Workgroup: Network Working Group
Internet-Draft: draft-wang-nof-framework-01
Published: 11 July 2022
Intended Status: Informational
Expires: 12 January 2023

Authors: H. Wang F. Qin L. Zhao S. Chen
Huawei China Mobile Huawei Huawei

Framework of Fast Fault Detection for IP-baesd SANs

Abstract

NVMe over Fabrics defines a common architecture that supports a range of storage networking fabrics for NVMe block storage protocol over a storage networking fabric, such as Ethernet, Fibre Channel and InfiniBand. For IP-based network, RDMA or TCP technology can be used to transport NVMe commands. When a network fault occurs, NVMe connections need to be switched over. Currently, no effective method is available for quick detection, switchover is performed only based on KA timeout, resulting in low performance.

This document defines the basic framework of how network-assisted hosts and storage devices can quickly detect NVMe connection failures caused by network faults for NVMe IP-based SANs.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

For a long time, the key storage applications and high performance requirements were mainly based on FC networks. With the increase of transmission rates, the medium has evolved from HDDs to solid-state storage, and the protocol has evolved from SCSI to NVMe. The emergence of new NVMe technologies brings new opportunities.

IP-based SANs is an implementation of NVMe over Fabrics that best fits NVMe semantics. It is the development trend of high-speed storage networks in the future. Ethernet-based NVMe has been defined in NVM Express. The specification defined in this document optimizes network control in terms of ease of use, maintainability, and reliraft ability, making Ethernet-based NVMe more suitable for high reliability requirements of key applications. This feature improves system usability and maintainability.

The [I-D.guo-nof-requirement] describes the problems of the current NVMe solution. On an IP-based SAN, if the access link of a storage device is faulty, hosts cannot access the storage device. Because the host cannot directly detect the fault, the host has to wait for the KA timeout. To speed up the detection, hosts and storage devices can utilize fast KA or BFD to perform fast detection. However, this solution introdueced additional load on hosts and storage devices and is hard to use in large-scale IP-based SAN. In fact, the IP network can directly detect the fault. Then the IP network can notify the necessary hosts or storage devices of the fault.

2. Terminology

NoF : NVMe of Fabrics

FC : Fiber Channel

NVMe : Non-Volatile Memory Express

SAN: Storage Area Network

3. Reference Models

An IP-based SAN mainly includes three types of roles: an initiator (referred to as a host), a switch, and a target (referred to as a storage device). Initiators and targets are also referred to as endpoint devices. Hosts and storage devices use the Ethernet-based NVMe protocol to transmit data over the network to provide high-performance storage services.

3.1. Small-scale SAN

```
               +--+       +--+
    Host       |H1|       |H2|
 (Initiator)   +-,+       +_.+
                | `',   _-` |
                |    _-`    |
                | _-`   `', |
      IP     +----+       +----+
   Network   | SW1|       | SW2|
             +---,+       +_.--+
                | `',   _-` |
                |    `',    |
                | _-`   `', |
   Storage     +-`+       +`'+
   (Target)    |S1|       |S2|
               +--+       +--+
```

Figure 1 : Small-scale SAN

This is the basic model for small-scale storage access networks. Hosts and storage devices are dual-homed to different switches.

When the access link of the storage device is faulty, the host needs to quickly detect the fault so that the NVMe connection can be quickly switched to the standby path.

3.2. Large-scale SAN

```
                          +--+      +--+      +--+      +--+
              Host        |H1|      |H2|      |H3|      |H4|
           (Initiator)    +/-+      +-,+      +.-+      +/-+
                           |         | '.   ,-`|         |
                           |         |   `',   |         |
                           |         | ,-`  '. |         |
                         +-\--+    +--`-+    +`'--+    +-\--+
                         | SW1|    | SW2|    | SW3|    | SW4|
                         +--,-+    +---,,    +,.--+    +-.--+
                             `.          `'.,`         .`
                               `.   _,-'`    ``'.,   .`
               IP              +--'`+            +`-`-+
             Network           | SW5|            | SW6|
                               +--,,+            +,.,-+
                               .`   `'.,     ,.-``   ',
                             .`         _,-'`          `.
                         +--`-+    +--'`+    `'---+    +-`'-+
                         | SW7|    | SW8|    | SW9|    |SW10|
                         +-.,-+    +-..-+    +-.,-+    +-_.-+
                           | '.   ,-` |        | `.,   .' |
                           |   `',    |        |    '.`   |
                           | ,-`  '.  |        | ,-`  `', |
             Storage      +-`+      `'\+      +-`+      +`'+
             (Target)     |S1|      |S2|      |S3|      |S4|
                          +--+      +--+      +--+      +--+
```

Figure 2 : Large-scale SAN

This is a relatively large-scale storage network which applies to a large-scale storage device access network.

When the access link of the storage device is faulty, the host needs to quickly detect the fault so that the NVMe connection can be quickly switched to the standby path.

4. Functional Components

The NVMe IP-based SANs consists of storage devices, hosts and switches. Hosts and storage devices need to obtain required fault information from the IP network. Switches need to synchronize locally detected fault information on the IP network so that other switches can obtain the faults and notify hosts or storage devices that require the fault infomation.

