

IBM Spectrum Virtualize inter-node communication support over Ethernet-based RDMA

Enabling Spectrum Virtualize Ethernet data centers for high availability

Overview

Challenge

Until now Spectrum virtualize ethernet data centers had dependencies on Fibre Channel interconnect that made it practically difficult to deploy, especially dual site cluster configurations.

Solution

With the introduction of inter-node communication support over Ethernet-based RDMA for Spectrum Virtualize, Fibre Channel dependencies are removed. That made Spectrum Virtualize Ethernet datacenters deployment possible for Business Resiliency services; also, a mixed configuration can be achieved that includes a Primary FC site and Secondary Ethernet site connected using IP link.

The objective of this white paper is to describe IBM® Spectrum® Virtualize solutions for high availability clusters in an Ethernet environment.

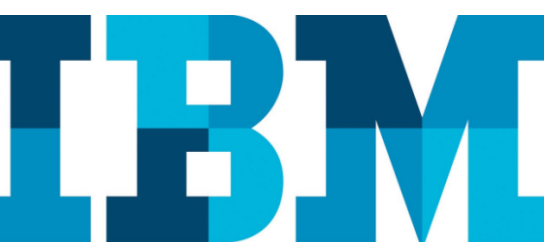
Previously Spectrum Virtualize based Ethernet data centers have never seen themselves as a complete fit-in for business continuity solutions that demand business resiliency that is, high availability. Spectrum Virtualize Ethernet data centers had dependency on Fibre Channel (FC) interconnect and Fibre Channel over IP (FCIP) routers for IP links in long-distance dual-site use cases (that is IBM HyperSwap® and enhanced stretched cluster).

Expansion of Fibre Channel based primary datacenters to long distance dual site Fibre Channel datacenters also seems to be a farfetched dream for small and medium business as it demands FCIP router for IP links, which is not cost-effective.

Inter-node communication support over Ethernet based RDMA is introduced in IBM Spectrum Virtualize version 8.2.1 that addressed above challenges. Such challenges to Spectrum Virtualize ethernet data centers and solutions are described in detail below.

Introduction

Until recently, Spectrum Virtualize was never looked upon as a complete high availability solution in an Ethernet environment. This is due to FC dependency for inter-node communication. This dependency imposed several limitations on deploying high availability Spectrum Virtualize clustering solutions in Ethernet-based data centers. Such limitations are described in the following sections.



Previous challenges with Spectrum Virtualize Ethernet data centers

In the past, IBM Spectrum Virtualize Ethernet data centers were subjected to many limitations in high availability configurations, such as:

- **Standard topology:** IBM System Storage® SAN Volume Controller (SVC) clusters previously needed Fibre Channel interconnect for inter-node communication. Clustered IBM Storwize® configurations with multiple I/O group environment also required Fibre Channel interconnect.
- **Short-distance inter-node communication:** Setting up dual-site clusters in campus like (short distance) environment, only Fibre Channel interconnect was allowed for Inter-node communication, despite the presence of Ethernet network in between.
- **Long-distance inter-node communication:** For long-distance multisite clusters, the demand for Fibre Channel connectivity was typically fulfilled by adding FCIP routers for IP links.

FCIP routers are needed in setting up long-distance Fibre Channel inter-node IP links which is obviously a costly affair. Only enterprise-level business industries could afford it. This made long-distance dual-site setup a farfetched dream for low- and mid-segment businesses.

Also following use case scenarios were not possible earlier with Spectrum Virtualize:

- **Migration of existing FCIP-based dual site clusters to Ethernet-based higher bandwidth clusters:** For long distance dual site Fibre Channel data center, when existing FCIP router's maximum supported bandwidth became obsolete, it was not possible to switch to an Ethernet-based dual-site setup without any dependency on Fibre Channel interconnect. For example, an existing 1G FCIP router needs a replacement with a 10G FCIP router. Buying a new FCIP router that supports higher bandwidth might cost too much.
- **Upgradation of existing FCIP-based links with higher bandwidth RDMA-based Ethernet links:** For a long-distance dual-site environment that has FCIP links between sites, it was not possible to switch to Ethernet-based Remote Direct Memory Access (RDMA) links. This causes customers to buy higher bandwidth FCIP routers with additional significant cost.

Till now, a major factor that is limiting the adoption of Ethernet for storage area network (SAN) is performance as compared to that of Fibre Channel infrastructure. A primary reason for such difference in performance is that the TCP-IP stack consumes processor cycles while copying transferred data

buffers. This activity is Processor-intensive and causes higher latencies in I/O than that of Fibre Channel.

1.2. Solution

Introduced in release 8.2.1, Spectrum Virtualize now enables complete Ethernet-based data centers without any dependencies on Fibre Channel interconnect. Inter-node communication over Ethernet based RDMA is available for deployment in customer environment, since release 8.2.1 (end of 2018).

RDMA is a technology that allows computers in a network to exchange data in the main memory without involving the processor, cache, or operating system of either computer. Similar to Direct Memory Access (DMA), RDMA improves throughput and performance because it does not use a lot of resources. RDMA also facilitates a faster data transfer rate and low-latency networking. RDMA technology supports a feature called zero-copy networking that makes it possible to read data directly from the main memory of one computer and write that data directly to the main memory of another computer. Using RDMA technology, Spectrum Virtualize Ethernet based data centers certainly have overcome major performance issues over traditional TCP/IP based data transfer and removed Fibre Channel dependency completely.

Now with inter-node communication support over Ethernet-based RDMA, Ethernet data centers support the following high availability configurations:

- Standard topology
- IBM HyperSwap (SVC and Storwize) and enhanced stretched cluster (SVC only)
 - No inter switch link (ISL)
 - Short distance ISL
 - Long distance ISL

Also, following scenarios are now possible with Spectrum Virtualize:

- Migration of existing FCIP-based dual-site cluster to Ethernet-based higher bandwidth cluster
- Upgradation of existing FCIP based links with higher bandwidth RDMA-based Ethernet links

Note: HyperSwap, enhanced stretched cluster, and migration solutions are supported only through SCORE request. Contact your local IBM support to open a SCORE request.

Advantages

Inter-node communication over Ethernet based RDMA uses standard Ethernet switches and RDMA-capable network interface cards (NICs). The performance gained by RDMA-capable NICs outweighs the cost of NICs and this cost is comparable with that of Fibre Channel capable NICs. Besides, it can be deployed over low cost 10G infrastructure as well. This makes it a low-cost solution and its performance is far beyond contemporary Ethernet protocols by means of buffer copy avoidance.

Planning for deployment

Inter-node communication over Ethernet-based RDMA

Software supported

- Spectrum Virtualize version 8.2.1 and later

Hardware supported

- IBM FlashSystem 9100, SV1, Storwize V7000, Storwize V5100 (supported HW)

Network requirements

- 25G RDMA capable Chelsio and Mellanox adapters
- Ethernet switches for example, Cisco, Mellanox, Arista, and so on.
- 10G, 25G, and 100G network environments are supported.

While planning for Spectrum Virtualize Ethernet data centers deployment, it is of utmost importance to understand network deployment configurations and use cases.

Network deployment configurations

This section provides a few examples of network configurations that are applicable for both SVC and Storwize nodes.

Direct-attach network configuration

Figure 1 illustrates direct-attach network configuration required for inter-node Ethernet communication. In this configuration, nodes are inter-connected with each other directly over Ethernet based RDMA links and so are hosts with nodes. (Optionally, hosts' connectivity can also be established by switching fabrics.) Both nodes have connectivity, over management Ethernet ports, to an IP quorum instance with configuration data stored. Connectivity to IP Quorum is mandatory when SVC nodes or Storwize I/O groups are present. Such configurations are scalable with introduction of additional network switches.

