OPEN POSSIBILITIES.

Managing Ethernet-Attached Drives using Swordfish

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

Managing Ethernet-Attached Drives Using Swordfish

Mark Carlson
Principal Engineer, KIOXIA

OPEN POSSIBILITIES.

OPEN PLATINUM™

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

The Evolution of Storage Networks

- Direct attached storage: Single host owns storage

- Storage area networks: Multiple hosts share storage
  - Avoid "silos" of storage and enables storage efficiencies
  - Examples include Fibre Channel & iSCSI storage networks
  - But require “storage controllers" to front storage

- Hyperscale: DAS storage on commodity systems
  - Special software manages many hyperscale nodes in a solution

- Industry moving to NVMe® technology

- Emergence of NVMe-oF™ technologies enables emergence of Ethernet as fabric for network based NVMe storage systems, but 'last foot' is still PCIe.

- Now, systems AND devices on native Ethernet as a storage network

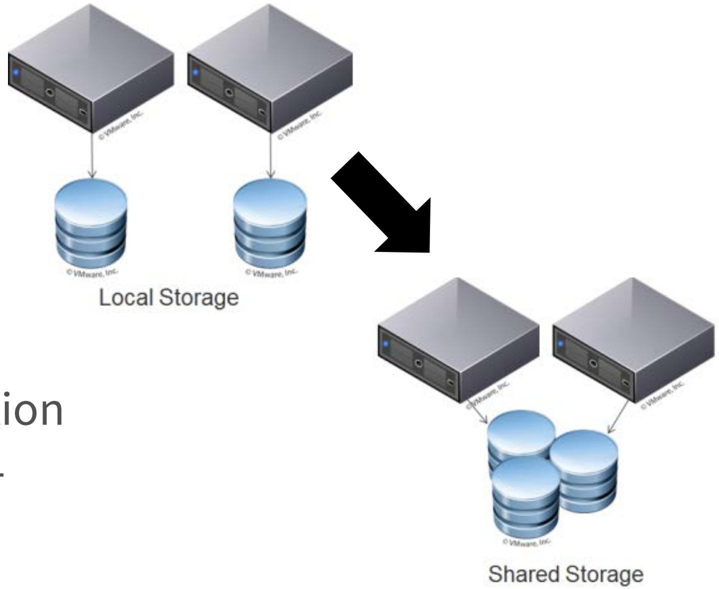
Ethernet as a Storage Network

- Initially, just a transport
- End points performed all the storage services (iSCSI)
- Use of Ethernet matured: Specialized protocols
  - Key/value protocol to access data in mainframe context
  - Object protocol to access massive amounts of unstructured data

- Now, NVMe® over Ethernet: Storage in a queuing paradigm
- High performance / low latency / few or no processing blockages
- No longer gated by transaction paradigm (wait for ACK)
- Next step, NVMe over Ethernet to the drive
- Removes “storage controller" processing bottleneck

NVMe® over Fabrics (NVMe-oF™) Technology

- Sharing NVMe based storage across a network
  - Better utilization: capacity, rack space, power
  - Better scalability: management, fault isolation
- NVMe-oF standard at NVMexpress.org
  - 50+ contributors
  - Version 1.0 released in 2016; NVMe-oF specification merged with the NVMe 2.0 specifications in 2021
  - Fabrics: Ethernet, InfiniBand, Fibre Channel
- Products now in the market from most major storage system vendors

