SDC 19
September 23-26, 2019
Santa Clara, CA

A Quantum Leap: NVMe over Fabrics with Fibre Channel, FC-NVMe2

Craig W. Carlson
Senior Technologist
Marvell

Rupin Mohan
Director R&D, CTO (SAN)
Hewlett Packard Enterprise

About the presenter

- Presented by: Craig W. Carlson
  - Senior Technologist, Marvell
  - Member of SNIA Technical Council
  - Chair of FC-NVMe working group within T11
  - Vice Chair T11 Fibre Channel
  - Chair T11.3 Committee on Fibre Channel Protocols
  - FCIA Board Member
  - NVMe Board Member

Agenda

- FC-NVMe-2
- Why Use FC-NVMe?
- 128GFC
- FCIA Roadmap
- Summary

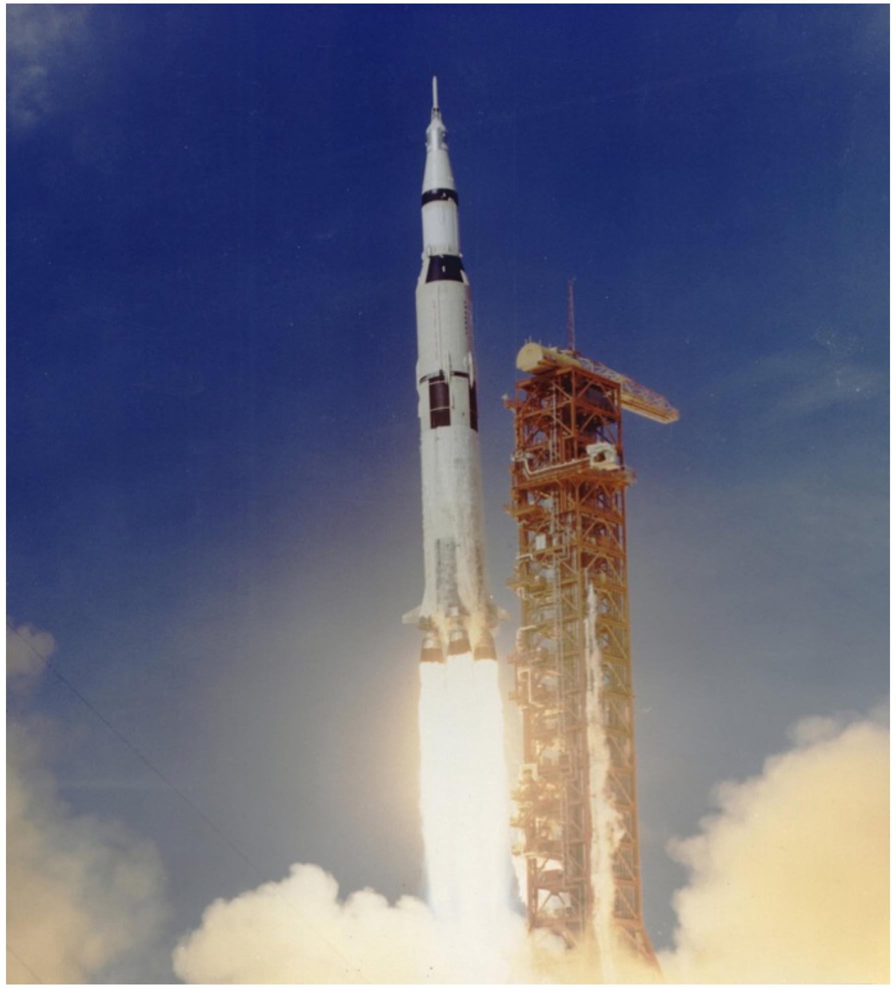
September 23-26, 2019
Santa Clara, CA

FC-NMVe REFRESHER

FC-NVMe

- **Design Goals**
  - Comply with NVMe over Fabrics Spec
  - High Performance/Low Latency
  - Use existing HBA and Switch hardware
    - Don't want to require new ASICs to be spun to support FC-NVMe
  - Fit into the existing FC infrastructure as much as possible, with very little real-time software management
  - Maintain Fibre Channel Fabric Services
    - Name Server
    - Zoning
    - Management

Performance



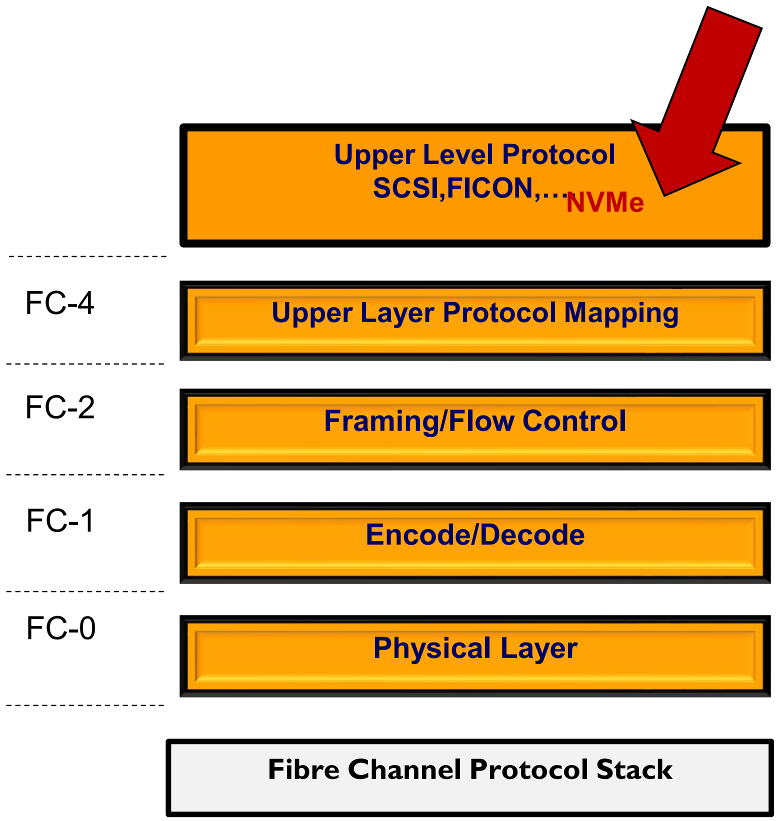
- **The Goal of High Performance/Low Latency**
  - Means that FC–NVMe needs to use an existing hardware accelerated data transfer protocol
  - Use FCP as the data transfer protocol
    - Currently both SCSI and FC-SB (FICON) use FCP for data transfers
    - FCP is deployed as hardware accelerated
    - Like FC, FCP is a connectionless protocol
      - Any FCP based protocols provide a way of creating a “connection”, or association between participating ports

