高可用性架构
ONTAP Select

NetApp
January 29, 2026

目录

高可用性架构

ONTAP Select高可用性配置

探索高可用性选项，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运行在商用硬件上的基于软件的解决方案，但对弹性和容错能力的期望和需求并未改变。提供零恢复点目标 (RPO) 的高可用性 (HA) 解决方案可以保护客户免受基础架构堆栈中任何组件故障导致的数据丢失。

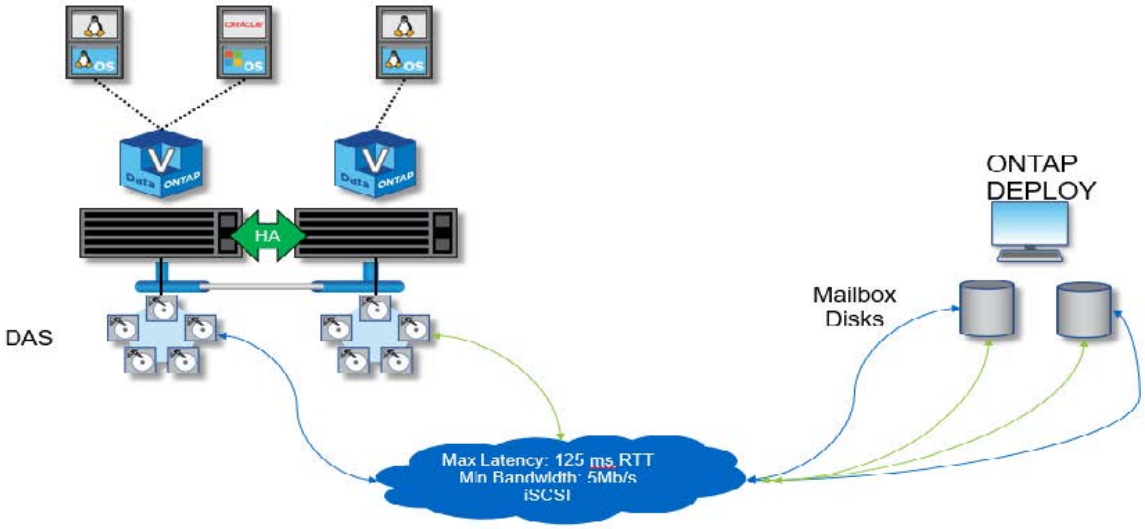
很大一部分 SDS 市场建立在无共享存储的概念之上，软件复制通过在不同的存储孤岛中存储多个用户数据副本来提供数据弹性。ONTAPONTAP Select基于此前提构建，使用ONTAP提供的同步复制功能 (RAID SyncMirror) 在集群内存储额外的用户数据副本。这发生在 HA 对的环境中。每个 HA 对都存储两个用户数据副本：一个在本地节点提供的存储上，另一个在 HA 合作伙伴提供的存储上。在ONTAP Select集群中，HA 和同步复制绑定在一起，两者的功能不能分离或独立使用。因此，同步复制功能仅在多节点产品中可用。

> 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。

ONTAP Select HA 部署模型有两种：多节点集群（四节点、六节点或八节点）和双节点集群。双节点ONTAP Select集群的显著特点是使用外部调解器服务来解决脑裂问题。ONTAPDeployONTAP机充当其配置的所有双节点 HA 对的默认调解器。

