INFINIDAT
STORING THE FUTURE

INFINIBOX 4.0.X BEST PRACTICES GUIDE FOR SETTING UP THE REPLICATION SERVICE

Lat updated: 02/24/2019

Table of Contents

1 Introduction

1.1 Scope of document

This document provides guidelines and instructions for setting up the InfiniBox Replication service.

1.1.1 Related Documentation

This document is part of a series that includes:

- InfiniBox best practices for setting up the Ethernet network
    - This document describes how to set up the physical network prior to setting up the services
- InfiniBox best practices for setting up the iSCSI service
- InfiniBox best practices for setting up the NAS service
- InfiniBox best practices for setting up the Replication service
    - (This document)

1.1.2 Obsolete documentation

- Deployment consideration for setting up services
    - InfiniBox 4.0 user documentation, and former InfiniBox releases, include an article named **Deployment considerations for setting up services**. This article is now obsolete. The instructions provided in this article were merged into this document.

2 Replication Planning

2.1 Overview

InfiniBox allows customers to replicate volumes, consistency groupss and filesystems across multiple InfiniBox systems. For each dataset, customers can choose whether to use synchronous (**sync**) or asynchronous (**async**) replication.

- **Note:** synchronous replication is not available for filesystems.

InfiniBox asynchronous replication is a snapshot-based solution that allows users to protect their data by replicating it to a remote site, without adding latency to the host I/Os. Users can set an RPO (Recovery Point Objective) as low as 4 seconds if the link quality requirements between the sites are fulfilled.

InfiniBox synchronous replication allows users to protect the data with zero RPO, by sending the I/O to the remote site before acknowledging the host. Synchronous replication has an impact on the latency of the write operations since the acknowledge to the host will be sent only after the data was written in both sites.

- **Note:** read operations will be served locally from the source system so no latency will be added.

InfiniBox synchronous replication depends significantly on the quality of the link between the sites, and may fail due to a network outage. Instead of requiring a full resynchronization of the data, or generating inflated journals of the data that needs to be sent to the DR site once connectivity is restored, InfiniBox uses its InfiniSnap technology to keep track of these changes.

This means that once connectivity is restored, InfiniBox will use Asynchronous replication to bring the DR system up to date, and then switch back to Synchronous replication automatically, without dropping IO.

2.2 Replicating over IP

With the increase in Ethernet performance and reliability, Fibre Channel (FC) is no longer a single option as a replication infrastructure:

- FC requires either a dedicated fiber optics between sites or a xWDM channel dedicated to the Fiber Channel fabric. It also requires another one for redundancy.
- IP WAN links are cheaper, can easily be shared, and are always deployed between production sites and DR sites to facilitate administration, monitoring, clustering, etc.
- All of these differences drive the cost of FC higher than comparable Ethernet.

In addition:

- IP provides a robust framework for devising an optimized protocol for replication
- The ubiquity of IP & Ethernet provides a rich set of algorithms, toolkits and expertise for optimizing various line conditions, as well as troubleshooting

To achieve all of these advantages, InfiniBox uses Ethernet networks to replicate data between sites.

However, just like with FC, using Ethernet for replication does not remove the need to think ahead about proper bandwidth sizing. Customers must plan and test their network to make sure it can sustain the additional bandwidth requirements of replication. This is especially important when using Synchronous replication.

2.3 Understanding the requirements

To allow a successful replication deployment, administrators need to make sure the environment conforms with the three types of requirements:

- **Physical connectivity**
  All the settings required to allow the source and target systems to communicate with each other. These include the LACP interfaces, routers, firewall rules.
- **Logical network settings**
  The logical settings required to allow the source and target system to communicate. These include IP addresses, default gateway / routing rules, etc.
- **Network performance requirements**
  These requirements focus on the “quality” of the network, to support the required bandwidth, latency, stability, etc.

3 Physical connectivity

3.1 Connecting the two sites

The two sites must be connected using any form of physical connectivity, that can sustain the required bandwidth and latency (See more below). Each system will be connected to a local switch fabric, using one or more ports in an LACP (Link Aggregation Control Protocol). The switch will allow it to access the remote site, usually passing through a combination of routers, firewalls and VPNs.

3.2 Connecting InfiniBox to the replication network

The InfiniBox on each site must be connected to the switch that has access to the link between the sites. This connectivity must be resilient to overcome single point failures. The following instructions should be following on each site.

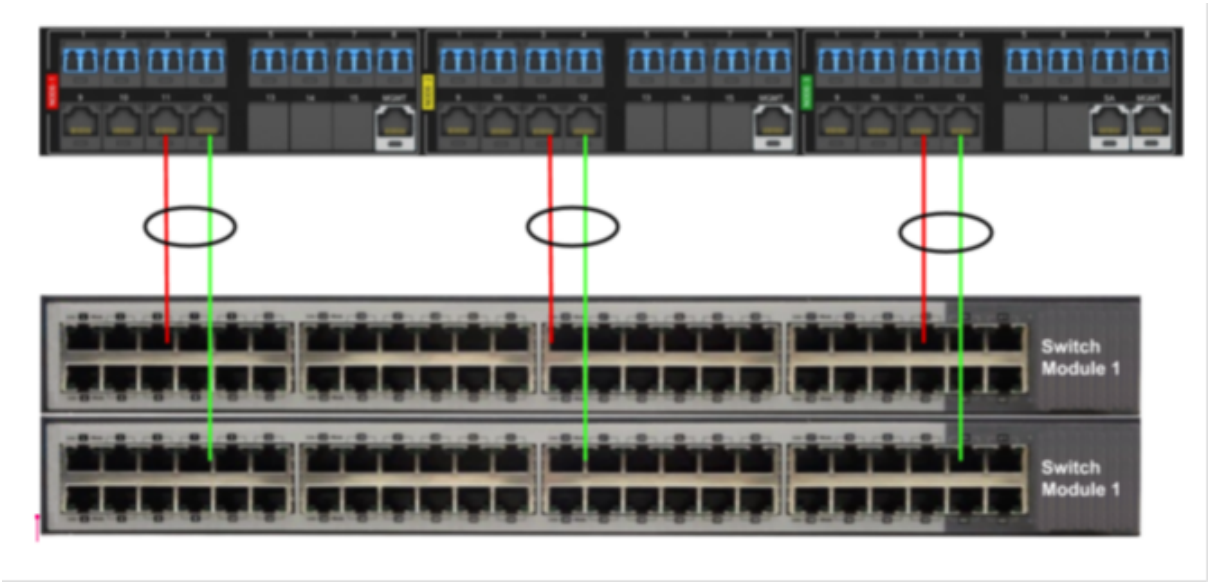
3.3 Switch requirements

It is also recommended to set the ports connected to InfiniBox to account for changes using a spanning-tree algorithm. For many network vendor implementations, this will be called “portfast”, “edge” or “edge-port” in the switch configuration semantics.