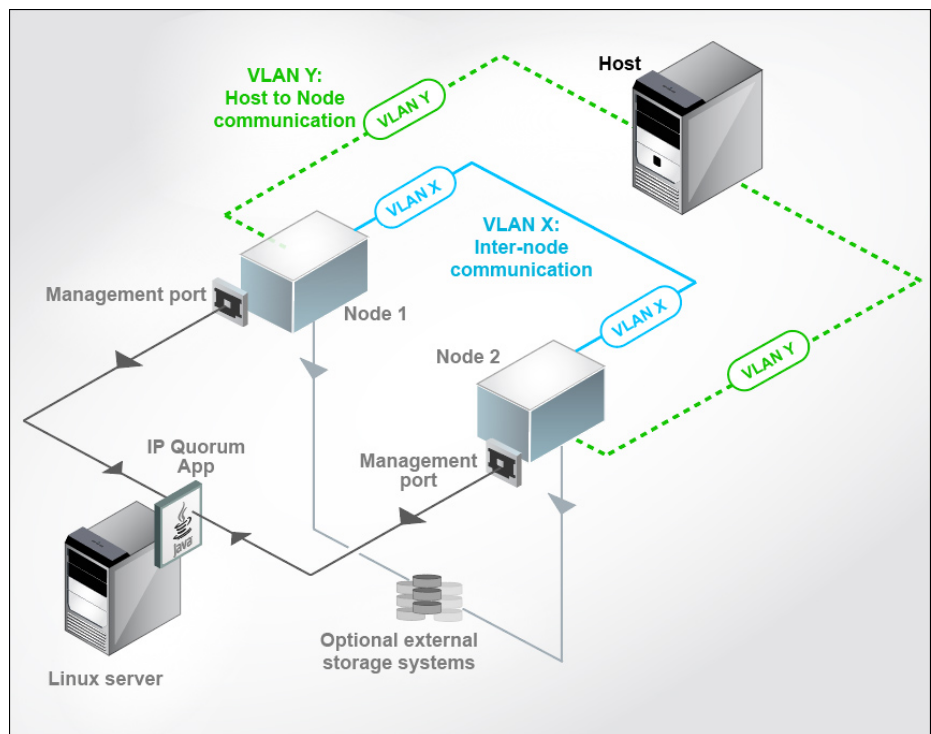


Figure 1. Direct-attach network configuration

Medium SAN network configuration

Inter-node connections are separated from the *host to node* connection using virtual LANs (VLANs). Also, it is worth mentioning that inter-node Ethernet connectivity can be done *only* over identical ports. For example, Node 1's port 4 can make an inter-node connection with Node 2's port 4 only, and not with port 5 or 6. This is applicable in every network configuration.

Figure 2 illustrates a medium SAN configuration. Two Ethernet switches are used to provide redundant fabrics. Both switches are *not* connected with each other by ISLs or any other means. The control enclosures and each host system are connected to both Ethernet switches over RDMA Ethernet links in a redundant fashion. This is to ensure that a connectivity issue in one fabric would not affect the other redundant fabric, providing continuous high availability.

If an external storage is used with the system, you can connect as shown in this illustration. Both nodes have connectivity to an IP quorum instance, with configuration data stored over management Ethernet ports. Connectivity to the IP quorum is mandatory when SVC nodes or the Storwize I/O groups are present. Under such configurations, an individual fabric might face network congestion, if a huge number of hosts are driving the I/O groups in the fabric.

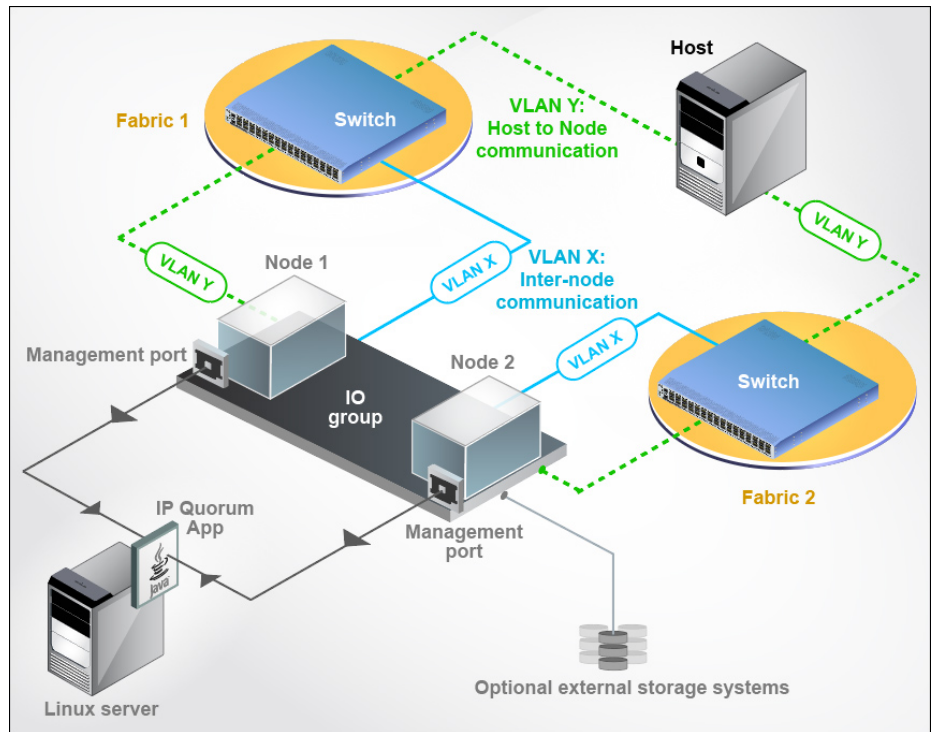


Figure 2. Medium SAN network configuration

Larger SAN network configuration

Figure 3 illustrates a larger sized SAN configuration with external storage systems. The Ethernet SAN fabric consists of switches which are interconnected with ISLs in a *leaf and spine* fashion.

For redundancy, connect each control enclosure and external storage system to two similar fabrics. This is to ensure that a connectivity issue in one fabric would not affect the other redundant fabric and provides continuous high availability. The example fabric attaches the control enclosures and the storage systems to the spine switch.

Both nodes have connectivity to an IP quorum instance, with configuration data stored over management Ethernet ports. Connectivity to the IP quorum is mandatory when SVC nodes or the Storwize I/O groups are present.

