White Paper

Demystifying Breakthrough Oracle Database Storage Technologies

Oracle Exadata Database Machine X8M: Delivering Performance in a League of Its Own

By Brian Garrett, VP Validation Services

December 2019

This ESG White Paper was commissioned by Oracle and is distributed under license from ESG.

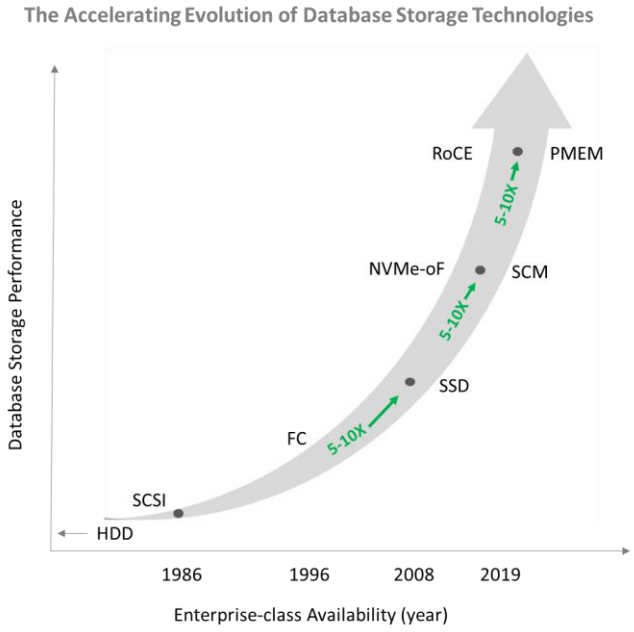
Contents

Demystifying Breakthrough Oracle Database Storage Technologies

This report explores several emerging storage technologies that are delivering breakthrough performance benefits for Oracle Database applications. Our goal is to explain confusing vendor claims about the order of magnitude performance benefits that can be achieved with an "alphabet soup" of new storage access protocols (e.g., RoCE and NVMe-oF), media options (e.g., PMEM and SCM), and architectures.

The Accelerating Evolution of Database Storage Technologies

Database storage performance has crept along an evolutionary path over the past couple of decades as it struggled to keep pace with exponential improvements in CPU and server technologies. It took more than twenty years for database storage latency to improve by a factor of 5x to 10x due to evolutionary advances in storage access protocols and mechanically spinning hard disk drive technologies (e.g., from 5 MB/sec parallel SCSI and 4,800 RPM hard drives in 1986 to 16 Gbps Fibre Channel and 15,000 RPM hard drives in 2006).

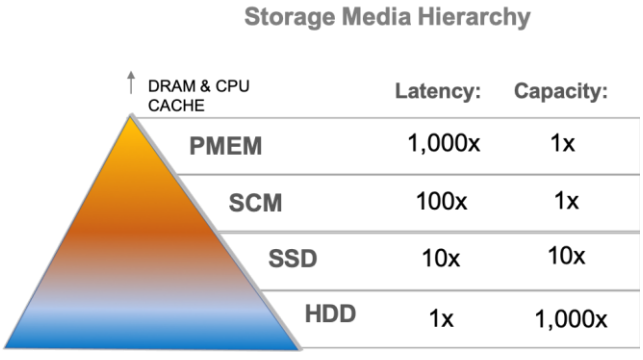


A big bang in database storage performance improvements started around 2008 as solid-state drives (SSDs) began working their way into enterprise-class storage solutions. Storage latencies dropped by a factor of 5x to10x over the next ten years as legacy storage architectures evolved to eliminate new bottlenecks, and one to three millisecond database storage latencies became routine in 2018 thanks to all-flash arrays with new architectures.

2019 marked the beginning of a double big bang of groundbreaking improvements, each with the potential to increase database storage performance by another 5x to 10x. The speed of new storage media options such as persistent memory (PMEM) is approaching that of DRAM, and remote direct memory access (RDMA) technologies such as RDMA over Converged Ethernet (RoCE) are turbocharging storage access.