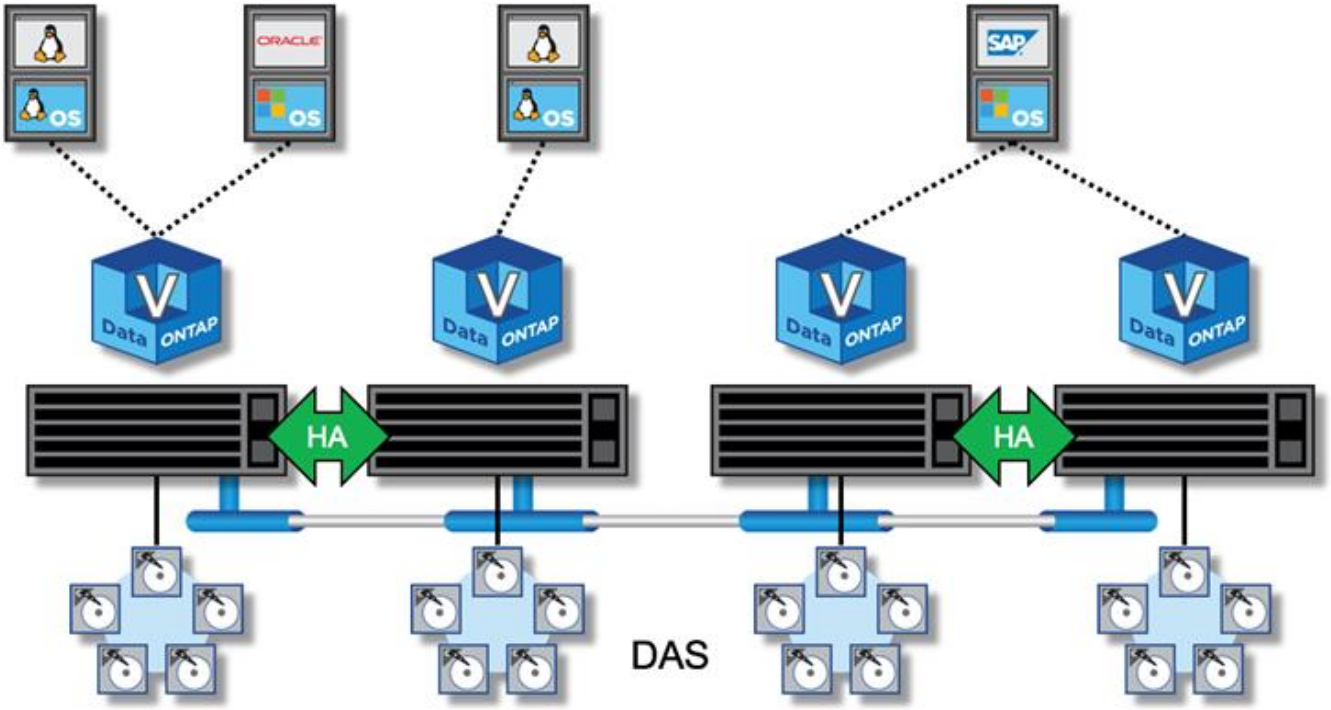
下图表示了这两种架构。

带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同步镜像，因此即使发生故障转移，也不会丢失数据。

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同步镜像，因此即使发生故障转移，也不会丢失数据。
- 使用 DAS 存储时，物理服务器上只能存在一个ONTAP Select实例。ONTAPONTAP Select需要对系统的本地 RAID 控制器进行非共享访问，并且旨在管理本地连接的磁盘，而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 网络存在单点故障 (SPOF)，因此，防止网络分区和脑裂情况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更多幸存节点建立。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 中介服务之间的心跳网络流量极小且具有弹性，因此ONTAP Deploy VM 可以托管在与ONTAP Select双节点集群不同的数据中心。

> ONTAP Deploy 虚拟机在充当双节点集群的调解器时，将成为该集群不可或缺的一部分。如果调解器服务不可用，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。为了确保集群仲裁正常运行，最低带宽要求为 5Mbps，最大往返时间 (RTT) 延迟要求为 125 毫秒。

如果充当调解器的ONTAP Deploy 虚拟机暂时或可能永久不可用，则可以使用辅助ONTAP Deploy 虚拟机来恢复双节点集群仲裁。这会导致新的ONTAP Deploy 虚拟机无法管理ONTAP Select节点，但可以成功参与集群仲裁算法。ONTAPONTAP Select节点与ONTAP Deploy 虚拟机之间的通信是通过 IPv4 上的 iSCSI 协议完成的。ONTAPONTAP Select节点管理 IP 地址是启动器， ONTAP Deploy 虚拟机 IP 地址是目标。因此，在创建双节点集群时，节点管理 IP 地址无法支持 IPv6 地址。在创建双节点集群时，系统会自动创建ONTAP Deploy 托管的邮箱磁盘，并将其屏蔽为正确的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建议您在每个双节点集群实例化后备份ONTAP Deploy 数据库。

双节点 HA 与双节点延伸 HA (MetroCluster SDS)

可以将双节点主动/主动 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）之间的唯一区别在于节点之间的网络连接距离。

双节点集群定义为两个节点位于同一数据中心且相距 300 米以内的集群。通常，两个节点都具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 是指节点物理上相距超过 300 米（不同房间、不同建筑和不同数据中心）的集群。此外，每个节点的上行链路连接分别连接到单独的网络交换机。MetroClusterMetroCluster不需要专用硬件。但是，该环境应满足延迟（RTT 最大 5 毫秒，抖动最大 5 毫秒，总计 10 毫秒）和物理距离（最大 10 公里）的要求。

