Network configuration

Enterprise applications

NetApp
April 25, 2024

This PDF was generated from https://docs.netapp.com/us-en/ontap-apps-dbs/oracle/oracle-network-config-lifs.html on April 25, 2024. Always check docs.netapp.com for the latest.

Table of Contents

Network configuration

Logical interface design for Oracle databases

Oracle databases need access to storage. Logical interfaces (LIFs) are the network plumbing that connects a storage virtual machine (SVM) to the network and in turn to the database. Proper LIF design is required to ensure sufficient bandwidth exists for each database workload, and failover does not result in a loss of storage services.

This section provides an overview of key LIF design principles. For more comprehensive documentation, see the ONTAP Network Management documentation. As with other aspects of database architecture, the best options for storage virtual machine (SVM, known as a vserver at the CLI) and logical interface (LIF) design depend heavily on scaling requirements and business needs.

Consider the following primary topics when building a LIF strategy:

- **Performance.** Is the network bandwidth sufficient?
- **Resiliency.** Are there any single points of failure in the design?
- **Manageability.** Can the network be scaled nondisruptively?

These topics apply to the end-to-end solution, from the host through the switches to the storage system.

LIF types

There are multiple LIF types. ONTAP documentation on LIF types provide more complete information on this topic, but from a functional perspective LIFs can be divided into the following groups:

- **Cluster and node management LIFs.** LIFs used to manage the storage cluster.
- **SVM management LIFs.** Interfaces that permit access to an SVM through the REST API or ONTAPI (also known as ZAPI) for functions such as snapshot creation or volume resizing. Products such as SnapManager for Oracle (SMO) must have access to an SVM management LIF.
- **Data LIFs.** Interfaces for FC, iSCSI, NVMe/FC, NVMe/TCP, NFS, or SMB/CIFS data.

> ⓘ A data LIF used for NFS traffic can also be used for management by changing the firewall policy from `data` to `mgmt` or another policy that allows HTTP, HTTPS, or SSH. This change can simplify network configuration by avoiding the configuration of each host for access to both the NFS data LIF and a separate management LIF. It is not possible to configure an interface for both iSCSI and management traffic, despite the fact that both use an IP protocol. A separate management LIF is required in iSCSI environments.

SAN LIF design

LIF design in a SAN environment is relatively simple for one reason: multipathing. All modern SAN implementations allow a client to access data over multiple, independent, network paths and select the best path or paths for access. As a result, performance with respect to LIF design is simpler to address because SAN clients automatically load-balance I/O across the best available paths.

If a path becomes unavailable, the client automatically selects a different path. The resulting simplicity of design makes SAN LIFs generally more manageable. This does not mean that a SAN environment is always more easily managed, because there are many other aspects of SAN storage that are much more complicated

than NFS. It simply means that SAN LIF design is easier.

Performance

The most important consideration with LIF performance in a SAN environment is bandwidth. For example, a two-node ONTAP AFF cluster with two 16Gb FC ports per node allows up to 32Gb of bandwidth to/from each node.

Resiliency

SAN LIFs do not fail over on an AFF storage system. If a SAN LIF fails because of controller failover, then the client's multipathing software detects the loss of a path and redirects I/O to a different LIF. With ASA storage systems, LIFs will be failed over after a short delay, but this does not interrupt IO because there are already active paths on the other controller. The failover process occurs in order to restore host access on all defined ports.

Manageability

LIF migration is a much more common task in an NFS environment because LIF migration is often associated with relocating volumes around the cluster. There is no need to migrate a LIF in a SAN environment when volumes are relocated within the HA pair. That is because, after the volume move has completed, ONTAP sends a notification to the SAN about a change in paths, and the SAN clients automatically reoptimize. LIF migration with SAN is primarily associated with major physical hardware changes. For example, if a nondisruptive upgrade of the controllers is required, a SAN LIF is migrated to the new hardware. If an FC port is found to be faulty, a LIF can be migrated to an unused port.

Design recommendations

NetApp makes the following recommendations:

- Do not create more paths than are required. Excessive numbers of paths make overall management more complicated and can cause problems with path failover on some hosts. Furthermore, some hosts have unexpected path limitations for configurations such as SAN booting.
- Very few configurations should require more than four paths to a LUN. The value of having more than two nodes advertising paths to LUNs is limited because the aggregate hosting a LUN is inaccessible if the node that owns the LUN and its HA partner fail. Creating paths on nodes other than the primary HA pair is not helpful in such a situation.
- Although the number of visible LUN paths can be managed by selecting which ports are included in FC zones, it is generally easier to include all potential target points in the FC zone and control LUN visibility at the ONTAP level.
- In ONTAP 8.3 and later, the selective LUN mapping (SLM) feature is the default. With SLM, any new LUN is automatically advertised from the node that owns the underlying aggregate and the node's HA partner. This arrangement avoids the need to create port sets or configure zoning to limit port accessibility. Each LUN is available on the minimum number of nodes required for both optimal performance and resiliency. *In the event a LUN must be migrated outside of the two controllers, the additional nodes can be added with the `lun mapping add-reporting-nodes` command so that the LUNs are advertised on the new nodes. Doing so creates additional SAN paths to the LUNs for LUN migration. However, the host must perform a discovery operation to use the new paths.
- Do not be overly concerned about indirect traffic. It is best to avoid indirect traffic in a very I/O-intensive environment for which every microsecond of latency is critical, but the visible performance effect is negligible for typical workloads.

NFS LIF design

In contrast to SAN protocols, NFS has a limited ability to define multiple paths to data. The parallel NFS (pNFS) extensions to NFSv4 address this limitation, but as ethernet speeds have reached 100Gb and beyond there is rarely value in adding additional paths.

Performance and resiliency

Although measuring SAN LIF performance is primarily a matter of calculating the total bandwidth from all primary paths, determining NFS LIF performance requires taking a closer look at the exact network configuration. For example, two 10Gb ports can be configured as raw physical ports, or they can be configured as a Link Aggregation Control Protocol (LACP) interface group. If they are configured as an interface group, multiple load balancing policies are available that work differently depending on whether traffic is switched or routed. Finally, Oracle direct NFS (dNFS) offers load-balancing configurations that do not exist in any OS NFS clients at this time.

Unlike SAN protocols, NFS file systems require resiliency at the protocol layer. For example, a LUN is always configured with multipathing enabled, meaning that multiple redundant channels are available to the storage system, each of which uses the FC protocol. An NFS file system, on the other hand, depends on the availability of a single TCP/IP channel that can only be protected at the physical layer. This arrangement is why options such as port failover and LACP port aggregation exist.

In an NFS environment, both performance and resiliency are provided at the network protocol layer. As a result, both topics are intertwined and must be discussed together.

Bind LIFs to port groups

To bind a LIF to a port group, associate the LIF IP address with a group of physical ports. The primary method for aggregating physical ports together is LACP. The fault-tolerance capability of LACP is fairly simple; each port in an LACP group is monitored and is removed from the port group in the event of a malfunction. There are, however, many misconceptions about how LACP works with respect to performance:

- LACP does not require the configuration on the switch to match the endpoint. For example, ONTAP can be configured with IP-based load balancing, while a switch can use MAC-based load balancing.
- Each endpoint using an LACP connection can independently choose the packet transmission port, but it cannot choose the port used for receipt. This means that traffic from ONTAP to a particular destination is tied to a particular port, and the return traffic could arrive on a different interface. This does not cause problems, however.
- LACP does not evenly distribute traffic all the time. In a large environment with many NFS clients, the result is typically even use of all ports in an LACP aggregation. However, any one NFS file system in the environment is limited to the bandwidth of only one port, not the entire aggregation.
- Although robin-robin LACP policies are available on ONTAP, these policies do not address the connection from a switch to a host. For example, a configuration with a four-port LACP trunk on a host and a four-port LACP trunk on ONTAP is still only able to read a file system using a single port. Although ONTAP can transmit data through all four ports, no switch technologies are currently available that send from the switch to the host through all four ports. Only one is used.

The most common approach in larger environments consisting of many database hosts is to build an LACP aggregate of an appropriate number of 10Gb (or faster) interfaces by using IP load balancing. This approach enables ONTAP to deliver even use of all ports, as long as enough clients exist. Load balancing breaks down when there are fewer clients in the configuration because LACP trunking does not dynamically redistribute load.

When a connection is established, traffic in a particular direction is placed on only one port. For example, a database performing a full table scan against an NFS file system connected through a four-port LACP trunk reads data though only one network interface card (NIC). If only three database servers are in such an environment, it is possible that all three are reading from the same port, while the other three ports are idle.

Bind LIFs to physical ports

Binding a LIF to a physical port results in more granular control over network configuration because a given IP address on a ONTAP system is associated with only one network port at a time. Resiliency is then accomplished through the configuration of failover groups and failover policies.

Failover policies and failover groups

The behavior of LIFs during network disruption is controlled by failover policies and failover groups. Configuration options have changed with the different versions of ONTAP. Consult the ONTAP network management documentation for failover groups and policies for specific details for the version of ONTAP being deployed.

