高可用性架构
ONTAP Select

NetApp
May 21, 2024

目录

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。

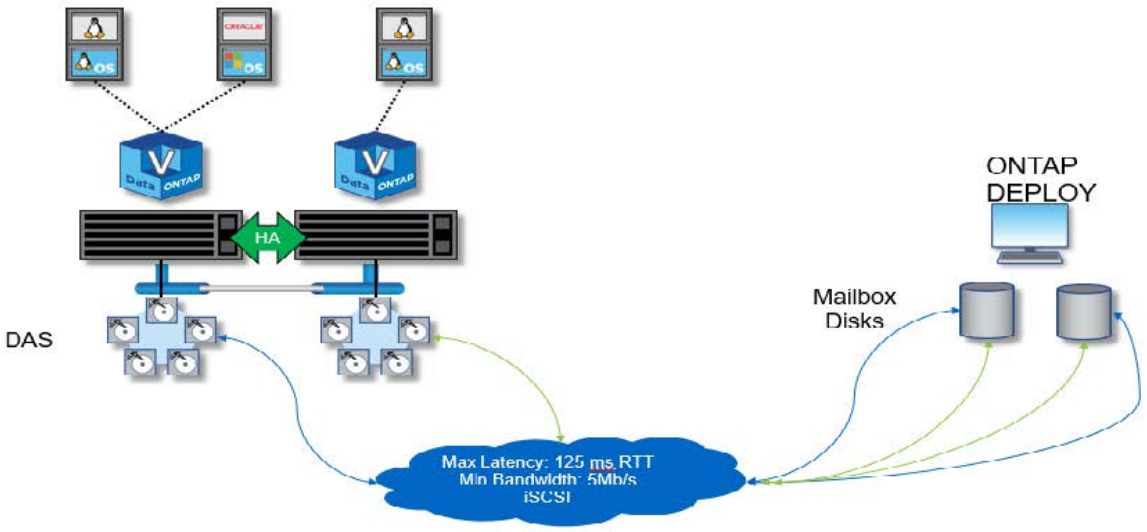
SDS 市场的很大一部分是基于无共享存储的概念构建的，软件复制可通过在不同存储孤岛之间存储多个用户数据副本来提供数据故障恢复能力。ONTAP Select 在此前提下构建，可使用 ONTAP 提供的同步复制功能（ RAID SyncMirror ）在集群中存储一份额外的用户数据副本。此问题发生在 HA 对的上下文中。每个 HA 对都会存储两个用户数据副本：一个位于本地节点提供的存储上，一个位于 HA 配对节点提供的存储上。在 ONTAP Select 集群中， HA 和同步复制绑定在一起，两者的功能不能分离或单独使用。因此，同步复制功能仅在多节点产品中可用。

