
- 卷实时重复数据删除
- 卷后台重复数据删除
- 自适应实时压缩
- 实时数据缩减
- 聚合实时重复数据删除
- 聚合后台重复数据删除

要验证 ONTAP Select 是否已启用所有默认存储效率策略，请在新创建的卷上运行以下命令：

```
<system name>::> set diag
Warning: These diagnostic commands are for use by NetApp personnel only.
Do you want to continue? {y|n}: y
twonode95IP15::*> sis config
Vserver:                                SVM1
Volume:                                 _export1_NFS_volume
Schedule                                -
Policy:                                 auto
Compression:                            true
Inline Compression:                     true
Compression Type:                       adaptive
Application IO Si                       8K
Compression Algorithm:                  lzopro
Inline Dedupe:                          true
Data Compaction:                        true
Cross Volume Inline Deduplication:      true
Cross Volume Background Deduplication:  true
```

> 对于从9.6及更高版本升级的ONTAP Select 、您必须使用高级许可证在DAS SSD存储上安装ONTAP Select。此外、在使用ONTAP Deploy进行初始集群安装期间、您必须选中*启用存储效率*复选框。在未满足先前条件的情况下，要在 ONTAP 升级后启用类似于 AFF 的特性，需要手动创建启动参数并重新启动节点。有关更多详细信息，请联系技术支持。

**ONTAP Select** 存储效率配置

下表汇总了各种可用的存储效率选项、默认情况下启用的选项或默认情况下不启用但建议使用的选项、具体取决于介质类型和软件许可证。

| **ONTAP Select** 功能 | **DAS SSD** (高级或高级**XL** [1]) | **DAS HDD** (所有许可证) | **vNAS** (所有许可证) |
|---|---|---|---|
| 实时零检测 | 是（默认） | 是由用户按卷启用 | 是由用户按卷启用 |
| 卷实时重复数据删除 | 是（默认） | 不可用 | 不支持 |
| 32 K 实时压缩（二级压缩） | 是由用户逐卷启用。 | 是由用户按卷启用 | 不支持 |
| 8 K 实时压缩（自适应压缩） | 是（默认） | 是由用户逐卷启用 | 不支持 |
| 后台数据压缩 | 不支持 | 是由用户逐卷启用 | 是由用户按卷启用 |
| 数据压缩扫描程序 | 是的。 | 是的。 | 是由用户按卷启用 |
| 实时数据缩减 | 是（默认） | 是由用户逐卷启用 | 不支持 |
| 数据缩减扫描程序 | 是的。 | 是的。 | 不支持 |
| 聚合实时重复数据删除 | 是（默认） | 不适用 | 不支持 |
| 卷后台重复数据删除 | 是（默认） | 是由用户逐卷启用 | 是由用户按卷启用 |
| 聚合后台重复数据删除 | 是（默认） | 不适用 | 不支持 |

[1]ONTAP Select 9.6支持新许可证(高级XL)和新的VM大小(大型)。但是，只有使用软件 RAID 的 DAS 配置才支持大型 VM 。在9.6版中、大型ONTAP Select VM不支持硬件RAID和vNAS配置。

有关 **DAS SSD** 配置的升级行为的注释

升级到ONTAP Select 9.6或更高版本后、请等待 `system node upgrade-revert show` 命令以指示升级已完成、然后再验证现有卷的存储效率值。

在升级到ONTAP Select 9.6或更高版本的系统上、在现有聚合或新创建的聚合上创建的新卷与在全新部署中创建的卷具有相同的行为。已进行 ONTAP Select 代码升级的现有卷与新创建的卷具有大多数相同的存储效率策略，但存在一些变体：

场景 **1**

如果在升级之前未在卷上启用存储效率策略、则：

- 具有的卷 `space guarantee = volume` 未启用实时数据缩减、聚合实时重复数据删除和聚合后台重复数据删除。这些选项可以在升级后启用。
- 具有的卷 `space guarantee = none` 未启用后台数据压缩。此选项可在升级后启用。
- 升级后，现有卷上的存储效率策略将设置为 auto 。

方案**2.**

如果在升级之前已在卷上启用了某些存储效率、则：

- 具有的卷 `space guarantee = volume` 升级后看不到任何差异。
- 具有的卷 `space guarantee = none` 启用聚合后台重复数据删除。
- 具有的卷 `storage policy inline-only` 将其策略设置为auto。
- 具有用户定义的存储效率策略的卷不会更改策略、但具有的卷除外 `space guarantee = none`。这些卷已启用聚合后台重复数据删除。

# 网络

## 网络连接：一般概念和特征

首先，熟悉适用于 ONTAP Select 环境的一般网络概念。然后，了解单节点和多节点集群的具体特征和选项。

### 物理网络

物理网络主要通过提供底层第二层交换基础架构来支持 ONTAP Select 集群部署。与物理网络相关的配置包括虚拟机管理程序主机和更广泛的交换网络环境。

#### 主机 **NIC** 选项

每个 ONTAP Select 虚拟机管理程序主机都必须配置两个或四个物理端口。您选择的确切配置取决于多种因素，包括：

- 集群包含一个或多个 ONTAP Select 主机

- 使用的是什么虚拟机管理程序操作系统
- 如何配置虚拟交换机
- 链路是否使用 LACP

#### 物理交换机配置

您必须确保物理交换机的配置支持 ONTAP Select 部署。物理交换机与基于虚拟机管理程序的虚拟交换机集成在一起。您选择的确切配置取决于多种因素。主要注意事项包括：

- 如何在内部网络和外部网络之间保持隔离？
- 您是否会在数据网络和管理网络之间保持隔离？
- 如何配置第二层 VLAN ？

### 逻辑网络连接

ONTAP Select 使用两个不同的逻辑网络，根据类型分隔流量。具体而言，流量可以在集群中的主机之间流动，也可以流向存储客户端和集群外的其他计算机。虚拟机管理程序管理的虚拟交换机有助于支持逻辑网络。

#### 内部网络

在多节点集群部署中，各个 ONTAP Select 节点使用隔离的 " 内部 " 网络进行通信。此网络不会在 ONTAP Select 集群中的节点之外公开或可用。

> 只有多节点集群存在内部网络。

内部网络具有以下特征：

- 用于处理 ONTAP 集群内流量，包括：
  - 集群
  - 高可用性互连（ HA-IC ）
  - RAID 同步镜像（ RSM ）
- 基于 VLAN 的单个第 2 层网络
- 静态 IP 地址由 ONTAP Select 分配：
  - 仅 IPv4
  - 未使用 DHCP
  - 链路本地地址
- 默认情况下， MTU 大小为 9000 字节，可在 7500-9000 范围（包括在内）内进行调整

#### 外部网络

外部网络处理 ONTAP Select 集群节点与外部存储客户端以及其他计算机之间的流量。外部网络是每个集群部署的一部分，具有以下特征：

- 用于处理 ONTAP 流量，包括：

    - 数据（ NFS ， CIFS ， iSCSI ）
    - 管理（集群和节点；可选 SVM ）
    - 集群间（可选）
- 也可以支持 VLAN ：
    - 数据端口组
    - 管理端口组
- 根据管理员的配置选择分配的 IP 地址：
    - IPv4或IPv6
- 默认情况下， MTU 大小为 1500 字节（可调整）

外部网络包含所有大小的集群。

虚拟机网络环境

虚拟机管理程序主机可提供多种网络功能。

ONTAP Select 依靠虚拟机提供的以下功能：

虚拟机端口

ONTAP Select 可使用多个端口。它们根据进行分配和使用
包括集群大小在内的多种因素。

虚拟交换机

虚拟机管理程序环境中的虚拟交换机软件、无论是vSwitch (VMware)还是
Open vSwitch (KVM)将虚拟机公开的端口与物理以太网连接起来
NIC端口。您必须根据需要为每个ONTAP Select主机配置vSwitch
环境。

## 单节点和多节点网络配置

## ONTAP Select 既支持单节点网络配置，也支持多节点网络配置。

单节点网络配置

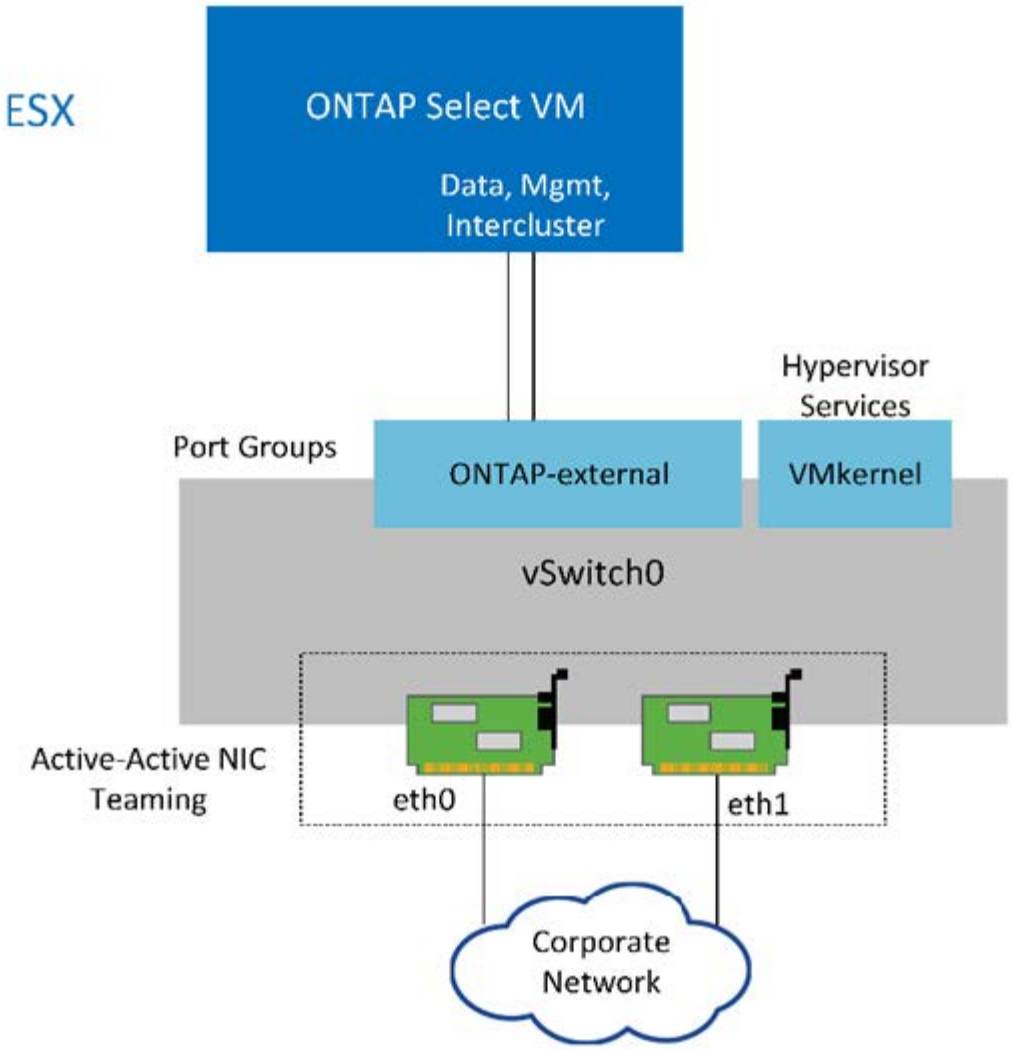
单节点 ONTAP Select 配置不需要 ONTAP 内部网络，因为不存在集群， HA 或镜像流量。

与多节点版本的 ONTAP Select 产品不同，每个 ONTAP Select VM 都包含三个虚拟网络适配器，它们提供给 ONTAP 网络端口 e0a ， e0b 和 e0c 。

这些端口用于提供以下服务：管理，数据和集群间 LIF 。

下图显示了这些端口与底层物理适配器之间的关系，其中显示了 ESX 虚拟机管理程序上的一个 ONTAP Select 集群节点。

- 单节点 ONTAP Select 集群的网络配置 *




即使两个适配器足以用于单节点集群，仍需要 NIC 绑定。

**LIF** 分配

如本文档的多节点 LIF 分配一节所述， ONTAP Select 使用 IP 空间将集群网络流量与数据和管理流量分开。此平台的单节点变体不包含集群网络。因此，集群 IP 空间中不存在任何端口。

集群和节点管理 LIF 会在 ONTAP Select 集群设置期间自动创建。其余 LIF 可在部署后创建。

管理和数据 **LIF** （ **e0a** ， **e0b** 和 **e0c** ）

ONTAP 端口 e0a ， e0b 和 e0c 作为传输以下类型流量的 LIF 的候选端口进行委派：

- SAN/NAS 协议流量（ CIFS ， NFS 和 iSCSI ）
- 集群，节点和 SVM 管理流量
- 集群间流量（ SnapMirror 和 SnapVault ）

多节点网络配置

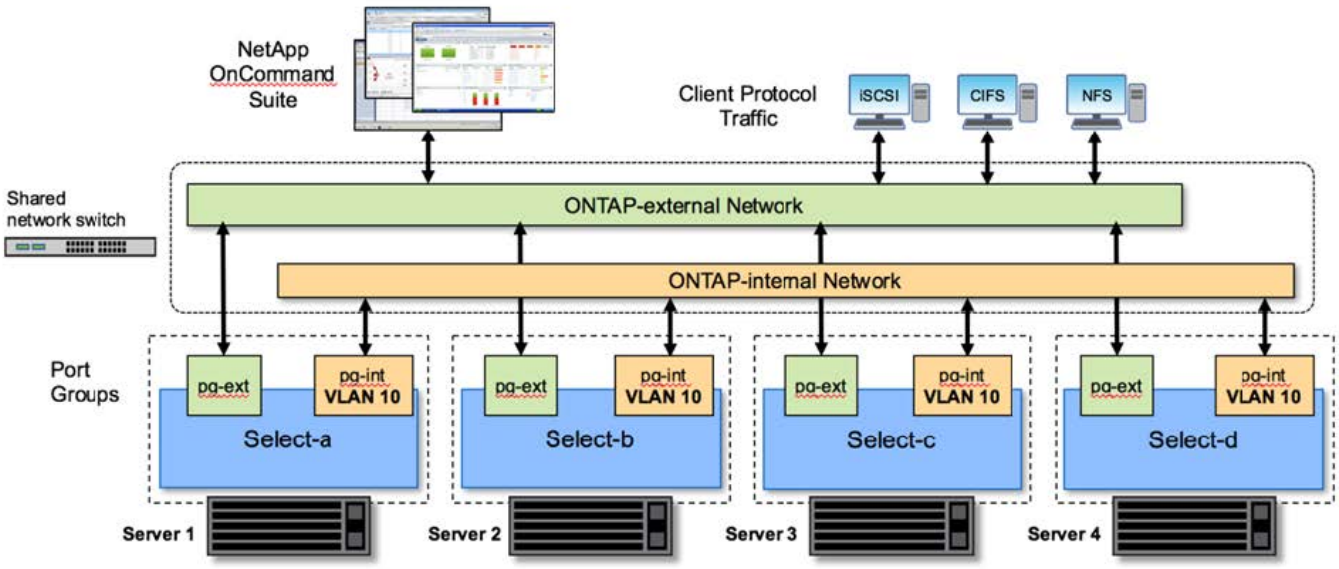
多节点 ONTAP Select 网络配置由两个网络组成。

这些网络是一个内部网络，负责提供集群和内部复制服务，并是一个外部网络，负责提供数据访问和管理服务。对这两个网络中的流量进行端到端隔离对于构建适合集群故障恢复能力的环境来说极为重要。

下图显示了这些网络，其中显示了在 VMware vSphere 平台上运行的四节点 ONTAP Select 集群。六节点和八节点集群具有类似的网络布局。

> 每个 ONTAP Select 实例都驻留在一个单独的物理服务器上。内部和外部流量使用单独的网络端口组进行隔离，这些端口组分配给每个虚拟网络接口，并允许集群节点共享同一个物理交换机基础架构。

- ONTAP Select 多节点集群网络配置概述 *




每个 ONTAP Select VM 都包含七个虚拟网络适配器，这些适配器作为一组七个网络端口（ e0a 到 e0g ）提供给 ONTAP 。虽然 ONTAP 将这些适配器视为物理 NIC ，但它们实际上是虚拟的，并通过虚拟化网络层映射到一组物理接口。因此，每个托管服务器不需要六个物理网络端口。