ONTAP 8.3 and higher allow management of LIF failover based on broadcast domains. Therefore, an administrator can define all of the ports that have access to a given subnet and allow ONTAP to select an appropriate failover LIF. This approach can be used by some customers, but it has limitations in a high-speed storage network environment because of the lack of predictability. For example, an environment can include both 1Gb ports for routine file system access and 10Gb ports for datafile I/O. If both types of ports exist in the same broadcast domain, LIF failover can result in moving datafile I/O from a 10Gb port to a 1Gb port.

In summary, consider the following practices:

1. Configure a failover group as user-defined.
2. Populate the failover group with ports on the storage failover (SFO) partner controller so that the LIFs follow the aggregates during a storage failover. This avoids creating indirect traffic.
3. Use failover ports with matching performance characteristics to the original LIF. For example, a LIF on a single physical 10Gb port should include a failover group with a single 10Gb port. A four-port LACP LIF should fail over to another four-port LACP LIF. These ports would be a subset of the ports defined in the broadcast domain.
4. Set the failover policy to SFO-partner only. Doing so makes sure that the LIF follows the aggregate during failover.

Auto-revert

Set the `auto-revert` parameter as desired. Most customers prefer to set this parameter to `true` to have the LIF revert to its home port. However, in some cases, customers have set this to `false `so that an unexpected failover can be investigated before returning a LIF to its home port.

LIF-to-volume ratio

A common misconception is that there must be a 1:1 relationship between volumes and NFS LIFs. Although this configuration is required for moving a volume anywhere in a cluster while never creating additional interconnect traffic, it is categorically not a requirement. Intercluster traffic must be considered, but the mere presence of intercluster traffic does not create problems. Many of the published benchmarks created for ONTAP include predominantly indirect I/O.

For example, a database project containing a relatively small number of performance-critical databases that only required a total of 40 volumes might warrant a 1:1 volume to LIF strategy, an arrangement that would require 40 IP addresses. Any volume could then be moved anywhere in the cluster along with the associated

LIF, and traffic would always be direct, minimizing every source of latency even at microsecond levels.

As a counter example, a large, hosted environment might be more easily managed with a 1:1 relationship between customers and LIFs. Over time, a volume might need to be migrated to a different node, which would cause some indirect traffic. However, the performance effect should be undetectable unless the network ports on the interconnect switch are saturating. If there is concern, a new LIF can be established on additional nodes and the host can be updated at the next maintenance window to remove indirect traffic from the configuration.

TCP/IP and ethernet configuration for Oracle databases

Many Oracle on ONTAP customers use ethernet, the network protocol of NFS, iSCSI, NVMe/TCP, and especially the cloud.

Host OS settings

Most application vendor documentation include specific TCP and ethernet settings intended to ensure the application is working optimally. These same settings are usually sufficient to also deliver optimal IP-based storage performance.

Ethernet flow control

This technology allows a client to request that a sender temporarily stop data transmission. This is usually done because the receiver is unable to process incoming data quickly enough. At one time, requesting that a sender cease transmission was less disruptive than having a receiver discard packets because buffers were full. This is no longer the case with the TCP stacks used in OSs today. In fact, flow control causes more problems than it solves.

Performance problems caused by Ethernet flow control have been increasing in recent years. This is because Ethernet flow control operates at the physical layer. If a network configuration permits any host OS to send an Ethernet flow control request to a storage system, the result is a pause in I/O for all connected clients. Because an increasing number of clients are served by a single storage controller, the likelihood of one or more of these clients sending flow control requests increases. The problem has been seen frequently at customer sites with extensive OS virtualization.

A NIC on a NetApp system should not receive flow-control requests. The method used to achieve this result varies based on the network switch manufacturer. In most cases, flow control on an Ethernet switch can be set to `receive desired` or `receive on`, which means that a flow control request is not forwarded to the storage controller. In other cases, the network connection on the storage controller might not allow flow-control disabling. In these cases, the clients must be configured to never send flow control requests, either by changing to the NIC configuration on the host server itself or the switch ports to which the host server is connected.

NetApp recommends making sure that NetApp storage controllers do not receive Ethernet flow-control packets. This can generally be done by setting the switch ports to which the controller is attached, but some switch hardware has limitations that might require client-side changes instead.

MTU Sizes

The use of jumbo frames has been shown to offer some performance improvement in 1Gb networks by reducing CPU and network overhead, but the benefit is not usually significant.

NetApp recommends implementing jumbo frames when possible, both to realize any potential performance benefits and to future-proof the solution.

Using jumbo frames in a 10Gb network is almost mandatory. This is because most 10Gb implementations reach a packets-per-second limit without jumbo frames before they reach the 10Gb mark. Using jumbo frames improves efficiency in TCP/IP processing because it allows the OS, server, NICs, and the storage system to process fewer but larger packets. The performance improvement varies from NIC to NIC, but it is significant.

