Workgroup: Netowork Working Group
Internet-Draft: draft-guo-nof-requirement-01
Published: 11 July 2022
Intended Status: Informational
Expires: 12 January 2023

Authors: L. Guo Y. Feng J. Zhao F. Qin
CAICT China Mobile China Telecom China Mobile
L. Zhao H. Wang
Huawei Huawei

Requirement of Fast Fault Detection for IP-based 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, but the network fault detection is weak.

This document describes the solution requirements for fast fault detection to improve reliability.

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 are 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 SATA to NVMe. The emergence of new NVMe technologies brings new opportunities. With the development of the NVMe protocol, the application scenario of the NVMe protocol is extended from PCIe to other fabrics, solving the problem of NVMe extension and transmission distance. The block storage protocol uses NoF to replace SCSI, reducing the number of protocol interactions from application hosts to storage systems. The end-to-end NVMe protocol greatly improves performance.

Fabrics of NoF includes Ethernet, Fibre Channel and InfiniBand. Comparing FC-NVMe to Ethernet- or InfiniBand-based Network alternatives generally takes into consideration the advantages and disadvantages of the networking technologies. Fibre Channel fabrics are noted for their lossless data transmission, predictable and consistent performance, and reliability. Large enterprises tend to favor FC storage for mission-critical workloads. But Fibre Channel requires special equipment and storage networking expertise to operate and can be more costly than IP-based alternatives. Like FC, InfiniBand is a lossless network requiring special hardware. IP-based NVMe storage products tend to be more plentiful than FC-NVMe-based options. Most storage startups focus on IP-based NVMe. But unlink FC, The Ethernet switch does not notify the Change of device status. When the device is faulty, relying on the NVMe link heartbeat message mechanism , the host takes tens of seconds to complete service failover.

```
+--------------------------------------+
|          NVMe Host Software          |
+--------------------------------------+
+--------------------------------------+
|   Host Side Transport Abstraction    |
+--------------------------------------+

   /\      /\      /\      /\      /\
  /  \    /  \    /  \    /  \    /  \
   FC      IB     RoCE    iWARP   TCP
  \  /    \  /    \  /    \  /    \  /
   \/      \/      \/      \/      \/

+--------------------------------------+
|Controller Side Transport Abstraction |
+--------------------------------------+
+--------------------------------------+
|          NVMe SubSystem              |
+--------------------------------------+
```

This document describes the application scenarios and capability requirements of the IP-based NVMe that implements fast fault detection similar to FC. The proposal is already under discussion in working group of NVMe organization.

2. Terminology

IP-based NVMe: using RDMA or TCP to transport NVMe through Ethernet

FC: Fiber Channel

NVMe: Non-Volatile Memory Express

NoF: NVMe of Fabrics

3. Use Case

The NVMe over RDMA or TCP IP-based network in storage is as follows, the network 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.

```
                        +--+      +--+      +--+      +--+
            Host        |H1|      |H2|      |H3|      |H4|
         (Initiator)    +/-+      +-,+      +.-+      +/-+
                         |         | '.   ,-`|         |
                         |         |   `',   |         |
                         |         | ,-`  '. |         |
                       +-\--+    +--`-+    +`'--+    +-\--+
                       | SW |    | SW |    | SW |    | SW |
                       +--,-+    +---,,    +,.--+    +-.--+
                           `.          `'.,`         .`
                             `.   _,-'`    ``'.,   .`
                IP           +--'`+            +`-`-+
           Network           | SW |            | SW |
                             +--,,+            +,.,-+
                             .`   `'.,     ,.-``   ',
                           .`         _,-'`          `.
                       +--`-+    +--'`+    `'---+    +-`'-+
                       | SW |    | SW |    | SW |    | SW |
                       +-.,-+    +-..-+    +-.,-+    +-_.-+
                         | '.   ,-` |        | `.,   .' |
                         |   `',    |        |    '.`   |
                         | ,-`  '.  |        | ,-`  `', |
           Storage      +-`+      `'\+      +-`+      +`'+
           (Target)     |S1|      |S2|      |S3|      |S4|
                        +--+      +--+      +--+      +--+
```

Hosts and storage devices are connected to the network separately and In order to achieve high reliability, each host and storage device are connected to dual network planes simultaneously. The host can read and write data services when an NVMe connection is established between the host and the storage device.

When a storage device link is faulty during running, the host cannot detect the fault status of the indirectly connected device at the transport layer. Based on the IP-based NVMe protocol, the host uses the NVMe heartbeat to detect the status of the storage device. The heartbeat message interval is 5s. Therefore, it takes tens of seconds to determine whether the storage device is faulty and perform service switchover using the multipath software. Failure tolerance time for core applications cannot be reached. In order to obtain the best customer experience and business reliability requirement, we need to enhance fault detection and failover for IP-based NVMe.

In this proposl, a fast fault detection solution with switch participation is proposed. This scheme utilizes the ability of switches to detect faults quickly at the physical layer and link layer, and allows the switch to synchronize the detected fault information in the IP network, and then notify the fault status to the endpoint devices.

Fault detection procedure: The host can detect the fault status of the storage device and quickly switch to the standby path.

1. If a storage fault occurs, the access switch detects the fault at the storage network layer or link layer.

2. The switch synchronizes the status to other switches on the network.

3. The switch notifies the storage fault information to the hosts.

4. Quickly disconnect the connection from the storage device and trigger the multipathing software to switch services to the redundant path. The fault is detected within 1s.

```
+----+       +-------+     +-------+    +-------+
|Host|       |Switch |     |Switch |    |Storage|
+----+       +-------+     +-------+    +-------+
   |             |            |-+          |
   |             |            |1|          |
   |             |            |-+          |
   |             |<----2------|            |
   |             |            |            |
   |<----3-------|            |            |
   |             |            |            |
   |<----4-------|------------|-----------> |
   |             |            |            |
```

4. References

4.1. 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>.

4.2. Informative References

[ODCC-2020-05016] Open Data Center Committe, "NVMe over RoCEv2 Network Control Optimization Technical Requirements and Test Specifications", 2020.

Authors' Addresses

Liang Guo

CAICT
No.52, Hua Yuan Bei Road, Haidian District,
Beijing
Beijing, 100191
China

Email: guoliang1@caict.ac.cn

Yi Feng
China Mobile
12 Chegongzhuang Street, Xicheng District
Beijing
Beijing,
China

Email: fengyiit@chinamobile.com

Jizhuang Zhao
China Telecom
South District of Future Science and Technology in Beiqijia Town,
Changping District
Beijing
Beijing,
China

Email: zhaojzh@chinatelecom.cn

Fengwei Qin
China Mobile
12 Chegongzhuang Street, Xicheng District
Beijing
China

Email: qinfengwei@chinamobile.com

Lily Zhao
Huawei
No. 3 Shangdi Information Road, Haidian District
Beijing
Beijing,
China

Email: Lily.zhao@huawei.com

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

Email: rainsword.wang@huawei.com