高可用性架构
ONTAP Select

NetApp
May 07, 2026

目录

高可用性架构

ONTAP Select 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转移到基于软件的商用硬件上运行的解决方案，但围绕弹性和容错的期望和需求并没有改变。提供零恢复点目标 (RPO) 的高可用性解决方案可保护客户免受基础设施堆栈中任何组件故障造成的数据丢失。

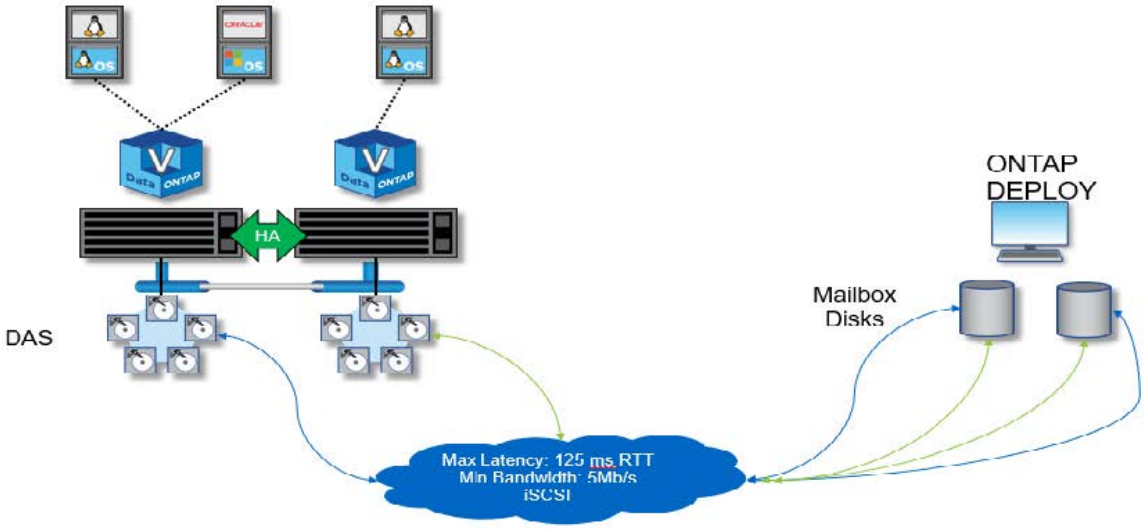
SDS 市场的很大一部分建立在无共享存储的概念之上，软件复制通过在不同的存储孤岛上存储多个用户数据副本来提供数据弹性。ONTAP Select 基于此前提，使用 ONTAP 提供的同步复制功能（RAID SyncMirror）在集群中存储用户数据的额外副本。这发生在 HA 对的上下文中。每个 HA 对存储两个用户数据副本：一个在本地节点提供的存储上，另一个在 HA 合作伙伴提供的存储上。在 ONTAP Select 集群中，HA 和同步复制联系在一起，两者的功能不能分离或独立使用。因此，同步复制功能仅在多节点产品中可用。

> 在 ONTAP Select 集群中，同步复制功能是 HA 实现的功能，而不是异步 SnapMirror 或 SnapVault 复制引擎的替代品。同步复制不能独立于 HA 使用。

有两种 ONTAP Select HA 部署模型：多节点集群（四个、六个、八个、十个或十二个节点）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场景。ONTAP Deploy VM 充当其配置的所有双节点 HA 对的默认调解器。