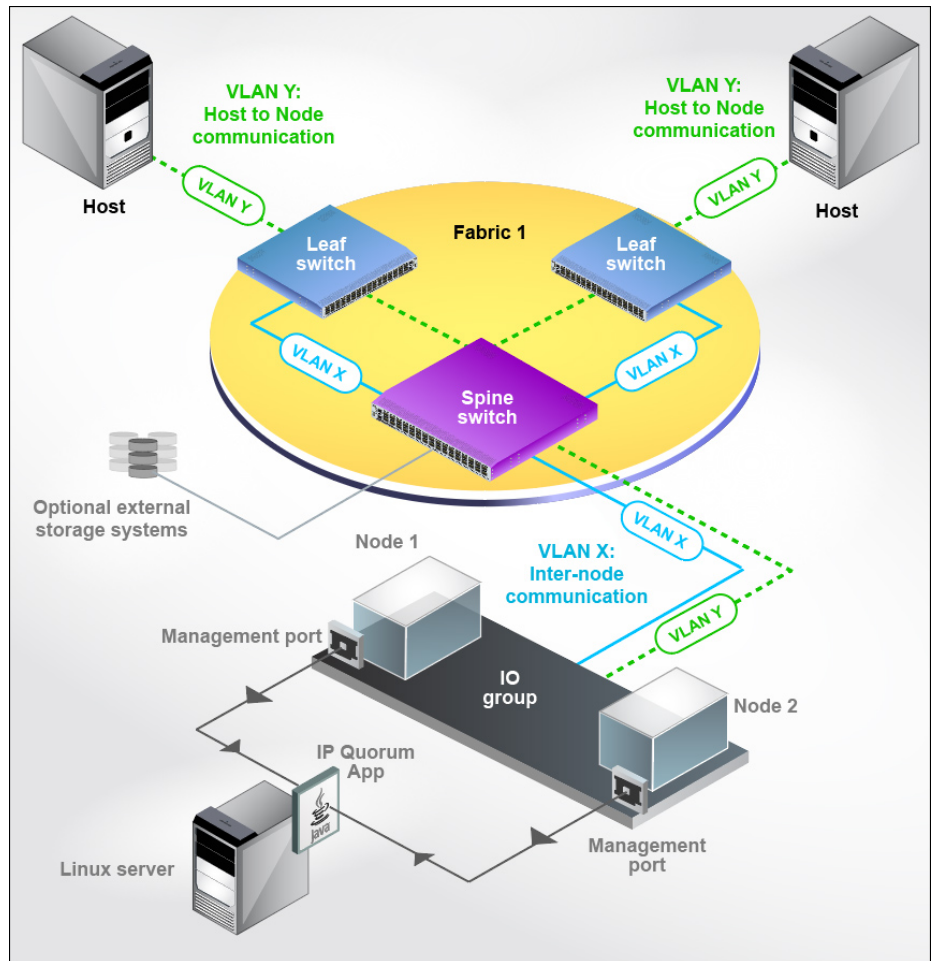


Figure 3. Large SAN network configuration

Use cases

Some of the most common use cases that can be thought of are in terms of sites and distance in between. A standard topology use case dictates a single-site Ethernet data center, while HyperSwap and enhanced stretched cluster use cases mandate dual-site Ethernet data centers. Dual-site Ethernet data centers may or may not have ISLs in between and this depends upon the distance between the sites.

Standard topology

The standard topology refers to single-site Ethernet data centers that might include switching fabrics.

Figure 4 shows the use case of an Ethernet based inter-node communication over RDMA. This setup has redundancy with two hosts, two Ethernet switches, two nodes, and a backend storage.

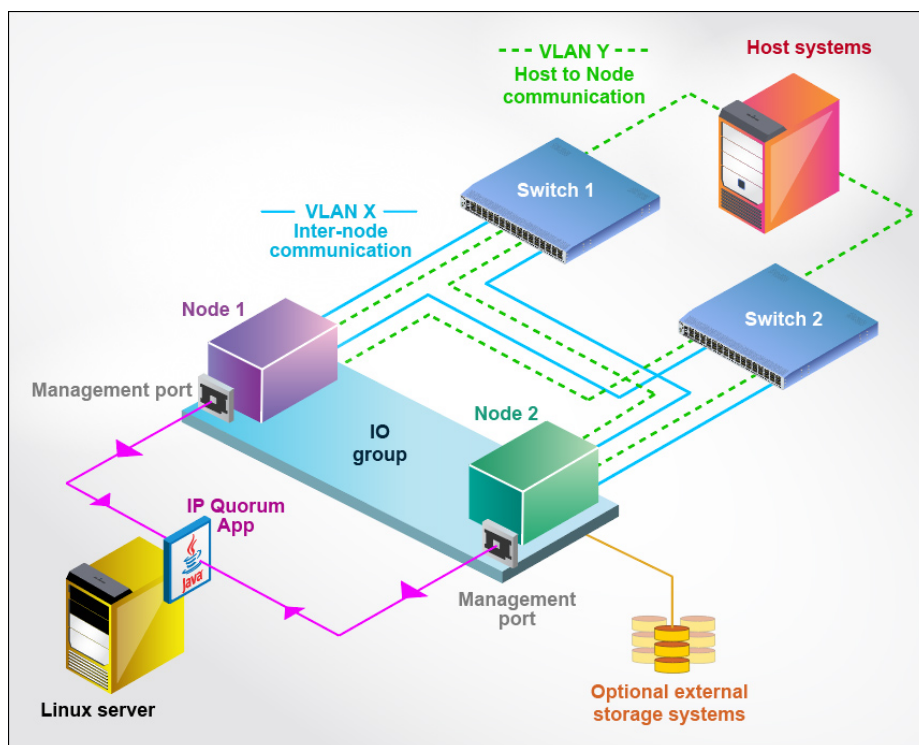


Figure 4. Standard topology

HyperSwap (SVC and Storwize) and stretched cluster (SVC)

HyperSwap and stretched cluster both topologies cover dual-site environment. Based on distance between sites, it can further be categorized as:

- No ISL
- Short-distance ISL
- Long-distance ISL

Note: The following diagrams depict connectivity between networking components only. However, it is required to consult with IBM support for actual connectivity.

No ISL

An environment without ISL is showcased by Figure 5, where each host system and all SVC nodes are connected with each of the switches present in the fabric at both sites. Each control enclosure has connectivity to the IP quorum instance that has the configuration data stored using management ports, and optionally, have storage systems as the backend disks. Switches present in the fabric for each site are not inter-connected with ISLs. The maximum distance between sites, under this use case, is up to 300 meters.

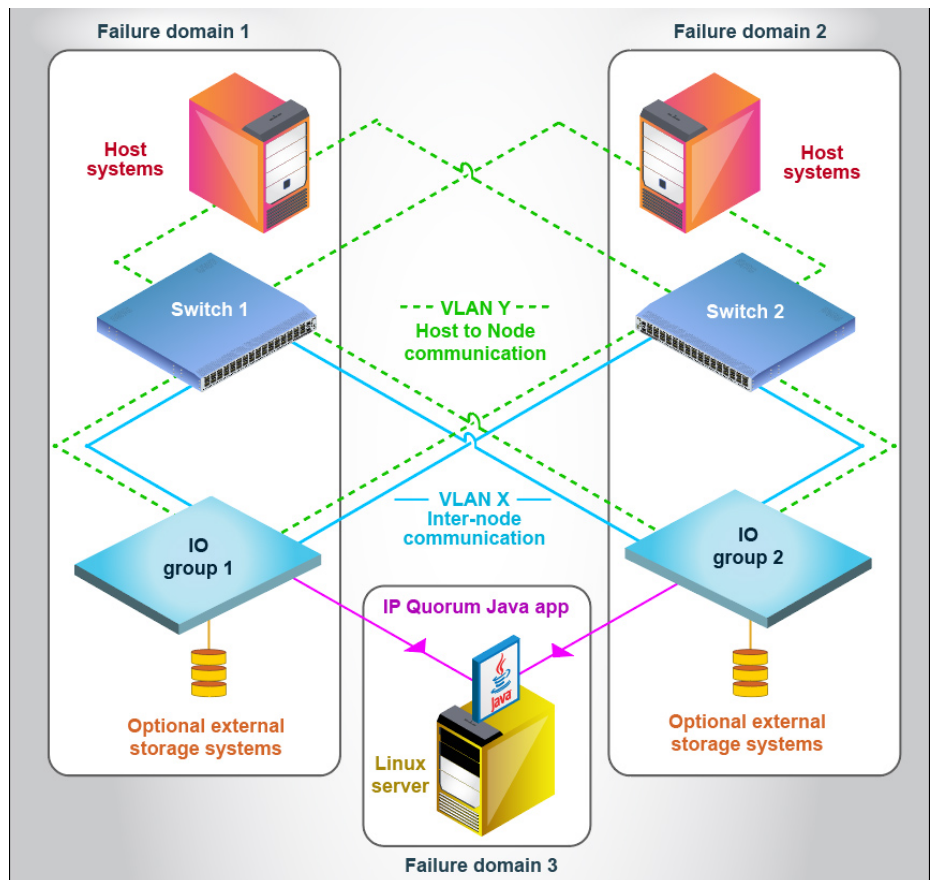


Figure 5. No ISL configuration

Short-distance ISL

Figure 6 showcases a short-distance ISL environment where switches present in each site's fabric are connected with each other over ISLs. The general guideline for ISL connectivity is that there must be as many ISLs between switches as connected RDMA-capable Ethernet source ports are present. The maximum distance considered for such an environment is up to 10 km.

