White Paper: Data Centre Storage and Optical Multi-layer Coordination

OIF-SOC-WP-01.0

Date Approved – February 2025

Technical White Paper created and approved
OIF
www.oiforum.com

The OIF is an international non-profit organization with over 150 member companies, including the world's leading carriers and vendors. Being an industry group uniting representatives of the data and optical worlds, OIF's purpose is to accelerate the deployment of interoperable, cost-effective and robust optical internetworks and their associated technologies. Optical internetworks are data networks composed of routers and data switches interconnected by optical networking elements.

With the goal of promoting worldwide compatibility of optical internetworking products, the OIF actively supports and extends the work of national and international standards bodies. Working relationships or formal liaisons have been established with EA, IEEE 802.3, INCITS T11, Infiniband, IPEC, ITU SG-15, PCI-SIG, SNIA SFF.

For additional information contact:

OIF

39221 Paseo Padre Pkwy, Suite J

Fremont, CA 94538 USA

510-392-4903 – info@oiforum.com

www.oiforum.com

Working Group: **Networking & Operations WG**

TITLE: **Technical White Paper: Data Centre Storage and Optical Multi-layer Coordination**

SOURCE:

TECHNICAL EDITOR
Yi Lin
Huawei Technologies Co., Ltd.
Email: yi.lin@huawei.com

TECHNICAL EDITOR
Italo Busi
Huawei Technologies Co.,Ltd.
Email: Italo.Busi@huawei.com

WORKING GROUP CHAIR

Jia He
Huawei Technologies Co., Ltd.
Email: hejia@huawei.com

ABSTRACT: Transaction failures caused by jitters, intermittent disconnection and bit errors over the unstable Data Center Interconnection (DCI) links occur frequently in the financial industry. Fast fault detection and recovery of optical connections can be leveraged to address these issues by coordinating the storage application and optical connections in the DCI network. This white paper will delineate the anticipated scenarios, technical requirements, and potential solutions, and provide a comprehensive gap analysis regarding the storage and optical coordination.

Notice: This technical white paper ("White Paper") has been created by the Optical Internetworking Forum (OIF). This document is offered to the OIF members solely as a convenience and is not binding on any person or entity, including but not limited to, the OIF, its members, or the companies listed as resources above. The OIF reserves the rights to at any time to add, amend, or withdraw statements contained herein.

The user's attention is called to the possibility that implementation of the technical content of this White Paper ("Content") may require the use of inventions covered by the patent rights held by third parties.

THIS DOCUMENT AND THE CONTENT ARE PROVIDED ON AN "AS IS" BASIS. THE OIF EXPRESSLY DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED WITH RESPECT TO THIS DOCUMENT AND THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY IMPLIED WARRANTIES OF MERCHANTABILITY, TITLE OR FITNESS FOR A PARTICULAR PURPOSE, ANY REPRESENTATION OR WARRANTY THAT THE USE OR IMPLEMENTATION OF THIS DOCUMENT OR THE CONTENT WILL NOT INFRINGE ANY THIRD PARTY RIGHTS, AND ANY REPRESENTATION OR WARRANTY WITH RESPECT TO ANY CLAIM THAT HAS BEEN OR MAY BE ASSERTED BY ANY THIRD PARTY IN CONNECTION WITH THE WHITE PAPER OR SUCH CONTENT, THE VALIDITY OF ANY PATENT RIGHTS RELATED TO ANY SUCH CLAIM, OR THE EXTENT TO WHICH A LICENSE TO USE ANY SUCH RIGHTS MAY OR MAY NOT BE AVAILABLE OR THE TERMS HEREOF.

Copyright © 2025 Optical Internetworking Forum

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction other than the following: (1) the above copyright notice and this paragraph must be included on all such copies and derivative works, and (2) this document itself may not be modified in any way, such as by removing the copyright notice or references to the OIF, except as needed by OIF for the purpose of developing OIF work product.

By downloading, copying, or using this document in any manner, the user agrees to the terms and conditions of this notice.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

Optical network (e.g. OTN (Optical Transport Network), DWDM (Dense Wavelength Division Multiplexing)) is used for data center interconnection (DCI)/production center interconnection, with the merits of guaranteed high bandwidth, ultra-low latency and high availability. An example is shown in Figure 1.

