What's New in FC-NVMe-2

Live Webcast
October 15, 2020
11:00 AM PT/2:00 PM ET

About the Presenters

Mark Jones
Director Technical Marketing, Broadcom
FCIA Board of Directors

Craig W. Carlson
Senior Technologist with Marvell, Vice Chair T11, Chair T11 FC-NVMe-2 Working Group,
FCIA Board of Directors

Marcus Thordal
Broadcom
Principal Solution Architect

About the Fibre Channel Industry Association (FCIA)

25+ Years
Promoting Fibre Channel Technology

Industry Leading
Member Companies

142M+ FC Ports
Shipped Since 2001

Key Tenants of Fibre Channel

- Purpose-built as network fabric for storage and standardized in 1994, Fibre Channel (FC) is a complete networking solution, defining both the physical network infrastructure and the data transport protocols. Features include:
  - **Lossless, congestion free systems**—A credit-based flow control system ensures delivery of data as fast as the destination buffer can receive, without dropping frames or losing data.
  - **Multiple upper-layer protocols**—Fibre Channel is transparent and autonomous to the protocol mapped over it, including SCSI, TCP/IP, ESCON, and NVMe.
  - **Multiple topologies**—Fibre Channel supports point-to-point (2 ports) and switched fabric (224 ports) topologies.
  - **Multiple speeds**—Products are available supporting 8GFC, 16GFC, and 32GFC today.
  - **Security**—Communication can be protected with access controls (port binding, zoning, and LUN masking), authentication, and encryption.
  - **Resiliency**—Fibre Channel supports end-to-end and device-to-device flow control, multi-pathing, routing, and other features that provide load balancing, the ability to scale, self-healing, and rolling upgrades.

Agenda

- Background
- Fibre Channel Terminology
- Fibre Channel Basics
- The problem(s)…
- But, I thought it was reliable?
- The solution ☺
- Sequence level error recovery (SLER) example diagrams
- Summary

Background

Background

- Initial NVM Express over Fabrics spec(s) require an association to be terminated if a transport connection is lost; Admin or I/O ☹ i.e., big hammer
  - An association exists until the controller is shutdown, a Controller Level Reset, or the NVMe Transport connection is lost between the host and controller for the Admin or any I/O Queue.
- Ability to disconnect single I/O connection was added to NVMe-oF 1.1
  - An association between a host and controller is terminated if:
    - the controller is shutdown;
    - a Controller Level Reset occurs;
    - the NVMe Transport connection is lost between the host and controller for the Admin Queue; or
    - an NVMe Transport connection is lost between the host and controller for any I/O Queue and the host or controller does not support individual I/O Queue deletion

Fibre Channel Terminology

Terminology

- BLS – Basic Link Service (see FC-FS-6)
  - Basic Fibre Channel protocol Function
  - ABTS – Abort Sequence
    - BA_ACC – Basic Accept
    - BA_RJT – Basic Reject
- ELS – Extended Link Service (see FC-LS-5)
  - Extended Fibre Channel protocol function
  - LS_ACC – Link Service Accept
  - LS_RJT – Link Service Reject
  - REC – Read Exchange Concise ELS

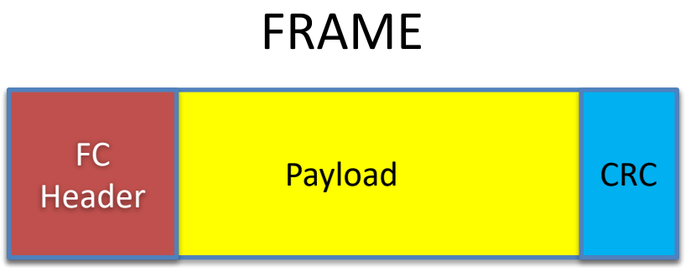
Fibre Channel Basics

Fibre Channel Basics – Frames, Sequences, and Exchanges

- **Fibre Channel data transfer has 3 fundamental constructs**
  - Frames – A “packet” of data
  - Sequences – A set of frames for larger data transfers
  - Exchanges – An associated set of commands and responses that make up a single command

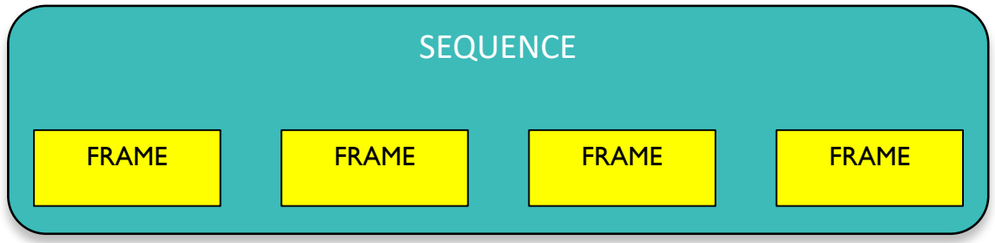
Frames

- **Each unit of transmission is called a “frame”**
  - A frame can be up to 2112 bytes
  - Each frame consists of a FC Header, payload, and CRC



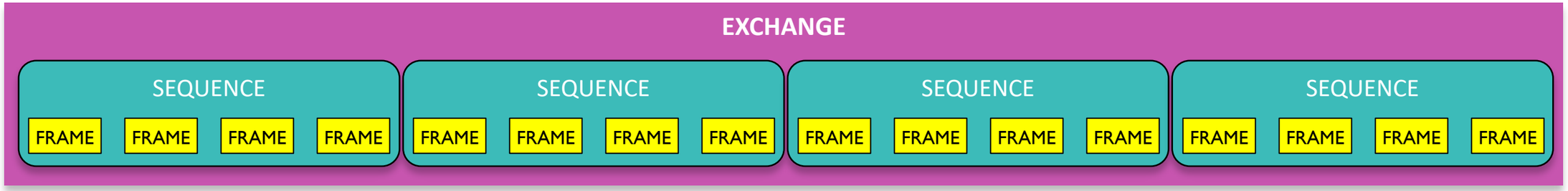
Sequences

- **Multiple frames can be bundled into a “Sequence”**
  - A Sequence can be used to transfer a large amounts of data
    - possibly up to multi-megabytes (instead of 2112 bytes for a single frame)



