STORAGE DEVELOPER CONFERENCE
SDC 23
BY Developers FOR Developers

NVM Express State of the Union and an overview of Live Migration

Presented by
Ross Stenfort, Meta
Mike Allison, Samsung

Speakers

Ross Stenfort
Hardware Systems Engineer

Meta

Mike Allison
Sr. Director NAND Product Planning

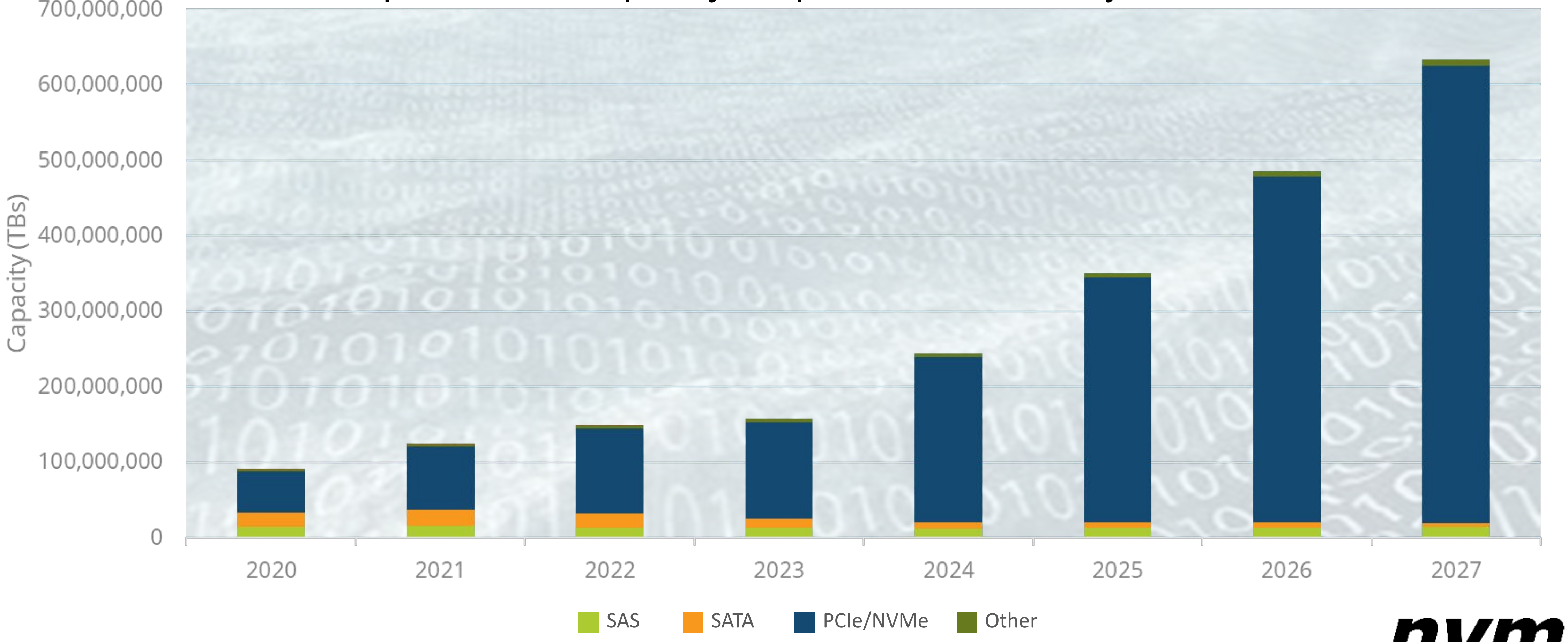
SAMSUNG

nvm EXPRESS®

SDC 23

NVMe® Specifications – The Language of Storage

Enterprise SSD Capacity Shipment Forecast by Interface



Source: IDC Worldwide Solid State Drive Forecast, 2023-2027 Doc # US49401623, Apr 2023

NVMe® Technology Powers the Connected Universe

Petabytes	2021	2022	2023	2024	2025	2026	2027
Enterprise	32,483	42,973	37,094	48,602	65,701	82,499	106,106
Cloud	73,191	86,307	53,534	89,678	128,164	175,730	237,949
Client	145,610	157,304	200,391	274,530	350,518	437,054	517,991

Consumer · Client · Embedded · Enterprise · Cloud

Cell Phones   Tablets   Laptops   Desktops   Storage Arrays   Data Centers

Source: Data and projections provided by Forward Insights Q2'23

NVM Express Organization

Board of Directors

Chair: Amber Huffman
Treasurer: Curtis Ballard
Secretary: Dave Landsman

Technical Workgroup

Chair: Peter Onufryk

Marketing Workgroup

Chair: Cameron Brett, Kerry Munson

SUBGROUPS

Computational Storage

Chairs: Kim Malone, Bill Martin

Fabric & Multi-Domain Subsystem

Chair: Fred Knight, Erik Smith

Management Interface

Chairs: Austin Bolen, John Geldman

Interoperability and Compliance

Chair: Ryan Holmqvist

NVMe-oF™ Boot

Chairs: Phil Cayton, Rob Davis, Doug Farley

Errata

Chair: Mike Allison

Board of Directors

Elections occur yearly

DELL EMC

intel

Hewlett Packard Enterprise

Google

Microsoft

Western Digital

Meta

NetApp

SAMSUNG

MARVELL

Micron

SEAGATE

SOLIDIGM

nvm EXPRESS

Organizational **Enhancements**

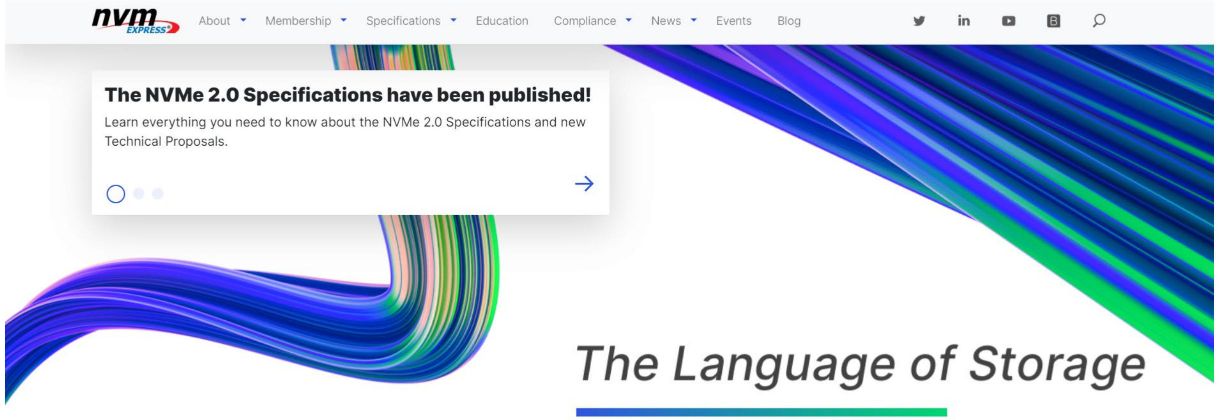
Tooling Updates
Zoom, Causeway, Bugzilla

Errata Taskgroup

Website Redesign

Software Taskgroup Framework

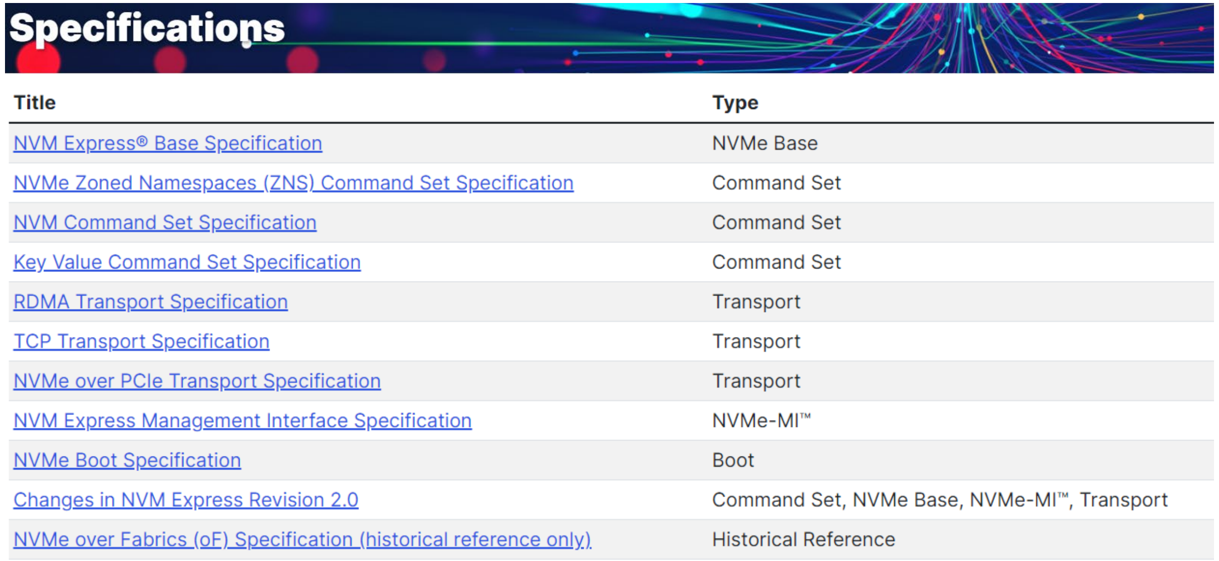
Modernizing the NVM Express Website



The NVMe 2.0 Specifications have been published!

Learn everything you need to know about the NVMe 2.0 Specifications and new Technical Proposals.