> 不支持向 ONTAP Select VM 添加虚拟网络适配器。

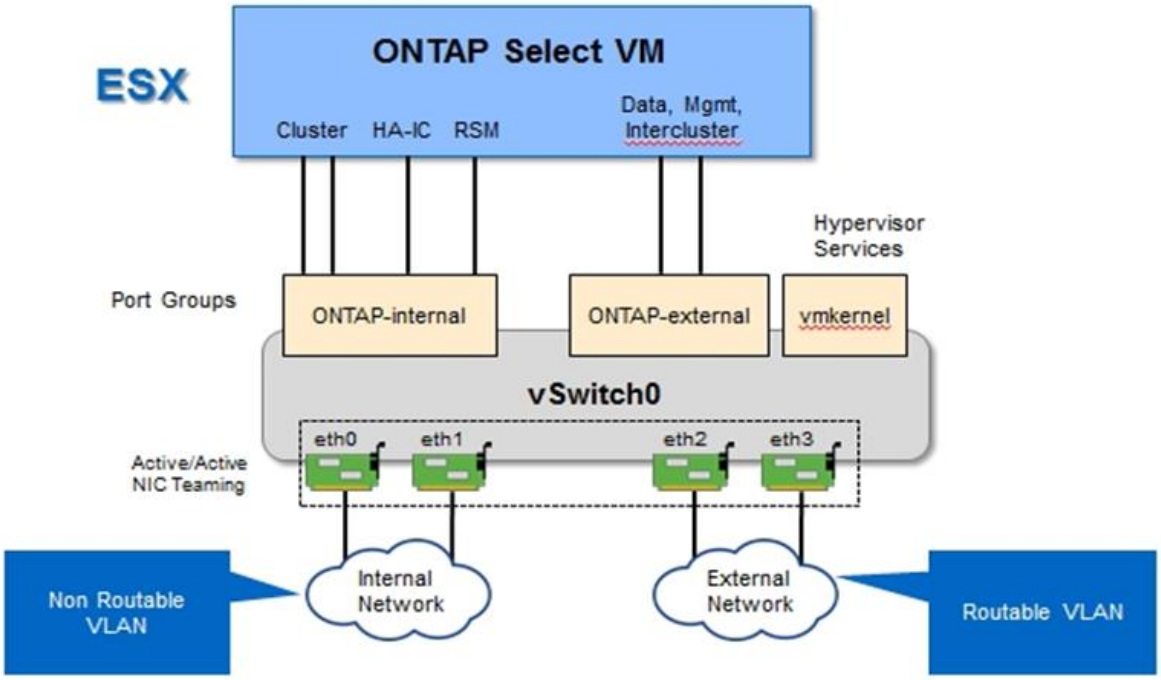
这些端口经过预配置，可提供以下服务：

- e0a ， e0b 和 e0g 。管理和数据 LIF
- e0c ， e0d 。集群网络 LIF
- e0e 。RSM
- e0f。HA interconnect

端口 e0a ， e0b 和 e0g 位于外部网络上。虽然端口 e0c 到 e0f 执行多种不同的功能，但它们共同构成内部 Select 网络。在制定网络设计决策时，应将这些端口放置在一个第 2 层网络上。无需在不同网络之间分隔这些虚拟适配器。

下图显示了这些端口与底层物理适配器之间的关系，其中显示了 ESX 虚拟机管理程序上的一个 ONTAP Select 集群节点。

- 多节点 ONTAP Select 集群中单个节点的网络配置 *




在不同物理 NIC 之间隔离内部和外部流量可防止因对网络资源的访问不足而导致系统出现延迟。此外，通过 NIC 绑定进行聚合可确保单个网络适配器出现故障不会阻止 ONTAP Select 集群节点访问相应的网络。

请注意，外部网络端口组和内部网络端口组均以对称方式包含所有四个 NIC 适配器。外部网络端口组中的活动端口是内部网络中的备用端口。相反，内部网络端口组中的活动端口是外部网络端口组中的备用端口。

#### LIF 分配

随着 IP 空间的推出， ONTAP 端口角色已弃用。与 FAS 阵列一样， ONTAP Select 集群也包含默认 IP 空间和集群 IP 空间。通过将网络端口 e0a ， e0b 和 e0g 置于默认 IP 空间中，将端口 e0c 和 e0d 置于集群 IP 空间中，这些端口实际上已与托管不属于的 LIF 隔离。ONTAP Select 集群中的其余端口将通过自动分配提供内部服务的接口来使用。它们不会像 RSM 和 HA 互连接口那样通过 ONTAP shell 公开。

> 并非所有 LIF 都可通过 ONTAP 命令 Shell 查看。HA 互连和 RSM 接口在 ONTAP 中隐藏，并在内部用于提供各自的服务。

以下各节将详细介绍网络端口和 LIF 。

##### 管理和数据生命周期(e0a、e0b和e0g)

ONTAP 端口 e0a ， e0b 和 e0g 会委派为传输以下类型流量的 LIF 的候选端口：

- SAN/NAS 协议流量（ CIFS ， NFS 和 iSCSI ）
- 集群，节点和 SVM 管理流量

- 集群间流量（ SnapMirror 和 SnapVault ）

> 集群和节点管理 LIF 会在 ONTAP Select 集群设置期间自动创建。其余 LIF 可在部署后创建。

集群网络 **LIF** （ **e0c** ， **e0d** ）

ONTAP 端口 e0c 和 e0d 已委派为集群接口的主端口。在每个 ONTAP Select 集群节点中， ONTAP 设置期间会使用链路本地 IP 地址（ 169.254.x.x ）自动生成两个集群接口。

> 不能为这些接口分配静态 IP 地址，也不应创建其他集群接口。

集群网络流量必须流经低延迟的非路由第 2 层网络。由于集群吞吐量和延迟要求， ONTAP Select 集群的物理位置应接近（例如多件包，单个数据中心）。不支持通过在 WAN 或远距离的地理位置之间分隔 HA 节点来构建四节点，六节点或八节点延伸型集群配置。支持使用调解器的延伸型双节点配置。

有关详细信息，请参见一节 "双节点延伸型 HA （ MetroCluster SDS ）最佳实践"。

> 为了确保集群网络流量的最大吞吐量，此网络端口配置为使用巨型帧（ 7500 到 9000 MTU ）。要使集群正常运行，请验证是否已在向 ONTAP Select 集群节点提供内部网络服务的所有上游虚拟和物理交换机上启用巨型帧。

**RAID SyncMirror** 流量（ **e0e** ）

使用网络端口 e0e 上的内部网络接口在 HA 配对节点之间同步复制块。此功能会使用集群设置期间由 ONTAP 配置的网络接口自动执行，不需要管理员进行任何配置。

> 端口 e0e 由 ONTAP 预留用于内部复制流量。因此，端口和托管 LIF 在 ONTAP 命令行界面或 System Manager 中均不可见。此接口已配置为使用自动生成的链路本地 IP 地址，不支持重新分配备用 IP 地址。此网络端口需要使用巨型帧（ 7500 到 9000 MTU ）。

**HA** 互连（ **e0f** ）

NetApp FAS 阵列使用专用硬件在 ONTAP 集群中的 HA 对之间传递信息。但是，软件定义的环境往往没有这种类型的设备可用（例如 InfiniBand 或 iWARP 设备），因此需要使用备用解决方案 。尽管考虑了多种可能性，但对互连传输提出的 ONTAP 要求要求要求在软件中模拟此功能。因此，在 ONTAP Select 集群中， HA 互连的功能（传统上由硬件提供）已通过以太网作为传输机制设计到操作系统中。

每个 ONTAP Select 节点都配置有一个 HA 互连端口 e0f 。此端口托管 HA 互连网络接口，该接口负责两项主要功能：

- 在 HA 对之间镜像 NVRAM 的内容
- 在 HA 对之间发送 / 接收 HA 状态信息和网络检测信号消息

HA 互连流量通过在以太网数据包中对远程直接内存访问（ RDMA ）帧进行分层来使用单个网络接口通过此网络端口进行传输。

> 以类似于 RSM 端口（ e0e ）的方式，用户既不能通过 ONTAP 命令行界面也不能通过 System Manager 看到物理端口和托管网络接口。因此，无法修改此接口的 IP 地址，也无法更改端口的状态。此网络端口需要使用巨型帧（ 7500 到 9000 MTU ）。

## ONTAP Select 内部和外部网络

ONTAP Select 内部和外部网络的特征。

### ONTAP Select 内部网络

内部 ONTAP Select 网络仅存在于产品的多节点变体中，负责为 ONTAP Select 集群提供集群通信， HA 互连和同步复制服务。此网络包括以下端口和接口：

- * e0c ， e0d.* 托管集群网络 LIF
- 托管 RSM LIF 的 * e0e.*
- 托管 HA 互连 LIF 的 * e0f.*

此网络的吞吐量和延迟对于确定 ONTAP Select 集群的性能和故障恢复能力至关重要。为了确保集群安全并确保系统接口与其他网络流量分开，需要进行网络隔离。因此，此网络必须由 ONTAP Select 集群独占使用。

> 不支持对 Select 集群流量以外的流量使用 Select 内部网络，例如应用程序或管理流量。ONTAP 内部 VLAN 上不能存在其他 VM 或主机。

遍历内部网络的网络数据包必须位于一个专用的 VLAN 标记第 2 层网络上。可通过完成以下任务之一来完成此操作：

- 将带有 VLAN 标记的端口组分配给内部虚拟 NIC （ e0c 到 e0f ）（ VST 模式）
- 使用上游交换机提供的原生 VLAN ，其中原生 VLAN 不用于任何其他流量（分配一个没有 VLAN ID 的端口组，即 EST 模式）

在所有情况下，内部网络流量的 VLAN 标记都是在 ONTAP Select VM 之外进行的。

> 仅支持 ESX 标准和分布式 vSwitch 。不支持其他虚拟交换机或 ESX 主机之间的直接连接。内部网络必须完全打开；不支持 NAT 或防火墙。

在 ONTAP Select 集群中，内部流量和外部流量使用称为端口组的虚拟第 2 层网络对象进行分隔。正确分配这些端口组的 vSwitch 非常重要，尤其是对于负责提供集群， HA 互连和镜像复制服务的内部网络而言。如果这些网络端口的网络带宽不足，则发生原因 性能可能会下降，甚至会影响集群节点的稳定性。因此，四节点，六节点和八节点集群要求内部 ONTAP Select 网络使用 10 Gb 连接；不支持 1 Gb NIC 。但是，可以对外部网络进行权衡，因为限制传入 ONTAP Select 集群的数据流不会影响其可靠运行的能力。

双节点集群可以使用四个 1 Gb 端口传输内部流量，也可以使用一个 10 Gb 端口，而不是四节点集群所需的两个 10 Gb 端口。如果环境中的条件使服务器无法安装四个 10 Gb NIC 卡，则可将两个 10 Gb NIC 卡用于内部网络，并将两个 1 Gb NIC 用于外部 ONTAP 网络。

#### 内部网络验证和故障排除

可以使用网络连接检查程序功能验证多节点集群中的内部网络。可以从运行的Deploy命令行界面调用此功能 `network connectivity-check start` 命令：

运行以下命令以查看测试的输出：

```
network connectivity-check show --run-id X (X is a number)
```

此工具仅适用于对多节点 Select 集群中的内部网络进行故障排除。不应使用此工具对单节点集群（包括 vNAS 配置）， ONTAP Deploy 到 ONTAP Select 连接或客户端连接问题进行故障排除。

集群创建向导（ ONTAP Deploy GUI 的一部分）包含内部网络检查程序，作为创建多节点集群期间可用的可选步骤。鉴于内部网络在多节点集群中发挥的重要作用，将此步骤加入集群创建工作流可提高集群创建操作的成功率。

从 ONTAP Deploy 2.10 开始，内部网络使用的 MTU 大小可以设置为 7， 500 到 9， 000 之间。此外，还可以使用网络连接检查程序测试介于 7， 500 和 9， 000 之间的 MTU 大小。默认 MTU 值设置为虚拟网络交换机的值。如果环境中存在 VXLAN 等网络覆盖，则必须将此默认值替换为较小的值。

#### ONTAP Select 外部网络

ONTAP Select 外部网络负责集群的所有出站通信，因此，无论是单节点配置还是多节点配置都存在。尽管此网络没有对内部网络严格定义的吞吐量要求，但管理员应注意不要在客户端和 ONTAP VM 之间创建网络瓶颈，因为性能问题可能会被错误地描述为 ONTAP Select 问题。

> 可以采用与内部流量类似的方式在 vSwitch 层（ VST ）和外部交换机层（ EST ）标记外部流量。此外， ONTAP Select VM 本身也可以在一个称为 VGT 的过程中对外部流量进行标记。请参见一节 "数据和管理流量隔离" 了解更多详细信息。

下表重点介绍了 ONTAP Select 内部网络与外部网络之间的主要区别。

- 内部网络与外部网络快速参考 *

| Description | 内部网络 | 外部网络 |
|---|---|---|
| 网络服务 | 集群<br>HA/IC<br>RAID SyncMirror (RSM) | 数据管理<br>集群间<br>(SnapMirror和SnapVault) |
| 网络隔离 | Required | 可选 |
| 帧大小（ MTU ） | 7， 500 到 9， 000 | 1、500 (默认)<br>9、000 (支持) |
| IP 地址分配 | 已自动生成 | 用户定义的 |
| DHCP支持 | 否 | 否 |

##### NIC 绑定

为了确保内部和外部网络具有提供高性能和容错能力所需的带宽和故障恢复能力特性，建议使用物理网络适配器绑定。支持使用单个 10 Gb 链路的双节点集群配置。但是， NetApp 建议的最佳实践是在 ONTAP Select 集群的内部和外部网络上使用 NIC 绑定。

##### MAC 地址生成

分配给所有 ONTAP Select 网络端口的 MAC 地址由随附的部署实用程序自动生成。该实用程序使用 NetApp 专用于平台的组织唯一标识符（ Organizationally Unique Identifier， OUI ），以确保与 FAS 系统不存在冲突。然后，此地址的副本将存储在 ONTAP Select 安装虚拟机（ ONTAP Deploy ）的内部数据库中，以防止在将来的

节点部署期间意外重新分配。管理员不应修改为网络端口分配的 MAC 地址。

## 支持的网络配置

### 选择最佳硬件并配置网络以优化性能和故障恢复能力。

服务器供应商深知客户有不同的需求和选择至关重要。因此，在购买物理服务器时，在做出网络连接决策时，可以选择多种方式。大多数商用系统都提供各种 NIC 选项，可提供单端口和多端口选项，其速度和吞吐量各不相同。这包括在VMware ESX中支持25 Gb/秒和40 Gb/秒NIC适配器。

由于 ONTAP Select VM 的性能与底层硬件的特性直接相关，因此，通过选择速度更高的 NIC 来增加 VM 的吞吐量可提高集群性能并改善整体用户体验。可以使用四个 10 Gb NIC 或两个高速 NIC （ 25/40 Gb/ 秒）来实现高性能网络布局。此外，还支持许多其他配置。对于双节点集群，支持 4 个 1 Gb 端口或 1 个 10 Gb 端口。对于单节点集群，支持 2 个 1 Gb 端口。