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。PremiumXL 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虚拟机与共享存储之间的网络，vNAS 配置通常具有较高的固有延迟。MetroClusterSDS 配置必须在节点之间提供最多 10 毫秒的延迟，其中包括共享存储延迟。换句话说，仅测量 Select 虚拟机之间的延迟是不够的，因为对于这些配置来说，共享存储延迟不可忽略。

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高可用性 (HA) 模型建立在高可用性合作伙伴的概念之上。ONTAPONTAP Select将此架构扩展到非共享商用服务器领域，利用ONTAP中提供的 RAID SyncMirror (RSM) 功能在集群节点之间复制数据块，从而在高可用性 (HA) 对中提供两个用户数据副本。

带有调解器的双节点集群可以跨越两个数据中心。有关更多信息，请参阅"双节点扩展 HA (MetroCluster SDS) 最佳实践" 。

镜像聚合

ONTAP Select集群由 2 到 8 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。此镜像对用户透明，并且是数据聚合的一个属性，会在数据聚合创建过程中自动配置。

ONTAP Select集群中的所有聚合都必须进行镜像，以便在发生节点故障转移时确保数据可用性，并在发生硬件故障时避免出现单点故障 (SPOF)。ONTAPONTAP Select集群中的聚合基于 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 伙伴提供）

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（分别表示本地镜像对和远程镜像对）。实际的丛编号在您的安装中可能有所不同。

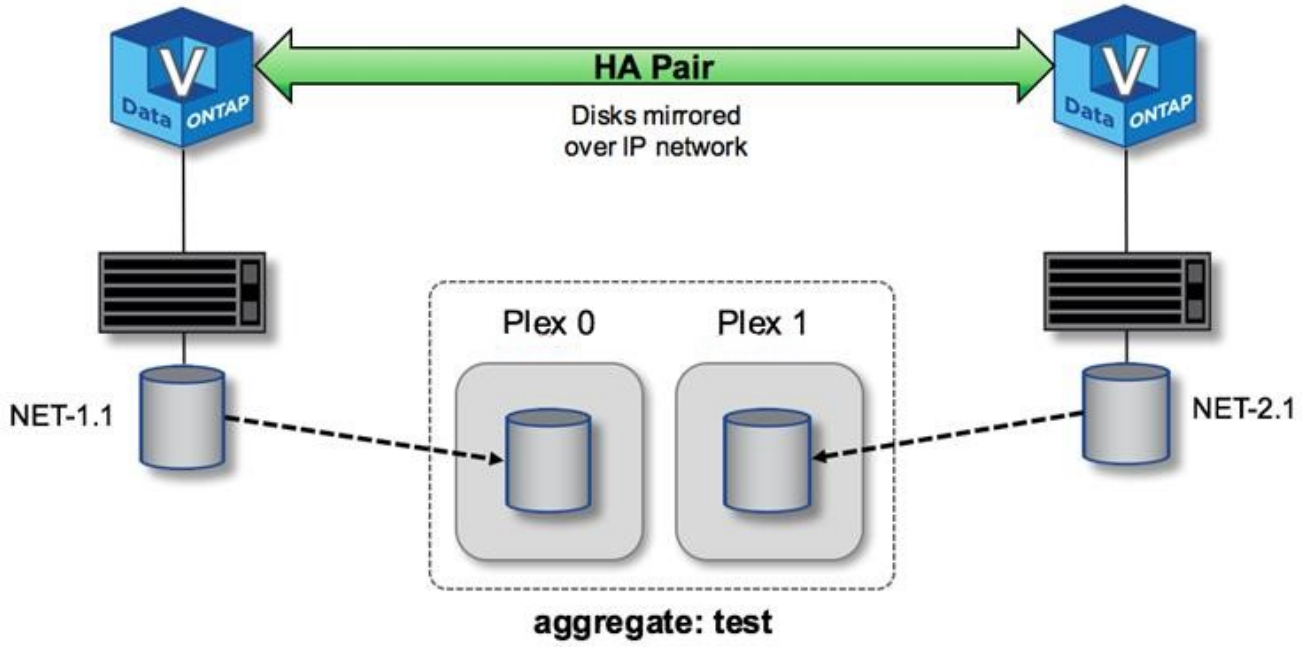
这种方法与标准ONTAP集群的工作方式有着根本的不同。这适用于ONTAP Select集群中的所有根磁盘和数据磁盘。聚合包含数据的本地副本和镜像副本。因此，包含 N 个虚拟磁盘的聚合可提供相当于 N/2 个磁盘的唯一存储，因为第二个数据副本位于其自己的唯一磁盘上。

