Hewlett Packard Enterprise

HPE Persistent Memory for HPE ProLiant Servers

The performance of memory with the persistence of storage

Technical white paper

Contents

Abstract

Businesses today need systems that provide faster access to critical and frequently used data. HPE ProLiant servers strive to remain on the cutting edge of technologies that accelerate performance, guard against data loss, and minimize downtime. This paper describes HPE Persistent Memory, a new technology that offers a high performance compute solution for HPE servers and discusses the benefits and limitations of the technology for IT administrators.

Introduction

Today's businesses demand real-time data to realize faster business outcomes. These businesses need systems that offer uncompromising performance and workload-optimized applications—systems that make data available as quickly and reliably as possible. For decades, the computing industry has been researching ways to reduce the performance gap between the low-latency processor and higher (longer) latency storage devices. The progression of storage technology has thus evolved both in how data is accessed (SAS/SATA-to-PCI bus), and how data is stored (magnetic-to-solid state media).

A recent trend has emerged of moving storage functionality to the memory bus, thus taking advantage of that interconnect's low latency and fast performance. Placing storage devices on the memory bus offers something more—the prospect of byte-addressable storage, a new semantic that cuts through cumbersome software layers and offers sub-microsecond device latencies.

HPE is introducing Persistent Memory, a new category of server-storage devices that reside on the server memory bus. HPE Persistent Memory delivers the performance of standard memory but with the added persistence of traditional server-storage. Persistent memory devices fall into two subcategories: performance-optimized persistent memory and capacity-optimized persistent memory, both compared in Table 1.

Table 1. Comparison of Performance-optimized and Capacity-optimized Persistent Memory.

ATTRIBUTE	PERFORMANCE-OPTIMIZED PERSISTENT MEMORY	CAPACITY-OPTIMIZED PERSISTENT MEMORY
Byte Addressable	Yes	Yes
Persistent	Yes	Yes
Latencies Under 100ns	Yes	No
Capacity over 100 GiB	No	Yes
Endurance	Long life	Limited

The HPE 8GB Non-Volatile DIMM (NVDIMM) Single Rank x4 DDR4-2133 Module is a performance-optimized offering and the first product in HPE's Persistent Memory category.. These NVDIMMs operate as high performance DRAM-based memory during normal operation but use NAND flash memory to retain data in the event of power failure or system crash. When the system reboots, HPE NVDIMMs restore the data to its previous availability state. HPE NVDIMMs deliver breakthrough performance for real world applications.

The HPE NVDIMM memory module combines 8GB of DRAM and 8GB of flash in a single module that fits in a standard server DIMM slot. DRAM operates at high speed but it is relatively expensive, and if a server shuts down unexpectedly, any data in DRAM is lost. Flash is slower but its nonvolatile, meaning it retains data when the power source is removed.

Server-based storage can often be described using a tiered hierarchy, largely defined by data availability requirements. A typical server-based storage hierarchy has the top tier (Tier 0) using workload accelerators for caching and accelerating the most frequently used data. Workload accelerators typically use NAND flash technology and operate off the PCI bus. Subsequent tiers involve slower, more traditional storage media operating through SAS and SATA interfaces.

HPE NVDIMMs, with their high level of performance, become the new top tier in the server-storage hierarchy, handling the most frequently accessed data (Figure 1). HPE NVDIMMs do not need to replace existing server-storage devices such as workload accelerators but complement their functionality. The benefit of NVDIMMs is two-fold: access to the most frequently used data is substantially faster, and since many of the writes occur with DRAM-based NVDIMMs, the endurance of NAND flash-based workload accelerators and SSDs is increased.