Use two switches that support creating LACP port-group (AKA a LAG) that spread across the switches. This is also known as Port Channel. Typically, stacked switches support such configuration, but some non-stacked switches also support this (e.g. Cisco Nexus Virtual Port Channel).

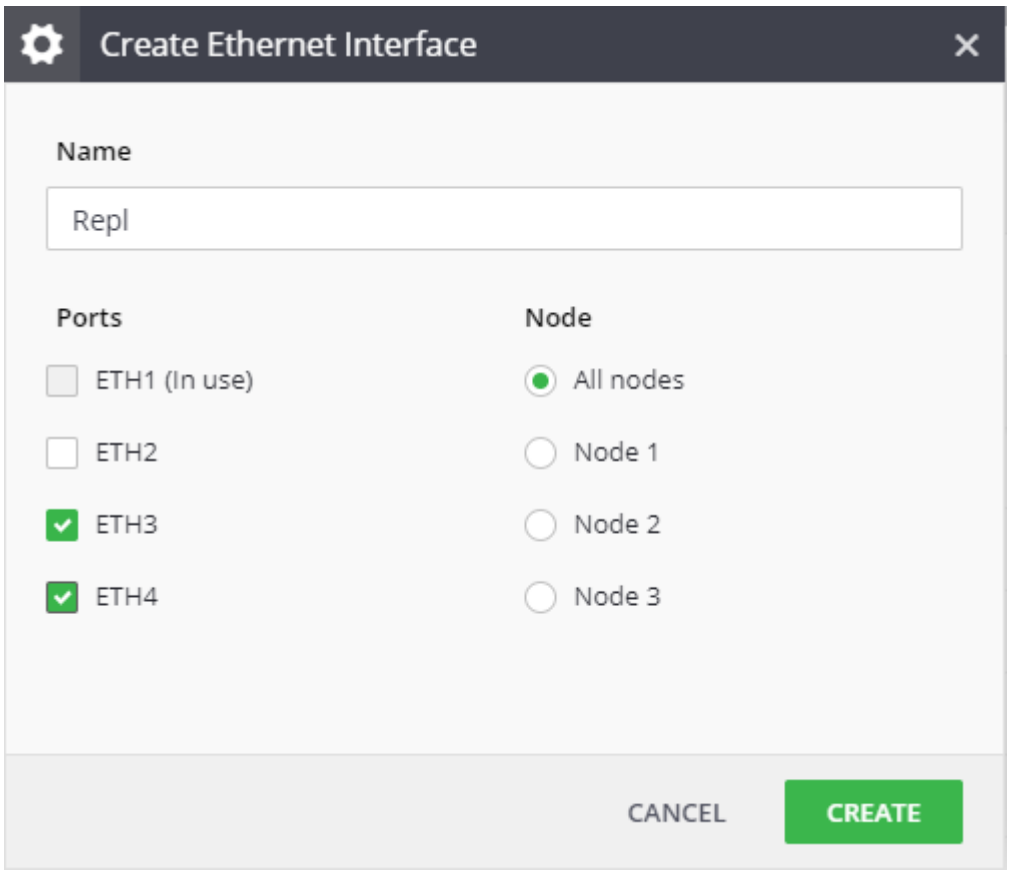
3.3.1 Steps

1. Connect to the switches:



   a. Select one port from each node (ETH3 in the example), and connect it to switch module 1
   b. Select one port from each node (ETH4 in the example), and connect it to switch module 2
2. Configure a LAG for each couple of ports in the switch, as indicated above.

3. Create a new Port Group in InfiniBox, with ETH3 and ETH4 ports from every node.



3.4 WAN acceleration

Some customers deploy network accelerators to compress data over the WAN. While this is acceptable for Asynchronous replication, it is harmful in synchronous replication scenarios as it adds latency to the IO (Even in environments that are below the maximum supported latency).

4 Logical network settings

4.1 IP addresses

Each InfiniBox system uses a set of IP addresses for replication, some for data plane others for control plane. These IP addresses will be specified when you create the Network Space, and are used to allow one system to communicate directly with a remote system. The number of IP addresses allocated to each system is tightly coupled with the types of replication used.

Replication Type	IP Address for	Minimum number of addresses	Recommended number of addresses
Asynchronous replication only	Management	1	1
	Data	3	6
Synchronous and Asynchronous replication	Management	1	1
	Data	6	9

4.2 Routing

Some networks require routing to cross from one site to the other. InfiniBox supports both a “default gateway” configuration (simple, common) and a static route table (advanced, more flexible).

The storage administrator should get the routing definitions from the network team in advance to prevent delays during the replication configuration.

4.3 Firewall rules

InfiniBox Replication uses TCP/IP for both Asynchronous and Synchronous replications, over the following TCP ports:

