White Paper

BRIDGING EMC ISILON NAS ON IP TO INFINIBAND NETWORKS WITH MELLANOX SWITCHX

Abstract

This white paper explains how to configure a Mellanox SwitchX Series switch to bridge the external network of an EMC Isilon cluster to an InfiniBand network.

November 2013

Part Number H12523

Table of contents

Introduction

InfiniBand is a common networking interconnect used in many high-performance computing (HPC) environments as well as within the enterprise. Typically, resources connected with InfiniBand have not been able to access NAS storage platforms, because such platforms operate exclusively over Ethernet. Bridging the two interconnects has also been difficult because of as software-only solutions, the reduced bandwidth of Ethernet, and other protocol incompatibilities. However, recently introduced switches from Mellanox—the SwitchX Series—provide the capability to bridge InfiniBand networks to 10 Gigabit Ethernet (GbE) networks. This bridging capability now makes it feasible to connect hosts using InfiniBand to NAS devices, while retaining performance to transfer data quickly and efficiently.

This white paper describes how to configure a Mellanox SwitchX and an EMC® Isilon® cluster to bridge an InfiniBand network to a 10GbE network.

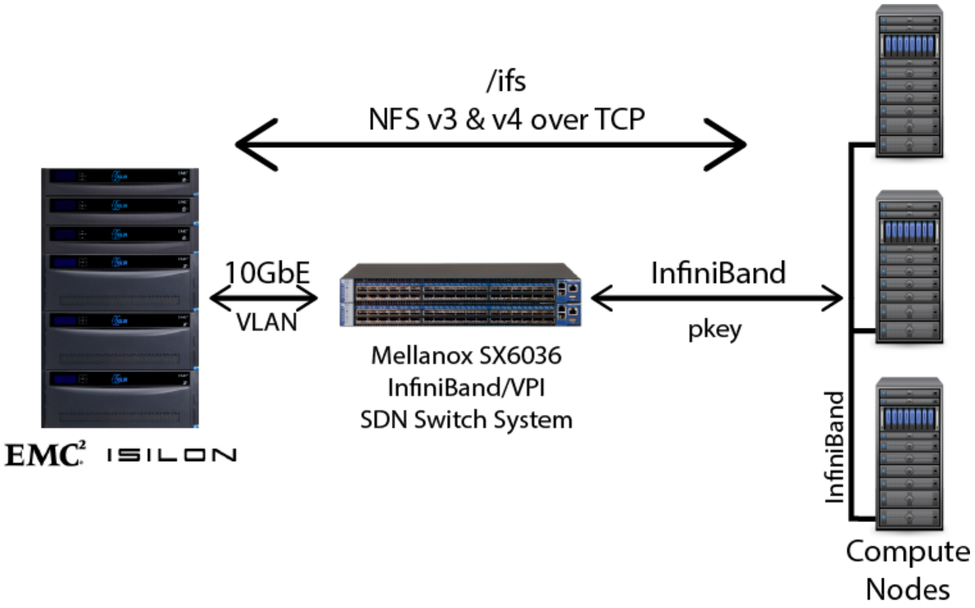

Figure 1. Bridging an InfiniBand network to an Isilon storage cluster

Considerations

Internet Protocol over InfiniBand (IPoverIB) must be running on the hosts that will access the EMC Isilon cluster. The IPoverIB interface should have a network address assigned to it that can reach the interface on the Isilon cluster. Although the two interfaces do not have to be on the same network, having them on the same network simplifies the configuration. The solution uses IPv4 over Ethernet.

The cabling connects common 10GbE copper (QSFP+) connections to standard InfiniBand QDR. Various cable lengths are available (consult Mellanox for part numbers and availability).

Any SwitchX Series switch from Mellanox is capable of acting as a gateway for InfiniBand to IP networks using the Virtual Protocol Interconnect (VPI). VPI is a licensed feature of Mellanox SwitchX switches (contact Mellanox for availability). The switch evaluated for this guide was an SX6036 with 36 ports, all of which can support either Ethernet or InfiniBand as a signaling protocol.