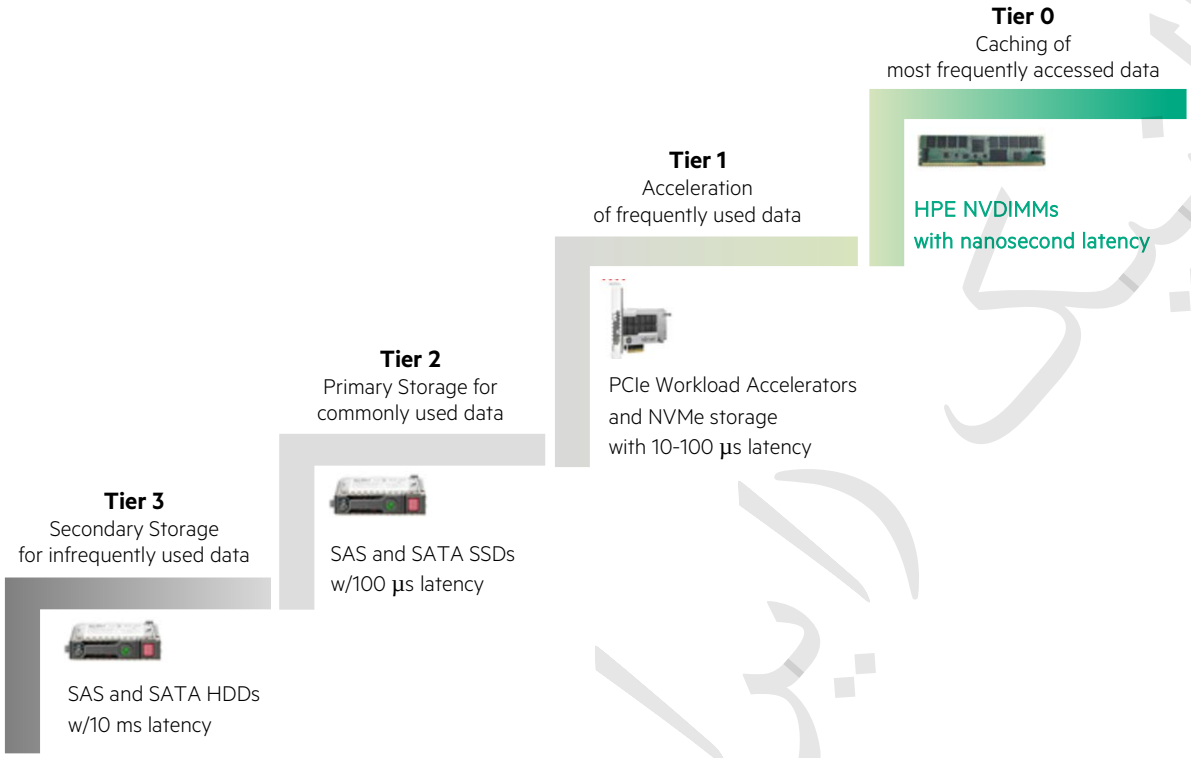

Figure 1. Server-storage hierarchy with HPE NVDIMMs.

At this time, HPE plans to offer HPE NVDIMMs as an option for the HPE ProLiant DL380 Gen9 and DL360 Gen9 Servers.

HPE NVDIMM overview

The HPE 8GB NVDIMM Single Rank x4 DDR4-2133 Module is the first fully integrated Performant Persistent Memory solution for ProLiant servers. This memory module has many of the same attributes of emerging memory media, but is built with trusted components based on very mature technologies.

NVDIMM architecture

The HPE NVDIMM combines standard DRAM-based system memory with NAND flash-based Persistency Components to achieve both high performance and power-off storage persistency (Figure 2). HPE NVDIMMs install in HPE server DDR4 memory sockets that include the I^2C bus, the Asynchronous DRAM Refresh (ADR) signal, and both main and backup (BU) power busses.

HPE NVDIMMs meets the JEDEC standard for NVDIMM-N devices. As a DRAM-based storage media, the HPE NVDIMM offers both a byte-addressable and block storage interface. As a backup storage device, the HPE NVDIMM uses NAND flash-based components in backup situations.

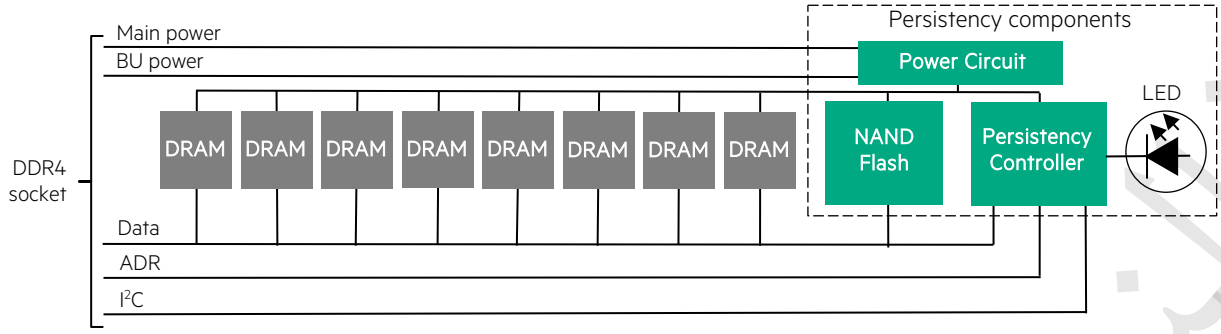

Figure 2. NVDIMM architecture.

