SNIA NSF | NETWORKING STORAGE

Notable Updates in NVMe-oF 1.1

Live Webcast

June 30, 2020

10:00 am PT

Today's Presenters

John Kim
NVIDIA

Phil Cayton
Intel

Ilker Cebeli
Samsung

David Peterson
Broadcom

Tim Lustig
NVIDIA

SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced in their entirety without modification
  - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

  NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

SNIA-At-A-Glance

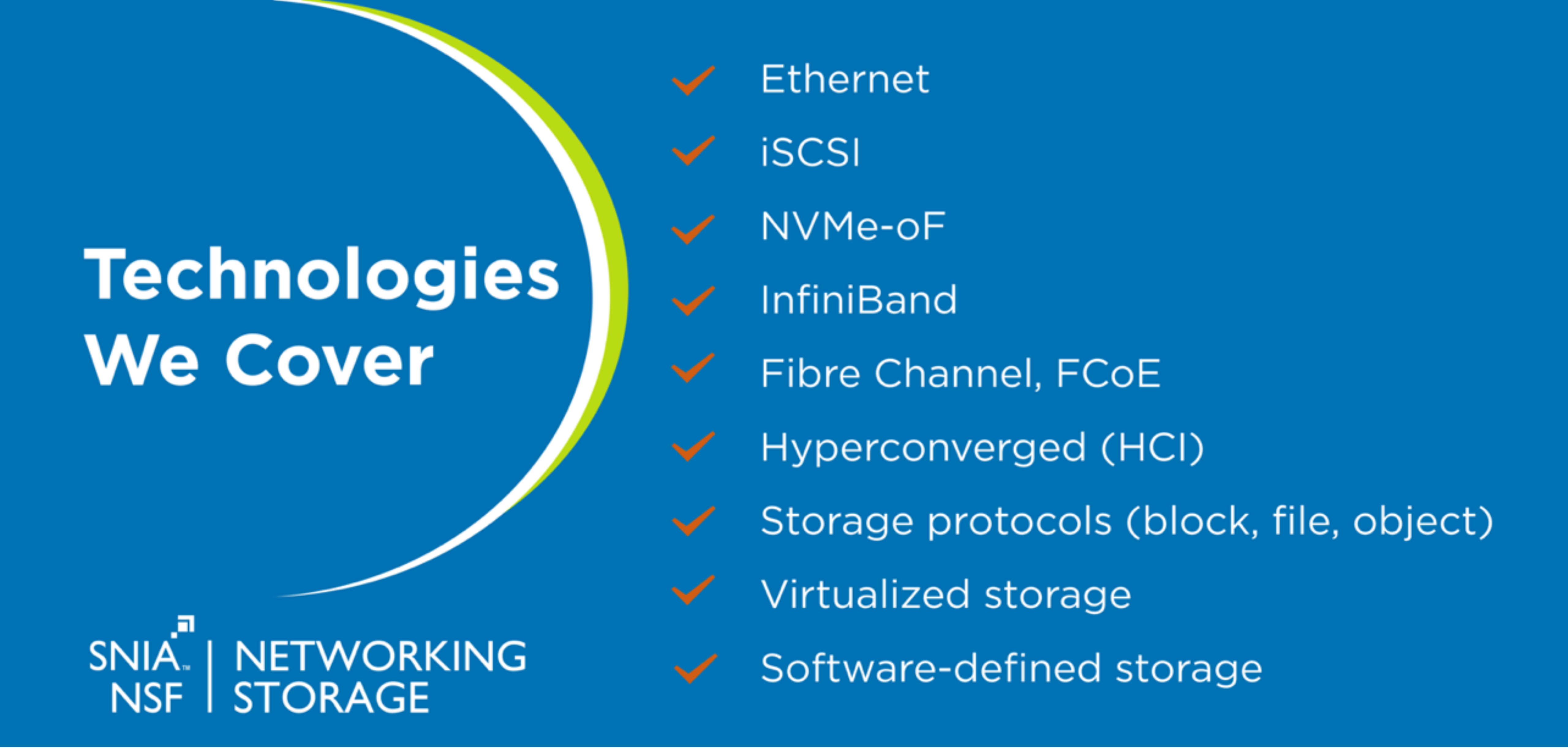
SNIA-at-a-Glance

185
industry leading organizations

2,000
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: **snia.org/technical** **@SNIA**



Technologies We Cover

- ✓ Ethernet
- ✓ iSCSI
- ✓ NVMe-oF
- ✓ InfiniBand
- ✓ Fibre Channel, FCoE
- ✓ Hyperconverged (HCI)
- ✓ Storage protocols (block, file, object)
- ✓ Virtualized storage
- ✓ Software-defined storage

Agenda

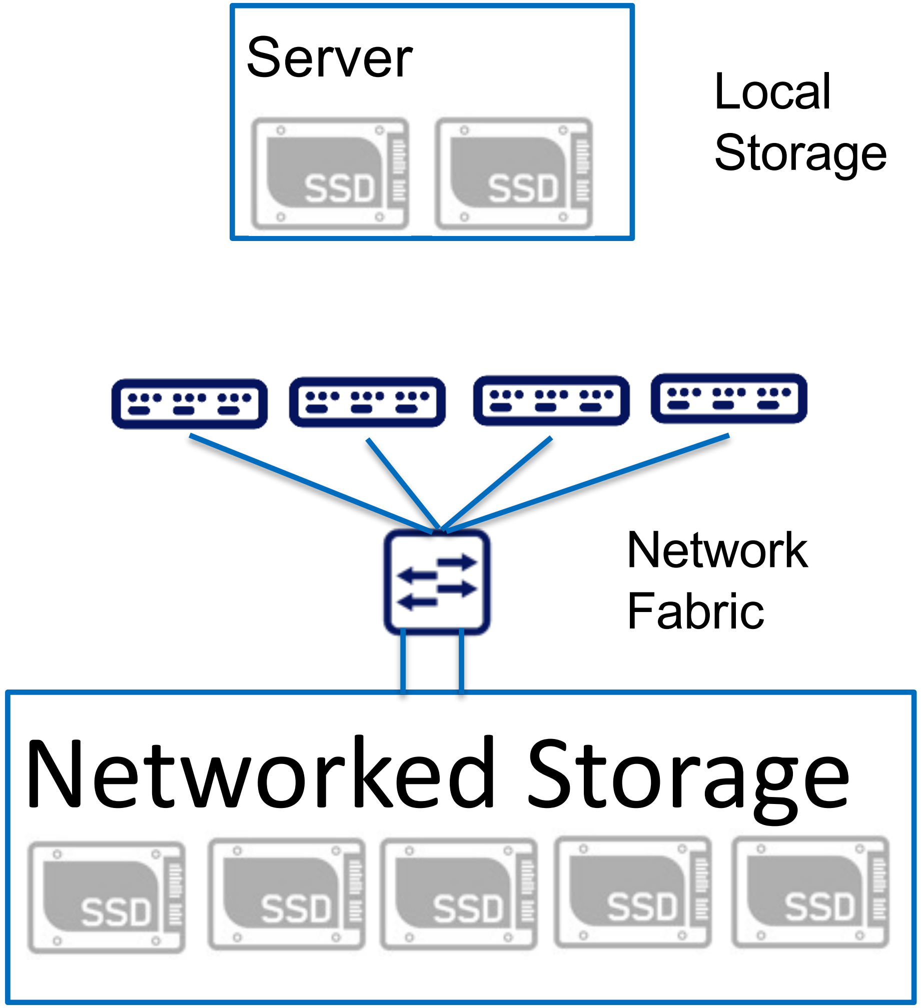
- Brief History of NVMe over Fabrics™
- What's new in NVMe-oF including support for CMB and PMR
- Managing and Provisioning NVMe-oF Devices with Swordfish
- Background and Problem Statement for FC-NVMe-2
- Fibre Channel Transport Connection over NVMe-oF
- Sequence Level Error Recovery for FC
- Current Status of FC-NVMe-2 Standard

NVMe-oF 1.1 Overview

John F. Kim

Why the Need for NVMe over Fabrics™?

- NVMe over PCIe limited to local use
- Constant desire to network storage
  - Sharing, provisioning,
  - Cloud / virtualization / containers
  - Data / workload migration
  - Better efficiency and data protection
- Desire to keep NVMe semantics
  - Lower latency, multi-queue, etc.



Brief History of NVMe over Fabrics

- NVMe Protocol
  - 1.0 in March 2011; 1.1 in October 2012
  - 1.2 in November 2014; 1.3 in May 2017
  - 1.4 in June 2019
- NVMe-oF 1.0 first approved June 2016 (link)
  - 1.0 included RDMA
  - FC-NVMe approved August 2017 (with NVMe-oF 1.0a)
  - NVMe TCP published November 2018
  - NVMe-oF 1.1 October 2019 (following NVMe 1.4)

What’s New in NVMe-oF 1.1

- TCP transport option (TP 8000)
- Multi-pathing improvements
  - Asymmetric Namespace Access (ANA)
  - Domains and Divisions for easier maintenance and upgrades
- Discovery and Transport
  - Persistent Discovery Controller Connections (TP 8002)
  - Fabric I/O Queue Deletion (TP 8001)
  - End-to-end flow control made optional (TP 8005)
  - RDMA controller ID in queue pairs (TP 8008)
- CMB and PMR support in SSDs
- FC-NVMe changes
- Redfish/Swordfish manageability changes

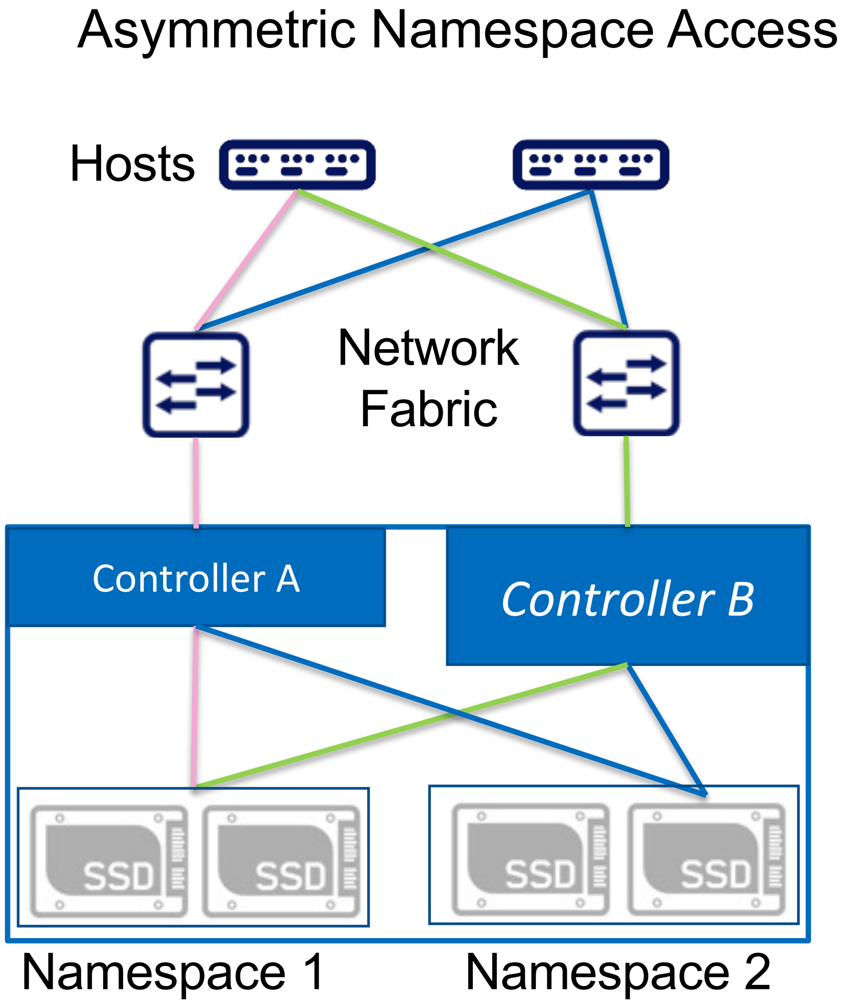
Why Do We Need NVMe™ on TCP?