The Isilon OneFS® operating system, version 7.0.2.1, was used in the evaluation for this white paper. No minimum version of OneFS is required when deploying this configuration. The features used are common to all versions.

This paper assumes a basic familiarity with the configuration of Mellanox SwitchX switches and the administration of an Isilon cluster. Most of the configuration will be done while connected to the switch through a terminal and to the Isilon cluster through the OneFS command-line administration interface.

Client setup

The following configuration uses the MLNX_OFED driver stack (which was the only stack evaluated). This driver stack is available from Mellanox for various distributions of Linux and other operating systems. The following client configuration was developed and tested on CentOS 6.4.

On any client connected to the switch, the ports of the InfiniBand Host Bus Adapter (HBA) must be set to AutoSense. Run `connectx_port_config`, a utility included with the MLNX_OFED driver stack (normally `/sbin/connectx_port_config`) to set each port on the HBA to AutoSense. For systems with a single HBA, you can run `connectx_port_config -c auto,auto`. If the client is already connected to an InfiniBand fabric, this command is non-disruptive.

If IPoverIB is not configured, configure it on each client that you intend to connect to Isilon storage through the switch. Once the MLNX_OFED stack is installed, it automatically provides, or plumbs, an ib* interface for all the HBA ports. Configuring any of the interfaces can be done through ifconfig or through another networking setup utility.

Here is an example of how to configure IPoverIB with ifconfig:

```
    ifconfig ib0 inet 172.28.9.140 netmask 255.255.255.0
```

The solution discussed in this guide uses an OpenSM InfiniBand subnet manager process running on a client node. Once configured for bridging between Ethernet and InfiniBand, the OpenSM process available on the switch will no longer be active or available to start. Instead the subnet manager process will need to run on a client machine that is part of the existing InfiniBand network to be bridged with the Ethernet network. The manager, which is called “opensm,” is included with all MLNX_OFED distributions and must be started by the root account on at least one of the clients. As a best practice, you should, for redundancy, run the manager on two or more clients.

Switch setup

You must change the system mode of the switch to VPI from Ethernet or InfiniBand. This configuration requires a license; contact Mellanox to obtain a license key.

The following steps take place using the command line of the switch. To access the command line, ssh to the switch with the admin account, enter the 'enable' mode and then enter the following configuration commands after entering 'config term':

Set the system profile:

```
system profile vpi-single-switch
```

At this point, the group of ports required to use Ethernet—that is, the ports connected to the 10GbE external interfaces on the Isilon nodes—should be enumerated to the switch. For example:

```
switch (config)# interface infiniband 1/1-1/8 shutdown
switch (config)# port 1/1-1/8 type ethernet
switch (config)# interface ethernet 1/1-1/8 no shutdown
```

The Isilon network ports do not auto-negotiate the port speed with the switch; as a result, you must set the speed of each port to 10GbE with the OneFS Web administration interface or command-line interface.

Client ports connected to compute systems that will be mounting shares from the Isilon cluster should remain connected as InfiniBand. However, if a port’s transport protocol appears incorrect either from the OneFS Web administration interface or from the 'show ports' command on the switch, a similar procedure can be followed to fix incorrectly assigned ports. For example, to set a port from Ethernet back to InfiniBand, run the procedure in reverse:

```
switch (config)# interface ethernet 1/9 shutdown
switch (config)# port 1/9 type infiniband
switch (config)# interface infiniband 1/9 no shutdown
```

General switch settings

On the switch, verify that the gateway (GW) feature is active (`Supported`) and that both InfiniBand and Ethernet are available:

```
switch-641b04 [standalone: master] (config) # show system capabilities
IB: Supported
Ethernet: Supported, Full L2, L3
GW: Supported
Max SM nodes: 648
IB Max licensed speed: FDR
Ethernet Max licensed speed: 40Gb
```

Disable IP routing:

```
switch-641b04 [standalone: master] (config) # no ip routing
switch-641b04 [standalone: master] (config) # show ip routing
IP routing: disabled
```

Disable IGMP for IP and the SM on InfiniBand:

```
    switch-641b04 [standalone: master] (config) # no ip igmp snooping
    switch-641b04 [standalone: master] (config) # no ib sm
```