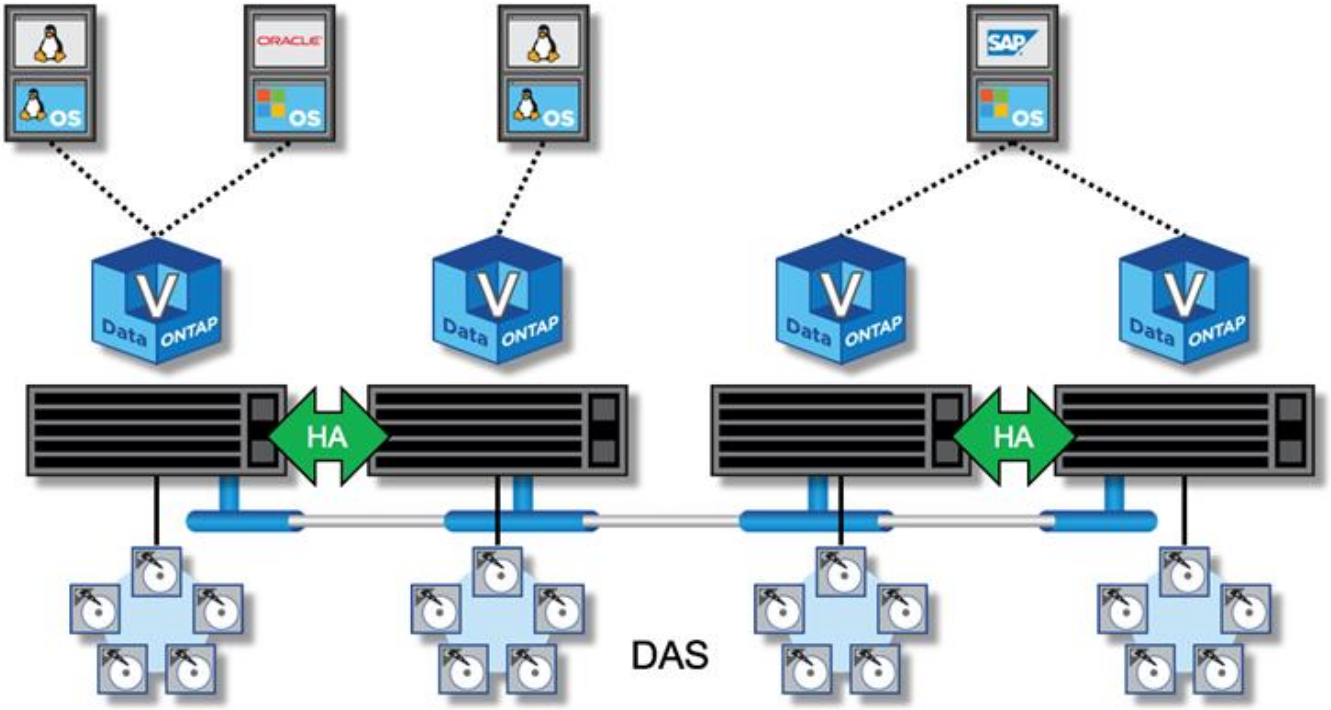
这两个架构如下图所示。

双节点 **ONTAP Select** 集群，带有远程介体并使用本地连接存储



> 双节点 ONTAP Select 集群由一个 HA 对和一个中介器组成。在 HA 对中，每个集群节点上的数据聚合会同步镜像，并且在发生故障转移时不会丢失数据。

四节点 **ONTAP Select** 集群使用本地连接存储



- 四节点 ONTAP Select 集群由两个 HA 对组成。六节点、八节点、十节点和十二节点集群分别由三个、四个、五个和六个 HA 对组成。在每个 HA 对中，每个集群节点上的数据聚合都会同步镜像，并且在发生故障转移时不会丢失数据。
- 使用 DAS 存储时，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，旨在管理本地连接的磁盘，如果没有对存储的物理连接，这是不可能的。

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 网络是一个单点故障（SPOF），防止网络分区和分裂脑场景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，因为集群仲裁可以由三个或多个幸存的节点建立。双节点集群依靠 ONTAP Deploy VM 托管的中介服务实现相同的结果。

ONTAP Select 节点和 ONTAP Deploy 中介服务之间的心跳网络流量最小且具有弹性，因此 ONTAP Deploy VM 可以托管在不同于 ONTAP Select 双节点集群的数据中心。

> ONTAP Deploy VM 在充当双节点集群的中介时，成为该集群不可或缺的一部分。如果中介服务不可用，则双节点集群继续服务数据，但 ONTAP Select 集群的存储故障转移功能被禁用。因此，ONTAP Deploy 中介服务必须与 HA 对中的每个 ONTAP Select 节点保持恒定的通信。最小带宽为 5Mbps，最大往返时间（RTT）延迟为 125ms，才能使集群仲裁正常运行。