The Language of Storage



Specifications

Title	Type
NVM Express® Base Specification	NVMe Base
NVMe Zoned Namespaces (ZNS) Command Set Specification	Command Set
NVM Command Set Specification	Command Set
Key Value Command Set Specification	Command Set
RDMA Transport Specification	Transport
TCP Transport Specification	Transport
NVMe over PCIe Transport Specification	Transport
NVM Express Management Interface Specification	NVMe-MI™
NVMe Boot Specification	Boot
Changes in NVM Express Revision 2.0	Command Set, NVMe Base, NVMe-MI™, Transport
NVMe over Fabrics (oF) Specification (historical reference only)	Historical Reference

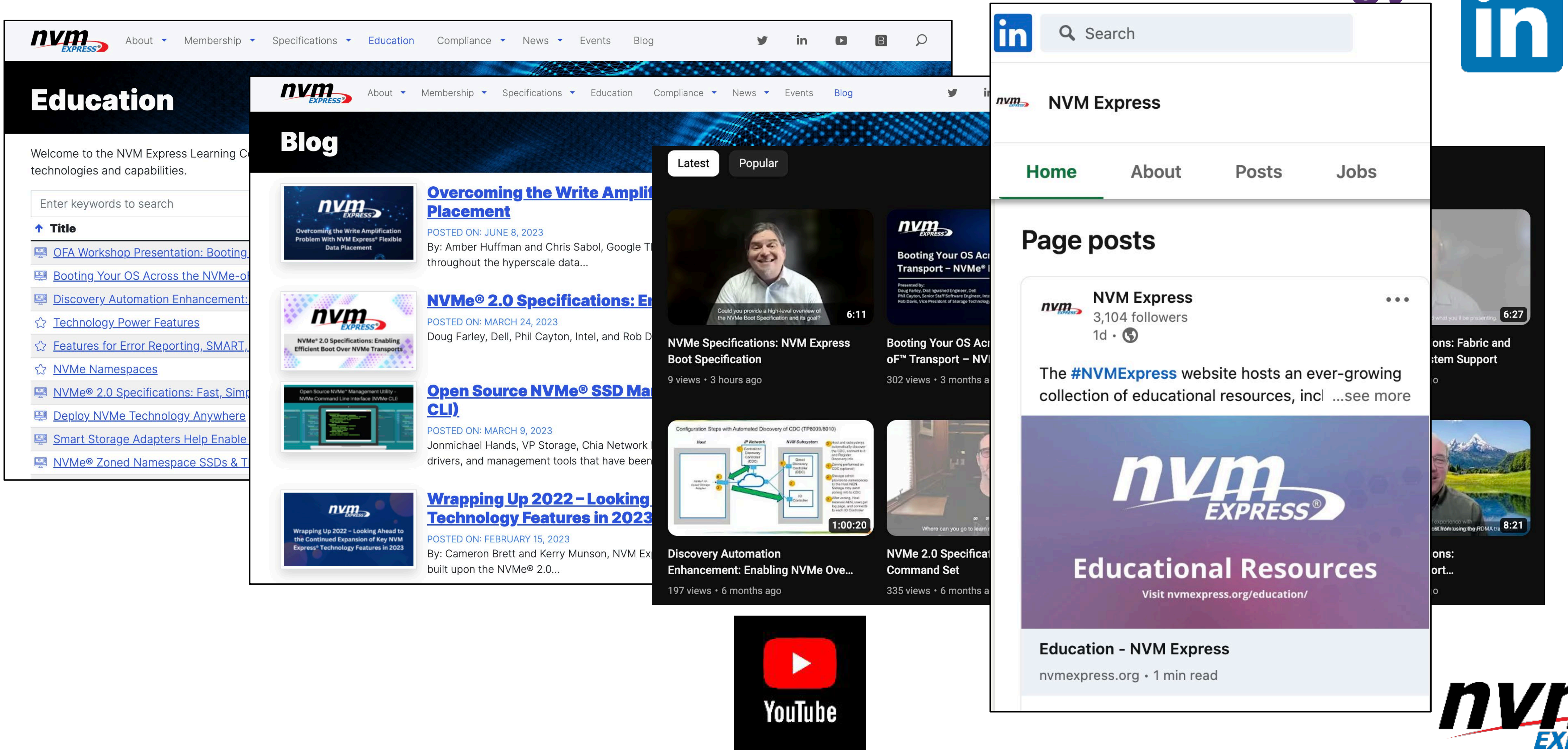
Refreshed pages

Updated user interface

Consolidated & reorganized

- *Specifications*
- *Blogs*
- *Webinars*

Resources to Learn About NVMe® Technology



Education

Welcome to the NVM Express Learning C
technologies and capabilities.

Enter keywords to search

Title

- OFA Workshop Presentation: Booting
- Booting Your OS Across the NVMe-of
- Discovery Automation Enhancement:
- Technology Power Features
- Features for Error Reporting, SMART,
- NVMe Namespaces
- NVMe® 2.0 Specifications: Fast, Simp
- Deploy NVMe Technology Anywhere
- Smart Storage Adapters Help Enable
- NVMe® Zoned Namespace SSDs & T

Blog

Overcoming the Write Amplif Placement
POSTED ON: JUNE 8, 2023
By: Amber Huffman and Chris Sabol, Google T throughout the hyperscale data...

NVMe® 2.0 Specifications: E
POSTED ON: MARCH 24, 2023
Doug Farley, Dell, Phil Cayton, Intel, and Rob D

Open Source NVMe® SSD Ma CLI)
POSTED ON: MARCH 9, 2023
Jonmichael Hands, VP Storage, Chia Network drivers, and management tools that have been

Wrapping Up 2022 – Looking Technology Features in 2023
POSTED ON: FEBRUARY 15, 2023
By: Cameron Brett and Kerry Munson, NVM Ex built upon the NVMe® 2.0...

NVMe Specifications: NVM Express Boot Specification
9 views • 3 hours ago

Booting Your OS Acr oF™ Transport – NVI
302 views • 3 months a

ons: Fabric and tem Support

Discovery Automation Enhancement: Enabling NVMe Ove...
197 views • 6 months ago

NVMe 2.0 Specificat Command Set
335 views • 6 months a

ons: ort...

NVM Express

Page posts

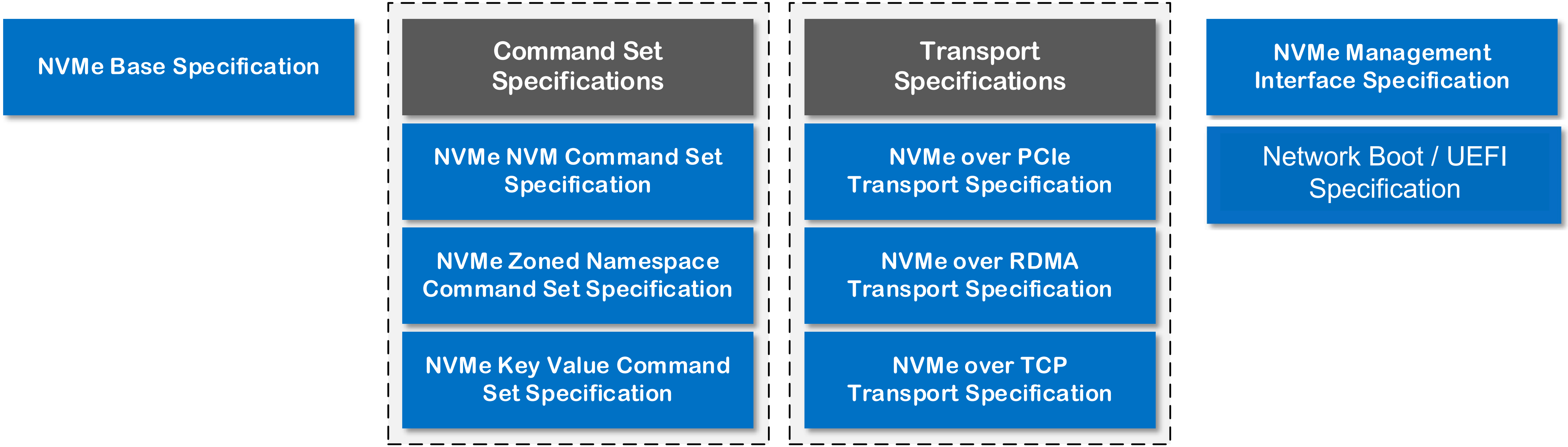
NVM Express
3,104 followers
1d •

The #NVMExpress website hosts an ever-growing collection of educational resources, incl ...see more