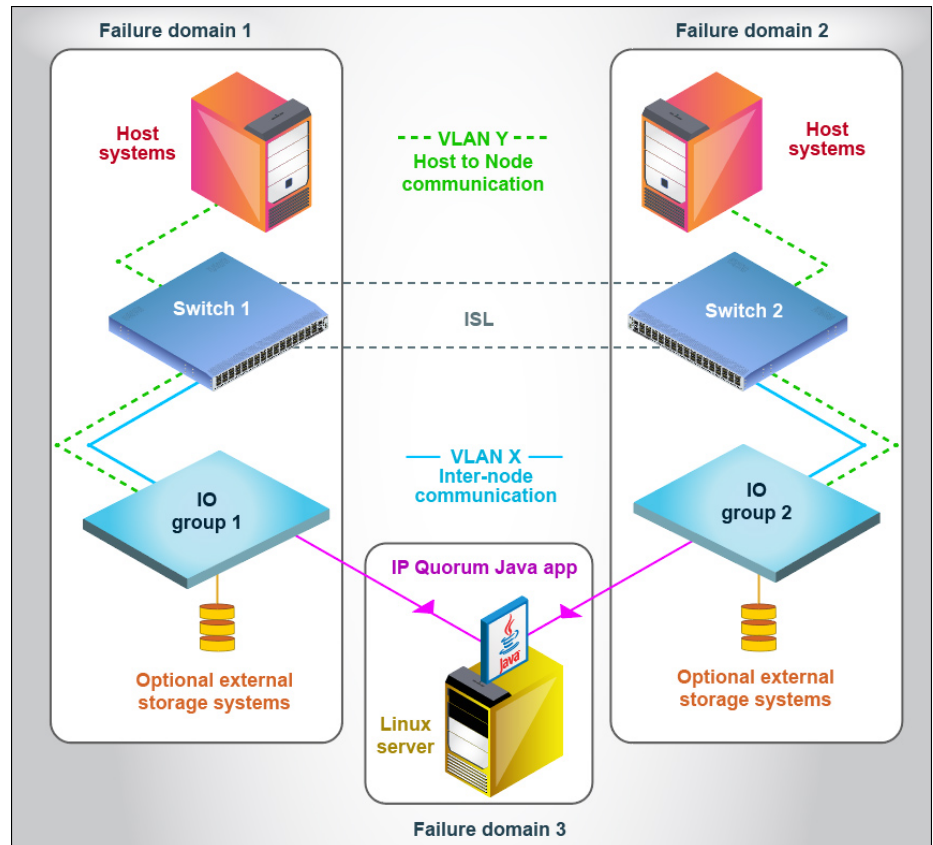


Figure 6. Short distance ISL configuration

Long-distance ISL

A long-distance ISL use case is depicted in Figure 7, where ISLs are connected over either stretched layer 2 network or layer 3 network in between two sites. This ISL connectivity could consist of coarse wavelength division multiplexing (CWDM) or dense wavelength division multiplexing (DWDM), packet-switched, or Virtual Extensible LAN (VXLAN) networks. That would be driven by customer's existing environment or preferences.

DWDM and CWDM methods are applicable for L2 network.
Packet switched or VXLAN methods are deployed for L3 network.

Maximum distance covered under this configuration is up to 100 km.

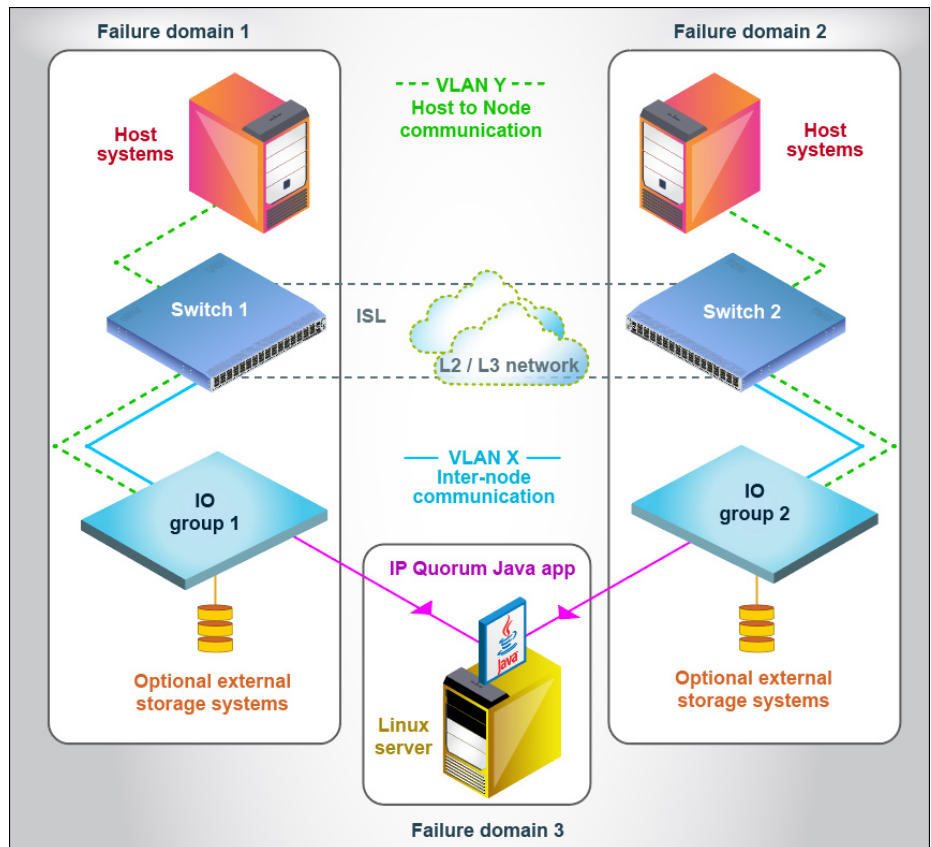


Figure 7. Long distance ISL configuration

Now that we have gone through the various network configurations and use-cases, a solution designer should consider the following factors and streamline storage infrastructure.

IP Quorum connectivity

Ethernet data centers mandates connectivity with IP quorum. IP quorum is a Java™ application which is used in Ethernet data centers to resolve failure scenarios (split brain) where half of the nodes or the enclosures on the system become unavailable. In addition to resolving split brain, IP quorum also stores cluster configuration data which can be used in cluster recovery scenarios, that is tier 3 and tier 4 recovery.

For more information, visit [here](#):

1. Navigate to [IBM Knowledge Center](#) and click **Select a product**.
2. Search for a Spectrum Virtualize product (for example, **IBM Storwize V7000**) and click it.
3. Select version 8.2.1 or later.
4. Click **Configuring**.
5. Click **Configuration details**.
6. Click **Quorum disk**.
7. Click on **IP quorum configuration**.