下图显示了四节点ONTAP Select集群中的 HA 对。此集群中有一个聚合（测试），它使用来自两个 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 (Plex 0) 提供；一组远程磁盘，由故障转移配对节点 (Plex 1) 提供。

Plex 0 是用于存放所有本地磁盘的存储桶。Plex1 是用于存放镜像磁盘（即负责存储用户数据第二个复制副本的磁盘）的存储桶。拥有聚合的节点会将磁盘提供给 Plex 0，而该节点的 HA 配对节点会将磁盘提供给 Plex 1。

下图中有一个包含两个磁盘的镜像聚合。此聚合的内容在两个集群节点之间进行镜像，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，聚合测试归左侧的集群节点所有，并使用本地磁盘 NET-1.1 和 HA 伙伴镜像磁盘 NET-2.1。

- ONTAP Select镜像聚合*



部署ONTAP Select集群时，系统上的所有虚拟磁盘都会自动分配给正确的 Plex，无需用户执行任何与磁盘分配相关的额外步骤。这可以防止磁盘意外分配给错误的 Plex，并提供最佳的镜像磁盘配置。

写入路径

集群节点之间的数据块同步镜像以及系统故障时不丢失数据的要求对传入写入操作在ONTAP Select集群中传播时所采用的路径有重大影响。此过程包含两个阶段：

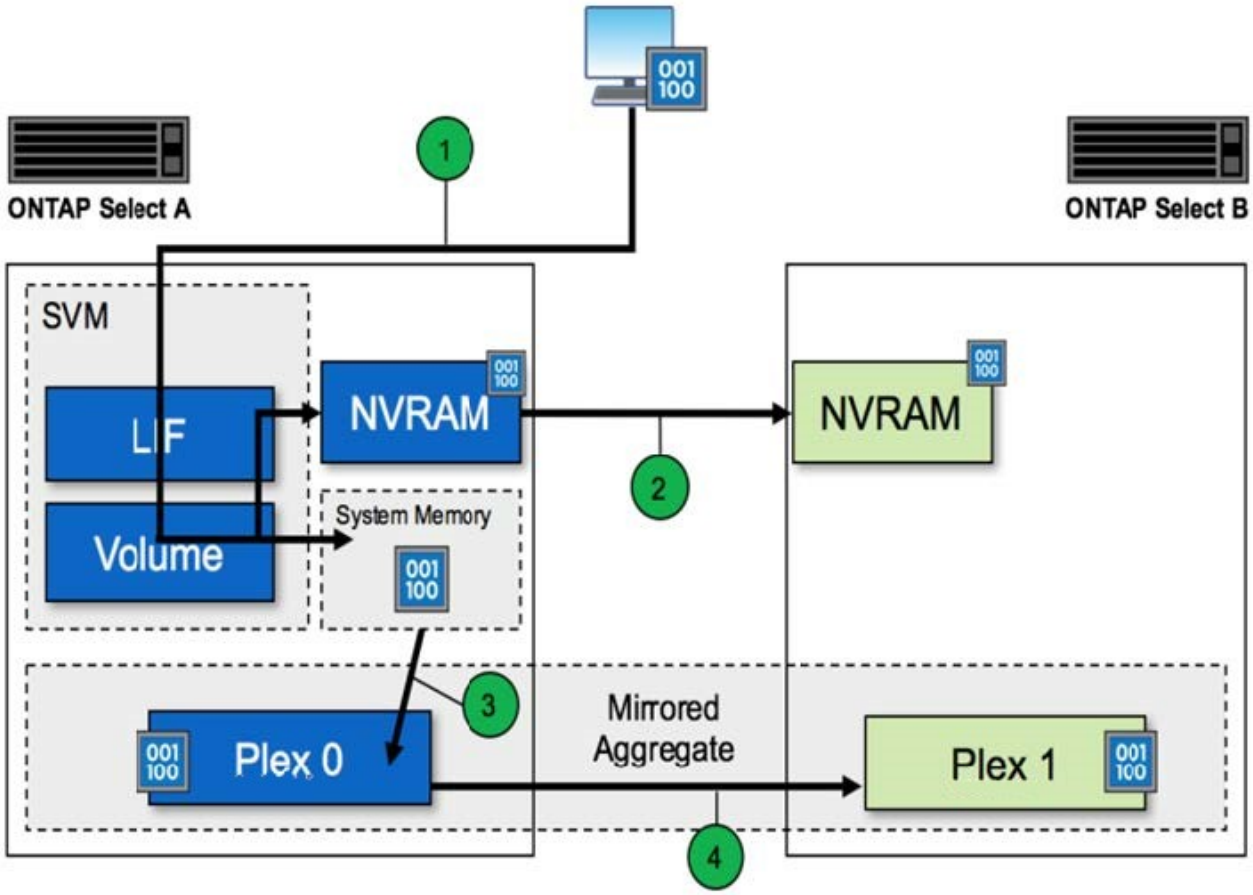
- 致谢
- 降级

对目标卷的写入操作通过数据 LIF 进行，并提交到ONTAP Select节点系统磁盘上的虚拟化NVRAM分区，然后再向客户端确认。在 HA 配置中，还会执行一个额外的步骤，因为这些NVRAM写入操作会在确认之前立即镜像到目标卷所有者的 HA 配对节点。此过程可确保在原始节点发生硬件故障时，HA 配对节点上的文件系统保持一致。

将写入内容提交至NVRAM后， ONTAP会定期将此分区的内容移动到相应的虚拟磁盘，此过程称为降级转储。此过程仅在拥有目标卷的集群节点上发生一次，不会在 HA 配对节点上发生。

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：

- 写入通过ONTAP Select节点 A 拥有的逻辑接口进入系统。
- 写入操作提交到节点 A 的NVRAM并镜像到 HA 伙伴节点 B。
- 当两个 HA 节点都出现 I/O 请求后，该请求就会被确认回客户端。

