通过 RDMA 管理 NFS

ONTAP 9

NetApp
April 24, 2024

目录

通过 RDMA 管理 NFS

基于 RDMA 的 NFS

基于 RDMA 的 NFS 利用 RDMA 适配器，可以在存储系统内存和主机系统内存之间直接复制数据，从而避免 CPU 中断和开销。

基于 RDMA 的 NFS 配置专为具有延迟敏感型或高带宽工作负载（例如机器学习和分析）的客户而设计。NVIDIA 已通过 RDMA 扩展 NFS ，以启用 GPU 直接存储（ GDS ）。GDS可通过完全绕过CPU和主内存、使用RDMA直接在存储系统和GPU内存之间传输数据、进一步加速支持GPU的工作负载。

从ONTAP 9.14.1开始、NFSv4.1协议支持基于RDMA的NFS配置。

从ONTAP 9.10.1开始、如果与使用RoCE协议版本2的RDMA提供支持的迈乐诺克斯CX-5或CX-6适配器结合使用、则NFSv4.0协议支持基于RDMA的NFS配置。只有使用 NVIDIA Tesla 和 Ampere 系列 GPU 以及 Mellanox NIC 卡和 MoFED 软件时，才支持 GDS 。

基于 RDMA 的 NFS 支持仅限于节点 - 本地流量。支持所有成分卷位于同一节点上的标准 FlexVol 或 FlexGroup ，并且必须从同一节点上的 LIF 进行访问。如果 NFS 挂载大小超过 64k ，则会导致 NFS over RDMA 配置的性能不稳定。

要求

- 存储系统必须运行ONTAP 9.10.1或更高版本
  - 从ONTAP 9.12.1开始、您可以使用System Manager配置基于RDMA的NFS。在ONTAP 9.10.1和9.11.1中、您需要使用命令行界面配置基于RDMA的NFS。
- HA对中的两个节点必须为相同版本。
- 存储系统控制器必须支持RDMA

正在**ONTAP**中开始…	以下控制器支持**RDMA**…
9.10.1及更高版本	• A400 • a700 • A800
ONTAP 9.14.1及更高版本	• AFF C系列 • A900

- 配置了RDMA支持的硬件的存储设备(例如 Mellanox CX-5或CX-6)。
- 必须配置数据 LIF 以支持 RDMA 。
- 客户端必须使用支持 Mellanox RDMA 的 NIC 卡和 Mellanox OFED （ MoFED ）网络软件。

> 基于RDMA的NFS不支持接口组。

下一步行动

- 为基于 RDMA 的 NFS 配置 NIC
- 通过 RDMA 为 NFS 配置 LIF
- 基于 RDMA 的 NFS 的 NFS 设置

相关信息

- "RDMA"
- NFS中继概述
- "RFC 7530： NFS 版本 4 协议"
- "RFC 8166： 适用于远程操作步骤调用版本 1 的远程直接内存访问传输"
- "RFC 8167： 基于 RDMA 的 RPC 传输的双向远程操作步骤调用"
- "RFC 8267： NFS 上层绑定到 RDMA 上的 RPC 版本 1"

为基于 RDMA 的 NFS 配置 NIC

基于 RDMA 的 NFS 需要为客户端系统和存储平台配置 NIC 。

存储平台配置

需要在服务器上安装 X1148 RDMA 适配器。如果您使用的是 HA 配置，则故障转移配对节点上必须具有相应的 X1148 适配器，以便 RDMA 服务可以在故障转移期间继续运行。NIC 必须支持 ROCE 。

从ONTAP 9.10.1开始、您可以使用以下命令查看RDMA卸载协议列表： `network port show -rdma -protocols roce`

客户端系统配置

客户端必须使用支持 Mellanox RDMA 的 NIC 卡（例如 X1148 ）和 Mellanox OFED 网络软件。有关支持的型号和版本，请参见 Mellanox 文档。尽管客户端和服务器可以直接连接，但由于交换机的故障转移性能有所提高，因此建议使用交换机。

客户端，服务器和任何交换机以及交换机上的所有端口都必须使用巨型帧进行配置。此外，还应确保优先级流量控制在任何交换机上有效。

确认此配置后，您可以挂载 NFS 。

System Manager

您必须使用ONTAP 9.12.1或更高版本使用System Manager通过RDMA使用NFS配置网络接口。

步骤

1. 检查是否支持RDMA。导航到*网络>以太网端口*、然后在组视图中选择相应的节点。展开节点时、请查看给定端口的* RDMA protocols*字段：值* RoCE*表示支持RDMA；短划线(-)表示不支持RDMA。
2. 要添加VLAN、请选择*+ VLAN*。选择相应的节点。在*端口*下拉菜单中、如果可用端口支持RDMA、则会显示文本*已启用RoCE *；如果不支持RDMA、则不会显示任何文本。
3. 按照中的工作流进行操作 使用 NFS 为 Linux 服务器启用 NAS 存储 配置新的NFS服务器。