#### 网络最低配置和建议配置

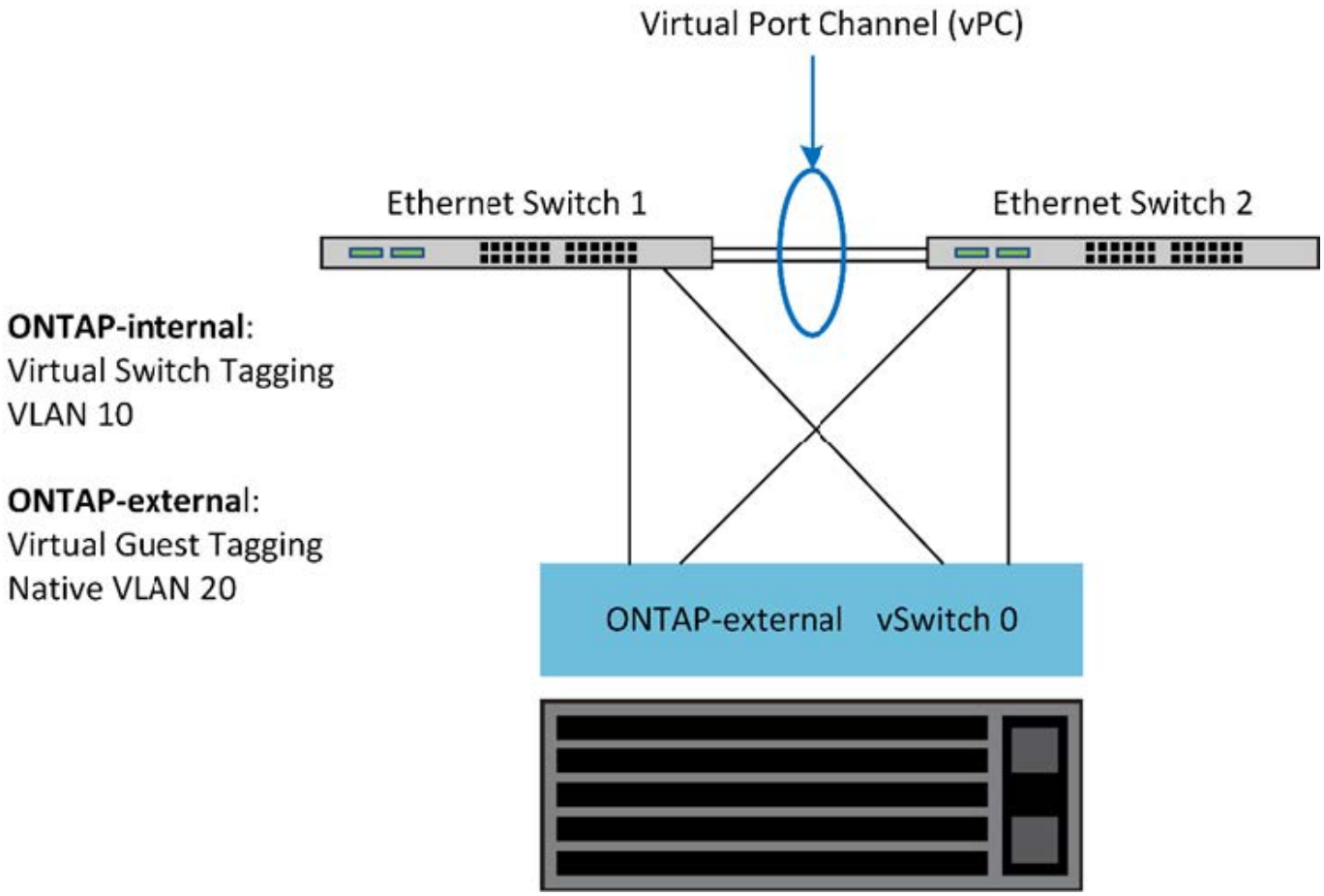
根据集群大小、可以使用多种受支持的以太网配置。

| 集群大小 | 最低要求 | 建议 |
| --- | --- | --- |
| 单节点集群 | 2个1GbE | 2个10GbE |
| 双节点集群或MetroCluster SDS | 4个1GbE或1个10GbE | 2个10GbE |
| 4/6/8节点集群 | 2个10GbE | 4个10GbE或2个25/40GbE |

> 不支持在正在运行的集群中的单链路拓扑和多链路拓扑之间进行转换、因为可能需要在每个拓扑所需的不同NIC绑定配置之间进行转换。

#### 使用多个物理交换机进行网络配置

如果有足够的硬件可用、NetApp建议使用下图所示的多交换机配置、因为这样可以增强保护、防止物理交换机出现故障。




## VMware vSphere vSwitch配置

### 双 NIC 和四 NIC 配置的 ONTAP Select vSwitch 配置和负载平衡策略。

ONTAP Select 支持使用标准 vSwitch 配置和分布式 vSwitch 配置。分布式 vSwitch 支持链路聚合构造（ LACP ）。链路聚合是一种常见的网络构造，用于在多个物理适配器之间聚合带宽。LACP 是一种与供应商无关的标准，可为网络端点提供开放式协议，将物理网络端口组捆绑到一个逻辑通道中。ONTAP Select 可以与配置为链路聚合组（ LAG ）的端口组配合使用。但是， NetApp 建议使用各个物理端口作为简单上行链路（中继）端口，以避免使用 LAG 配置。在这些情况下，标准和分布式 vSwitch 的最佳实践是相同的。

本节介绍双 NIC 和四 NIC 配置中应使用的 vSwitch 配置和负载平衡策略。

在配置 ONTAP Select 要使用的端口组时，应遵循以下最佳实践；端口组级别的负载平衡策略是基于源虚拟端口 ID 的路由。VMware 建议在连接到 ESXi 主机的交换机端口上将 STP 设置为 PortFast 。

所有 vSwitch 配置都要求至少将两个物理网络适配器捆绑到一个 NIC 组中。ONTAP Select 支持双节点集群使用一个 10 Gb 链路。但是， NetApp 的最佳实践是通过 NIC 聚合确保硬件冗余。

在 vSphere 服务器上， NIC 组是一种聚合构造，用于将多个物理网络适配器捆绑到一个逻辑通道中，从而可以在所有成员端口之间共享网络负载。请务必记住，在没有物理交换机支持的情况下，可以创建 NIC 组。负载平衡和故障转移策略可以直接应用于 NIC 组，而 NIC 组不知道上游交换机配置。在这种情况下，策略仅应用于出站流量。

> ONTAP Select 不支持静态端口通道。分布式 vSwitch 支持启用了 LACP 的通道，但使用 LACP LAG 可能会导致 LAG 成员之间的负载分布不均匀。

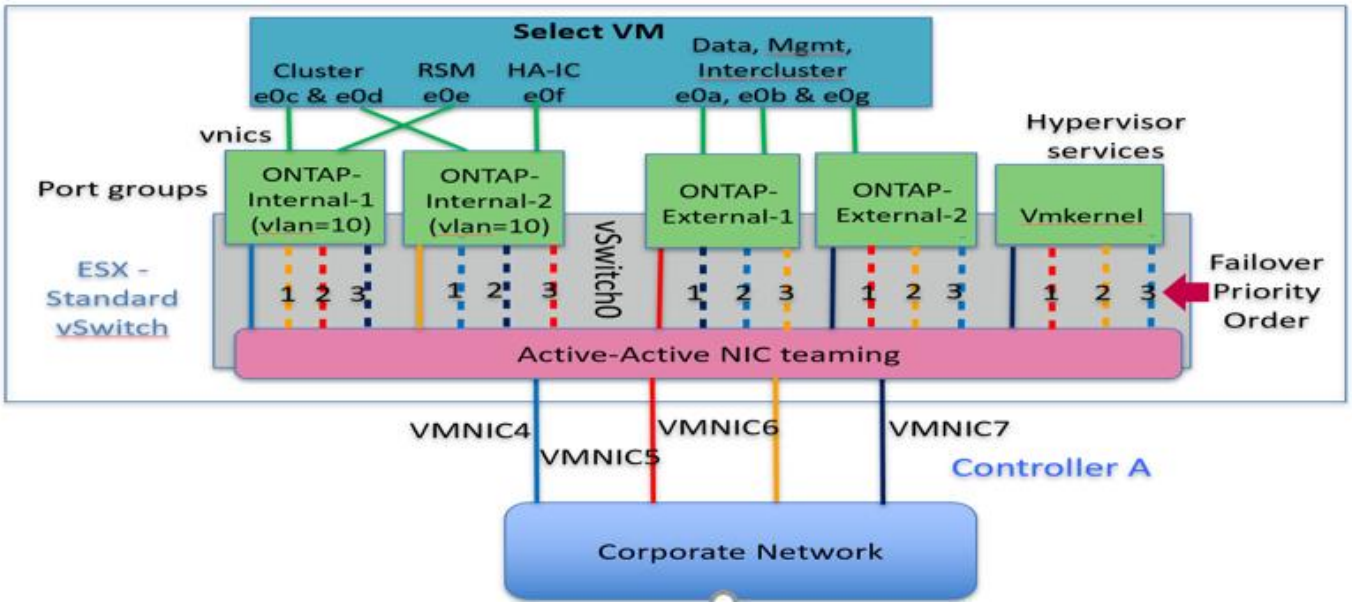
对于单节点集群， ONTAP Deploy 会将 ONTAP Select VM 配置为对外部网络使用端口组，并对集群和节点管理流量使用相同的端口组或（可选）不同的端口组。对于单节点集群，可以将所需数量的物理端口作为活动适配器添加到外部端口组中。

对于多节点集群， ONTAP Deploy 会将每个 ONTAP Select VM 配置为对内部网络使用一个或两个端口组，而对外部网络单独使用一个或两个端口组。集群和节点管理流量可以使用与外部流量相同的端口组，也可以使用单独的端口组。集群和节点管理流量不能与内部流量共享同一端口组。

标准或分布式 **vSwitch** 以及每个节点四个物理端口

可以为多节点集群中的每个节点分配四个端口组。每个端口组都有一个活动物理端口和三个备用物理端口，如下图所示。

每个节点具有四个物理端口的 * vSwitch *



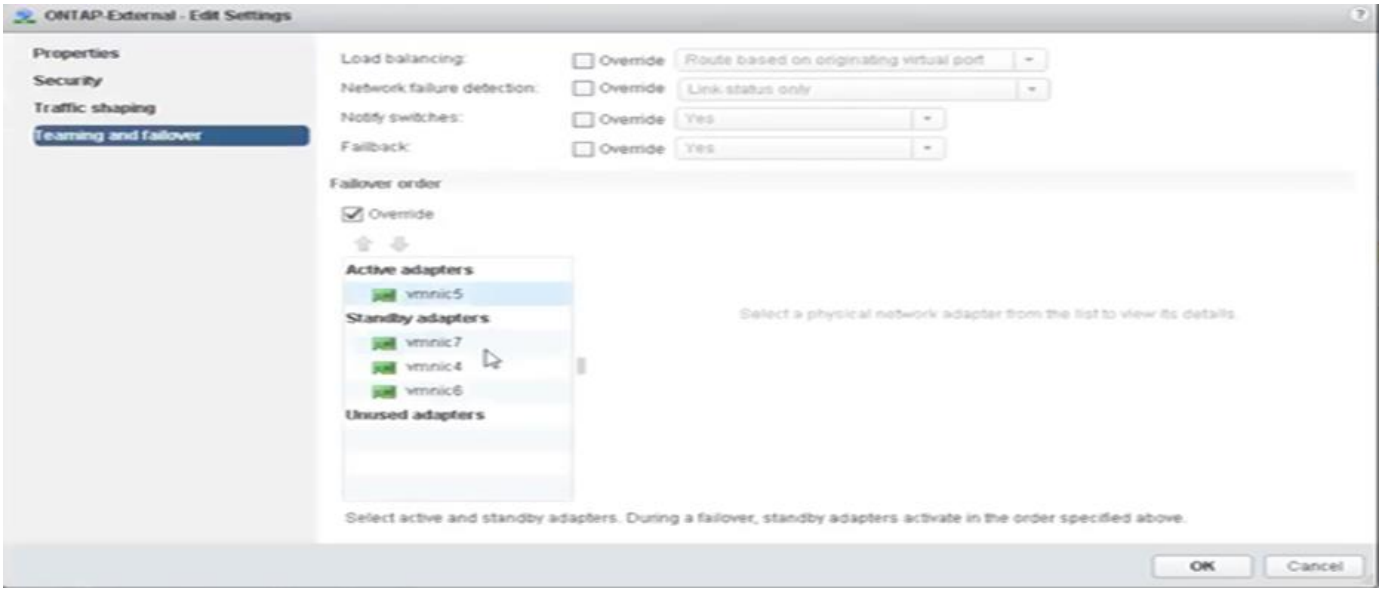

端口在备用列表中的顺序非常重要。下表提供了四个端口组之间的物理端口分布示例。

- 网络最低配置和建议配置 *

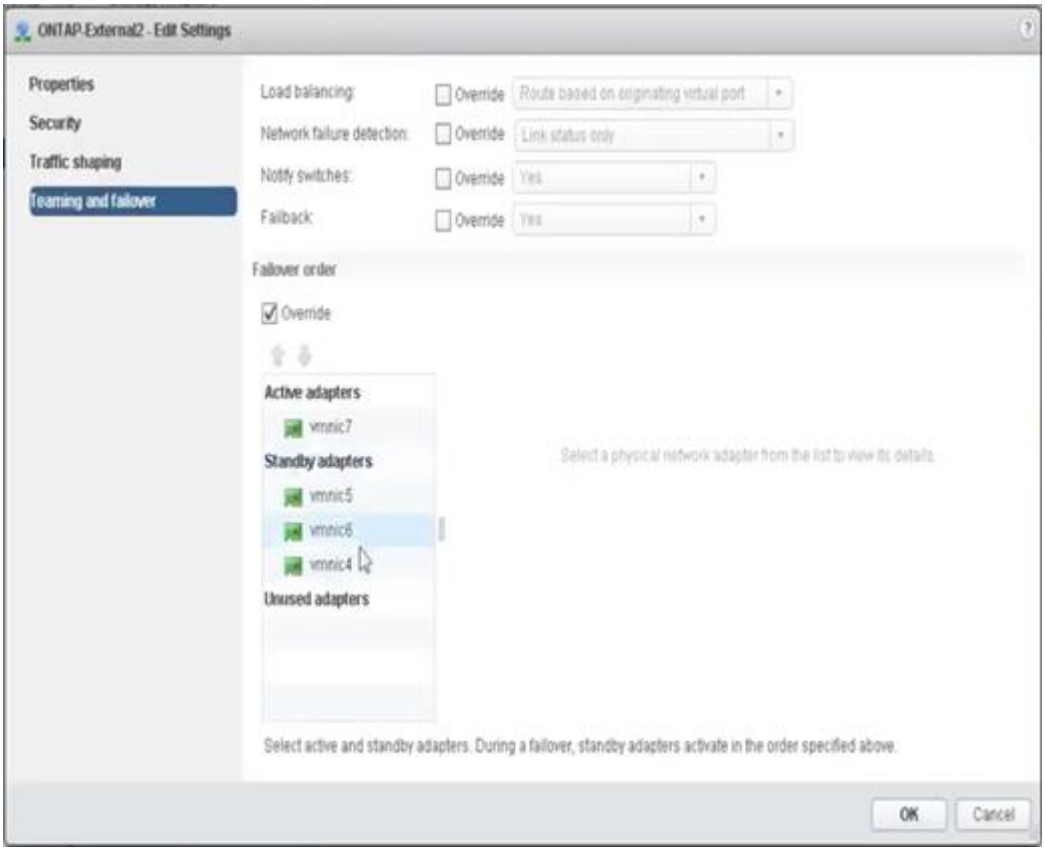
| 端口组 | 外部 **1.** | 外部**2.** | 内部 **1.** | 内部**2.** |
|---|---|---|---|---|
| 活动 | vmnic0 | vmnic1. | vmnic2. | vmnic3. |
| 备用 1 | vmnic1. | vmnic0 | vmnic3. | vmnic2. |
| 备用 2. | vmnic2. | vmnic3. | vmnic0 | vmnic1. |
| 待机3. | vmnic3. | vmnic2. | vmnic1. | vmnic0 |

下图显示了 vCenter GUI （ ONTAP 外部和 ONTAP 外部端口 2 ）中外部网络端口组的配置。请注意，活动适配器来自不同的网卡。在此设置中， vmnic 4 和 vmnic 5 是同一物理 NIC 上的双端口，而 vmnic 6 和 vminc 7 是同一个 NIC 上的类似双端口（本示例不使用 vnmic 0 到 3 ）。备用适配器的顺序提供了一个分层故障转移，内部网络中的端口也是最后一个。备用列表中的内部端口顺序在两个外部端口组之间进行类似的交换。

