The Evolution, Functionality and Benefits of NVDIMMs for Storage and Server Applications

Arthur Sainio
Bob Frey
June 30th, 2015
Agenda

- Memory / Storage Hierarchy
- Evolution of NVDIMMs
- Taxonomy
- NVM Trend/Adoption
- NVM Roadmap
- NVM Performance
- How NVDIMMs Work
- NVDIMM Features
- Applications
- Ecosystem and Industry Standardization
Memory Hierarchy

- Data-Intensive applications need fast access to storage
- Large performance gap between main memory and HDD
- SSDs have narrowed the gap, but a ~1000X gap still exists until a SCM becomes viable for mainstream adoption

Performance Gap

![Diagram showing the memory hierarchy and access times](image)

- CPU Cache: SRAM
- New Memory Technology: MRAM / ReRAM
- SSD
- HDD

ACCESS TIME (ns)

10^0 10^1 10^2 10^3 10^4 10^5 10^6

- Lowest Latency (Highest Cost): Load/Store (Application Direct Access)
- Highest Latency (Lowest Cost): Block (Application Indirect Access)

- NVDIMM
- NVRAM
- Nanotube, Other
- PCM
- STT-MRAM
- ReRAM

Adapted from SNIA presentations by Viking, HP
Storage Hierarchy – Persistent Memory

Legacy Storage

**Tier 0**
- Access to Host
- Storage Tier: Hot Data
- Latency: 0.01 µs

**Tier 1**
- Network Accessible
- Storage Tier: Hot Data
- Latency: 10 µs

**Tier 2**
- Mail Servers
- VOD Media Streaming
- Data Warehouse
- Backup
- Latency: 100 µs

**Tier 3**
- Archive
- Content Delivery
- Surveillance
- CRM
- OLTP
- Indexing
- Analytics
- Web Hosting
- Database
- OLTP
- Analytics
- In-memory
- Caching
- Indexing
- Latency: 1,000 µs

Leading Edge Storage

**NVDIMM**
- Memory Channel

**NVRAM**
- PCIe

SNIA presentations by Netlist
The Evolution of NVM

- **Origin**: to replace battery-backed for NVM cache on storage array controllers

- **NVRAM (Hybrid)**
  - PCIe interface, stores in DRAM, backs up into NAND only on a power loss, Supercap based, eliminates battery, maps in MMIO

- **NVDIMM (Hybrid)**
  - NVDIMM-N Stores in DRAM, backs up into NAND only on a power loss, Supercap based, eliminates battery
  - NVDIMM-F – maps NAND into memory address space
  - NVDIMM-P - maps NAND and DRAM into memory address space
  - And Beyond! FPGA on NVDIMM
    - Accelerate dynamically changing workloads

- **SCM – Storage Class Memories**
  - Next Gen memory technologies (ReRAM, STT-MRAM, PCM)
  - Inherently persistent and removes the supercap module
JEDEC NVDIMM Taxonomy

Single letter designator - combines the media technology (NAND, etc) and the access mechanism (byte, block, etc.)

**NVDIMM-N**
- One Access Method: direct byte-oriented access to DRAM and a persistence backup/restore function on power fail
- Memory mapped DRAM. No system access to Flash
- Capacity: DRAM DIMM (8GB, 16GB, 32GB) – uses DRAM and Flash
- Latency: DRAM (10’s of nanoseconds)
- Energy source for backup
- JEDEC type and electrical mechanical definition completed

**NVDIMM-F**
- One Access Method: block-oriented access through a shared command buffer, i.e. a mounted drive.
- Memory mapped Flash. DRAM is not system mapped
- Capacity: NAND (100’s GB – 1’s TB) - uses NAND flash; the command buffer may be DRAM, SRAM, etc.
- Latency: NAND (10’s of microseconds)
- JEDEC type definition completed