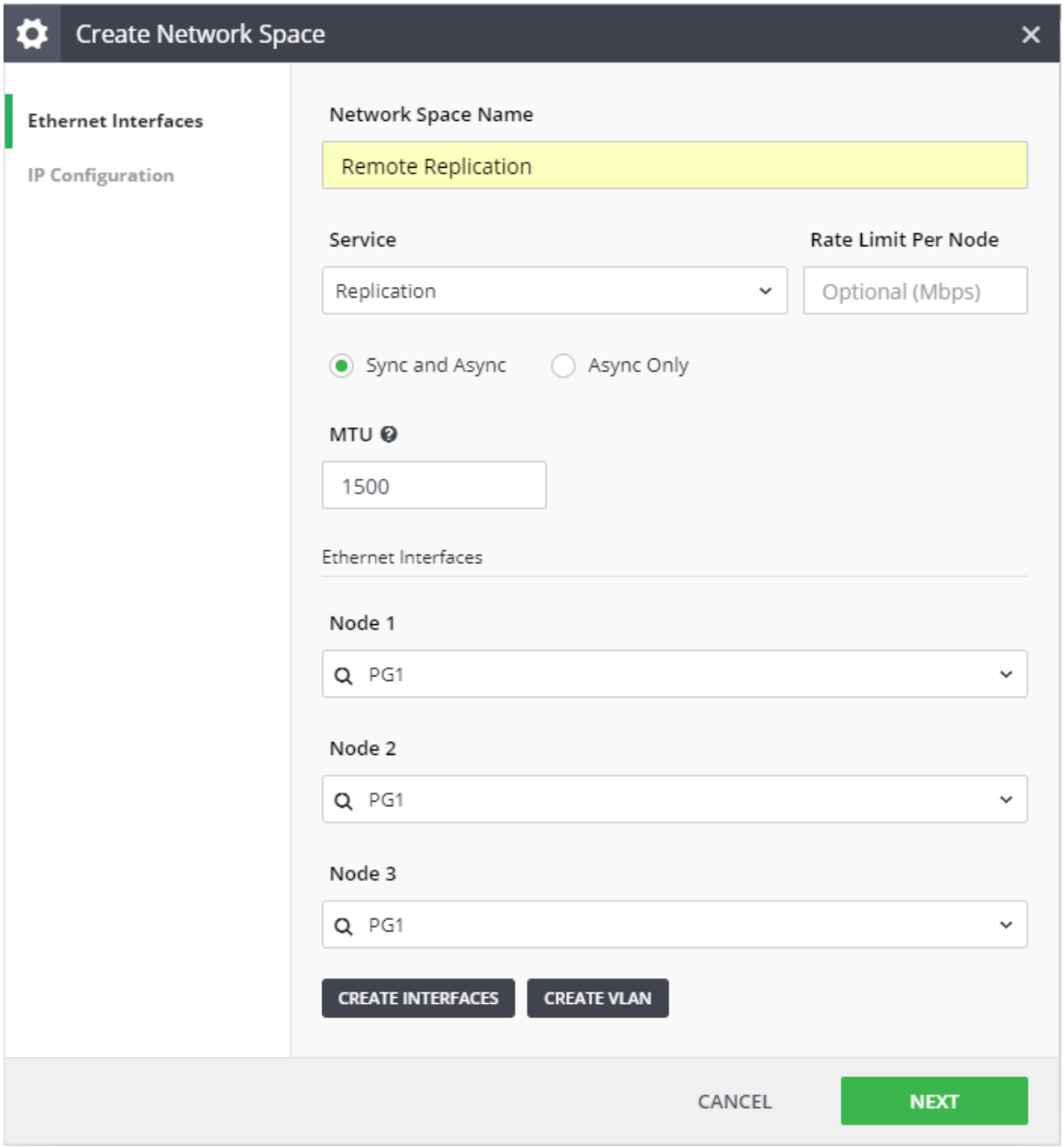
- Management: 80 (HTTP) or 443 (HTTPS)
- Data: 8067

InfiniBox systems need to communicate bidirectionally. This happens in the event of a primary site failure and recovery, when data needs to be sent from the DR site back to the production site. This requires firewall rules to be open in both directions.

The firewall rules must also allow access between any IP address on the network space on one system with all IP addresses on the network space on the remote system.

4.4 Creating a Replication Network Space

1. On the InfiniBox left menu, click **Settings**
   - Click the **Network Spaces** tab
   - Click **Create**
2. The **Create Network Space** screen opens

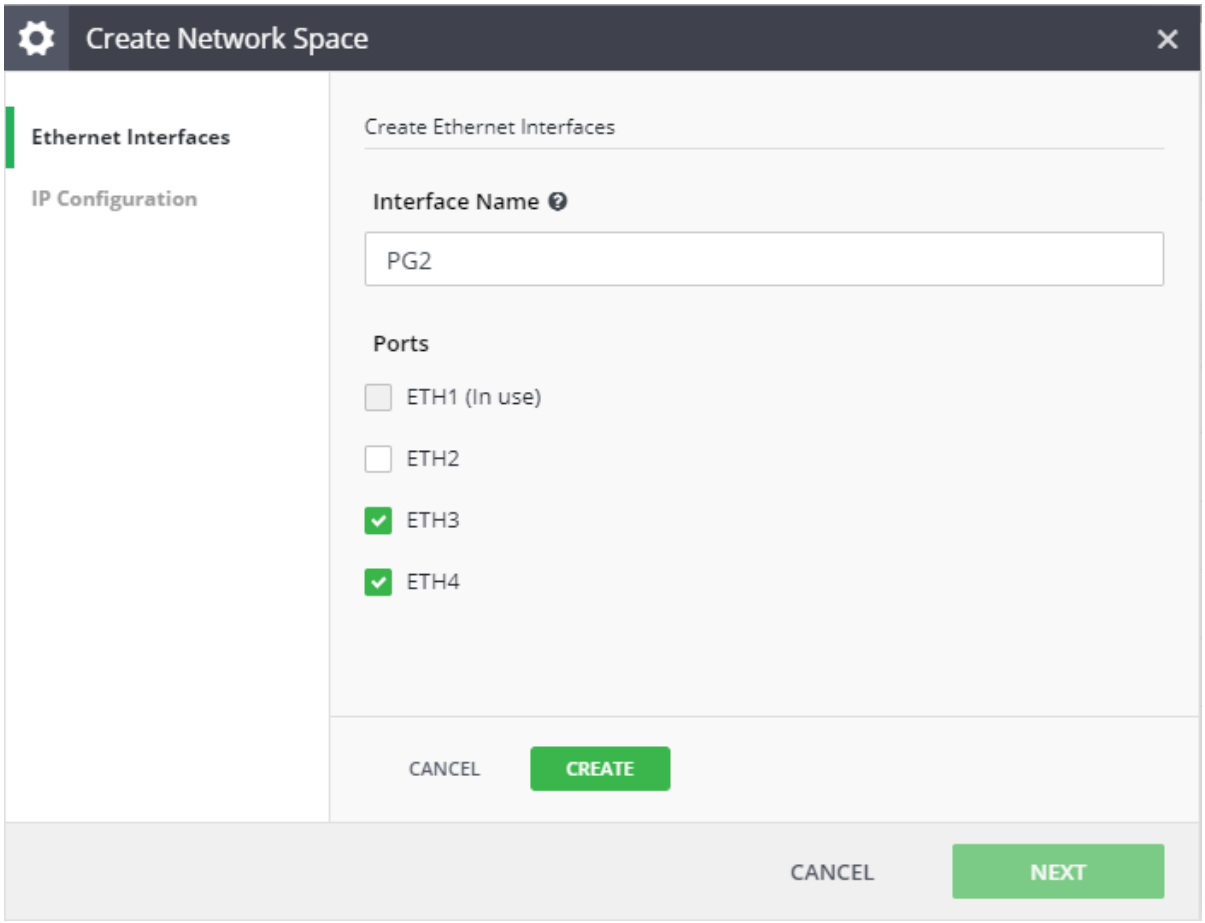


3. Fill in the network space name and MTU.

   Select Replication from the **Service** drop-down list.

