Workgroup: Netowork Working Group
Internet-Draft: draft-nof-requirement-00
Published: 7 March 2022
Intended Status: Experimental
Expires: 8 September 2022
Authors: L. Guo Y. Feng J. Zhao L. Zhao
CAICT China Mobile China Telecom Huawei
H. Wang
Huawei

NVMe over Fabric Network Requirement

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 Ethernet-based network, RDMA or TCP technology can be used to transport NVMe, but the network management mechanism is simple, and fault detection is weak.

This document describes the solution requirements for automatic device discovery to improve usability and quick switchover 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

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 8 September 2022.

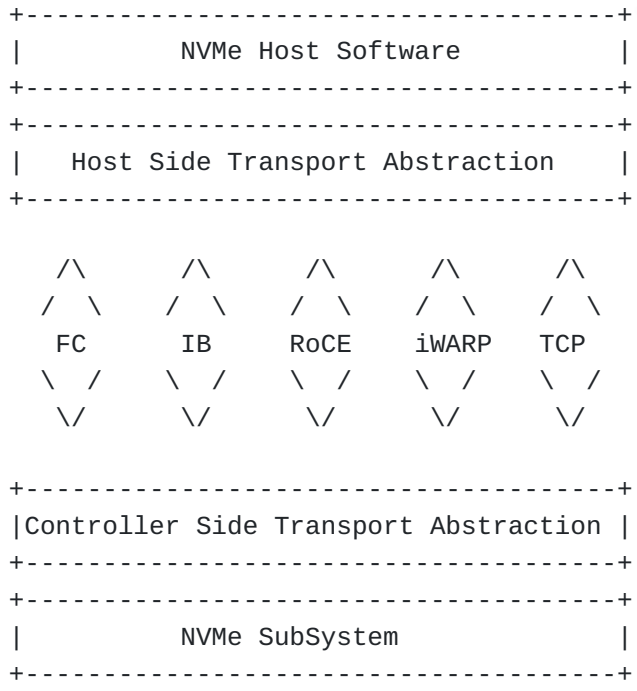
Table of Contents

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 Ethernet-based alternatives. Like FC, InfiniBand is a lossless network requiring special

hardware. Ethernet-based NVMe storage products tend to be more plentiful than FC-NVMe-based options. Most storage startups focus on Ethernet-based NVMe. But unlike FC, InfiniBand and Ethernet lack a discovery service that enables the automatic addition of nodes to the fabric. And unlink FC, The Ethernet switch does not have zone management and 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 switchover.



This document describes the application scenarios and capability requirements of the Ethernet-based NVMe that implements automatic device discovery, domain management, and fault notification similar to FC.

2. Terminology

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

FC: Fiber Channel

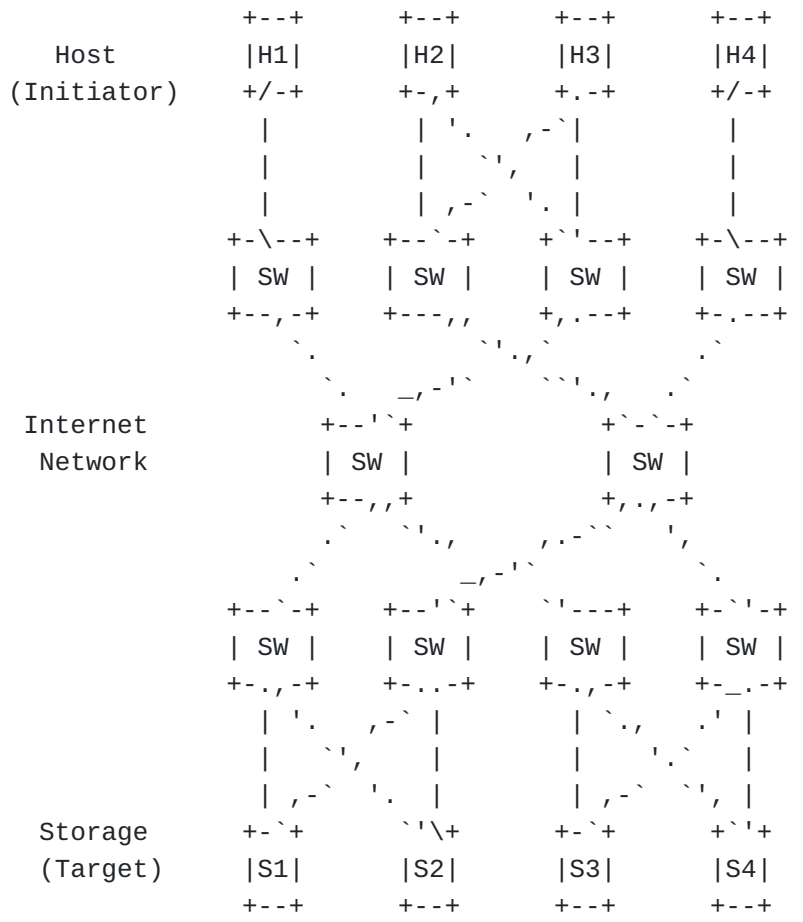
NVMe: Non-Volatile Memory Express

NoF: NVMe of Fabrics

CDC: Centralized Discovery Controller

3. Use Case

The NVMe over RDMA or TCP Ethernet-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. Hosts and storage devices use the Ethernet-based NVMe to transmit data over the network to provide high performance storage services.