Normal operation

In normal operation (Figure 3), the NVDIMM uses high speed DRAM for servicing all data accesses, and operates like a standard system memory DIMM.

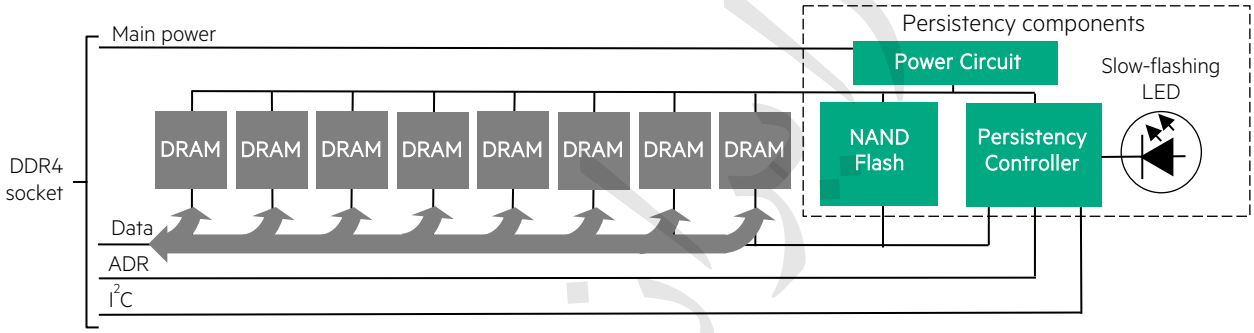

Figure 3. Normal operation of NVDIMM.

During normal operation, the Persistency Controller constantly monitors the I²C bus and ADR signal for notification of events that threaten the data in DRAM, such as:

- Sudden power-loss
- Shutdown/restart initiated from the OS
- Catastrophic system errors
- Operating system crashes

Notification of any of these events will result in the persistency controller initiating a backup operation. The notification functionality required by the persistency controller is available only on specific HPE ProLiant platforms that include the Complex Programmable Logic Device (CPLD) and an Intel® Broadwell series processor. These components use the ADR signal, which indicates that the processor should flush the write-protected data buffers and place DRAM into self-refresh.

Note
NVDIMM functionality requires HPE ProLiant platforms that include the CPLD and Intel Broadwell processors. Without these components, HPE NVDIMMs will not function properly. Refer to the section NVDIMM deployment for HPE platforms and operating systems that offer NVDIMM functionality.

Backup operation

The backup operation begins as soon as the persistency controller receives notification of a backup trigger event. The persistency controller systematically transfers all the contents within the volatile DRAM to the onboard NAND flash device (Figure 4). The transfer operation can take over a minute to complete. During this time, LEDs indicate that the backup operation is in progress and that NVDIMMs should not be removed from the system.

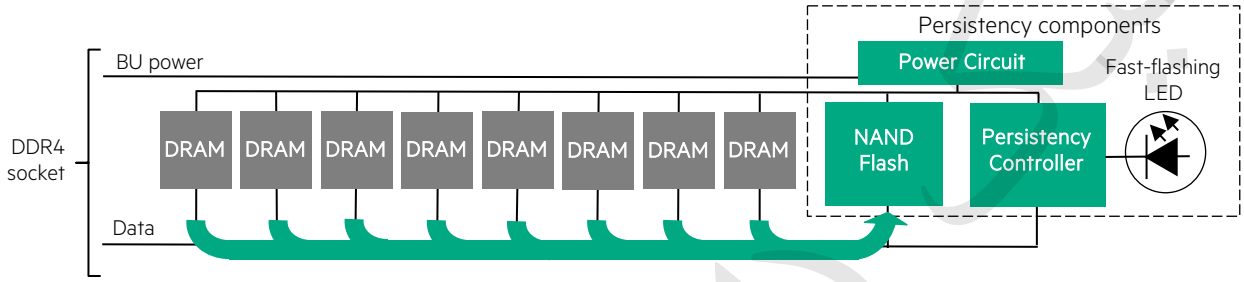

Figure 4. Backup operation of NVDIMM.