Educational Resources
Visit nvmexpress.org/education/

Education - NVM Express
nvmexpress.org • 1 min read

NVMe® 2.0 Family of Specifications



NVMe Base Specification

Command Set Specifications

NVMe NVM Command Set Specification

NVMe Zoned Namespace Command Set Specification

NVMe Key Value Command Set Specification

Transport Specifications

NVMe over PCIe Transport Specification

NVMe over RDMA Transport Specification

NVMe over TCP Transport Specification

NVMe Management Interface Specification

Network Boot / UEFI Specification

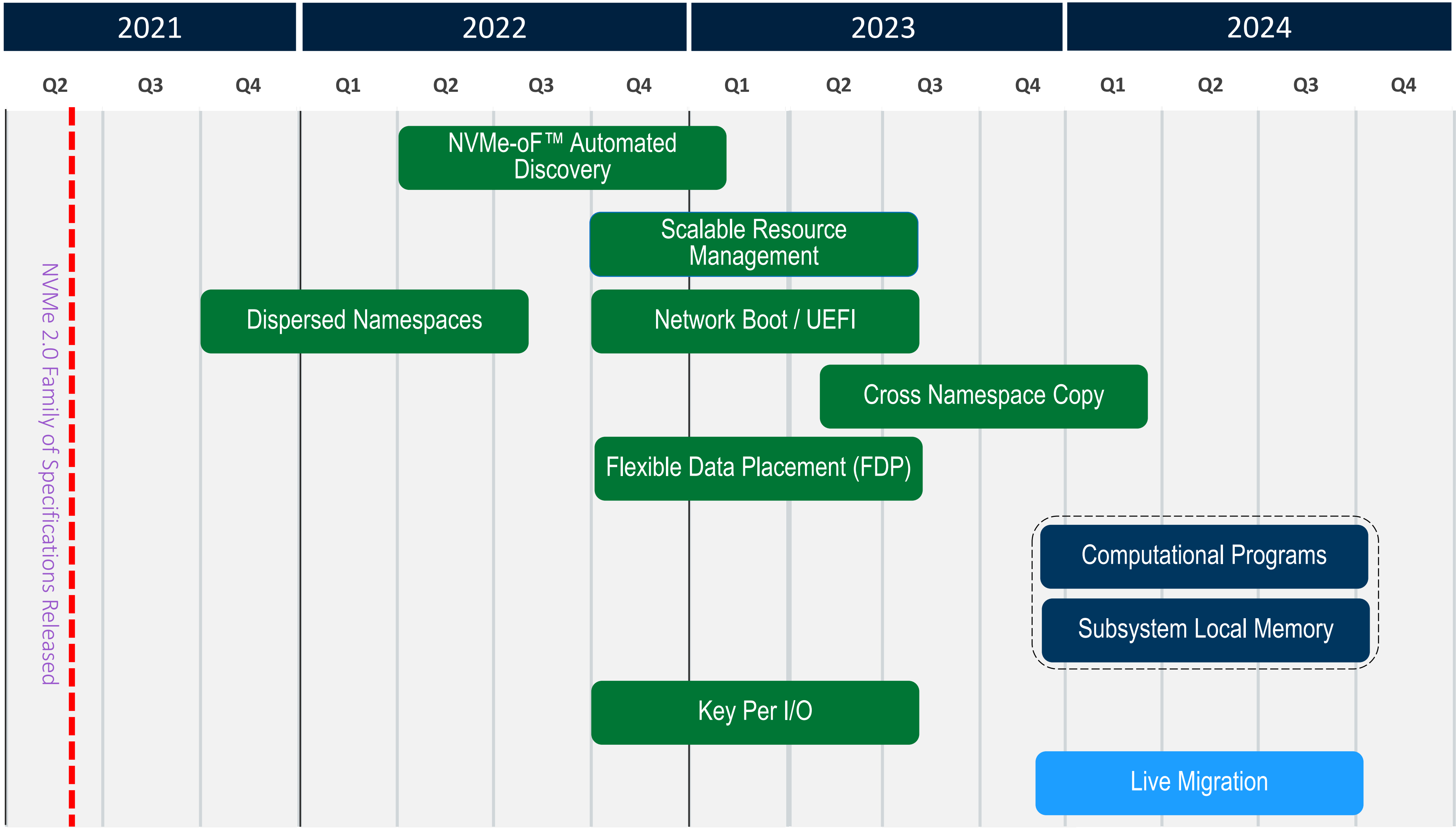
NVMe 2.0 specifications were released on June 3, 2021
Refer to nvmexpress.org/developers

Activity Since Release of NVMe® 2.0 Family of Specifications*

New Authorized Technical Proposals	Ratified Technical Proposals	Ratified ECNs
60	69	13

[*] Activity as of 7/28/2023

NVMe® Specifications Feature Roadmap



2021 | 2022 | 2023 | 2024

Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4 Q1 Q2 Q3 Q4

NVMe 2.0 Family of Specifications Released

NVMe-oF™ Automated Discovery

Scalable Resource Management

Dispersed Namespaces

Network Boot / UEFI

Cross Namespace Copy

Flexible Data Placement (FDP)

Computational Programs

Subsystem Local Memory

Key Per I/O

Live Migration

🟩 **Ratified Feature** (left edge indicates ratification quarter)

🟦 **Planned Feature** (left edge indicates planned ratification quarter)

■ **Planned New Specification** (left edge indicates planned ratification quarter)

🟦 **Ratified New Specification** (left edge indicates planned ratification quarter)

Specification **Advancements**

Flexible Data Placement

Reducing Write Amplification

Network Boot / UEFI

New Network Storage Functionality

Computational Storage

Executing Programs within a Device

Live Migration *new feature!*

Seamlessly Move Data across vMachines

NVMe® Live Migration

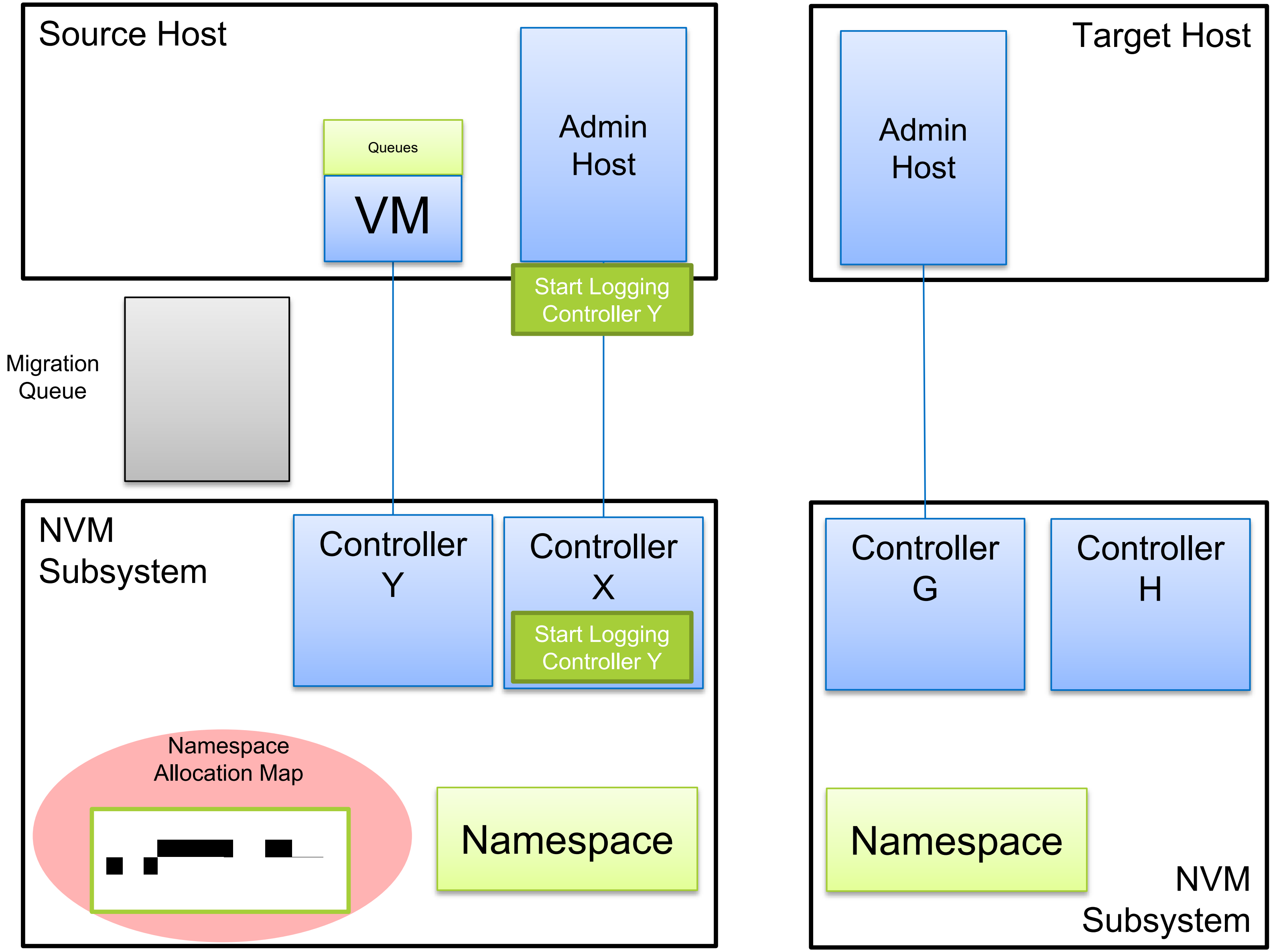
Benefits

- NVM Express® is adding capabilities to allow host to manage the migrating VM from one NVM subsystem to a different NVM subsystem by supporting the migration of the controller being used by the VM which includes the attached namespaces and the controller state.
- Pre-Copy Phase Host Actions
  - Requests the controller to track LBA changes (dirty LBAs) of the attached namespaces
  - Migrate the allocated LBAs of the attached namespaces
  - Migrate the dirty LBAs
  - Host may use a new mechanism to throttle commands processing by migrating controller to slow down changes
- Stop-and-Copy Phase Host Actions
  - Requests the controller to pause causing all fetched commands to be completed
  - Migrate any remaining dirty LBAs
- Post-Copy Phase
  - Migrate controller state
  - Resume the migrated controller