SDC 19

FC-NVMe STATUS

FC-NVMe Standards and Partner Update

- FC-NVMe Standard (T11) ratified on 8/10/2017
- Use existing FC HBA and FC switch hardware
  - Co-existence of FCP SCSI and FC-NVMe traffic
- Performance:
  - Demonstrated low latency & high performance
- Availability
  - Linux based host drivers available now
- FC-NVMe-2 Started development spring ’18
  - Focusing on Enhanced Error Recovery
- FC-NVMe-2 Currently in approval phase



FC-NVMe Ecosystem Readiness

Element	FC-NVMe Support Status
FC Switches	Available today NOTE: Just ensure you right the right F/W version that supports FC-NVMe
HBAs	Host Side: • Linux Unified Driver available for download today Target Side: • User mode (SPDK), Kernel mode - alpha drivers available today • FW available today
Operating Systems	Linux Community and OS Vendors: • SLES12SP3, RHEL7.5 support FC-NVMe (Tech Preview) today • SLES12SP4/SLES15, RHEL7.6 support FC-NVMe GA Q3/Q4,2018 • VMware and Microsoft – engaged
Storage	Multiple vendors to support 2H 2018

September 23-26, 2019
Santa Clara, CA

FC-NVMe-2

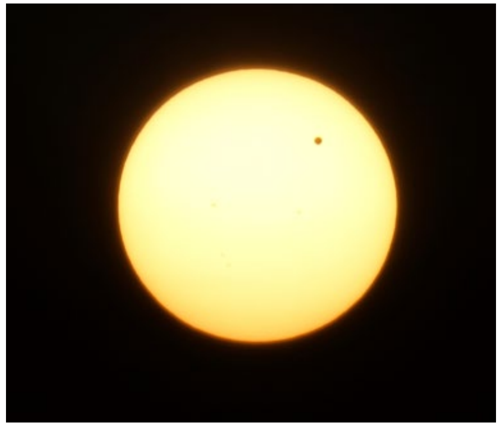
The next step

- The big new item in FC-NVMe-2 is Enhanced Error Recovery
- Allows errors (missing or corrupt frames) to be detected and recovered at the transport layer before the protocol layer knows anything was amiss

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}

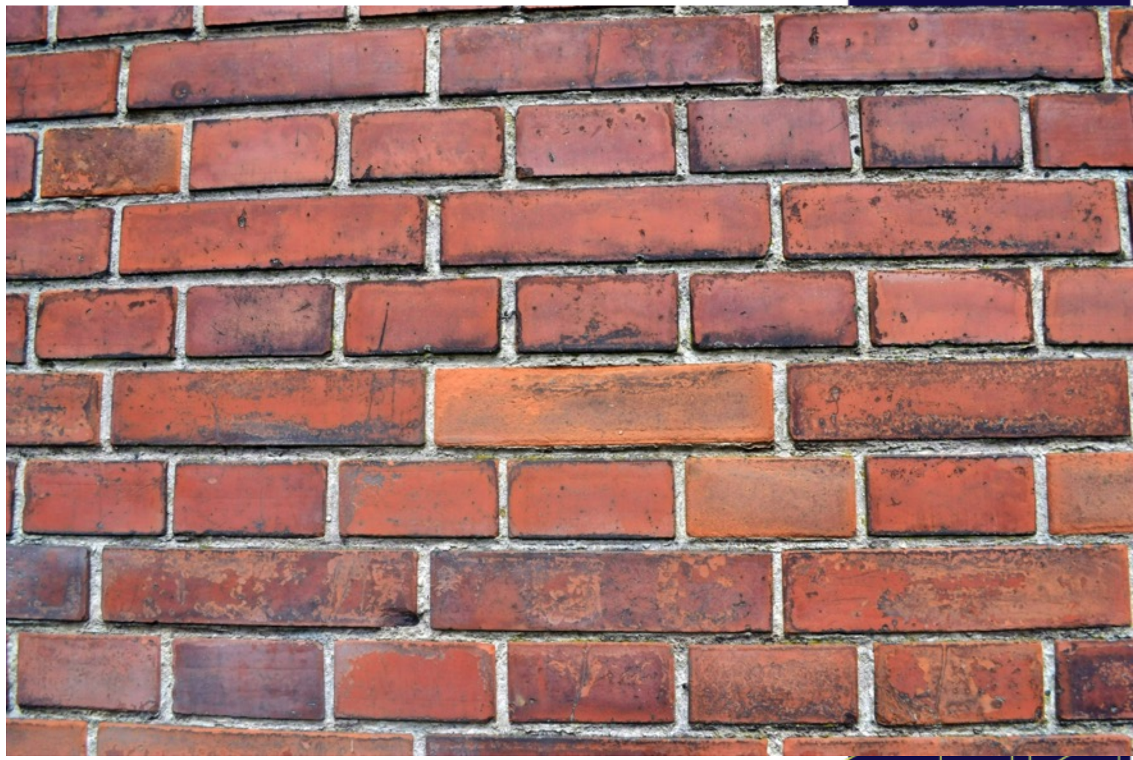
Actual bit error rate is often much better, but with theoretical rate, bits could occur multiple times per hour

