Workgroup: Network Working Group
Internet-Draft: draft-wang-ffd-framework-02
Published: 1 March 2024
Intended Status: Informational
Expires: 2 September 2024
Authors: H. Wang (Huawei) F. Qin (China Mobile) L. Zhao (Huawei) S. Chen (Huawei) H. Huang (Huawei)

Framework of Fast Fault Detection for IP-based Network

Abstract

The IP-based distributed system and software application layer often use heartbeat to maintain the network topology status. However, the heartbeat setting is long, which prolongs the system fault detection time. IP-based storage network is the typical usage of that scenario. When the IP-based storage 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 keepalive timeout, resulting in low performance.

This document defines the basic framework of how network assisted host devices can quickly detect application connection failures caused by network faults.

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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 2 September 2024.

Table of Contents

1. Introduction

Today, distributed systems based on network communication are widely used. In order to ensure that both ends of the distributed system can perceive faults, heartbeat is a common technology. However, relying on the heartbeat to detect whether the peer is faulty also faces challenges: if the heartbeat is set too short, it may be misjudged by network disturbances; if the heartbeat is set too long, when a fault occurs, it will not be found for a long time.

Application scenarios such as IP-based NVMe, distributed storage, and cluster computing are typical scenarios for such technologies.

The [I-D.guo-ffd-requirement] describes the problems of the current IP-based NVMe solution. On an IP-based storage area network, 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 fault detection, hosts and storage devices can implement fast KA or BFD. However, this solution introdueced additional cost on hosts and storage devices and is hard to use in large-scale IP-based storage area network. In fact, the IP network can directly detect these faults, so we can use the IP network to assist these access endpoints to quickly perceive the fault, so as to perform quickly service recovery.

2. Terminology

NoF : NVMe of Fabrics

FC : Fiber Channel

NVMe : Non-Volatile Memory Express

SAN: Storage Area Network

3. Reference Models

The frame solution here is applicable to the system where the terminals are directly connected to the IP network.

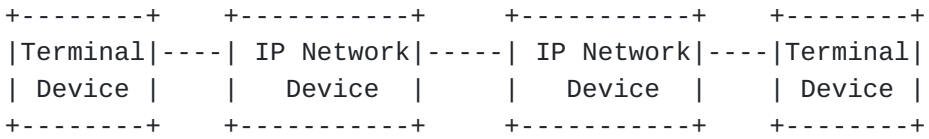

Figure 1 : Basic framework

Terminals are connected to the IP network, and they establish IP connections through the reachability provided by the IP network. When the connection path fails, they cannot be detected quickly. They can only detect it after the keep-alive timeout, and then can carry out service protection processing. This time may be relatively long. Therefore, it is necessary to notify the terminal device of some failures in the network, such as access port failures and internal network failures that will cause IP connection failures between terminals, so that the terminal device can respond quickly and perform corresponding service processing.

As introduced in Introduction, there are scenarios such as IP-based NVMe, distributed storage, and cluster computing. Here we take IP-

based NVME as a typical scenario for introduction, and the processing behavior of other scenarios is similar.

An IP-based storage area network mainly includes three types of roles:

o Initiator, the terminal device, is also called the host.

o Switch, which is a network device used to access terminal devices.

o Target is also a terminal device, also known as a storage device.

The host and storage devices use the Ethernet-based NVMe protocol to transmit data through the IP network to provide high-performance storage services.