For jumbo-frame implementations, there is the common but incorrect belief that all connected devices must support jumbo frames and that the MTU size must match end-to-end. Instead, the two network end points negotiate the highest mutually acceptable frame size when establishing a connection. In a typical environment, a network switch is set to an MTU size of 9216, the NetApp controller is set to 9000, and the clients are set to a mix of 9000 and 1514. Clients that can support an MTU of 9000 can use jumbo frames, and clients that can only support 1514 can negotiate a lower value.

Problems with this arrangement are rare in a completely switched environment. However, take care in a routed environment that no intermediate router is forced to fragment jumbo frames.

NetApp recommends configuring the following:

- Jumbo frames are desirable but not required with 1Gb Ethernet (GbE).
- Jumbo frames are required for maximum performance with 10GbE and faster.

TCP parameters

Three settings are often misconfigured: TCP timestamps, selective acknowledgment (SACK), and TCP window scaling. Many out-of-date documents on the Internet recommend disabling one or more of these parameters to improve performance. There was some merit to this recommendation many years ago when CPU capabilities were much lower and there was a benefit to reducing the overhead on TCP processing whenever possible.

However, with modern OSs, disabling any of these TCP features usually results in no detectable benefit while also potentially damaging performance. Performance damage is especially likely in virtualized networking environments because these features are required for efficient handling of packet loss and changes in network quality.

NetApp recommends enabling TCP timestamps, SACK, and TCP window scaling on the host, and all three of these parameters should be on by default in any current OS.

FC configuration for Oracle databases

Configuring FC SAN for Oracle databases is primarily about following everyday SAN best practices.

This includes typical planning measures such as ensuring sufficient bandwidth exists on the SAN in between the host and storage system, checking that all SAN paths exist between all required devices, using the FC port settings required by your FC switch vendor, avoiding ISL contention, and using proper SAN fabric monitoring.

Zoning

An FC zone should never contain more than one initiator. Such an arrangement might appear to work initially, but crosstalk between initiators eventually interferes with performance and stability.

Multitarget zones are generally regarded as safe, although in rare circumstances the behavior of FC target ports from different vendors has caused problems. For example, avoid including the target ports from both a NetApp and an non-NetApp storage array in the same zone. In addition, placing a NetApp storage system and a tape device in the same zone is even more likely to cause problems.

Oracle database and direct-connect ONTAP connectivity

Storage administrators sometimes prefer to simplify their infrastructures by removing network switches from the configuration. This can be supported in some scenarios.

iSCSI and NVMe/TCP

A host using iSCSI or NVMe/TCP can be directly connected to a storage system and operate normally. The reason is pathing. Direct connections to two different storage controllers results in two independent paths for data flow. The loss of path, port, or controller does not prevent the other path from being used.

NFS

Direct-connected NFS storage can be used, but with a significant limitation - failover will not work without a significant scripting effort, which would be the responsibility of the customer.

The reason nondisruptive failover is complicated with direct-connected NFS storage is the routing that occurs on the local OS. For example, assume a host has an IP address of 192.168.1.1/24 and is directly connected to an ONTAP controller with an IP address of 192.168.1.50/24. During failover, that 192.168.1.50 address can fail over to the other controller, and it will be available to the host, but how does the host detect its presence? The original 192.168.1.1 address still exists on the host NIC that no longer connects to an operational system. Traffic destined for 192.168.1.50 would continue to be sent to an inoperable network port.

The second OS NIC could be configured as 19 2.168.1.2 and would be capable of communicating with the failed over 192.168.1.50 address, but the local routing tables would have a default of using one **and only one** address to communicate with the 192.168.1.0/24 subnet. A sysadmin could create a scripting framework that would detect a failed network connection and alter the local routing tables or bring interfaces up and down. The exact procedure would depend on the OS in use.

In practice, NetApp customers do have direct-connected NFS, but normally only for workloads where IO pauses during failovers are acceptable. When hard mounts are used, there should not be any IO errors during such pauses. The IO should hang until services are restored, either by a failback or manual intervention to move IP addresses between NICs on the host.

FC direct connect

It is not possible to directly connect a host to an ONTAP storage system using the FC protocol. The reason is the use of NPIV. The WWN that identifies an ONTAP FC port to the FC network uses a type of virtualization called NPIV. Any device connected to an ONTAP system must be able to recognize an NPIV WWN. There are no current HBA vendors who offer an HBA that can be installed in a host that would be able to support an NPIV target.

Copyright information

Copyright © 2024 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP “AS IS” AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b) (FEB 2014).

Trademark information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.