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。

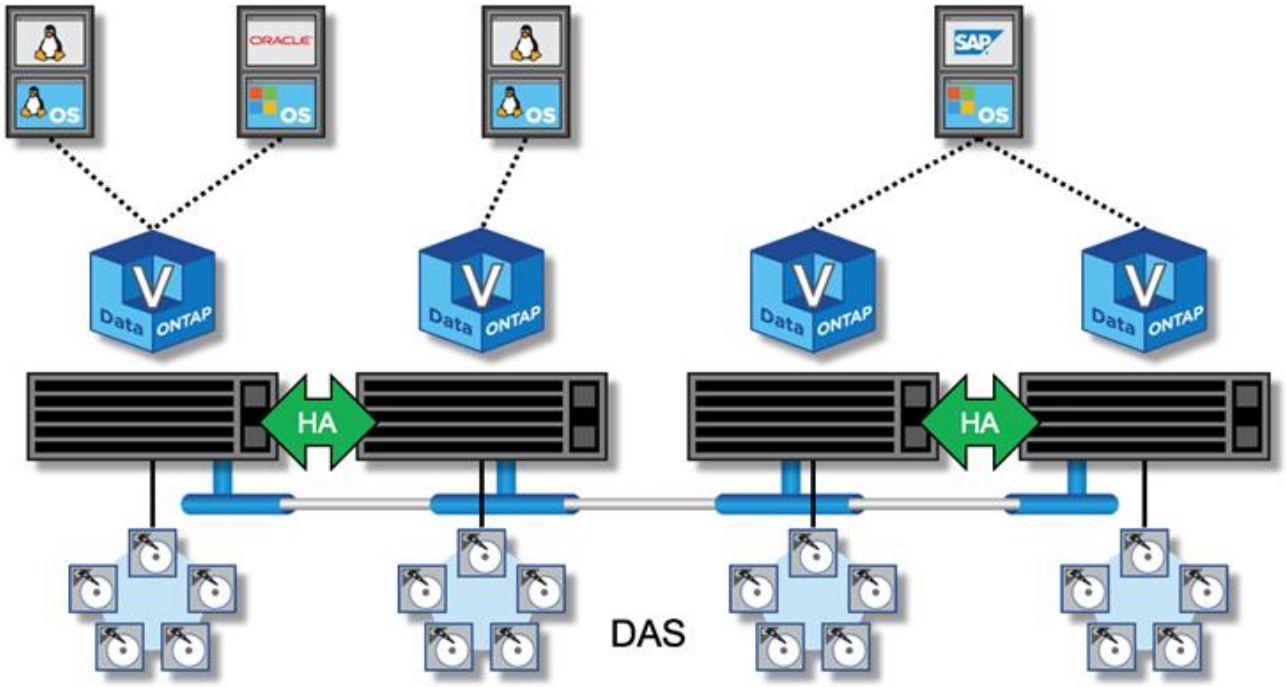
下图显示了这两种架构。

- 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。

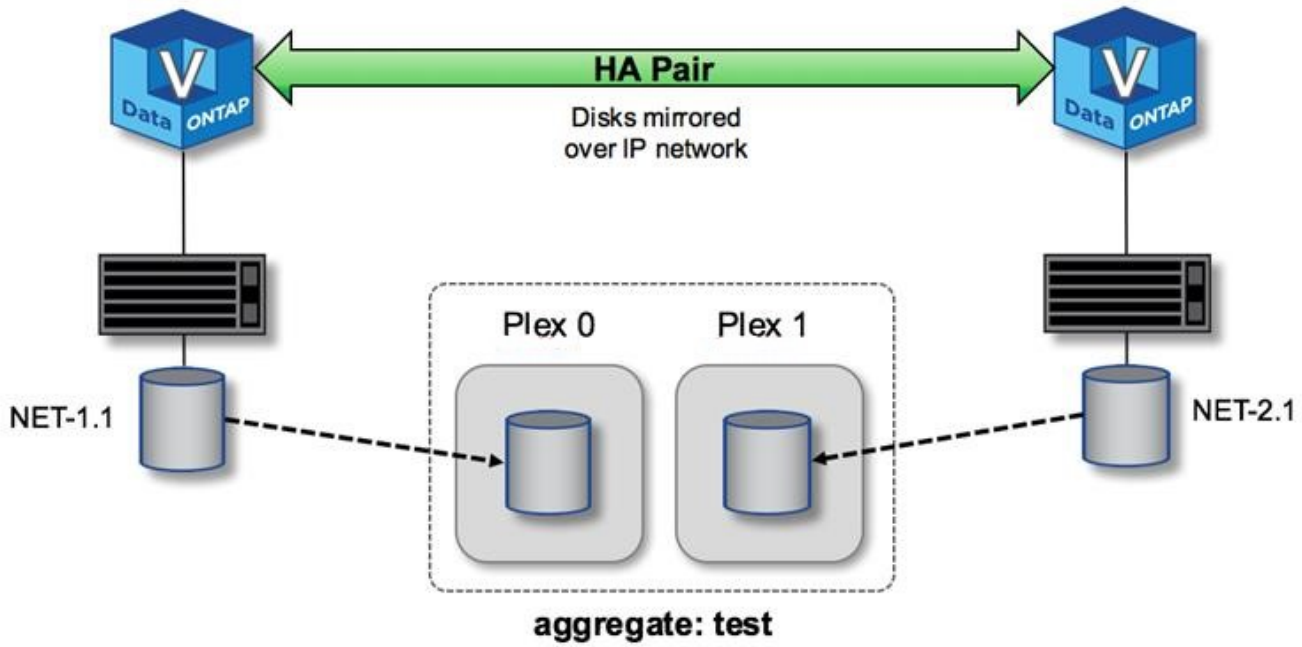
这种方法与标准 ONTAP 集群的工作方式有着根本的不同。此适用场景 将对 ONTAP Select 集群中的所有根磁盘和数据磁盘执行。聚合同时包含数据的本地副本和镜像副本。因此，包含 N 个虚拟磁盘的聚合可提供相当于 N/2 个磁盘的唯一存储，因为第二个数据副本驻留在其自身的唯一磁盘上。