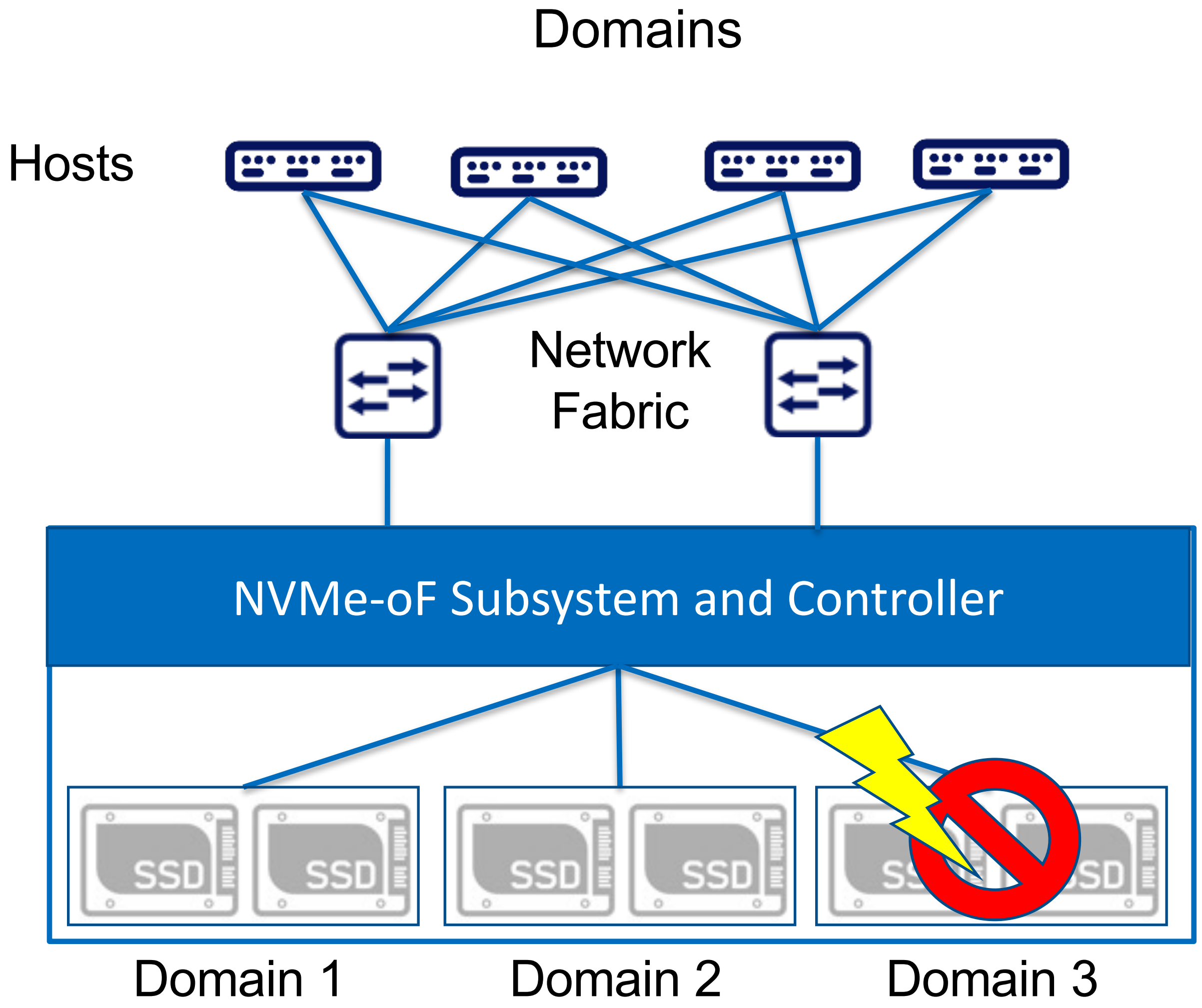
- Not all customers can use RDMA or Fibre Channel
  - PCIe has limitations on distance and scale
  - RDMA requires special hardware for best performance
  - Fibre Channel requires a FC network
- TCP is everywhere
  - Can provide good performance with proper network setup
- See SNIA NSF webcast “What NVMe/TCP Means for Networked Storage” from January 2019
  - https://www.brighttalk.com/webcast/663/344698

NVMe-oF 1.1 Multi-Pathing Improvements

- **NVMe 1.1 added multi-path**
  - Requires multiple controllers
  - Assumed each controller is identical
- **NVMe 1.4 added asymmetric namespace access (ANA)**
  - Controllers can be different
  - Access paths can be given preference, including with NVMe-oF
- **Domains and Divisions**
  - NVMe subsystem can be sub-divided into domains
  - Each domain can be powered, managed, or faulted separately
  - Good for large NVM subsystems, non-disruptive operation

Illustration of ANA and Domains





NVMe-oF 1.1 Discovery and Transport

- **Persistent controller discovery**
  - In NVMe-oF 1.0, discovery is a one-time deal
  - Now can persist discovery controller connections, be notified of changes
- **I/O Queue Deletion (Graceful Disconnect)**
  - Can delete an I/O queue without terminating host-controller association
  - Only for NVMe-oF; Good for large storage systems
- **Other minor changes**
  - **Submission queue flow control** now optional (can do at lower level)
  - **RDMA queue pair** can specify controller ID during creation

CMB, PMR, NVMe-oF

Ilker Cebeli

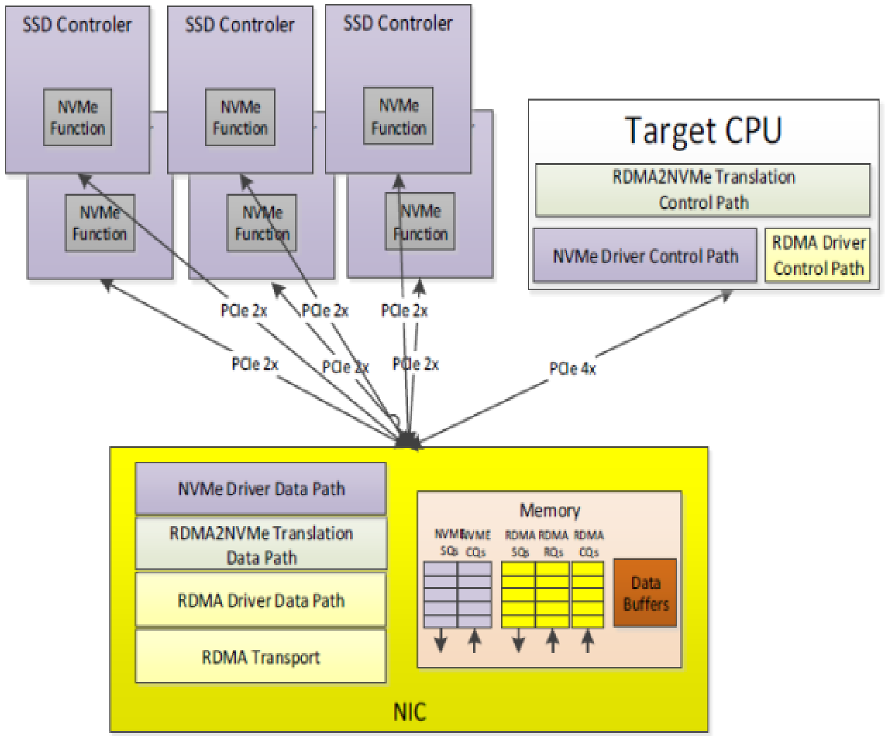
NVMe-oF 1.1 and NVMe 1.4 Changes

- Controller Memory Buffer and Persistent Memory Region Enhancements **(if CMB is supported – mandatory)**
  - Fixes the potential for DMA misrouting with Controller Memory Buffer (CMB) and Persistent Memory Region (PMR).
  - New requirement / incompatible change in section 3.1.1:
    - “Controller Memory Buffer Supported (CMBS): If set to ‘1’, this bit indicates that the controller supports the Controller Memory Buffer, and that addresses supplied by the host are permitted to reference the Controller Memory Buffer only if the host has enabled the Controller Memory Buffer’s controller memory space. If the controller supports the Controller Memory Buffer, this bit shall be set to ‘1’.”

CMB Benefits for NVMe over Fabrics

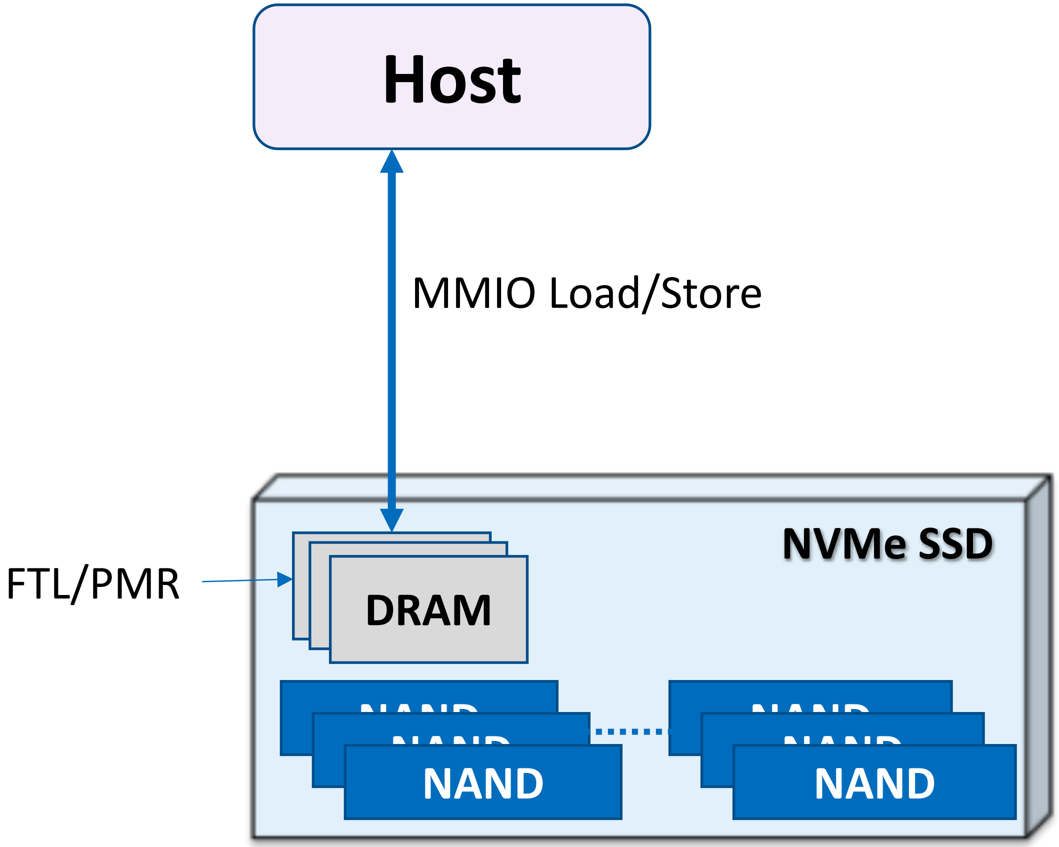
- Latency Improvements
- Disaggregation from CPU PCI Bandwidth
    - Allow utilizing PCI Gen4/Gen5
    - Cost optimizations
- Memory utilization