It is assumed the system may lose AC power during the backup operation. Since a power source is required to keep the NVDIMM operating during backup, the Smart Storage Battery used to power Smart Array cache modules also provides BU power to the DIMM slots for the NVDIMMs. Up to 16 NVDIMMs can be installed in a system (depending on Smart Storage Battery capacity) and receive backup power without the need for auxiliary cabling or more risky supercap technology.

Restore operation

The restore operation (Figure 5) is the reverse of the backup operation. As a system with NVDIMMs boots, the following operations occur:

- The persistency controller moves the contents of the Flash device back to the DRAM.
- The memory is scanned for any errors that may have occurred during the backup operation.
- The ProLiant System BIOS measures the charge level of the Smart Storage Battery to ensure that it contains sufficient charge to perform a backup of all NVDIMMs in the system.

The system boot process does not complete until both the DRAM contents are restored and the Smart Storage Battery contains sufficient charge to handle a subsequent power loss. If there are any errors in this process, notifications are sent as POST messages and logged in the IML Log.

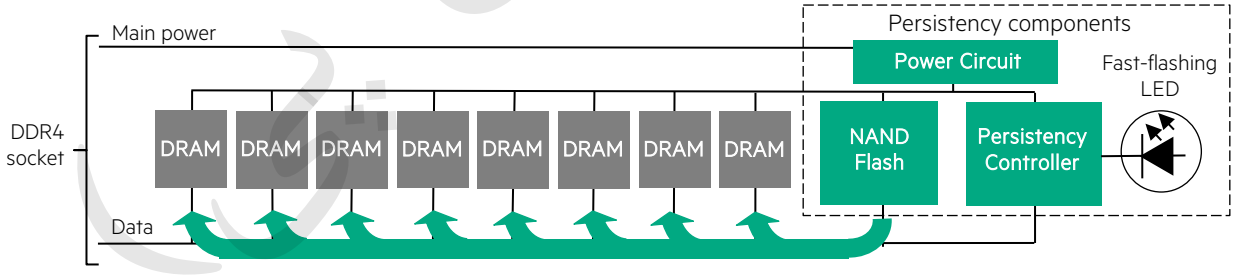

Figure 5. Restore operation of NVDIMM.

HPE NVDIMM deployment

HPE Persistent Memory is not just new hardware technology but a complete hardware/software ecosystem designed to work with today's applications and workloads. The hardware contingent includes HPE ProLiant servers optimized for HPE NVDIMM functionality through the inclusion of the Complex Programmable Logic Device (CPLD), the Intel Broadwell processor, and the Smart Storage Battery Backup system. The software contingent includes the server firmware, operating system drivers and software applications. We have fine-tuned HPE ProLiant server BIOS and Integrated Lights Out (iLO) firmware for Persistent Memory. We have also worked with industry-standard operating system providers to deliver drivers that natively support HPE Persistent Memory. The following sections define the hardware and software requirements for deploying HPE NVDIMMs.

Hardware/firmware requirements

HPE Persistent Memory can be deployed on the following HPE systems:

- HPE ProLiant DL380 Gen9 Server factory-configured with Intel Broadwell (Xeon® E5-2600 v4) processor(s)
- HPE ProLiant DL360 Gen9 Server factory-configured with Intel Broadwell (Xeon E5-2600 v4) processor(s)
- System with HPE Smart Storage Battery
- System with a minimum of one registered DIMM (RDIMM) per processor
- HPE System ROMPaq firmware v2.40 or later
- HPE iLO 4 Firmware v2.40 or later

Operating system requirements

The following sections identify operating systems that support HPE Persistent Memory.

Microsoft Windows Server 2012 R2 drivers

The Windows Server 2012 R2 out-of-box driver allows individual NVDIMM devices and the Host hardware interleaving of NVDIMM devices. Windows Server components like Storage Spaces can be used on NVDIMM devices. NVDIMM devices can be grouped together and create larger simple or mirrored SCM volumes.

The initial HPE Persistent Memory enablement offered by HPE Windows Server 2012 R2 drivers allows HPE NVDIMMs to be implemented two different ways:

- With a block-interface overlay, using SCM as block-storage devices applications can readily use no differently than SATA HDDs or SAS SSDs.
- With a byte-addressable memory interface, allowing applications to directly access physical memory locations on the SCM device as a "Memory Mapped Interface," enabled on a per-volume basis or interleave access for multiple NVDIMMs

For more information refer to the white paper *Deploying HPE Persistent Memory on HPE ProLiant servers running Microsoft Windows Server 2012 R2*.