Exchanges

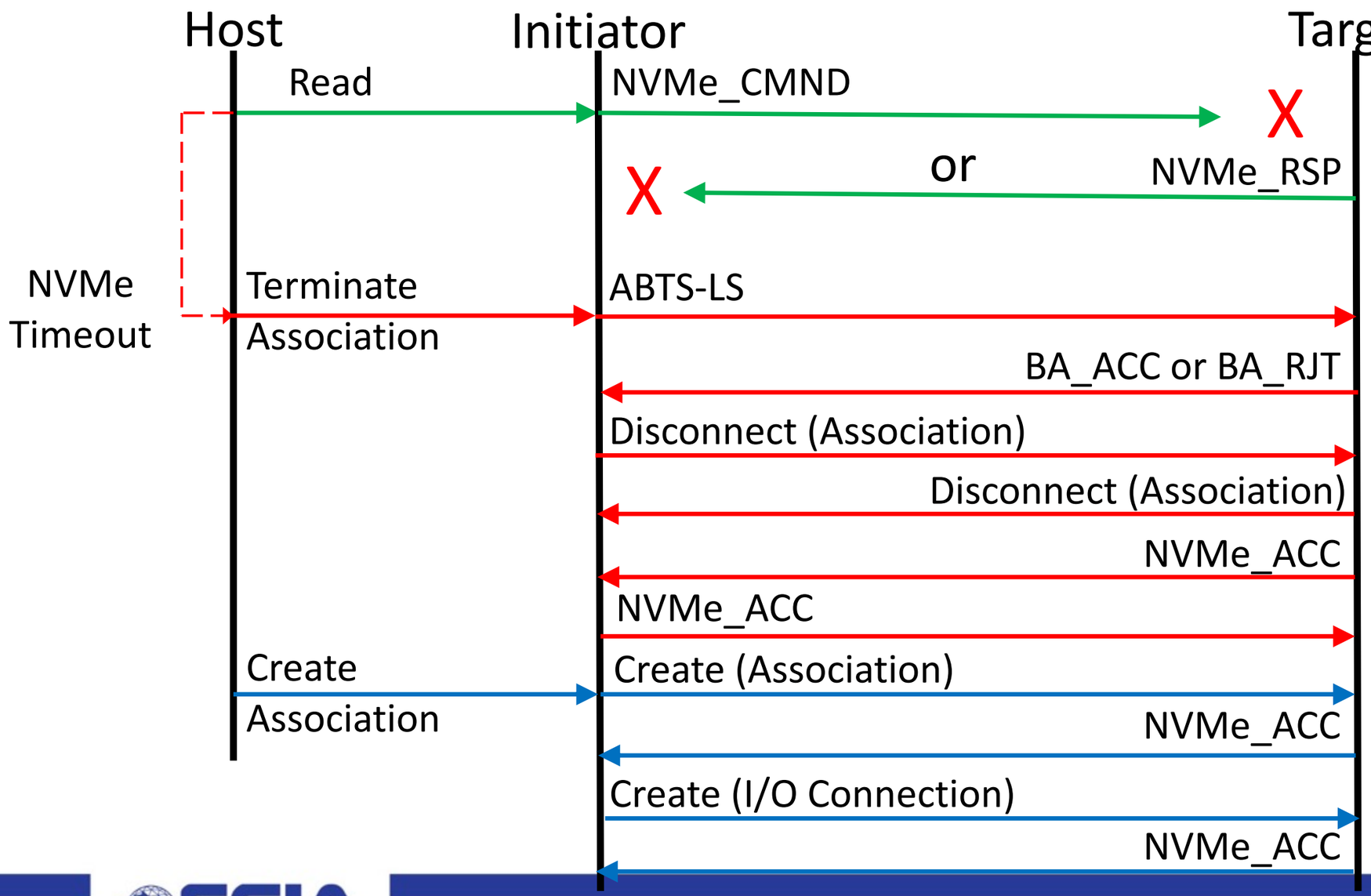
- **An interaction between two Fibre Channel ports is termed an "Exchange"**
  - Many protocols (including SCSI and FC-NVMe) use an Exchange as a single command/response
  - Individual frames within the same Exchange are guaranteed to be delivered in-order
  - Individual exchanges may take different routes through the fabric
    - This allows the Fabric to make efficient use of multiple paths between individual Fabric switches



*not to scale

The Problem

The Problem(s) –

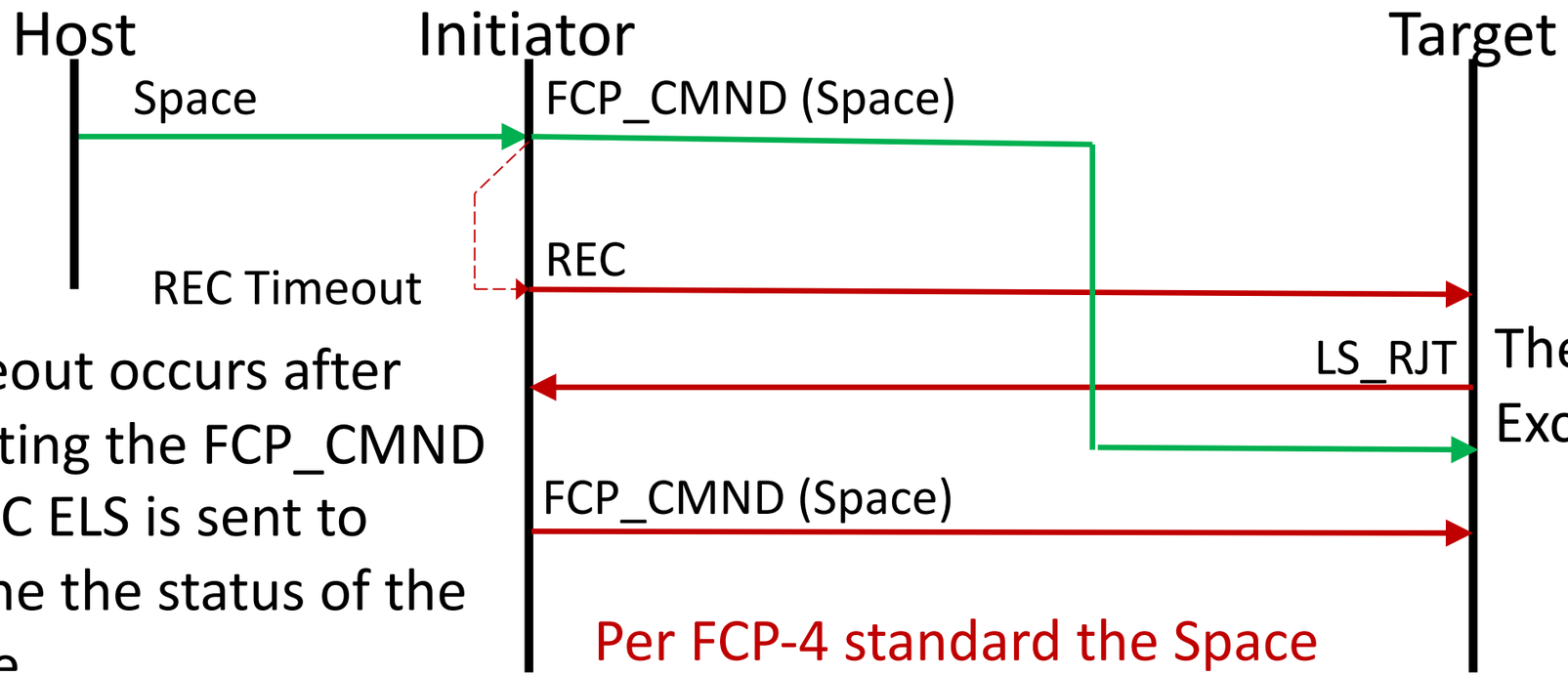


- First FC-NVMe (NVMeoFC) standard specified no capability to recover from an error during an Exchange i.e., big hammer
  - NVMeoFC association is terminated if any NVMeoFC connection for the NVMeoFC association is terminated