System Example



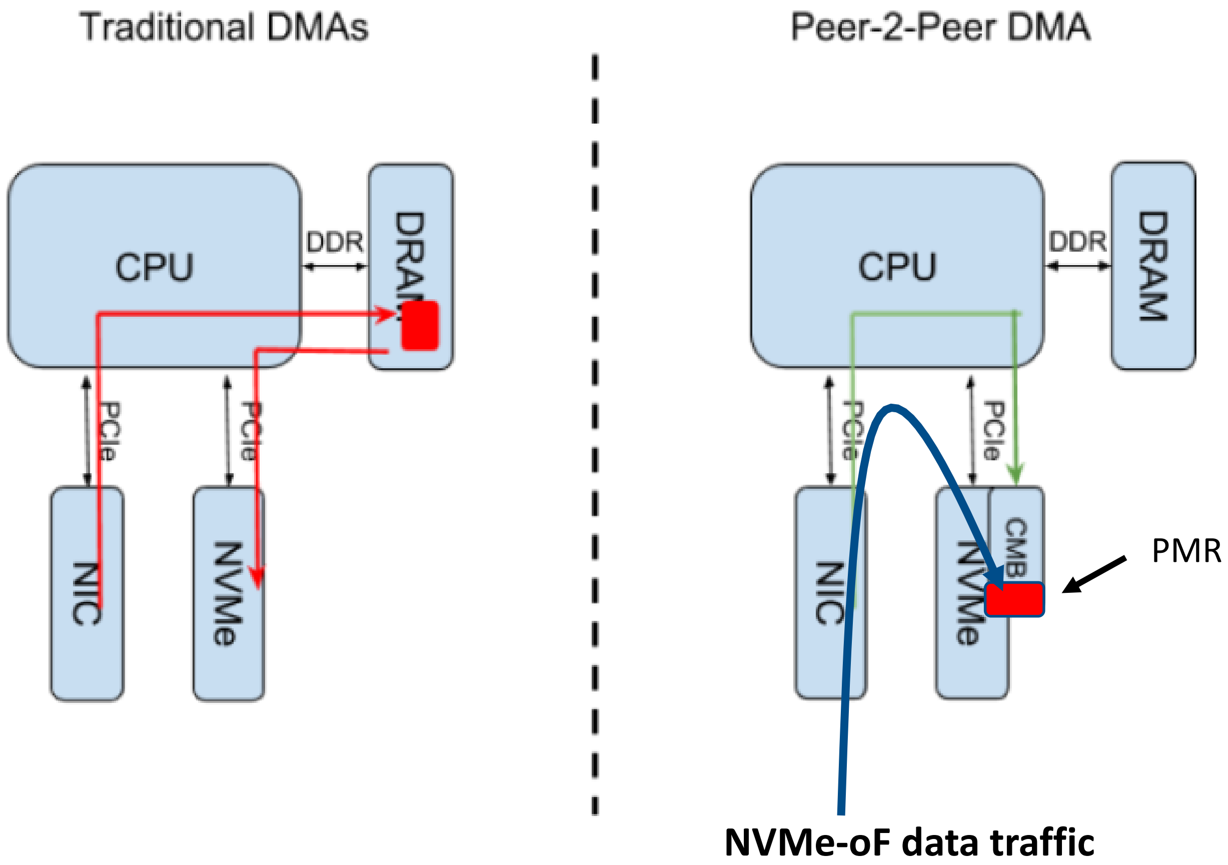
PMR (Persistent Memory Region) in NVMe SSDs

- PMR provide SSD's internal memory to hosts and guarantee persistency
  - Enables Memory Mapped I/O for byte-accessible volume (Load/Store) via PCIe interface
- Power Loss Protection
  - In case of power loss, PMR data saved to NAND
- Persistent Memory Regions add non-volatile CMBs
  - Log for software RAID & erasure coding systems
  - Commit log device for NOSQL databases as well
  - Journal for file systems, Metadata
  - RDMA transactions



Some Use Cases for CMB and NVMe-oF

- Placing some (or all) of your NVMe queues in CMB rather than host memory. Reduce latency
- Using the CMB as a DMA buffer allows for offloaded NVMe copies. This can improve performance and offloads the host CPU
- Using the CMB as a DMA buffer allows RDMA NICs to directly place NVMe-oF data into the NVMe SSD. Reduce latency and CPU load
- Using the PMR as DMA buffer allows RDMA NICs to directly write to Persistent memory region of NVMe SSDs

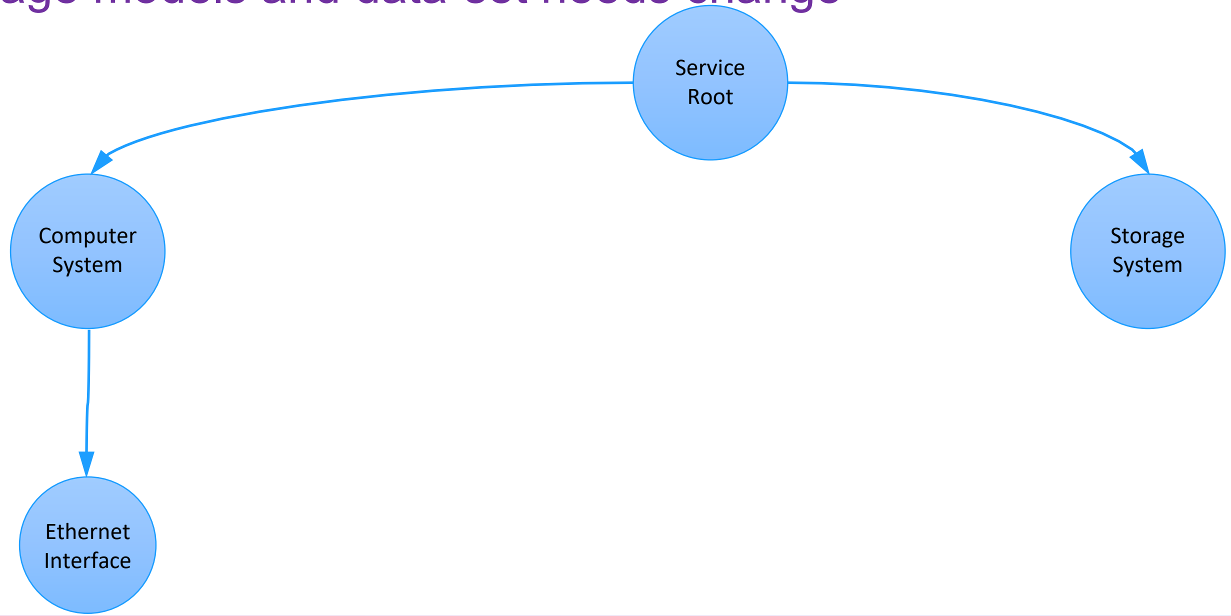


Managing and Provisioning NVMe-oF Devices with Swordfish

Phil Cayton

What’s the Big Deal

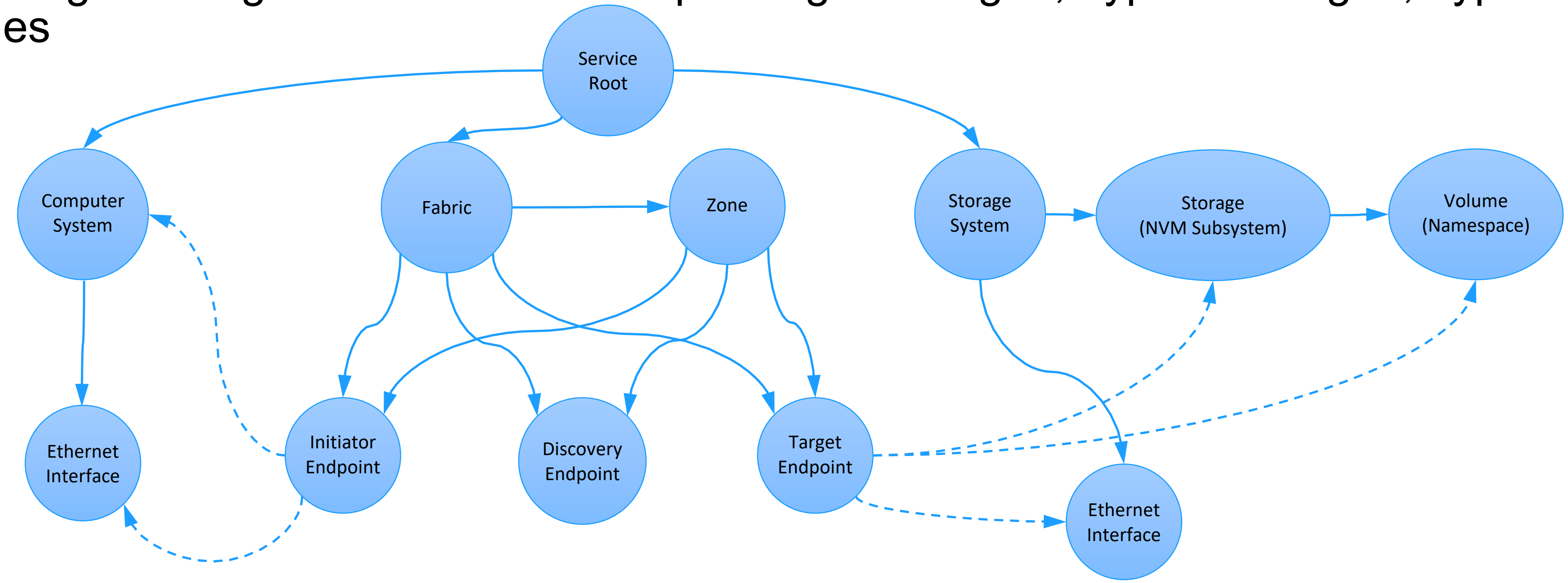
- NVMe/NVMe-oF complex lots of moving parts and components
- Difficult to configure, provision, administrate for even very simple configurations
- When you scale to large, complex, dynamic installations administration becomes even more difficult
  - Where components come and go and change configuration
  - Where your network infrastructure and paths change
  - Where usage models and data-set needs change