Proxy-ARP interface setup

To allow connectivity between the two networks, you must set up a proxy-ARP interface on the switch. Proxy-ARP allows the switch to pass physical addresses from one network to another so that the addresses can be resolved and the two networks can be bridged. Here is an example of how to set up a proxy-ARP interface:

First enable proxy-ARP functionality:

```
    switch-641b04 [standalone: master] (config) # ip proxy-arp
    switch-641b04 [standalone: master] (config) # show ip proxy-arp
    Proxy-arp: enabled
```

Then create the interface. Use an IP address on the network on which all hosts (Isilon nodes and IP over InfiniBand clients) will reside:

```
    switch-641b04 [standalone: master] (config) # interface proxy-arp 1
    switch-641b04 [standalone: master] (config interface proxy-arp 1) # ip
    address 172.28.9.149 /24
    switch-641b04 [standalone: master] (config interface proxy-arp 1) # no
    shutdown
    switch-641b04 [standalone: master] (config interface proxy-arp 1) # exit
```

A new VLAN interface must be set up so that ports can be selectively assigned to it. The default VLAN interface (VLAN 1) is not configured to support proxy-ARP:

```
    switch-641b04 [standalone: master] (config) # vlan 10
```

Assign to the VLAN both the switch ports connected to the Isilon nodes and the switch ports connected to the IP over InfiniBand clients:

```
    switch-641b04 [standalone: master] (config) # interface ethernet 1/7
    switch-641b04 [standalone: master] (config interface ethernet 1/7) #
    switchport access vlan 10
    switch-641b04 [standalone: master] (config interface ethernet 1/7) # exit
```

Repeat the process for each connected switch port.

Assign the proxy-ARP interface to the VLAN interface:

```
    switch-641b04 [standalone: master] (config vlan 10) # exit
    switch-641b04 [standalone: master] (config) # interface proxy-arp 1
    switch-641b04 [standalone: master] (config interface proxy-arp 1) #
    switchport access vlan 10
```

Map the proxy-ARP interface to the VLAN:

```
    switch-641b04 [standalone: master] (config) # interface vlan 10
    switch-641b04 [standalone: master] (config interface vlan 10) # ip proxy-arp-
    map 1
    switch-641b04 [standalone: master] (config interface vlan 10) # no shutdown
```

Create a PKEY interface and map it to the proxy-ARP interface:

```
switch-641b04 [standalone: master] (config) # interface pkey 0x7777
switch-641b04 [standalone: master] (config interface pkey 0x7777) # ip proxy-
arp-map 1
switch-641b04 [standalone: master] (config interface pkey 0x7777) # no
shutdown
```

A PKEY interface for InfiniBand is the rough analog to a VLAN interface for Ethernet. This process created a PKEY interface and bridged it with the VLAN by way of the proxy-ARP interface. At this stage, the broadcast domains of the VLAN and the InfiniBand PKEY are joined and broadcast traffic will transit between each network by way of the proxy-ARP interface.

Isilon setup

From the perspective of any Isilon node connected to a SwitchX switch from Mellanox, once the switch port is set to use Ethernet it is indistinguishable from any other Ethernet switch. Because of the significant bandwidth available to InfiniBand, it is strongly recommended as a best practice to use the 10GbE interfaces on the Isilon nodes when connecting to this solution. Using the gigabit interfaces even as a single bonded interface is not recommended and will severely limit the overall bandwidth and transfer speed available to any client system and the utility of the overall solution.

Before you set the IPv4 address of any of the ports connected to the solution, you must create a subnet through the OneFS command-line interface.

```
isi networks create subnet -n ipoibsub --netmask 255.255.255.0 --gateway
172.28.9.1 --sc-service-addr 172.28.9.159 --mtu=1500
```

No VLAN tagging is necessary when connecting directly from the Isilon node to the Mellanox switch, because the port on the switch is already assigned to the VLAN that is being bridged to the InfiniBand network. However, it is recommended that you specify a VLAN tag if the Isilon interface is connected to an aggregating network switch that is then connected to the Mellanox switch. You specify a VLAN tag by adding `--vlan-id <vlan id>` to the end of the command above. For more information on how to create a subnet, see the OneFS Command Reference for your version of OneFS.

