Managing Capacity in NVM Express SSDs

Paul Suhler
Micron Technology

Mark Carlson
Toshiba Memory
Abstract

- Customers of SSDs have very different requirements. While a single uniform space of non-volatile storage media may satisfy some, others require independent areas with performance isolation. Still others may require a small amount of low-latency storage and a much larger amount of higher-latency storage. To enable a single SSD type to be configured by the customer to satisfy their particular requirements, NVM Express has defined a mechanism to configure non-volatile media in not just SSDs but also in storage arrays.

- Learning Objectives:
  1. Attendees will better understand NVMe Capacity Management
  2. Attendees will learn some of the use cases and requirements coming from Hyperscalers for SSDs
  3. Attendees will understand how SSDs can accommodate these new requirements
Speakers

Paul Suhler

Mark Carlson

Toshiba Memory
Agenda

- NVMe Capacity Entities
- SSD Organizations (Use Cases)
- Management Methods
Use cases:

- More flexible IO Determinism
  - SSD vendors currently ship static configurations
- Address the need to divide work between host and drive
- Enable Endurance Groups for Storage Systems
  - Capacity management

Goal:

- Enable one SKU to be configured by customer for their use case
NVMe Capacity Entity Hierarchy

- **Namespaces** – Contain an array of logical blocks
- **NVM Sets** – Contain namespaces
- **Endurance Groups** – Contain NVM Sets
- **Domains** – Contain Endurance Groups, controllers, etc.
Original NVMe SSD Organization

- Media Units (e.g., dies) are connected to the controller by channels.
- Endurance is managed across all Media Units.
IO Determinism Use Case

- Need to create NVM Sets according to their capacity requirements (e.g., 1 TB sets)
  - Typically once at the beginning of drive life
  - Supported Media Unit configurations are available indicating NVM Sets formed from Media Units along channels for isolation
  - Drive may only support two configurations (e.g. ½ TB and 1 TB sets) for this market

![Diagram of NVM Subsystem (SSD) with four separate, isolated NVM Sets]
Host Managed Media Users

Need to closely manage placement of data and accommodate append behavior

- Big concerns about Write Amplification and managing wear
- 1 EG / 1 NVM Set / 1 MU
- No predictable latency
- RAW UBER
- 1 Namespace / MU

Get Log Page – Media Unit Status
Mixed-Mode NAND Operation

- NAND cells allow operation at a maximum number of bits per cell (e.g., QLC), as well as at a smaller number (e.g., TLC, SLC).
- Different Endurance Groups can have different bits per cell.
- One SSD can use some Media Units for a small amount of fast capacity and the remaining Media Units at a much higher density.
Storage Systems Users

• Need to create, resize and delete Endurance Groups within a Domain
  • No need to implement Media Unit Management
  • Primarily tied to domains and partitions (TP 4009) - in review

• Variable Capacity Management
  • Capacity is drawn from the Domain
  • NVM Set is created as well
  • Deletion of Endurance Group also Deletes NVM Set(s), Namespace(s)
Management Methods

• Two methods:
  • Fixed Capacity Management
  • Variable Capacity Management

• Fixed Capacity Management would be for drives.
  • The operation will select from a fixed set of complete configurations; the selected configuration typically will be for the lifetime of the NVM subsystem.
  • This should satisfy the requirements of Hyperscalers.
    • Incrementally configuring endurance groups / NVM sets will not be supported for this method (not needed). Changing the configuration after the media has been used will not be supported in this method. If a use case for this is found, it could be the basis for a future TP.
Capacity Management command

- **Operation primarily for SSDs:**
  - Select Capacity Configuration: Selects one of the supported configurations.

- **Operations primarily for Storage Arrays:**
  - Create Endurance Group: Creates an Endurance Group of a specified size in a Domain.
  - Delete Endurance Group: Deletes a specified Endurance Group and all its contents.
  - Create NVM Set: Creates an NVM Set of a specified size in an Endurance Group.
  - Delete NVM Set: Deletes a specified NVM Set and all its namespaces.
Capacity Configuration Descriptor

- Endurance Group Configuration Descriptors
  - Endurance Group Information
    - Capacity Adjustment Factor
    - Total Endurance Group Capacity
    - Spare Endurance Group Capacity
    - Endurance Estimate
  - NVM Set Identifier List
  - Channel Descriptor List
    - Media Units on each channel
Media Unit Status Descriptor

- Media Unit Identifier
- Domain Identifier
- Endurance Group Identifier
- NVM Set Identifier
- Capacity Adjustment Factor
- Available Spare
- Percentage Used
- Number of Channels
- Channel Identifiers: Channels attached to this Media Unit
Variable Capacity Management

- Variable Capacity Management is for systems to dynamically create Endurance Groups and NVM Sets.
- The operation specifies a capacity for Endurance Groups and NVM sets without understanding of the underlying Media Units.
Questions?