More than Meets the Eye

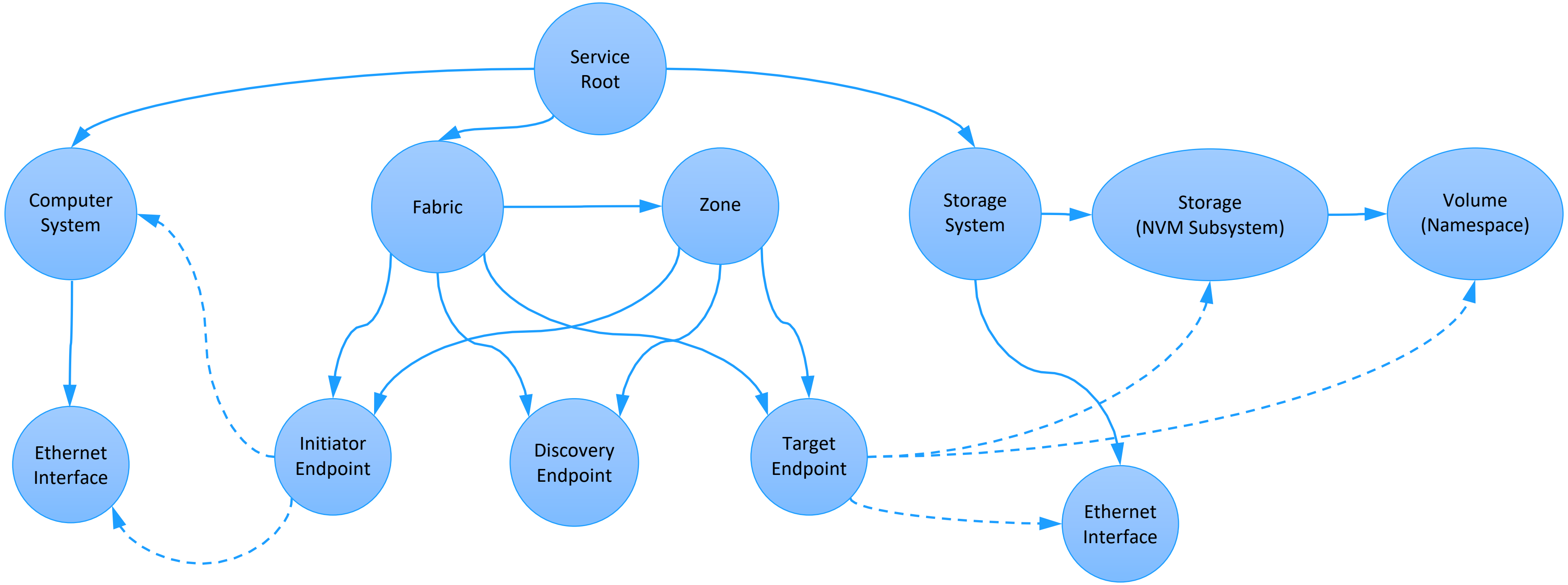
- NVMe/NVMe-oF complex lots of moving parts and components
- Difficult to configure, provision, administrate for even very simple configurations
- When you scale to large, complex, dynamic installations administration becomes even more difficult

- Need a storage management model encompassing converged, hyperconverged, hyperscale, cloud, etc., usages



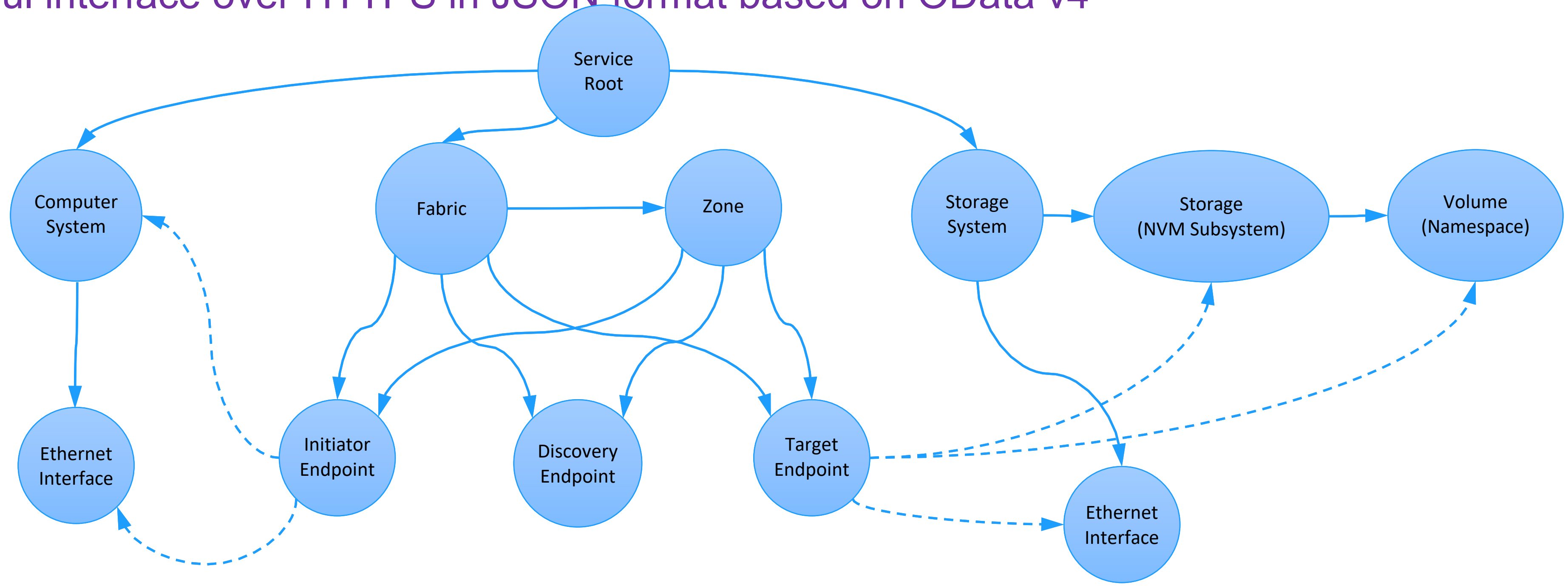
How Do We Solve This?

- Enable management of NVMe/NVMe-oF devices in a large scale environment
- Extend traditional storage domain coverage to converged environments (servers, storage, fabric)
- Develop a common management model:
  - from the point-of-view of what a client wants to accomplish
  - only provide information a client needs



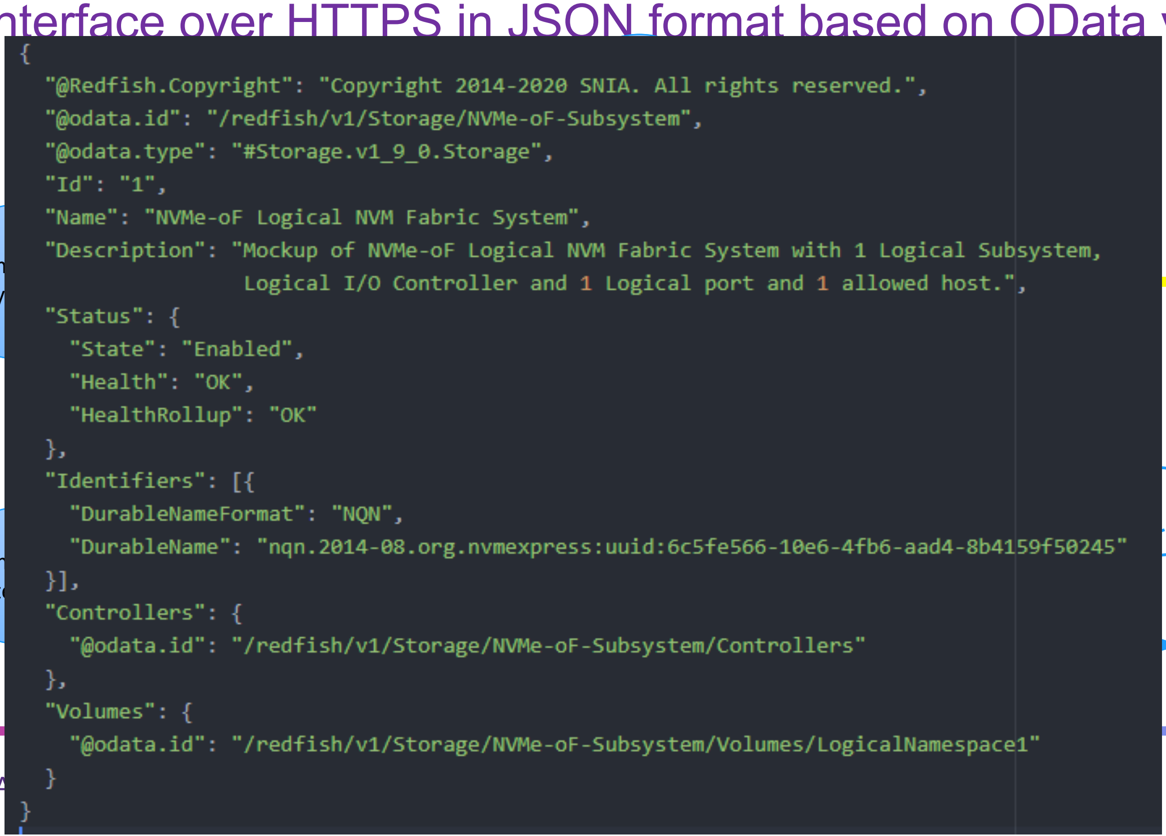
The DMTF Redfish^{TM} + SNIA Swordfish^{TM} Approach

- Swordfish specification defines logical fabric management model that can be used for NVMe and NVMe over Fabrics management
  - Single model allows management of various NVMe over Fabrics types: RDMA, TCP, Fibre Channel
  - Same model can be used for management of all fabric connected system and services
- Swordfish API is a seamless ***extension*** of the Redfish API
  - RESTful interface over HTTPS in JSON format based on OData v4

