网络配置：
Enterprise applications

NetApp
May 09, 2024

目录

网络配置：

Oracle数据库的逻辑接口设计

Oracle数据库需要访问存储。逻辑接口(Logical Interface、Logical Interface、Logical Interface)是将Storage Virtual Machine (SVM)连接到网络并进而连接到数据库的网络管道。要确保每个数据库工作负载都有足够的带宽、并且故障转移不会导致存储服务丢失、需要正确的LIF设计。

本节概述了LIF的主要设计原则。有关更全面的文档、请参见 "ONTAP网络管理文档"。与数据库架构的其他方面一样、Storage Virtual Machine (SVM、在CLI中称为Vserver)和逻辑接口(LIF)设计的最佳选项在很大程度上取决于扩展要求和业务需求。

在制定LIF策略时、请考虑以下主要主题：

- *性能。*网络带宽是否足够?
- *故障恢复能力。*设计中是否存在单点故障?
- *易管理性。*网络能否无干扰地扩展?

这些主题适用于从主机到交换机再到存储系统的端到端解决方案。

LIF类型

LIF类型有多种。 "有关LIF类型的ONTAP文档" 请提供有关此主题的更完整信息、但从功能角度来看、可以将这些生命周期表分为以下几组：

- *集群和节点管理Lifs.*用于管理存储集群的Lifs。
- * SVM管理LIF.*允许通过REST API或ONTAPI (也称为ZAPI)访问SVM的接口、用于执行创建快照或调整卷大小等功能。SnapManager for Oracle (SMO)等产品必须能够访问SVM管理LIF。
- 数据**Lifs.** FC、iSCSI、NVMe/FC、NVMe/TCP、NFS接口 或SMB/CCIFS数据。

> 用于NFS流量的数据LIF也可通过从更改防火墙策略来进行管理 `data` to `mgmt` 或其他允许HTTP、HTTPS或SSH的策略。此更改可以避免配置每个主机以访问NFS数据LIF和单独的管理LIF、从而简化网络配置。无法为iSCSI和管理流量配置接口、尽管两者都使用IP协议。在iSCSI环境中、需要使用单独的管理LIF。

SAN LIF设计

SAN环境中的LIF设计相对简单、原因之一是：多路径。所有现代SAN实施都允许客户端通过多个独立的网络路径访问数据、并选择最佳访问路径。因此、与LIF设计相关的性能更易于解决、因为SAN客户端会自动在最佳可用路径之间对I/O进行负载平衡。

如果某个路径不可用、则客户端会自动选择其他路径。由此带来的设计精简性使SAN的生命周期通常更易于管理。这并不意味着SAN环境始终可以更轻松地进行管理、因为SAN存储的许多其他方面都比NFS复杂得多。这只是意味着SAN LIF的设计更简单。

性能

在SAN环境中、LIF性能最重要的考虑因素是带宽。例如、每个节点具有两个16 Gb FC端口的双节点ONTAP AFF集群允许每个节点之间最多32 Gb的带宽。

故障恢复能力

SAN AFF不会在SAN存储系统上进行故障转移。如果SAN LIF因控制器故障转移而失败、则客户端的多路径软件会检测到路径丢失、并将I/O重定向到其他LIF。对于ASA存储系统、将在短暂延迟后对lifs进行故障转移、但这不会中断IO、因为其他控制器上已存在活动路径。执行故障转移过程是为了恢复所有定义端口上的主机访问。

易管理性

在NFS环境中、LIF迁移是一项更为常见的任务、因为LIF迁移通常与在集群中重新定位卷相关联。在HA对中重新定位卷后、无需在SAN环境中迁移LIF。这是因为、在卷移动完成后、ONTAP会向SAN发送有关路径更改的通知、并且SAN客户端会自动重新优化。使用SAN迁移LIF主要与重大物理硬件更改相关。例如、如果需要无中断升级控制器、则会将SAN LIF迁移到新硬件。如果发现FC端口出现故障、则可以将LIF迁移到未使用的端口。