The Problem(s) –

- FCP-2/3/4 error detection and recovery was developed back in the days…when in-order delivery was from N_Port to N_Port
  - frames are delivered in-order
  - Exchanges are not delivered out-of-order
- And then came “Exchange-based routing”
  - frames are delivered in-order within an Exchange
  - Exchanges may be delivered out-of-order

The Problem(s) –



REC timeout occurs after transmitting the FCP_CMND and a REC ELS is sent to determine the status of the Exchange.

The LS_RJT indicates the Exchange is unknown.

Per FCP-4 standard the Space command is retransmitted in a new Exchange.

Houston, We Have a Problem…

- Big hammer error recovery approach does not work well for most FC deployments
- FCP-4 based error detection and recovery “is problematic”
- Thus new functionality was needed…

But, I Thought It Was Reliable?

Buy Why?

- I thought it was reliable?
  - Bit errors do happen
    - Actual bit errors tend to be much lower than theoretical occurrences
  - Software/hardware errors can also lead to frame loss

What Causes Bit Errors

Cosmic Rays from the sun and other sources.

Studies by IBM in the 1990s suggest that computers typically experience about one cosmic-ray-induced error per 256 megabytes of RAM per month.

Radiation from local environment

For modern chips care must be taken to minimize radiation from components

RF and power line noise from local equipment

Even changing generators at local power company can induce low frequency noise

What Causes Bit Errors

Software/hardware bugs

Need I say more?

Common specified Bit Error Rate is 10^{-12} to 10^{-15}

Actual bit error rate is often much better, but with theoretical rate, bits could occur multiple times per hour

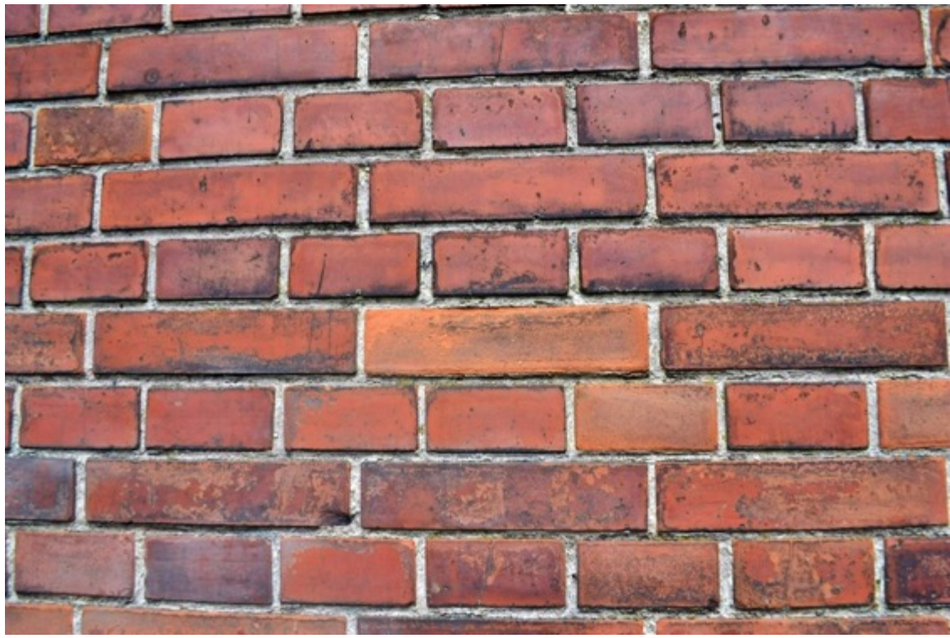
How Did This Work Before?

- Limited Error Recovery on the link
  - Low level error detection
  - FEC (Forward Error Correction) on high speed links
- Protocol Level Error Recovery
  - Both SCSI and NVMe have their own recovery mechanisms

The Solution

Enhanced Error Recovery

- Goal
  - Don’t let the protocol layer see any errors
    - Don’t want to rely on protocol level error recovery
- Enhanced Error Recovery
  - Detect and recover from errors before they reach the protocol layer
  - Protocol layer doesn’t even know anything happened



The Solution for Fibre Channel

- Perform all steps needed to recover from error(s) within the Exchange ☺
- Thus SLER – Sequence level error recovery
  - FLUSH request & response Basic Link Services (BLSs)
    - *No BA_ACC for either
    - *Sequence Initiative bit meaning changed – see FC-FS-6
  - Sequence Retransmission request & response (NVMe_SR/NVMe_SR_RSP) Information Units (IUs)
    - Generalized for other FC-4 protocols to use
  - RED Basic Link Service
    - Earlier indication of error sent from Target
  - Capability negotiated via Process Login (PRLI)

The Solution for Fibre Channel

- FLUSH request is periodically transmitted by Initiator Port to poll each outstanding Exchange to determine if
  - command is progressing properly
  - any Sequences have been received incorrectly
- Polling interval is controlled by FLUSH_TOV (default value $\geq 2\ \mathrm{sec}$)
- FLUSH response information from Target Port is compared with the expected state information at/known by Initiator Port
- If information is inconsistent (e.g., Target Port indicates it sent NVMe_RSP but the Initiator Port did not receive it), then error recovery actions may be performed to complete the Exchange