   添加网络接口时、您可以选择*使用RoCE端口*。对于要使用基于RDMA的NFS的任何网络接口、请选择此选项。

命令行界面

1. 使用命令检查 NFS 服务器上是否启用了 RDMA 访问：

   ```
   vserver nfs show-vserver SVM_name
   ```

   默认情况下、 `-rdma` 应启用。如果不是，请在 NFS 服务器上启用 RDMA 访问：

   ```
   vserver nfs modify -vserver SVM_name -rdma enabled
   ```

2. 通过 NFSv4.0 通过 RDMA 挂载客户端：
   a. proto 参数的输入取决于服务器 IP 协议版本。如果为IPv4、请使用 `proto=rdma`。如果使用IPv6、请使用 `proto=rdma6`。
   b. 将NFS目标端口指定为 `port=20049` 而不是标准端口2049：

      ```
      mount -o vers=4,minorversion=0,proto=rdma,port=20049 Server_IP_address
      :/volume_path mount_point
      ```

3. 可选：如果需要卸载客户端、请运行命令 `unmount mount_path`

更多信息

- 创建 NFS 服务器
- 使用 NFS 为 Linux 服务器启用 NAS 存储

通过 RDMA 为 NFS 配置 LIF

要使用基于RDMA的NFS、必须将LIF (网络接口)配置为与RDMA兼容。LIF及其故障转移对都必须能够支持RDMA。

创建新的 LIF

System Manager

要使用System Manager通过RDMA为NFS创建网络接口、必须运行ONTAP 9.12.1或更高版本。

步骤

1. 选择*网络>概述>网络接口*。
2. 选择 … + Add 。
3. 如果选择* NFS、SMB/CIFS、S3*、则可以选择*使用RoCE端口*。选中*使用RoCE端口*复选框。
4. 选择Storage VM和主节点。分配一个名称。输入IP地址和子网掩码。
5. 输入IP地址和子网掩码后、System Manager会将广播域列表筛选为具有支持RoCE的端口的广播域列表。选择广播域。您可以选择添加网关。
6. 选择 * 保存 * 。

命令行界面

步骤

1. 创建 LIF：

   ```
   network interface create -vserver SVM_name -lif lif_name -service-policy
   service_policy_name -home-node node_name -home-port port_name {-address
   IP_address -netmask netmask_value | -subnet-name subnet_name} -firewall
   -policy policy_name -auto-revert {true|false} -rdma-protocols roce
   ```

   - 服务策略必须为 default-data-files 或包含 data-nfs 网络接口服务的自定义策略。
   - 。 `-rdma-protocols` 参数接受默认为空的列表。时间 `roce` 作为一种价值、只能在支持RoCE卸载的端口上配置LIF、从而影响爬虫程序LIF迁移和故障转移。

修改 LIF

System Manager

要使用System Manager通过RDMA为NFS创建网络接口、必须运行ONTAP 9.12.1或更高版本。

步骤

1. 选择*网络>概述>网络接口*。
2. 选择 … ⋮ 要更改的网络接口旁边的*>编辑*。
3. 选中*使用RoCE端口*以启用基于RDMA的NFS、或者取消选中此复选框以将其禁用。如果网络接口位于支持RoCE的端口上、您将看到*使用RoCE端口*旁边的复选框。
4. 根据需要修改其他设置。
5. 选择*保存*以确认所做的更改。

命令行界面

1. 您可以使用检查您的生命周期管理器的状态 `network interface show` 命令：服务策略必须包含 data-nfs 网络接口服务。。 `-rdma-protocols` 列表应包括 `roce`。如果上述任一条件不正确，请修改 LIF 。
2. 要修改 LIF ，请运行：

```
network interface modify vserver SVM_name -lif lif_name -service-policy
service_policy_name -home-node node_name -home-port port_name {-address
IP_address -netmask netmask_value | -subnet-name subnet_name} -firewall
-policy policy_name -auto-revert {true|false} -rdma-protocols roce
```

> 如果当前未将 LIF 分配给支持该协议的端口，则修改 LIF 以要求使用特定的卸载协议会产生错误。

迁移 LIF

ONTAP 还允许您迁移网络接口(LIF)以利用基于RDMA的NFS。执行此迁移时、必须确保目标端口支持RoCE。从ONTAP 9.12.1开始、您可以在System Manager中完成此操作步骤。在为网络接口选择目标端口时、System Manager将指定端口是否支持RoCE。

只有在以下情况下、才能将LIF迁移到基于RDMA的NFS配置：

- 它是一个NFS RDMA网络接口(LIF)、托管在支持RoCE的端口上。
- 它是一个NFS TCP网络接口(LIF)、托管在支持RoCE的端口上。
- 它是一个NFS TCP网络接口(LIF)、托管在不支持RoCE的端口上。

有关迁移网络接口的详细信息、请参见 迁移 LIF。

更多信息

- 创建 LIF
- 创建 LIF
- 修改 LIF

- 迁移 LIF

修改 NFS 配置

大多数情况下、您不需要修改已启用NFS的Storage VM的配置、以便通过RDMA使用NFS。

但是，如果您要处理与 Mellanox 芯片和 LIF 迁移相关的问题，则应增加 NFSv4 锁定宽限期。默认情况下，宽限期设置为 45 秒。从ONTAP 9.10.1开始、宽限期的最大值为180 (秒)。

步骤

1. 将权限级别设置为高级：

   ```
   set -privilege advanced
   ```

2. 输入以下命令：

   ```
   vserver nfs modify -vserver SVM_name -v4-grace-seconds number_of_seconds
   ```

有关此任务的详细信息，请参见 指定 NFSv4 锁定宽限期。