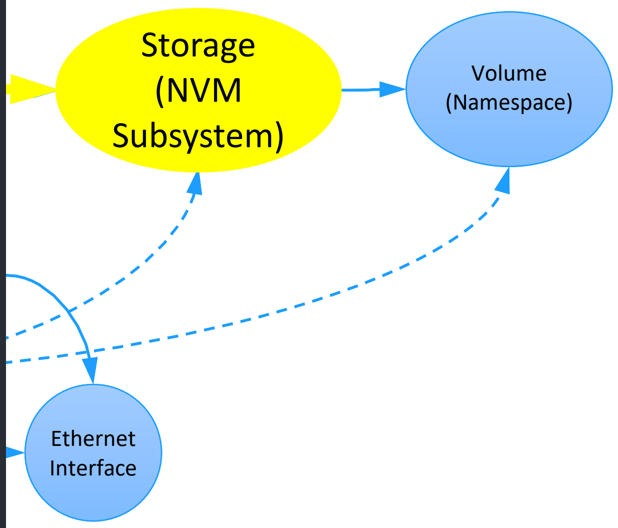


The DMTF Redfish™ + SNIA Swordfish™ Approach

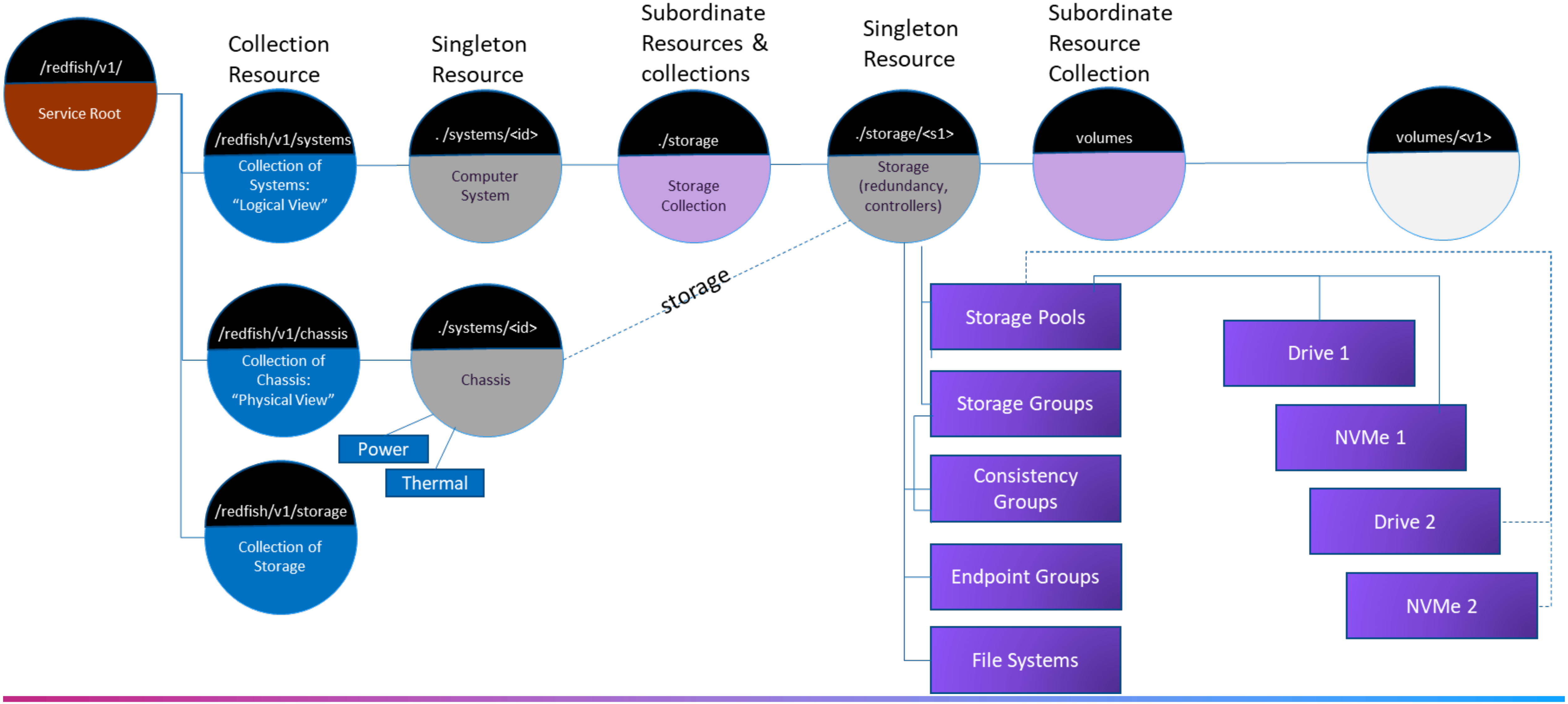
- Swordfish specification defines logical fabric management model that can be used for NVMe and NVMe over Fabrics management
  - Single model allows management of various NVMe over Fabrics types: RDMA, TCP, Fibre Channel
  - Same model can be used for management of all fabric connected system and services
- Swordfish API is a seamless ***extension*** of the Redfish API
  - RESTful interface over HTTPS in JSON format based on OData v4



```
{
  "@Redfish.Copyright": "Copyright 2014-2020 SNIA. All rights reserved.",
  "@odata.id": "/redfish/v1/Storage/NVMe-oF-Subsystem",
  "@odata.type": "#Storage.v1_9_0.Storage",
  "Id": "1",
  "Name": "NVMe-oF Logical NVM Fabric System",
  "Description": "Mockup of NVMe-oF Logical NVM Fabric System with 1 Logical Subsystem,
                 Logical I/O Controller and 1 Logical port and 1 allowed host.",
  "Status": {
    "State": "Enabled",
    "Health": "OK",
    "HealthRollup": "OK"
  },
  "Identifiers": [{
    "DurableNameFormat": "NQN",
    "DurableName": "nqn.2014-08.org.nvmexpress:uuid:6c5fe566-10e6-4fb6-aad4-8b4159f50245"
  }],
  "Controllers": {
    "@odata.id": "/redfish/v1/Storage/NVMe-oF-Subsystem/Controllers"
  },
  "Volumes": {
    "@odata.id": "/redfish/v1/Storage/NVMe-oF-Subsystem/Volumes/LogicalNamespace1"
  }
}
```