The Solution for Fibre Channel

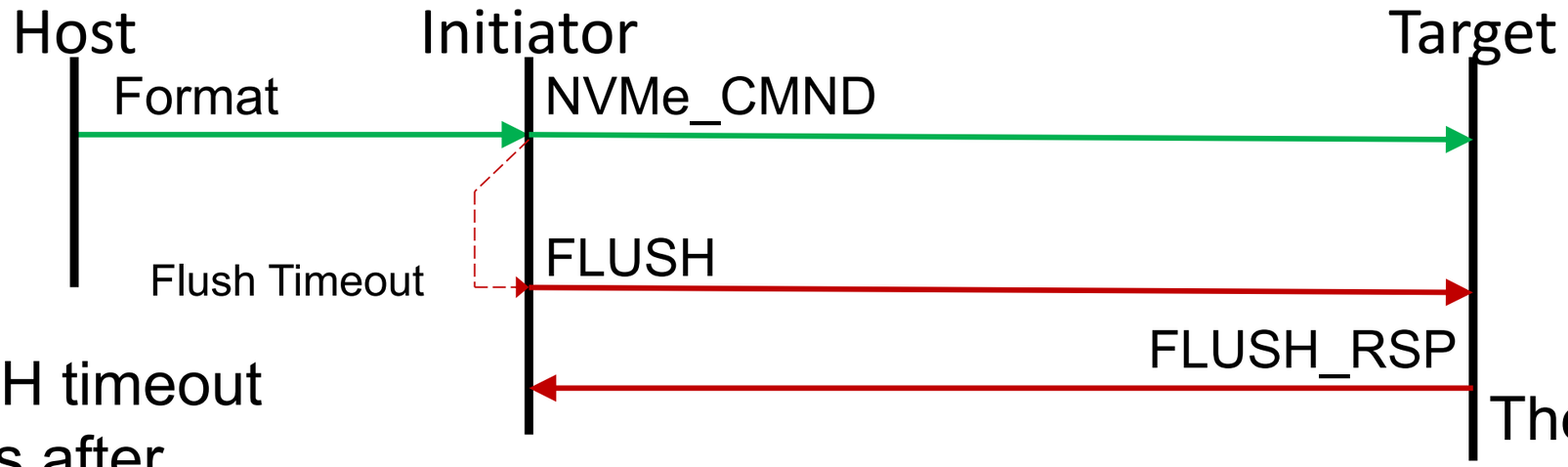
- Flush Exchange and verify status (FLUSH) request BLS
  - Sent from Initiator Port, or Target Port to determine if NVMe_CONF was sent
  - Transmitted within an Open Exchange
  - May be transmitted without Sequence Initiative
  - Parameter field – specified by FC-4 standard
    - For NVMeoFC - set to SLER qualifier associated with the Exchange

The Solution for Fibre Channel

- RED BLS
  - Sent by Target Port to Initiator Port to indicate a Sequence error was detected in an open Exchange
  - After sending RED the Target Port waits for FC-4 specific event before further processing of the Exchange
    - For NVMeoFC – the FC-4 specific event is an NVMe_SR IU
  - No Payload
  - No Reply Sequence

Sequence Level Error Recovery (SLER) Example Diagrams

Long Running Command

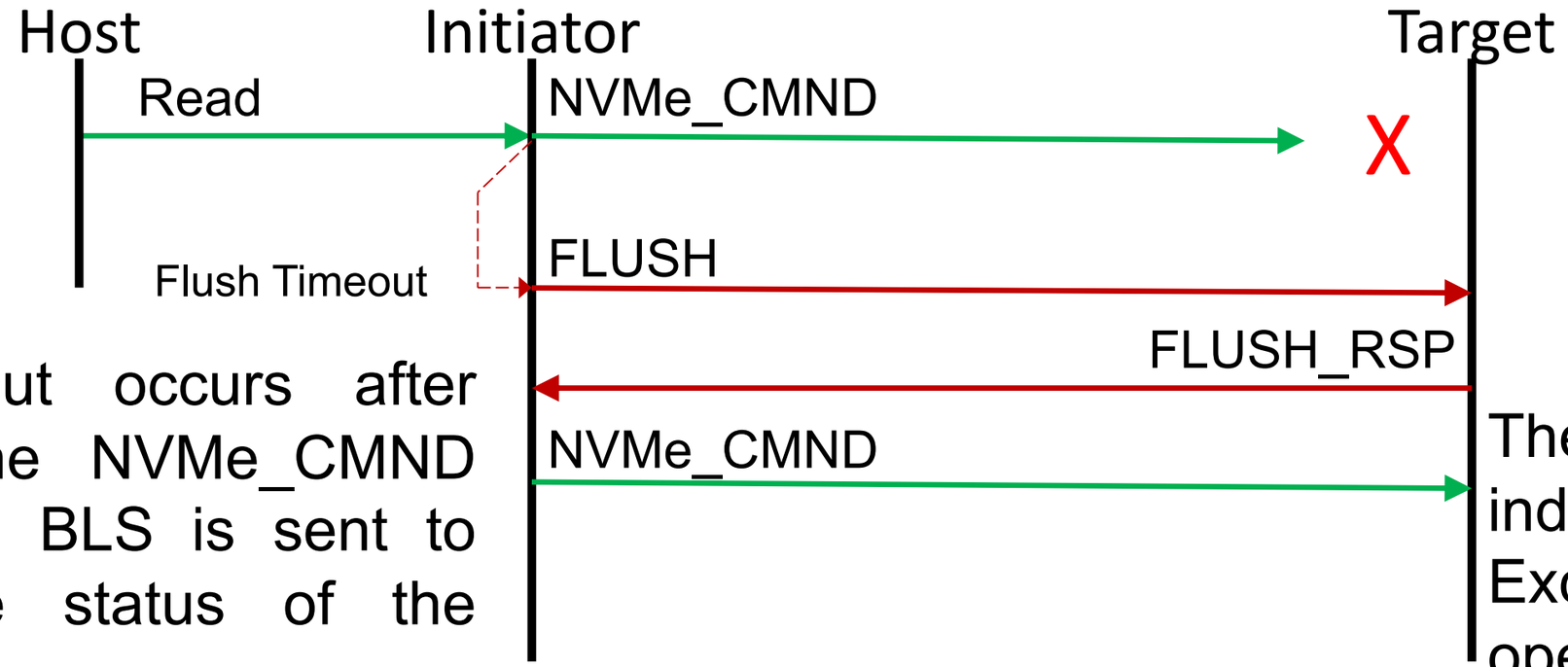


FLUSH timeout occurs after transmitting the NVMe_CMND and a FLUSH BLS is sent to determine the status of the Exchange.

No error recovery is required.

The FLUSH_RSP indicates the Exchange is open and the target NVMe_Port holds Sequence Initiative.