Building the Pieces

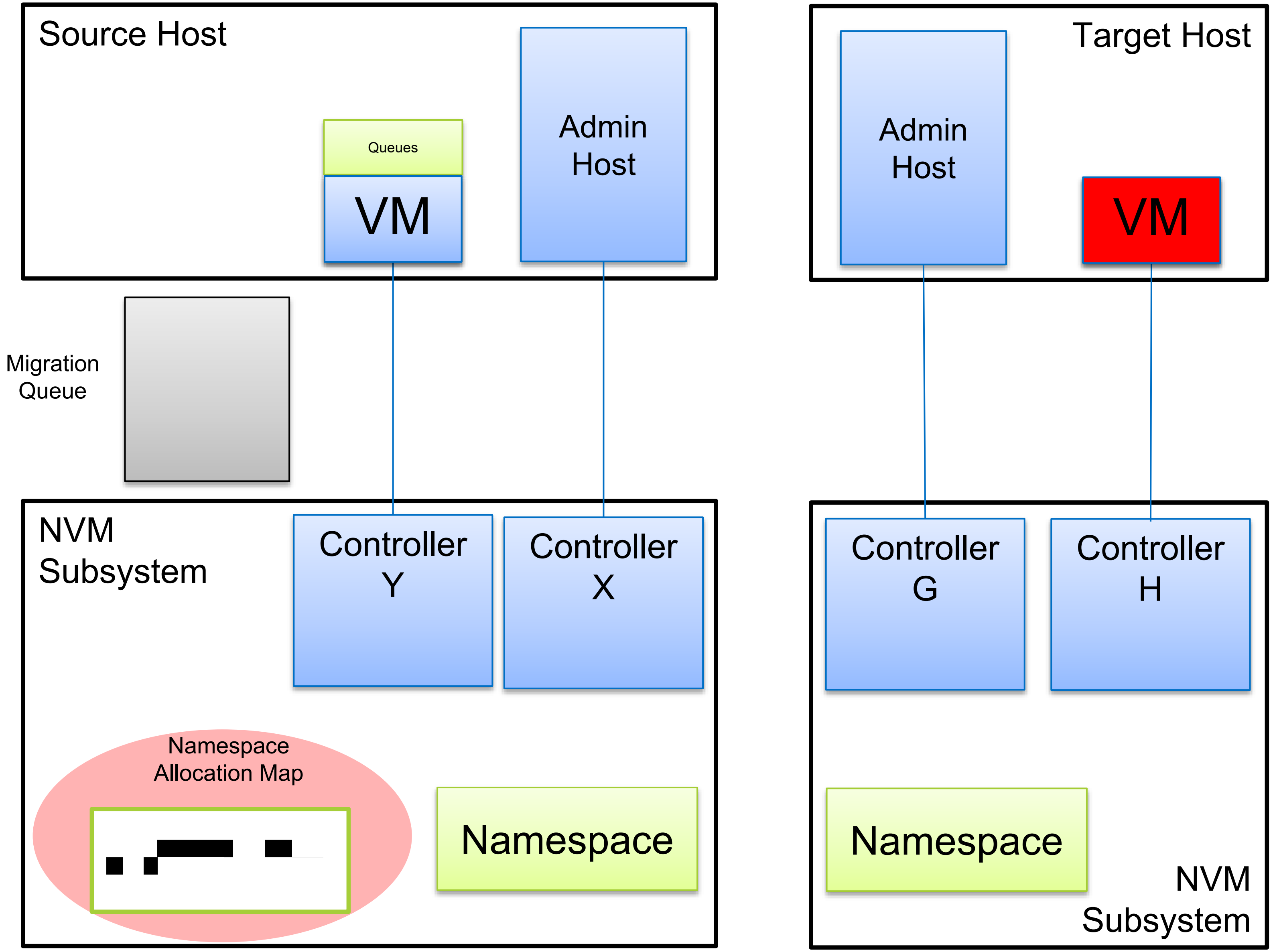
- TP4165 Tracking LBA Allocation with Granularity
  - Reporting of allocated LBAs within a namespace for migrating a namespace
  - Usable in Snapshot use cases
- TP4159 PCIe® Infrastructure for Live Migration
  - Developing the theory of operation
- A TPAR to:
  - Support limit the BW and IOPS of a controller to allow slowing down of command processing on a migrating controller

Pre-Copy Phase Start



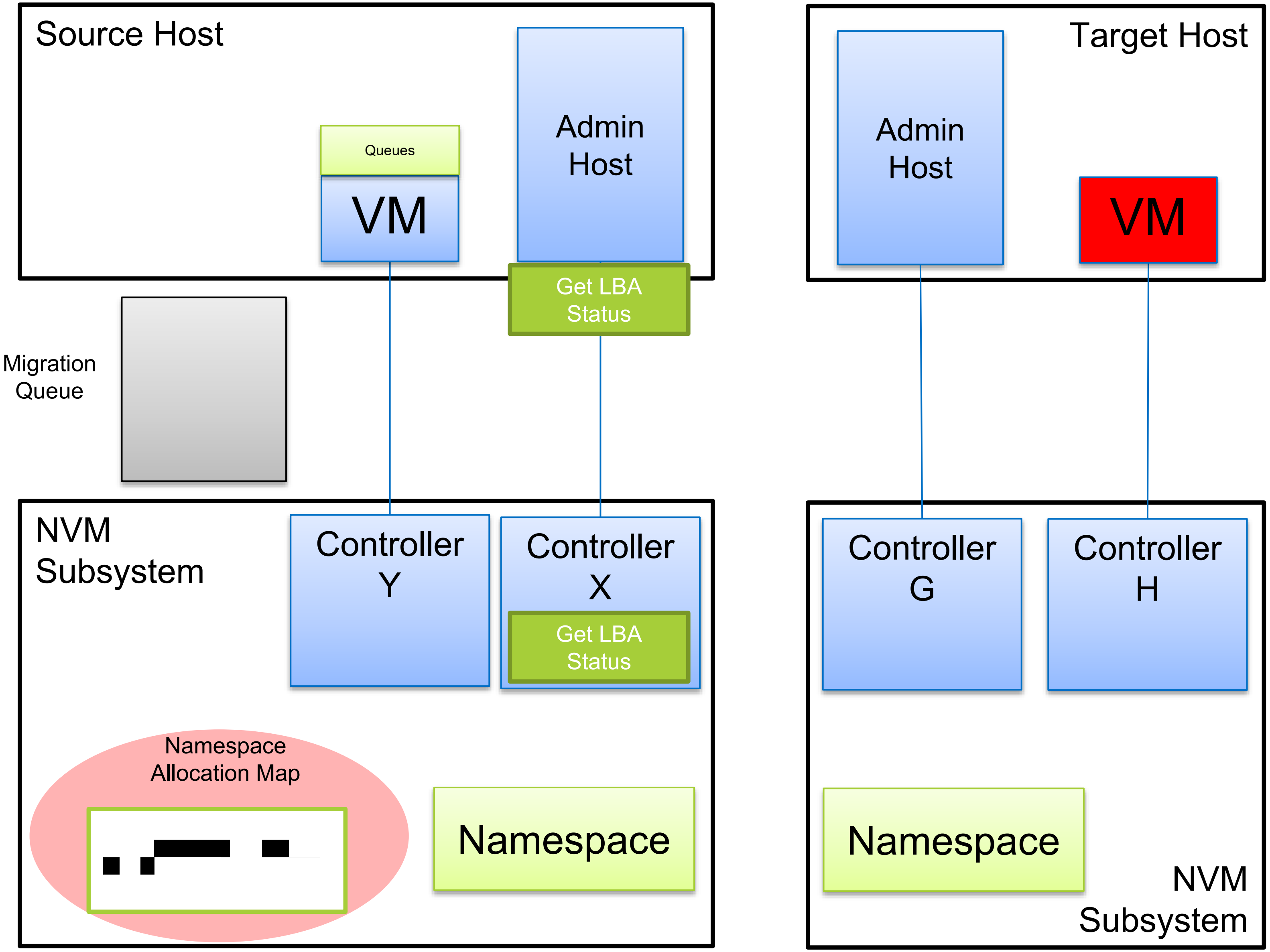
- Source Admin Host initiates a migration of a controller by requesting to log LBA changes (dirty LBAs)
- A Migration Queue is established

Pre-Copy Phase Start



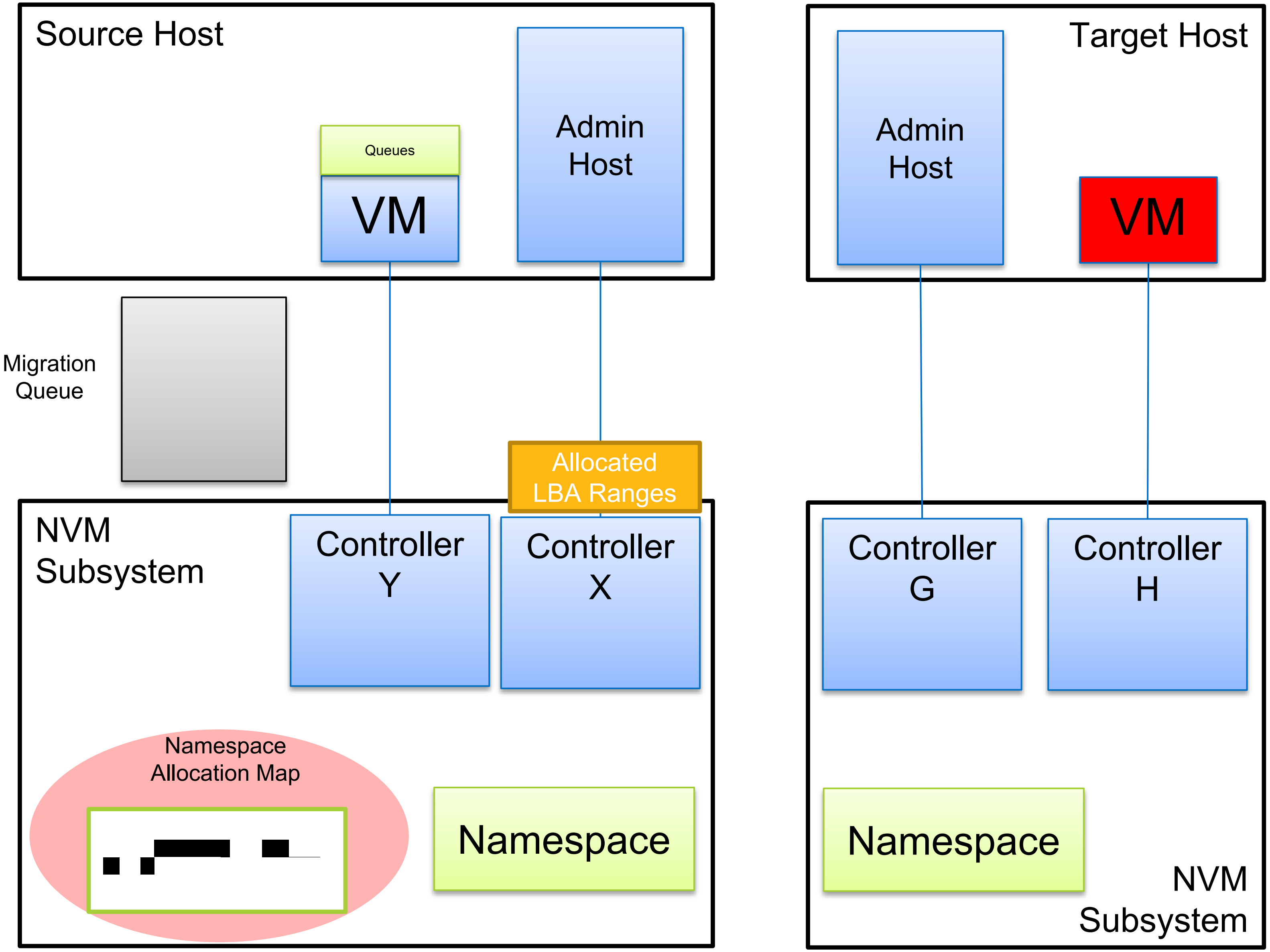
- Source Admin Host initiates a migration of a controller by requesting to log LBA changes (dirty LBAs)
- A Migration Queue is established
- The memory associated with the migrating VM can be moved anytime by the Source Admin Host