RDMA adapters, ports, and VLANs

Before deployment of Ethernet data centers, you need to thoroughly plan how many RDMA-capable Ethernet adapters are needed to support the data center that includes inter-node and host-to-node connectivity. It is expected to find the number of RDMA-capable Ethernet adapters first, then work out on the selection of offered platforms which have sufficient PCIe slots for RDMA adapters.

Refer to the following general guidelines and recommendations:

- Inter-node connectivity is established between identical ports of nodes.
- The protocol technology [such as RMDA over Converged Ethernet (RoCE) or iWARP] on the source and destination adapters must be the same.
- Use different subnets for host-to-node and inter-node connectivity.
- For inter-node connectivity in layer 2 network, all identical Ethernet ports on each node *must* have IP addresses from the same subnet.
- As inter-node connectivity can be established between identical ports only, such ports must be connected within the same switching fabric.
- Inter-node traffic within a single site *must not* pass through ISL in any way.
- Consider the following recommendations for VLAN configuration:
 - The local and remote port virtual LAN identifiers must be the same.
 - Use a VLAN to create a physical separation of networks for unrelated systems, wherever possible. All identical ports across nodes that are used for inter-node communication must be assigned with the same VLAN ID and ports that are used for host attachment must have a different VLAN ID. If you plan to

use VLAN to create this separation, you must configure VLAN support on all the Ethernet switches in your network before you configure RDMA-capable Ethernet ports on nodes present in the system.

- On each switch in your network, set VLAN to the Trunk mode and specify the VLAN ID for RDMA ports that will be in the same VLAN.
- In addition, if VLAN settings for a RDMA-capable Ethernet port needs to be updated, these settings cannot be updated independently of other configuration settings. Before updating VLAN settings on specific RDMA-capable Ethernet ports, you must unconfigure the port, make any necessary changes to the switch configuration, then reconfigure the RDMA-capable Ethernet ports on each node in the system.
- A minimum of two dedicated RDMA-capable Ethernet ports are required for inter-node communications to ensure best performance and reliability. These ports must be configured for inter-node traffic only and must not be used for host attachment, virtualization of Ethernet-attached external storage, or IP replication traffic on a sharing basis.

Inter-node connections are separated from a host-to-node connection using VLANs. Also, it is worth mentioning that an inter-node Ethernet connectivity can be done *only* over identical ports. For example, Node 1's port 4 can make inter-node connection to Node 2's port 4 only, not to port 5 or 6. This is applicable on every network configuration.

Note: The rules mentioned in this section represent fully tested configurations as of today (October 2019). In future, there might be more relaxed configurations.

1. Navigate to [IBM Knowledge Center](#) and click **Select a product**.
2. Search for a Spectrum Virtualize product (for example, **IBM Storwize V7000**) and click it.
3. Select version 8.2.1 or later.
4. Click **Configuring**.
5. Click **Configuration details**.
6. Click **Configuration details for using RDMA-capable Ethernet ports for node-to-node communications**.

Switching infrastructure

Switching infrastructure is a crucial element in any Storage Area Network solutions. Switching infrastructure provides powerful features that can help administrators organize and manage the storage network, including storage virtualization, provisioning, inter switch link trunking, performance planning etc. Thus, it is much needed to consider following factors while planning for it.

Dedicated vs shared switches in fabric

Ethernet data center might need a dedicated switching infrastructure for node-to-node communication rather than a shared one. Factors such as I/O traffic driven by application hosts would govern the design of the fabric. For example, if an application workload is expected to overload the switch bandwidth in a system, the host-to-storage connectivity is recommended to have a redundant fabric separate from that used by the inter-node connectivity. Deploying such dedicated switches for host to storage and inter-node connectivity would make sure that application driven I/O workload does not congest switch ports dedicated to inter-node connectivity in fabric. This setup would keep inter-node connectivity isolated from host connectivity because both application-driven workload and inter-node traffic would have their own dedicated redundant fabrics.

If it is intended to use a shared switching fabric, customers would need to ensure that quality of service (QoS) guarantees that inter-node traffic is not affected by any other traffic present in the same switching fabric.

L2 versus L3 switches

Switch selection is totally based out of customer environment. For example, standard topology or small-scale setups could make use of L2 switches because MAC-based broadcasting is expected to be faster for relatively smaller office networks. This stands true for networks without ISL and short-distance ISL-based networks as well for up to 10 km of distance where the round-trip time (RTT) does not exceed 1 ms. L3 switches can also be used for short distance ISL-based networks.

L3 switches and routers are recommended for long-distance ISL-based networks because they are equipped with deeper buffer pockets.

ISL connectivity

It is essential to understand application workload for both HyperSwap and enhanced stretched clusters, as workload is going to be transferred over ISLs. Normally when both sites are working, host write operations requested at any site would eventually be served by active site with primary volumes and sync up data would be transferred over ISLs. When storage failure in one site occurs, the host application might fail over to another site or host read requests would be served from the surviving site over ISLs.

Host throttling can be used to make sure that ISL does not get exhausted. At the same time, it is also required to keep the ISL bandwidth big enough to handle host traffic as well as inter-node traffic. If deployed, the ISL's bandwidth is too low to handle the application host's and inter-node traffic, and multiple ISLs are expected to be established between sites.

Link quality

There are a few more factors such as RTT, packet drops, network jitter, and so on that affect the link quality and performance. The maximum allowed RTT for inter-node connectivity is up to 5 ms. Maximum percentage of packet drops supported are up to 0.5 %. The link quality governs the overall performance and reliability of the HyperSwap/Stretched Systems solution.

It is also worth mentioning that proposed solution works well not only with 25G network fabric but also can be deployed in 10G network fabrics. That means such configurations can be deployed using an existing 10G network infrastructure.

Firewall settings

Deployment environment's firewall configuration must ensure that traffic is open for the TCP port **21455** and UDP ports **4791**, **21451** and **21452**. Inter-node communication over RDMA-capable Ethernet connections, uses the TCP port **21455** for data traffic and UDP port **21451** and **21452** for discovery services on the system. If deployed RDMA adapters adhere to the RoCE V2 protocol, ensure that traffic is also open for the UDP port **4791**.

Additionally, RDMA-capable Ethernet ports use Internet Group Management Protocol (IGMP) for group multicast communication for discovery service among nodes. This is applicable for L2 networks. Thus, the firewall configuration must enable IGMP traffic for redundant site configurations.

Configuration

Things to remember

While configuring IP for RDMA-based inter-node connectivity:

- Set the switch port to the Trunk mode for VLAN.
- Set the VLAN ID first on the switch side and then on the node port.
- Remove the VLAN setting on the node first and then on the switch, if needed.
- Perform add, update or remove VLAN operations for one port at a time and as a best practice, maintain 15 seconds of time interval for adding, updating, or removing the VLAN ID.

This section talks about the configuration of three use cases of the Ethernet-based RDMA inter-node communication:

- Standard topology
- HyperSwap
- Enhanced stretched cluster.

Here you can find the actual CLI configuration and the GUI based configuration along with screen captures to find how they look and how they are deployed.

Ethernet-based RDMA inter-node connectivity can be configured using the CLI or a service assistant GUI.

Configuration using CLI

This section explains the flow of tasks that should be performed for configuring inter-node connectivity over Ethernet-based RDMA for IBM Spectrum Virtualize family products.