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

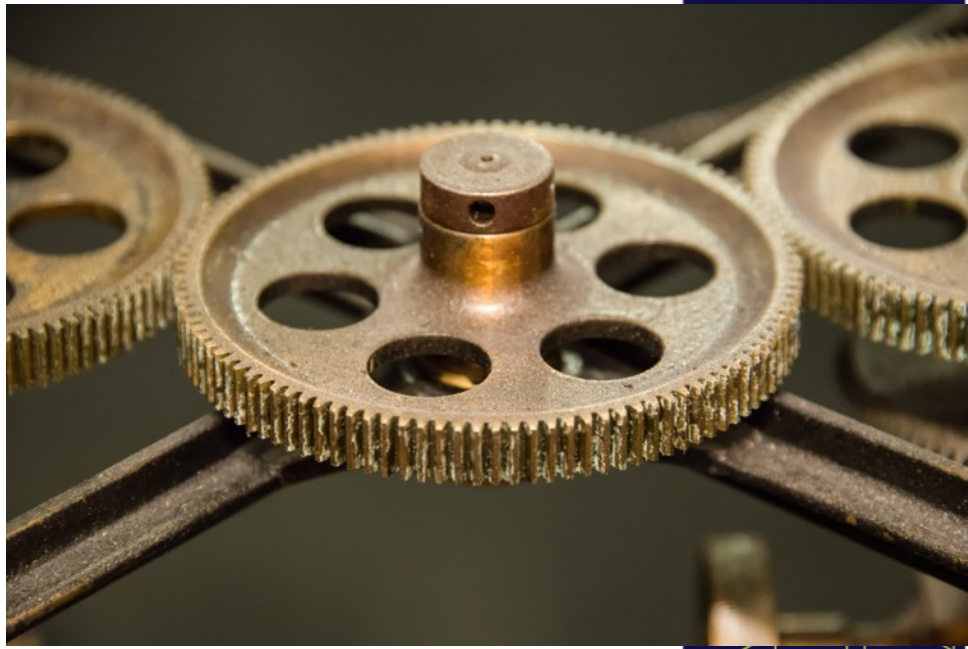
Enhanced Error Recovery

- 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



More details on Enhanced Error Recovery

- Error recovery takes place at FC Frame level
  - Missing frames timeout and are retransmitted
  - Defined new FC-NVMe servies and FC Basic Link Services for fast recovery
  - Protocol layer does not know anything happened



New Stuff

- New Basic Link Services
  - FLUSH
  - RED
- New FC-NVMe IUs
  - NVMe_SR

FLUSH

- New FLUSH BLS defined
  - (A BLS – Basic Link Service – is a low level Fibre Channel command frame)
  - Used to determine if a command is still active
    - Can be used to detect lost frames so that error recovery can start