- 第 1 部分： ONTAP Select 外部端口组配置 *




第2部分：**ONTAP Select**外部端口组配置



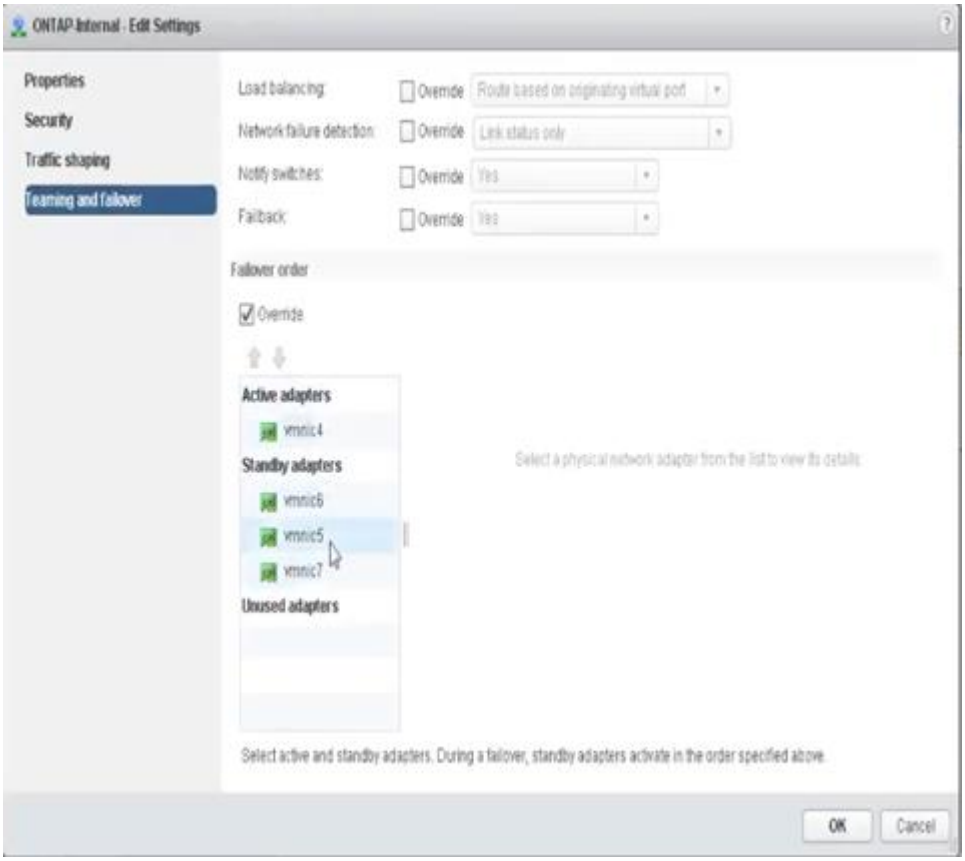

为便于阅读，分配如下：

| **ONTAP** 外部 | **ontap-External2.** |
|---|---|
| 活动适配器：vmnic5.<br>备用适配器：vmnic7、vmnic4、vmnic6 | 活动适配器：vmnic7.<br>备用适配器：vmnic5、vmnic6、vmnic4 |

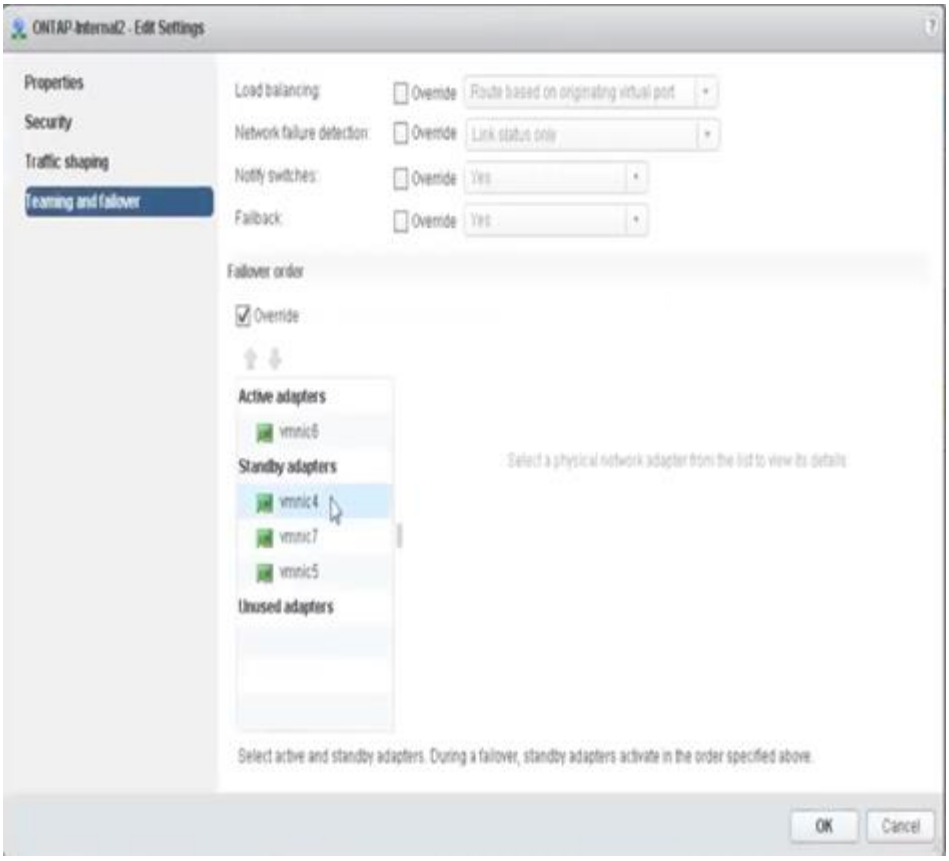
下图显示了内部网络端口组（ONTAP 内部和 ONTAP 内部 2）的配置。请注意，活动适配器来自不同的网卡。

在此设置中， vmnic 4 和 vmnic 5 是同一物理 ASIC 上的双端口，而 vmnic 6 和 vmnic 7 则是同一个 ASIC 上的类似双端口。备用适配器的顺序提供了一个分层故障转移，外部网络中的端口也是最后一个。备用列表中外部端口的顺序在两个内部端口组之间进行类似的交换。

- 第 1 部分： ONTAP Select 内部端口组配置 *




- 第 2 部分： ONTAP Select 内部端口组 *




为便于阅读，分配如下：

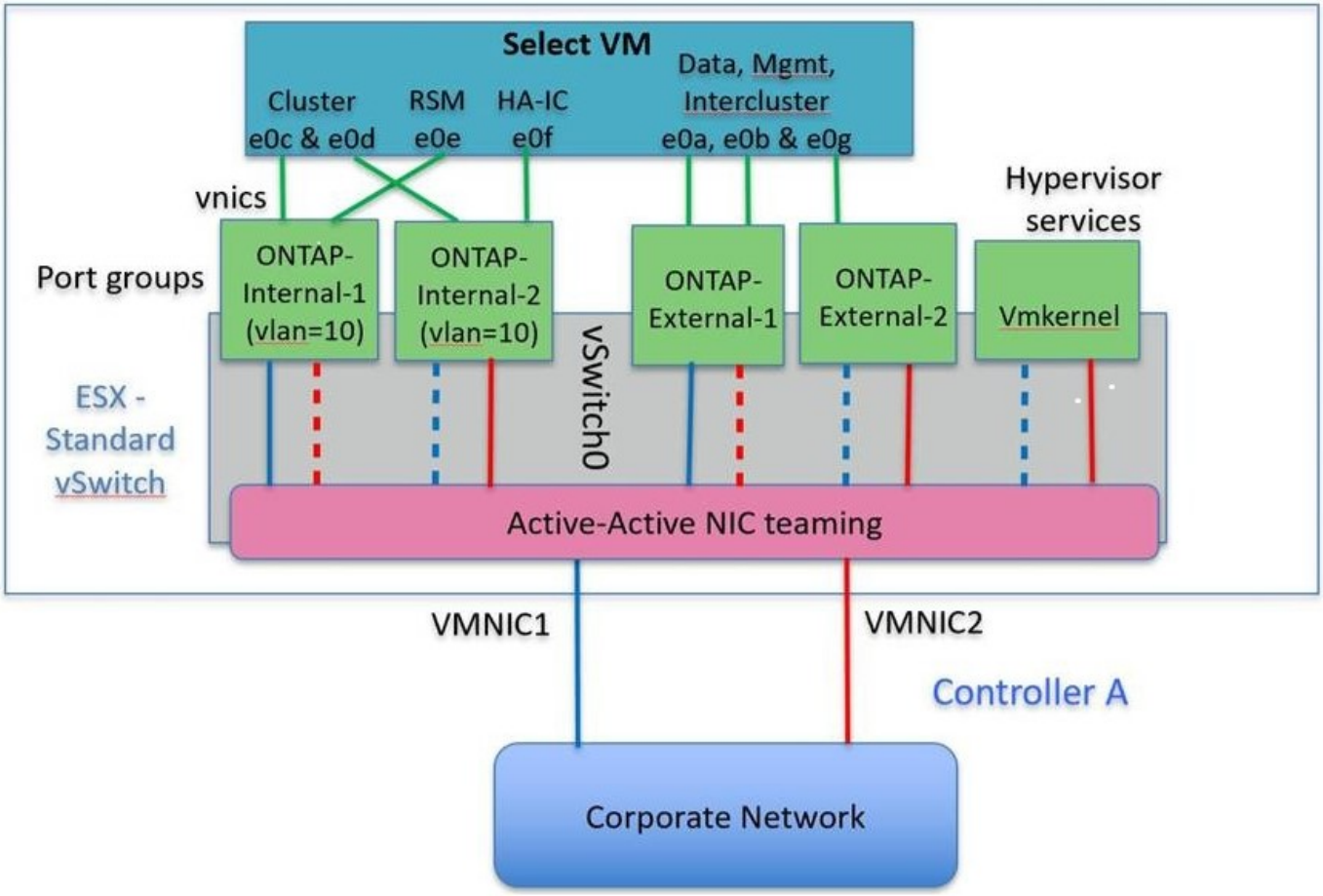
| **ONTAP** 内部 | **ONTAP** 内部 **2.** |
|---|---|
| 活动适配器：vmnic4.<br>备用适配器：vmnic6、vmnic5、vmnic7 | 活动适配器：vmnic6<br>备用适配器：vmnic4、vmnic7、vmnic5 |

标准或分布式 **vSwitch** 以及每个节点两个物理端口

使用两个高速（25/40 Gb） NIC 时，建议的端口组配置在概念上与使用四个 10 Gb 适配器的配置非常相似。即使仅使用两个物理适配器，也应使用四个端口组。端口组分配如下：

| 端口组 | 外部 **1** （**e0a**， **e0b**） | 内部 **1** （**e0c**， **e0e**） | 内部**2 (e0d**、**e0f)** | 外部 **2** （**e0g**） |
|---|---|---|---|---|
| 活动 | vmnic0 | vmnic0 | vmnic1. | vmnic1. |
| 备用 | vmnic1. | vmnic1. | vmnic0 | vmnic0 |

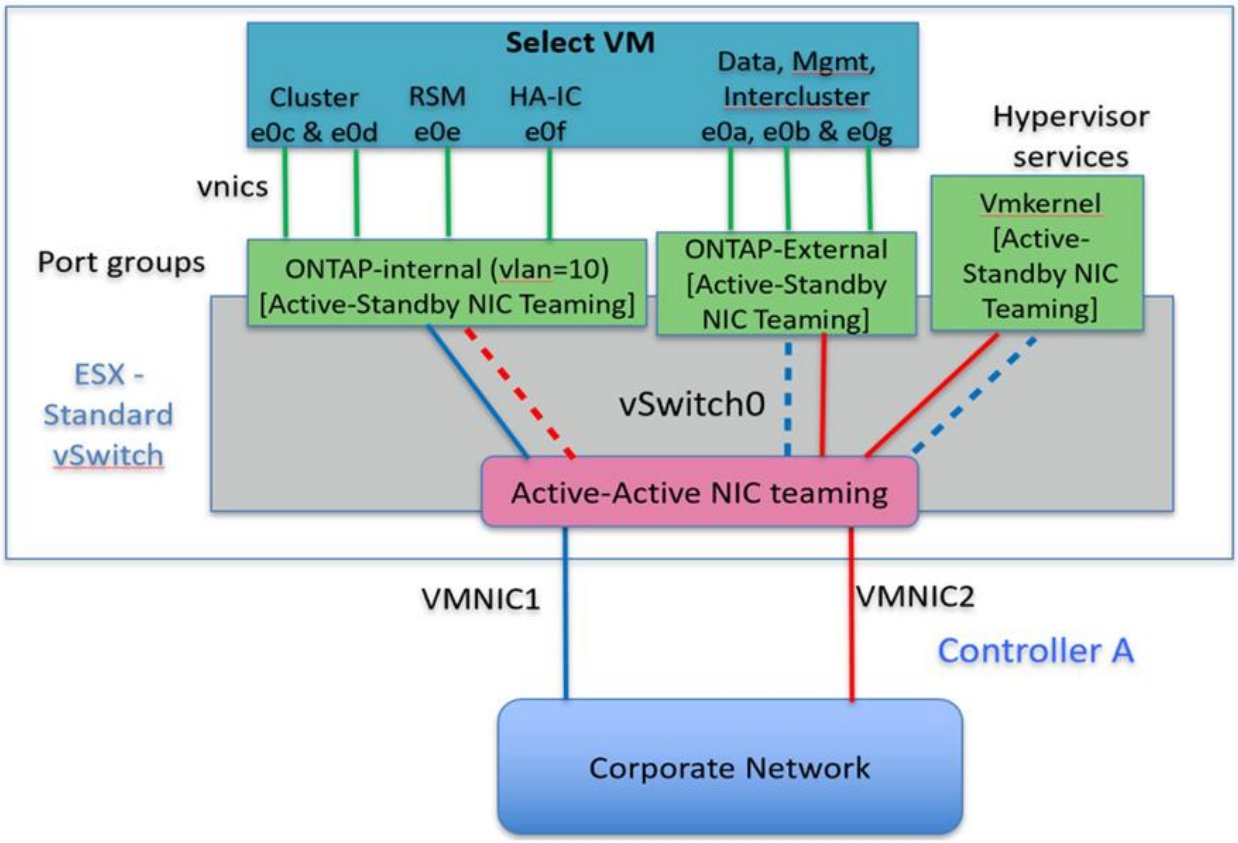
- 每个节点具有两个高速（25/40 Gb）物理端口的 vSwitch *




使用两个物理端口（ 10 Gb 或更少）时，每个端口组应配置一个活动适配器和一个备用适配器，使其彼此相对。内部网络仅适用于多节点 ONTAP Select 集群。对于单节点集群，可以将这两个适配器配置为外部端口组中的活动适配器。

以下示例显示了 vSwitch 的配置以及负责处理多节点 ONTAP Select 集群的内部和外部通信服务的两个端口组。如果网络发生中断，外部网络可以使用内部网络 vmnic ，因为内部网络 vmnic 属于此端口组并配置为备用模式。外部网络的情况正好相反。在两个端口组之间交替使用活动和备用 vmnic 对于在网络中断期间正确地对 ONTAP Select VM 进行故障转移至关重要。

