WHITE PAPER

Data Center
Software-Defined Memory

System Memory at a Fraction of the DRAM Cost

Intel® Optane™ SSDs with Intel® Memory Drive Technology Offers Memory Expansion Solution.

Abstract

Contemporary application software requires much more memory than was imagined a decade ago, and requirements for larger memory appear to increase with time. Server applications demand the most memory, but system memory capacity is limited, and specialized solutions are proprietary and expensive.

Intel® Memory Drive Technology, coupled with high-performing Intel® Optane™ solid state drives (SSDs), enables an alternative to expensive dynamic random-access memory (DRAM) and proprietary systems: software-defined memory (SDM) coupled with high-performance non-volatile memory (NVM). This white paper demonstrates how the Intel® Memory Drive Technology SDM implementation — using the latest NVM innovations, namely, storage-class memory (SCM) — delivers excellent performance, flexibility, and scale at a reasonable cost. Relevant use cases will be covered to help illustrate this point.

Background

In an effort to keep pace with Moore's Law, the number of cores on a processor has significantly increased in the last decade. At the same time, data and workloads have grown, taking advantage of the increasing compute power. With the introduction of commercial NAND-based SSDs in 2009, the I/O subsystem also saw significant improvements: SSDs grew in density and size and had much lower power consumption, delivering economies of scale that were unimaginable before then.

However, one key component of the computer system has not kept pace with such advancements: system memory. Typically made of DRAM, system memory's performance and capacity improvements lag. At the same time, the price of DRAM remains very high. As of 2017 (Intel® Xeon® Scalable Processors), dual processor server system memory capacities reached 1TB-1.5TB with standard processor SKUs, and quad-processor server models based on specialized "M" SKUs reached 3TB-6TB for overall DRAM capacity.

Server Memory Requirements

Current computing workloads can be classified into two key use cases from the perspective of system memory requirements:

- **Compute-centric workloads,** ranging from modest memory requirements (where multiple workloads can share the same Intel-based system) to cases where a single workload requires the full server's capacity. Most workloads fit within this "Compute-centric" category.
- **Memory-demanding workloads**, which require more RAM than a server can provide and require specialized systems with high DRAM capacity to execute. These include in-memory databases, bioinformatics, graph-processing, etc.

Memory-demanding workloads obviously drive the need for more memory, but so do many compute-centric workloads. Memory can become the constrained resource in compute-centric workloads when the number of compute cores scales faster than the memory capacity, while the memory usage per core either remains the same or increases. From 2005 to 2017, for example, Intel® Xeon® processors grew from one core to 28 cores per processor. Even more workloads can be put on the same system using containers, virtual machines, or other multi-tenancy paradigms, further driving the need for large memory to utilize the added compute power.

The growth in compute power of a single system and of workload memory usage has only been partially met by developments in the DRAM market. Server systems remain limited in the number of DRAM slots per processor, with a maximum of 12 slots per processor on Intel® Xeon® Scalable Processors. The DIMM capacity is also limited. As of late 2017, the optimal $/GB is achieved by using 32GB DIMMs; 64GB DIMMs are more expensive per GB; and 128GB DIMMs are in short supply and come at a hefty premium of hundreds of percent points per GB, in comparison. This means that a dual-socket server with 768GB of memory provides optimal $/GB value.

Storage-Class Memory (SCM)

With the growing popularity of NAND flash in the consumer storage market, manufacturing volumes have grown and technology has improved. Since 2009, NAND-based SSDs have evolved and are widely used, forcing the storage industry to innovate with non-volatile enterprise storage. These include both controllers and the storage media itself, which is no longer limited to NAND.

Intel® Optane™ SSDs combine the attributes of memory and storage. With a combination of high throughput, low latency, high QoS, and ultra-high endurance, Intel® Optane™ SSDs provide an innovative solution for breaking through the data access bottleneck by providing a new data storage tier.

Intel first publicly demonstrated Intel® Optane™ SSDs in 2015. With the product launch in March 2017, they have improved the performance of solid state I/O devices and narrowed the performance gap between NVM and DRAM.