Lab setup used for the configuration:

- Nodes: Two SV1 (Cayman)
- Adapters: Two 25G Mellanox per node
- Fabric: Cisco Nexus3000

Perform the following steps to configure inter-node connectivity over Ethernet-based RDMA using CLI and service assistant GUI. You can use the same steps to configure HyperSwap as well as enhanced stretched cluster:

1. Install RDMA capable adapters supporting identical technology (either iWARP or RoCE) across all nodes, in the same PCI slot.
2. Install IBM Spectrum Virtualize 8.2.1.0 or later code version on all nodes in a system.
3. Verify that all the installed adapters are listed under the `sainfo lshardware` CLI.
4. Verify that all Ethernet adapter ports are listed under the `sainfo lsnodeip` CLI.

CLI example:

```
IBM_2145:ibm-svc:superuser> sainfo lsnodeip
port_id  rdma_type  port_speed  vlan  link_state  state  node_IP_address  gateway  subnet_mask
1        1Gb/s      1Gb/s      1     active      unconfigured
2        1Gb/s      1Gb/s      1     inactive   unconfigured
3        1Gb/s      1Gb/s      1     inactive   unconfigured
4        RoCE       25Gb/s     1     active      unconfigured
5        RoCE       25Gb/s     1     active      unconfigured
6        RoCE       25Gb/s     1     active      unconfigured
7        RoCE       25Gb/s     1     active      unconfigured
IBM_2145:ibm-svc:superuser>
```

Figure 8. Ethernet adapter ports CLI view

Service assistant GUI example:

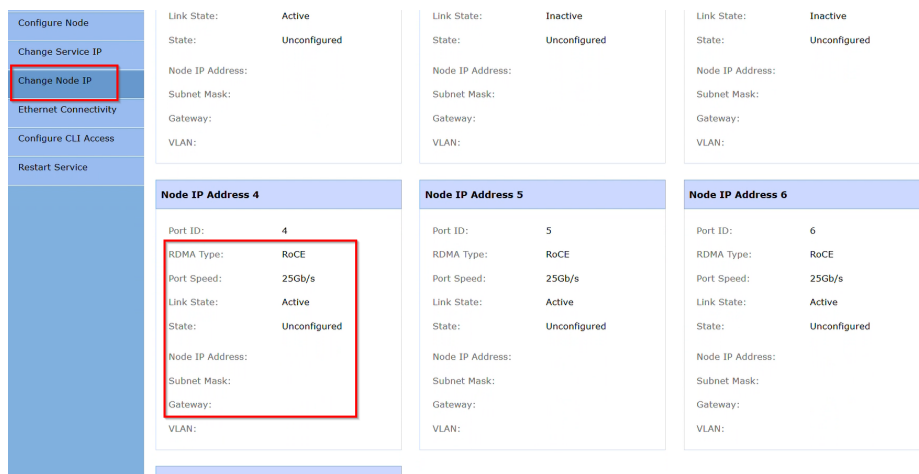


Figure 9. Service assistant GUI change node IP view

Note: Here all the RoCE adapters are listed under the **Change Node IP** tab with the default status as unconfigured.

5. Assign IP to these adapters on all nodes in the system to establish node-to-node links. VLAN can also be configured for each link to isolate from others.
 - a. Configure IPs with or without VLAN on all nodes in a system:

```
satask chnodeip -ip <IPv4 address> -mask <subnet mask> -gw <gateway> -vlan <vlan id> -port_id <RDMA port for establishing Inter-node links>
```

CLI example for configuring IP without VLAN:

```
IBM_2145:ibm-svc:superuser> satask chnodeip -ip 192.168.100.50 -mask 255.255.255.0 -gw 192.168.100.1 -port_id 5
IBM_2145:ibm-svc:superuser>
```

Figure 10. Node IP assignment CLI task

GUI example for configuring IP with VLAN:

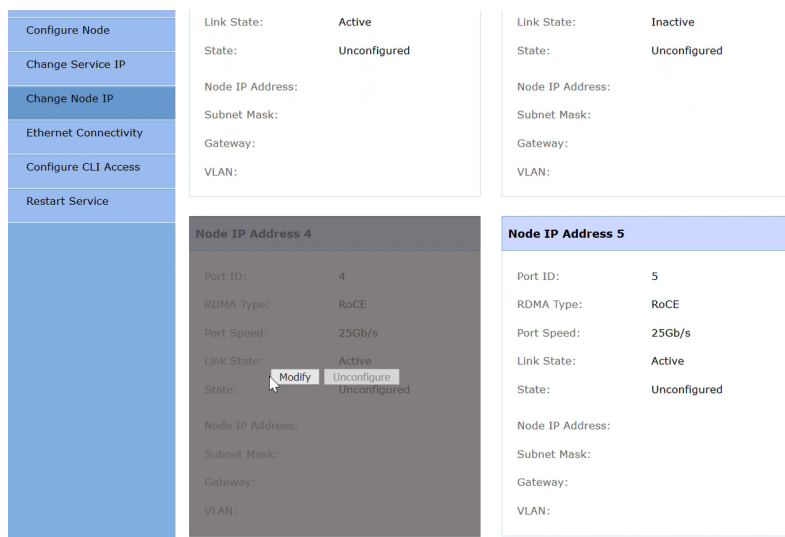


Figure 11. Service assistant GUI change node IP configuration

- b. Select the adapter port and click **Modify**, and enter the IP details, and click **Save**.

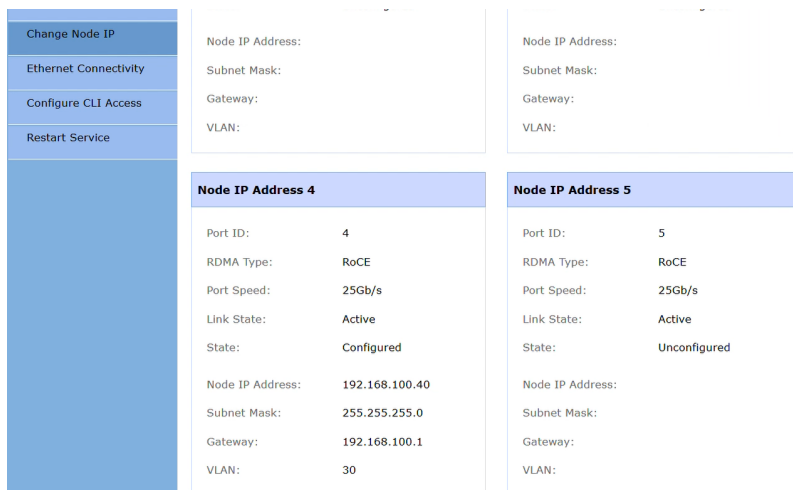


Figure 11. Service assistant GUI change node IP updated view

- c. List the assigned IPs using the `sainfo lsnodeip` CLI or from the same **Change Node IP** tab from the service assistant GUI.

CLI example:

```
IBM_2145:ibm-svc:superuser>sainfo lsnodeip
port_id  rdma_type port_speed vlan link_state state      node_IP_address gateway      subnet_mask
1                1Gb/s      active  unconfigured
2                inactive  unconfigured
3                inactive  unconfigured
4          RoCE    25Gb/s    30   active  configured 192.168.100.40 192.168.100.1 255.255.255.0
5          RoCE    25Gb/s    active  configured 192.168.100.50 192.168.100.1 255.255.255.0
6          RoCE    25Gb/s    active  configured 192.168.100.150 192.168.100.1 255.255.255.0
7          RoCE    25Gb/s    active  unconfigured
IBM_2145:ibm-svc:superuser>
```

Figure 12. Ethernet adapter ports CLI view

- d. Assign IP addresses on remote nodes if at least one connection is established among the local and remote nodes.

CLI example for assigning IP to a remote node:

```
IBM_2145:ibm-svc:superuser>satask chnodeip -ip 192.168.100.150 -mask 255.255.255.0 -gw 192.168.100.1 -port_id 6 78FNM0
IBM_2145:ibm-svc:superuser>
```

Figure 13. Remote node IP assignment CLI task

- e. List the assigned IPs for the remote node using the `sainfo lsnodeip <panel name>` CLI, if at least one connection is established among the local and remote nodes.