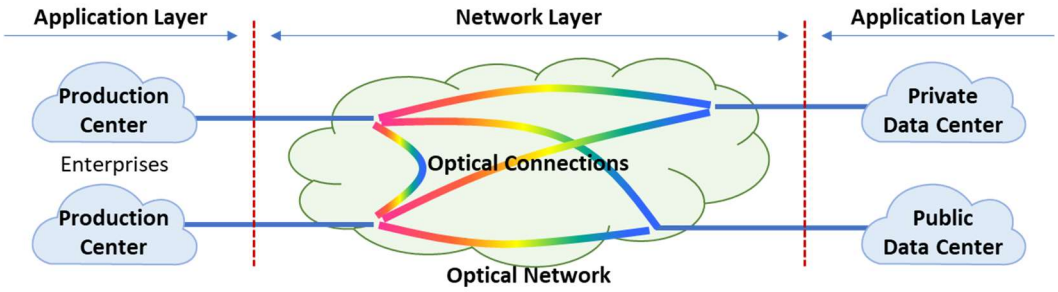

Figure 1 Optical Network Used in Data Center Interconnection/Production Center Interconnection

Public infrastructure construction, quality degradation of the link, etc. may cause jitter, intermittent disconnection and bit errors over the Data Center Interconnection (DCI) links. Critical services, for example in the financial production system, are sensitive to the problems caused by the unstable links. In the financial disaster recovery system, network jitter and bit errors may result in increased storage I/O latency and therefore storage read/write failures, which will eventually evolve into transaction failures. Fast fault detection of optical connection failures or degradations can be leveraged to address these issues by coordinating the storage applications and optical connections in the DCI network.

The white paper will focus on the coordination between the optical network and the storage application.

The whitepaper will include the following content:

- Scenarios where the coordination between storage application and optical connections is required.
- Key technical requirements and assumptions
- Applicability of generic multi-layer protection solutions
- Gap analysis to enable the coordination.

The detailed protocols is for further study, which will NOT be defined in this whitepaper.

2 Scenarios and Technical Requirements

Figure 2 shows a disaster recovery scenario for the financial production system. In this scenario, optical connections may become unstable due to unexpected events, while the storage devices cannot detect the network interruption in time.

This is due to the fact that failure detection mechanisms defined for Storage Area Network (SAN) technologies, such as Fibre Channel (FC), do not support fast failure detection of network failures (the detection time is typically a multiplier of 5s) nor detection of Signal Degrade (SD) conditions of the optical networks.

As a result, services between active systems may be suspended for a long period, leading to timeout and failure of bank transactions.

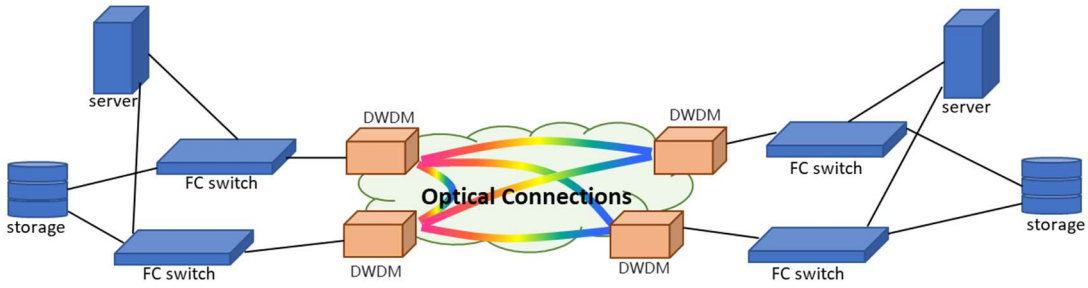

Figure 2 Disaster Recovery Scenario for the Financial Production System

To solve the problems, the key technical requirements are summarized as follows:

- **Awareness of network status**: Visibility of network status (e.g., connection degradation, bit error, intermittent disconnection, fiber cut, …) by the application layer, to enable integrated management of the storage application and optical connections.

- **High service reliability**: Coordination between the storage application and optical connections, to enable fast storage switchover.