Pre-Copy Phase – Initial Namespace Migration



Source Admin Host issues Get LBA status command to obtain the allocated LBAs

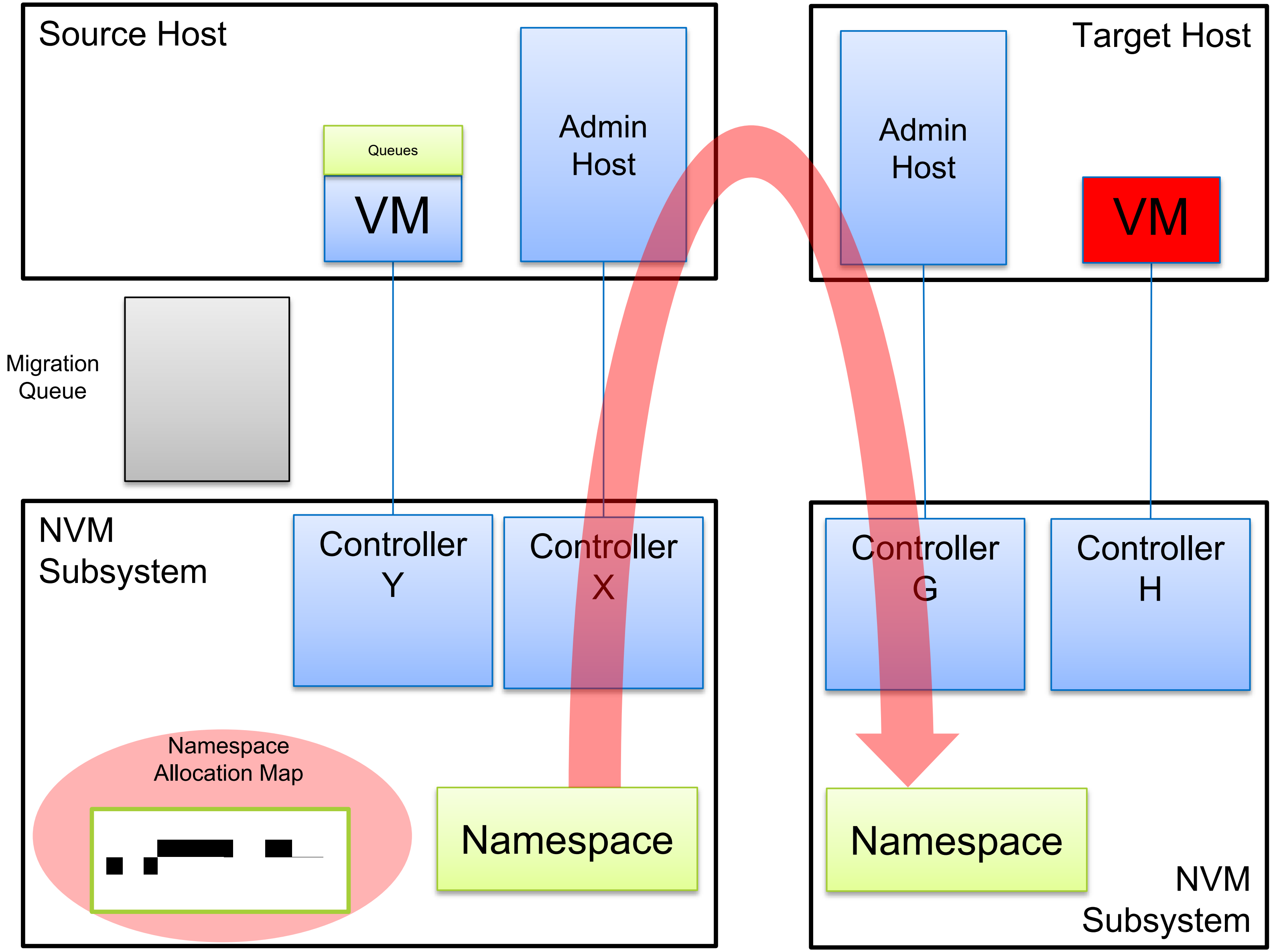
Pre-Copy Phase – Initial Namespace Migration



Source Admin Host issues Get LBA status command to obtain the allocated LBAs

- Controller returns a list of descriptors. Each descriptor indicates an LBA range

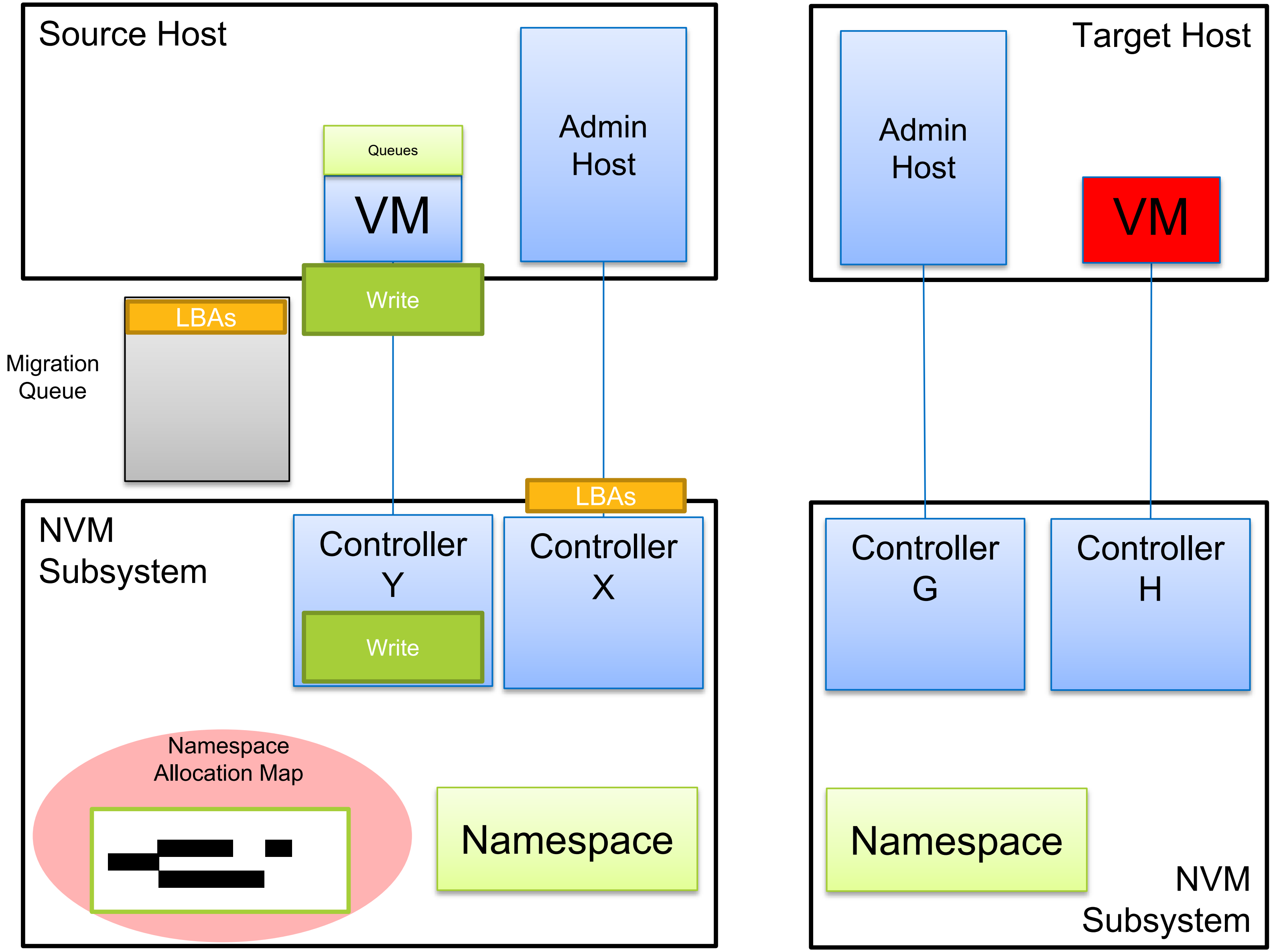
Pre-Copy Phase – Initial Namespace Migration



Source Admin Host issues Get LBA status command to obtain the allocated LBAs

- Controller returns a list of descriptors. Each descriptor indicates an LBA range
- The Source Admin Host uses these LBA ranges to issue read commands to copy the allocated LBAs to the destination