CLI example of listing IP of remote node:

```
IBM_2145:ibm-svc:superuser>sainfo lsnodeip 78FNM0
port_id  rdma_type port_speed vlan link_state state      node_IP_address gateway      subnet_mask
1                1Gb/s      active  unconfigured
2                inactive  unconfigured
3                inactive  unconfigured
4          RoCE    25Gb/s    30   active  configured 192.168.100.40 192.168.100.1 255.255.255.0
5          RoCE    25Gb/s    active  configured 192.168.100.50 192.168.100.1 255.255.255.0
6          RoCE    25Gb/s    active  configured 192.168.100.150 192.168.100.1 255.255.255.0
7          RoCE    25Gb/s    active  unconfigured
IBM_2145:ibm-svc:superuser>
```

Figure 14. Remote Ethernet adapter ports CLI view

GUI example:

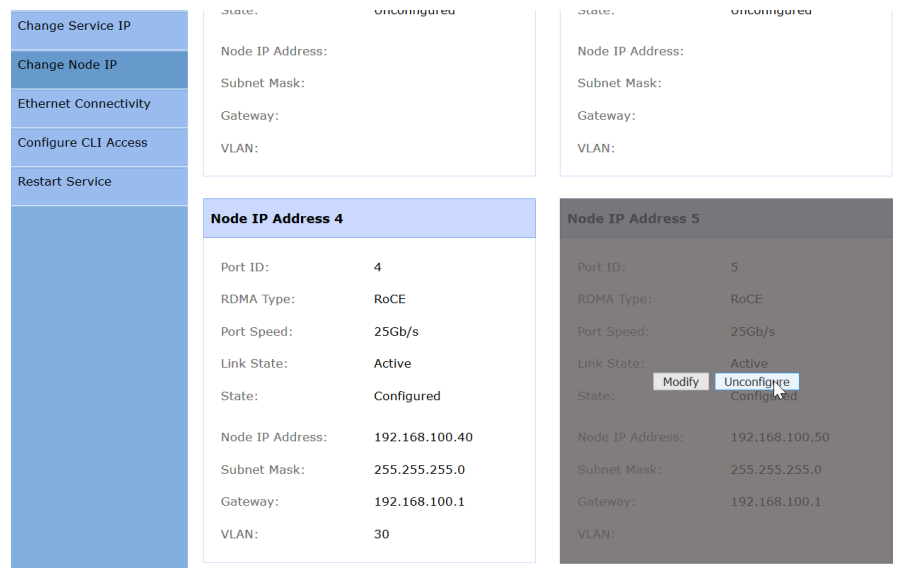


Figure 15. Service assistant GUI change remote node IP update view

- f. Configure IP using the `satask chnodeip -noip -port_id <RDMA port for establishing Inter-node links>` CLI or just by clicking **Unconfigure** from the service assistant GUI.

CLI example:

```
IBM_2145:ibm-svc:superuser> satask chnodeip -noip -port_id 6
IBM_2145:ibm-svc:superuser>
```

Figure 15. Node IP de-assignment CLI task

GUI example:

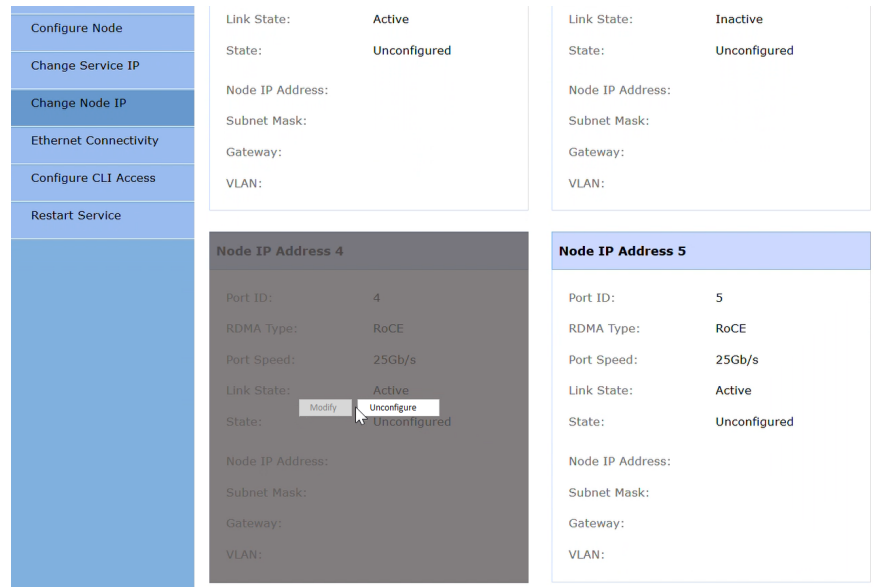


Figure 16. Service assistant GUI change node IP unconfigure view

- After assigning IP addresses on all nodes [with identical network settings (such as IP subnet, VLAN ID) on identical port IDs], verify that they can communicate with each other from the configured ports using the `sainfo lsnodeipconnectivity` CLI.

CLI example:

```
IBM_2145:ibm-svc:superuser>sainfo lsnodeipconnectivity
status local_port_id local_vlan local_rdma_type local_ip_addr remote_port_id remote_vlan remote_rdma_type remote_ip_addr remote_wwnn
Connected:RoCE 4 30 RoCE 192.168.100.40 4 30 RoCE 192.168.100.41 509507680C088B06 78FNMTO
000002032BA11602
Connected:RoCE 5 RoCE 192.168.100.50 5 RoCE 192.168.100.51 509507680C088B06 78FNMTO
000002032BA11602
IBM_2145:ibm-svc:superuser>
```

Figure 17. Ethernet adapter ports CLI view

GUI example for listing the IP link connectivity between nodes of a system using the **Ethernet Connectivity** tab:

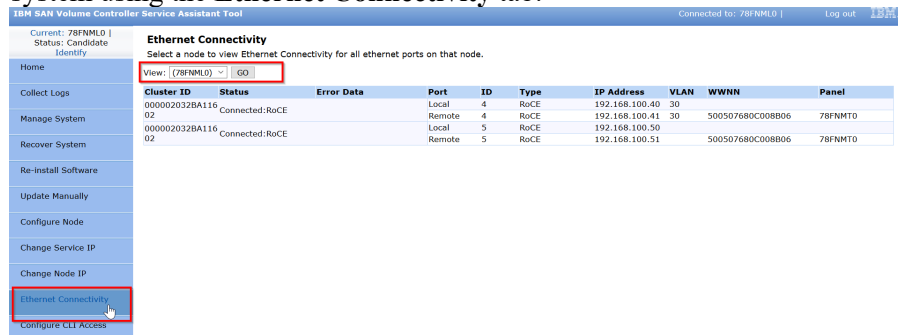


Figure 18. Service assistant GUI Ethernet connectivity

7. If node-to-node port links are not listed, then check the connectivity using the `svctask ping` CLI. Or, check the IP configuration on ports or follow the “Troubleshooting” section of this paper.
8. Entries under `sainfo lsnodeipconnectivity` having the status as *Connected* indicates that all the links are up for node-to-node communication and can see each other using RDMA connection. If any of the links has an error data, refer to the “Troubleshooting” section of this paper.
9. After all the nodes are visible to each other, list all the nodes using the `sainfo lsservicenodes` CLI to view them in the candidate state until the cluster is created.
10. Create a cluster using the usual procedure. This cluster can be of any topology: Standard, HyperSwap, or enhanced stretched systems.