New Storage Media Options for Oracle Databases

PMEM and storage class memory (SCM) are the generally accepted names for two variants of emerging non-volatile storage media options that were designed with a goal of delivering the random-access benefits of RAM with the non-volatile benefits of traditional SSD and HDD technologies.



PMEM versus SCM

PMEM is typically packaged as an NVDIMM module that plugs into a memory slot, like DRAM.

SCM is PMEM technology typically packaged like an SSD module with a PCIe connection to the storage system. While this asynchronous protocol is up to 10x slower than PMEM, it's up to 10x *faster* than the latest NVMe-attached SSDs, and up to 1,000x faster than legacy HDDs.

Memory Mode versus Direct Mode PMEM

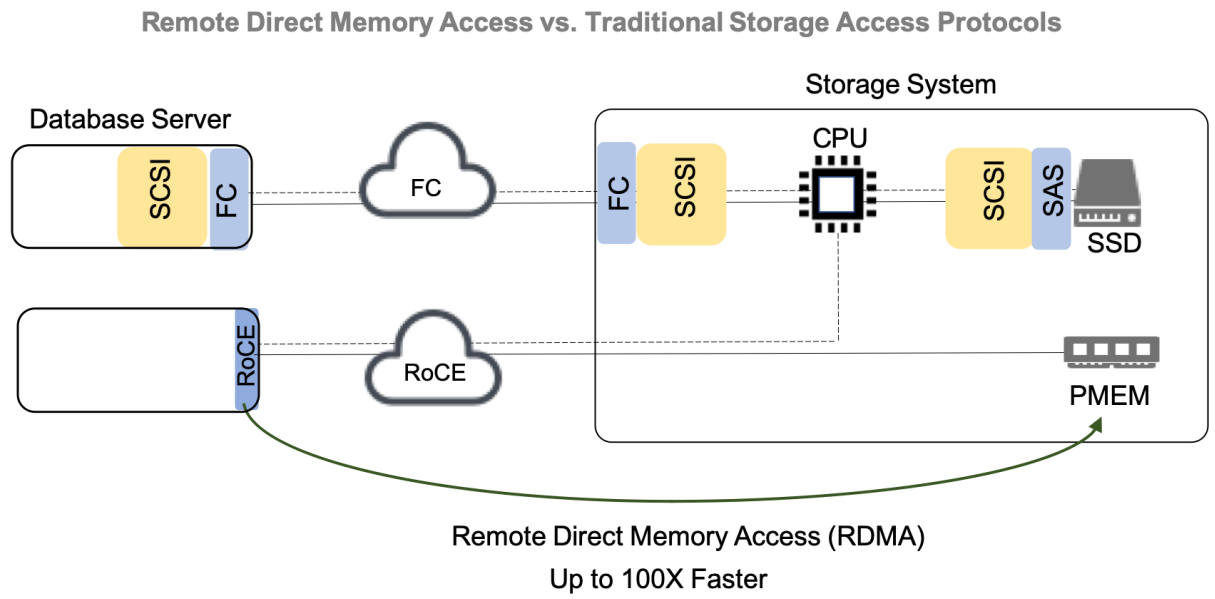
PMEM supports two addressing modes: *memory mode,* which uses the same block-based addressing scheme as traditional HDDs and SSDs; and *direct mode,* which mimics the byte-level addressing scheme of DRAM memory. Memory mode PMEM uses DRAM which acts like a cache in front of larger PMEM modules. Direct mode PMEM, which is commonly referred to as AppDirect mode, provides faster access through direct access to persistent storage on the memory bus. The byte level addressing scheme of AppDirect PMEM is ideally suited for non-volatile application storage requests that are smaller than a traditional block-based I/O (e.g., a 64-byte metadata write instead of an 8KB read/modify/write). Legacy databases and operating systems need to be modified to maximize the performance benefits of AppDirect PMEM, but applications that utilize a database platform that incorporates AppDirect PMEM (like the Oracle Exadata X8M) can run unaltered because the database platform vendor has already done the conversion work.