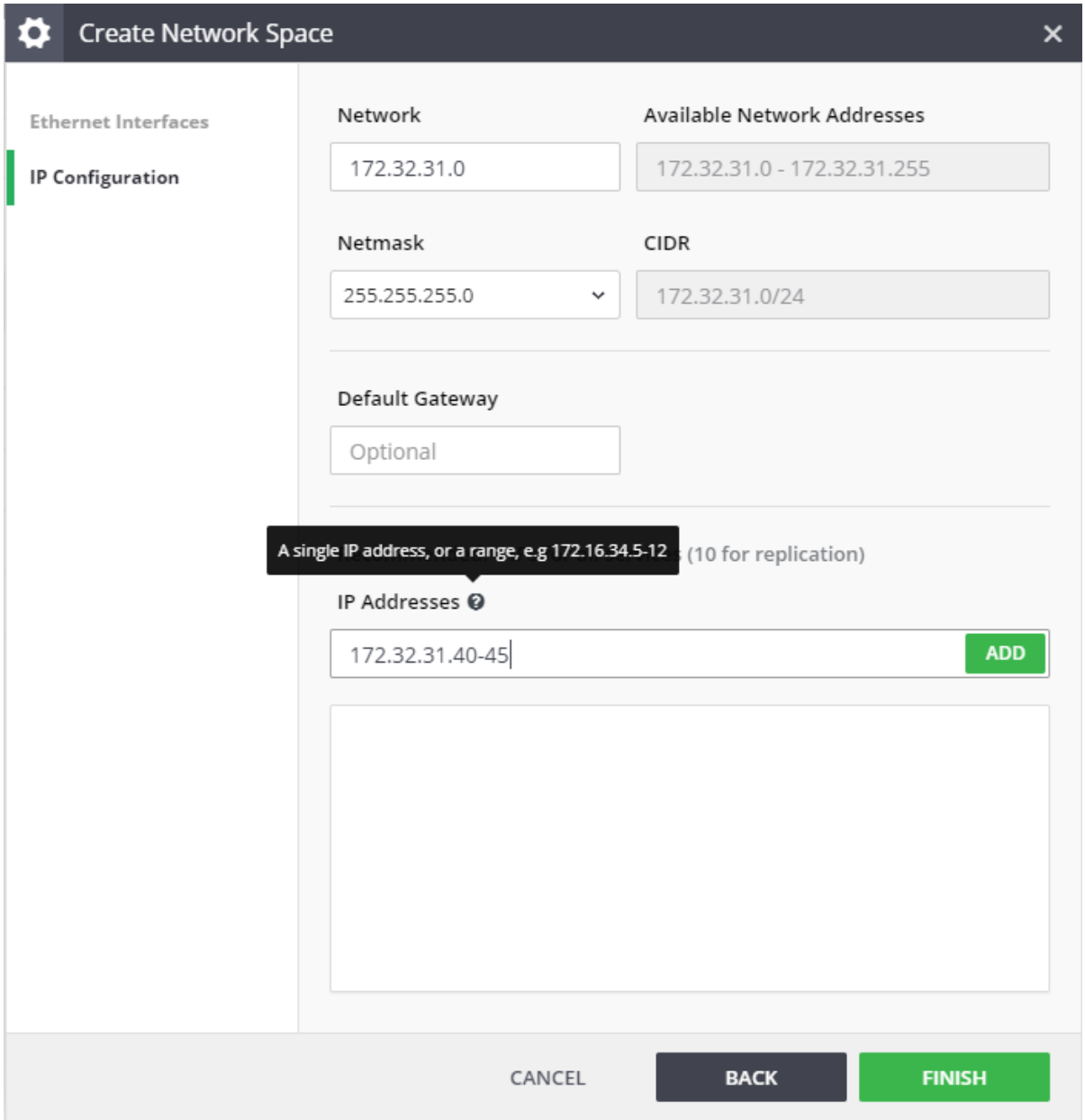
4. Select either **Sync and Async** (if you plan to use synchronous replication), or **Async Only** (otherwise).
5. Either select an Ethernet interface or click **Create Interfaces**.

   Note: The rate limit of the replication network space only affects the Asynchronous replication, the resynchronization of Synchronous replication failover, and the Initialization phase of the replication.



   a. You can rename the default Interface name.
   b. Select two ports for Replication from the available Ethernet ports.
      (ports that are already taken by other interfaces are greyed-out).
   c. Click **Create.**
   d. You are returned to the Create Network Space screen.
      Note: The Ethernet interfaces can be created beforehand, from the Network Interfaces tab.
6. (Optional)
   Click **Create VLAN** in order to group interfaces into a Virtual LAN.
7. Click **Next** to move to the **IP Configuration** screen.

8. On the **IP Configuration** screen, fill in the networking data:
    - Network
    - Netmask
    - IP addresses - click **Add** to verify the validity of the IP addresses
9. Define 10 IP addresses for a service that will run sync and async replications, or 7 IP addresses for a service that will run async-only replications.



10. Click **Finish**.

    The network space is created and is visible on the screen.

5 Network performance requirements for Asynchronous Replication

5.1 Latency and reliability

For Asynchronous replication, latency has little effect as InfiniBox leverages TCP/IP in highly optimized fashion.