For more information, about the configuration:

1. Navigate to [IBM Knowledge Center](#) and click **Select a product**.
2. Search for a Spectrum Virtualize product (for example, **IBM Storwize V7000**) and click it.
3. Select version 8.2.1 or later.

After performing these steps, based on the use case, you can perform the steps shown in the following table to find additional information about it.

Use case	Action
Standard topology	<ol style="list-style-type: none"> 1. Click Configuring. 2. Click Configuration details. 3. Click Configuration details for using RDMA-capable Ethernet ports for node-to-node communications.
HyperSwap	<ol style="list-style-type: none"> 1. Click Configuring. 2. Click Configuration details. 3. Click HyperSwap system configuration details. 4. Click Configuring an IBM HyperSwap topology system.
Enhanced stretched systems	<ol style="list-style-type: none"> 1. Click Configuring. 2. Click Configuration details. 3. Click Stretched system configuration details
Ethernet port configuration for inter-node communication	<ol style="list-style-type: none"> 1. Click Administering. 2. Click Managing nodes. 3. Click Managing nodes that use RDMA-capable Ethernet ports. 4. Click Changing the IP address on a RDMA-capable Ethernet port.

Troubleshooting

In Ethernet environments, most of the common issues are related with network adapters and network settings.

The `sainfo lsnodeipconnectivity` CLI shows inter-node connectivity links for RDMA-capable Ethernet ports. Issues in inter-node connectivity are displayed in the `error_data` field.

Such error is detectable only when the discovery between at least one identical port of the peer nodes is successful. For cases where discovery between all identical ports of peer nodes failed, the `sainfo lsnodeipconnectivity` CLI will not display any output for inter-node connected links. There could be multiple reasons for node discovery failure.

The following checks might help to troubleshoot issues with adapter or network settings:

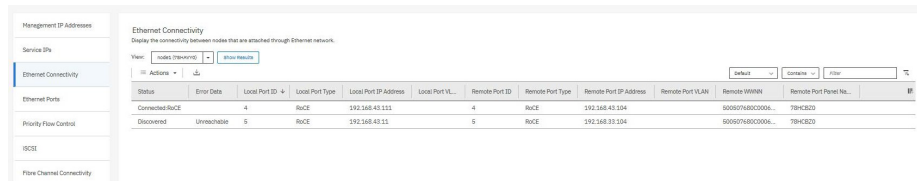
- Check that firewall, if applicable, allows traffic for TCP port 21455 and UDP ports 4791, 21451 and 21452.
- Check if local and remote ports with identical port IDs are configured using same subnet IPs.
- Check the output of the `sainfo lsnodeip` CLI. The `link_state` field displays the link status of the RDMA-capable port. For a working set up, the link status must be **active**. In case it is inactive, check if the port connectivity with the fabric is present or not.
- Check the output of the `sainfo lsnodeip` CLI. The `vlan` field displays the same VLAN ID for identical RDMA-capable ports on both the nodes of the system.
- Check if the `svctask ping` CLI can ping from the IPs assigned at a local port to a remote port IP.

After successful (inter-node) discovery, the `sainfo lsnodeipconnectivity` CLI can display connectivity issues under the `error_data` field.

Sample output of this CLI and GUI View are displayed as below:

```
IDM_2145:ibm-svc:supervisor:sainfo lsnodeipconnectivity
status      local_port_id local_vlan local_rdma_type local_ip_addr remote_port_id remote_vlan remote_rdma_type remote_ip_addr remote_vnwn remote_panel_name
cluster_id  error_data
Connected:RoCE 4          RoCE      192.168.43.104 4          RoCE      192.168.43.111 500507680C00B8A9 78HAY0
0000020325E17152
Discovered  5          RoCE      192.168.33.104 5          RoCE      192.168.43.11 500507680C00B8A9 78HAY0
0000020325E17152 Unreachable
```

Figure 19. Ethernet port connectivity CLI view



The screenshot shows a web interface for 'Ethernet Connectivity'. It includes a sidebar with navigation options like 'Management IP Address', 'Service IPs', 'Ethernet Connectivity', 'Ethernet Ports', 'Priority Flow Control', 'iSCSI', and 'Fibre Channel Connectivity'. The main content area displays a table with columns for Status, Error Data, Local Port ID, Local Port Type, Local Port IP Address, Local Port VLAN, Remote Port ID, Remote Port Type, Remote Port IP Address, Remote Port VLAN, Remote WWNN, and Remote Port Fabric No. The table contains two rows: one for a 'Connected RoCE' connection and one for a 'Disconnected Unreachable' connection.

Status	Error Data	Local Port ID	Local Port Type	Local Port IP Address	Local Port VLAN	Remote Port ID	Remote Port Type	Remote Port IP Address	Remote Port VLAN	Remote WWNN	Remote Port Fabric No.
Connected RoCE		4	RoCE	192.168.43.111		4	RoCE	192.168.43.104		50050768000006...	79HC820
Disconnected Unreachable		5	RoCE	192.168.43.11		5	RoCE	192.168.33.104		50050768000006...	79HC820

Figure 20: Ethernet port connectivity GUI view

The CLI output might display the following possible errors:

- Protocol mismatch
- Unreachable
- Duplicate IP addresses
- Degraded
- VLAN ID mismatch

You can fix these errors by changing the settings for the RDMA-capable ports. For all these errors, you can find the cause and possible actions in IBM Knowledge Center.

For more information on Ethernet port connectivity for inter-node communication, perform the following steps:

1. Navigate to [IBM Knowledge Center](#) and click **Select a product**.
2. Search for a Spectrum Virtualize product (for example, **IBM Storwize V7000**) and click it.
3. Select version 8.2.1 or later.
4. Click **Command-line interface**.
5. Click **Service information commands**.
6. Click **lnodeipconnectivity**.

Summary

To achieve the inter-node communication support over Ethernet-based RDMA, IBM Spectrum Virtualize offers cost-effective and optimized performance solutions for Ethernet environments.

Get more information

To learn more about the Spectrum Virtualize product, contact your IBM representative or IBM Business Partner, or visit the following websites:

IBM Support:
<https://www.ibm.com/support/knowledgecenter/>

About the authors

Shrirang Bhagwat is a lead developer in the Spectrum Virtualize team at IBM Systems Labs, India. He specializes in block storage for Ethernet environment. He can be reached at: shbhagwa@in.ibm.com.

Abhishek Jaiswal is senior developer in the Spectrum Virtualize team at IBM Systems Labs, India. He is actively working in Ethernet environment (RDMA Clustering, iSER/iSCSI HA and so on) for Spectrum Virtualize. He can be reached at: ajaiswa9@in.ibm.com.

Sanjay Tripathi is developer in the Spectrum Virtualize team at IBM Systems Labs, India. He has actively worked in RDMA clustering, Secure Remote Access (SRA) and so on. He can be reached at: satripa2@in.ibm.com.

Aakanksha Mathur is a test lead in Spectrum Virtualize team at IBM System Labs, India. She has expertise in block storage, storage protocols like iSCSI, iSER, FC and is working on Ethernet mission for IBM SVC. She can be reached at: aakanksha@in.ibm.com.



© Copyright IBM Corporation 2019
IBM Systems
3039 Cornwallis Road
RTP, NC 27709

Produced in the United States of America

IBM, the IBM logo and ibm.com are trademarks or registered trademarks of the Internal Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked items are marked on their first occurrence in the information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at ibm.com/legal/copytrade.shtml

Other product, company or service names may be trademarks or service marks of others.

References in the publication to IBM products or services do not imply that IBM intends to make them available in all countries in the IBM operates.



Please recycle