New Storage Access Protocol Options for Oracle Databases

The database application performance benefits of PMEM and SCM are magnified when combined with emerging front-end storage access protocols that leverage direct memory access technology.

RDMA versus Traditional Database Server Storage Access Protocols

To illustrate the benefits of RDMA versus traditional storage access protocols, let's examine the differences between a legacy FC-attached storage system and a RoCE-attached database server. In the traditional example shown below (top), the database reads and log writes need to wait for logical and physical transport FC processing (SCSI/FC) before the I/O is forwarded to the storage system. Then the CPU inside the storage system needs to set up—and wait for—front-end processing (FC/SCSI) and back-end processing (SCSI/SAS) before the database application can continue. This traditional access method is up to 100x slower than the RDMA over Converged Ethernet (RoCE) method, which provides database applications with non-volatile AppDirect PMEM storage at nearly the same speed as local memory.



RoCE versus NVMe over Fabric (NVMe-oF)

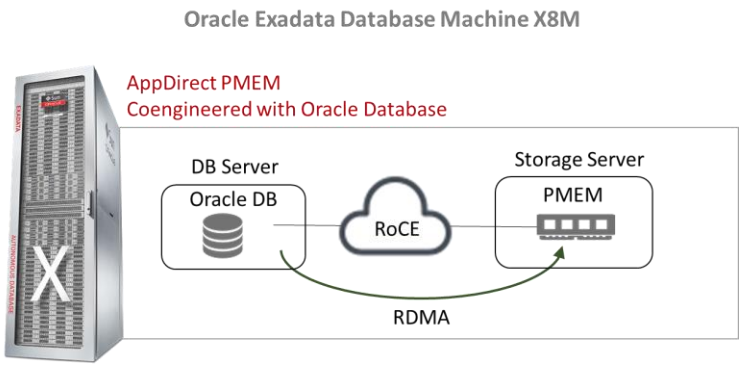
RoCE and NVMe-oF are two of the emerging front-end storage access protocol options that leverage DMA technology to improve the performance of database applications. RoCE leverages a 40 or 100Gbps Ethernet network and RoCE adapters; NVMe-oF can be deployed in a similar manner with 40 or 100Gbps Ethernet network and host adapters and can also be configured to leverage legacy FC networking (FC-NVMe). While RoCE and NVMe-oF provide similar networking performance benefits for front-end database I/O requests, how they are used in an end-to-end storage architecture can lead to database and application performance differences of up to 10x or more.

Exploring the Oracle Database Performance Advantages of Emerging Storage Architectures

Now that we've looked at the potential performance impact of some exciting new front- and back-end storage technologies, let's take a look at the performance differences between architectural approaches that combine those technologies to accelerate Oracle Database application performance.