设计建议

NetApp提出以下建议：

- 请勿创建超出所需数量的路径。路径数量过多会使整体管理变得更加复杂、并且某些主机上的路径故障转移可能会出现发生原因问题。此外、对于SAN启动等配置、某些主机存在意外的路径限制。
- 很少有配置需要一个LUN具有四个以上的路径。如果有两个以上的节点向LUN公布路径、则其价值会受到限制、因为如果拥有LUN的节点及其HA配对节点发生故障、则托管LUN的聚合将无法访问。在这种情况下、在主HA对以外的节点上创建路径毫无用处。
- 虽然可以通过选择要包含在FC分区中的端口来管理可见LUN路径的数量、但通常在FC分区中包含所有潜在目标点并在ONTAP级别控制LUN可见性会更容易。
- 在ONTAP 8.3及更高版本中、默认使用选择性LUN映射(SLM)功能。通过SLM、任何新的LUN都会自动从拥有底层聚合的节点以及该节点的HA配对节点公布。这种安排无需创建端口集或配置分区来限制端口可访问性。为了获得最佳性能和故障恢复能力、每个LUN可在所需的最少节点上使用。
*如果必须将LUN迁移到两个控制器之外、则可以使用添加其他节点 `lun mapping add-reporting-nodes` 命令、以便在新节点上公布LUN。这样会为LUN创建更多SAN路径以进行LUN迁移。但是、主机必须执行发现操作才能使用新路径。
- 不要过分关注间接流量。在I/O密集型环境中、最好避免间接流量、因为在这种环境中、每微秒的延迟都至关重要、但对于典型工作负载、可见的性能影响可以忽略不计。

NFS LIF设计

与SAN协议不同、NFS定义多个数据路径的能力有限。NFSv4的并行NFS (pNFS)扩展解决了这一限制、但由于以太网速度已达到100 GB甚至超过100 GB、因此添加额外路径很少有价值。

性能和故障恢复能力

虽然衡量SAN LIF性能主要是计算所有主路径的总带宽、但要确定NFS LIF性能、需要更深入地了解确切的网络配置。例如、可以将两个10 Gb端口配置为原始物理端口、也可以将其配置为链路聚合控制协议(Link Aggregation Control Protocol、LACP)接口组。如果将其配置为接口组、则可以使用多个负载平衡策略、这些策略的工作方式会有所不同、具体取决于流量是交换流量还是路由流量。最后、Oracle Direct NFS (DNFS)提供了目前在任何操作系统NFS客户端中都不存在的负载平衡配置。

与SAN协议不同、NFS文件系统要求在协议层具有故障恢复能力。例如、LUN始终配置为启用多路径、这意味着存储系统可以使用多个冗余通道、每个通道都使用FC协议。另一方面、NFS文件系统取决于单个TCP/IP通道的可用性、该通道只能在物理层进行保护。这种安排就是为什么存在端口故障转移和LACP端口聚合等选项的原因。

在NFS环境中、性能和故障恢复能力均在网络协议层提供。因此，这两个主题是相互交织的，必须一起讨论。

将**LIP**绑定到端口组

要将LIF绑定到端口组、请将LIF IP地址与一组物理端口相关联。将物理端口聚合在一起的主要方法是LACP。LACP的容错功能相当简单；LACP组中的每个端口都会受到监控、并在发生故障时从端口组中删除。但是、对于LACP在性能方面的工作原理、存在许多误解：