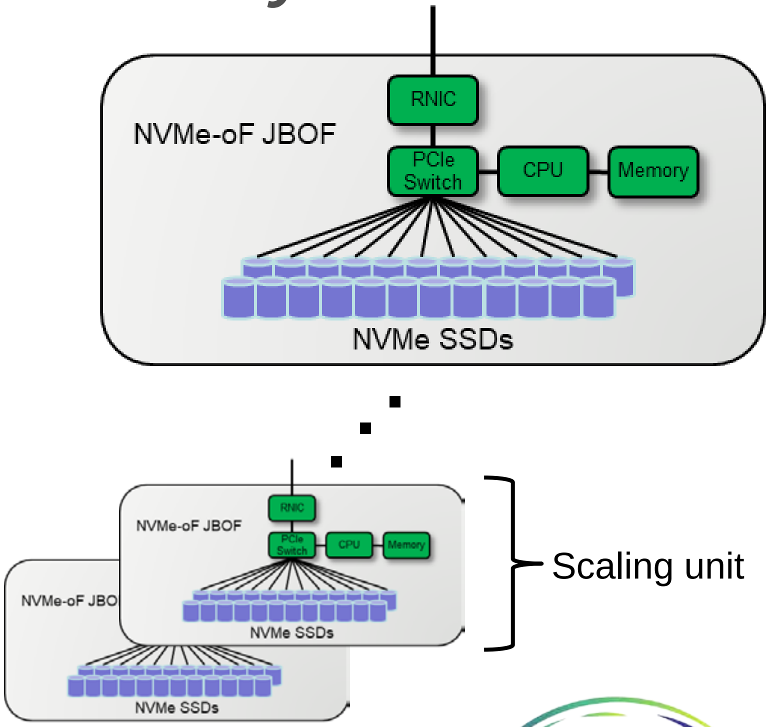


OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT
NOVEMBER 9-10, 2021

NVMe-oF™ Storage Targets Today

- Systems terminate the NVMe-oF architecture connection and use PCIe® based SSDs internally
  - SSDs behind an array/JBOF controller
- Performance Limits
  - SSD performance increasing faster than CPU NVMe-over-Ethernet-to-drive use cases
  - NIC performance
  - Latency - Store and Forward architecture
- Cost – CPU, SoC/rNICs, Switches, Memory don't scale well to match increasing SSD performance