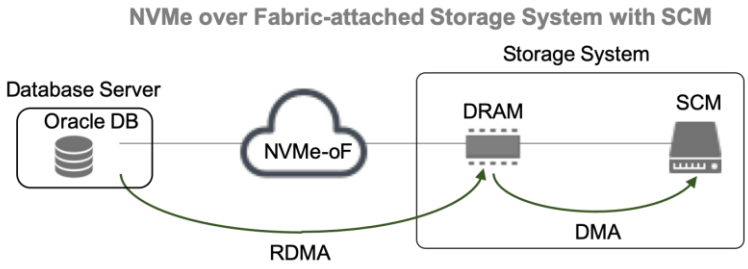
RoCE with AppDirect PMEM

The RoCE with AppDirect PMEM architecture of the Oracle Exadata Database Machine X8M is summarized in the diagram below. The Oracle Database application uses a single RDMA transfer to access PMEM in a storage server. Due to the native AppDirect PMEM support that was recently added to the Oracle Database application, an RDMA PMEM I/O request is accessed the same way as local memory: at a byte level with no CPU overhead. Scale-out is enabled with support for up to 18 storage servers in a single Exadata rack, with each storage server holding 1.5TB of PMEM and the customer's choice of SSD permanent storage for maximum performance or HDD for cost-effective permanent storage. Larger Oracle Exadata X8M systems can be created by adding additional RoCE switches to connect up to 17 additional racks. With this approach, all the complexity associated with leveraging a purpose-built platform with RDMA and AppDirect PMEM that's coengineered with Oracle Database is hidden inside the Oracle Exadata X8M.



NVMe-oF Storage System with SCM

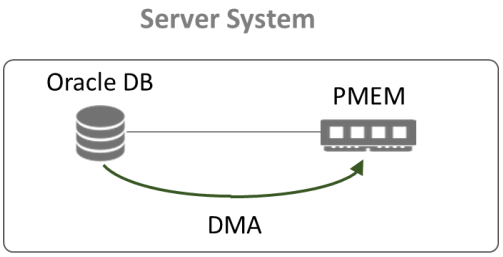
Enterprise-class storage solutions from market leading vendors are in the process of adding support for NVMe-oF and PMEM, which is packaged in a storage form factor and is generally known as SCM. The first solutions became available in 2019 and others have been pre-announced but weren't generally available when this report was written. While the architecture between the front- and back-ends of the various storage systems may differ, they share the approach of using two DMA transfers: 1) NVMe-oF for RDMA transfer between the database server and a DRAM cache buffer with battery backup, and 2) DMA transfer between DRAM and SCM media. This approach is significantly



faster than previous generation FC front-end and NVMe back-end architectures due to the use of memory-mapped DMA, ***but it's up to 10x slower than the Oracle Exadata X8M*** approach due mostly to the performance penalty associated with how RDMA read requests are handled. Oracle Exadata X8M reads are fast due to a special algorithm that publishes the data in the PMEM cache. This eliminates the performance overhead of NVMe-oF SCM solutions, which need to do a messaging handshake to look up the remote memory address before the RDMA can start. Interoperability can also be a challenge with this approach due to the lack of native NVMe-oF operating system support (e.g., Windows Server) and generally available enterprise-class storage systems with NVMe-oF support when this report was written.

Server System with PMEM

Server systems that leverage the DMA performance benefits of PMEM are starting to emerge. Most of these solutions don't support AppDirect PMEM yet. Server systems suffer from ***scalability and high availability challenges*** due to server DIMM slot limits (12 slots when this report was written) and the fact that the PMEM is not shared across servers, creating the traditional single-server high-availability challenge. Hyperconverged infrastructure (HCI) holds the promise of overcoming these limitations with a scale-out approach that leverages RDMA between server nodes, but that requires that the storage (and the cluster metadata traffic) be carefully managed across multiple servers so a single device or HCI node failure does not render data unavailable.



The Breakthrough Storage Performance Advantages of the Oracle Exadata X8M Architecture

The RoCE and AppDirect PMEM architecture of the Oracle Exadata X8M delivers extremely fast database I/O response times of less than 19 microseconds for 8K OLTP database requests, and up to 16M OLTP read IOPS and 560 GB/sec of uncompressed analytics throughput from a single rack.[1] A single platform can scale up to 18 racks with RoCE cables and internal switches, and more than 18 racks with external RoCE switches.

Table 1 summarizes the end-to-end performance advantages and considerations associated with emerging storage architectures for Oracle Database.

Table 1. Oracle Database Storage Technology Considerations

	Storage System	Server System	Oracle Exadata X8M
I/O latency	Fast	Faster	**Fastest**
Considerations	NVMe-oF and SCM support	Scalability and DIMM slot limits	**Low Latency, High Throughput**
	5-10x slower	HA complexity	**HA Coengineered with Oracle Database**

I/O latency:

- Storage systems with NVMe-oF and SCM are fast (up to 10x faster than previous generation SSDs).
- Server systems with PMEM are faster.