Lost Command

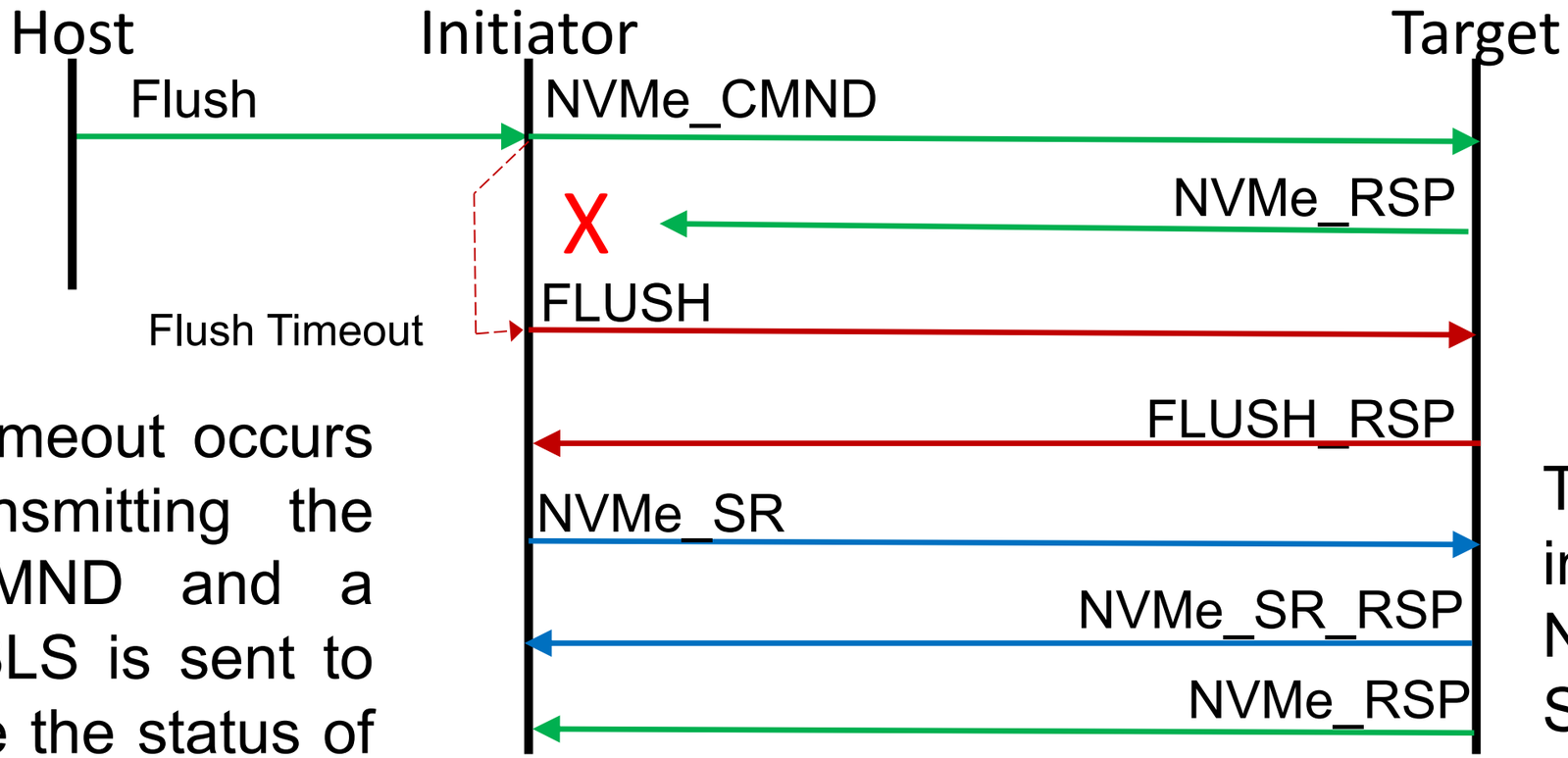


FLUSH timeout occurs after transmitting the NVMe_CMND and a FLUSH BLS is sent to determine the status of the Exchange.

The FLUSH_RSP indicates the Exchange is not open at the target, and the Exchange is closed.

The NVMe_CMND IU is retransmitted using the same OX_ID, SLER qualifier, and CSN.

Lost Response

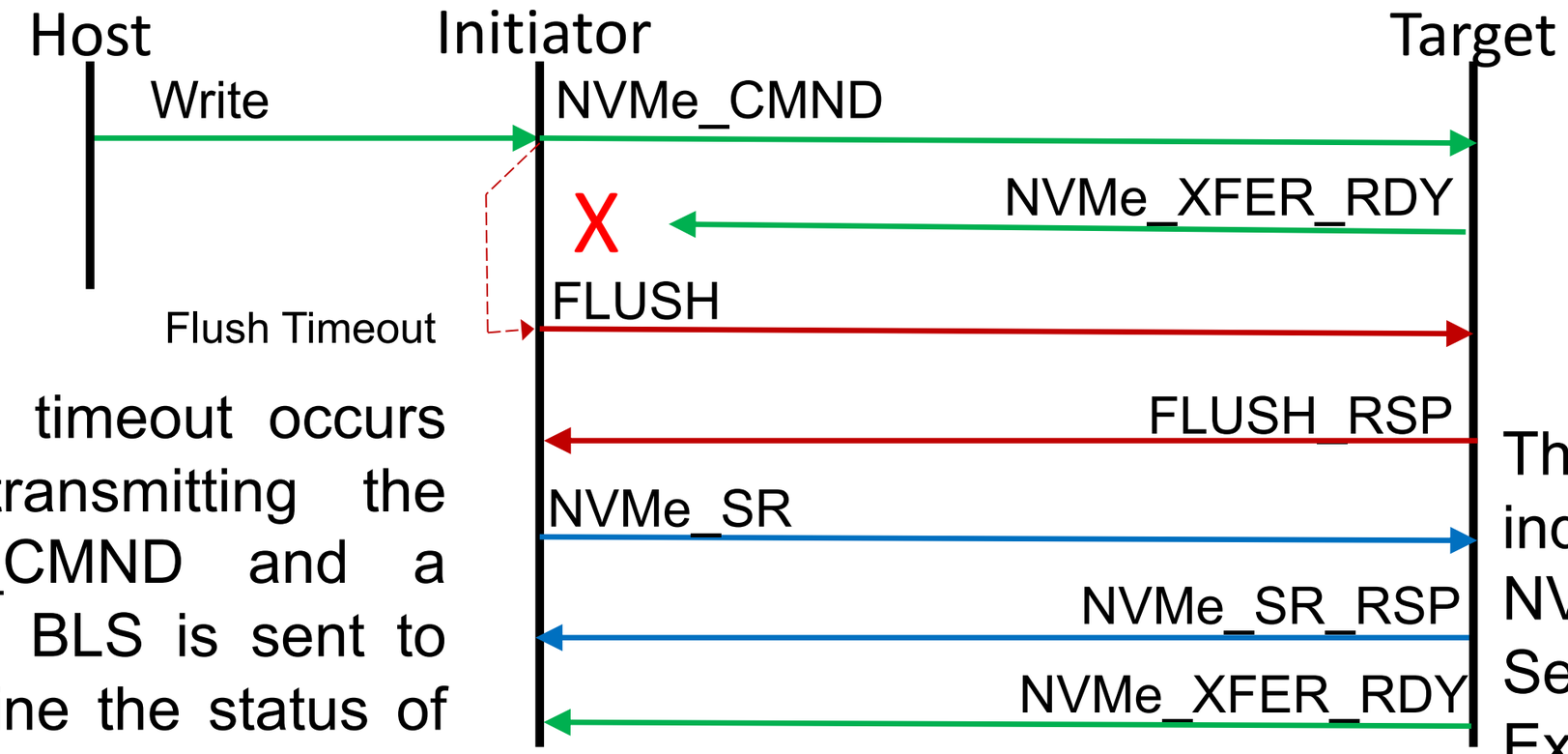


FLUSH timeout occurs after transmitting the NVMe_CMND and a FLUSH BLS is sent to determine the status of the Exchange.