NOTE – As described in section 7.2.3 (example 3), 1+1 protection mechanism at the client layer has some limitations which can be addressed by improving the detection mechanisms in the SAN technology or by some DCN signaling, as described in section 7.2.4 (example 4).

In order to achieve high service reliability, the storage devices can be dual-homed to one or two Fiber Channel (FC) switches (two FC switches are shown in Figure 2) and, at the edge between the DC network and the Optical network, one or two FC switches (two FC switches are shown in Figure 2) can be dual-homed to one or two Optical devices (two Optical devices are shown in Figure 2).

3 Solution Analysis

3.1 Architecture

Connecting clients in different data centers through an optical network is architecturally a multi-l

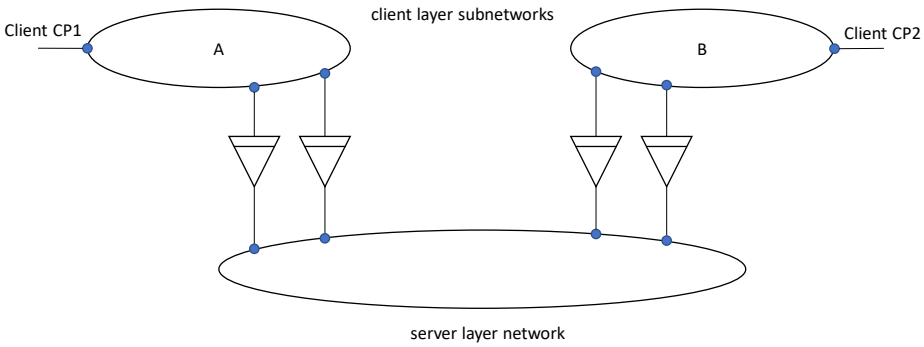

Figure 3: Client subnetworks connected by server layer

There are two client layer subnetworks connected by a server layer network. A connection from client Connection Point (CP) CP1 to client CP2 is not possible completely in the client layer network but the traffic from client CP1can be adapted to the server layer, placed into a server layer connection, adapted back to the client layer, and then delivered across client subnetwork B to client CP2.

NOTE – In case of redundant interconnections, the working and protection paths can be carried either over a common server layer network (as shown in Figure 3) or over different server layer networks (A/B plane split).

The generic multi-layer architecture of Figure 3 can be applied recursively to describe multi-layer scenarios with three or more layers.

In the Storage and OTN use case, there are three layers involved (e.g., Storage, Fibre Channel (FC) and OTN), as shown in Figure 4 (common OTN network) and Figure 5 (split OTN networks).