[1] https://blogs.oracle.com/exadata/exadata-x8m

- Oracle Exadata X8M with RoCE and AppDirect PMEM is the fastest architecture for Oracle Database: 5-10x faster than storage systems with NVMe-oF and SCM, up to 100x faster than previous generation all-flash storage systems with SSD, and 1,000x faster than HDD.

NVMe-oF and SCM support: The availability of storage and operating systems with native NVMe-oF and SCM support was limited when this report was written. This limits the choices currently available for getting the performance advantages of an RDMA architecture and the latest storage media. This also introduces risk as the industry works its way through interoperability and support challenges over the next couple of years.

Scalability and DIMM slot limits: The DIMM slot limits of industry-standard servers (12 when this report was written) limits the capacity scalability of emerging PMEM server system architectures. Clustering and hyperconverged server system architectures can be used to break this capacity scalability limit, but these add latency that throttles the potential performance gains of PMEM due to the performance penalty associated with east-west traffic between nodes.

HA complexity: The Oracle Exadata X8M and enterprise-class storage systems were purpose-built with mission-critical levels of high availability in mind. For example, the Oracle Exadata X8M uses hardware-level mirroring to make sure that database write commits are simultaneously written to two sets of PMEM devices. Implementing similar levels of high availability and performance is impossible with the industry-standard servers that were shipping when this report was written.

Low Latency, High Throughput: Oracle Exadata X8M combines Intel Optane DC persistent memory and 100 gigabit RDMA over Converged Ethernet to remove storage bottlenecks and dramatically increase performance for the most demanding workloads including OLTP, analytics, IoT, fraud, network intrusion detection, and high frequency trading.

HA Coengineered with Oracle Database: The performance benefits of the Oracle Exadata X8M architecture are magnified with the simplicity and future-proof benefits of being coengineered with Oracle Database.

Dig Deeper into Vendor Claims

Now that we've explored some of the emerging architectures that are taking advantage of breakthrough storage technologies for Oracle Database performance, let's clarify some of the vendor claims that have begun to emerge.

"Our NVMe-oF (or RoCE) attached storage system uses RDMA to deliver the same performance benefits as Oracle Exadata X8M."

NVMe-oF speeds up the front-end access to database storage, but it requires a messaging handshake to get the remote memory address before the RDMA can start. The architecture between the RDMA front-end and back-end of the storage system often adds latencies as well. Ask your storage vendor for latency measurements with an 8KB OLTP workload and compare them with Oracle Exadata X8M (less than 19 microseconds).

"Our storage system uses the latest SCM technology to deliver the same performance benefits as Oracle PMEM."

SCM and PMEM accelerate Oracle Database workloads using similar technologies that provide DMA access to storage media, but SCM, which stores data in non-volatile memory accessed over the PCI bus, is 5-10x slower than Oracle Exadata X8M PMEM that's accessed like memory in a DIMM slot.

"Our 'NVMe-ready' storage architecture can deliver the same performance benefits as Oracle Exadata X8M."

Enterprise-class storage system architectures have evolved over the past couple of years to take advantage of the performance benefits of NVMe DMA on the back-end (e.g., NVMe-attached flash) and more recently on the front-end (e.g., NVMe-oF). And as we've explored in this report, the architecture between the front- and back-ends matters as much, if not more, if you want to take advantage of emerging SCM and PMEM technology. Ask your storage vendor to clarify whether "NVMe-ready" means front-end, back-end, or both, and when it will be generally available and supported.

"Our server system (or hypervisor or HCI platform) supports PMEM, which delivers the same performance benefits as Oracle Exadata X8M PMEM."