每个节点具有两个物理端口（ 10 Gb 或更少）的 * vSwitch *

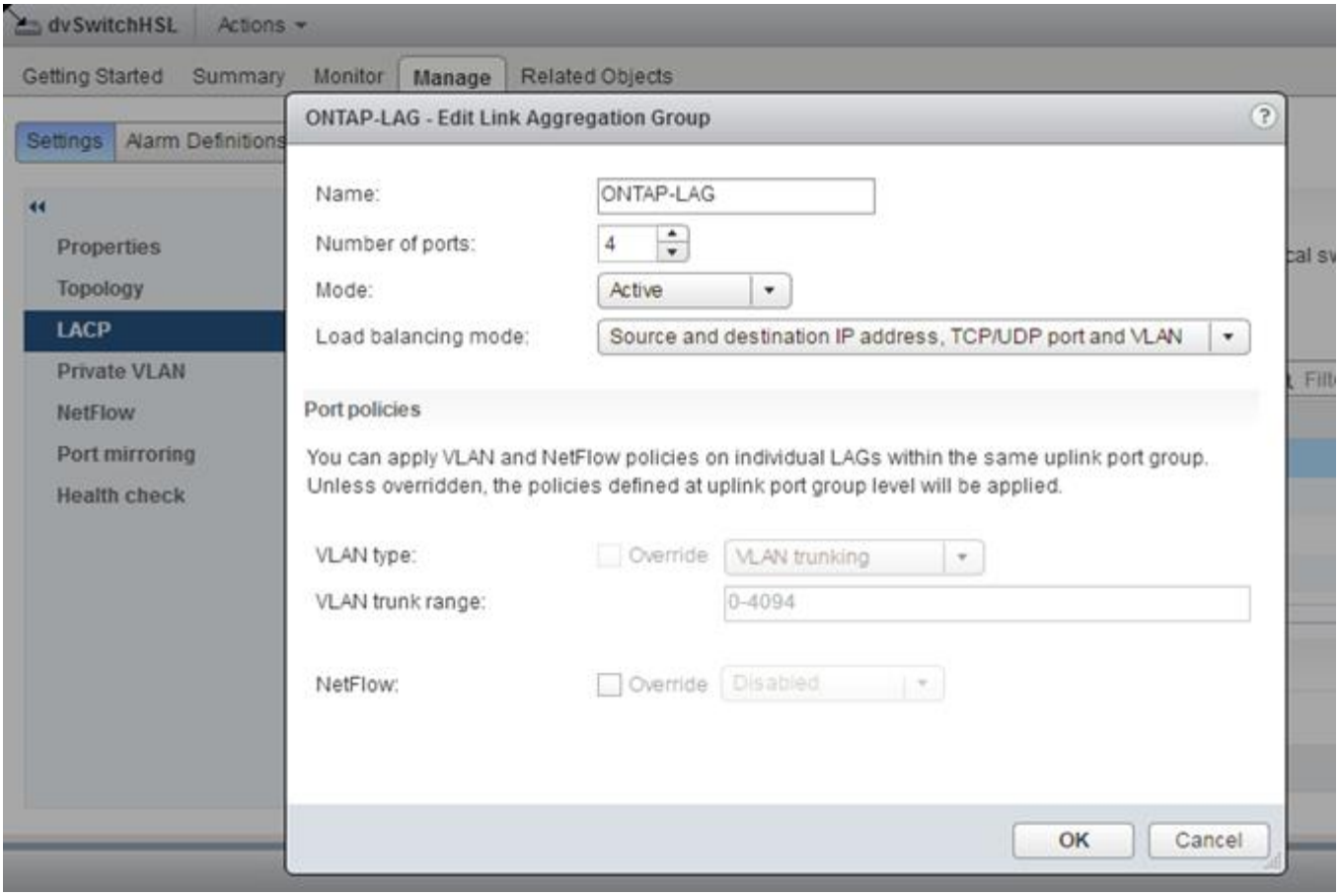



采用 **LACP** 的分布式 **vSwitch**

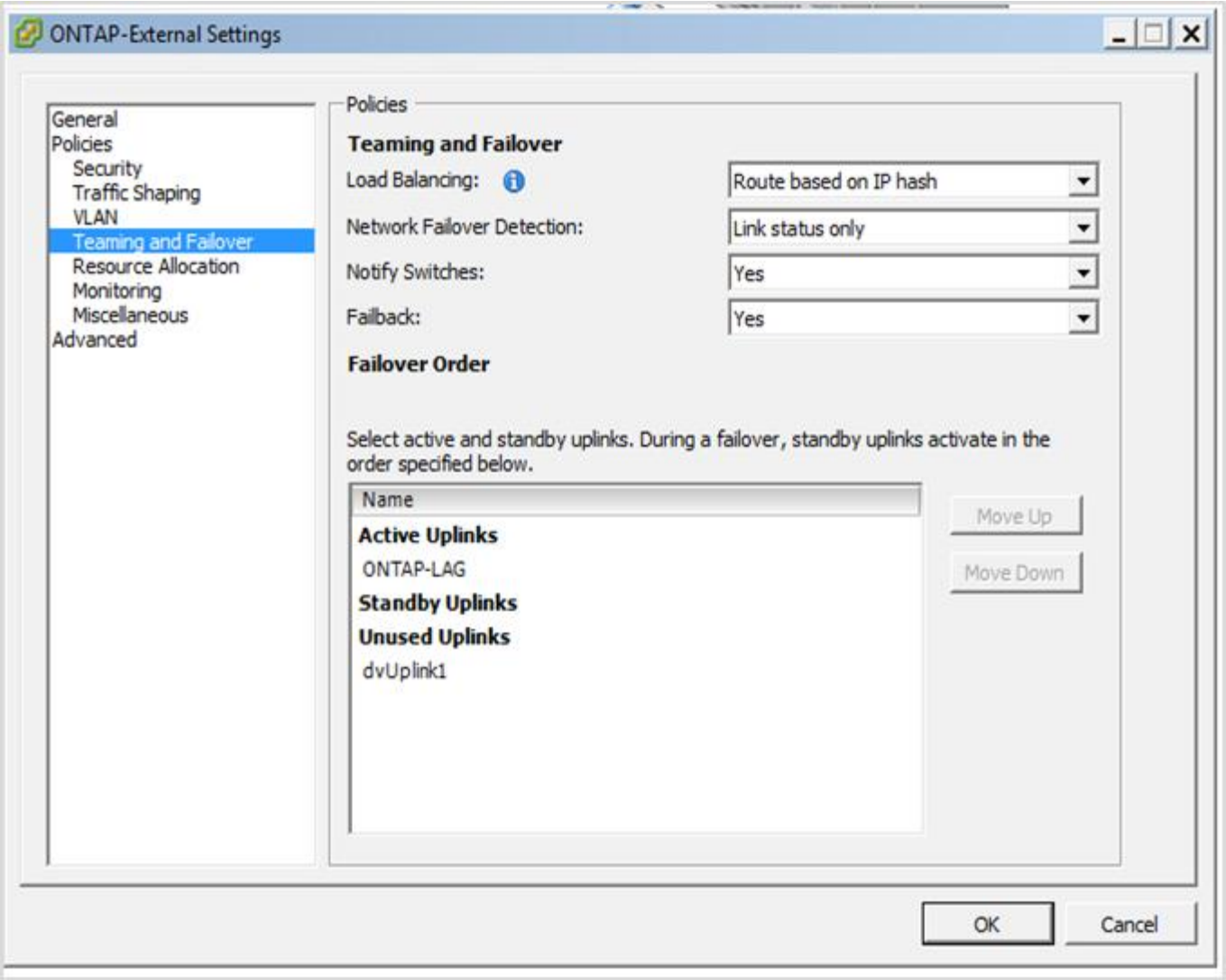
在配置中使用分布式 vSwitch 时，可以使用 LACP （尽管这不是最佳实践）来简化网络配置。唯一受支持的 LACP 配置要求所有 vmnic 都位于一个 LAG 中。上行链路物理交换机在通道中的所有端口上必须支持介于 7 ， 500 到 9 ， 000 之间的 MTU 大小。内部和外部 ONTAP Select 网络应在端口组级别隔离。内部网络应使用不可路由（隔离）的 VLAN 。外部网络可以使用 VST ， EST 或 VGT 。

以下示例显示了使用 LACP 的分布式 vSwitch 配置。

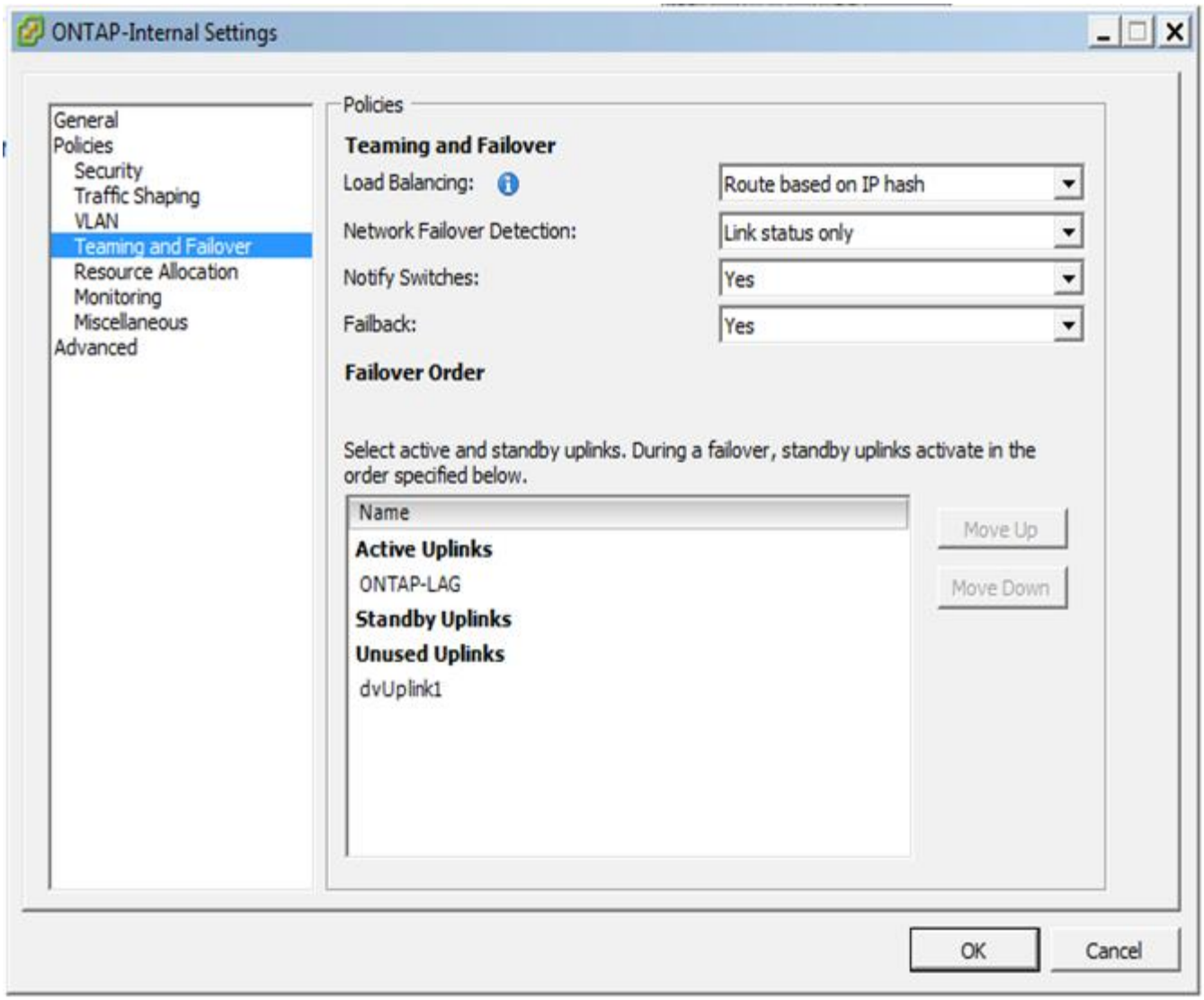
使用 LACP* 时的 * LAG 属性




- 使用已启用 LACP 的分布式 vSwitch 的外部端口组配置 *




- 使用启用了 LACP 的分布式 vSwitch 的内部端口组配置 *




LACP 要求您将上游交换机端口配置为端口通道。在分布式 vSwitch 上启用此功能之前，请确保已正确配置启用了 LACP 的端口通道。

## 物理交换机配置

### 基于单交换机和多交换机环境的上游物理交换机配置详细信息。

在决定从虚拟交换机层到物理交换机的连接时，应仔细考虑。内部集群流量与外部数据服务的隔离应通过第 2 层 VLAN 提供的隔离扩展到上游物理网络层。

物理交换机端口应配置为中继端口。ONTAP Select 外部流量可以通过以下两种方式之一在多个第 2 层网络之间进行分隔。一种方法是，将 ONTAP VLAN 标记的虚拟端口与一个端口组结合使用。另一种方法是，在 VST 模式下将单独的端口组分配给管理端口 e0a 。您还必须根据 ONTAP Select 版本以及单节点或多节点配置为 e0b 和 e0c/e0g 分配数据端口。如果外部流量在多个第 2 层网络之间隔离，则上行链路物理交换机端口应在其允许的 VLAN 列表中包含这些 VLAN 。

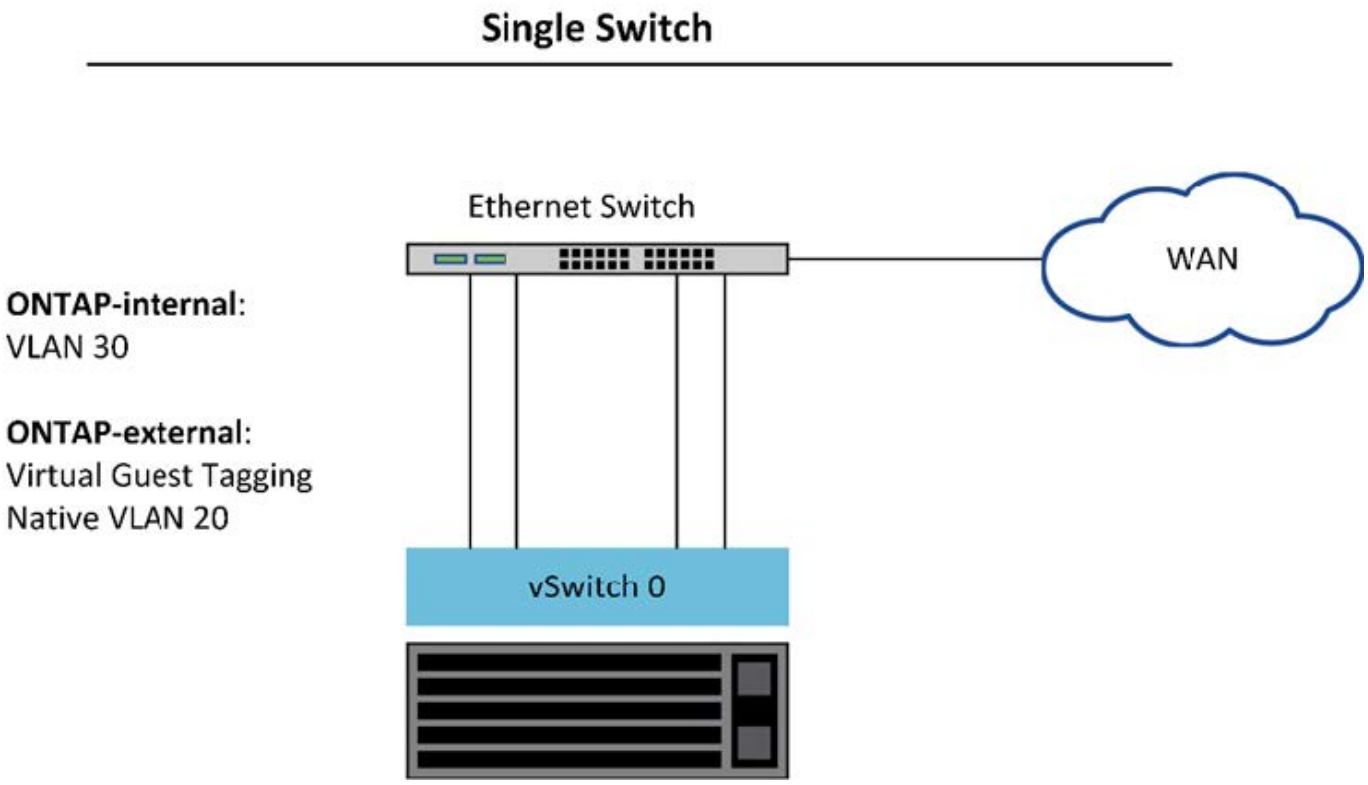
ONTAP Select 内部网络流量使用使用使用链路本地 IP 地址定义的虚拟接口进行。由于这些 IP 地址不可路由，因此集群节点之间的内部流量必须流经一个第 2 层网络。不支持 ONTAP Select 集群节点之间的路由跃点。

共享物理交换机

下图显示了多节点 ONTAP Select 集群中的一个节点可能使用的交换机配置。在此示例中，托管内部和外部网络端口组的 vSwitch 使用的物理 NIC 连接到同一个上游交换机。交换机流量使用不同 VLAN 中的广播域保持隔离。