- LACP不要求交换机上的配置与端点匹配。例如、可以为ONTAP配置基于IP的负载平衡、而交换机可以使用基于MAC的负载平衡。
- 使用LACP连接的每个端点都可以独立选择数据包传输端口、但不能选择用于接收的端口。这意味着、从ONTAP到特定目标的流量会绑定到特定端口、而返回流量可能会到达其他接口。但是、这不会造成发生原因问题。
- LACP不会始终均匀分布流量。在具有许多NFS客户端的大型环境中、结果通常甚至会使用LACP聚合中的所有端口。但是、环境中的任何一个NFS文件系统都仅限于一个端口的带宽、而不是整个聚合的带宽。
- 尽管ONTAP上提供了robin-robin LACP策略、但这些策略不会处理从交换机到主机的连接。例如、如果配置中的一个主机上有一个四端口LACP中继、而ONTAP上有一个四端口LACP中继、则仍然只能使用一个端口读取文件系统。虽然ONTAP可以通过所有四个端口传输数据、但目前尚无可通过所有四个端口从交换机发送到主机的交换机技术。仅使用一个。

在包含许多数据库主机的大型环境中、最常见的方法是使用IP负载平衡构建一个包含适当数量10 Gb (或更快)接口的LACP聚合。通过这种方法、只要存在足够多的客户端、ONTAP就可以均匀地使用所有端口。如果配置中的客户端较少、则负载平衡会中断、因为LACP中继不会动态重新分配负载。

建立连接后、特定方向的流量仅会放置在一个端口上。例如、对通过四端口LACP中继连接的NFS文件系统执行完整表扫描的数据库仅通过一个网络接口卡(Network Interface Card、NIC)读取数据。如果在此类环境中只有三个数据库服务器、则这三个服务器都可能从同一端口读取数据、而其他三个端口则处于空闲状态。

将**Lifs**绑定到物理端口

将LIF绑定到物理端口可以更精细地控制网络配置、因为ONTAP系统上的给定IP地址一次只与一个网络端口相关联。然后、可通过配置故障转移组和故障转移策略来实现故障恢复能力。

故障转移策略和故障转移组

故障转移策略和故障转移组控制了在网络中断期间的故障转移。配置选项已随ONTAP的不同版本而发生更改。请参见 "有关故障转移组和策略的ONTAP网络管理文档" 有关要部署的ONTAP版本的具体详细信息、请参见。

ONTAP 8.3及更高版本支持基于广播域管理LIF故障转移。因此、管理员可以定义可访问给定子网的所有端口、并允许ONTAP选择适当的故障转移LIF。某些客户可以使用这种方法、但由于缺乏可预测性、在高速存储网络环境中这种方法存在一些限制。例如、一个环境可以包括用于例行文件系统访问的1 Gb端口和用于数据文件I/O的10 Gb端口如果两种类型的端口都位于同一广播域中、则LIF故障转移可能会导致数据文件I/O从10 Gb端口移动到1 Gb端口。

概括地说、请考虑以下做法：

1. 将故障转移组配置为用户定义的组。

2. 使用存储故障转移(SFR)配对控制器上的端口填充故障转移组、以便在存储故障转移期间、这些LUN跟随聚合。这样可以避免产生间接流量。
3. 使用性能特征与原始LIF匹配的故障转移端口。例如、单个10 Gb物理端口上的LIF应包含一个具有单个10 Gb端口的故障转移组。一个四端口LACP LIF应故障转移到另一个四端口LACP LIF。这些端口将是广播域中定义的端口的子集。
4. 将故障转移策略设置为仅SFo-Partner。这样可以确保LIF在故障转移期间跟随聚合。

自动还原

设置 `auto-revert` 参数。大多数客户倾向于将此参数设置为 `true` 以使LIF还原到其主端口。但是、在某些情况下、客户会将此值设置为`false `so、以便在将LIF返回到其主端口之前可以调查意外故障转移。

LIF与卷的比率

一个常见的误解是、卷和NFS Sifs之间必须有1：1的关系。虽然要在集群中的任何位置移动卷而不创建额外的互连流量、都需要使用此配置、但这绝对不是一项要求。必须考虑集群间流量、但仅存在集群间流量并不会造成问题。为ONTAP创建的许多已发布基准主要包括间接I/O