**NVDIMM-P**
- Two Access Methods: persistent DRAM (–N) and also block-oriented drive access (–F)
- Memory mapped Flash and memory mapped DRAM
- Supported -> Load/Store, Emulated Block
- Capacity: NVM (100’s GB – 1’s TB) – uses DRAM and NAND flash
- Latency: NVM (100’s of nanoseconds)
- No JEDEC definition yet
NVDIMM Market Trends

• NVDIMM market expected to reach $100’s of millions in 3-4 years
  • Multiple NVDIMM vendors

• NVDIMMs are used in server and storage applications (data centers and cloud computing)

• The market may further take off as the Software ecosystem matures
  • Industry work underway to develop the load/store stack

• The market may follow the PCIe trend starting 2018 with 3.3X growth, 2018-$0.6B to 2021-$5.5B

Source; IHS, Netlist, Mar 2015
NVM Adoption by System Application

RAID Card Application
NVM in RAID

CPU
Memory
Memory
Memory
Memory
Storage
NVDIMM
RAID Controller
PCIe Slot

PCle Application
NVM in PCIe

CPU
Memory
Memory
Memory
Memory
NVM
PCIe Slot

DIMM Application
NVM in Memory Channel

CPU
Memory
Memory
Memory
NVDIMM
Storage

Rack Scale Architecture

SNIA presentations by Netlist
NVM Roadmap – Enterprise Storage

Tick
Haswell / Grantley
1.2 V 2133 MT/s
16 GB DDR4 RDIMM

DD4 NVDIMM-N
1.2 V 2133 MT/s
8 GB - RDIMM

Tock
Grantley
1.2 V 2400 MT/s
16 GB DDR4 RDIMM

DD4 NVDIMM-N
1.2 V 2400 MT/s
16 GB - RDIMM

DD4 NVDIMM-F
1.2 V 2400 MT/s
400 GB - RDIMM

Tick
1.2 V 2667 MT/s
32 GB DDR4 R/LRDIMM

DDR4 NVDIMM-N
1.2 V 2DPC 2667 MT/s
32 GB Type N – R/LRDIMM

DDR4 NVDIMM-P
1.2 V 2667 MT/s
16 GB Type N + 400 GB Type F R/LRDIMM

NV-LRDIMM (Type-N, -F, -P)
JEDEC Standard

JEDEC Standard

Product Availability

DDR3 / DDR4
Cost Crossover

8Gb DDR4 2400 MT/s

Spec Closed
DDR4 NVDIMM SPD

Spec Closed
NV Register Set

Spec Closed
DDR4 Gen2 LRDIMM Chipset

Spec Closed
DDR4 Gen2 LRDIMM Raw Cards

8/16GB-

Validated
DDR4 Gen2 LRDIMM Raw Cards

32GB-L

SNIA presentations by Netlist
Performance Benchmark – NVDIMM- Type N DDR3 - IOPs & Response

- **Performance - IOPs**
  - Symmetric Read & Write Performance (1.6M IOPS)
  - Very low Average Response Times (0.02 mS)

- **Bandwidth – MB/s**
  - Very Fast Throughput (70GB/s Read / 21GB/s Write - Random)
  - Very Low Average Response Time (1.8 mS R, 5.8 mS W)

- **Response Time (Latency at single OIO or T1Q1)**
  - RND 4K 100% Write - .004 mS Average Response Time
  - RND 4K 100% Write - .120 mS Maximum Response Time

SNIA presentations by Netlist
**NVDIMM(-F) – How it Works**

- Form factor is RDIMM/LRDIMM JEDEC module
- Device model is “SSD on DDR bus”; that is, storage device not a memory device
  - Typically visible to system as block storage device
  - Flash capacity exposed directly to host, as in any SSD

Adapted from SNIA presentations by SanDisk
NVDIMM(-N) – How it Works

- RDIMM/LRDIMM JEDEC module with power backup
- Data stored into NAND on power loss to achieve persistence
- Host only addresses the DRAM and has no direct access to the flash
- NVDIMM controller moves data from DRAM to flash upon power loss or other trigger; can back up portions or all of DRAM upon command
- When power fails a backup power source provides power to the NVDIMM while DRAM is backed up to Flash
- MRC (Memory Reference Code) configures NVDIMM controller to move data back from Flash to DRAM when recovery is needed
NVDIMM(-P) – How it Works