> 对于 ONTAP Select 内部网络，在端口组级别进行标记。虽然以下示例对外部网络使用 VGT ，但该端口组同时支持 VGT 和 VST 。

- 使用共享物理交换机的网络配置 *

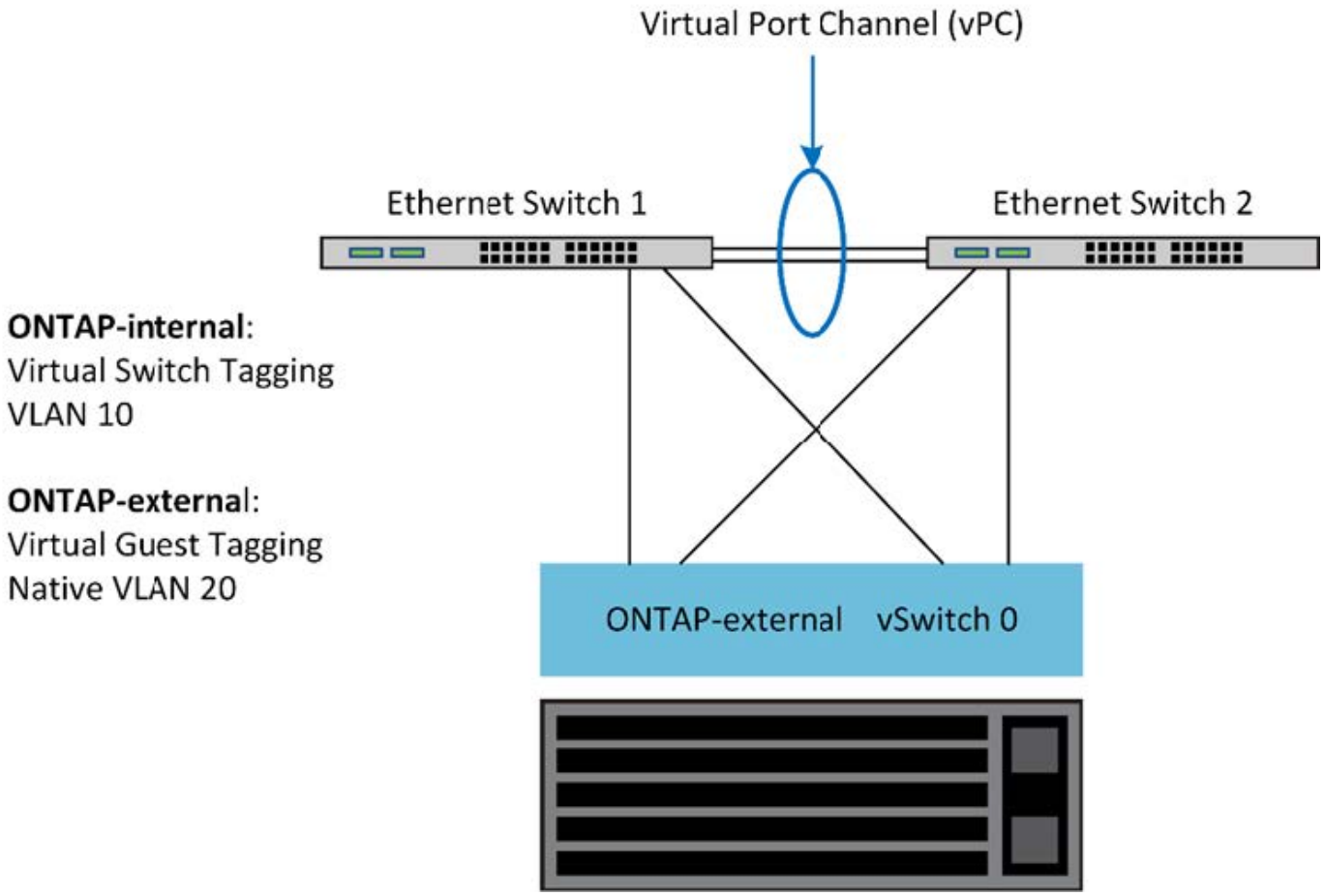



> 在此配置中，共享交换机将成为单点故障。如果可能，应使用多个交换机来防止物理硬件故障导致集群网络中断。

多个物理交换机

需要冗余时，应使用多个物理网络交换机。下图显示了一个多节点 ONTAP Select 集群中的一个节点所使用的建议配置。内部端口组和外部端口组中的 NIC 均通过缆线连接到不同的物理交换机，从而保护用户免受单个硬件交换机故障的影响。交换机之间配置了虚拟端口通道，以防止出现生成树问题。

- 使用多个物理交换机的网络配置 *




## 数据和管理流量隔离

### 将数据流量和管理流量隔离到单独的第 2 层网络中。

ONTAP Select 外部网络流量是指数据（ CIFS ， NFS 和 iSCSI ），管理和复制（ SnapMirror ）流量。在 ONTAP 集群中，每种流量都使用一个单独的逻辑接口，该接口必须托管在虚拟网络端口上。在 ONTAP Select 的多节点配置中，这些端口指定为端口 e0a 和 e0b/e0g 。在单节点配置中，这些端口指定为 e0a 和 e0b/e0c ，而其余端口则保留用于内部集群服务。

NetApp 建议将数据流量和管理流量隔离到单独的第 2 层网络中。在 ONTAP Select 环境中，可以使用 VLAN 标记来完成此操作。为此，可以将一个带 VLAN 标记的端口组分配给网络适配器 1 （端口 e0a ），用于管理流量。然后，您可以为端口 e0b 和 e0c （单节点集群）以及 e0b 和 e0g （多节点集群）分配一个单独的端口组以传输数据流量。

如果本文档前面所述的 VST 解决方案 还不够，则可能需要将数据和管理 LIF 同时托管在同一个虚拟端口上。要执行此操作，请使用一个称为 VGT 的过程， VM 将在该过程中执行 VLAN 标记。

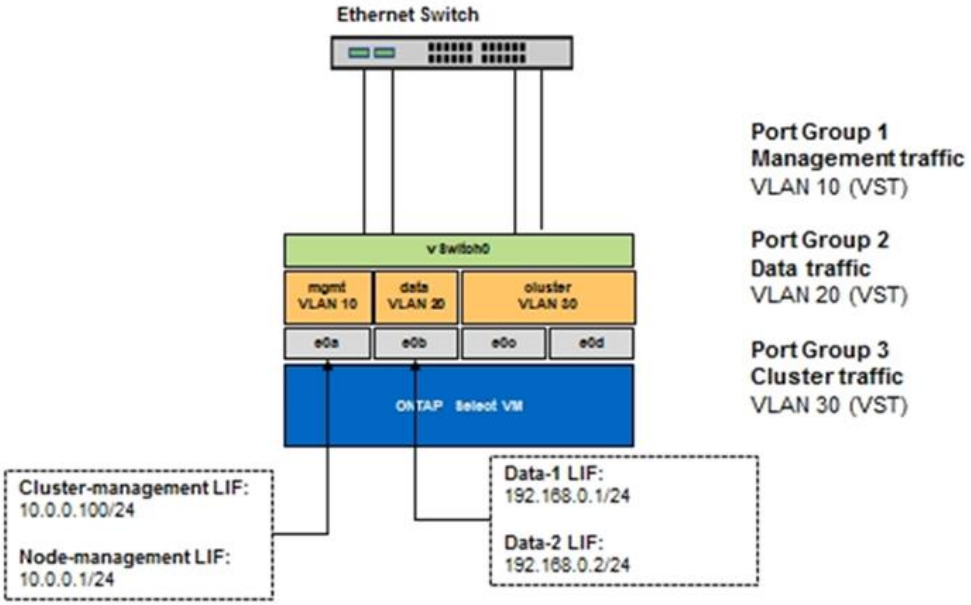
> 使用 ONTAP Deploy 实用程序时，无法通过 VGT 实现数据和管理网络隔离。此过程必须在集群设置完成后执行。

使用 VGT 和双节点集群时，还需要注意其他事项。在双节点集群配置中，在 ONTAP 完全可用之前，使用节点管理 IP 地址与调解器建立连接。因此，映射到节点管理 LIF （端口 e0a ）的端口组仅支持 EST 和 VST 标记。此外，如果管理流量和数据流量使用同一端口组，则整个双节点集群仅支持 EST/VST 。

VST 和 VGT 这两种配置选项均受支持。下图显示了第一种方案 VST ，其中流量通过分配的端口组在 vSwitch

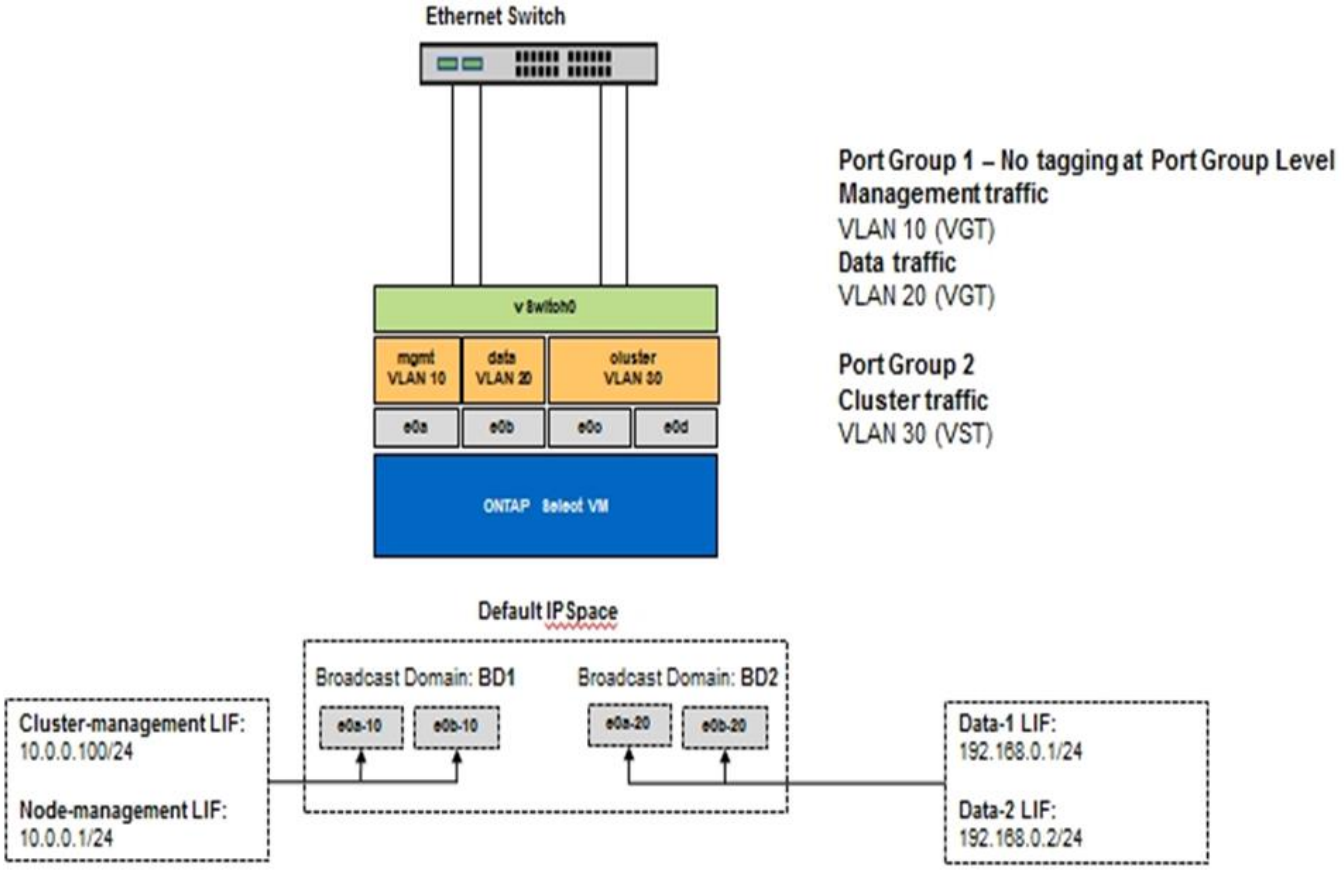
层进行标记。在此配置中，集群和节点管理 LIF 会分配给 ONTAP 端口 e0a ，并通过分配的端口组使用 VLAN ID 10 进行标记。数据 LIF 会分配给端口 e0b 以及 e0c 或 e0g ，并使用第二个端口组为其分配 VLAN ID 20 。集群端口使用第三个端口组，并且位于 VLAN ID 30 上。

- 使用 VST* 进行数据和管理隔离




下图显示了第二种方案 VGT ，在这种情况下， ONTAP VM 会使用放置在不同广播域中的 VLAN 端口对流量进行标记。在此示例中，虚拟端口 e0a-10/e0b-10/ （ e0c 或 e0g ） -10 和 e0a-20/e0b-20 位于 VM 端口 e0a 和 e0b 的顶部。此配置允许直接在 ONTAP 中执行网络标记，而不是在 vSwitch 层执行。管理和数据 LIF 放置在这些虚拟端口上，从而可以在一个 VM 端口中进一步细分第 2 层。集群 VLAN （ VLAN ID 30 ）仍会在端口组上进行标记。

- 注： *
- 使用多个 IP 空间时，这种配置方式尤其有用。如果需要进一步进行逻辑隔离和多租户，请将 VLAN 端口分组到单独的自定义 IP 空间中。
- 要支持 VGT ， ESXi/ESX 主机网络适配器必须连接到物理交换机上的中继端口。连接到虚拟交换机的端口组必须将其 VLAN ID 设置为 4095 ，才能在端口组上启用中继。
- 使用 VGT 实现数据和管理分离 *




# 高可用性架构

## 高可用性配置

## 发现高可用性选项，为您的环境选择最佳的 HA 配置。

尽管客户开始将应用程序工作负载从企业级存储设备迁移到在商用硬件上运行的基于软件的解决方案，但对故障恢复能力和容错的期望和需求并未改变。提供零恢复点目标（ RPO ）的 HA 解决方案 可保护客户免受因基础架构堆栈中任何组件出现故障而导致的数据丢失的影响。

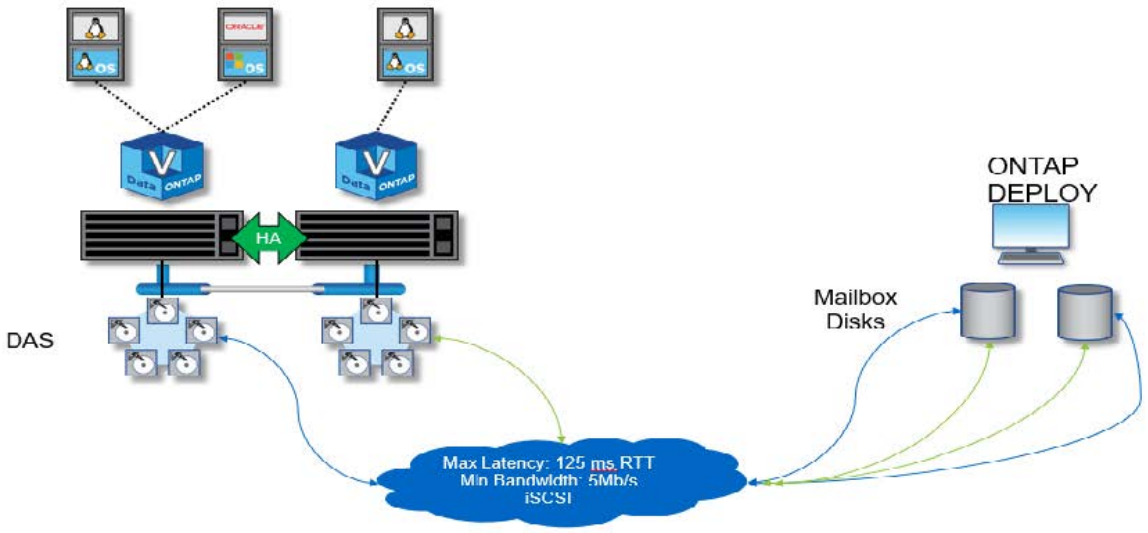
SDS 市场的很大一部分是基于无共享存储的概念构建的，软件复制可通过在不同存储孤岛之间存储多个用户数据副本来提供数据故障恢复能力。ONTAP Select 在此前提下构建，可使用 ONTAP 提供的同步复制功能（ RAID SyncMirror ）在集群中存储一份额外的用户数据副本。此问题发生在 HA 对的上下文中。每个 HA 对都会存储两个用户数据副本：一个位于本地节点提供的存储上，一个位于 HA 配对节点提供的存储上。在 ONTAP Select 集群中， HA 和同步复制绑定在一起，两者的功能不能分离或单独使用。因此，同步复制功能仅在多节点产品中可用。