4.1. Storage Device

As the server side, storage devices provide storage access services for hosts. If a storage device is connected to an IP network and is interested in the status of other devices, the storage device can initiate a subscription request to the connected switch to obtain status notifications of other devices from the access switch.

To reduce the complexity of storage devices, it's suggest to extend the LLDP protocol to support subscription from storage devices to switches and use the new L2-based protocol to notify the switch of status to the storage device.

```
  +-------+                 +------+
  |Storage|                 |Switch|
  +-------+                 +------+
      |      Subscribe Msg      |
      | ----------------------->|
      |                         |
      |     Notification Msg    |
      | <-----------------------|
      |                         |
      |                         |
```

Figure 3 : Storage Device

4.2. Host

The host is the client of the storage device. As the client side, a host needs to quickly obtain the service status of the storage device that provides services. When the host receives a notification message from the switch indicating that the storage device is faulty, the host will quickly disconnect from the storage device and switch to a redundant one.

The recommended protocol on the host side is the same as that on the storage device.

```
+-------+                 +------+
|  HOST |                 |Switch|
+-------+                 +------+
    |      Subscribe Msg      |
    | ----------------------->|
    |                         |
    |     Notification Msg    |
    | <-----------------------|
    |                         |
    |                         |
```

Figure 4 : Host Device

4.3. Network Device

Switches can quickly detect local faults and synchronize the faults to other switches on the IP network. After detecting a fault, the switch needs to notify the required host or storage device of the fault.

```
+------+                  +------+
|Switch|                  |Switch|
+------+                  +------+
   |    Information Sync     |
   | ----------------------->|
   |                         |
   |                         |
   |                         |
```

Figure 5 : Network Device

5. Procedures

5.1. Network Deployment

The IP-based SAN uses the standard Ethernet technolog. Network deployments typically use the current IP technologies. For example, OSPF is usually deployed as an underlay protocol.

5.2. Storage and Host Access

Hosts and storage devices are connected to the ethernet network. The administrator assigns access IP addresses to the hosts and storage devices. In most scenarios, these routes can be advertised through the underlay protocol. In addition, after hosts and storage devices go online, they needs to send subscription requests to the switch to obtain the status information of the target device.

To prevent hosts or storage devices from being aware of extra IP address, it is recommended that LLDP be used to implement this message.

5.3. Status Infomation Sync And Notification

When hosts and storage devices go online, the switch can calculates an initial state of these devices and synchronizes the state on the IP network.

After detecting a local fault, the switch needs to notify other access devices that need the fault information. In addition, the switch needs to synchronize the fault information to other switches on the network. To ensure that synchronization messages can be reliably synchronized to other switches, a reliable transmission protocol, such as TCP or Quic, must be used. For large-scale IP networks, hierarchical synchronization can be used to reduce the number of sessions between switches.

The synchronization information about the host and storage devices belongs to the application layer's information.

```
+-------+           +----+      +------+      +----+         +-------+
|  HOST |-----------|TOR1|------|Spine1|------|TOR3|---------|Storage|
+---/---+           +-/--+      +--/---+      +-/--+         +---/---+
    |---------------->|  Info Sync |  Info Sync |<---------------|
    |  SubscribeMsg   |----------->|<-----------|  Subscribe Msg |
    |                 |<-----------|----------->|                |
    |<----------------|  Info Sync |  Info Sync |                |
    |Notification Msg |            |            |                |
    |                 |            |            |                |
```

Figure 7 : Information Advertisement

5.3.1. Access Link Failure

When an access link is faulty, the access switch detects the fault. Based on the faulty link, the access switch can calculate the devices whose IP addresses are affected. The access switch advertises the faulty IP address information on other access links. The switch synchronizes the faulty IP address information on the IP network based on the computation result. After receiving the synchronized fault information, other switches notify the access host or storage device of the fault information.

5.3.2. Network Link or Device Failure

ECMP or redundant link protection is usually deployed to prevent this failure.

6. Security Considerations

NA

7. IANA Considerations

This document makes no request of IANA.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

8.2. Informative References

[I-D.guo-nof-requirement]

Guo, L., Feng, Y., Zhao, J., Qin, F., Zhao, L., and H. Wang, "Requirement of Fast Fault Detection for IP-based SANs", Work in Progress, Internet-Draft, draft-guo-nof-requirement-01, 11 July 2022, <https://www.ietf.org/archive/id/draft-guo-nof-requirement-01.txt>.

Authors' Addresses

Haibo Wang
Huawei
No. 156 Beiqing Road
Beijing
100095
P.R. China

Email: rainsword.wang@huawei.com

Fengwei Qin
China Mobile
Beijing
China

Email: qinfengwei@chinamobile.com

Lily Zhao
Huawei
No. 3 Shangdi Information Road
Beijing
100085
P.R. China

Email: Lily.zhao@huawei.com

Shuanglong Chen
Huawei
No. 156 Beiqing Road
Beijing
100095
P.R. China

Email: chenshuanglong@huawei.com