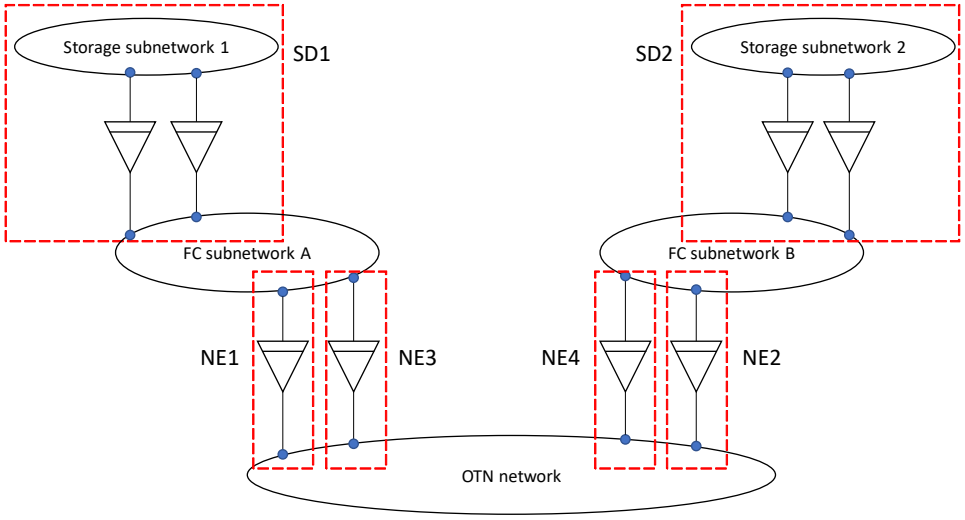

Figure 4: Storage and FC subnetworks connected over a common OTN network

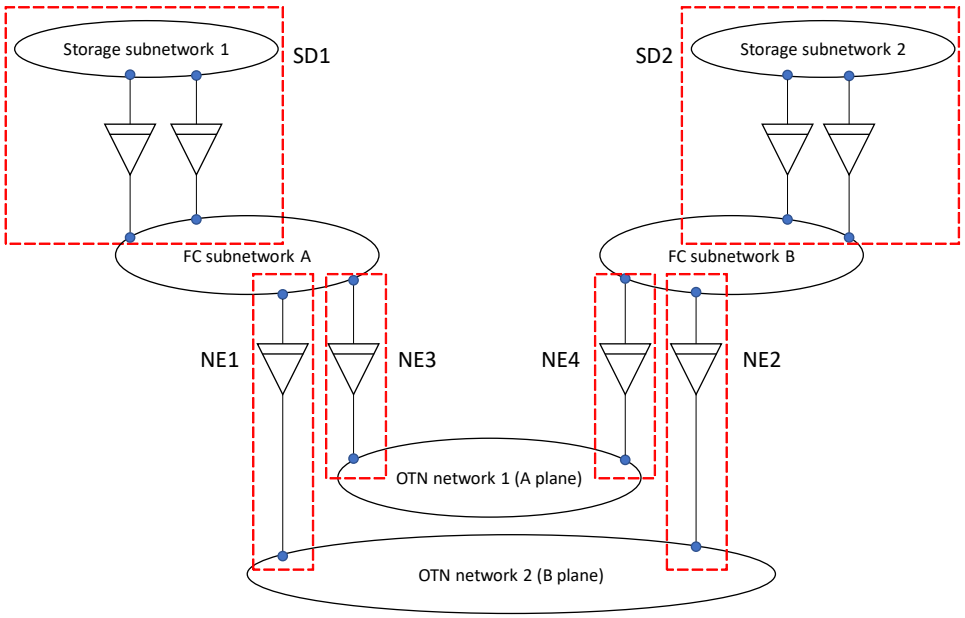

Figure 5: Storage and FC subnetworks connected over split OTN networks

It is worth noting that any SAN technology can be used as an intermedia layer network between Storage and OTN layers.

A specific interface like FC-1600 (16G Fibre Channel) at the edge between the FC subnetwork and the OTN network would be supported by the adaptation function to OPUflex as described in clause 17.9 of [ITU-T G.709]. The OPUflex frame is then adapted to ODUflex then to an OTU layer such as OTU4.

3.2 Reliability Scenarios

To survive failures, a solution framework should consider reliability methods using forwarding (or "data") plane functions as well as methods that use external functions acting on the data plane. The formers are generally known as protection methods using pre-assigned capacity between nodes and the latters as management/control methods. A useful protection architecture with multiple types is described in [ITU-T G.808]. Management/control methods are used to provide re-routing to establish connections and use of any capacity available between nodes to do this is known as "restoration" in [ITU-T G.808].

Communication to alert functions of resource changes (e.g., failures), to configure (e.g., forwarding tables), monitor, etc. may be supported within the layer network itself (e.g., if packet based) or might be an external communication network. An external communication network will be referred to as a Data Communication Network, or DCN, as described in [ITU-T G.7712].

Resiliency in a multi-layer network scenario, as shown in Figure 4, can be a combination of resiliency mechanisms in each of the layers (e.g., Storage, FC and OTN layers as shown in Figure 4).

A framework to analyze resilience solutions in a multi-layer scenario should incorporate at least the client and server layers, reliability methods and DCN as primary factors. This is given in Table 1:

Table 1: Resilience solution framework overview

	Data Plane Protection	**Management/Control Plane Restoration**	**DCN**
Storage Subnetworks	Yes/no	Yes/no	Yes/no
FC Subnetworks	Yes/no	Yes/no	Yes/no
OTN network	Yes/no	Yes/no	Yes/no

3.2.1 Example 1

Table 2 exhibits an example (Example 1) of primary factor combination for analyzing resilience solution for Storage and FC subnetworks connected over an OTN network.