> 在 ONTAP Select 集群中，同步复制功能是 HA 实施的一项功能，而不是异步 SnapMirror 或 SnapVault 复制引擎的替代功能。同步复制不能独立于 HA 使用。

ONTAP Select HA 部署模式有两种：多节点集群（四个，六个或八个节点）和双节点集群。双节点 ONTAP Select 集群的突出特点是使用外部调解器服务来解决脑裂问题。ONTAP Deploy 虚拟机用作其配置的所有双节点 HA 对的默认调解器。

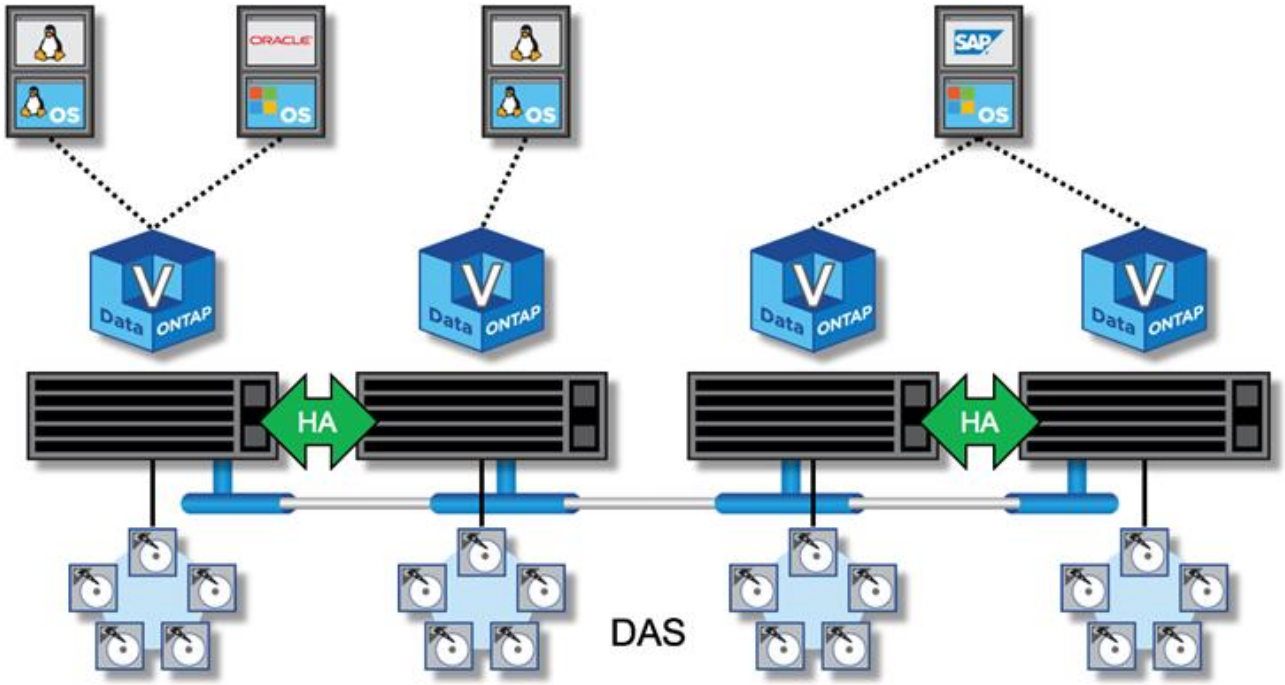
下图显示了这两种架构。

- 具有远程调解器并使用本地连接存储的双节点 ONTAP Select 集群 *




双节点 ONTAP Select 集群由一个 HA 对和一个调解器组成。在 HA 对中，每个集群节点上的数据聚合都会进行同步镜像，如果发生故障转移，则不会丢失任何数据。

- 使用本地连接存储的四节点 ONTAP Select 集群 *




- 四节点 ONTAP Select 集群由两个 HA 对组成。六节点和八节点集群分别由三个和四个 HA 对组成。在每个 HA 对中，每个集群节点上的数据聚合都会进行同步镜像，如果发生故障转移，则不会丢失任何数据。
- 使用 DAS 存储时，一个物理服务器上只能存在一个 ONTAP Select 实例。ONTAP Select 需要对系统的本地 RAID 控制器进行非共享访问，并可用于管理本地连接的磁盘，如果没有与存储的物理连接，则无法实现这一点。

双节点 **HA** 与多节点 **HA**

与 FAS 阵列不同， HA 对中的 ONTAP Select 节点仅通过 IP 网络进行通信。这意味着 IP 网络是单点故障（ SPOF ），防止网络分区和脑裂情形成为设计的一个重要方面。多节点集群可以承受单节点故障，因为三个或更多正常运行的节点可以建立集群仲裁。双节点集群依靠 ONTAP Deploy 虚拟机托管的调解器服务来实现相同的结果。

ONTAP Select 节点和 ONTAP Deploy 调解器服务之间的检测信号网络流量极少，并且具有故障恢复能力，因此 ONTAP Deploy 虚拟机可以托管在与 ONTAP Select 双节点集群不同的数据中心中。

> 当充当双节点集群的调解器时， ONTAP Deploy 虚拟机将成为该集群不可或缺的一部分。如果调解器服务不可用，则双节点集群将继续提供数据，但 ONTAP Select 集群的存储故障转移功能将被禁用。因此， ONTAP Deploy 调解器服务必须与 HA 对中的每个 ONTAP Select 节点保持持续通信。要使集群仲裁正常运行，至少需要 5 Mbps 的带宽和 125 毫秒的最大往返时间（ RTT ）延迟。

如果充当调解器的 ONTAP Deploy 虚拟机暂时或可能永久不可用，则可以使用二级 ONTAP Deploy 虚拟机来还原双节点集群仲裁。这会导致新的 ONTAP Deploy 虚拟机无法管理 ONTAP Select 节点，但它已成功参与集群仲裁算法。ONTAP Select 节点与 ONTAP Deploy 虚拟机之间的通信可通过使用基于 IPv4 的 iSCSI 协议来实现。ONTAP Select 节点管理 IP 地址为启动程序， ONTAP Deploy VM IP 地址为目标。因此，在创建双节点集群时，节点管理 IP 地址不能支持 IPv6 地址。在创建双节点集群时，系统会自动创建 ONTAP Deploy 托管邮箱磁盘，并将其屏蔽到正确的 ONTAP Select 节点管理 IP 地址。整个配置会在设置期间自动执行，无需执行进一步的管理操作。创建集群的 ONTAP Deploy 实例是该集群的默认调解器。

如果必须更改原始调解器位置，则需要执行管理操作。即使原始 ONTAP Deploy 虚拟机丢失，也可以恢复集群仲裁。但是， NetApp 建议您在实例化每个双节点集群后备份 ONTAP Deploy 数据库。

双节点 **HA** 与双节点延伸型 **HA** （ **MetroCluster SDS** ）

可以将双节点主动 / 主动 HA 集群延伸到更远的距离，并可能将每个节点放置在不同的数据中心中。双节点集群与双节点延伸型集群（也称为 MetroCluster SDS ）之间的唯一区别是节点之间的网络连接距离。

双节点集群定义为一个集群，其中两个节点位于同一数据中心，距离 300 米。通常，两个节点都具有指向同一网络交换机或一组交换机间链路（ ISL ）网络交换机的上行链路。

双节点 MetroCluster SDS 的定义是一个集群，其节点（不同的机房，不同的建筑物和不同的数据中心）物理隔离超过 300 米。此外，每个节点的上行链路连接都连接到不同的网络交换机。MetroCluster SDS 不需要专用硬件。但是，环境应遵守延迟（ RTT 最长为 5 毫秒，抖动最大为 5 毫秒，总共为 10 毫秒）和物理距离（最长为 10 公里）的要求。

MetroCluster SDS是一项高级功能、需要高级版许可证或高级尊享版许可证。高级版许可证支持创建中小型 VM 以及 HDD 和 SSD 介质。高级XL许可证还支持创建NVMe驱动器。

> 本地连接存储（ DAS ）和共享存储（ vNAS ）均支持 MetroCluster SDS 。请注意，由于 ONTAP Select VM 和共享存储之间的网络， vNAS 配置的固有延迟通常较高。MetroCluster SDS 配置必须在节点之间提供最长 10 毫秒的延迟，包括共享存储延迟。换言之，仅测量 Select VM 之间的延迟是不够的，因为对于这些配置，共享存储延迟并不可忽略。

## **HA RSM** 和镜像聚合

使用 RAID SyncMirror （ RSM ），镜像聚合和写入路径防止数据丢失。

同步复制

ONTAP HA 模式基于 HA 配对节点的概念构建。ONTAP Select 可通过使用 ONTAP 中的 RAID SyncMirror （ RSM ）功能在集群节点之间复制数据块，从而将此架构扩展到非共享商用服务器环境中，从而为分布在 HA 对中的用户数据提供两个副本。

具有调解器的双节点集群可以跨越两个数据中心。有关详细信息，请参见一节 "双节点延伸型 HA （ MetroCluster SDS ）最佳实践"。

镜像聚合

一个 ONTAP Select 集群由两到八个节点组成。每个 HA 对包含两个用户数据副本，这些副本通过 IP 网络在节点之间同步镜像。此镜像对用户是透明的，它是数据聚合的一个属性，在数据聚合创建过程中会自动配置。

必须镜像 ONTAP Select 集群中的所有聚合，以便在发生节点故障转移时提供数据，并避免发生硬件故障时出现 SPOF 。ONTAP Select 集群中的聚合使用 HA 对中每个节点提供的虚拟磁盘构建，并使用以下磁盘：

- 一组本地磁盘（由当前 ONTAP Select 节点提供）
- 一组镜像磁盘（由当前节点的 HA 配对节点提供）

> 用于构建镜像聚合的本地磁盘和镜像磁盘的大小必须相同。这些聚合称为丛 0 和丛 1 （分别表示本地和远程镜像对）。实际丛编号在您的安装中可能有所不同。

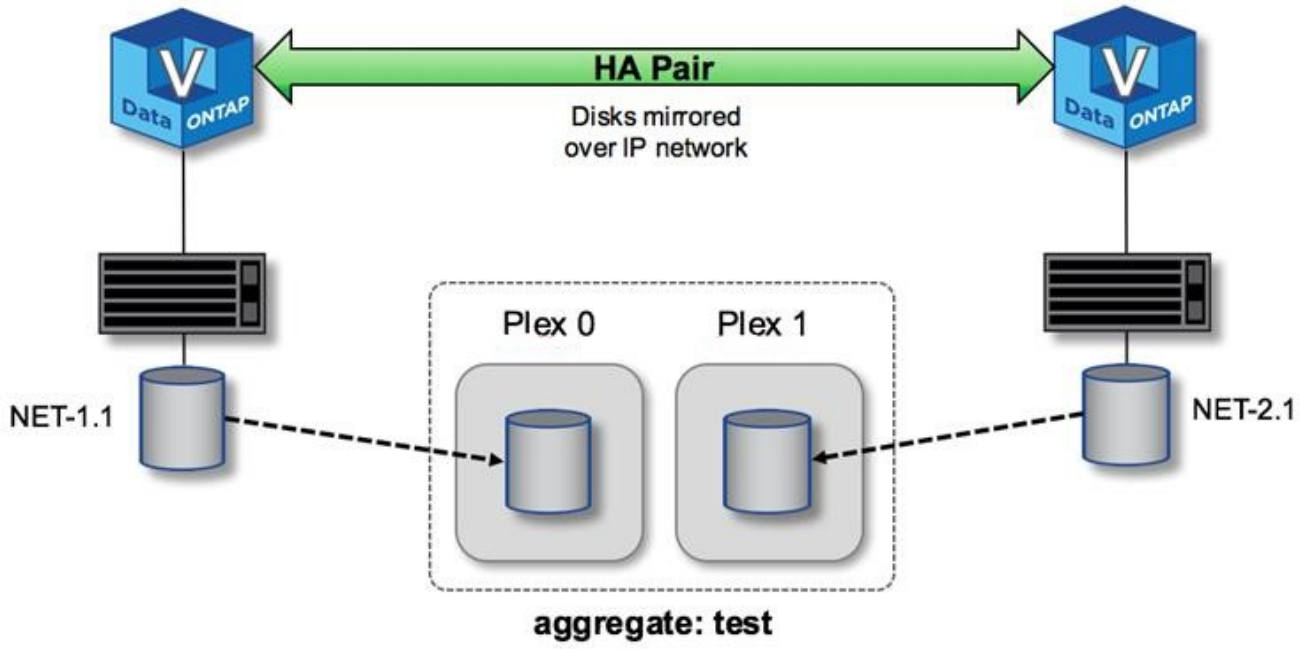
这种方法与标准 ONTAP 集群的工作方式有着根本的不同。此适用场景 将对 ONTAP Select 集群中的所有根磁盘和数据磁盘执行。聚合同时包含数据的本地副本和镜像副本。因此，包含 N 个虚拟磁盘的聚合可提供相当于 N/2 个磁盘的唯一存储，因为第二个数据副本驻留在其自身的唯一磁盘上。

下图显示了四节点 ONTAP Select 集群中的一个 HA 对。此集群中只有一个聚合（测试），该聚合使用两个 HA 配对节点的存储。此数据聚合由两组虚拟磁盘组成：一组本地磁盘，由 ONTAP Select 所属集群节点（丛 0 ）提供；另一组远程磁盘，由故障转移配对节点（丛 1 ）提供。

丛 0 是存放所有本地磁盘的分段。丛 1 是用于存放镜像磁盘或负责存储用户数据第二个复制副本的磁盘的存储分段。拥有聚合的节点将磁盘分配给 Plex 0 ，而该节点的 HA 配对节点将磁盘分配给 Plex 1 。

在下图中，存在一个包含两个磁盘的镜像聚合。此聚合的内容会在我们的两个集群节点之间进行镜像，并将本地磁盘 NET-1.1 置于 Plex 0 分段中，而将远程磁盘 NET-2.1 置于 Plex 1 分段中。在此示例中，聚合测试由左侧的集群节点拥有，并使用本地磁盘 NET-1.1 和 HA 配对镜像磁盘 NET-2.1 。

- ONTAP Select 镜像聚合 *




部署 ONTAP Select 集群后，系统上的所有虚拟磁盘都会自动分配给正确的丛，无需用户在磁盘分配方面执行额外步骤。这样可以防止意外将磁盘分配给不正确的丛，并提供最佳的镜像磁盘配置。

写入路径

在集群节点之间同步镜像数据块以及在发生系统故障时不丢失数据的要求会对传入写入在通过 ONTAP Select 集群传播时所采用的路径产生重大影响。此过程包括两个阶段：