In 2008, SCM was defined as:

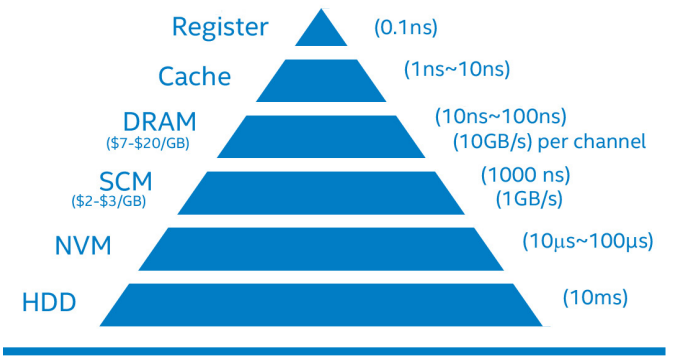

Figure 1: Key attributes of memory and storage classes

- Solid state, no moving parts
- Short access times (DRAM-like, within an order of magnitude)
- Low cost per bit (DISK-like, within an order of magnitude)
- Non-volatile (~10 years)

(IBM Almaden Research Center, Freitas, R.F., Wilcke, W.W.)

So on one hand, SCM is made up of all-silicon memory components with performance attributes close to those of memory components. On the other hand, it features non-volatile storage with the capacity and economics of legacy storage devices such as hard drives. For the sake of precision, with Intel® Optane™ SSDs, the latency for the highest-performing NVM technologies is on the order of 10µs, and therefore falls short of IBM's SCM definition, described in Figure 1. However, with Intel® Memory Drive Technology enhancing the actual average performance, we can already regard those technologies as SCM.

Without Intel® Memory Drive Technology, when users install an Intel® Optane™ SSD into their systems, they will see it as a storage device. The operating system can only operate those devices as a block device, which is typically used for storage, as they are not "byte addressable." This also means that they cannot be operable like standard system memory and cannot be transparently used as if they were DRAM. In this situation, the operating system could not access Intel® Optane™ SSDs at the byte level, only at the block level (for reference, blocks are typically hundreds or thousands of bytes).