- Form factor is RDIMM/LRDIMM JEDEC module
- Three Modes of Operation
  - Storage Mode: “SSD on DDR bus”; Storage device not a memory device, No application SW changes needed, BIOS & Driver required.
  - Memory Mode: DRAM Operation, Volatile
  - NVM Mode
  - Hybrid Mode: Combines NVM and Storage.
NVDIMM (−N) Features

Ultracapacitor Backup
- Provides power to move vital data contents from DRAM to flash during a system or power failure event
- Eliminates Battery
- Charged by NVDIMM via 12V power

Standard DDR4 RDIMM Interface

<table>
<thead>
<tr>
<th></th>
<th>DDR3</th>
<th>DDR4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>Up to 8GB</td>
<td>Up to 16GB</td>
</tr>
<tr>
<td>Power (V)</td>
<td>1.5 / 1.35</td>
<td>1.2</td>
</tr>
<tr>
<td>Performance</td>
<td>1DPC 1600 MT/s</td>
<td>1DPC 2400 MT/s</td>
</tr>
<tr>
<td></td>
<td>2DPC 1600 MT/s</td>
<td>2DPC 2133 MT/s</td>
</tr>
</tbody>
</table>

NV (Digital) Controller
- Facilitates backup/restore functionality
- Register access for temp monitoring, ultracapacitor, and flash statistics
NVDIMM –N Features

DDR3
- 4GB/8GB Capacity
- DDR3 1600
- DDR3 latency
- DDR3 Dual voltage 1.35V/1.5V operation
- Configuration: 1Rx8
- RDIMM Interface
- 12V Aux via a proprietary connection
- Availability - Now

DDR4
- 4GB/8GB/16GB Capacity
- DDR4 2400
- DDR4 latency
- DDR4 voltage 1.2V
- Configuration: 1R/2R
- 12V support via the DDR4 DIMM connector
- Availability - Now

ALL
- JEDEC Form Factor
- Support in Intel MRC
- Follows RDIMM Population Rules
- SLC or MLC NAND
- End-to-end Data integrity
- Backup initiated via SAVE Signal or I2C
  - OS SAVE as part of shutdown
  - SAVE during operation requires system re-initialization of DDR3 register & re-write DDR3 mode register
- Backup duration
  - 4GB ~ 16GB (Vendor Specific)
- Health Monitoring
  - Error reporting
  - Predictive failure warnings
  - Backup, restore, power, NAND life status
  - JEDEC or Proprietary
NVDIMM Stack

Adapted from SNIA presentations by AgigA Tech

[Diagram showing the NVDIMM Stack with various components and their interconnections, including Software, Application, Load Store, OS, BIOS, SMB, MRC + BIOS Modules, NVDIMM Driver, NVDIMM, Power Supply, Energy Module, CPU/Memory Controller, Power Supply, and Platform Hardware.]
Application Example: HP NVDIMM-N (DDR3) Acceleration – by HP

4X the Acceleration in Throughput and Latency

Source: SNIA 2/15
Application Example: Storage Bridge Bay (SBB)

Shadow Writes Required for Failover

Adapted from SNIA presentations by AgigA Tech
SBB: A Simpler/Better/Faster Way

Also a better alternative to Cache-to-Flash implementations:

- Separate failure domain
- No battery maintenance
- System hold-up requirements significantly less severe
- 4x write latency performance improvement

Adapted from SNIA presentations by AgigA Tech
Application Uses – Storage Main Memory & PCIe

- Store Metadata in memory for application acceleration
- Host Caching: As Cache for direct attached PCIe SSD
- Fast RAID Computation as block device for distributed storage
- Check-pointing state for fast sync and restore
- SSD Mapping Table
- Persistent RAM Disk
  - As Fast 4K block store
  - Store boot image for fast restore