The FLUSH_RSP indicates the initiator NVMe_Port holds Sequence Initiative and the Exchange is open.

The initiator NVMe_Port transmits an NVMe_SR IU specifying the NVMe_RSP be resent.

Lost Xfer_Rdy



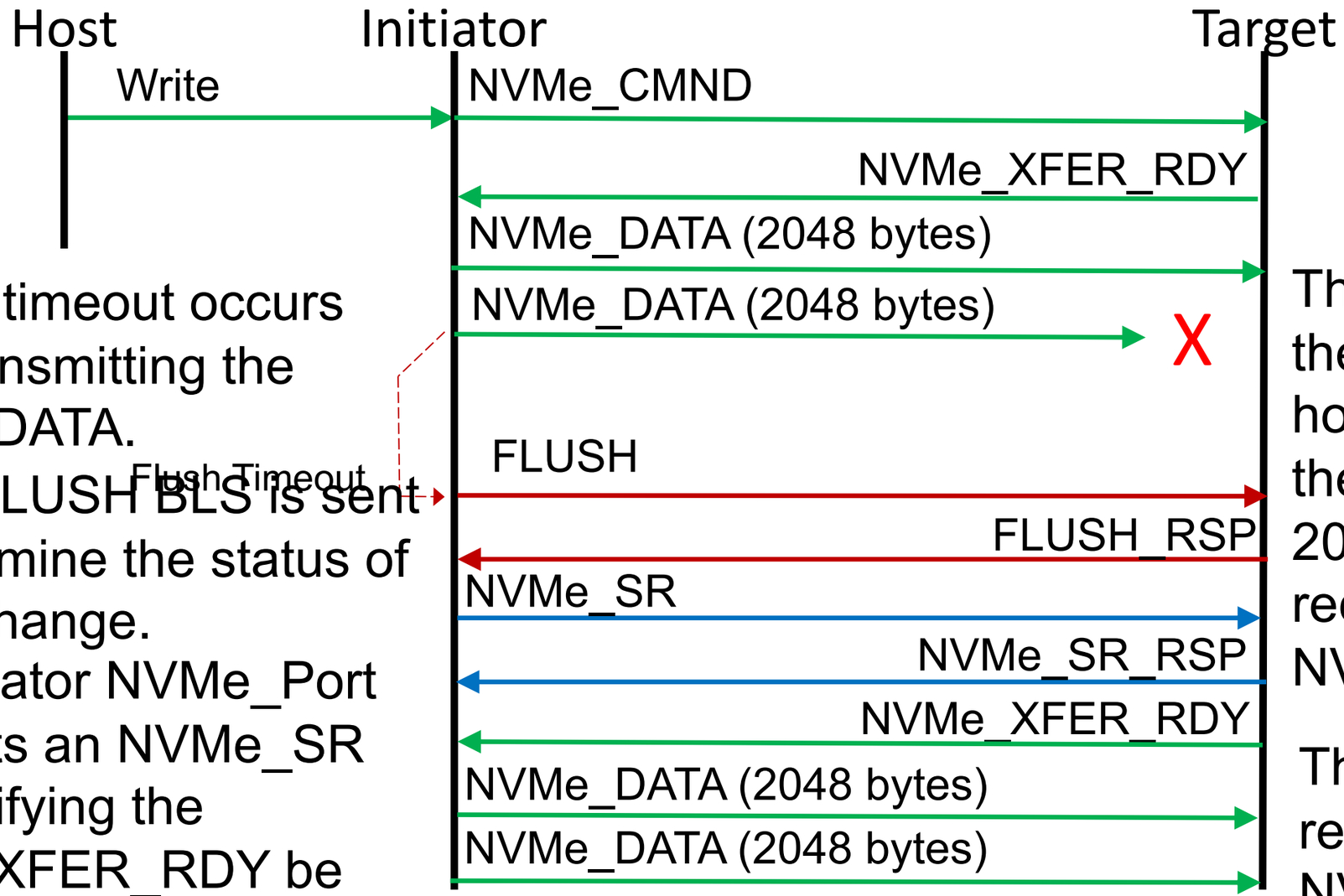
FLUSH timeout occurs after transmitting the NVMe_CMND and a FLUSH BLS is sent to determine the status of the Exchange.

The FLUSH_RSP indicates the initiator NVMe_Port holds Sequence Initiative and the Exchange is open.

The initiator NVMe_Port transmits an NVMe_SR IU specifying the NVMe_XFER_RDY be resent.

The target NVMe_Port retransmits the NVMe_XFER_RDY using Relative Offset zero.

Lost Write Data



FLUSH timeout occurs after transmitting the NVMe_DATA.

And a FLUSH BLS is sent to determine the status of the Exchange.