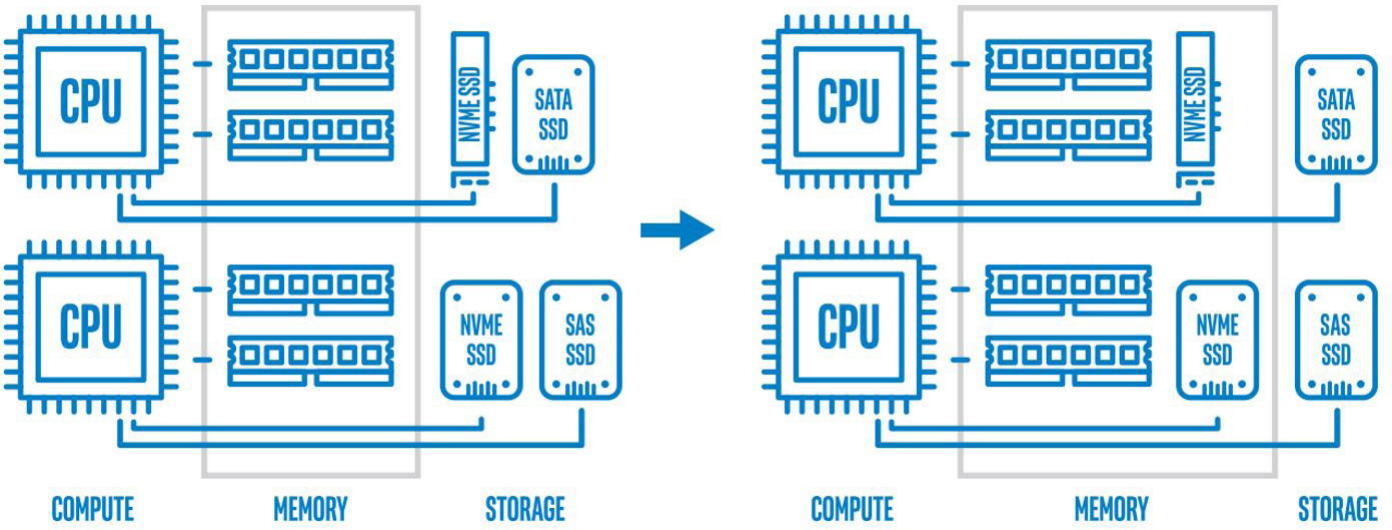

Figure 2: Intel® Memory Drive Technology modifies a part of the system's storage to transparently serve as system memory.

Software-Defined Memory for SCM (SDM-S)

Assembling memory of varying performance into a single system is a decades' old practice. Non-uniform memory access (NUMA) systems — typically defined as any contemporary multi-socket system — allow a processor to access memory at varying degrees of latency or "distance" (e.g. memory attached to another processor), over a network or fabric. In some cases, this fabric is purpose-built for such processor communication, like Intel® QuickPath and UltraPath Interconnects (Intel® QPI and UPI respectively). In other cases, standard fabrics such as PCI Express* or Intel® Omni-Path Fabric are used for the same purpose along with SDM: to provide memory coherency, operating as if additional memory was installed in the system (SDM for Fabrics, or SDM-F).

Accessing memory at varying lower performance over networks has proven to be feasible and useful by using predictive memory access technologies that support advanced caching and replication, effectively trading latency for bandwidth. This is exactly what Intel® Memory Drive Technology is doing to enable NVM to be used as system memory. Instead of doing it over fabric, however, it does so with SCM, creating software-defined memory for SCM (SDM-S). In SDM-S, most of the SCM capacity is transparently used as an extension to the DRAM capacity of the system.

By using Intel® Memory Drive Technology to perform advanced memory access prediction algorithms and implement an OS-transparent virtual machine, Intel® Optane™ SSDs can be used as part of the system memory to present the aggregated capacity of the DRAM and NVM installed in the system as one coherent, shared memory address space. No changes are required to the operating system, applications, or any other system components. With this approach, Intel® Memory Drive Technology brings new capabilities and flexibilities to IT.

SDM-S Use Cases

There are two key scenarios in which it is beneficial for an IT environment to use Intel® Memory Drive Technology:

- Where very large system memory is required — more than the DRAM capacity of an Intel® Xeon®-based server system
- Where the cost savings of replacing DRAM with NVM outweigh the performance difference between the two technologies

For the purposes of this paper, we shall refer to the former as "memory expansion" and the latter as "memory replacement."

Examples of Memory Expansion

Consider in-memory database (IMDB) engines, which require a shared-memory (non-distributed) system. Examples include SAP HANA*, Oracle* 12c with Column Store*, and IBM DB2 BLU*. With Intel® Memory Drive Technology, an IMDB of larger than 10TB can easily be run on a dual-socket Intel® Xeon® server system. In fact, IMDB deployments of any size larger than the server's optimal DRAM capacity (768GB at the time of writing) may benefit from it.

Additionally, in scientific computing and computer-aided engineering (CAE), there are many workloads that cannot be run without sufficient system memory. Examples include de-novo genome assembly, which typically uses De Bruijn graph in-memory, or pre-processing/meshing of a model for simulation in CAE, which today consumes hundreds of gigabytes to several terabytes of system memory for large models.

Examples of Memory Replacement

Multitenancy scenarios are very common in enterprise IT, and even more popular with cloud service providers. The more workloads (e.g. multiple VMs, multiple containers, multitenant databases, etc.) one can place into a single physical server, the better the utilization or yield of the infrastructure. In most multitenant cases, system memory limits the number of workloads per node, so there is reason to deploy systems with higher memory capacity. With Intel® Memory Drive Technology, less DRAM is used per node and is replaced by NVM. In some cases, even more NVM is deployed with SDM-S to further leverage the economic benefits – for example with in-memory data grids and in-memory distributed data processing frameworks such as Apache Spark*.