However, reliability of the connection is of utmost importance: intermittent failures such as packet drops and TCP re-transmissions will severely degrade the actual throughput and prevent Asynchronous replication from achieving the desired RPO. When you design and verify your network connection between the systems, make sure to test for these conditions over time: even periodic re-transmission rates of 1% could degrade performance and cause RPO lagging.

5.2 Bandwidth

Data written to the storage needs to be sent to the DR site. In Asynchronous replications, this happens periodically (at the interval set by the storage admin) and will usually take a short time to complete. This creates a “bursty” behavior.

The bandwidth required for async replication depends on the I/O pattern and the interval at which async replication is triggered for a particular dataset.

- Not every I/O that hosts send will be replicated eventually: if a host writes data to an LBA and shortly after than overwrites the LBA with new data or unmaps the LBA, the original write operation will not be sent (unless a replication is triggered in the middle).
- InfiniBox async replication also tries to identify specific portions of data that have changed, even if a host overwrites an existing LBA with similar content.

Because of this it is not easy to predict how much bandwidth with be required for async replication. Customers may begin with estimating this based on the throughput of their hosts, and test various interval settings to understand the application behavior.

As applications behavior changes over time and additional replicas may be added, it is important to measure the bandwidth utilization periodically.

5.3 Bandwidth baseline

Before starting to use a replication link it is important to create a baseline of the available bandwidth.

Testing your network for its available bandwidth should be done over a period of time and several times during the day as network traffic may vary in different times of the day. It is also **critically important** that you coordinate such a test with your network team, as the test may compete for bandwidth with other applications.

For more information about measuring the bandwidth between the sites see **Measuring the bandwidth between the sites** below.

6 Network performance requirements for Synchronous Replication

6.1 Latency and distance

For Synchronous replication, latency is important as each I/O has to traverse between the two storage arrays **before** it is acknowledged back to the host.

InfiniBox Synchronous replication requires up to 5ms round-trip latency and up to 100 kilometers maximum distance.

Note: When measuring distance, it is important to look at the actual length of fiber optic between the 2 sites and not the point-to-point distance between them.

Note: synchronous replication is available for volumes and consistency groupss only (it is not available for filesystems).

In addition, the reliability of the connection is of utmost importance: intermittent failures such as packet drops and TCP re-transmissions will severely degrade the actual latency and throughput and will cause high response times for your applications I/O. When you design and verify your network connection between the systems, make sure to test for these conditions over time: even periodic re-transmission rates of 1% could degrade performance.

6.1.1 Measuring the latency between the sites

It is highly recommended to test the latency between the sites ahead of the installation, checking for both the average latency and its “jitter” - the fluctuations in latency. The simplest way of testing this is using a ping command running 1,000 samples in each measurement, and repeating the test several times per day.

This can help uncover many potential issues such as:

- High level of jitter - the response time of each sample varies dramatically
- High latency - many samples take longer than 5ms to complete
- Different latency in different hours of the day - this usually indicates a bottleneck somewhere in the network (for example - the WAN link) at some hours of the day, which will affect the ability of the system to send data to the remote system and will result in high latency for synchronous replications
- Packet loss - packets get lost due to low link quality or congestion

An ideal environment will show consistent sub-5ms response time without losing any of the 1000 samples in the process. Some variance in the response times is expected, as long as the 5ms rule is kept.

6.2 Sizing the replication link bandwidth for Synchronous replication

For Synchronous replication, the required replication link bandwidth derives from the WRITE/XCOPY throughput on the volumes planned to be replicated.

It is recommended to size the replication link bandwidth to at least 130%-150% of the observed aggregate WRITE/XCOPY throughput to the volumes that are planned to be replicated.

The additional bandwidth throughput is to enable smooth handling of the following scenarios affecting needed replication bandwidth:

- Bursts of I/O that may occur normally as part of your workload
- Gradual increase in I/O in the future
- Synchronous replications that failed and are currently in re-synchronizing state, i.e. running the replication asynchronously in order to close the gap
- Minimal overhead that the replication protocol incurs

Note: Not sizing the replication link bandwidth may result in increased latency to the replicated volumes and potentially to inability to sustain Sync replication resulting in fallback to Async.

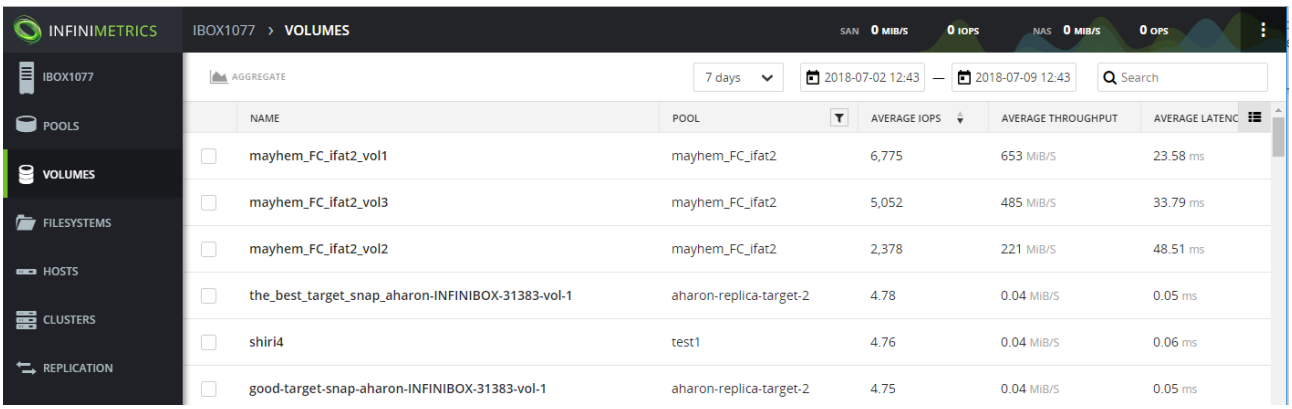
6.2.1 Link sizing example

Assume hosts are writing to a volume at the rate of 100MB/s, and that this volume needs to be replicated synchronously to a remote system.

The minimal bandwidth for the link dedicated to this volume replication should be **more** than 100MB/s, in order to allow for small bursts of I/O and for some future growth. It is important that the latency for sending data at a rate of 100MB/sec (over the WAN) to the remote system is 5ms or less.

Use InfiniMetrics to observe the I/O on the relevant volume.