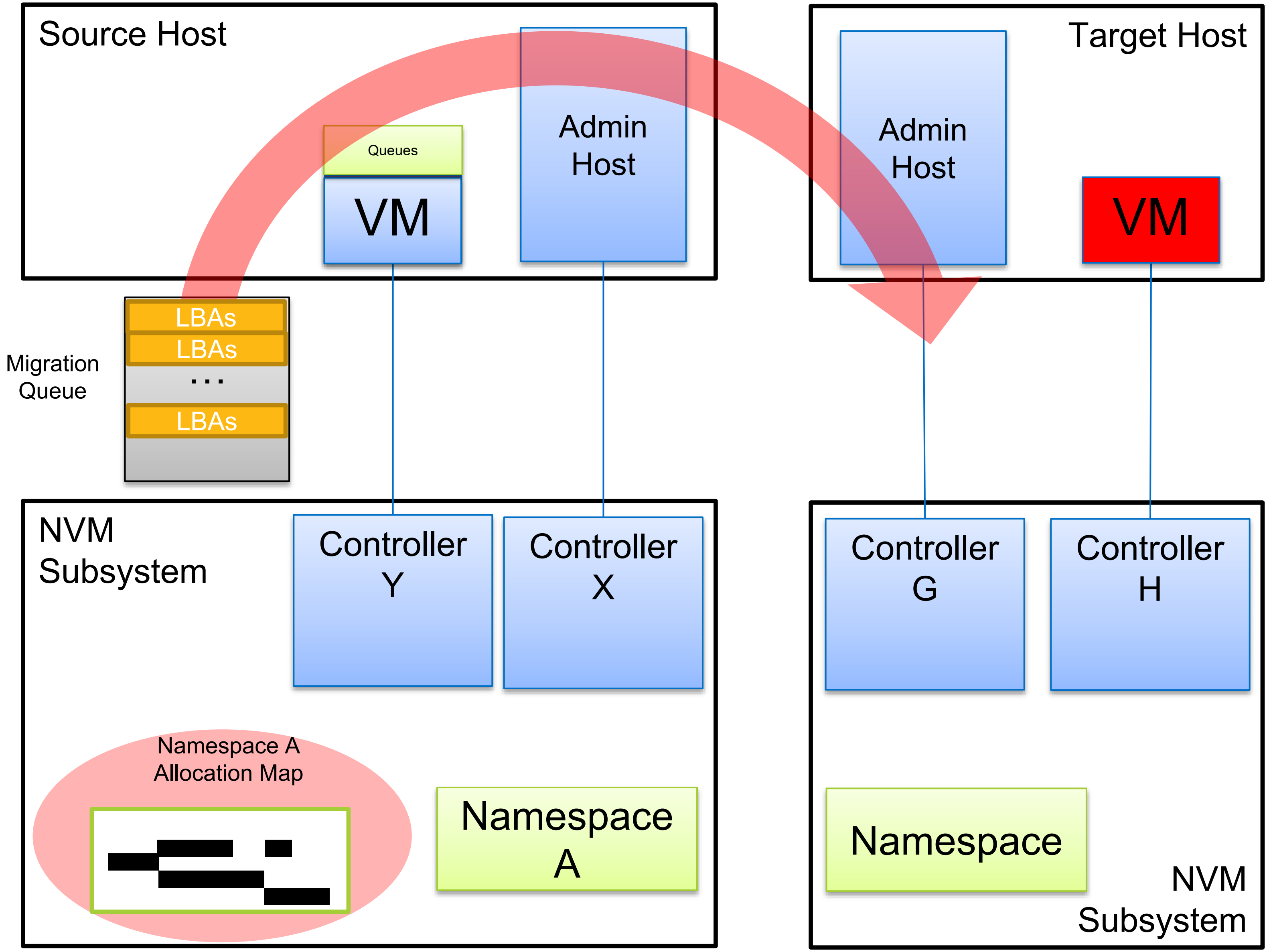
Pre-Copy Phase – Migrating Controller Continues



NVMe® commands that cause LBA changes to the namespace are logged in the Migration Queue

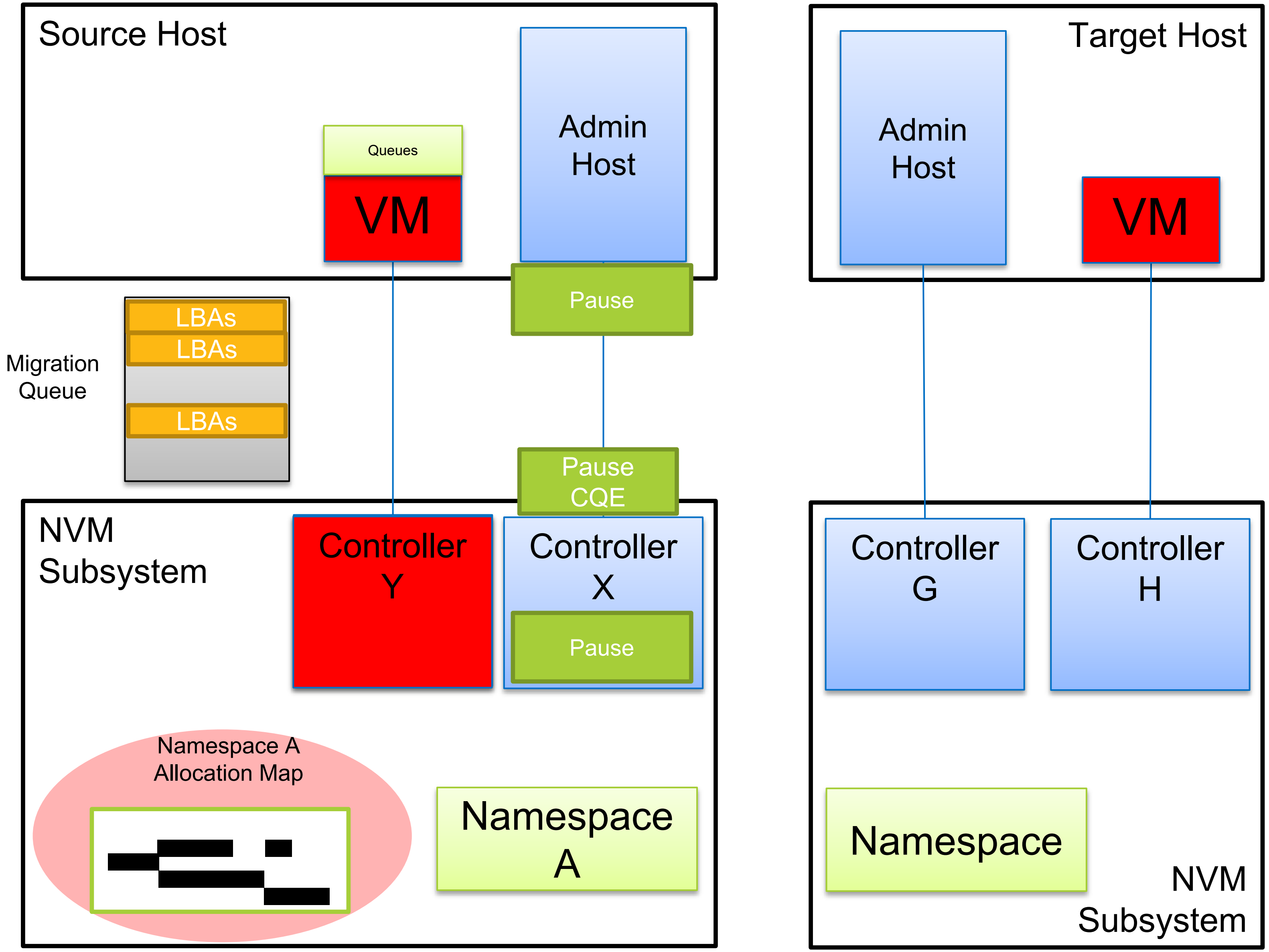
- Write commands
- LBA deallocation due to the Dataset Management command

Stop-and-Copy Phase – Pause Migrating Controller



After coping the allocated LBAs to the destination, the Source Admin Host may migrate the dirty LBAs

Stop-and-Copy Phase – Pause Migrating Controller

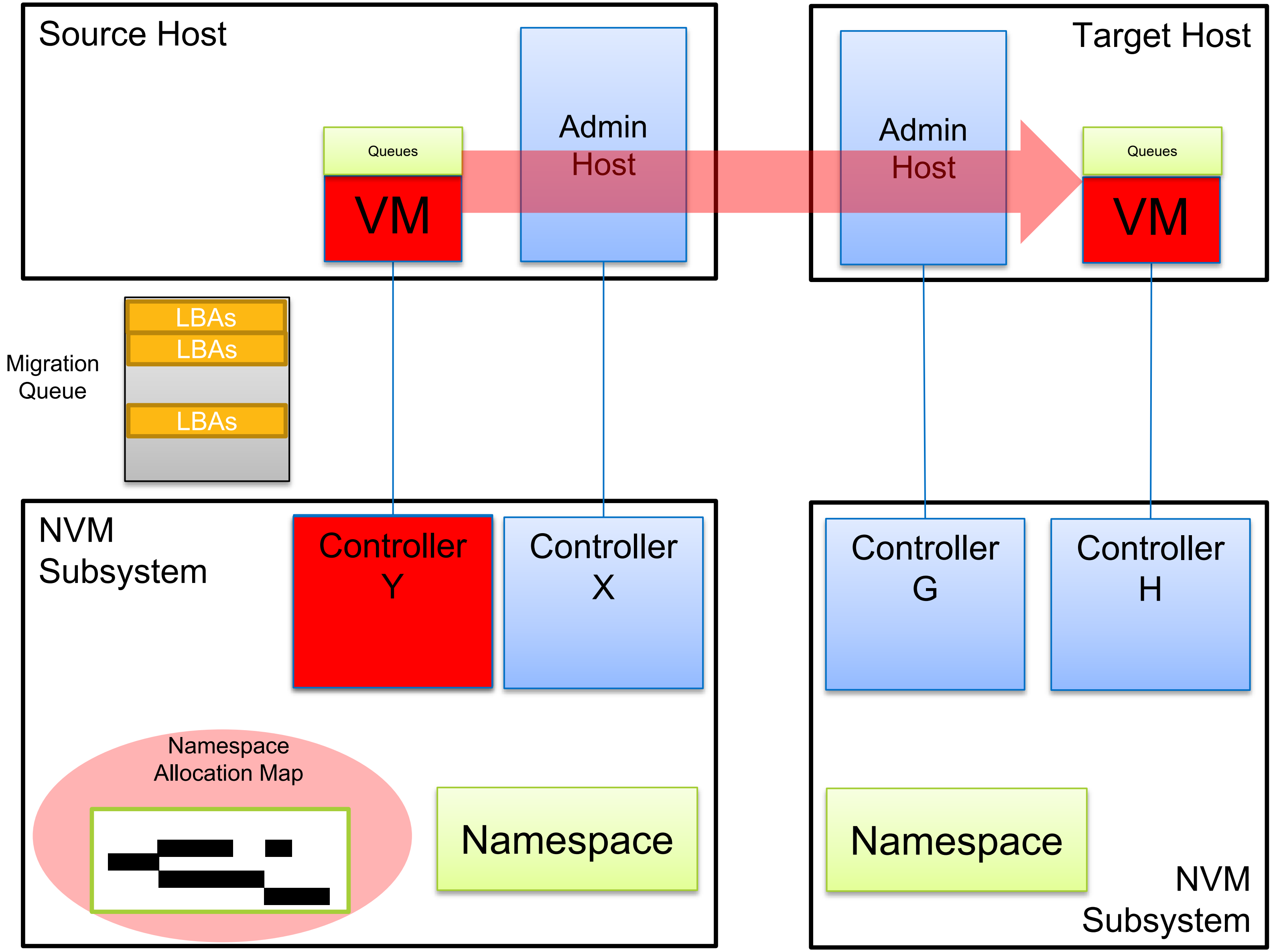


At some point the Source Admin Host pauses the VM

Issues a command to Pause the migrating controller to have the controller:

- Stop fetching commands
- Complete all previously fetched commands

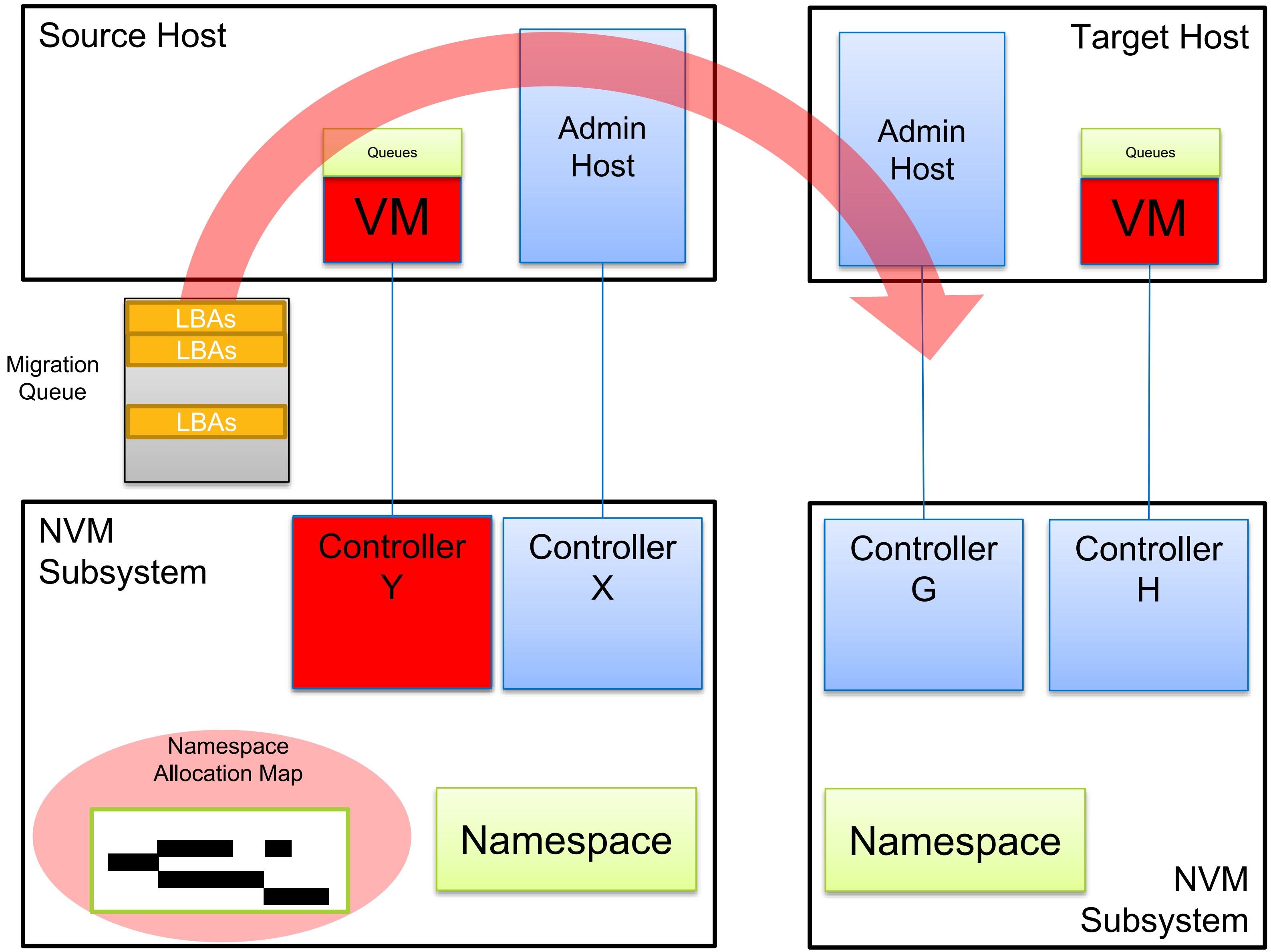
Stop-and-Copy Phase – Finish Migrating



Source Host

- Completes migration of VM

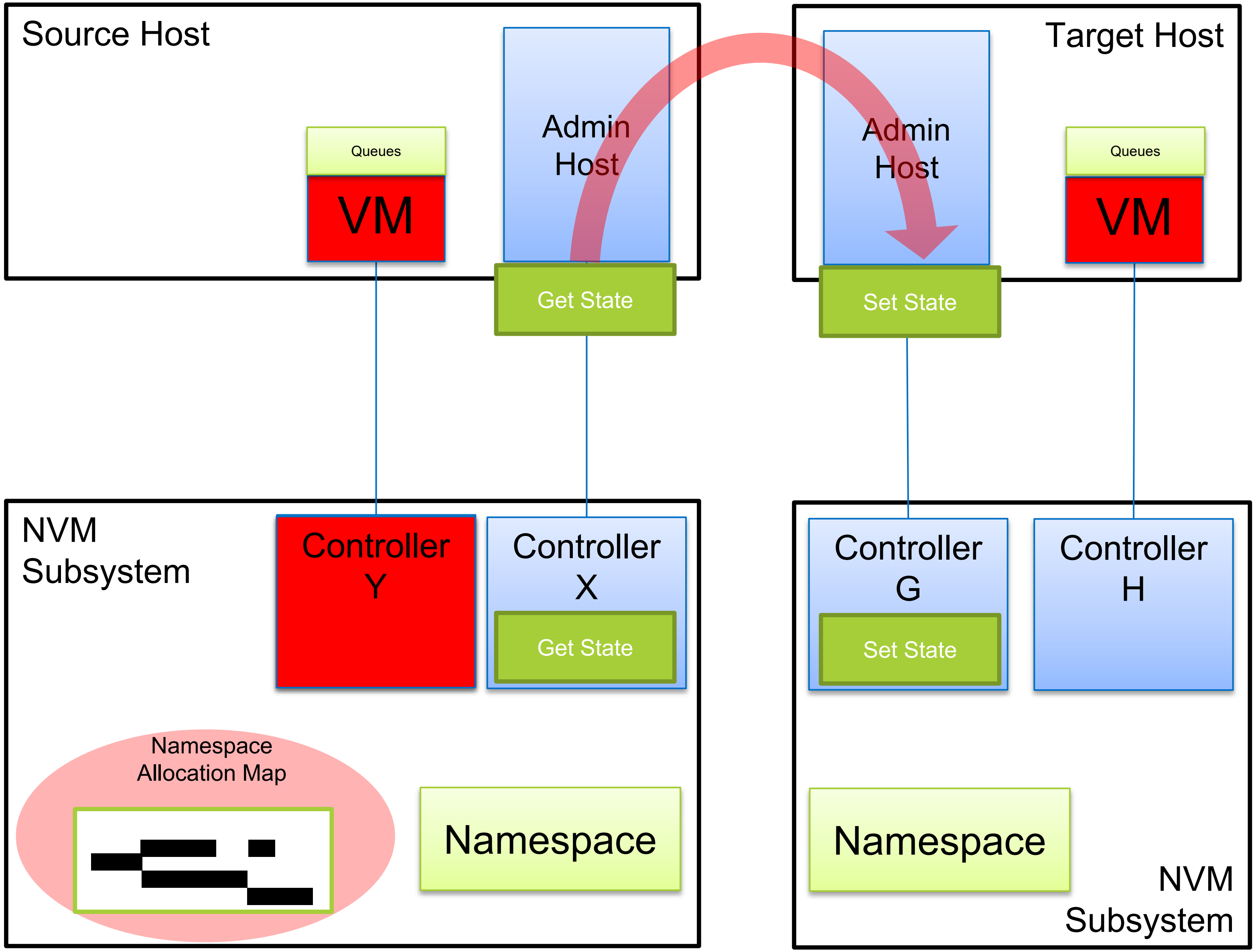
Stop-and-Copy Phase – Finish Migrating



Source Host

- Completes migration of VM
- Completes Migration of namespace dirty LBAs

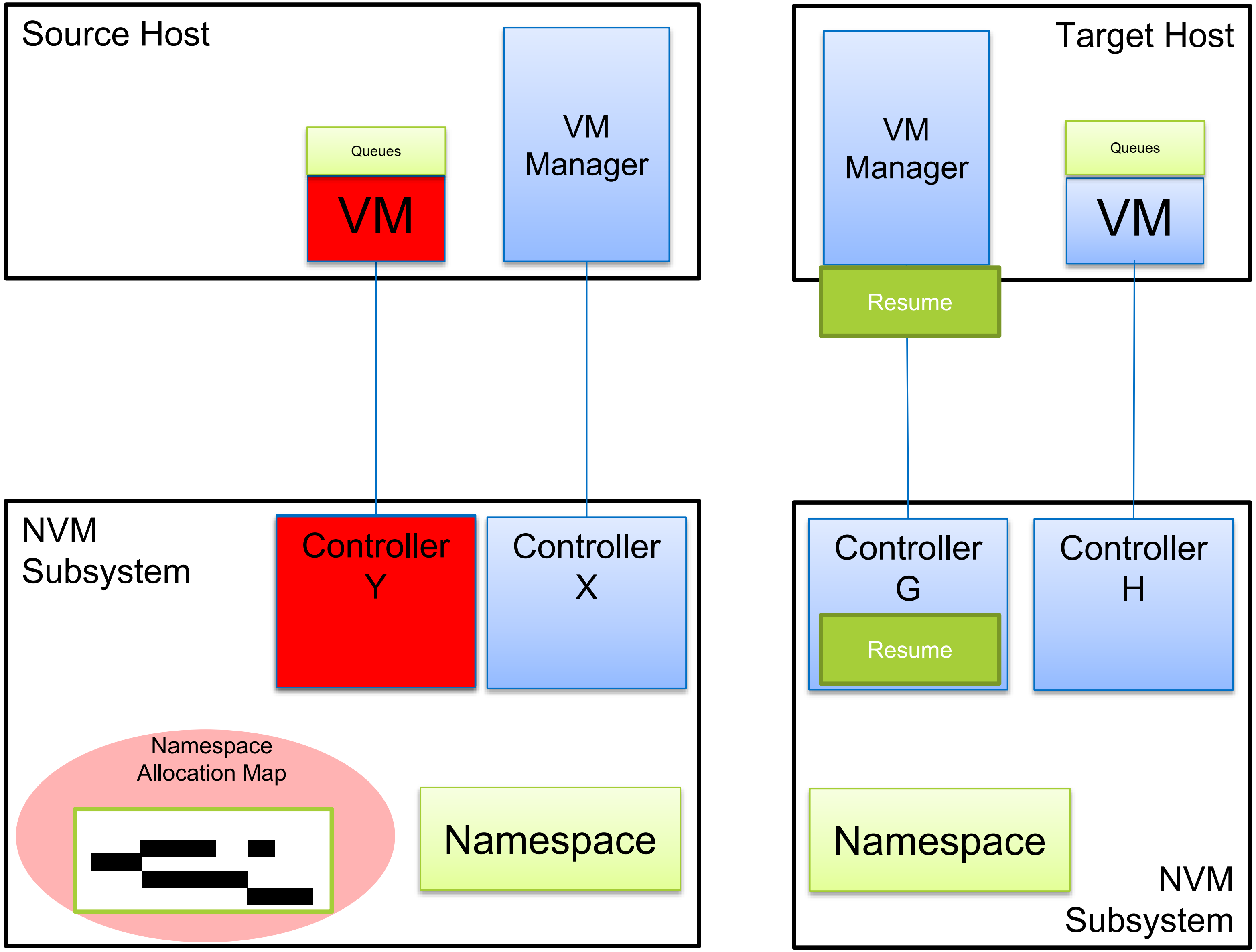
Post-copy Phase – Migrate Controller State



Source Admin Host

- Issuing command to get the migrating controller state and put that state into the destination controller

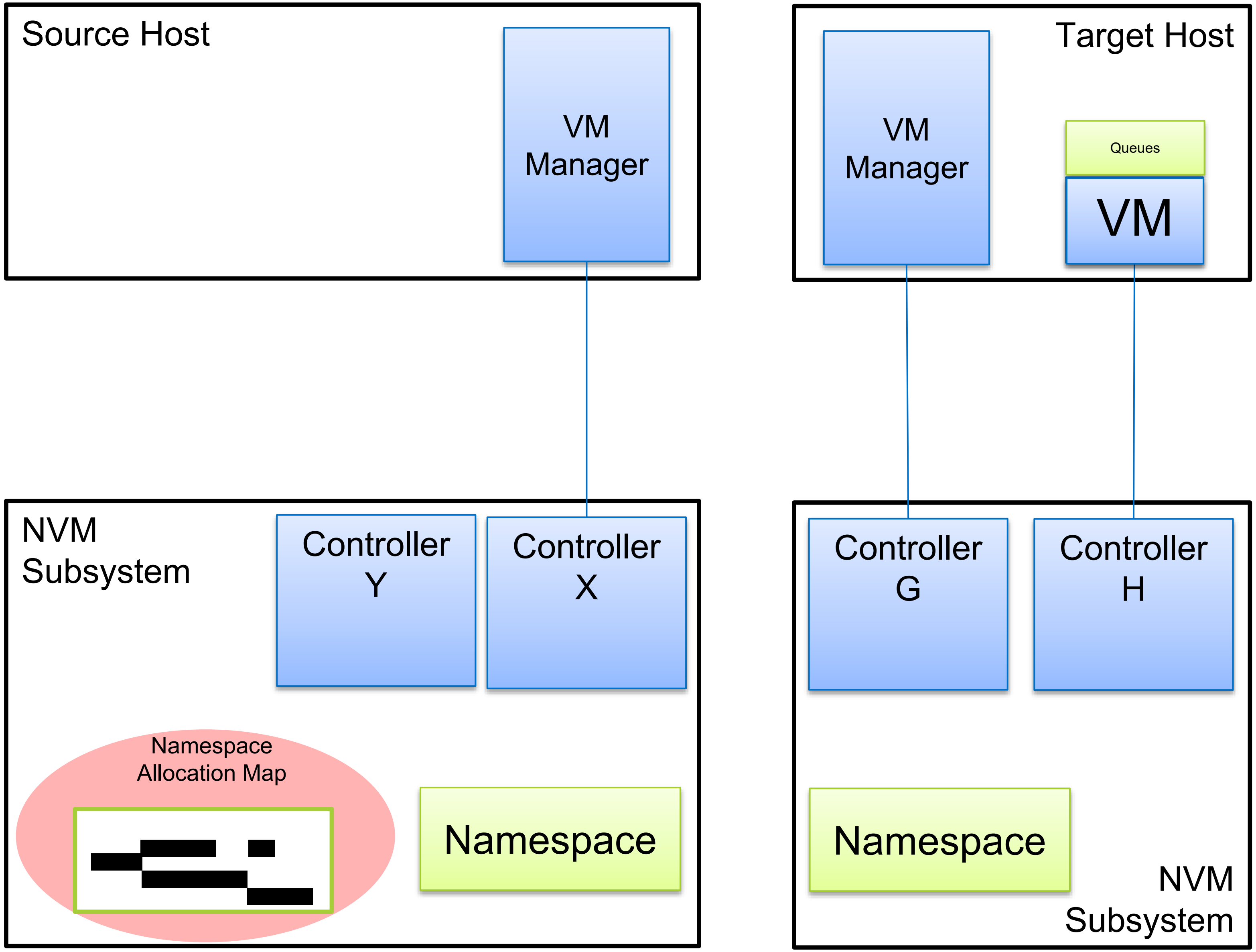
Post-copy Phase – Resuming Migrated Controller State



Target Admin Host

- Resume VM
- Issues a command to resume controller that was migrated

Post-copy Phase – Resuming Migrated Controller State