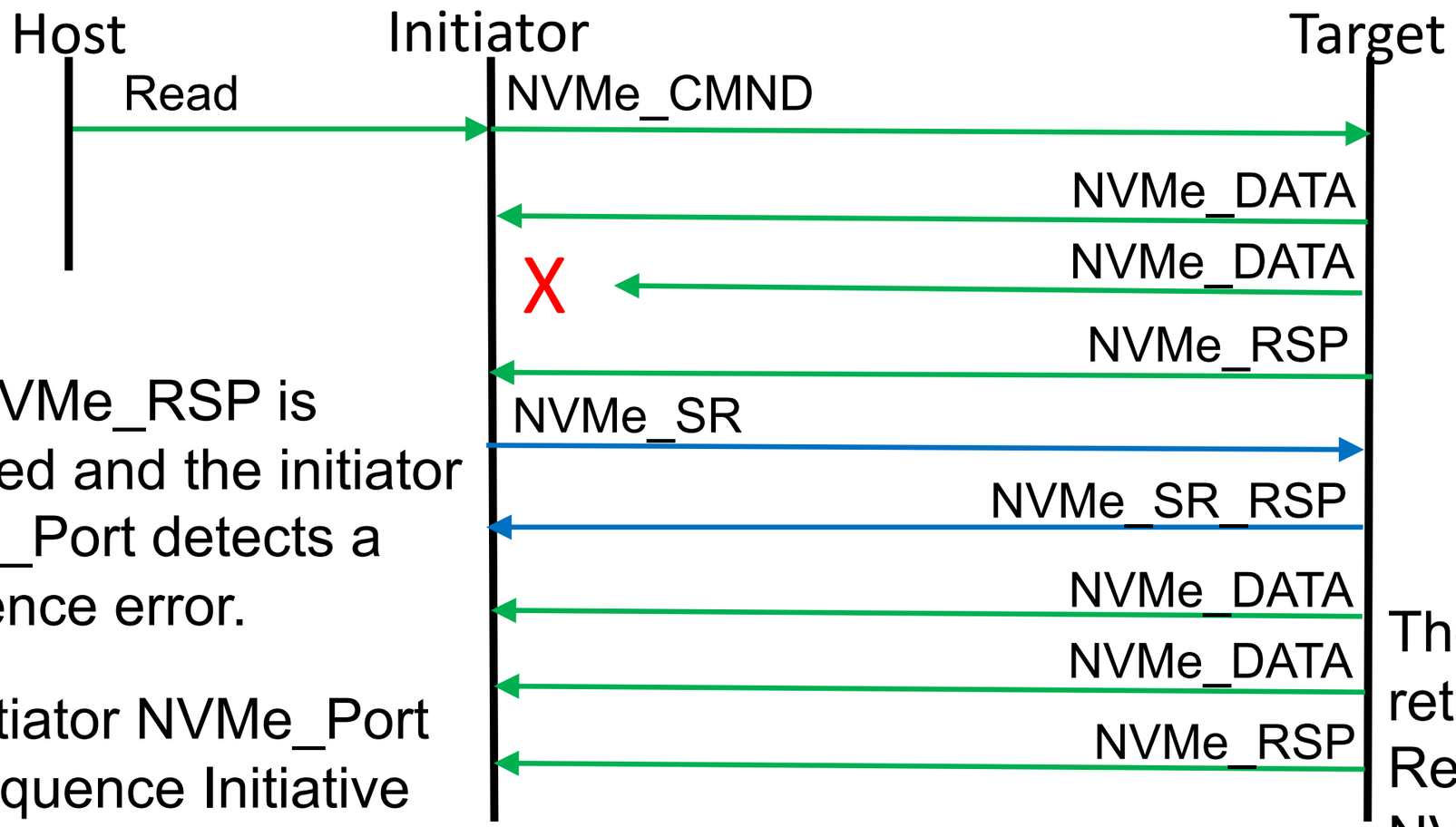
The initiator NVMe_Port transmits an NVMe_SR IU specifying the NVMe_XFER_RDY be resent.

The FLUSH_RSP indicates the initiator NVMe_Port holds Sequence Initiative, the Exchange is open, and 2048 bytes have been received by the target NVMe_Port.

The target NVMe_Port retransmits the NVMe_XFER_RDY with Relative Offset set to zero.

The initiator NVMe_Port retransmits the data.

Lost Read Data

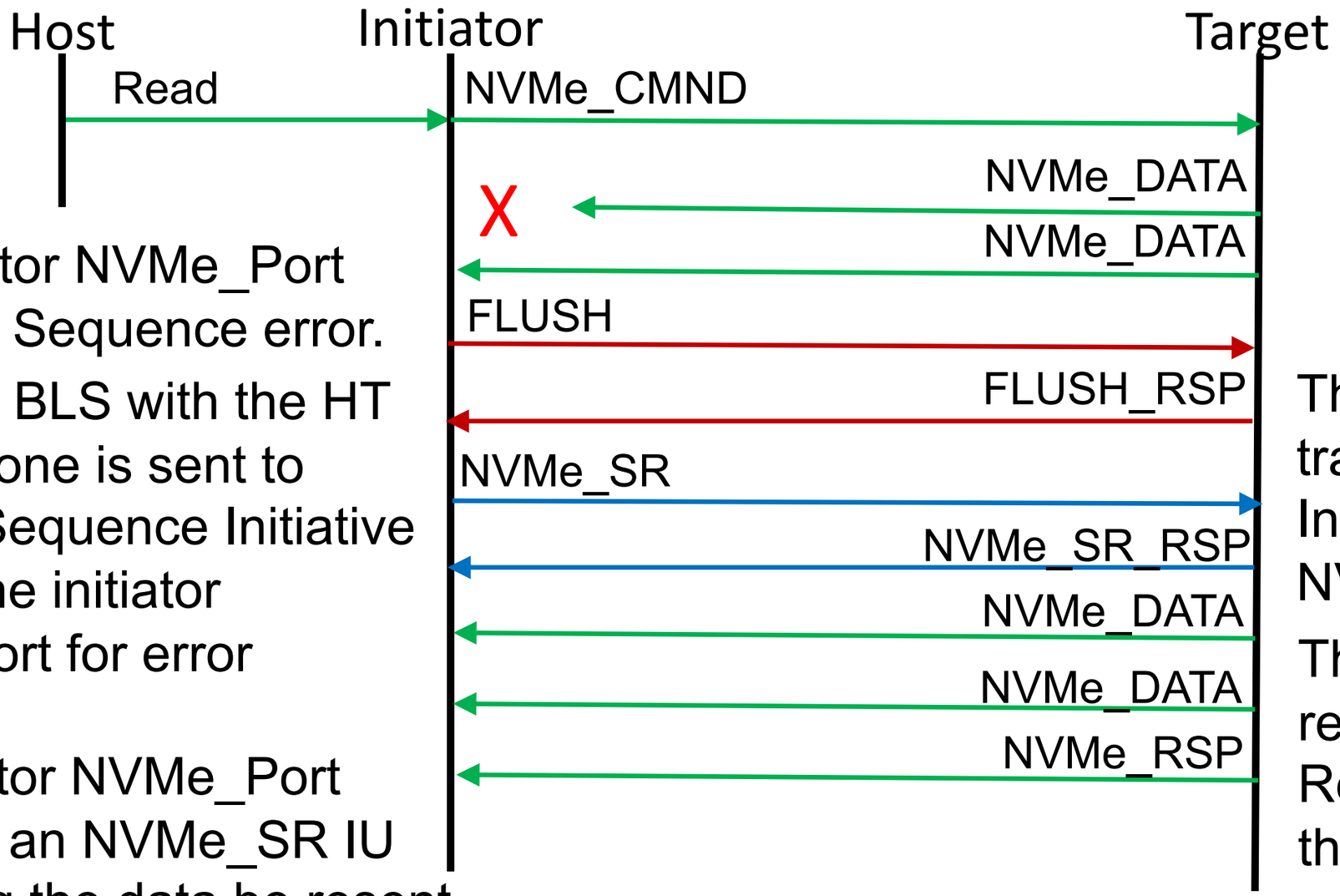


The NVMe_RSP is received and the initiator NVMe_Port detects a Sequence error.