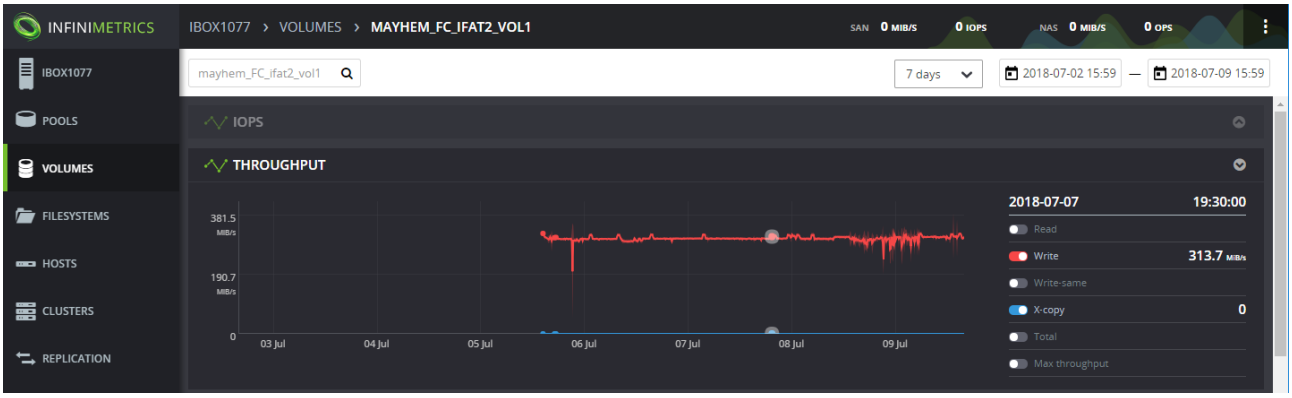
Select the system, and go to the VOLUMES tab which show all the sampled volume activity. You can filter the list using the search box to locate the relevant volume whose activity you want to examine.



Click on the volume name to drill-in and view its activity over time.

Select the time range in the bar: it is recommended to investigate the activity over an entire day or week.

Look at the THROUGHPUT graph, and the legend that appears on the right side. Remove the green line for READ operations by clicking on the toggle next to the Read, and add the blue line for XCOPY by clicking on the toggle next to the X-Copy.



The examples above show consistent write activity of 300MB/s doing WRITE operations and 0MB/s doing XCOPY, which total at 300MB/s,adding 50% overhead we get to 450MB/s.

As a result, for the replication link between the InfiniBox systems should be capable of sustained throughput of 450MB/s.

6.3 Contention between synchronous and asynchronous replications or other network consumers

Synchronous and Asynchronous may coexist and compete over a shared bandwidth. This is clearly the case when both Synchronous and Asynchronous replications are defined in the system. It can also happen no async replications are defined, when some replicas are synchronized but other have fallen out of sync.

In such cases, bursts of asynchronous replication may incur an increased latency for Synchronous replications. To prevent this, it's recommended to set the replication interval of the asynchronous replications very low, and to set a rate limit on the network space used for Asynchronous replication.

Note: The rate limit of the replication network space only affects the Asynchronous replication, the resynchronization of Synchronous replication failover, and the Initialization phase of the replication.

To calculate the rate limit, you will need to know the total verified available bandwidth for replication between the two sites. Make sure to verify this bandwidth beforehand (see Measuring the bandwidth between the sites)

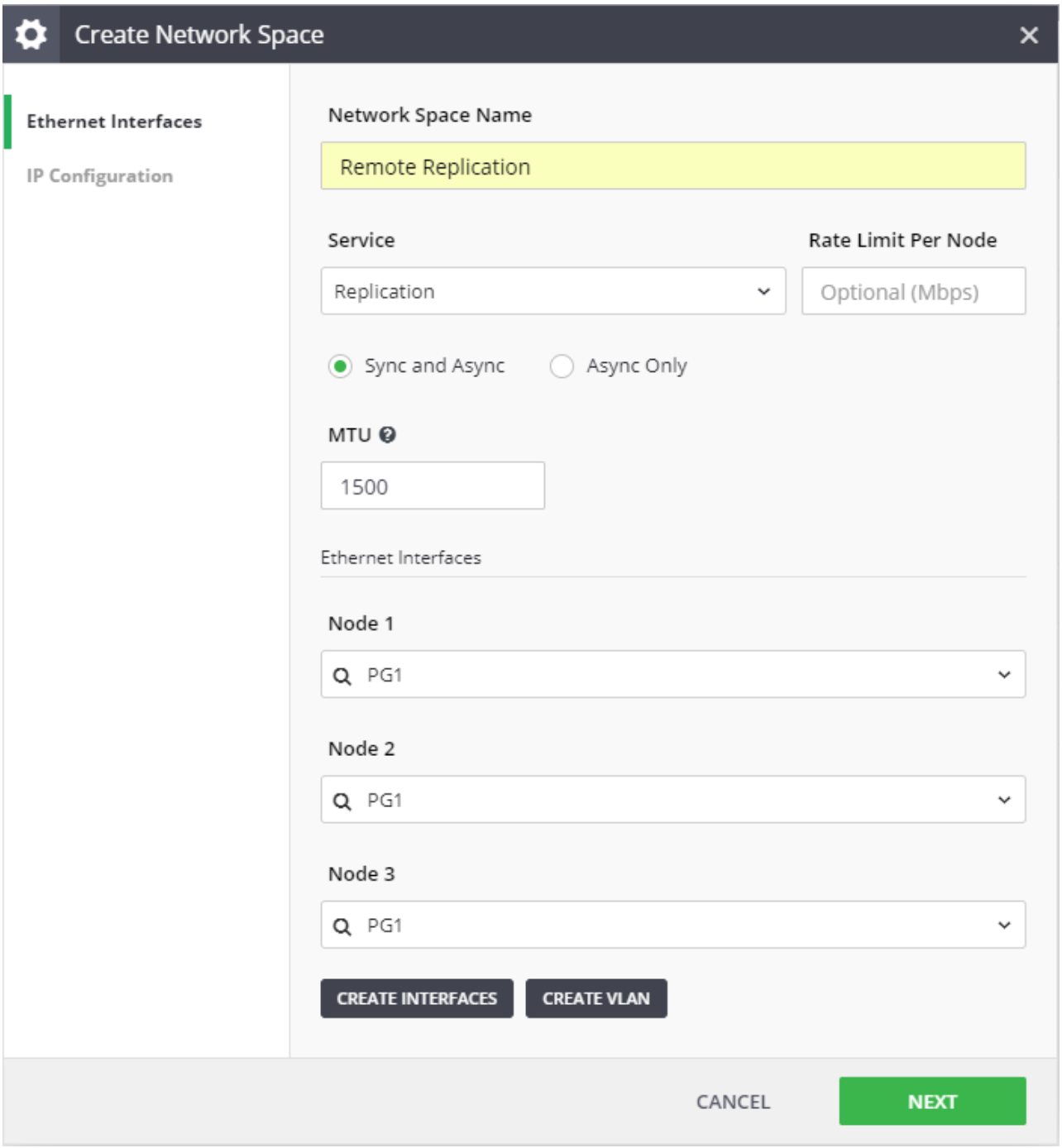
Also, you will need to know the total bandwidth required for sync replication of volumes and CGs, as measured by existing I/O to the volumes (see: Controlling the bandwidth).