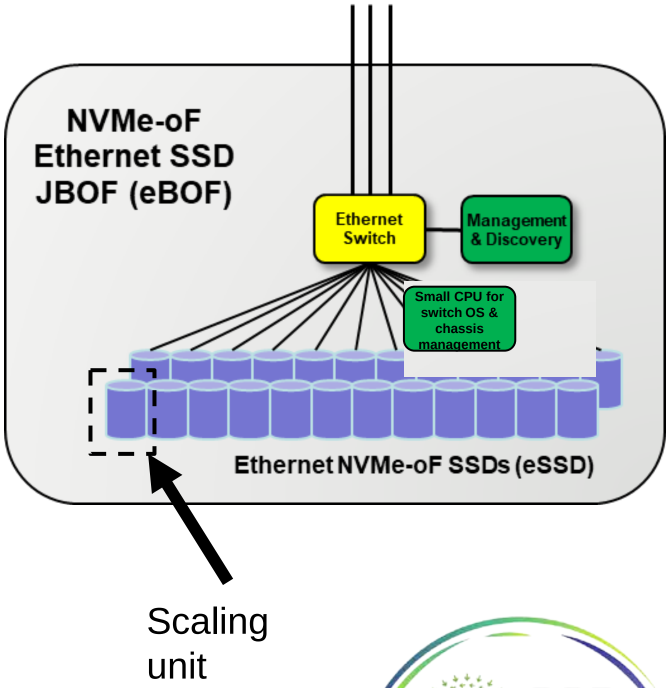


OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT NOVEMBER 9-10, 2021

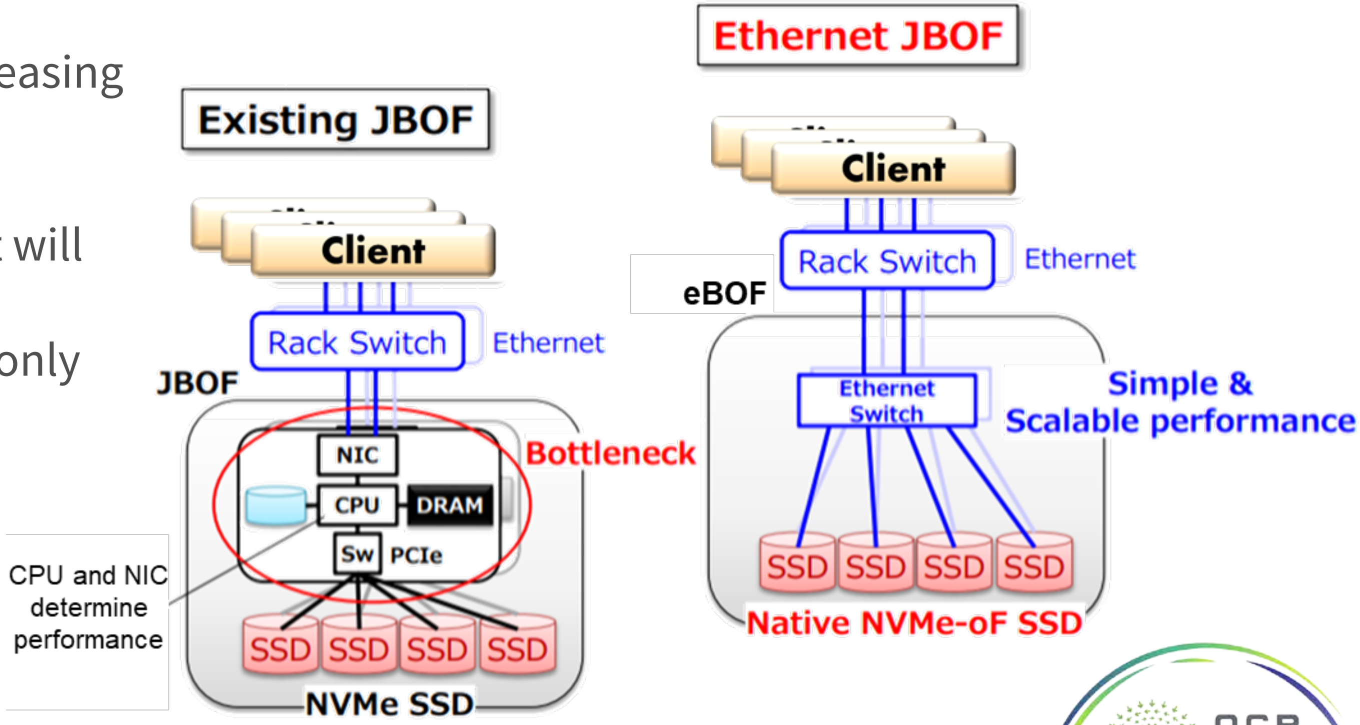
NVMe-oF ™ Ethernet SSDs - eSSDs

- With NVMe-oF technology termination on the drive itself, controller functionality is now distributed
  - Scaling point becomes a single drive in an inexpensive enclosure
  - Enables eBOFs (Ethernet-attached Bunch Of Flash)
    - Power, cooling, SSDs, and an Ethernet Switch
- Does this make each drive more expensive?
  - Maybe initially, but now customer buys their “controller” incrementally, as needed for new capacity
  - Efficiencies of scale now are applied to controller functionality
  - Lower cost/bandwidth and cost/IOPS



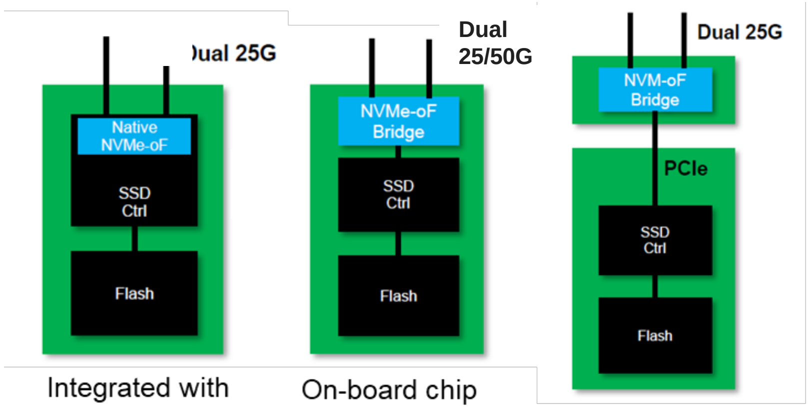
JBOF CPU/NIC Complex Can be a Bottleneck