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

- 写入操作从虚拟NVRAM转入虚拟数据聚合。
- 镜像引擎同步将块复制到两个 plex。

ONTAP Select HA 增强数据保护

高可用性 (HA) 磁盘心跳、HA 邮箱、HA 心跳、HA 故障转移和回馈功能可增强数据保护。

磁盘心跳

尽管ONTAP Select HA 架构充分利用了传统FAS阵列的许多代码路径，但也存在一些例外。其中一个例外是基于磁盘的心跳机制的实现，这是一种非基于网络的通信方法，集群节点使用这种方法来防止网络隔离导致裂脑行为。裂脑场景是集群分区的结果，通常由网络故障引起，导致集群两端都认为对方已宕机并试图接管集群资源。

企业级 HA 实施必须妥善处理此类情况。ONTAP通过一种基于磁盘的定制心跳机制来实现这一点ONTAP邮箱负责处理这一任务，它是集群节点用于传递心跳消息的物理存储位置。这有助于集群确定连接性，从而在发生故障转移时确定仲裁。

在使用共享存储 HA 架构的FAS阵列上， ONTAP通过以下方式解决裂脑问题：

- SCSI 永久预留
- 持久 HA 元数据
- 通过 HA 互连发送 HA 状态

然而，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，而看不到 HA 配对节点的本地存储。因此，当网络分区隔离 HA 对的每一侧时，上述确定集群仲裁和故障转移行为的方法将不可用。

虽然现有的裂脑检测和避免方法无法使用，但仍然需要一种能够适应无共享环境约束的调解方法。ONTAPONTAP Select进一步扩展了现有的邮箱基础架构，使其能够在网络分区时充当调解方法。由于共享存储不可用，因此调解是通过 NAS 访问邮箱磁盘来完成的。这些磁盘分布在整个集群中，包括双节点集群中的调解器，并使用 iSCSI 协议。因此，集群节点可以根据对这些磁盘的访问做出智能故障转移决策。如果一个节点可以访问其 HA 伙伴节点之外的其他节点的邮箱磁盘，则该节点很可能已启动且运行正常。

> 解决集群仲裁和裂脑问题的邮箱架构和基于磁盘的心跳方法是ONTAP Select多节点变体需要四个独立节点或双节点集群的调解器的原因。

HA邮箱发帖

HA 邮箱架构采用消息发布模型。集群节点会定期向集群中所有其他邮箱磁盘（包括中介节点）发布消息，表明该节点已启动并正在运行。在正常运行的集群中，任何时间点，集群节点上的单个邮箱磁盘都会收到来自所有其他集群节点的消息。