Table 2: Resiliency example 1 overview

	Data Plane Protection	**Management/Control Plane Restoration**	**DCN**
Storage Subnetworks	No.	No.	No.

FC Subnetworks	No.	Yes: an SDN controller reconfigures the FC switches and the storage devices in case of failures of the FC links within the FC subnetwork.	Yes: Support communications between the SDN controller and the FC switches and storage devices.
OTN network	Yes: OTN ring protection	No	No

In this example, OTN ring protection, as defined in [ITU-T G.873.2], would recover any single failure or degradation within the OTN network, while FC restoration would recover single or multiple failures within the FC subnetwork, as shown in Figure 6 below.

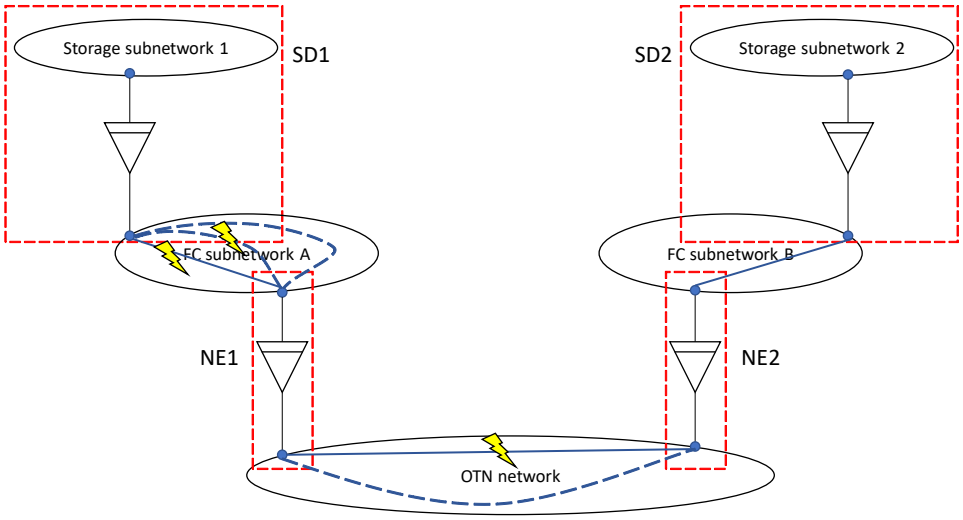

Figure 6: Recovered failures scenarios for example 1

However, in this example configuration, single failures at the edges between the FC subnetworks and the OTN networks (e.g., node failures of the edge FC switch or OTN switch, or failure of the FC link between them) cannot be recovered, as shown in Figure 7.

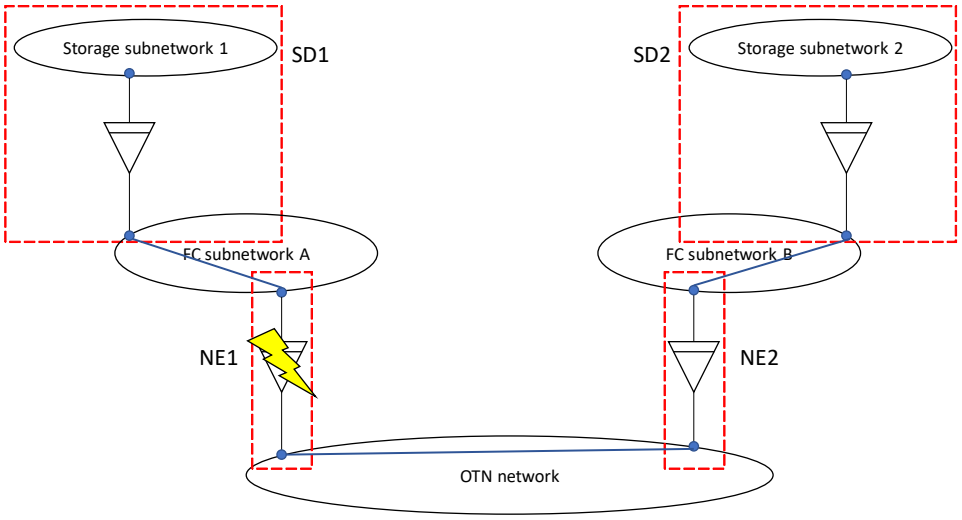

Figure 7: Unrecoverable failures scenarios for example 1

3.2.2 Example 2

Table 3 exhibits another example (Example 2) of primary factor combination for analyzing resilience solution for Storage and FC subnetworks connected over an OTN network.