- SSD throughput increasing faster than network bandwidth
  - SSD throughput will triple
  - Network speed only doubles

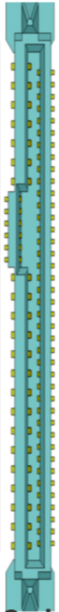


eSSDs

- Different eSSD designs today (largely NVMe-oF ™ /Ethernet)
- Some will support multiple interfaces and protocols
- RoCE, TCP



Name	Pin
GND	S1
S0T+ (A+)	S2
S0T- (A-)	S3
GND	S4
S0R- (B-)	S5
S0R+ (B+)	S6
GND	S7
RefClk1+	E1
RefClk1-	E2
3.3Vaux	E3
ePERst1#	E4
ePERst0#	E5
RSVD	E6
RSVD(Wake#) /SASAct2	P1
sPCIeRst/SAS	P2
RSVD(DevSLP#	P3
IfDet#	P4
Ground	P5
	P6
5 V	P7
	P8
	P9
PRSNT#	P10
Activity	P11
Ground	P12
12 V	P13
	P14
	P15



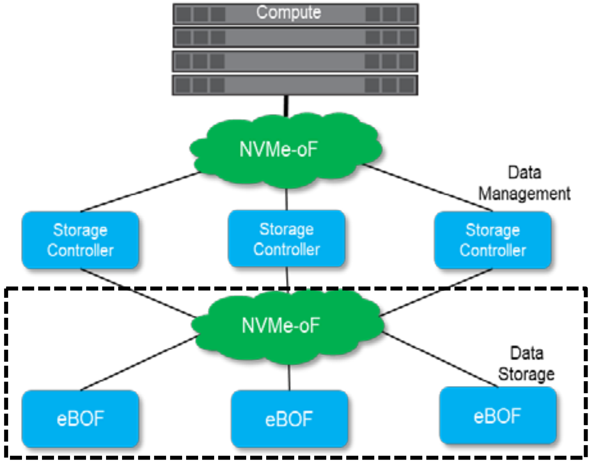
Pin	Name	SAS & Ethernet Signals proposal1	PCIe & Ethernet Signals proposal2
E7	RefClk0+		
E8	RefClk0-		
E9	GND		
E10	PETp0	TX1+	
E11	PETn0	TX1-	
E12	GND		
E13	PERn0		RX0-
E14	PERp0		RX0+
E15	GND		
E16	RSVD		
S8	GND		
S9	S1T+		
S10	S1T-		
S11	GND		
S12	S1R-	RX1-	
S13	S1R+	RX1+	
S14	GND		
S15	RSVD		
S16	GND		
S17	PETp1/S2T+		TX0+
S18	PETn1/S2T-		TX0-
S19	GND		
S20	PERn1/S2R-	RX0-	
S21	PERp1/S2R+	RX0+	
S22	GND		
S23	PETp2/S3T+		TX1+
S24	PETn2/S3T-		TX1-
S25	GND		
S26	PERn2/S3R-		
S27	PERp2/S3R+		
S28	GND		
E17	PETp3	TX0+	
E18	PETn3	TX0-	
E19	GND		
E20	PERn3		RX1-
E21	PERp3		RX1+
E22	GND		
E23	SMClk		
E24	SMDat		
E25	DualPortEn		