Another example could be a large grid running an embarrassingly parallel workload, such as "value at risk" (also known as "VAR") for a financial institution. Many independent Monte Carlo-based computations are executed concurrently, and the average memory-per-core requirement can be predicted based on the complexity of the simulation. This can be in the range of a few kilobytes up to terabytes per thread. However, the actual processes — typically many thousands of them — vary in memory consumption to the point that IT must over-provision memory into the nodes to avoid execution failures and job resubmissions.

With Intel® Memory Drive Technology, IT would provision just the amount of DRAM needed per core — even lower than the average — and augment the missing DRAM for peak use with Intel® Optane™ SSDs. No failures would occur, there would be no need to restart jobs and wait for results, and processes that do "spike" would still run at close to DRAM performance, most likely allowing the overall computation, across all nodes, to end at the same time as it would if only DRAM were used and overprovisioned.

For example, consider a VAR setup where average memory utilization per node is 256GB RAM, where about 80% of the nodes only pass the 192GB mark for less than 10% of the overall run, and the maximal exhibited memory usage is 700GB, on varying nodes. A bank using 1,000 servers to run this VAR computation would need to over-provision most, if not all, nodes with 768GB, to avoid the risk of job failure, which would mean restarting the job and the rest of the model waiting for those results from the "weakest link." Instead they could effectively deploy 192GB of DRAM per node, and have the rest served by Intel® Optane™ SSDs with Intel® Memory Drive Technology. While the savings per node could amount to $1,000-$1,500 — for 1,000 servers that means more than $1M, just for CapEx — OpEx savings also would add up as Intel® Optane™ SSDs use much less power per GB than DRAM.

Economic Model for Memory Expansion

In some cases, there are platform limitations that don't allow scaling memory capacity beyond a certain point. Scaling between 768GB and 1.5TB requires 64GB DIMMs, which are more expensive LRDIMMs. Dual-socket platform architecture has a limitation itself, and so scaling beyond 1.5TB requires using cost-prohibitive 128GB DIMMs or moving to a four-socket system. That means, for an application needing even just 2TB of addressable memory, organizations need to move away from the efficient dual-socket Intel® Xeon® server to a higher-scale system with more sockets, or using expensive high-density DIMMs.

The costs of upgrading to a high-end server are not insignificant: While a dual-socket server with 768GB of DRAM could cost in the $10,000 range, a quad-socket Intel® Xeon® system with double the capacity (double the cores and 1.5TB memory) could cost more than $35,000 as of late 2017. It would also typically require 4U of rack space, as well as much more power and cooling.

In comparison, a dual-socket system with 256GB DRAM and 1600GB of NVMe* drives, including Intel® Memory Drive Technology, could cost less than $20,000. The same system with 768GB DRAM and Intel® Memory Drive Technology adding 6400GB from NVMe SSDs could cost less than $50,000 — that's more than 6TB system memory for the cost of a 3TB DRAM quad-socket system.

Economic Model for Memory Replacement

The financial benefits of memory replacement can be attributed to two key aspects of TCO:

1. Acquisition cost: NVM costs up to 70% less than DRAM. A 32GB DDR4 ECC RDIMM currently costs a minimum of $350, or at least $11.00/GB, while enterprise-class drives retail for $3/GB - $4/GB, or under $6/GB with Intel® Memory Drive Technology. Thus, for a dual-socket Intel® Xeon® server, instead of buying 768GB of DRAM, one could buy 128GB of DRAM and two Intel® Optane™ SSD P4800X devices (320GB in memory mode), possibly saving $2,500. Furthermore, acquisition cost savings can grow if comparing to 64GB DIMM setups with 1.5TB total memory.
2. Operational cost: The current generation of Intel® Optane™ SSDs consumes up to 8 watts for active reads (~14 watts for burst peak I/O), while a DDR4 DIMM consumes 6 watts. In the first scenario — 20 DIMMs replaced by two Intel® Optane™ SSDs — the savings are at least 90 watts per server, which significantly reduces the lifetime cost of a cloud environment for power and cooling. For example, if power costs $0.15 per kWh, and assuming a PUE of 2.0, that's $700 saved per server for three years of non-stop operation.