A SmartConnect™ service address is specified at this point and used for connections to the Isilon cluster. SmartConnect distributes client connections across nodes in the cluster. SmartConnect tracks which interfaces are assigned to each subnet and network range, and distributes connections over them according to several different load-balancing algorithms, such as round robin.

When you create a subnet on the Isilon cluster, you must specify the MTU. The default is 1500 and should be explicitly specified at this value for the connection between the cluster and the Mellanox switch. You must use 1500 for two reasons. First, on the side of the Isilon cluster, the only values available through the OneFS administration interface are 1500 and 9000 bytes. Second, InfiniBand cannot pass frames larger than 4096 bytes, which means that IPv4 jumbo frames cannot be passed from the Ethernet network, and no reframing or fragmentation is available. As a result, the Isilon cluster should be configured with an MTU of 1500 bytes only so that the value will be represented correctly to the OneFS Web administration interface, and so that all traffic can be sent between both networks.

Once a subnet is present, a pool containing the network range for the Isilon interfaces can be specified:

```
isi networks create pool -n ipoibsub:ipoibpool1 --ranges 172.28.9.150-
172.28.9.155 --dynamic
```

Then the Isilon 10GbE interfaces on nodes connected to the switch can be attached to this network range:

```
isi networks modify pool -n ipoibsub:ipoibpool1 --add-ifaces 1:10gige-2
```

The command above adds the second 10GbE interface on the first Isilon node to the pool defined earlier. This will have the immediate result of an address being assigned to the interface from the range specified by the pool definition and the interface being activated. The Isilon interfaces that are connected to the Mellanox switch should be added to the pool.

For more information on how to create and modify a pool, see the OneFS Command Reference for your version of OneFS.

Mounting on client systems

Once the Isilon interfaces are configured and the Mellanox SwitchX switch is configured with VPI and a Proxy-ARP interface, traffic can begin flowing between the two networks. Assuming the ipoib interface has been configured on the client (see above), the client can then begin communicating with an Isilon cluster using the SmartConnect address:

```
showmount -e 172.28.9.159
```

And shares can be mounted on the client like this:

```
mount -t nfs  172.28.9.159:/ifs /ifscluster1
```

At this point, only TCP and UDP are available as transport protocols when mounting NFS shares from an Isilon cluster. RDMA is not supported, and an error will be returned when trying to use it as a transport protocol.

Observed performance

With clients connected using ConnectX HBAs to the Mellanox switch and the 10GbE interfaces of three Isilon X400 nodes, the observed performance over NFS was nominally 90 percent of the performance when just using 10GbE from end to end. In all instances of short duration, small file transfers for both read and write operations, the performance was identical to that of a pure 10GbE solution. With longer duration, large file transfers over NFS, SCP, and FTP, the overall performance was only somewhat degraded.

Conclusion

While the ability to bridge Ethernet and InfiniBand has existed in other out-of-band software solutions, Mellanox has provided an elegant, usable service in its SwitchX Series switches. With this capability embedded in the switch, InfiniBand clients can now easily access existing Isilon NAS clusters, which had previously only been available to Ethernet-connected clients.

Isilon provides best-in-class, highly scalable and high-throughput NAS solutions to meet a number of challenges in HPC, and is now accessible to all systems connected with InfiniBand or Ethernet. Using the solution described in this document, different storage tiers in HPC can finally be efficiently and effectively bridged. Tier 1 storage using Lustre can now accommodate Isilon NAS storage with a Lustre HSM through this solution, and compute nodes can access Isilon storage as a Tier 1 device.

About EMC

EMC Corporation is a global leader in enabling businesses and service providers to transform their operations and deliver IT as a service. Fundamental to this transformation is cloud computing. Through innovative products and services, EMC accelerates the journey to cloud computing, helping IT departments to store, manage, protect, and analyze their most valuable asset—information—in a more agile, trusted, and cost-efficient way. Additional information about EMC can be found at www.EMC.com.