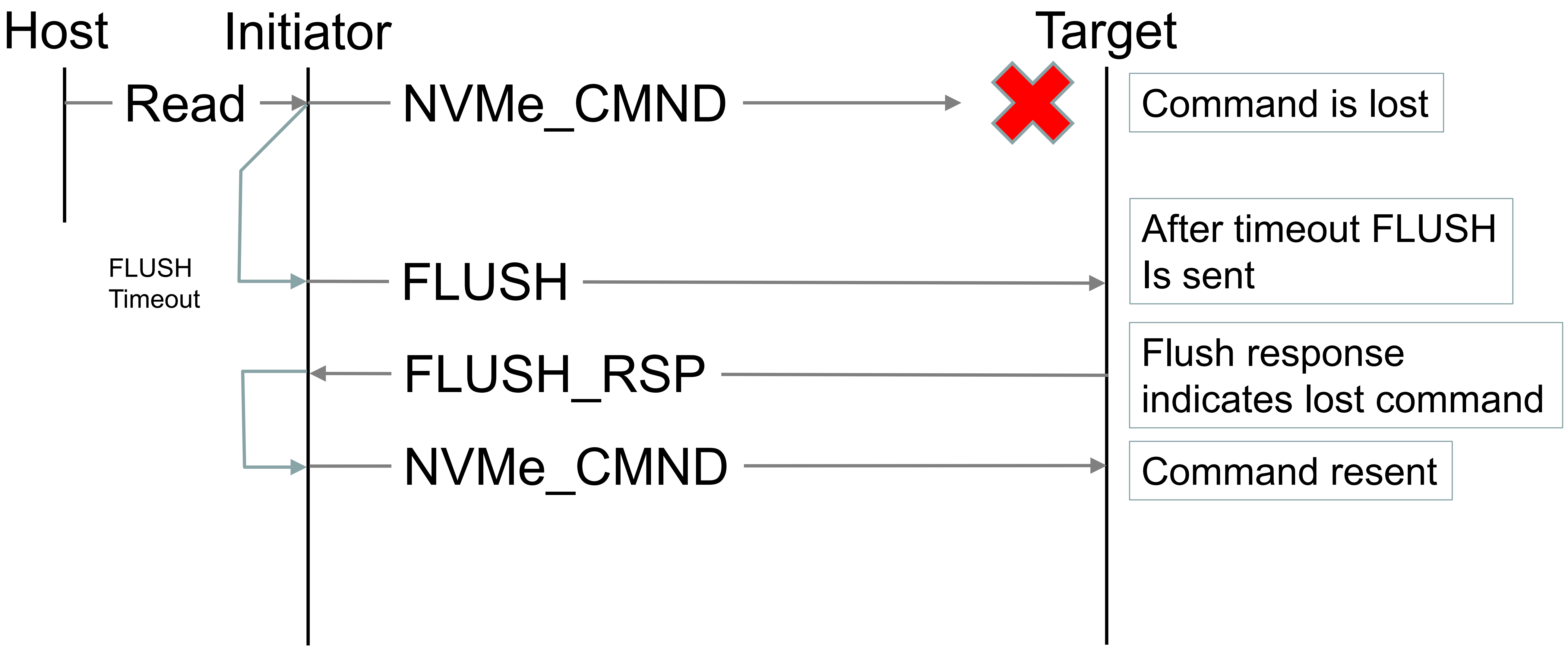
RED

- New RED BLS Defined
  - RED (Responder Error Detected) – Used by the recipient of a command to indicate that an error as been detected
  - Allows a receiver to indicate that error recovery should be initiated

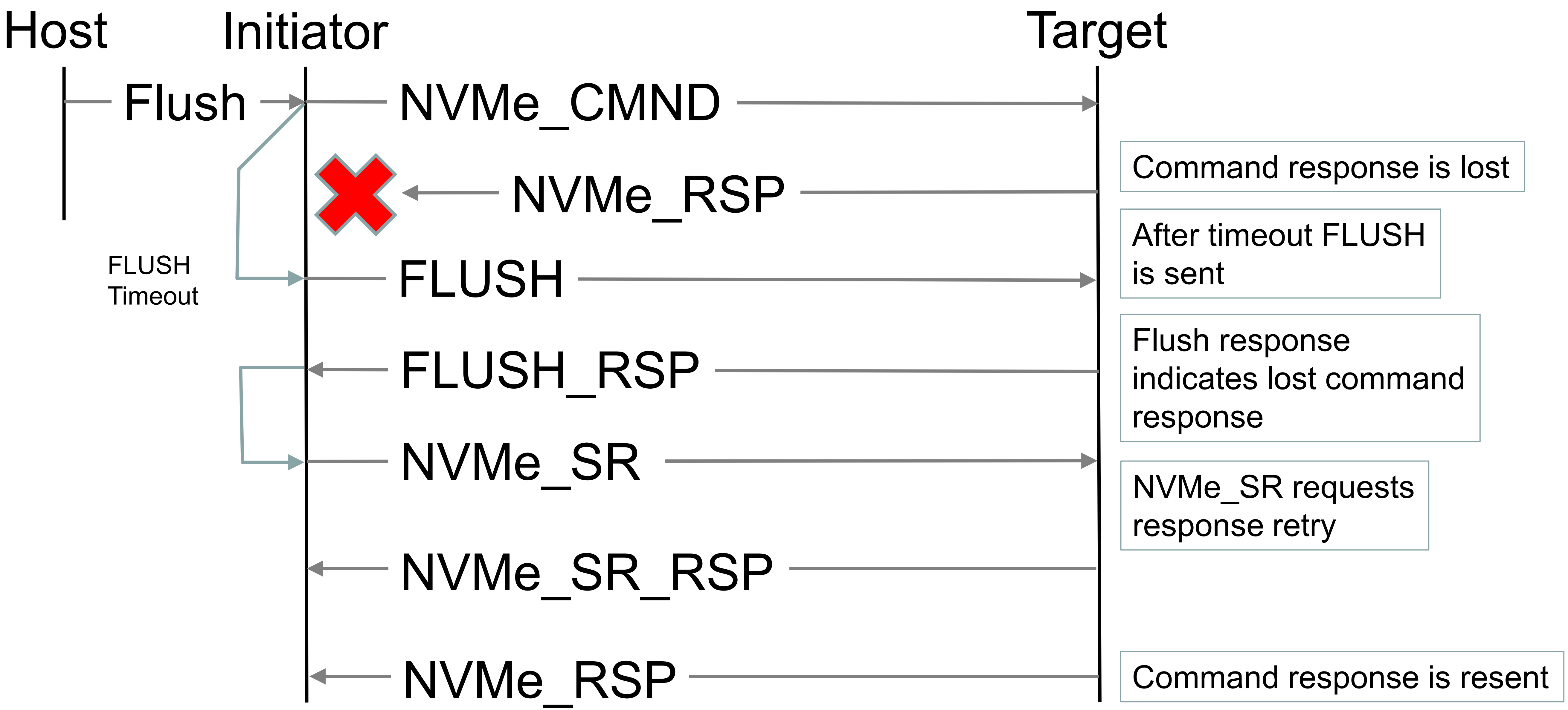
NVMe_SR IU

- New NVMe_SR IU defined
  - (An IU – Information Unit - is part of the basic command structure of FCP)
  - Tells the recipient to retransmit a command or data