Fig1. U.2 pin assignment

SFF-8639 connector

Use Case: Behind the Controller

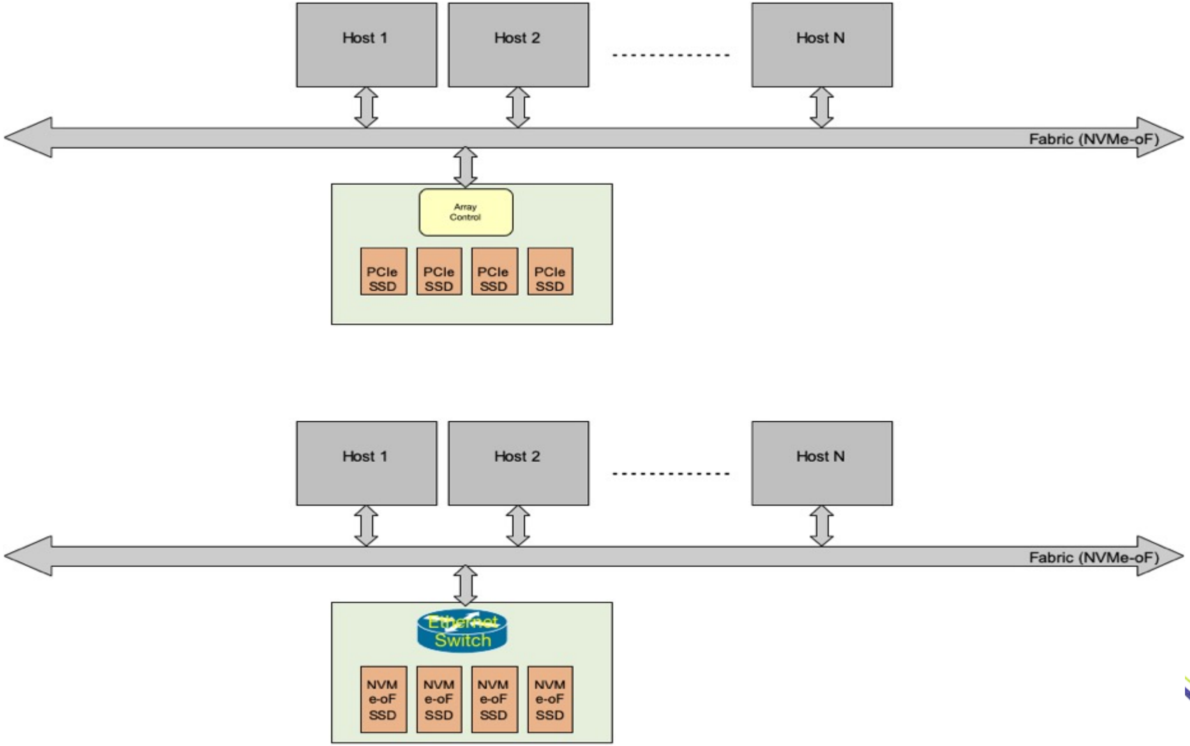
- Scale storage capacity with large pools of disks
  - Many NVMe® SSDs in many enclosures
  - PCIe® technology only scales so far and at JBOF increments
- Using eSSDs allows much higher scaling
  - Still allows hiding individual SSD management from users
- Data services in the storage controllers ➔ value add
  - Orchestration between hosts and large pools of disks
    - Whole disks or slices of disks that provide massive pools effectively
- Robust data protection schemes / distributed solution controllers



Use Case: Disaggregated SSD Storage

Today: Array controller handles conversion from NVMe-oF™ to PCIe® based drives

With eSSD: Ethernet drives only require an Ethernet Switch and fit into an eBOF for power and cooling



SNIA Native NVMe-oF ™ Drive Specification

- Discover and Configure: the drives, their interfaces, the speeds, the management capabilities
- Connectors
  - Some connectors may need to configure the PHY signals based on the type of drive interface
  - Survivability and mutual detection is important
- Pin-outs
  - For common connectors and form factors
- NVMe-oF technology integration
  - Discovery controllers / Admin controllers
- Management
  - Through Ethernet/TCP for Datacenter-wide management

