



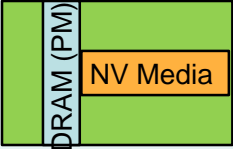

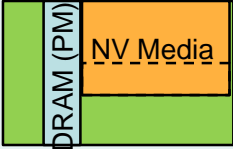
# Reliability, Availability, Serviceability (RAS) and Management for Non-Volatile Memory Storage

Mohan J. Kumar, Intel Corp  
Sammy Nachimuthu, Intel Corp  
Dimitris Ziakas, Intel Corp

# Agenda

- NVDIMM Types
- Storage vs. Memory Characteristics
- Storage vs. NVDIMM Characteristics
- NVDIMM expectation for storage use
- Storage RAS vs. Memory RAS
- Storage RAS vs. NVDIMM RAS
- Storage Management vs. Memory Management (DRAM)
- Storage Management vs. NVDIMM Management
- Summary

# NVDIMM Types

NVDIMM-N	NVDIMM-F	NVDIMM-P
		
<ul style="list-style-type: none"> <li>• Only DRAM is addressable by SW</li> <li>• NV Media acts as backup for DRAM</li> <li>• NV Media not addressable</li> <li>• At least 1:1 Capacity Ratio between DRAM &amp; NV Media</li> <li>• Tracks DRAM latency &amp; memory channel BW for Read and Write</li> </ul>	<ul style="list-style-type: none"> <li>• No DRAM</li> <li>• NV Media is directly addressable via Window mechanism</li> <li>• Tracks NV Media latency</li> <li>• Benefits from memory channel bandwidth</li> </ul>	<ul style="list-style-type: none"> <li>• Combination of NVDIMM-N and NVDIMM-F</li> <li>• Flash memory beyond that needed for persistence is accessible as block</li> </ul>



# Storage vs. Memory Characteristics

Attribute	Storage (e.g. PCIe)	Memory (e.g. DRAM)
<b>Unit of Access</b>	Block	Cacheline
<b>Latency for unit access</b>	~ uS to ms	~ of 10s of ns
<b>Bandwidth</b>	IO Channel Bandwidth (e.g. PCIe Gen3 x4 – 4GB/s)	Memory Channel Bandwidth (e.g. DDR4 2133 – 17.1GB/s)
<b>Interleave for higher perf using HW</b>	No	Yes
<b>Data Access</b>	Controller Mediated	Not Controller Mediated
<b>Application Access (Typical)</b>	Kernel Mediated	No Kernel mediation
<b>Access (Typical)</b>	DMA	Direct Access
<b>Expandability</b>	Yes (e.g. via switches)	No
<b>Attach</b>	IO Channel (e.g. PCIe)	Memory Channel



# Storage vs. NVDIMM Characteristics

Attribute	Storage (e.g. PCIe)	NVDIMM-N	NVDIMM-F
Unit of Access	Block	cacheline	Block
Latency for unit access	~ uS to ms	~ of 10s of ns	~ uS
Bandwidth	IO Channel Bandwidth (e.g. PCIe Gen3 x4 – 4GB/s)	Memory Channel Bandwidth (e.g. DDR4 2133 – 17.1GB/s)	~ Memory Channel Bandwidth
Interleave for higher perf using HW	No	Yes	No
Data Access	Controller Mediated	Not Controller Mediated	Controller Mediated
Application Access (Typical)	Kernel Mediated	No Kernel mediation	Kernel mediated
Access (Typical)	DMA	Direct Access	Direct access
Channel Expandability	Yes (e.g. via switches)	No	No
Attach	IO Channel (e.g. PCIe)	Memory Channel	Memory Channel

# NVDIMM expectation for storage use

- Data errors do not bring down the system
- Access to health information at boot and runtime
- Serviceable
- Data at rest security
- Ease of Migration (in spite of interleave)

# Storage vs. Memory Characteristics

## RAS

Attribute	Storage (e.g. PCIe)	Memory (e.g. DRAM)
Write Durability	Yes	NA (Volatile Media, not storage)
Impact of Error	Application wide	System wide*
Wear Level Management	Yes	No
Platform Single Point of Failure (SPOF) Protection	Multi-ported storage	No
Data Protection	Device and Controller level (ECC, RAID ...)	Device and Controller Level (ECC, PPR**, Platform Memory RAS)
Error Detection Granularity	Block	Cacheline
Write Cycle Limit	Yes	No

# Storage RAS

- Data Protection (device and software level)
- RAID Support - various RAID levels, Software vs. HW RAID
- Hot add/remove of Storage Device
- Health and Predictive Failure Reporting – SMART
- Multi-port capability (to overcome Platform as SPOF)





# Memory RAS

- ECC protected Memory
- SDDC (Single Device Data Correction)
- Data Scrub – Patrol, On Demand
- DIMM Sparing
- Memory Mirroring
- Poison
- Memory error recovery using MCA Recovery
- Hot add of Memory
- Memory Migration