Table 3: Resiliency example 2 overview

	Data Plane Protection	**Management/Control Plane Restoration**	**DCN**
Storage Subnetworks	No.	No.	No.
FC Subnetworks	No.	Yes: see Table 2.	Yes: see Table 2.
OTN network	Yes:1+1 OTN linear protection.	Yes: restoration as backup to protection using distributed OTN control plane.	Yes: support communications within the OTN control plane.

In this example, OTN linear protection, as defined in [ITU-T G.873.1], combined with OTN control plane restoration would recover single or multiple failures or degradations within the OTN network, while FC restoration would recover single or multiple failures within the FC subnetwork, as shown in Figure 8 below.

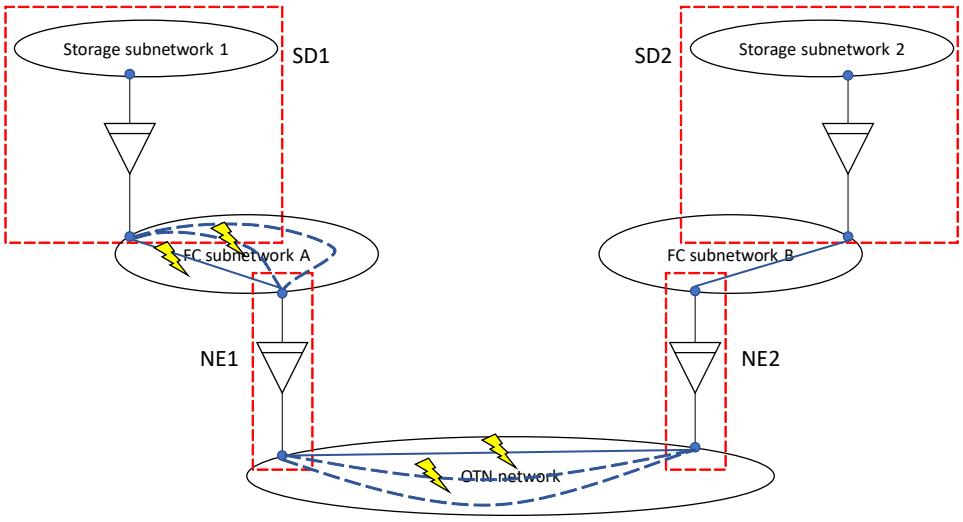

Figure 8: Recovered failures scenarios for example 2

However, as described for example 1, also in case of example 2, single failures at the edge between the FC subnetwork and the OTN network are not recovered.

3.2.3 Example 3

Table 4 exhibits a third example (Example 3) of primary factor combination for analyzing resilience solution for Storage and FC subnetworks connected over a common OTN network or split OTN networks.

Table 4: Resiliency example 3 overview

	Data Plane Protection	**Management/Control Plane Restoration**	**DCN**
Storage Subnetworks	Yes.	No.	No: end-to-end failure detection performed at application layer.
FC Subnetwork	No.	No.	No.
OTN network	No.	No.	No.

NOTE – This section describes only the case where the Storage and FC subnetworks are connected over a common OTN network but the same considerations can be applied when the Storage and FC subnetworks are connected over split OTN networks.

It is worth noting that in this case, unprotected connections are setup within the FC subnetworks and the OTN network.

In this example, the application-specific protection (at the storage layer) between the storage devices can recover any single failure scenario within either the OTN network, an

FC subnetwork or an edge between the OTN network and the FC subnetwork, as shown in Figure 9, Figure 10 and Figure 11.

The mechanisms used by the storage devices for the detection of the failure condition along the network and to perform protection switching at the application (storage) layer are implementation-specific. An example is provided in [VMWare ESXi/ESX].

Usually, using these mechanisms, failure conditions within the OTN network cannot be recovered in a timely manner.