例如、如果某个数据库项目包含的性能关键型数据库数量相对较少、并且总共只需要40个卷、则可能需要采用卷到LIF的1：1策略、这种安排需要40个IP地址。然后、可以将任何卷与关联的LIF一起移动到集群中的任何位置、流量将始终是直接的、即使是微秒级的延迟、也可以最大限度地减少每个源。

作为一个反例、客户与LI之间的1：1关系可能更易于管理大型托管环境。随着时间的推移、卷可能需要迁移到其他节点、这会对一些间接流量进行发生原因。但是、除非互连交换机上的网络端口饱和、否则不会检测到性能影响。如果存在问题、可以在其他节点上建立新的LIF、并可在下一个维护窗口更新主机、以便从配置中删除间接流量。

Oracle数据库的TCP/IP和以太网配置

许多基于ONTAP的Oracle客户使用以太网、即NFS、iSCSI、NVMe/TCP的网络协议、尤其是云。

主机操作系统设置

大多数应用程序供应商文档都包含特定的TCP和以太网设置、旨在确保应用程序以最佳状态运行。这些相同的设置通常足以提供基于IP的最佳存储性能。

以太网流量控制

此技术允许客户端请求发送方暂时停止数据传输。这通常是因为接收方无法足够快速地处理传入数据。一次、请求发送方停止传输比让接收方丢弃数据包造成的中断要少、因为缓冲区已满。如今、操作系统中使用的TCP堆栈已不再是这种情况。事实上、流量控制造成的问题比它解决的问题多。

近年来、以太网流量控制导致的性能问题不断增加。这是因为以太网流量控制在物理层运行。如果网络配置允许任何主机操作系统向存储系统发送以太网流量控制请求、则会导致所有已连接客户端的I/O暂停。由于单个存储控制器为越来越多的客户端提供服务、因此其中一个或多个客户端发送流量控制请求的可能性会增加。在广泛的操作系统虚拟化过程中、客户站点经常会出现此问题。

NetApp系统上的NIC不应接收流量控制请求。根据网络交换机制造商的不同、实现此结果的方法也会有所不同。在大多数情况下、可以将以太网交换机上的流量控制设置为 `receive desired` 或 `receive on`，表示流量控

制请求不会转发到存储控制器。在其他情况下、存储控制器上的网络连接可能不允许禁用流量控制。在这些情况下、必须将客户端配置为从不发送流量控制请求、方法是更改主机服务器本身的NIC配置或主机服务器所连接的交换机端口。

* NetApp建议*确保NetApp存储控制器不接收以太网流量控制数据包。这通常可以通过设置控制器所连接的交换机端口来实现、但某些交换机硬件存在一些限制、可能需要在客户端进行更改。

MTU大小

事实证明、使用巨型帧可以减少CPU和网络开销、从而在一定程度上提高1 Gb网络的性能、但其优势通常并不明显。

* NetApp建议*尽可能实施巨型帧、以实现任何潜在的性能优势并使解决方案适应未来需求。

在10 Gb网络中使用巨型帧几乎是强制性要求。这是因为大多数10 Gb实施在达到10 Gb标记之前都会达到每秒数据包数限制、而不会出现巨型帧。使用巨型帧可以提高TCP/IP处理的效率、因为它允许操作系统、服务器、NIC和存储系统处理的数据包数量较少、但数量较大。不同NIC的性能提升各不相同、但性能提升幅度很大。

对于巨型帧实施、人们普遍认为所有连接的设备都必须支持巨型帧、并且MTU大小必须端到端匹配、但这种看法并不正确相反、在建立连接时、这两个网络端点会协商双方可接受的最高帧大小。在典型环境中、网络交换机的MTU大小设置为9216、NetApp控制器设置为9000、客户端设置为9000和1514的混合。可以支持9000 MTU的客户端可以使用巨型帧、而只支持1514的客户端可以协商较低的值。

在完全交换的环境中、这种安排的问题很少见。但是、在路由环境中请注意、不会强制任何中间路由器对巨型帧进行分段。

- NetApp建议*配置以下内容：
- 巨型帧是需要的、但对于1 Gb以太网(GbE)则不需要巨型帧。
- 要在10GbE和更快的速度下实现最高性能、需要巨型帧。