Sub-Scenario 1: Initial Deployment

During initial system deployment, 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 only when an NVMe connection is established between the host and the storage device. To establish an NVMe connection, the host need to know the IP address of the storage device. However, Ethernet-based NVMe lacks a discovery service and cannot detect the status of access devices, manual configuration of storage IP addresses is required on host. Manual configuration is complex and error-prone.

Sub-Scenario 2: Expansion During expansion, hosts or storage devices need to be added. The problem is the same as that in sub-scenario 1. When a new host is mounted to a storage device, you need to manually configure the storage device, which is complex too.

Sub-scenario 3: Storage Faults When a storage device is faulty during running, no device proactively notifies the host of the fault status. Based on the Ethernet-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 the Ethernet-based NVMe, supports automatic device discovery and fault status notification. In the CDC(Centralized Discovery Controller) solution being discussed by NVMe organizations, hosts and storage devices report device information to the CDC, and the CDC synchronizes the information to the host. The host establishes an NVMe link to implement automatic device discovery. However, the CDC solution does not involve fault status notification. Therefore, the system cannot notify a fault status of a device, and the failover time is still long. For core services, the duration of service impact is critical.

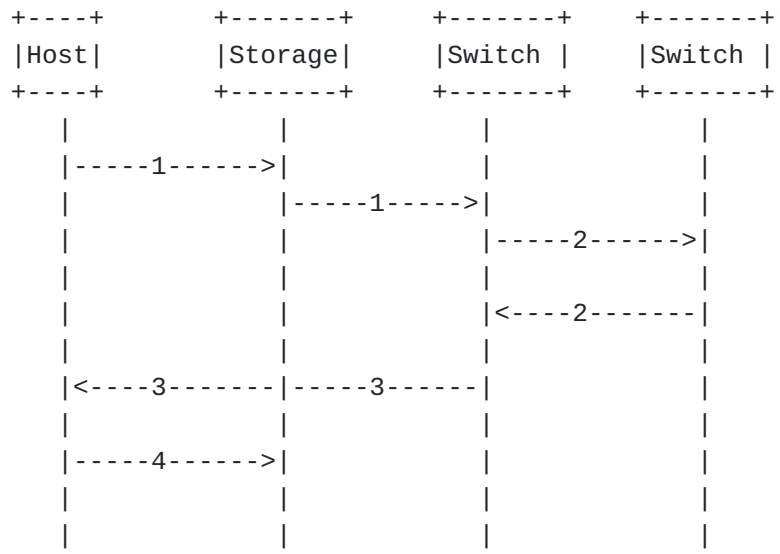
The solution proposed in this document is similar to the FC-NVMe system. The switch functions as the manager of the entire network, manages device information and status, and synchronizes information between switches on the entire network. In addition, to isolate storage services securely, a concept similar to a zone on a Fibre Channel network is introduced. Hosts and storage devices are planned in a zone and NVMe links can be established between the hosts and storage devices in the zone. The detailed requirements for hosts, switches, and storage devices on the network are as follows: Automatic device discovery: When a storage device is connected to a network, the host can automatically discover the storage device and establish an NVMe connection.

1. When the host accesses the network, the device access information is sent to the switch periodically. When the storage device is connected to the network, the device access information is sent to the switch periodically.

2. After receiving the host and storage access information, the switch synchronizes the information to other switches.

3. The switch identifies objects in the same zone and synchronizes host and storage device information to objects in the zone.

4. After receiving the storage device information provided by the switch, the host automatically establishes an NVMe connection.



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

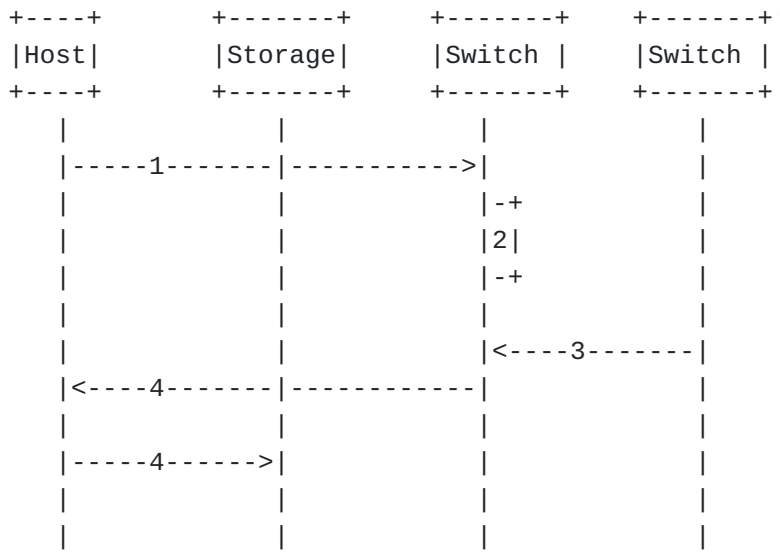
1. The host subscribes to the storage status information from the switch.

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

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

4. The switch identifies the hosts that subscribe to the storage status in the zone and synchronizes the storage fault information to the hosts.

5. 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.



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: yangzhiyong@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

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