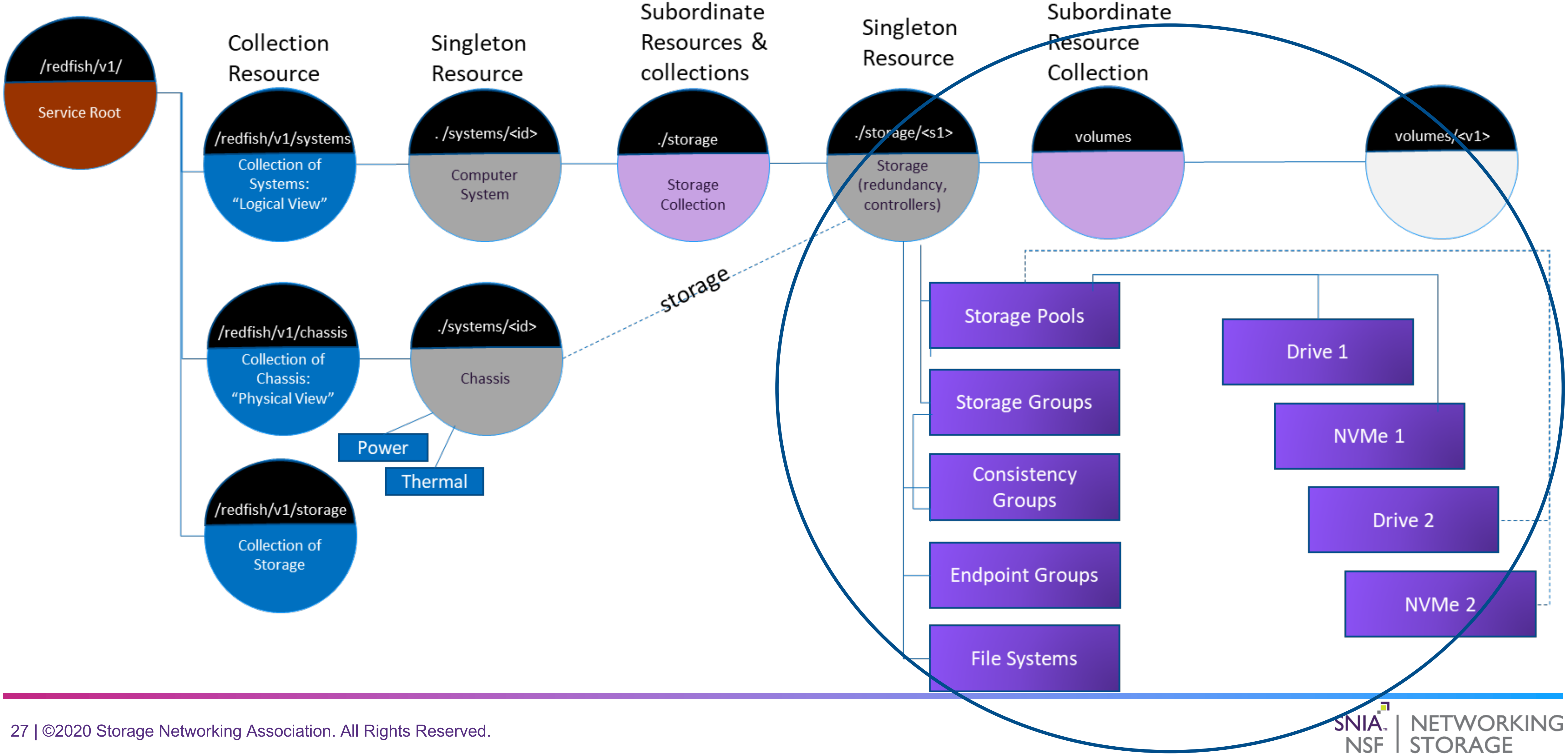


Managing with Swordfish: Extend Redfish Local Storage



SNIA NSF | NETWORKING STORAGE

Managing with Swordfish: Extend Redfish Local Storage

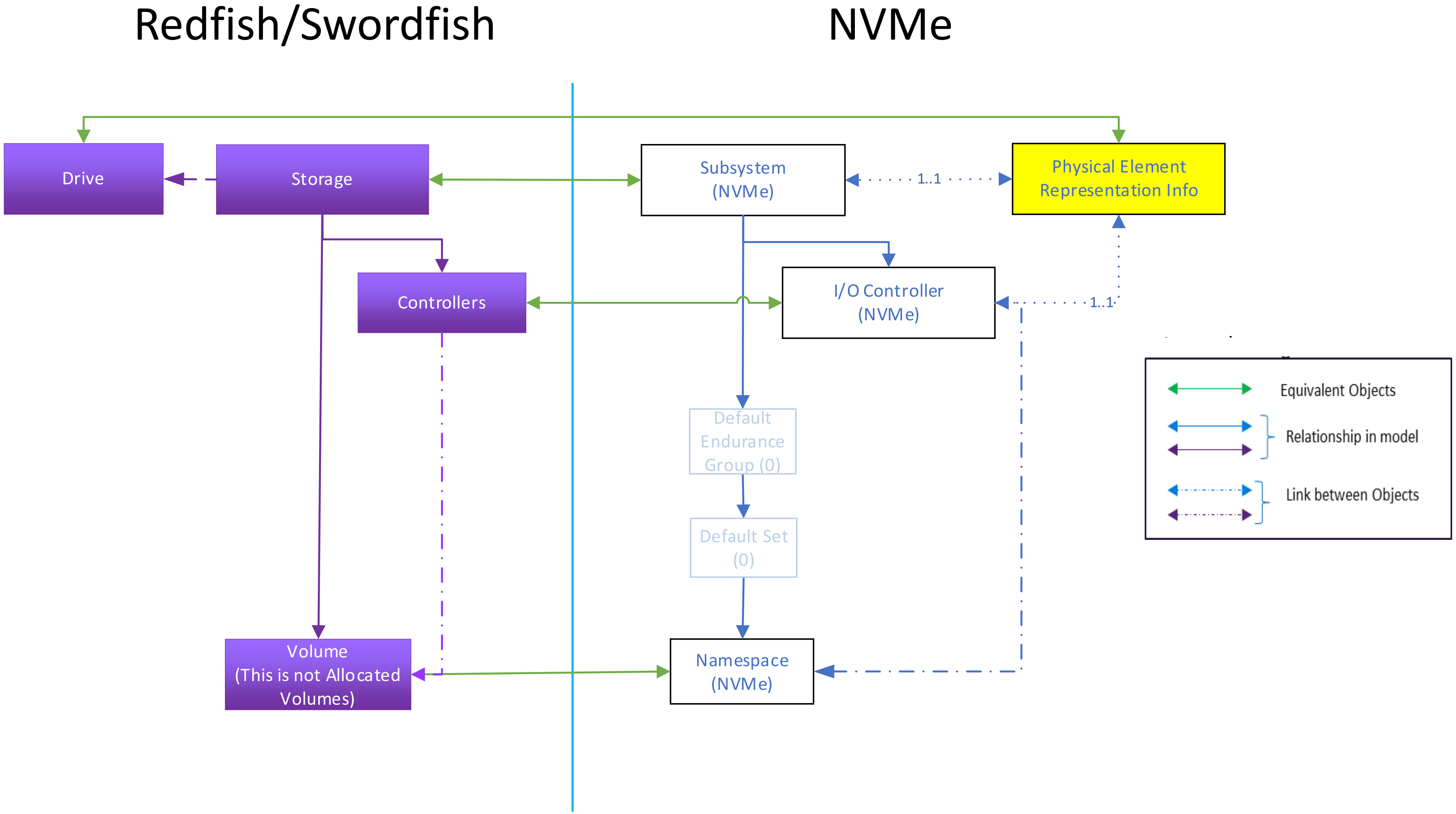


SNIA NSF | NETWORKING STORAGE

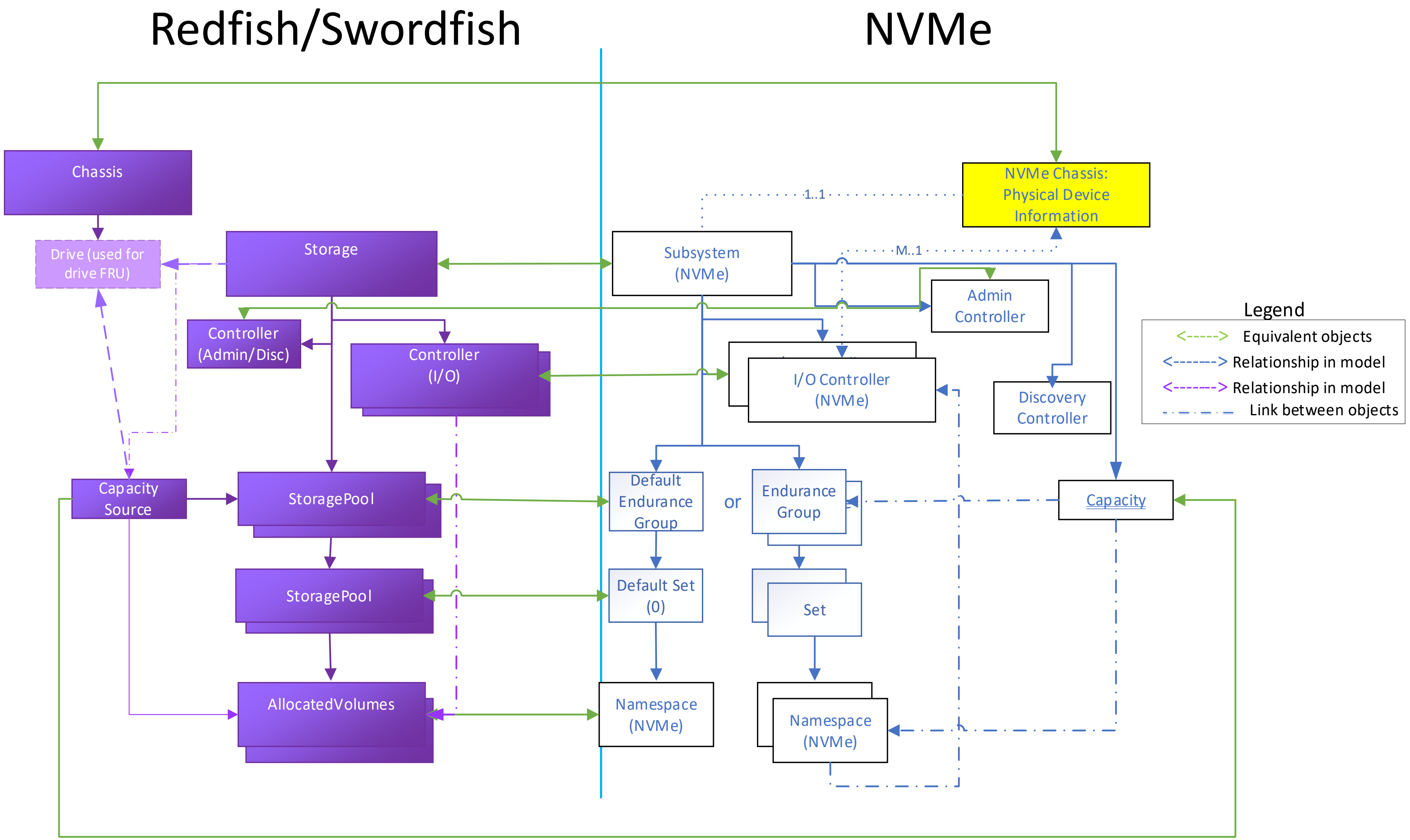
Diving into NVMe/NVMe-oF through Redfish/Swordfish

- Native NVMe/NVMe-oF management models do not provide a datacenter-level view to enable scale-out management
- Need to integrate NVMe/NVMe-oF management into environment already covered by Redfish/Swordfish
- Mapping exercise to manage NVMe/NVMe-oF through Redfish/Swordfish

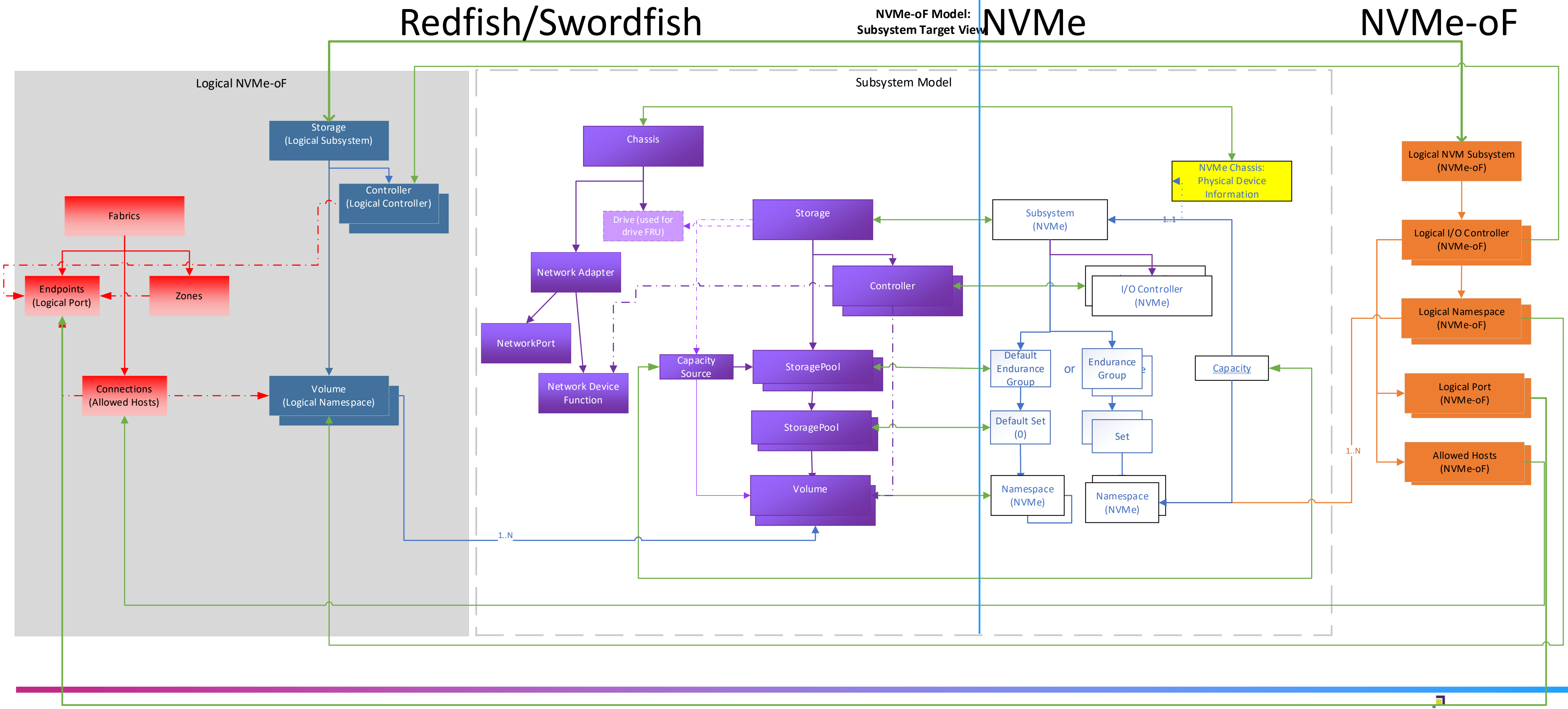
Managing NVMe using Swordfish: A Simple SSD



NVMe Subsystem Model: A More Complex SSD



Managing NVMe-oF using Swordfish: Subsystem Model



Status

- **Released Model Overview June 2020**
  - Contains mockups of use cases, schema with new NVMe specific properties, models for NVMe, NVMe-oF, ..
  - Resources:
    - https://www.snia.org/forums/smi/swordfish
    - http://swordfishmockups.com/
- **Next Steps:**
  - Develop detailed model overview and mapping document
    - Property mapping between RF/SF and NVMe/NVMe-oF specifications
    - Usage guidelines where appropriate
    - Being developed in SNIA Scalable Storage Management TWIG – in alignment with NVMe/DMTF
  - Develop RF/SF Profiles
    - for specific configurations with required / optional schema and properties for implementations
  - Conformance Test Program

FC-NVMe-2

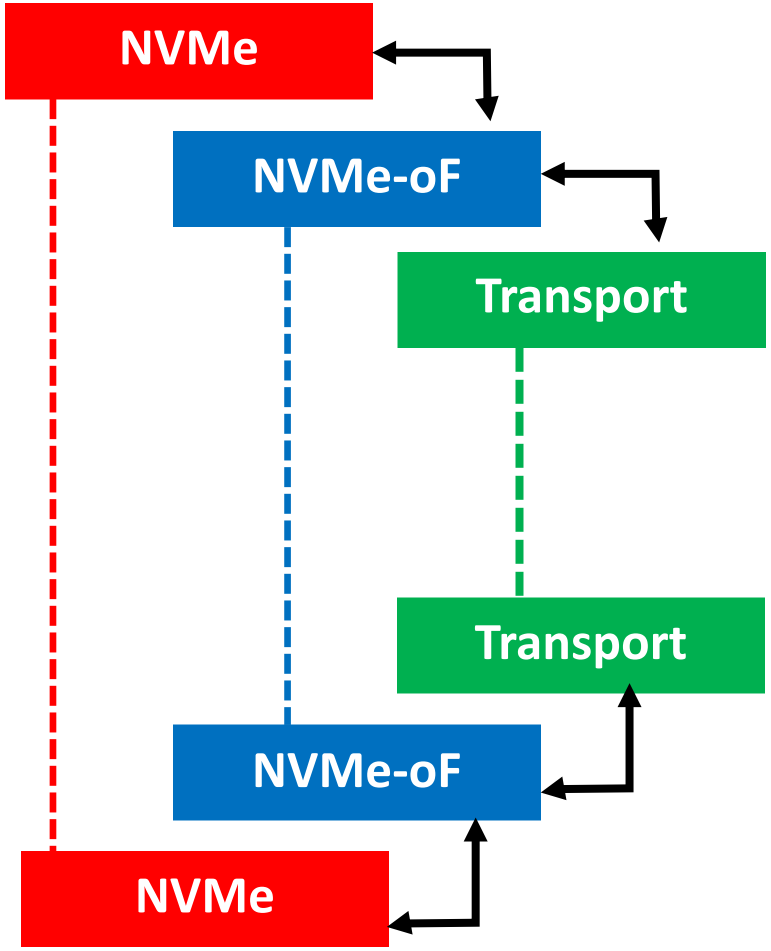
Sequence Level Error Recovery (SLER)