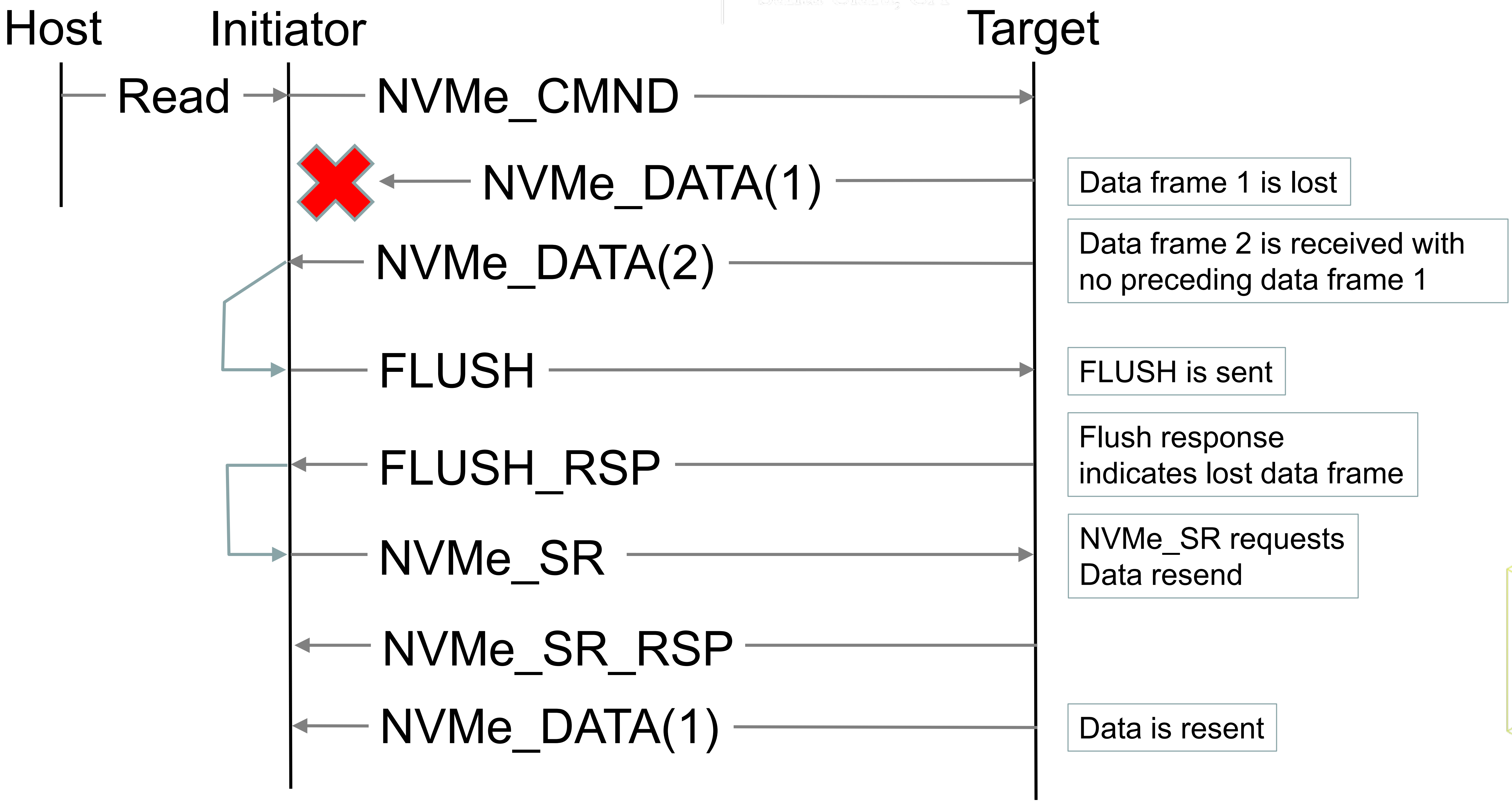
Detailed NVMe Command Loss Example



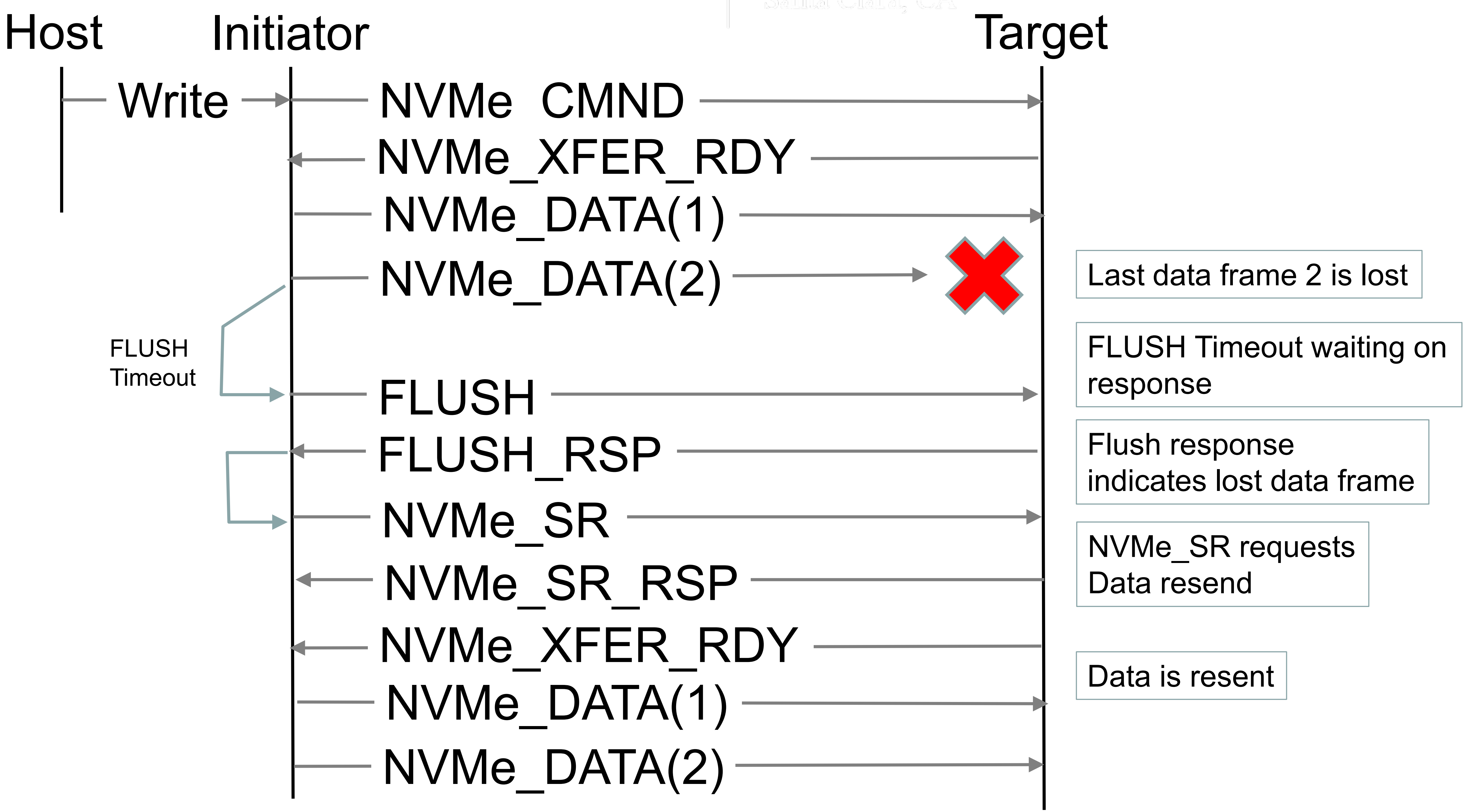
Detailed Lost NVMe Response example



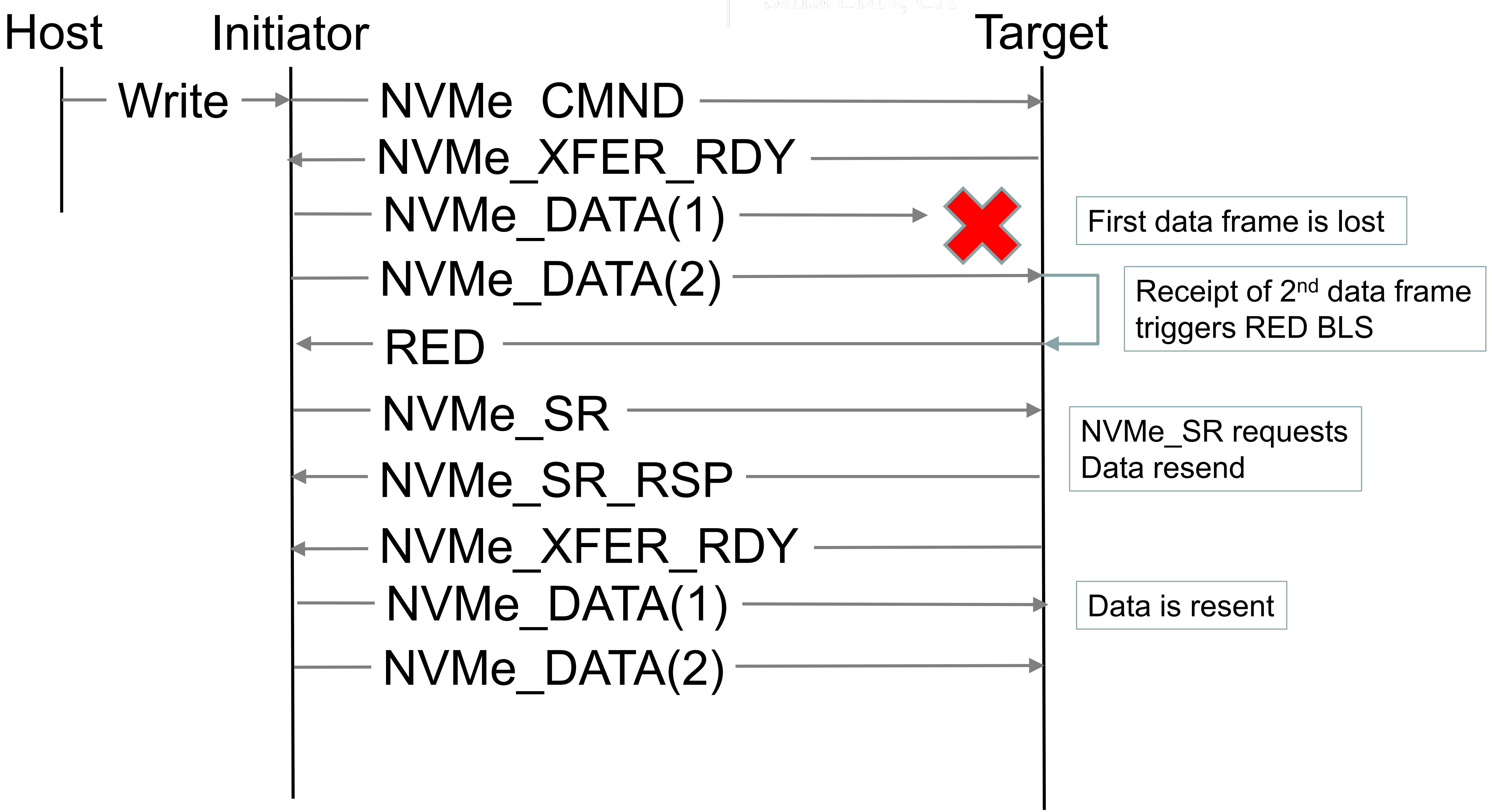
Lost Read Data frame example



Lost Write Data frame example



RED BLS example



Error Recovery Summary

- Goal is to recover from errors without upper level knowing anything happened
  - Recovery in 2 seconds or less
- This is going to be increasingly important as link speeds go up
- Starting with FC-NVMe and applying to SCSI FCP

September 23-26, 2019
Santa Clara, CA

ROADMAP UPDATE

Fibre Channel Physical Standards

- Fibre Channel physical layers defined in FC-PI series of Standards
  - Encoding and Protocol layers defined based on FC-FS series of standards
- Most Recent Standard ratified is FC-PI-7 – 64GFC
- Development has started on FC-PI-8 – 128GFC

Document	Represents…
FC-PI	1GFC 2GFC 4GFC
FC-PI-2	4GFC
FC-PI-4	8GFC
FC-PI-5	16GFC
FC-PI-6 FC-PI-6P	32GFC 128GFC (parallel)
FC-PI-7 FC-PI-7P	64GFC 256GFC (parallel)
FC-PI-8 FC-PI-8P	128GFC 512GFC (parallel)

FCIA Roadmap

Product Naming	Throughput (Mbytes/s)	Line Rate (Gbaud)	T11 Specification Technically Complete (Year)*	Market Availability (Year)*