<table>
<thead>
<tr>
<th>Persistent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata</td>
</tr>
<tr>
<td>Checkpoint State</td>
</tr>
<tr>
<td>Host Caching</td>
</tr>
<tr>
<td>RAMDisk</td>
</tr>
<tr>
<td>RAID Compute</td>
</tr>
<tr>
<td>Write buffer</td>
</tr>
<tr>
<td>SSD Mapping</td>
</tr>
<tr>
<td>Journaling</td>
</tr>
<tr>
<td>Logging</td>
</tr>
</tbody>
</table>

**Tier 0**
25ns
NVDIMM
Type 1 (N,F,P)

**Tier 1**
100us PCIe SSD
15us PCIe NVM

**Tier 2**
100us SATA/SAS
Advantages of NVDIMMs for Applications

Legacy HDD/SSD Solution
- Persistent data stored in HDD or SSD tiers
- Slow & unpredictable software stack

NVDIMM Solution
- Persistent data stored in fast DRAM tier
- Removes software stack from data-path

Accelerates SW-Apps!
- DRAM class latency & thru-put for persistent data
  - 1000X lower latency
  - 10X+ throughput increase
  - But, 10X lower capacity vs. SSD

• The value is in application acceleration
Low Write-Latency Persistent Storage

Value Proposition Specifics

- Accelerating Write-Latency
- Datacenter Workloads
- Transaction Commits, Logs & Journals

<table>
<thead>
<tr>
<th></th>
<th>400GB SLC PCIe SSD</th>
<th>8GB NVDIMM</th>
<th>Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write Latency (us)</td>
<td>&gt;15</td>
<td>&lt;0.1</td>
<td>&gt;150x</td>
</tr>
<tr>
<td>Read Latency (us)</td>
<td>&gt;47</td>
<td>&lt;0.1</td>
<td>&gt;470x</td>
</tr>
<tr>
<td>Endurance</td>
<td>Workload-dependent</td>
<td>Unlimited</td>
<td>Unlimited</td>
</tr>
<tr>
<td>PCIe Lanes Consumed</td>
<td>8</td>
<td>0</td>
<td>No PCIe used</td>
</tr>
<tr>
<td>Power (W)</td>
<td>&lt;25</td>
<td>&lt;10</td>
<td>&gt;2x</td>
</tr>
</tbody>
</table>

- >100x Lower Write Latency
- Unlimited Endurance
- Scalable
- No PCIe Resources Consumed
- Lower Total Cost of Ownership
  - $/Latency, $/IOPS & IOPS/W

Accelerating Write-Latency Datacenter Workloads

Transaction Commits, Logs & Journals

Adapted from SNIA presentations by Inphi
User Case - 5yr TCO Analysis
SATA/PCle vs. NVDIMM

Dell PowerEdge E-26xx CPU
DDR3 NVDIMM 1600 MTs
Write Intensive App

98% reduction
$2,723M saved

98% reduction
$695M saved

98% reduction
$3,418M saved

96% reduction
$1,771M saved

96% reduction
$1,424M saved

96% reduction
$347M saved

100,000 Servers
50,000 Servers
1,000 Servers

CAPEX
OPEX
TOTAL

CAPEX
OPEX
TOTAL

98%
96%

reduction
reduction

$0M
$1,000M
$2,000M
$3,000M
$4,000M

$0M
$1,000M
$2,000M
$3,000M
$4,000M

$0M
$1,000M
$2,000M
$3,000M
$4,000M

$0M
$1,000M
$2,000M
$3,000M
$4,000M
NVDIMM(-N) Ecosystem

Hardware Standardization
- System management, Power health
- System support H/W trigger (ADR)
- Mechanical (power source)
- JEDEC NVDIMM

Platform Support
- Off-the-shelf and OEM platform support for NVDIMM today
- System supported H/W trigger (ADR)
- Mechanical (power source)