如果充当中介的 ONTAP Deploy VM 暂时或可能永久不可用，则可以使用辅助 ONTAP Deploy VM 来恢复双节点集群仲裁。这会导致新的 ONTAP Deploy VM 无法管理 ONTAP Select 节点，但它已成功参与集群仲裁算法的配置。ONTAP Select 节点与 ONTAP Deploy VM 之间的通信是通过 IPv4 上的 iSCSI 协议完成的。ONTAP Select 节点管理 IP 地址为发起方，ONTAP Deploy VM IP 地址为目标方。因此，在创建双节点集群时，不可能为节点管理 IP 地址支持 IPv6 地址。在创建双节点集群时，ONTAP Deploy 托管邮箱磁盘将自动创建并屏蔽为正确的 ONTAP Select 节点管理 IP 地址。整个配置在设置过程中自动执行，无需执行进一步的管理操作。创建此集群的 ONTAP Deploy 实例是此集群的默认中介。

如果必须更改原始调解员位置，则需要执行管理操作。即使原始 ONTAP Deploy VM 丢失，也可以恢复群集仲

裁。但是，NetApp 建议在实例化每个双节点群集后备份 ONTAP Deploy 数据库。

双节点 HA 与双节点拉伸 HA（MetroCluster SDS）

可以将双节点、活动/活动 HA 集群扩展到更远的距离，并可能将每个节点放置在不同的数据中心。双节点集群和双节点延伸集群（也称为 MetroCluster SDS）之间的唯一区别是节点之间的网络连接距离。

双节点集群被定义为两个节点位于 300m 距离内的同一个数据中心的集群。一般来说，两个节点都有到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 被定义为具有物理分隔（不同房间、不同建筑物和不同数据中心）超过 300 米的节点的集群。此外，每个节点的上行链路连接都连接到单独的网络交换机。MetroCluster SDS 不需要专用硬件。但是，环境应遵守延迟要求（RTT 最多 5ms，抖动最多 5ms，总计 10ms）。

MetroCluster SDS 是一项高级功能，需要 Premium 许可证或 Premium XL 许可证。Premium 许可证支持创建中小型虚拟机以及 HDD 和 SSD 介质。Premium XL 许可证还支持创建 NVMe 驱动器。

> MetroCluster SDS 支持本地连接存储 (DAS) 和共享存储 (vNAS)。请注意，由于 ONTAP Select VM 和共享存储之间的网络，vNAS 配置通常具有更高的固有延迟。MetroCluster SDS 配置必须在节点之间提供最大 10ms 的延迟，包括共享存储延迟。换句话说，仅测量 Select VM 之间的延迟是不够的，因为共享存储延迟对于这些配置来说是不可忽视的。

ONTAP Select 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 模型建立在 HA 合作伙伴的概念之上。ONTAP Select 通过使用 ONTAP 中存在的 RAID SyncMirror (RSM) 功能在集群节点之间复制数据块，将此架构扩展到非共享商品服务器领域，提供分布在 HA 对中的用户数据的两个副本。

具有中介的双节点群集可以跨越两个数据中心。有关详细信息，请参阅部分"双节点拉伸 HA（MetroCluster SDS）最佳实践"。

镜像聚合

ONTAP Select 集群由 2 到 12 个节点组成。每个 HA 对包含两个用户数据副本，通过 IP 网络跨节点同步镜像。此镜像对用户是透明的，并且是数据聚合的属性，该属性在数据聚合创建过程中自动配置。

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获得数据可用性，并在发生硬件故障时避免 SPOF。ONTAP Select 集群中的聚合是从 HA 对中每个节点提供的虚拟磁盘构建的，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 合作伙伴提供）