1GFC	200	1.0625	1996	1997
2GFC	400	2.125	2000	2001
4GFC	800	4.25	2003	2005
8GFC	1,600	8.5	2006	2008
16GFC	3,200	14.025	2009	2011
32GFC	6,400	28.05	2013	2016
128GFC	25,600	4X28.05	2014	2016
64GFC	12,800	28.9 PAM-4 (57.8Gb/s)	2017	2019
256GFC	51,200	4X28.9 PAM-4 (4X57.8Gb/s)	2017	2019
128GFC	25,600	TBD	2020	Market Demand
256GFC	51,200	TBD	2023	Market Demand
512GFC	102,400	TBD	2026	Market Demand
1TFC	204,800	TBD	2029	Market Demand

Signaling Rate Abbreviations

Abbreviation	Signaling rate	Number of Lanes	Data rate
1GFC	1.0625 MBd	1	100 MB/s
2GFC	2.125 MBd	1	200 MB/s
4GFC	4.250 MBd	1	400 MB/s
8GFC	8.500 MBd	1	800 MB/s
16GFC	14.025 MBd	1	1600 MB/s
32GFC	28.050 MBd	1	3200 MB/s
64GFC	28.900 MBd	1	6400 MB/s
128GFC	112.200 MBd	1 or 4	12800 MB/s
256GFC	115.600 MBd	4	25600 MB/s

MB/s = Mega*bytes* per second
MBd = Mega*baud* per second

Just Completed…

- FC-PI-7 (64GFC) ratified mid 2018
- 64GFC had to be backward compatible to 32GFC and 16GFC.
  - Backward compatibility and "plug and play" to utilize existing infrastructure with new speeds is always a must have for FC development.
- Existing cable assemblies must plug into 64GFC capable products
  - LC (connector) and SFP+ (form factor)
- Reach goals
  - 100 meters for multi-mode short reach optical variant using OM4/OM5 cable plants
    - OM4 optical fibre has a higher optical bandwidth than OM3 fibre which leads to longer reach at a given speed.
  - 10KM for single mode optical variant
  - Electrical variant for backplane applications
- 64GFC will double the throughput of 32GFC
- Corrected bit-error-rate (BER) target of 1e-15
  - Advanced bit error recovery achieved through FEC

Forward Error Correction for 64GFC

- Forward Error Correction (FEC) is mandatory for all types of 64GFC links
- Transmitter encodes the data stream in a redundant way using an error correcting code
- 64GFC uses a block code called Reed Solomon.
  - 64GFC uses RS(544,514)
  - Allows correction of single bit errors or burst errors for 15 ten-bit symbols out of 5140 bits sent
- 64GFC uses terms such as uncorrected BER which is the minimum BER to be expected pre-FEC encoding/decoding
  - Uncorrected BER is 1e-04 range or lower
  - FEC-corrected BER is 1e-15 range or lower
  - These numbers help identify the usefulness of FEC in making 64GFC links robust