David Peterson

Agenda

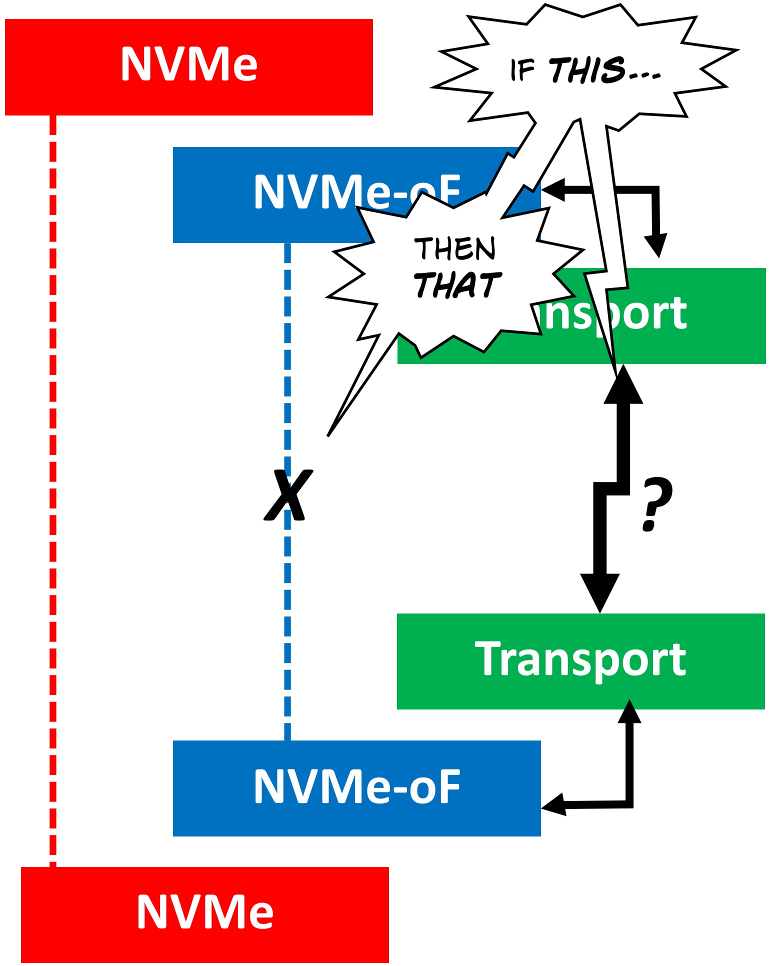
- Background
- The problem(s)…
- The solution ☺
- Current status

Before We Begin…



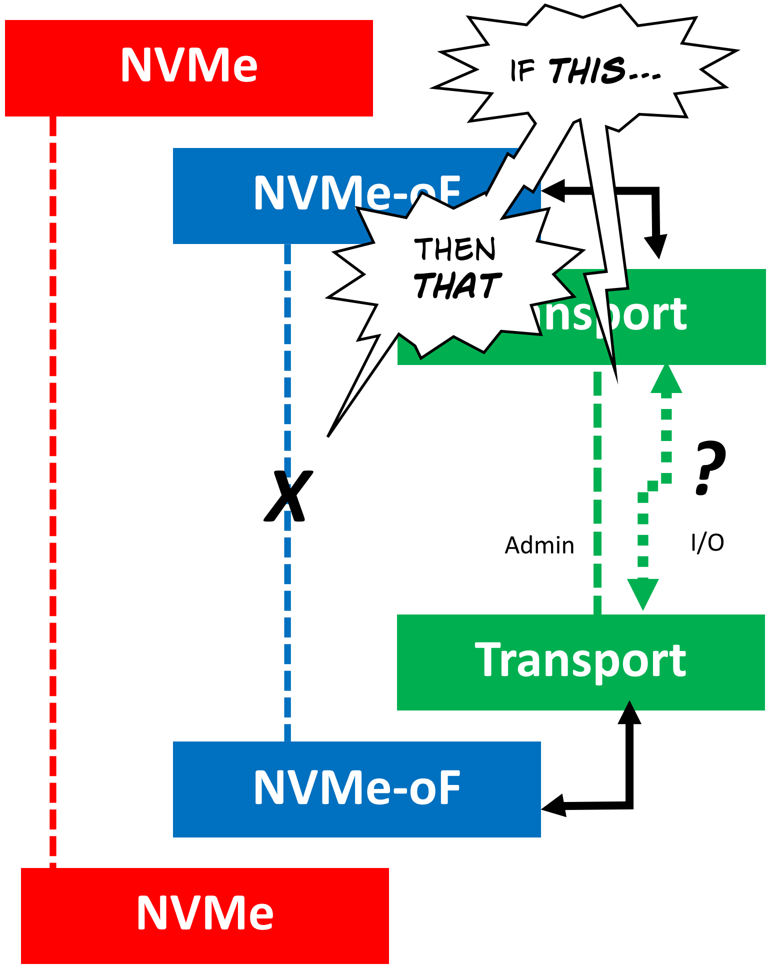
- NVMe relationships exist between a host and a target (the red line)
- ***Associations and I/O connections*** exist between NVMe-oF endpoints (the blue line)
- The transport provides the actual link between a host and a target (the green line)
- All the stars need to align for NVMe-oF to work!

Background



- Initial NVM Express over Fabrics spec(s) require an **association** to be terminated if a transport connection is lost
  - Admin Queue or I/O Queue
  - ☹ i.e., big hammer

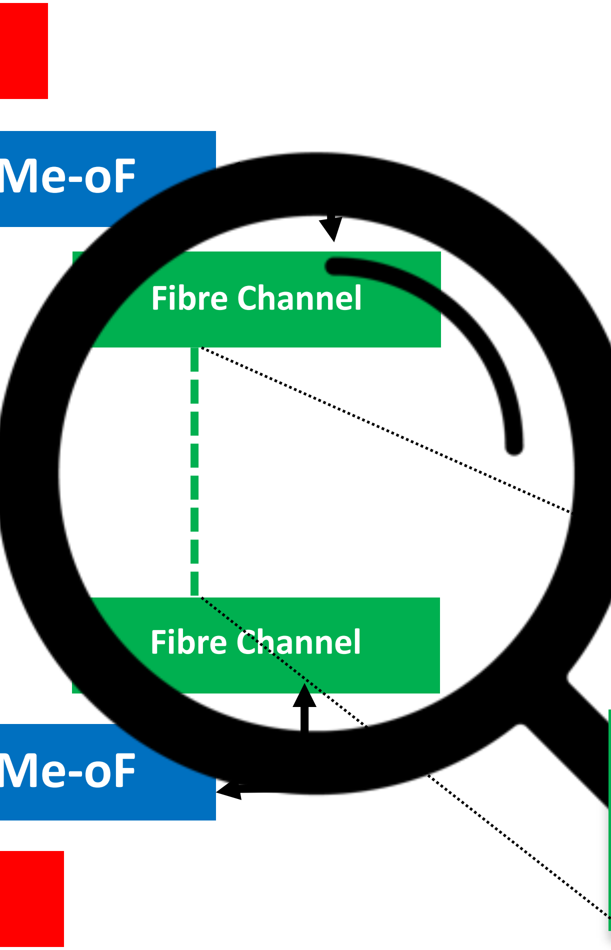
Background

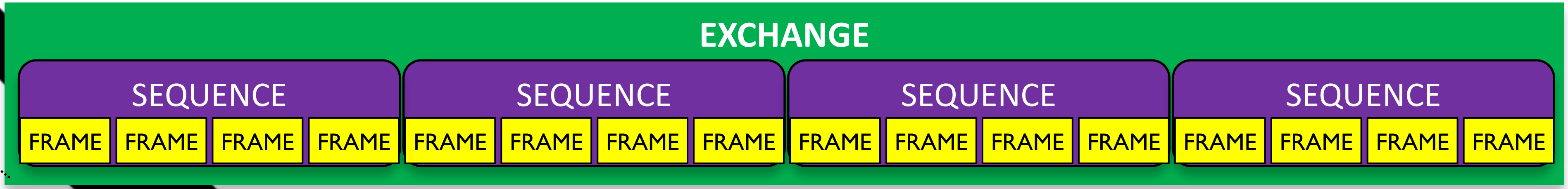


- NVMe-oF 1.1
  - More granular error recovery
  - Ability to disconnect single I/O connection was added
  - This allows the NVMe-oF association to remain in place if an error occurs on an associated I/O connection

What Makes a Fibre Channel Transport Connection?

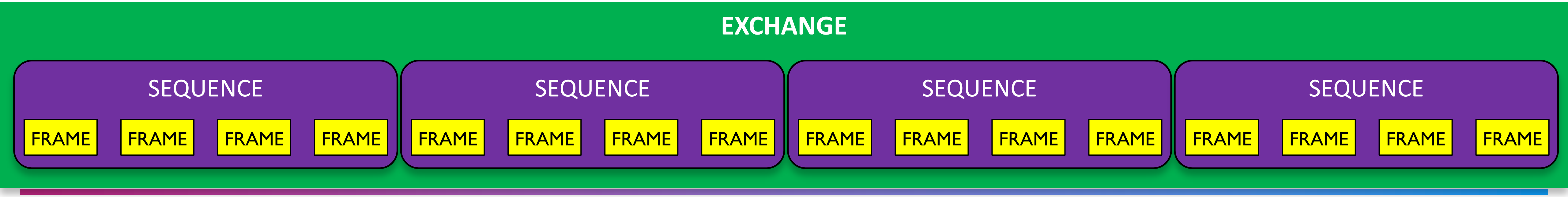
- In Fibre Channel, data transfer has 3 fundamental constructs
  - Frames – A “packet” of data
  - Sequences – A set of frames for larger data transfers
  - Exchanges – A set of Sequences for a single operation e.g., a command, data, and response





What’s So Special About An Exchange?