The formula for the rate limit is simple (divide by 3 because the rate limit is per node):

```
1.1 * 8 * ( Bandwidth between the systems  - sync replication bandwidth ) / 3
```

Note: that the above formula calculates the rate limit in bits/second. The rate limit on the network space is measured in bps, whereas the bandwidth calculations and measurements above are in Bps). Multiplying by 1.1 adds 10% to allow for some imbalance between nodes traffic.

Set the rate limit on the network space where replication is running using the results from this formula. Modify the network space to place the rate limit on the replication network:



If you share the replication network between InfiniBox and other devices, it is important to make sure the network provides consistent bandwidth and performance to every device. Note that this requirement is no "on average" or "over time", but rather needs to be accurate at ANY time. Typically, only network QoS (quality of service) can provide these capabilities.

7 Controlling the Replication Bandwidth

The mechanisms for controlling the bandwidth utilized for replication relies on limiting the bandwidth rate for network spaces in InfiniBox.

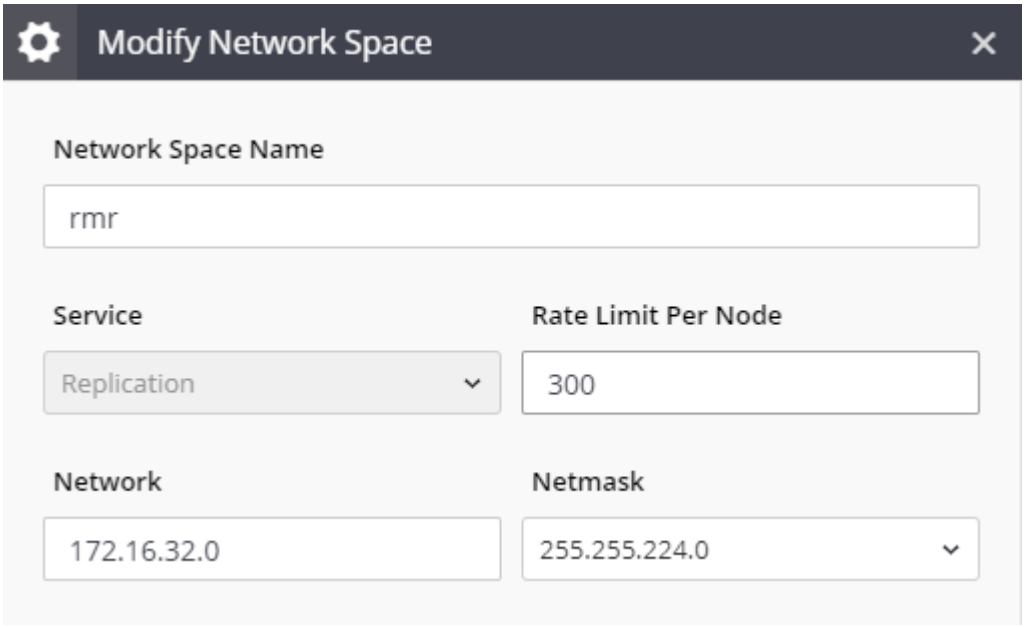
7.1 Rate limit for Network Spaces

Every network space, replication included, allows you to set a limit on the rate each InfiniBox node will utilize for sending data. This rate limit feature is a key element in controlling the bandwidth used for async-replication in many scenarios and avoid contention with other traffic on the same network.

One example is allowing you to avoid increased latencies in sync-replication caused by async-replication saturating the network line. Another example is allowing you to avoid congestion on the network when there are other (non-InfiniBox) traffic taking place.

The rate limit can be set when you create a network space, or modified later on.

- The rate limit is specified in mega-bits/seconds (not mega-bytes/seconds).
- The rate limit is specified per node. Set the limit to slightly more than 1/3rd of the actual bandwidth intended for the InfiniBox system.



The rate limit of the replication network space only affects the Asynchronous replication, the resynchronization of Synchronous replication failover, and the Initialization phase of the replication.

8 Measuring the bandwidth between the sites

The best tool for testing the available bandwidth isiPerf, a free traffic generating tool supporting multiple operating systems.

To use iPerf you will need two Linux hosts, one in the production site and another in the DR site. It is highly recommended to use a physical host or at least a dedicated network interface for this test to be sure `iPerf` does not compete for the bandwidth of that interface with anyone else.

Choose a host in the DR site to be the server, which has connectivity to the Network Space of the replication link.

Run the following command on that host:

```
iperf3 -s -p 8067
```

Choose a host in the production site to be the client, which has connectivity to the Network Space of the replication link.