Microsoft Windows Server 2016 Technical Preview 5

The initial HPE Persistent Memory enablement offered by Windows Server 2016 TP5 inbox driver allows individual (not interleaved) NVDIMM devices. Windows Server components such as Storage Spaces can group NVDIMM devices together and create larger simple or mirrored storage volumes. Windows Server 2016 allows volumes to be formatted for traditional block mode or for Direct Access Storage (DAX), a new byte-addressable mode that maximizes performance.

For more information, refer to the white paper *Deploying HPE persistent memory on HPE ProLiant servers running Microsoft Windows Server 2016 TP5*.

Linux

Support for NVDIMMs is available with the HPE Linux® Software Development Kit (SDK). The Linux SDK is designed for application providers and developers incorporating HPE 8GB NVDIMM technology into Linux applications. We are working with many major distributions to begin adding persistent memory functionality to the kernel. Updates on support in future Linux distributions will be announced on hp.com.

Application requirements

Applications will see immediate performance benefits utilizing HPE Persistent Memory as a very fast block storage device. However, the real performance benefits are unlocked once applications are updated to use HPE Persistent Memory in a byte-addressable (load/store) manner. This will translate into more efficient software application code and reduced latency that unlock the real performance potential of HPE Persistent Memory. Applications that would otherwise need significant time to save or restore large amounts of data during system crashes or power failures can now recall data as soon as the system is rebooted.

NVDIMM as a Block Device

The easiest way to deploy NVDIMMs is as a block device. Table 2 illustrates the extreme NVDIMM performance by comparing the 8 GiB NVDIMM device to the fastest SSD and PCIe Workload Accelerator cards in the HPE portfolio: Note that these numbers do scale when additional devices are added to the system.

Table 2. Comparison of NVDIMM performance to SAS SSDs and PCIe workload accelerators.

PARAMETER	NVDIMM VERSUS SAS SSD[1]	NVDIMM VERSUS PCIE WORKLOAD ACCELERATOR[2]
IOs Per Second (IOPS)	34x more IOPS	24x more IOPS
Bandwidth	16x more throughput	6x more throughput
Latency	81x lower latency	73x lower latency

Another way to deploy the HPE 8 GiB NVDIMM is to use it as a high-speed cache in front of a storage subsystem. As with any caching deployment, the expectation is that the performance of the subsystem will be somewhere between the cache's performance and the storage subsystem's performance. The effectiveness will vary based on the temporal/spatial locality of the application's access patterns. HPE is working to identify applications that have good access locality.

Persistent Memory as Byte-Addressable Storage

Block accesses have a lot of overhead in accessing storage. Figure 6 shows three scenarios. The left bar shows access to a NAND flash-based device. The green sections represent the time spent in the flash media, the grey the time spent in the media controller and on the PCI bus, and the black is time spent in software. The middle bar shows the time it takes to access an 8 GiB NVDIMM through the block layer. While significantly faster than the SSD, there is still a few microseconds of software overhead. The right bar shows a further reduction in latency if the software overhead is removed. As indicated, there are big benefits to moving from an SSD to an NVDIMM device, but there are still many orders of magnitude difference between block and byte-addressable accesses.

[1] For more information on SAS SSD devices go to hp.com/us/en/products/server-solid-state-drives/product-detail.html?oid=7605791
[2] For more information on the PCI Workload Accelerator go to hp.com/us/en/products/accelerators/product-detail.html?oid=7876061