3.1. Small-scale SAN

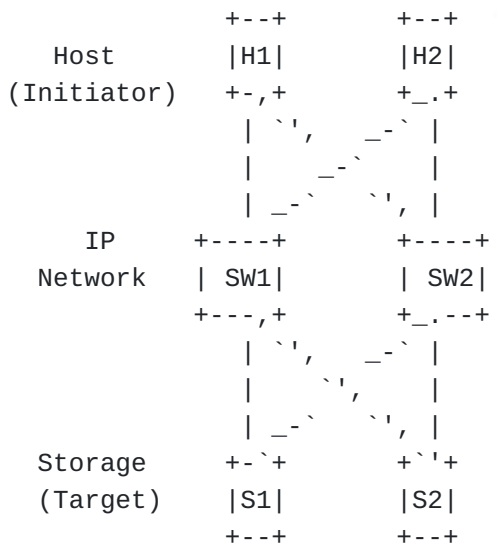

Figure 1 : Small-scale SAN

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

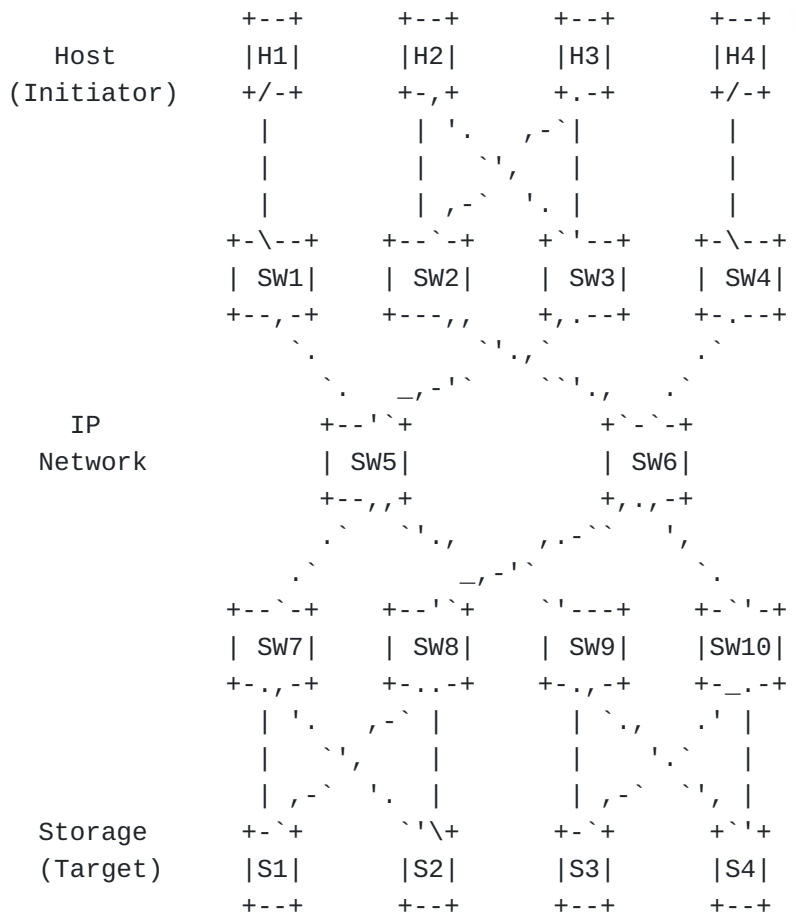

Figure 2 : Large-scale SAN

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.

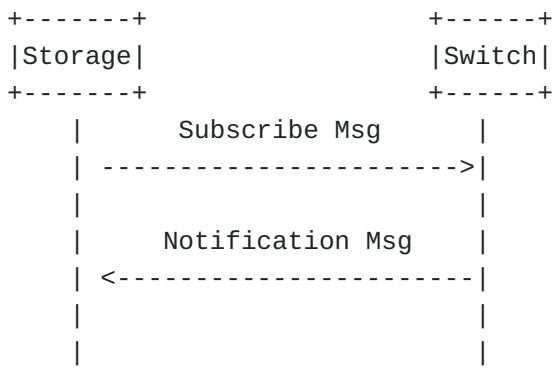

Figure 3 : Storage Device

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.

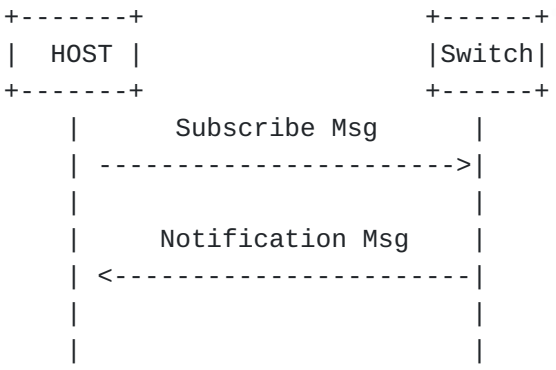

Figure 4 : Host Device

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.

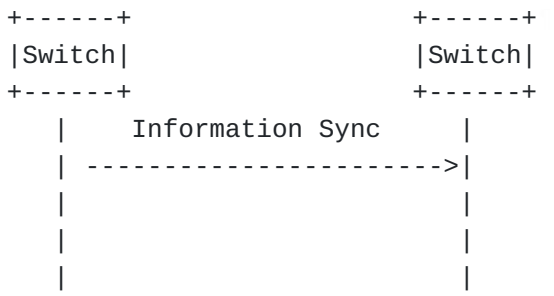

Figure 5 : Network Device

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.

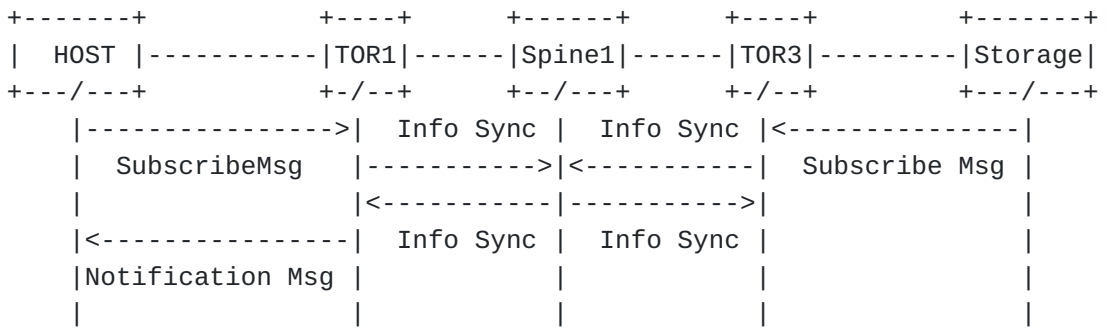

Figure 7 : Information Advertisement

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

In order to control the communication range of information and reduce the negative impact of possible information flooding, the Subscribe Msg and Notification Msg considered in this framework are suggested to be implemented through the L2 extension protocol, so that the sending and receiving of this information will only be controlled by the access network device within the domain. At the same time, the network device is not allowed to forward this message, only allowed to receive or send such message as needed.

For the communication protocol between network devices, in order to ensure its security, it can be encrypted by commonly used encryption technology, including but not limited to TCP-AO, TLS and other technologies.

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-ffd-requirement] Guo, L., Feng, Y., Zhao, J., Qin, F., Zhao, L., Wang, H., and W. Quan, "Requirement of Fast Fault Detection for IP-based Network", Work in Progress, Internet-Draft, draft-guo-ffd-requirement-01, 9 March 2023, <https://datatracker.ietf.org/doc/html/draft-guo-ffd-requirement-01>.

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

Hongyi Huang
Huawei
No. 156 Beiqing Road
Beijing
100095
P.R. China

Email: hongyi.huang@huawei.com