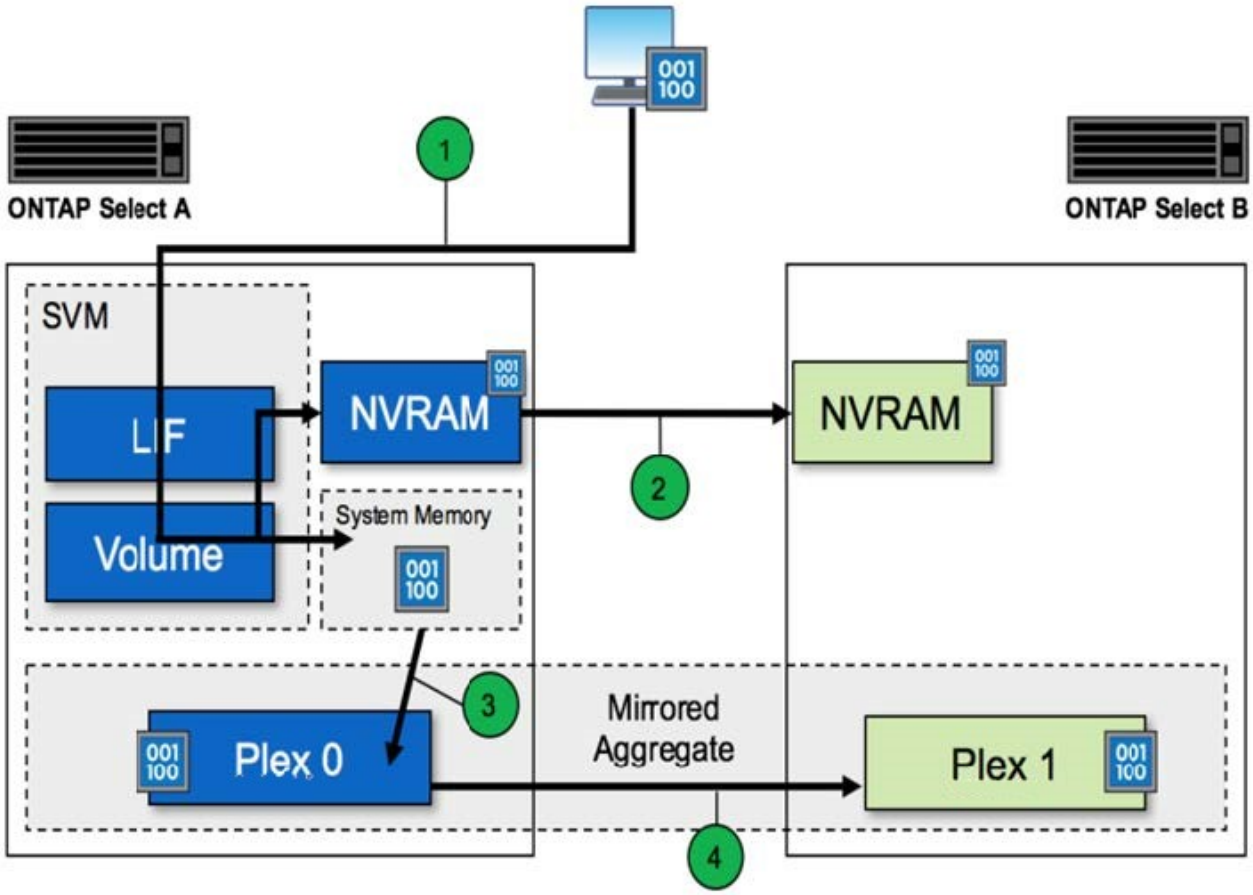
- 确认
- 转存

对目标卷的写入会通过数据 LIF 进行，并提交到 ONTAP Select 节点的系统磁盘上的虚拟化 NVRAM 分区，然后再确认回客户端。在 HA 配置中，还会执行另一个步骤，因为这些 NVRAM 写入操作会在被确认之前立即镜像到目标卷所有者的 HA 配对节点。如果原始节点出现硬件故障，此过程可确保 HA 配对节点上的文件系统一致性。

将写入提交到 NVRAM 后， ONTAP 会定期将此分区的内容移动到相应的虚拟磁盘，此过程称为转存。此过程仅在目标卷所属的集群节点上发生一次，而不会在 HA 配对节点上发生。

下图显示了传入写入请求到 ONTAP Select 节点的写入路径。

- ONTAP Select 写入路径工作流 *




传入写入确认包括以下步骤：

- 写入操作通过 ONTAP Select 节点 A 拥有的逻辑接口进入系统
- 写入将提交到节点 A 的 NVRAM 并镜像到 HA 配对节点 B
- 在两个 HA 节点上都存在 I/O 请求后，该请求将确认回客户端。

ONTAP Select 从 NVRAM 转存到数据聚合（ ONTAP CP ）包括以下步骤：

- 写入将从虚拟 NVRAM 转存到虚拟数据聚合。
- 镜像引擎将块同步复制到两个丛。

## HA 其他详细信息

### HA 磁盘检测信号， HA 邮箱， HA 检测信号， HA 故障转移和交还用于增强数据保护。

#### 磁盘检测信号

尽管 ONTAP Select HA 架构利用了传统 FAS 阵列使用的许多代码路径，但仍存在一些例外情况。其中一个例外情况是实施基于磁盘的检测信号，这是一种非基于网络的通信方法，集群节点使用此方法来防止网络隔离导致脑裂行为。脑裂情形是集群分区的结果，通常是由网络故障引起的，其中每一方都认为另一方已关闭并尝试接管集群资源。

企业级 HA 实施必须妥善处理此类情形。ONTAP 通过基于磁盘的自定义检测方法来实现这一点。这是 HA 邮箱

的作业， HA 邮箱位于物理存储上，集群节点使用此位置传递检测信号消息。这有助于集群确定连接，从而在发生故障转移时定义仲裁。

在使用共享存储 HA 架构的 FAS 阵列上， ONTAP 通过以下方式解决脑裂问题：

- SCSI 永久性预留
- 永久性 HA 元数据
- 通过 HA 互连发送的 HA 状态

但是，在 ONTAP Select 集群的无共享架构中，节点只能看到自己的本地存储，而不能看到 HA 配对节点的本地存储。因此，如果网络分区将 HA 对的每一侧隔离，则无法使用上述确定集群仲裁和故障转移行为的方法。

尽管无法使用现有的脑裂检测和避免方法，但仍然需要一种调解方法，一种可满足无共享环境限制的方法。ONTAP Select 进一步扩展了现有的邮箱基础架构，使其可以在发生网络分区时充当调解方法。由于共享存储不可用，因此可以通过 NAS 访问邮箱磁盘来完成调解。这些磁盘使用 iSCSI 协议分布在整个集群中，包括双节点集群中的调解器。因此，集群节点可以根据对这些磁盘的访问来做出智能故障转移决策。如果某个节点可以访问其 HA 配对节点以外其他节点的邮箱磁盘，则该节点可能已启动且运行状况良好。

> 解决集群仲裁和脑裂问题的邮箱架构和基于磁盘的检测信号方法是多节点 ONTAP Select 变体需要四个单独节点或一个双节点集群调解器的原因。

**HA** 邮箱发布

HA 邮箱架构使用消息发布模式。集群节点会定期向集群中的所有其他邮箱磁盘（包括调解器）发布消息，指出节点已启动且正在运行。在运行状况良好的集群中的任意时间点，集群节点上的单个邮箱磁盘会从所有其他集群节点发布消息。

连接到每个 Select 集群节点的虚拟磁盘专用于共享邮箱访问。此磁盘称为调解器邮箱磁盘，因为其主要功能是在发生节点故障或网络分区时充当集群调解的方法。此邮箱磁盘包含每个集群节点的分区，并由其他 Select 集群节点通过 iSCSI 网络挂载。这些节点会定期将运行状况发布到邮箱磁盘的相应分区。使用分布在整个集群中的可通过网络访问的邮箱磁盘，您可以通过可访问性表推断节点运行状况。例如，集群节点 A 和 B 可以发布到集群节点 D 的邮箱，但不能发布到节点 C 的邮箱此外，集群节点 D 无法发布到节点 C 的邮箱，因此节点 C 可能已关闭或与网络隔离，应接管。

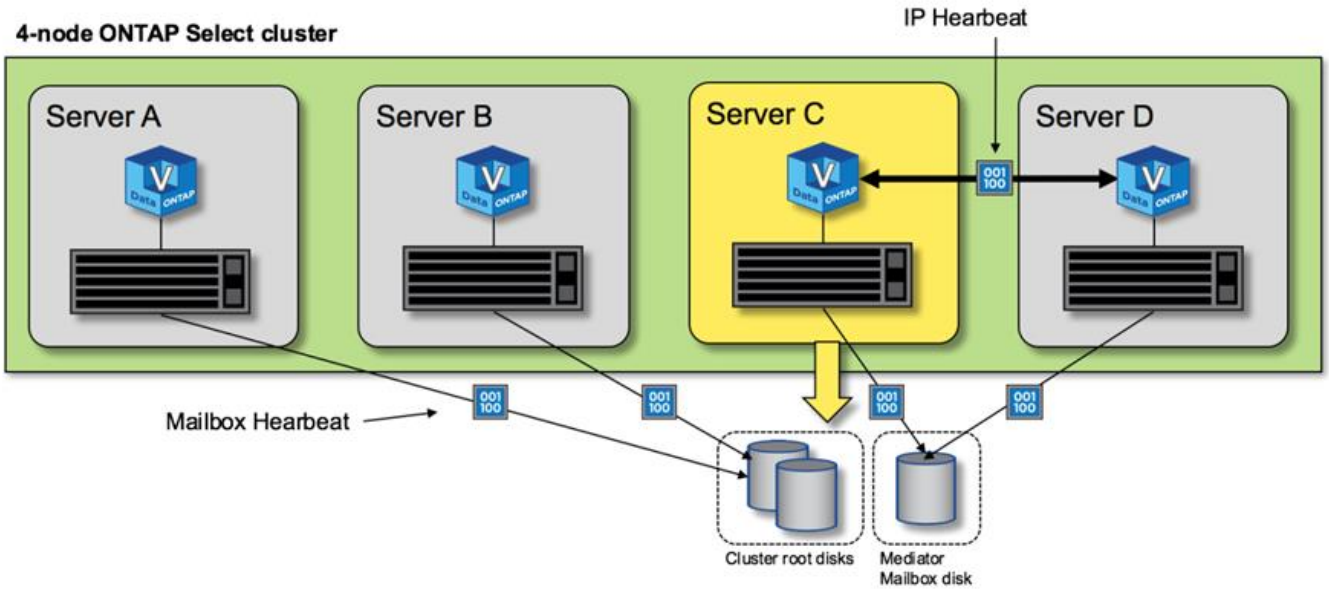
**HA** 检测信号

与 NetApp FAS 平台一样， ONTAP Select 会定期通过 HA 互连发送 HA 检测信号消息。在 ONTAP Select 集群中，此操作通过 HA 配对节点之间的 TCP/IP 网络连接来执行。此外，基于磁盘的检测信号消息会传递到所有 HA 邮箱磁盘，包括调解器邮箱磁盘。这些消息每隔几秒传递一次，并定期进行读回。通过发送和接收这些消息的频率， ONTAP Select 集群可以在大约 15 秒内检测 HA 故障事件，这与 FAS 平台上提供的窗口相同。如果不再读取检测信号消息，则会触发故障转移事件。

下图显示了从单个 ONTAP Select 集群节点节点节点 C 的角度通过 HA 互连和调解器磁盘发送和接收检测信号消息的过程

> 网络检测信号通过 HA 互连发送到 HA 配对节点 D ，而磁盘检测信号则在所有集群节点 A ， B ， C 和 D 上使用邮箱磁盘

四节点集群中的 * HA 检测信号：稳定状态 *




**HA** 故障转移和交还

在故障转移操作期间，运行正常的节点会使用其 HA 配对节点的本地数据副本为其对等节点提供数据。客户端 I/O 可以无中断继续，但必须先复制此数据的更改，然后才能进行交还。请注意， ONTAP Select 不支持强制交还，因为这会导致存储在正常运行的节点上的更改丢失。

重新启动的节点重新加入集群时，将自动触发同步回滚操作。同步回滚所需的时间取决于多个因素。这些因素包括必须复制的更改数，节点之间的网络延迟以及每个节点上磁盘子系统的速度。同步返回所需的时间可能会超过 10 分钟的自动交还窗口。在这种情况下，需要在同步回滚后手动交还。可以使用以下命令监控同步恢复的进度：

```
storage aggregate status -r -aggregate <aggregate name>
```

# 性能

## 性能

## 性能因硬件配置而异。

由于底层硬件和配置的特征， ONTAP Select 集群的性能可能会有很大差异。特定硬件配置是影响特定ONTAP Select 实例性能的最大因素。以下是影响特定 ONTAP Select 实例性能的一些因素：

- * 核心频率 * 。一般来说，最好使用较高的频率。
- * 单插槽与多插槽 * 。ONTAP Select 不使用多插槽功能，但支持多插槽配置所需的虚拟机管理程序开销会在整体性能方面造成一定程度的偏差。
- * RAID 卡配置和关联的虚拟机管理程序驱动程序 * 。虚拟机管理程序提供的默认驱动程序可能需要替换为硬件供应商驱动程序。
- * RAID 组中的驱动器类型和驱动器数量 * 。
- * 虚拟机管理程序版本和修补程序级别 * 。

## 性能：高级HA直连SSD存储

参考平台的性能信息。

参考平台

ONTAP Select (高级版XL)硬件(每个节点)

- Fujitsu PRIMERGY RX2540 M4：
  - 2.6 GHz 的 Intel （R） Xeon （R） Gold 6142b CPU
  - 32 个物理核心（16 个 2 插槽）， 64 个逻辑核心
  - 256 GB RAM
  - 每个主机的驱动器数： 24 个 960 GB SSD
  - ESX 6.5U1

客户端硬件

- 5 个 NFSv3 IBM 3550m4 客户端

配置信息

- 软件 RAID 1 x 9 + 2 RAID-DP （11 个驱动器）
- 22+1 RAID-5 （ONTAP 中的 RAID-0） /RAID 缓存 NVRAM
- 未使用存储效率功能（数据压缩，重复数据删除， Snapshot 副本， SnapMirror 等）

下表列出了根据使用软件RAID和硬件RAID的高可用性(HA) ONTAP Select 节点对上的读/写工作负载测量的吞吐量。性能测量是使用 SIO 负载生成工具进行的。

这些性能数据基于ONTAP Select 9.6。

*使用软件RAID和硬件RAID*的直连存储(DAS) SSD上单个节点(四节点中型实例的一部分) ONTAP Select 集群的性能结果

| **Description** | 顺序读取 **64KiB** | 顺序写入 **64KiB** | 随机读取 **8 KiB** | 随机写入 **8 KiB** | 随机 **WR/RD** （**50/50**） **8 KiB** |
|---|---|---|---|---|---|
| 采用DAS (SSD)软件RAID的ONTAP Select大型实例 | 2171 MiBps | 559 MiBps | 9554 MiBps | 294 MiBps | 5664 MiBps |
| 采用DAS (SSD)软件RAID的ONTAP Select中型实例 | 2090 MiBps | 592 MiBps | 每秒位为MIBps | 335 MiBps | 441 个 3 MiBps |

| Description | 顺序读取 64KiB | 顺序写入 64KiB | 随机读取 8 KiB | 随机写入 8 KiB | 随机 WR/RD （50/50） 8 KiB |
|---|---|---|---|---|---|
| 采用DAS (SSD)硬件RAID的ONTAP Select中型实例 | 2038 MiBps | 520 MiBps | 578 MiBps | 325 MiBps | 399 MiBps |

**64K** 顺序读取

详细信息：

- 已启用 SIO 直接 I/O
- 2 个节点
- 每个节点 2 个数据 NIC
- 每个节点 1 个数据聚合（2 TB 硬件 RAID），（8 TB 软件 RAID）
- 64 个 SIO 进程，每个进程 1 个线程
- 每个节点 32 个卷
- 每个进程 1 个文件；每个进程的文件大小为 12000 MB

**64K** 顺序写入

详细信息：

- 已启用 SIO 直接 I/O
- 2 个节点
- 每个节点2个数据网络接口卡(NIC)
- 每个节点 1 个数据聚合（2 TB 硬件 RAID），（4 TB 软件 RAID）
- 128个SIO进程、每个进程1个线程
- 每个节点的卷数：32 (硬件RAID)、16 (软件RAID)
- 每个进程 1 个文件；每个进程的文件大小为 30720 MB

**8 K** 随机读取

详细信息：

- 已启用 SIO 直接 I/O
- 2 个节点
- 每个节点2个数据NIC
- 每个节点 1 个数据聚合（2 TB 硬件 RAID），（4 TB 软件 RAID）
- 64 个 SIO 进程，每个进程 8 个线程
- 每个节点的卷数：32
- 每个进程 1 个文件；每个进程的文件大小为 12228MB

**8 K** 随机写入

详细信息：

- 已启用 SIO 直接 I/O
- 2 个节点
- 每个节点2个数据NIC
- 每个节点 1 个数据聚合（ 2 TB 硬件 RAID ），（ 4 TB 软件 RAID ）
- 64 个 SIO 进程，每个进程 8 个线程
- 每个节点的卷数：32
- 每个进程 1 个文件；每个进程的文件大小为 8192 MB

**8 K** 随机 **50%** 写入 **50%** 读取

详细信息：

- 已启用 SIO 直接 I/O
- 2 个节点
- 每个节点2个数据NIC
- 每个节点 1 个数据聚合（ 2 TB 硬件 RAID ），（ 4 TB 软件 RAID ）
- 每个进程 64 个 SIO 进程 208 个线程
- 每个节点的卷数：32
- 每个进程 1 个文件；每个进程的文件大小为 12228MB