> 用于构建镜像聚合的本地磁盘和镜像磁盘必须大小相同。这些聚合被称为丛 0 和丛 1（分别表示本地和远程镜像对）。在您的安装中，实际的丛编号可能会有所不同。

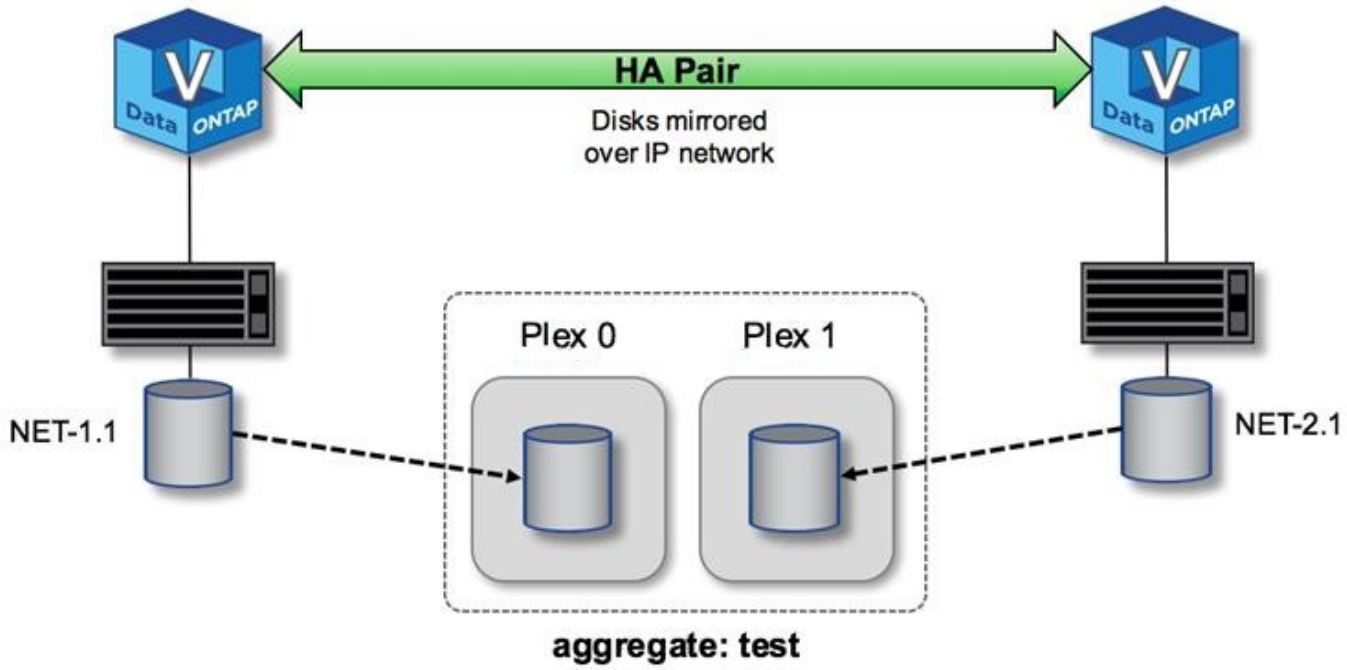
这种方法与标准 ONTAP 集群的工作方式截然不同。这适用于 ONTAP Select 集群中的所有根和数据磁盘。聚合包含数据的本地和镜像副本。因此，包含 N 个虚拟磁盘的聚合提供了 N/2 个磁盘的独特存储，因为数据的第二个副本驻留在其自己的独特磁盘上。

下图显示了四节点 ONTAP Select 集群中的 HA 对。在此集群中，有一个使用来自两个 HA 合作伙伴存储的单个聚合（test）。此数据聚合由两组虚拟磁盘组成：由拥有 ONTAP Select 集群节点贡献的本地集（Plex 0）和由故障转移合作伙伴贡献的远程集（Plex 1）。

Plex 0 是容纳所有本地磁盘的存储桶。Plex 1 是存储镜像磁盘的存储桶，或负责存储用户数据的第二个复制副本的磁盘。拥有聚合的节点向 Plex 0 提供磁盘，该节点的 HA 合作伙伴向 Plex 1 提供磁盘。

在下图中，有一个带有两个磁盘的镜像聚合。此聚合的内容在我们的两个集群节点上进行镜像，本地磁盘 NET-1.1 放置在 Plex 0 存储桶中，远程磁盘 NET-2.1 放置在 Plex 1 存储桶中。在此示例中，聚合 test 由左侧的集群节点所有，并使用本地磁盘 NET-1.1 和 HA 合作伙伴镜像磁盘 NET-2.1。