每个 Select 集群节点都附加一个虚拟磁盘，专门用于共享邮箱访问。此磁盘被称为中介邮箱磁盘，因为其主要功能是在发生节点故障或网络分区时充当集群中介。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网络可访问邮箱磁盘，您可以通过可达性矩阵推断节点的运行状况。例如，集群节点 A 和 B 可以向集群节点 D 的邮箱发送邮件，但不能向节点 C 的邮箱发送邮件。此外，集群节点 D 无法向节点 C 的邮箱发送邮件，因此节点 C 很可能已关闭或网络隔离，应该被接管。

HA 心跳

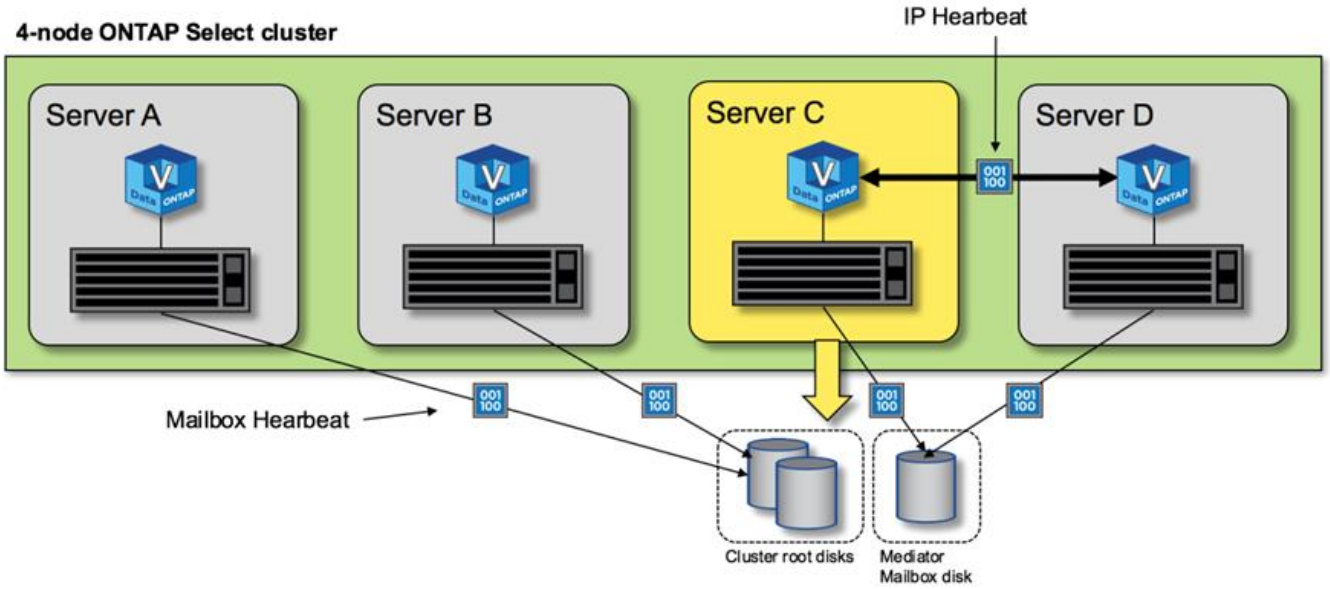
与NetApp FAS平台一样， ONTAP Select会定期通过 HA 互连发送 HA 心跳消息。在ONTAP Select集群中，此操作通过 HA 伙伴节点之间的 TCP/IP 网络连接执行。此外，基于磁盘的心跳消息会传递到所有 HA 邮箱磁盘，包括中介邮箱磁盘。这些消息每隔几秒传递一次，并定期读取。如此高的发送和接收频率使ONTAP Select集群

能够在大约 15 秒内检测到 HA 故障事件，这与FAS平台上的可用时间窗口相同。当不再读取心跳消息时，将触发故障转移事件。

下图从单个ONTAP Select集群节点（节点 C）的角度显示了通过 HA 互连和调解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 伙伴节点 D，而磁盘心跳使用跨所有集群节点 A、B、C 和 D 的邮箱磁盘。

四节点集群中的 **HA** 心跳：稳定状态



HA 故障转移和恢复

在故障转移操作期间，幸存节点将使用其 HA 伙伴节点数据的本地副本承担其对等节点的数据服务责任。客户端 I/O 可以继续不间断运行，但必须先复制对此数据的更改，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幸存节点上的更改丢失。

重启的节点重新加入集群时，会自动触发同步恢复操作。同步恢复所需的时间取决于多种因素。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。同步恢复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，需要在同步恢复后进行手动交还。您可以使用以下命令监控同步恢复的进度：

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