Run the following command where BANDWIDTH is replaced with the theoretical total allocated bandwidth for all replications between the two systems:

```
iperf3 -c HOSTNAME -p 8067 -b BANDWIDTH
```

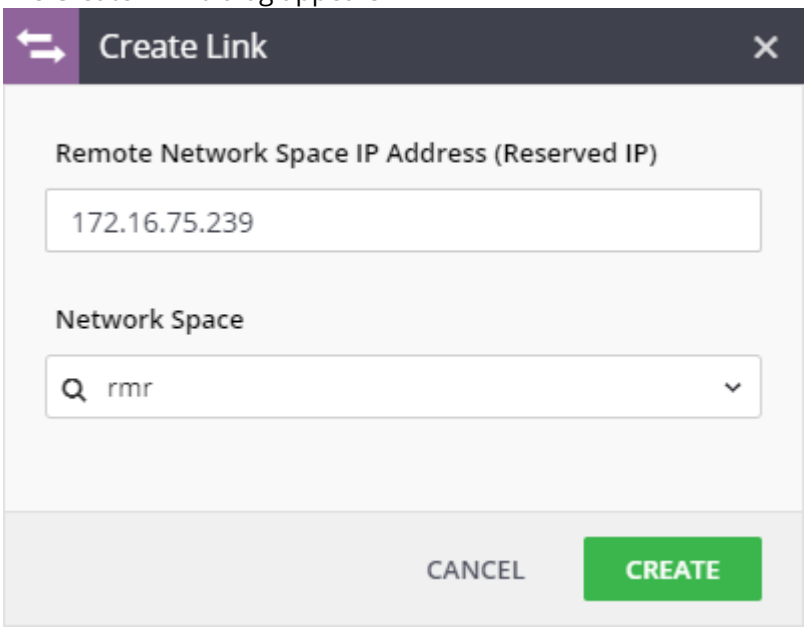
Use the `-b` parameter to make sure the iPerf test does not overuse the link capacity, which might affect other applications that use the same link adversely.

Make sure the connectivity between the two hosts is using the same switches and routes as the replication link.

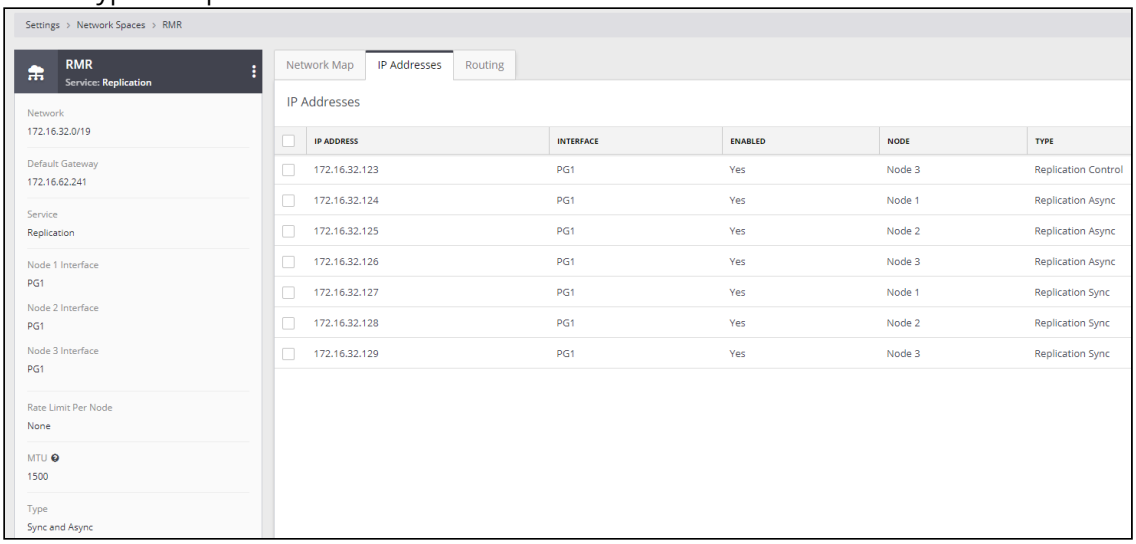
9 Creating the Replication Service

9.1 Creating a link

1. Click the Replication icon at the toolbar on the left. The **Replication** screen opens. Click the **Links** tab.
2. Click **Create Link**.
   a. The Create Link dialog appears



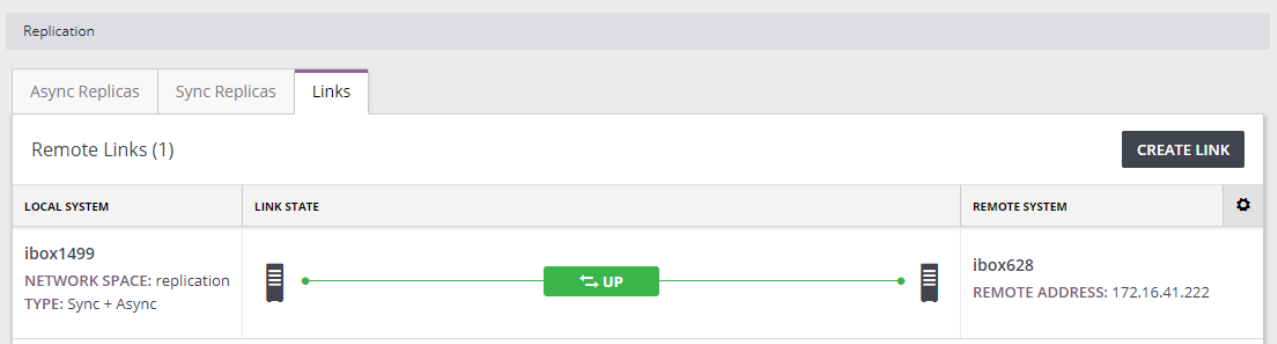
   b. Insert the IP address of the relevant replication network space of the **remote** system, whose type is Replication Control.



   c. Select the local network space that runs the Replication service.

d. In case that the remote system credentials are not identical to the credential of the local system, you will be asked to enter them.
e. Click **Create**.

The two systems are now connected. The Links tab displays the link status.



9.2 Creating a replica

Based on fulfilling of the steps above, you can now create a replica on the source system.

1. Click the **Replica** tab. Click **Create Replica**. Fill in the following details:
   - Replica type – either a volume or a consistency group
   - Source – the name of the volume or consistency group
   - Remote system – the system that stores the target of the replica
   - Remote target – select either of the following:
   - Create new – let InfiniBox to automatically create the replica entity on the target
   - Select existing – select an entity on the target InfiniBox
2. Remote name – filled automatically if Create new was selected
3. Remote pool – select the pool that will store the remote entity
4. Interval - the interval between two consecutive sync jobs
5. RPO (Recovery Point Objective) - the lag between the source and target sites
6. Click **Create**.
   The replica is created and is visible on the **Replication** screen.
   The replica automatically initializes and its replication progress is visible on screen.