# Storage vs. NVDIMM-N RAS

Attribute	Storage (e.g. PCIe)	NVDIMM-N
<b>Write Durability</b>	Yes	Possible with ADR or PCOMMIT
<b>Impact of Error</b>	Application wide	System wide**
<b>Wear Level Management</b>	Yes	NVDIMM controller implemented wear levelling
<b>Platform SPOF Protection</b>	Multi-ported storage	No
<b>Data Protection</b>	Device and Controller level (ECC, RAID ...)	Device and Controller Level (ECC, PPR, benefits from all Platform Memory RAS)
<b>Error Detection Granularity</b>	Block	Cacheline
<b>Write Cycle Limit</b>	Yes	No (Backing media only written on system fail conditions)

# Storage vs. NVDIMM-F RAS

Attribute	Storage (e.g. PCIe)	NVDIMM-F
<b>Write Durability</b>	Yes	Yes (by storage controller). Also benefits from PCOMMIT
<b>Impact of Error</b>	Application wide	System wide**
<b>Wear Level Management</b>	Yes	NVDIMM controller implemented wear levelling
<b>Platform SPOF Protection</b>	Multi-ported storage	No
<b>Data Protection</b>	Device and Controller level (ECC, RAID ...)	NVDIMM controller Level (does not benefit from Platform Memory RAS)
<b>Error Detection Granularity</b>	Block	Cacheline
<b>Write Cycle Limit</b>	Yes	Yes

# Storage vs. Memory Management

Attribute	Storage (e.g. PCIe)	Memory (e.g. DRAM)
Data at Rest Security	Capable of support	No
Health	SMART	NA (Platform test at each boot)
Serviceability	Typically does not require opening chassis	Requires opening the chassis
Migration across Platforms	Easy	Easy (Volatile Media)



# Storage Management

- Management integrated in the Storage Controller
- Exposed to software via SMART
- Out of band management interface is proprietary
- Data at rest security built-in
- Partition management and partition access protection implemented by storage controller
- Wear level management is required

# Memory Management

## Traditional View

- Typically, Platform Management software handles memory predictive failure
- OS-based memory predictive failure management at an address space level
- Otherwise, Memory Not managed entity
- Data at rest security does not apply (volatile memory)
- No need for wear level management



# Storage vs. NVDIMM-N Management

Attribute	Storage (e.g. PCIe)	NVDIMM-N
<b>Data at Rest Security</b>	Capable of support	No (controller mediation not possible since fronting media is DRAM)
<b>Health</b>	SMART	SMART
<b>Serviceability</b>	Typically does not require opening chassis	Requires opening the chassis
<b>Energy Source Management</b>	NA	Required
<b>Relative ease of Migration across Platforms</b>	Good	Difficult* (platform interleave impact)