Target Admin Host

- Resume VM
- Issues a command to resume controller that was migrated

Source Admin Host

- Remove VM
- Reset the migrated controller

Building the Pieces

- TP4165 Tracking LBA Allocation with Granularity
  - Reporting of allocated LBAs within a namespace for migrating a namespace
  - Usable in Snapshot use cases
- TP4159 PCIe® Infrastructure for Live Migration
  - Developing the theory of operation
- A TPAR to:
  - Support limit the BW and IOPS of a controller to allow slowing down of command processing on a migrating controller

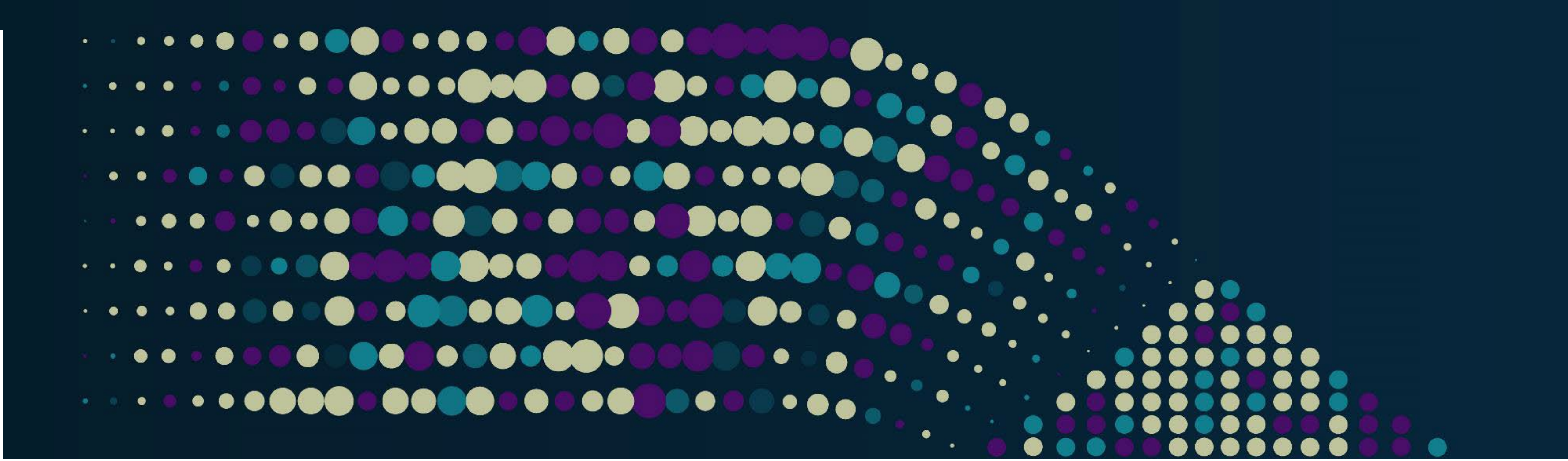
The Union is Strong and Delivering Value!

[redacted]

[redacted]

nvm EXPRESS®

Architected for Performance



Please take a moment to rate this session.

Your feedback is important to us.