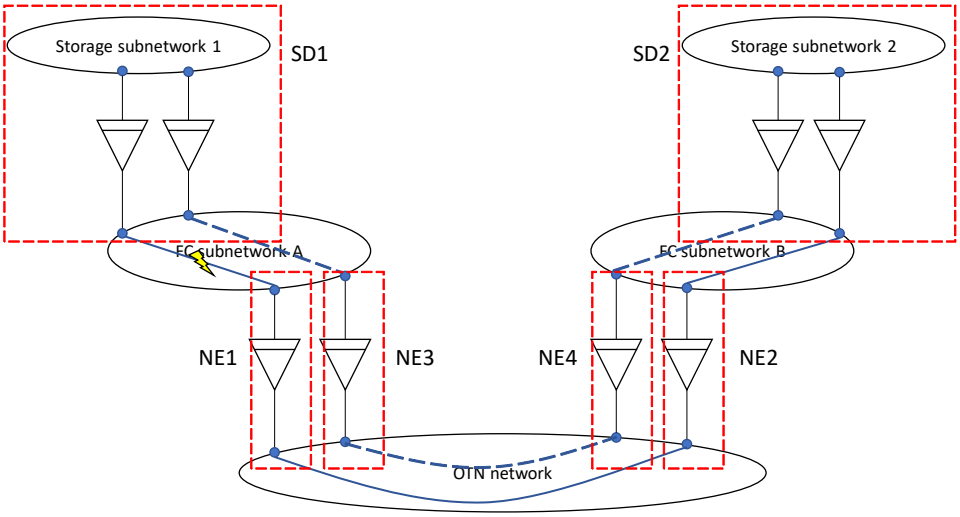

Figure 9: Recovered FC subnetwork failure scenario for example 3

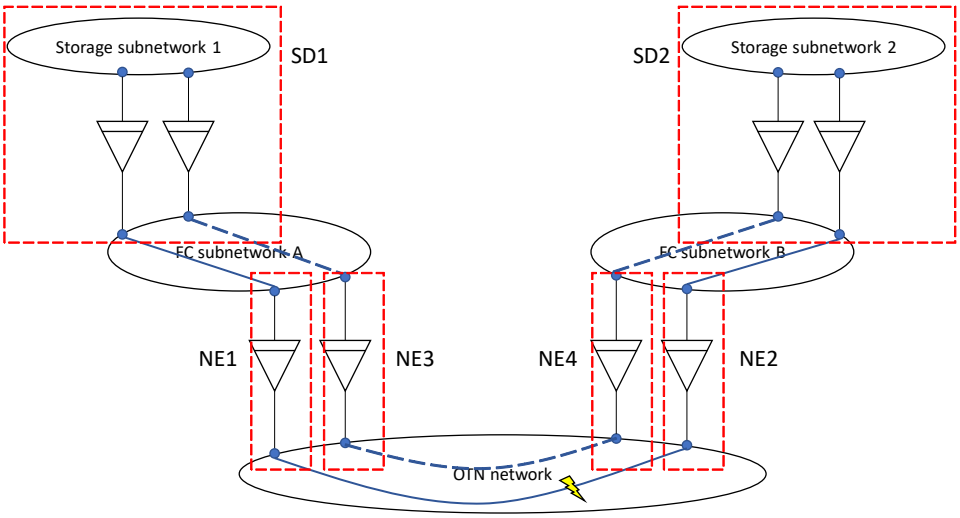

Figure 10: Recovered OTN failure scenario for example 3

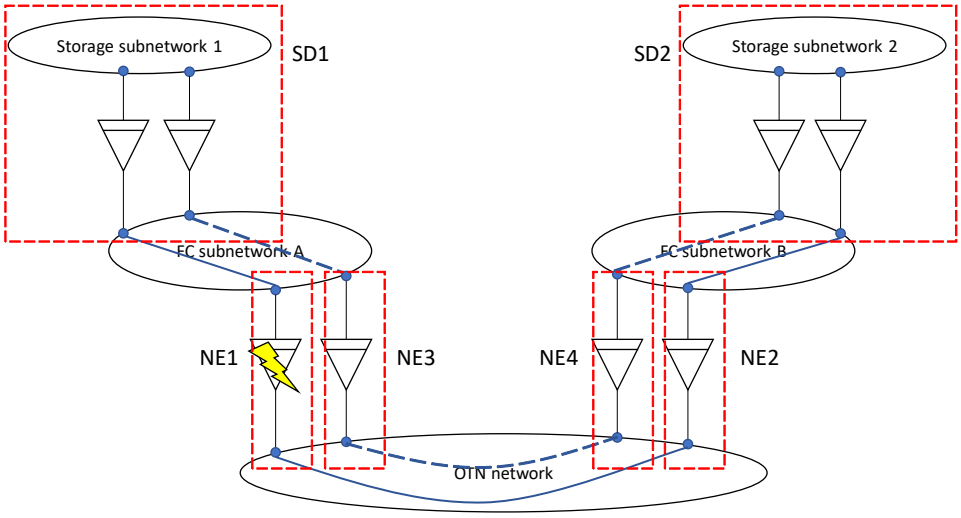

Figure 11: Recovered failure scenario at the FC/OTN edge for example 3

Depending on the mechanisms being implemented by the storage devices, degrade conditions within the OTN network cannot be recovered at all or cannot be recovered in a timely manner.

3.2.4 Example 4

Table 5 exhibits a fourth example (Example 4) of primary factor combination for analyzing resilience solution for Storage and FC subnetworks connected over common OTN network or split OTN networks.

Table 5: Resiliency example 4 overview

	Data Plane Protection	**Management/Control Plane Restoration**	**DCN**
Storage Subnetworks	Yes.	No.	Yes: for signal fail and degrade notifications from the edge OTN switches and the storage devices.
FC Subnetwork	No.	No.	No.
OTN network	No.	No.	No.

NOTE – This section describes only the case where the Storage and FC subnetworks are connected over a common OTN network but the same considerations can be applied when the Storage and FC subnetworks are connected over split OTN networks.

In this example, the application-specific linear protection (at the storage layer) between the storage devices can recover any single failure scenario within either the OTN

network, an FC subnetwork or an edge between the OTN network and the FC subnetwork, like in example 4, in a timely manner. Moreover, the application-specific linear protection can also recover any degrade condition within the OTN network in a timely manner.

It is worth mentioning that it is assumed that signaling messages sent on DCN from OTN to SAN are faster that the fault propagation and detection on the SAN data plane and that those mechanisms cannot distinguish the whether the faults happens in the local FC subnetwork or in the OTN server layer network.