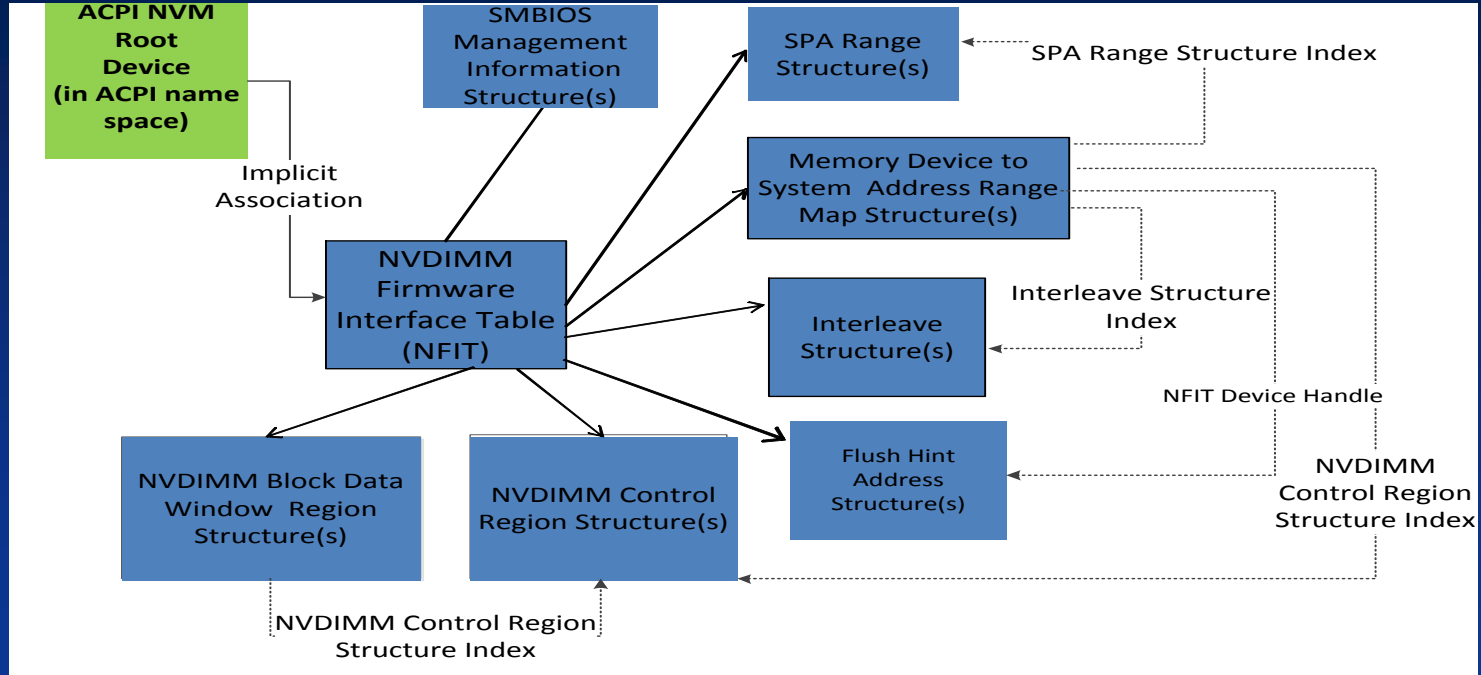


# Storage vs. NVDIMM-F Management

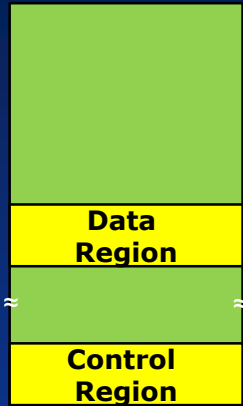
Attribute	Storage (e.g. PCIe)	NVDIMM-F
<b>Data at Rest Security</b>	Capable of support	Yes
<b>Health</b>	SMART	SMART
<b>Serviceability</b>	Typically does not require opening chassis	Requires opening the chassis
<b>Energy Source Management</b>	NA	NA
<b>Relative ease of Migration across Platforms</b>	Good	Medium* (platform interleave impact could be mitigated)



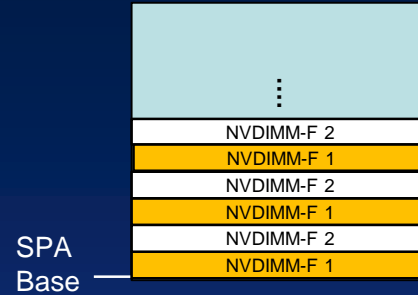
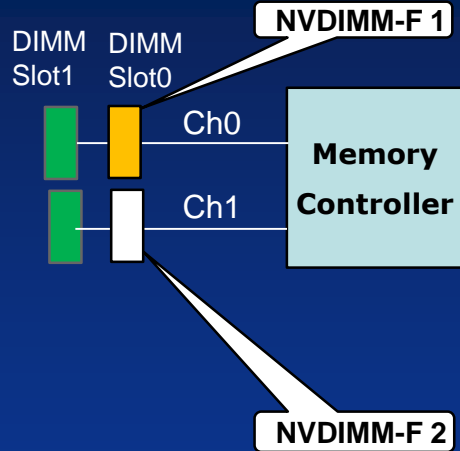
# NVDIMM Firmware Interface Table (NFIT)



# NVDMM-F Interleave - An example

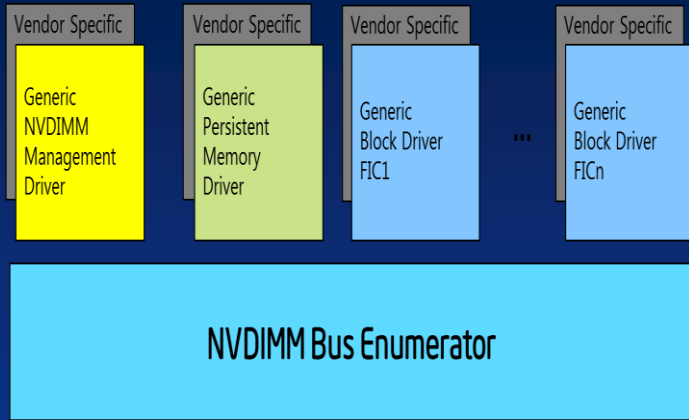


NVDIMM-F  
Device Space



- NFIT describes interleave across the channel to software
- NVDIMM-F control and data regions mapped in to system physical address (SPA) space
- Control and Data Regions exposed to drive via NFIT Tables
- NVDIMM-F also has SMBus control region for FW upgrades ...

# NVDIMM Management



**Exposing NVDIMM management to software**

- Management support integrated in the NVDIMM Controller
- Exposed to software via SMART
- Data at rest security is feasible for NVDIMM-F
- Wear level management is required for NV Media
- NVDIMM interleave management is required NVDIMMs depend on BIOS cooperation to access the data on subsequent boot (BIOS has to configure the interleaves and address space mapping identically across boots)
- For NVDIMM-N, platform has to monitor and manage the energy source (battery, supercap)
- Runtime management of NVDIMM exposed to software via ACPI \_DSM (Device Specific Method)

# Summary

- NVDIMM provides higher performance but requires solving the RAS and manageability to be on par with storage
- NVDIMM enumeration at hardware level is standardized by JEDEC
- Unlike Storage, NVDIMMs heavily dependent on Platform BIOS
- NVDIMM enumeration, configuration is now comprehended in firmware standards – E820, NFIT (Refer to [ACPI6.0 Specification](#))
- Runtime management of NVDIMM and vendor specific extension is abstracted via \_DSM (SMART, FW updates...)