Performance

When comparing alternatives, consider the tradeoffs of technology: cost and performance. Using DRAM-only configurations where possible would represent the highest-performing and most expensive alternative. We have shown that the combination of Intel® Optane™ SSDs and Intel® Memory Drive Technology is significantly lower in cost compared to DRAM, but what about performance? Is it close to DRAM? What is the tradeoff, and which workloads are a good fit?

We use two benchmarks for this comparison with Intel® Optane™ SSD P4800X with Intel® Memory Drive Technology. The results are to the right.

As expected, results differ for different workloads, but it is clear that with Intel® Optane™ SSDs and Intel® Memory Drive Technology one could reach close to DRAM performance.

Summary

Intel® Optane™ SSDs for PCIe*/NVMe allow for excellent I/O performance. Paired with Intel® Memory Drive Technology, these drives can be used to replace or expand DRAM at a fraction of the cost and deliver unmatched performance per dollar, potentially reducing the cost of a dual-socket server by more than $3,000 or doubling the system memory capacity of a quad-socket server for the same investment. Intel® Optane™ SSDs further narrow the performance gap between DRAM and SDM-S, and enable a wider range of use cases and workloads to benefit from the technology.

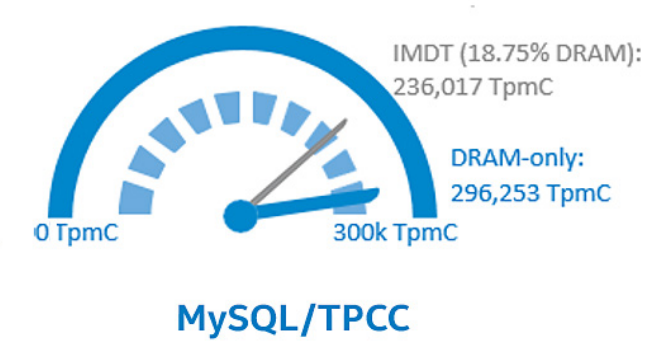

Figure 3: Performance comparison of standard DRAM vs. Intel® Optane™ SSD with Intel® Memory Drive Technology

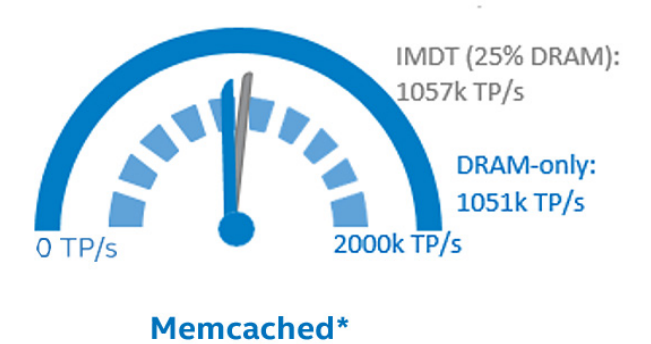

Figure 4: Performance effect of DRAM: NVM ratio in setup of Intel® Optane™ SSD with Intel® Memory Drive Technology

Learn more at intel.com/ssd

System configuration:

- DRAM-only system: motherboard S2600WT, 2x Intel® Xeon® E5-2699v4, 512GB DDR4
- NVMe+Intel® Optane™ system (MySQL*/TPCC*): motherboard S2600WT, 2x Intel® Xeon® E5-2699v4, DRAM 96GB, and 2 x Intel® SSD DC P4800X 375GB, tpcc-mysql (benchmark URL-https://github.com/Percona-Lab/tpcc-mysql), Percona Server for MySQL 5.7.19-17.
- NVMe+Intel® Optane™ system (memcached*): motherboard S2600WT, 2x Intel® Xeon® E5-2699v4, DRAM 128GB and 2 x Intel® SSD DC P4800X 320GB, memcached version 1.4.34, memaslap* version 1.0.

Printed in USA ♻ Please Recycle 1217/AK/NH 334784-003US