Take signal failure or signal degrade condition within the OTN network for example. The storage device generates and sends services to the FC switch. The OTN device receives services from the FC switch via the interface connected to that FC switch and send the services to a target OTN device via the interface towards the optical network. The OTN device is configured to obtain the state information (e.g. signal failure, signal degrade) of an optical transmission path for transmitting the services, and send the state information of the optical transmission path to the storage device. The storage device is dual connected to two OTN devices creating the working path and the protection path in the optical network to gain higher reliability or just two optical paths for load balancing. Therefore, the storage device can receive status information of the two optical paths (either working/protection paths or two paths for load balancing), which could be used by the processor of the storage device to determine whether to perform input and output IO link switching.

4 Summary

The analysis shows that Example 4 creates more resilience with a simple notification sent to the application layer from the network layer. This also enables reliability in a timely manner which is not well supported (e.g. signal degrade) currently.

5 References

5.1 Normative references

1. [ITU-T G.709] ITU-T Recommendation G.709, “Interfaces for the optical transport network”, June 2020
2. [ITU-T G.808] ITU-T Recommendation G.808 Amendment 1, “Terms and definitions for network protection and restoration”, March 2018
3. [ITU-T G.873.1] ITU-T Recommendation G.873.1 Amendment 1, “Optical transport network: Linear protection”, February 2022.
4. [ITU-T G.873.2] ITU-T Recommendation G.873.2, “ODUk shared ring protection”, August 2015
5. [ITU-T G.7712] ITU-T Recommendation G.7712, “Architecture and specification of data communication network”, August 2010

5.2 Informative references

1. [VMWare ESXi/ESX] “VMware Multipathing policies in ESXi/ESX”, July 2024, <https://knowledge.broadcom.com/external/article?legacyId=1011340>