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

Mike Allison
Sr. Director NAND Product Planning
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

3 ©2023 SNIA. All Rights Reserved.
NVMe® Technology Powers the Connected Universe

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

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
Organizational **Enhancements**

**Tooling Updates**
*Zoom, Causeway, Bugzilla*

**Errata Taskgroup**

**Website Redesign**

**Software Taskgroup**

**Framework**
Modernizing the NVM Express Website

Refreshed pages

Updated user interface

Consolidated & reorganized

Specifications

Blogs

Webinars

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVMe Express Base Specification</td>
<td>NVMe Base</td>
</tr>
<tr>
<td>NVMe Zoned Namespaces (ZNS) Command Set Specification</td>
<td>Command Set</td>
</tr>
<tr>
<td>NVM Command Set Specification</td>
<td>Command Set</td>
</tr>
<tr>
<td>Key-Value Command Set Specification</td>
<td>Command Set</td>
</tr>
<tr>
<td>RDMA Transport Specification</td>
<td>Transport</td>
</tr>
<tr>
<td>TCP Transport Specification</td>
<td>Transport</td>
</tr>
<tr>
<td>NVMe over PCIe Transport Specification</td>
<td>Transport</td>
</tr>
<tr>
<td>NVM Express Management Interface Specification</td>
<td>NVMe-MI™</td>
</tr>
<tr>
<td>NVMe Boot Specification</td>
<td>Boot</td>
</tr>
<tr>
<td>Changes in NVMe Express Revision 2.0</td>
<td>Command Set, NVMe Base, NVMe-MI™, Transport</td>
</tr>
<tr>
<td>NVMe over Fabrics IOE Specification (historical reference only)</td>
<td>Historical Reference</td>
</tr>
</tbody>
</table>
Resources to Learn About NVMe® Technology
### NVMe® 2.0 Family of Specifications

<table>
<thead>
<tr>
<th>Category</th>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NVMe Base Specification</strong></td>
<td>NVMe NVM Command Set Specification</td>
</tr>
<tr>
<td><strong>Command Set Specifications</strong></td>
<td>NVMe Zoned Namespace Command Set Specification</td>
</tr>
<tr>
<td></td>
<td>NVMe Key Value Command Set Specification</td>
</tr>
<tr>
<td><strong>Transport Specifications</strong></td>
<td>NVMe over PCIe Transport Specification</td>
</tr>
<tr>
<td></td>
<td>NVMe over RDMA Transport Specification</td>
</tr>
<tr>
<td></td>
<td>NVMe over TCP Transport Specification</td>
</tr>
<tr>
<td><strong>NVMe Management Interface</strong></td>
<td>Network Boot / UEFI Specification</td>
</tr>
</tbody>
</table>

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: 60
Ratified Technical Proposals: 69
Ratified ECNs: 13

* Activity as of 7/28/2023
NVMe® Specifications Feature Roadmap

- **2021**
  - Q2: NVMe-oF™ Automated Discovery
  - Q3: Dispersed Namespaces
  - Q4: Scalable Resource Management

- **2022**
  - Q1: Network Boot / UEFI
  - Q2: Cross Namespace Copy
  - Q3: Flexible Data Placement (FDP)
  - Q4: Key Per I/O

- **2023**
  - Q1: Computational Programs
  - Q2: Subsystem Local Memory
  - Q3: Live Migration
  - Q4: Cross Namespace Copy

- **2024**
  - Q1: Computational Programs
  - Q2: Subsystem Local Memory
  - Q3: Live Migration
  - Q4: Cross Namespace Copy

**NVM Express®**

- **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.

Source Host

- Admin Host
  - Get LBA Status
  - VM
  - Queues

NVM Subsystem

- Controller Y
- Controller X

Migration Queue

Target Host

- Admin Host
  - VM

NVM Subsystem

- Controller G
- Controller H

Namespace Allocation Map

Namespace

Namespace

NVM Subsystem
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 copying the allocated LBAs to the destination, the Source Admin Host may migrate the dirty LBAs.
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

Target Host

Source Host

- Completes migration of VM

Target Host

Stop-and-Copy Phase – Finish Migrating

Source Host

- Completes migration of VM

Target Host

Stop-and-Copy Phase – Finish Migrating

Source Host

- Completes migration of VM

Target Host
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

Source Host
- Admin Host
- Controller Y
- Controller X
- VM
- Get State

Target Host
- Admin Host
- Controller G
- Controller H
- VM
- Get State
- Set State

Namespace Allocation Map

NVM Subsystem

Source Host

Target Host

Queues

Namespace

NVM Subsystem
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!
Please take a moment to rate this session.

Your feedback is important to us.