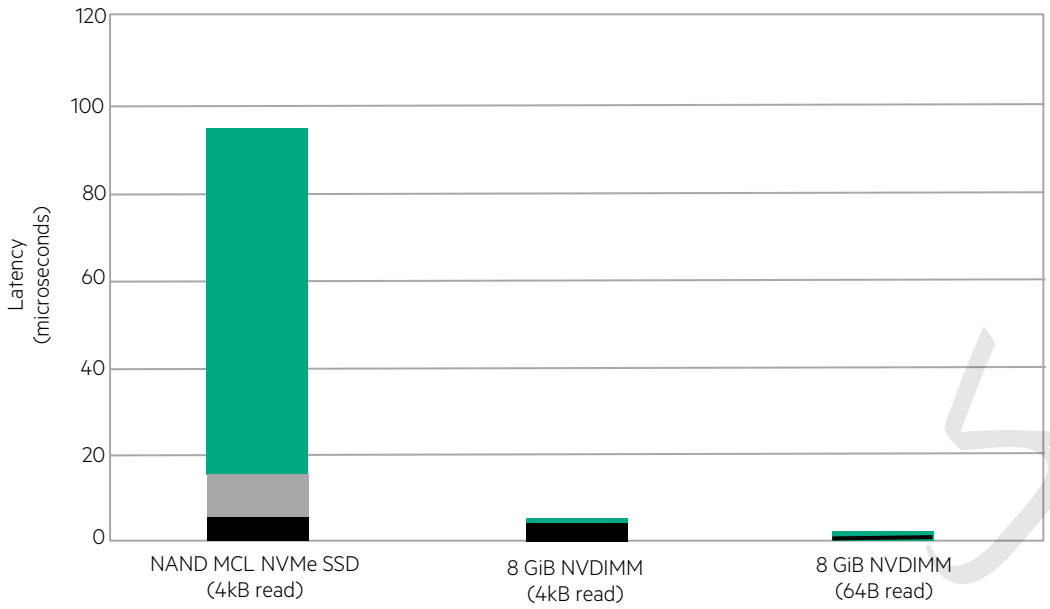

Figure 6. Comparison of NAND and NVDIMM access latencies.

Current applications do not know how to do byte addressable storage accesses and will need to be rewritten to take advantage of this new programming model. HPE is working with industry partners to make this enablement easier through enablers like the PMEM Library at http://www.pme.io. HPE is also working with HPE is working with application vendors to make these updates to their software.

Management

HPE NVDIMM functionality is fully integrated into the system firmware of compliant HPE ProLiant platforms. The UEFI System Utilities allows control of NVDIMM implementation including:

- Enabling/disabling,
- Memory interleaving
- Backup power policy
- Sanitization control
- Data integrity checking

The Integrate Lights Out (iLO) board management controller also provides health and alerting for NVDIMMs.

HPE NVDIMMs require certain considerations to achieve optimal performance and unique considerations regarding removal and relocation. The following guidelines should be followed by IT personnel when configuring and handling HPE NVDIMMs. For more information, refer to the HPE NVDIMM User Guide referenced at the end of this document.

Guidelines for optimal performance

To achieve optimal HPE NVDIMM performance, observe the following guidelines:

- Enable NVDIMMs for interleaving
- Install NVDIMMs evenly across memory channels (two per channel for 2133 MT/s)
- Change Power Profile setting to Maximum Performance in UEFI System Utilities
- Enable Intel Performance Monitoring in UEFI System Utilities

Guidelines for NVDIMM removal

To reduce the chance of loss of data or damage to electronic components the following guidelines should be observed:

- Take the same electrostatic precautions required for all electronic devices.
- Ensure the system is powered down
- Ensure no LEDs on the NVDIMM are illuminated before removing (removing a NVDIMM with an LED illuminated may result in loss of data)

Detailed information on removal procedures is provided in the HPE NVDIMM User Guide referenced at the end of this document.

Conclusion

HPE Persistent Memory has the performance you need to put data to work more quickly in your business and the resiliency you expect in the event of an unplanned system shutdown. As a DRAM-based, memory bus resident storage media, HPE NVDIMMs deliver outstanding performance… up to 34x more IOPs, 16x better bandwidth, and 81x lower latency than SAS SSDs, and up to 24x more IOPs, 6x better bandwidth, and 73x lower latency than PCIe Workload Accelerators. In addition, write-intensive workloads incorporating NVDIMMs increase the life of workload accelerators and SSDs as writes move to much higher endurance DRAM on the NVDIMMs. HPE NVDIMMs represent an initial step on the road to a new programming paradigm: where data is always persistent and we can throw away millions of lines of code that do nothing but move data between volatile working data structures and permanent storage.

This paper has described our initial offering of persistent memory devices and supportive platforms. We are working to increase persistent memory device capacity and performance and plan on expanding persistent memory capability throughout HPE server lines.

Resources, contacts, or additional links

HPE Persistent Memory information
hpe.com/servers/persistentmemory

HPE white paper “Deploying Persistent Memory on MS Windows 2012 R2”
hpe.com/V2/GetDocument.aspx?docname=4AA6-4680ENW

HPE white paper “Deploying Persistent Memory on MS Windows 2016 Technical Preview 5”
hpe.com/V2/GetDocument.aspx?docname=4AA6-4681NW

Sign up for updates

4AA6-4680ENW, October 2016

Hewlett Packard Enterprise