BIOS Support
- NVDIMM-aware BIOS
- Intel modifications to MRC to support NVDIMMs
- JEDEC NVDIMM I2C command set
- JEDEC SPD

New Applications

Software Standardization
- Applications
- Linux NVDIMM-aware kernel 4.1
- API’s

Mass Deployment
NVDIMM(-N) Standardization

**JEDEC Hybrid Memory Task Group**
- DDR4 12V Power Pins (1, 145) standardized
- DDR4 SAVE_n Pin (230) standardized
- 12V in DDR4 socket will simplify NVDIMM power circuitry and cable routing
- Under discussion: Standard system interface for NVDIMM
  - i2c register map for NVDIMM-N controller
  - Issued ballot on NVDIMM Controller Event Pin
  - SPD for NVDIMM representation

**SNIA NVDIMM SIG**
(Special Interest Group, >20 companies)
- Formed in 2014 as a SIG of the SNIA Solid State Storage Initiative
- Communicating existing industry standards, and areas for vendor differentiation
- Helping technology and solution vendors whose products integrate NVDIMMs to communicate their benefits and value to the greater market
- Developing vendor-agnostic user perspective case studies, best practices, and vertical industry requirements

Flow of NVDIMM Adoption and Support:
- Standards JEDEC
- BIOS
- NVDIMM Suppliers
- Motherboard ODMs, OEMs
- Platform Integrated Solutions

Platform Integrated Solutions
Eliminate File System Latency with Memory Mapped Files

Traditional

User
Application

Kernel
File System
Disk Driver

HW
Disk

New

User
Application

Kernel
Memory Mapped Files

HW
Persistent Memory

Load/Store

Use with disk-like NVM

**NVM.BLOCK Mode**
- Targeted for file systems and block-aware applications
- Atomic writes
- Length and alignment granularities
- Thin provisioning management

**NVM.FILE Mode**
- Targeted for file based apps.
- Discovery and use of atomic write features
- Discovery of granularities
Use with memory-like NVM

**NVM.PM.VOLUME Mode**
- Software abstraction to OS components for Persistent Memory (PM) hardware
- List of physical address ranges for each PM volume
- Thin provisioning management

**NVM.PM.FILE Mode**
- Describes the behavior for applications accessing persistent memory Discovery and use of atomic write features
- Mapping PM files (or subsets of files) to virtual memory addresses
- Syncing portions of PM files to the persistence domain
Building on the Basic PM Model

- NVM.PM.FILE programming model “surfaces” PM to application
- Refine API with additional libraries that evolve into language extensions
- Add compatible functionality to PM file systems
## Open Source Sub-initiatives

<table>
<thead>
<tr>
<th>Initiative</th>
<th>What will it do?</th>
<th>What is the status?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong> NVDIMM Aware Linux Kernel Support for NVDIMM-N modules (arch/x86/kernel/e820.c)</td>
<td>Motherboard vendors provide BIOS/MRC changes needed to recognize NVDIMM-N modules. Customers still need either Block or Load/Store Linux driver to enable NVDIMM-N modules.</td>
<td>Waiting for the ACPI Spec 6.0 to be published. Availability in Github - April/2015.</td>
</tr>
<tr>
<td><strong>2</strong> Block &amp; Load/Store Drivers</td>
<td>The load/store driver will likely come after the block driver into the Linux Kernel.</td>
<td>The PMEM initiative is awaiting ACPI related changes to be approved. General Availability in Linux Kernel - TBD</td>
</tr>
<tr>
<td><strong>3</strong> DAX Support</td>
<td>The Ext4 file system DAX support to NVDIMM-N modules eliminates the page cache layer completely. This requires the availability of [2] &amp; [3].</td>
<td>Q2, 2015, in the official 3.20 kernel</td>
</tr>
</tbody>
</table>

**Git:** [https://github.com/01org/prd](https://github.com/01org/prd)
Thank You!