ONTAP Select 镜像聚合



部署 ONTAP Select 群集时，系统上存在的所有虚拟磁盘将自动分配给正确的丛，无需用户进行有关磁盘分配的额外步骤。这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

集群节点之间数据块的同步镜像以及系统故障时无数据丢失的要求对传入写入在 ONTAP Select 集群中传播时所采取的路径具有重大影响。这个过程包括两个阶段：

- 确认
- 分段

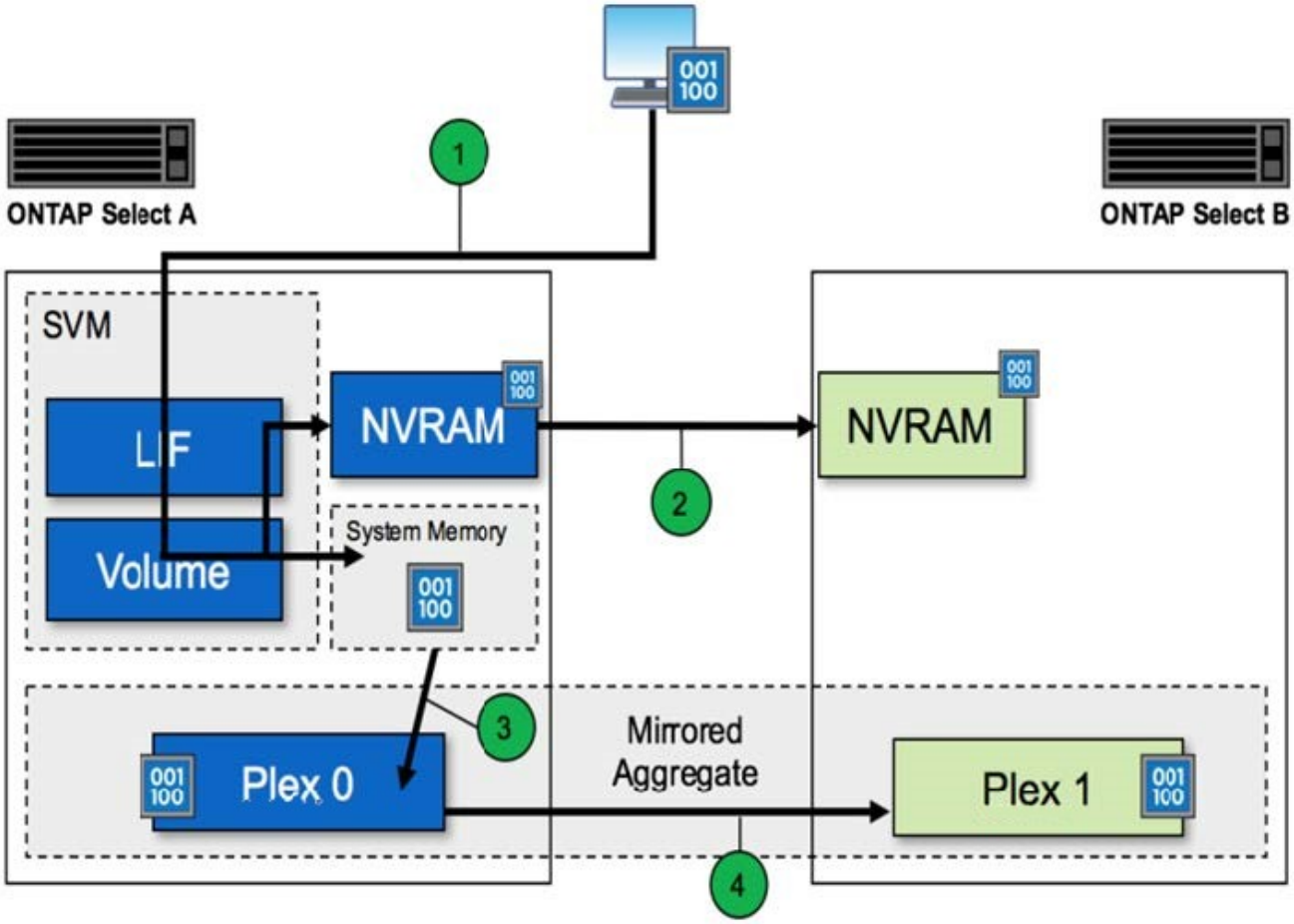
对目标卷的写入发生在数据 LIF 上，并提交到存在于 ONTAP Select 节点系统磁盘上的虚拟化 NVRAM 分区，然后才被确认返回到客户端。在 HA 配置上，会发生一个额外的步骤，因为这些 NVRAM 写入在被确认之前会立即镜像到目标卷所有者的 HA 合作伙伴。如果原始节点上存在硬件故障，此过程将确保 HA 合作伙伴节点上的