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：

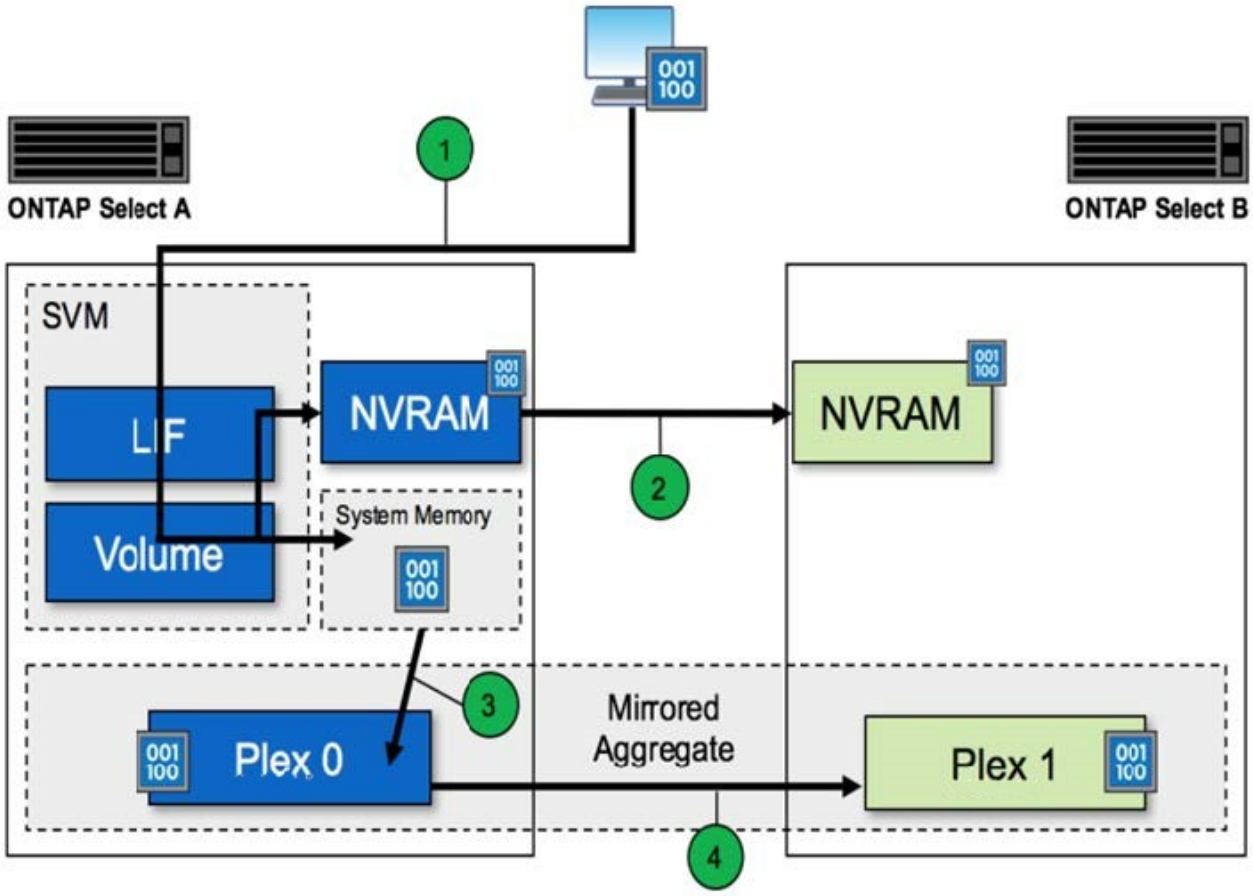
- 确认
- 转存

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

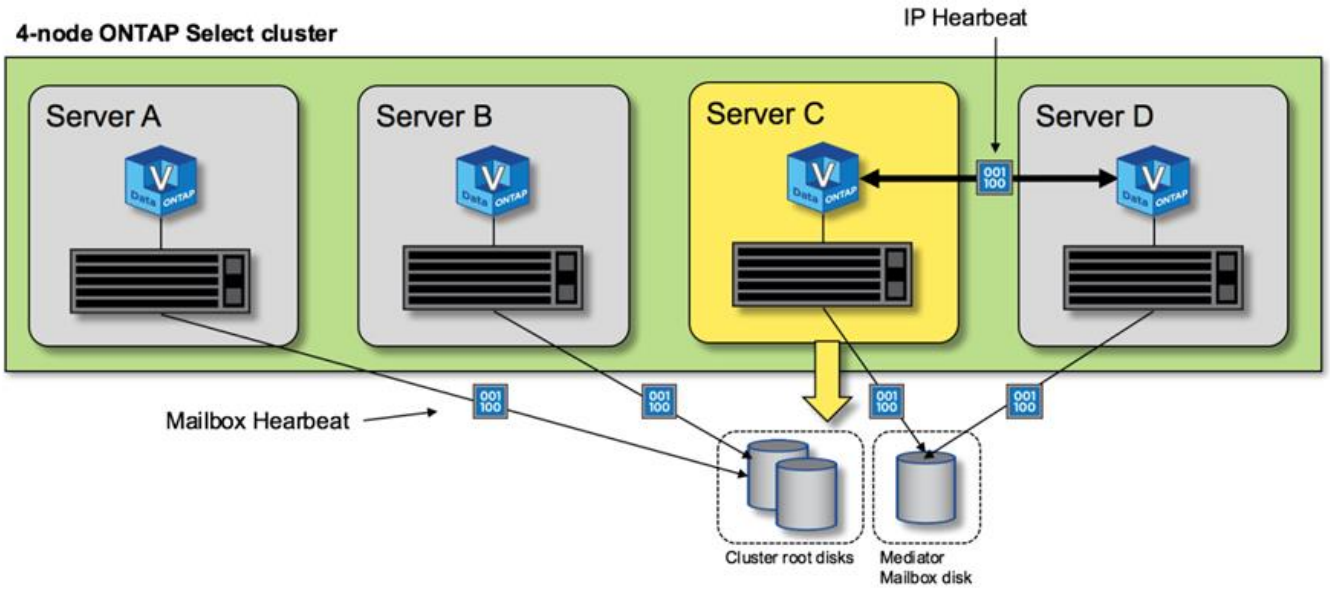
与 NetApp FAS 平台一样， ONTAP Select 会定期通过 HA 互连发送 HA 检测信号消息。在 ONTAP Select 集群中，此操作通过 HA 配对节点之间的 TCP/IP 网络连接来执行。此外，基于磁盘的检测信号消息会传递到所有

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>
```