The initiator NVMe_Port has Sequence Initiative and transmits an NVMe_SR IU specifying the data be resent.

The target NVMe_Port retransmits the data from Relative Offset zero, and the NVMe_RSP.

Lost Read Data



The initiator NVMe_Port detects a Sequence error.

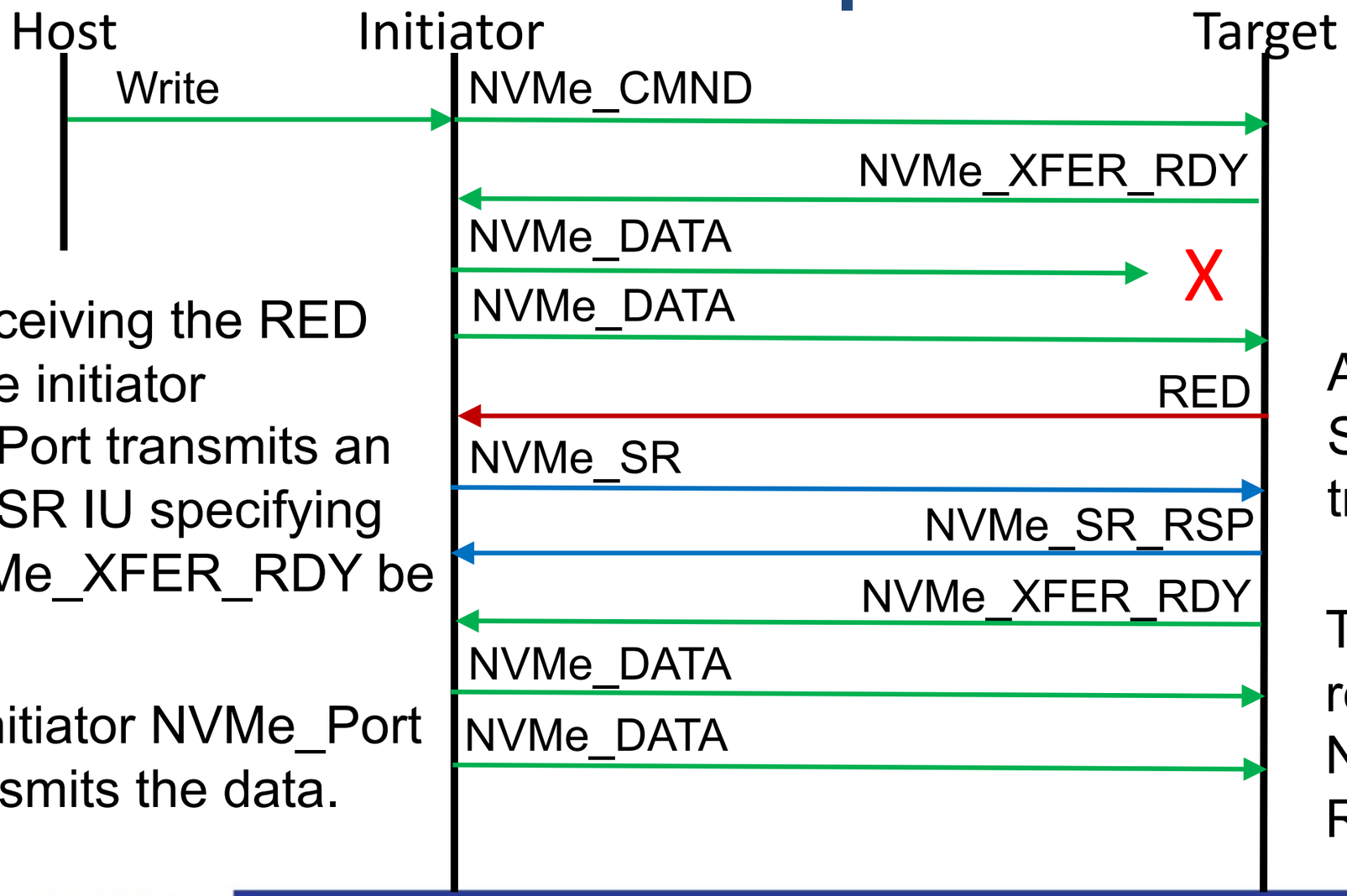
A FLUSH BLS with the HT bit set to one is sent to transfer Sequence Initiative back to the initiator NVMe_Port for error recovery.

The initiator NVMe_Port transmits an NVMe_SR IU specifying the data be resent.

The FLUSH_RSP transfers Sequence Initiative to the initiator NVMe_Port.

The target NVMe_Port retransmits the data from Relative Offset zero, and the NVMe_RSP.

Sequence Level Error Recovery (SLER) – Example RED



After receiving the RED BLS, the initiator NVMe_Port transmits an NVMe_SR IU specifying the NVMe_XFER_RDY be resent.

The initiator NVMe_Port retransmits the data.

After detecting a Sequence error the target transmits a RED BLS.

The target NVMe_Port retransmits the NVMe_XFER_RDY with Relative Offset set to zero.

Summary and Conclusion

Summary

- FC-NVMe-2 is Published and available via INCITS website
- FCP-5 rev 01 is posted on the INCITS T10 website
  - Arbitrated Loop and Class 2 support removed
  - SLER functionality being added as optional behavior

After this Webcast

- Please rate this event – we value your feedback
- We will post a Q&A blog at http://fibrechannel.org/ with answers to the questions we received today
- Follow us on Twitter @FCIAnews for updates on future FCIA webcasts
- Visit our library of on-demand webcasts at http://fibrechannel.org/webcasts/ to learn about:
  - Fibre Channel Fundamentals
  - FC-NVMe
  - Long Distance Fibre Channel
  - Fibre Channel Speedmap
  - FCIP (Extension): Data Protection and Business Continuity
  - Fibre Channel Performance
  - FICON
  - Fibre Channel Cabling
  - 64GFC
  - FC Zoning Basics
  - Fibre Channel Standards

Our Next Webcast

Fibre Channel Outlook – 2021 and Beyond

December 3, 2020

10:00 am PT/ 1:00 pm ET

Register at: https://www.brighttalk.com/webcast/14967/44646

Thank You