Management

- Scale out orchestration of 10's of thousands of drives possible by using a RESTful API such as DTMF Redfish™
- Redfish/SNIA Swordfish™ follow a principle that each element reports its own management information
  - Follow links in higher level management directly to the drive's management endpoint
  - HTTP/TCP/Ethernet based
- NVMe-oF ™ Drive Interoperability Profile
  - Mockups of typical configurations
  - Push new models through Swordfish contributions
  - Publish Interoperability Profile at DMTF
- The profile maps to NVMe® and NVMe-MI™ technologies properties and actions
  - Swordfish NVMe Model Overview and Mapping Guide

The Latest Joint Work:
Mapping NVMe® Technology to Redfish and Swordfish

- A three-way effort, hosted by the SNIA SSM TWG (develops Swordfish)
- Base manageability for NVMe storage devices (from RF/SF/NVM Discussions)
  - Managing individual and aggregate devices in environments at scale
  - Provide a clear "map" for NVMe technology folks that don't know RF/SF to understand
- Work in progress:
  - Provide detailed implementation guidance for RF/SF interfaces covering multiple NVMe / NVMe-oF ™ device types

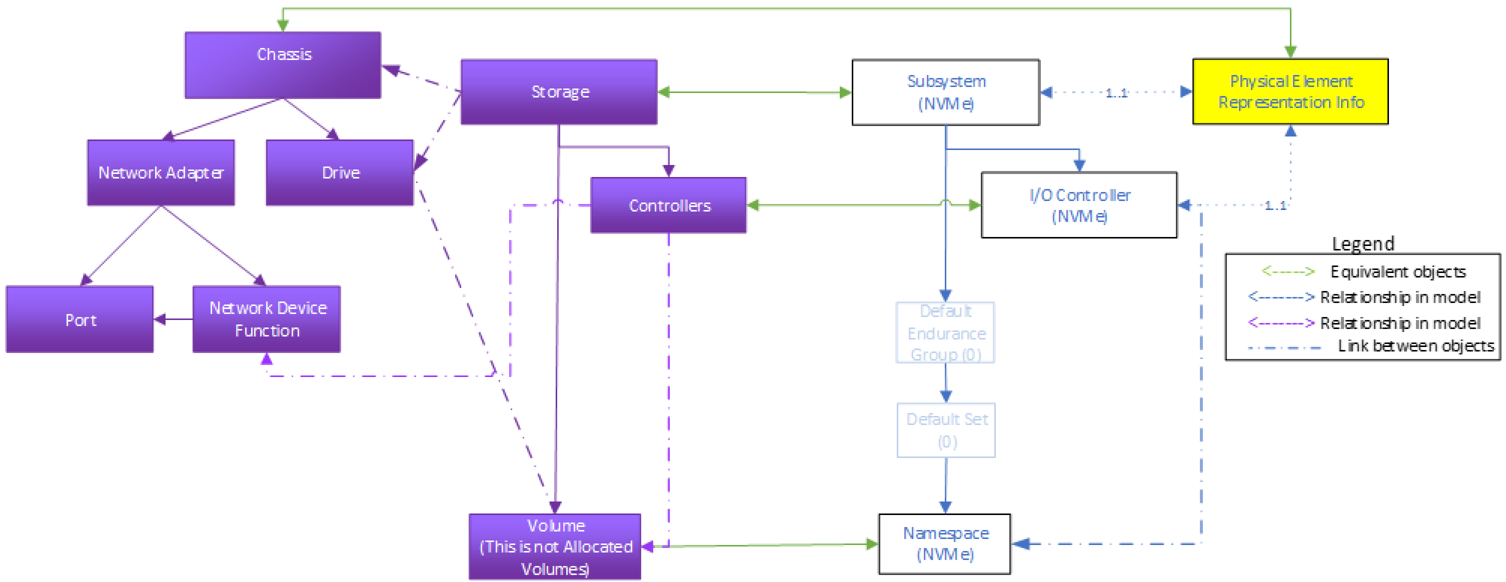
Fitting the Standards Together

- RF/SF use the available low-level transports to get device / transport specific information into the common models
  - RF/SF uses the commands that are provided in the NVMe® /NVMe-oF ™ /NVMe-MI ™ specifications
  - NVMe-MI specification can be used as the low-level to get the information into the high-level management environment as OOB access mechanism when appropriate
- Scope:
  - NVMe Subsystem, NVMe-oF and NVMe Domain Models