TCP参数

通常会有三种设置配置不当：TCP时间戳、选择性确认(SACK)和TCP窗口缩放。Internet上的许多过时文档建议禁用其中一个或多个参数以提高性能。这一建议在多年前具有一定的价值、那时CPU功能要低得多、尽可能减少TCP处理开销会有好处。

但是、在现代操作系统中、禁用任何这些TCP功能通常不会带来明显的优势、同时还可能会损害性能。在虚拟化网络环境中、性能可能会受到损害、因为要高效处理数据包丢失和网络质量变化、需要使用这些功能。

*TCP NetApp建议*在主机上启用TCP时间戳、SACK和TCP窗口缩放，在任何当前操作系统中，所有这三个参数都应默认为打开。

适用于Oracle数据库的FC配置

为Oracle数据库配置FC SAN主要是遵循日常SAN最佳实践。

这包括典型的规划措施、例如、确保主机和存储系统之间的SAN具有足够的带宽、检查所有必需设备之间是否存

在所有SAN路径、使用FC交换机供应商所需的FC端口设置、避免ISL争用、 并使用适当的SAN网络结构监控。

分区

一个FC分区不应包含多个启动程序。这种安排最初可能看起来有效、但启动程序之间的串扰最终会影响性能和稳定性。

虽然在极少数情况下、不同供应商的FC目标端口的行为会导致问题、但多目标区域通常被视为安全区域。例如、避免将NetApp和非NetApp存储阵列的目标端口都包含在同一分区中。此外、将NetApp存储系统和磁带设备置于同一分区更容易出现发生原因问题。

Oracle数据库和直连ONTAP连接

存储管理员有时倾向于通过从配置中删除网络交换机来简化其基础架构。在某些情况下、可以支持此功能。

iSCSI和NVMe/TCP

使用iSCSI或NVMe/TCP的主机可以直接连接到存储系统并正常运行。原因是路径问题。直接连接到两个不同的存储控制器会产生两条独立的数据流路径。丢失路径、端口或控制器不会阻止使用另一个路径。

NFS

可以使用直连NFS存储、但有一个重大限制—如果没有大量的脚本编写工作、故障转移将无法正常工作、这是客户的责任。

直连NFS存储的无中断故障转移之所以复杂、是因为本地操作系统上会发生路由。例如、假设主机的IP地址为192.168.1.1/24、并且直接连接到IP地址为192.168.1.50/24的ONTAP控制器。在故障转移期间、该192.168.1.50地址可以故障转移到另一个控制器、并且该地址可供主机使用、但主机如何检测到它的存在？原来的192.168.1.1地址仍然位于不再连接到操作系统的主机NIC上。发往192.168.1.50的流量将继续发送到无法运行的网络端口。

第二个操作系统NIC可配置为19 2.168.1.2、并且能够与故障转移的192.168.1.50地址通信、但本地路由表默认使用一个*且仅一个*地址与192.168.1.0/24子网通信。sysadmin可以创建一个脚本框架、用于检测失败的网络连接并更改本地路由表或启动和关闭接口。确切的操作步骤取决于所使用的操作系统。

在实践中、NetApp客户确实使用直连NFS、但通常仅适用于故障转移期间IO暂停的工作负载。使用硬挂载时、暂停期间不应出现任何IO错误。在服务还原之前、IO应挂起、可以通过故障恢复或手动干预在主机上的NIC之间移动IP地址。

FC直连

不能使用FC协议将主机直接连接到ONTAP存储系统。原因是使用了NPIV。用于向FC网络标识ONTAP FC端口的WWN使用一种称为NPIV的虚拟化类型。连接到ONTAP系统的任何设备都必须能够识别NPIV WWN。目前没有HBA供应商提供可安装在能够支持NPIV目标的主机中的HBA。