- An interaction between two Fibre Channel ports is termed an “Exchange”
  - Many protocols (including SCSI based Fibre Channel 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 with Exchange-based routing
    - This allows the Fabric to make efficient use of multiple paths between individual Fabric switches



EXCHANGE

SEQUENCE

FRAME FRAME FRAME FRAME

SEQUENCE

FRAME FRAME FRAME FRAME

SEQUENCE

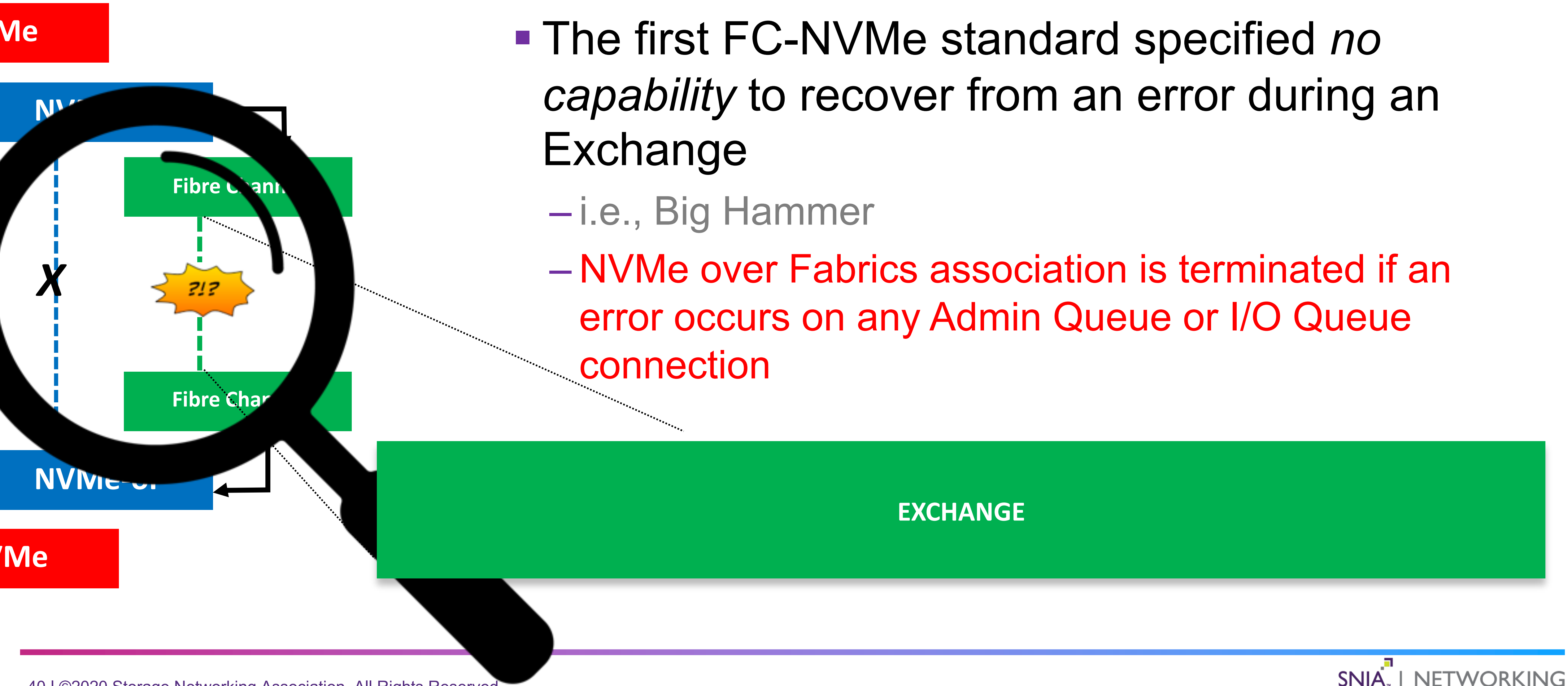
FRAME FRAME FRAME FRAME

SEQUENCE

FRAME FRAME FRAME FRAME

How does Fibre Channel Do It?

- The first FC-NVMe standard specified *no capability* to recover from an error during an Exchange
  - i.e., Big Hammer
  - NVMe over Fabrics association is terminated if an error occurs on any Admin Queue or I/O Queue connection

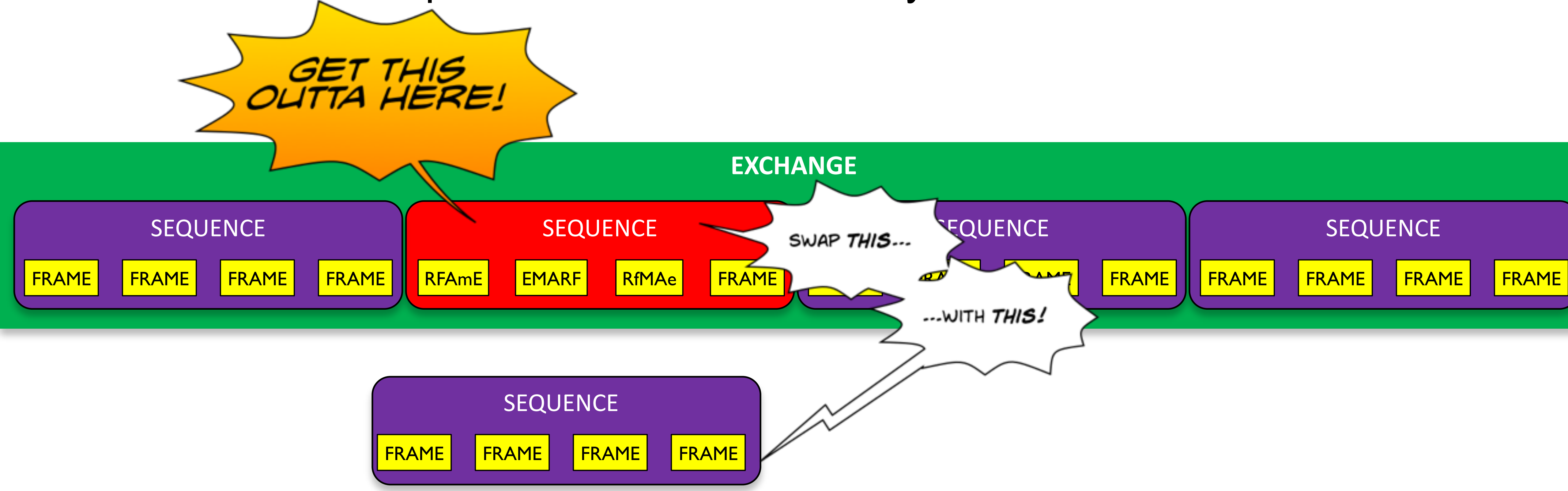


Houston, We Have a Problem…

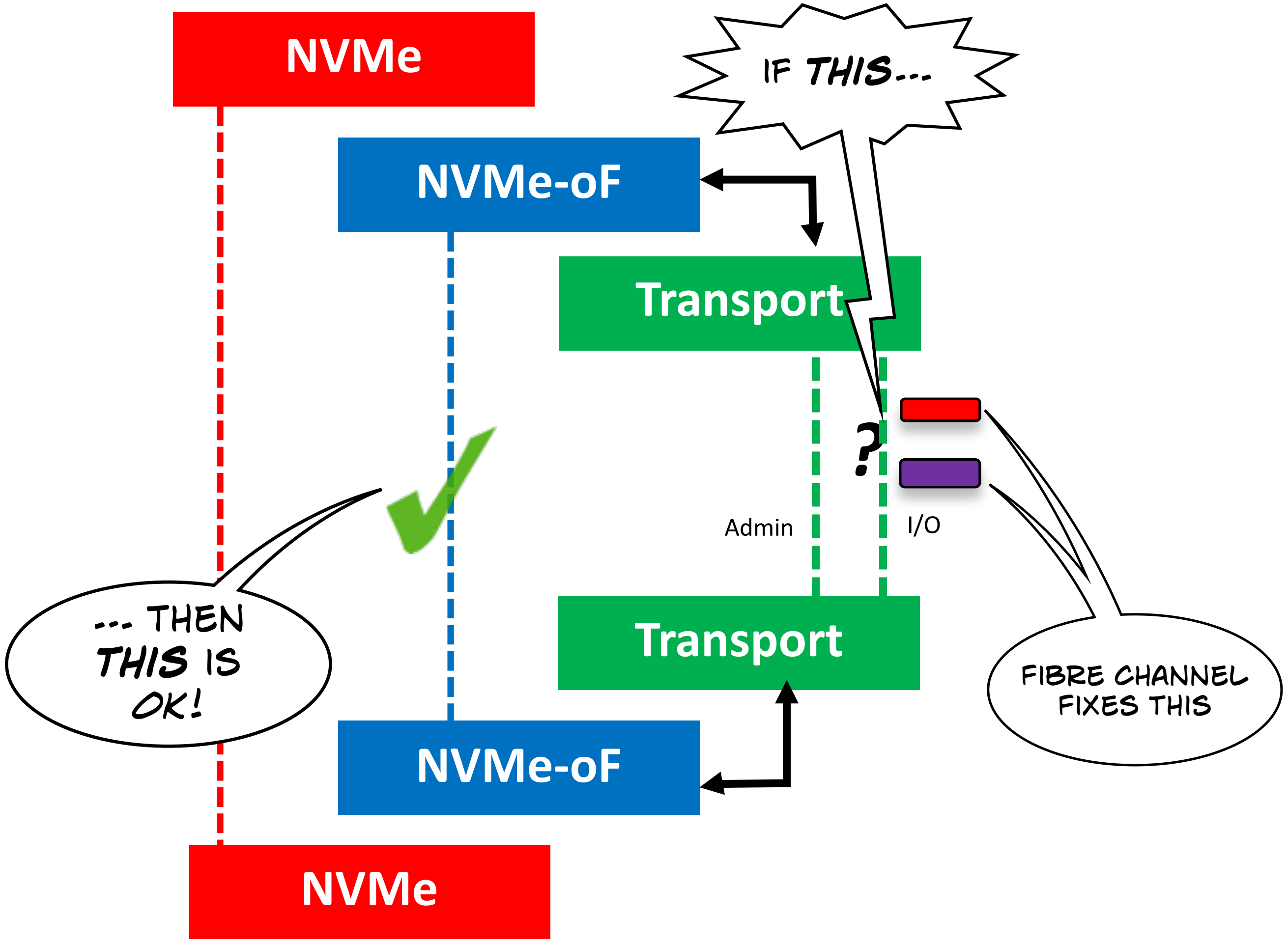
- Big hammer error recovery approach does not work well for most FC deployments
- Exchanges may be delivered out of order with Exchange-based routing
- Thus new functionality was needed…

The Solution for FC

- Perform all steps needed to recover from error(s) within the Exchange ☺
- Thus SLER – Sequence level error recovery



What This Means for NVMe-oF



- No more big hammer ☺
  - Can hit a little nail with a little hammer
- Can recover errors at a much smaller level i.e., within the Exchange
- Before: Even resilient multiple paths couldn't prevent the association from going down
- Now: Can correct errors at the Sequence level, keeping multiple paths resilient, and the NVMe-oF association up and running!

Current Status

- The ability to seamlessly recover from a (Fibre Channel) transport level error in an Exchange-based routing environment has been standardized for use in NVMe over Fabrics environments ☺
- FC-NVMe-2 Standard is Published and available via INCITS
- Draft Standard document is posted on INCITS T10 website
  - Document is called FCP-5 rev 01
  - Sequence level error recovery (SLER) functionality being added as optional behavior

Summary

- Reviewed the basics of NVMe over Fabrics
- Understanding of how CMB and PMR Roles
- Swordfish/Redfish for Managing NVMe-oF Devices
- Introduction to FC-NVMe-2
- Fibre Channel Transport Connection over NVMe-oF
- Sequence Level Error Recovery
- Current Status of FC-NVMe-2 Specification

After this Webcast

- Please rate this webcast and provide us with your feedback
- This webcast and a copy of the slides will be available at the SNIA Educational Library https://www.snia.org/educational-library
- A Q&A from this webcast, including answers to questions we couldn’t get to today, will be posted on our blog at https://sniansfblog.org/
- Follow us on Twitter @SNIANSF