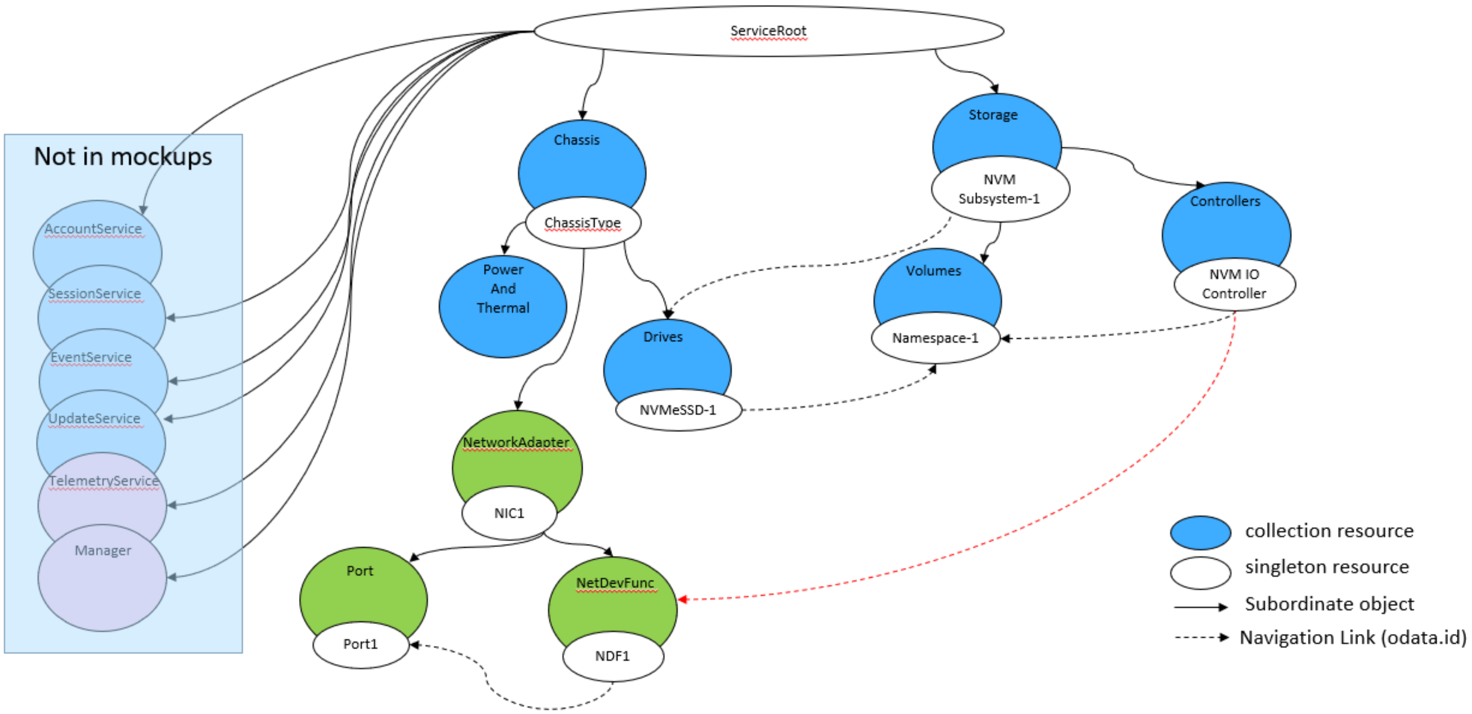
Major NVM Objects Mapped to RF/SF

- NVM Subsystem
  - An NVM subsystem includes one or more controllers, zero or more namespaces, and one or more ports. Examples of NVM subsystems include Enterprise and Client systems that utilize PCI Express based solid state drives and/or fabric connectivity.
- NVM Controller (IO, Admin and Discovery)
  - The interface between a host and an NVM subsystem
  - Admin controller: controller that exposes capabilities that allow a host to manage an NVM subsystem
  - Discovery: controller that exposes capabilities that allow a host to retrieve a Discovery Log Page
  - I/O: controller that implements I/O queues and is intended to be used to access a non-volatile memory storage medium
- Namespace
  - A quantity of non-volatile memory that may be formatted into logical blocks. When formatted, a namespace of size n is a collection of logical blocks with logical block addresses from 0 to (n-1)
- Endurance Group
  - A portion of NVM in the NVM subsystem whose endurance is managed as a group
- NVM Set
  - An NVM Set is a collection of NVM that is separate (logically and potentially physically) from NVM in other NVM Sets.
- NVM Domain
  - A domain is the smallest indivisible unit that shares state (e.g., power state, capacity information).
  - Domain members can be NVM controllers, endurance groups, sets or namespaces

NVMe® Subsystem Model: eSSD Use Case



Instance View: eSSD



Who is Developing Redfish and Swordfish*?

Redfish (DMTF)

American Megatrends, Inc.
Artesyn Embedded Technologies
Atos
Eaton
Ericsson AB
Google
Insyde Software Corp.
Mellanox Technologies
New H3C Technologies Co., Ltd
NVIDIA
Quanta Computer
Supermicro
Vertiv
Xilinx

Both

Arm Limited
Broadcom Inc.
Cisco
Dell Inc.
Fujitsu
Hewlett Packard Enterprise
Huawei
IBM
Intel Corporation
KIOXIA
Lenovo
Microchip Technology Inc.
NetApp
Seagate
Western Digital
VMware Inc