文件系统一致性。

写入提交到 NVRAM 后，ONTAP 会定期将此分区的内容移动到适当的虚拟磁盘，这一过程称为转储。此过程仅在拥有目标卷的集群节点上发生一次，在 HA 合作伙伴上不会发生。

下图显示了传入 ONTAP Select 节点的写入请求的写入路径。

ONTAP Select 写入路径工作流程



传入写入确认包括以下步骤：

- 写操作通过 ONTAP Select 节点 A 拥有的逻辑接口进入系统。
- 写入提交到节点 A 的 NVRAM 并镜像到 HA 合作伙伴节点 B。
- 在两个 HA 节点上都存在 I/O 请求后，该请求将被确认回客户端。

ONTAP Select 从 NVRAM 到数据聚合（ONTAP CP）的转储包括以下步骤：

- 写入从虚拟 NVRAM 转储到虚拟数据聚合。
- 镜像引擎将块同步复制到两个丛。

ONTAP Select HA 增强了数据保护

高可用性 (HA) 磁盘心跳、HA 邮箱、HA 心跳、HA 故障转移和 Giveback 工作，以增强数

据保护。

磁盘心跳

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。其中一个例外是基于磁盘的心跳的实现，这是一种非基于网络的通信方法，由集群节点使用，以防止网络隔离导致脑裂行为。脑裂场景是集群分区的结果，通常是由网络故障引起的，其中双方都认为对方已关闭并试图接管集群资源。

企业级 HA 实现必须优雅地处理此类方案。ONTAP 通过自定义的基于磁盘的心跳方法来实现这一点。这是 HA 邮箱的工作，HA 邮箱是物理存储上的一个位置，群集节点使用它来传递心跳消息。这有助于群集确定连接，从而在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 永久预留
- 持久 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 合作伙伴的本地存储。因此，当网络分区隔离 HA 对的每一侧时，确定集群仲裁和故障转移行为的前述方法不可用。

虽然不能使用现有的脑裂检测和避免方法，但仍然需要一种仲裁方法，这种方法符合无共享环境的约束。ONTAP Select 扩展了现有的邮箱基础架构，使其能够在发生网络分区时充当仲裁方法。由于共享存储不可用，因此通过 NAS 访问邮箱磁盘来实现仲裁。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仲裁器。因此，集群节点可以根据对这些磁盘的访问情况做出智能故障转移决策。如果某个节点可以访问其 HA 合作伙伴之外的其他节点的邮箱磁盘，则该节点可能正常运行且运行状况良好。

解决集群仲裁和脑裂问题的邮箱架构和基于磁盘的心跳方法是 ONTAP Select 的多节点变体需要四个独立节点或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型。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，说明该节点已启动且正在运行。在健康集群中的任何时间点，集群节点上的单个邮箱磁盘都具有从所有其他集群节点发布的消息。

连接到每个 Select 群集节点的是专门用于共享邮箱访问的虚拟磁盘。此磁盘称为中介邮箱磁盘，因为其主要功能是在发生节点故障或网络分区时充当群集中介的方法。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 网络装载。这些节点定期将运行状况状态发布到邮箱磁盘的相应分区。通过使用遍布整个集群的网络可访问邮箱磁盘，您可以通过可访问性矩阵推断节点健康状况。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 很可能已关闭或网络隔离，应被接管。