Persistent memory in a DIMM slot in a server system accelerates the performance of Oracle Database applications, but you need to ask questions to learn whether it can deliver the sub-20 microsecond latencies of Oracle Exadata X8M: Is AppDirect mode supported? Can I accelerate redo logs with PMEM and restart a database in seconds like I can with Oracle Exadata X8M? Will DIMM slot limits block me from meeting the capacity and performance needs of my Oracle Database applications? What happens if a PMEM module fails? Is the solution as highly available as the Oracle Exadata X8M?

"Applications aren't ready for AppDirect PMEM."

It's true that general-purpose applications, operating systems, and hypervisors need to be modified to take advantage of the performance benefits of AppDirect PMEM, and most haven't been ported yet. However, since Oracle has made the required modifications inside Oracle Exadata X8M, the applications that benefit from running Oracle Database on it do not require any modifications. Oracle has done all the "heavy lifting" required to make the benefits of PMEM and RoCE available to customers without any modifications on their part.

"Our solution delivers more IOPS so it's faster than Oracle Exadata X8M."

While IOPS measures the maximum number of I/Os possible under ideal conditions, latency—that is, how long it takes for a database application I/O request to finish—is the metric that matters the most when you're comparing the speed of two storage solutions. Latency is the metric that impacts application users. It's like the zero-to-sixty rating for a sports car. The faster each I/O completes, the faster you'll get to the finish line. Like IOPS, the power of the engine under ideal conditions (e.g., on a dynamometer) is helpful, but like latency, the zero-to-sixty ratings are a better indication of how fast you'll get to the finish line when driving in real-world conditions. Ask questions about the real-world conditions that were used for competitive performance claims: Was the size of each I/O request smaller than a typical real-world database I/O request (e.g., 512 byte reads versus 4KB OLTP)? Was the latency measured from the actual storage media itself (e.g., Oracle Exadata X8M PMEM versus a DRAM cache buffer)? Are the big IOPS and bandwidth claims due to the turbocharged latency of the storage architecture or to a big pile of equipment that's more than you can afford?

The Bigger Truth

Sometimes choosing a technology solution doesn't require a deep dive into technical details. But digging a little deeper can reveal the underlying reasons for performance differences between solutions. For that reason, we have looked at why emerging innovations in storage media, access protocols, and architectures are driving greater performance in the Oracle Exadata X8M than in other storage and server solutions.

Database storage performance improvements have evolved at an accelerating pace, and in 2019, two key groundbreaking improvements appeared: 1) new storage media (such as PMEM and SCM) that approach the speed of DRAM, and 2) direct memory access technologies that turbocharge storage access (such as RoCE and NVMe-oF).

The Oracle Exadata X8M combines these technology advancements for extreme performance, while enabling scalability, high availability, and cost-effective tiering. Other storage and server solutions that leverage PMEM or SCM suffer from a variety of performance penalties and tradeoffs compared to an Oracle Exadata X8M that's coengineered with Oracle Database.

Other vendors claim that because they use similar emerging technologies, they can deliver the same performance as the Oracle Exadata X8M. That is not the case; imagine wearing the same running shoes as Usain Bolt….it does not mean you can run as fast as him! Dig a little deeper and ask a few pointed questions so you can make a truly informed decision based on the performance, scalability, availability, and interoperability requirements of your Oracle Database applications. In our current assessment, the Oracle Exadata X8M is in a league of its own due to a simply elegant architecture that leverages the power of the latest emerging storage media and access technologies in a purpose-built solution that's coengineered with Oracle Database.

The bottom line—to extend our Usain Bolt analogy—is that the daylight that Oracle puts between itself and the other contenders in the "database performance race" is not a matter of opinion but is instead a matter of fact. When it comes to accelerating the reads and writes of Oracle Database applications, that "head start" means that there aren't any genuine challengers to Exadata X8M amongst other current storage array, server, or HCI offerings. They can buy the same running shoes, for sure, but in ESG's opinion, only Oracle can win the race.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides actionable insight and intelligence to the global IT community.