Swordfish (SNIA)

Hitachi
Inspur
Kalray
Microsoft
NEC Corporation
NGD Systems, Inc.
Pure Storage
Samsung
Silicon Motion, Inc.
SK Hynix
StarWind

*as of February 2020

Where to Find More Info..

SNIA Swordfish™

Swordfish Standards

- Schemas, Specs, Mockups, User and Practical Guide's, https://www.snia.org/swordfish

Swordfish Specification Forum

- Ask and answer questions about Swordfish
- http://swordfishforum.com/

Scalable Storage Management (SSM) TWG

- Technical Work Group that defines Swordfish
- Influence the next generation of the Swordfish standard
- Join SNIA & participate: https://www.snia.org/member_com/join-SNIA

Join the SNIA Storage Management Initiative

- Unifies the storage industry to develop and standardize interoperable storage management technologies
- https://www.snia.org/forums/smi/about/join

DMTF Redfish™

Redfish Standards

- Specifications, whitepapers, guides,… https://www.dmtf.org/standards/redfish

Open Fabric Management Framework

- OFMF Working Group (OFMFWG)
  - Description & Links https://www.openfabrics.org/working-groups/
- OFMFWG mailing list subscription
  - https://lists.openfabrics.org/mailman/listinfo/ofmfwg
- Join the Open Fabrics Alliance
  - https://www.openfabrics.org/membership-h

NVM Express

- Specifications https://nvmexpress.org/developers/
- Join: https://nvmexpress.org/join-nvme/

OPEN POSSIBILITIES.

OCP GLOBAL SUMMIT NOVEMBER 9-10, 2021

Disclaimer

- The NVM Express Logo® and the NVM Express®, NVMe®, NVMe-oF™, and NVMe-MI™ word marks are registered or unregistered service marks of the NVM Express organization in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited.
- The Redfish wordmark is an unregistered trademark of the DMTF organization.
- The DMTF & Redfish logo is a registered trademark of DMTF.
- The Swordfish logo and wordmark are unregistered service marks of the SNIA organization.

Thank you!

OCP GLOBAL SUMMIT

NOVEMBER 9-10, 2021