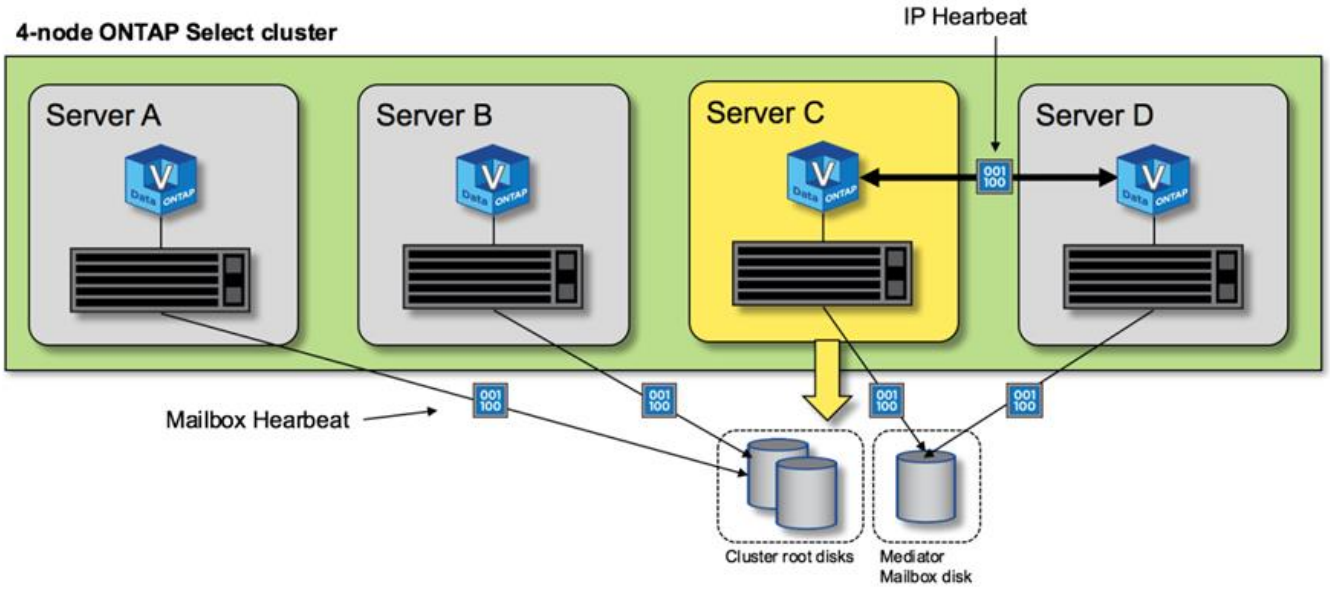
HA 心跳

与 NetApp FAS 平台一样，ONTAP Select 定期通过 HA 互连发送 HA 心跳消息。在 ONTAP Select 集群中，这是通过 HA 合作伙伴之间存在的 TCP/IP 网络连接执行的。此外，基于磁盘的心跳消息将传递到所有 HA 邮箱磁盘，包括中介邮箱磁盘。这些消息每隔几秒钟传递一次，并定期读回。发送和接收这些数据的频率允许 ONTAP Select 集群在大约 15 秒内检测到 HA 故障事件，这与 FAS 平台上可用的窗口相同。当心跳消息不再被读取时，将触发故障转移事件。

下图显示了从单个 ONTAP Select 集群节点（节点 C）的角度通过 HA 互连和调解器磁盘发送和接收心跳消息的过程。

网络心跳通过 HA 互连发送到 HA 合作伙伴节点 D，而磁盘心跳在所有群集节点 A、B、C 和 D 上使用邮箱磁盘。

四节点集群中的 **HA** 心跳：稳定状态



HA 故障转移和交还

在故障转移操作期间，幸存的节点使用其 HA 合作伙伴数据的本地副本为其对等节点承担数据服务责任。客户端 I/O 可以不间断地继续，但必须先复制对此数据的更改，然后才能进行回馈。请注意，ONTAP Select 不支持强制回馈，因为这会导致存储在幸存节点上的更改丢失。

当重新启动的节点重新加入集群时，会自动触发同步回操作。同步回所需的时间取决于几个因素。这些因素包括必须复制的更改数量、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，需要在同步回之后进行手动交还。可以使用以下命令监控同步回的进度：

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