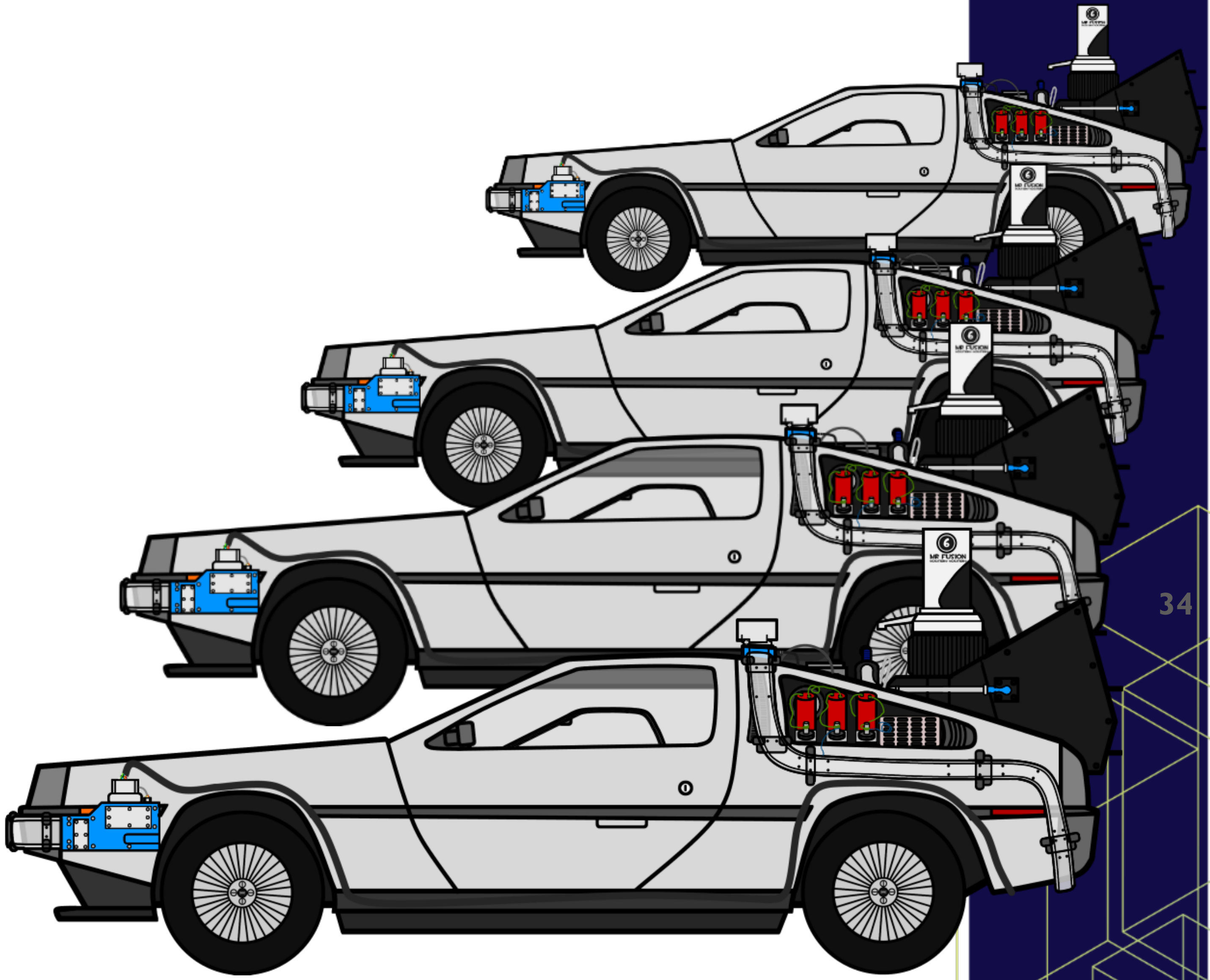
Forward Error Correction (FEC)

A set of algorithms that perform corrections that allow for recovery of one or more bit errors

- SNIA Dictionary

256GFC (Parallel Four Lane)

- FC-PI-7P will describe a four lane 64GFC variant that has a throughput of 256GFC (4x64GFC)
  - Standard currently in development
- Data striped across the four lanes
- MRD requested the following variants
  - 100m on multi-mode cable OM4/OM5
  - 2km single mode variant
- Backward compatibility with 128GFC (4x32GFC) is also a requirement



128GFC FC-PI-8 Planned Requirements

- Backward compatible to 64GFC and 32GFC
- Same external connectors as 32/64GFC
- Existing cable assemblies will work with 128GFC
- Multi-mode cable plant reach is 100 meters on OM4/OM5
- Single mode cable plant reach of 10KM
- 128GFC links should double the throughput in MB/sec of 64GFC links
- Corrected BER target of 1e-15
- Reduce latency of 64GFC by up to 20%
- A four lane parallel 512GFC is also planned

WHY USE FC-NVMe?

Top 6 Reasons FC-NVMe Might Be The Right Choice

1. Leverage existing dedicated Storage Network (SAN)
2. Run NVMe and SCSI Side-by-Side
3. Robust and battle-hardened discovery and name service
4. Zoning and Security
5. Integrated Qualification and Support
6. With FC-NVMe-2 Industry leading error detection/recovery

SUMMARY

FC-NVMe

- **Wicked Fast!**
- **Builds on 20 years of the most robust storage network experience**
- **Can be run side-by-side with existing SCSI-based Fibre Channel storage environments**
- **Inherits all the benefits of Discovery and Name Services from Fibre Channel**
- **Capitalizes on trusted, end-to-end Qualification and Interoperability matrices in the industry**

TABLE OF CONTENTS

Download: https://fibrechannel.org/

More Info

- **FCIA**
  - www.fibrechannel.org

- **Contacts**
  - cwcarlson@marvell.com
  - rupin.mohan@hpe.com

Thank you!

FCIA
FIBRE CHANNEL INDUSTRY ASSOCIATION