White paper
Performance Report of Intel® Optane™ Persistent Memory on PRIMEFLEX® for SAP HANA®

Intel Optane persistent memory is blurring the line between DRAM and persistent storage for in-memory computing. Unlike DRAM, Intel Optane persistent memory retains its data if power to the server is lost or the server reboots, but it still provides near-DRAM performance. In SAP HANA operation this results in tangible benefits, which are described in this document based on real measurement results.

Content

Preface	2
What Is PMEM and Why Do Companies Have Use for It?	2
Intel Optane PMEM and FUJITSU Server PRIMERGY and PRIMEQUEST	4
Intel Optane Persistent Memory Operating Modes	4
SAP HANA and Intel Optane PMEM	5
FUJITSU Integrated System PRIMEFLEX for SAP HANA	5
Benchmark Achievements on PRIMEFLEX for SAP HANA with PMEM	6
Performance Test on PRIMEFLEX for SAP HANA with PMEM	7
Intel Select Solutions	9
Summary	10

Preface

Intel Optane Persistent Memory is non-volatile memory that provides much larger capacities than DRAM and higher performance than conventional storage media (i.e. SSDs and HDDs). This report offers general insights into the Intel Optane Persistent Memory technology, outlines its benefits for SAP HANA, shows the performance range and scale-up capabilities with the SAP BW edition for SAP HANA Standard Application Benchmark and compares performance test results of SAP HANA running on Intel Optane Persistent Memory and DRAM. Intel Optane Persistent Memory is abbreviated to Intel Optane PMEM or PMEM in the following. The previously known name was Intel Optane DC Persistent Memory, abbreviated to DCPMM.

- Target audience: This report assumes readers have basic knowledge of SAP HANA.
- Platforms used for SAP BW edition for SAP HANA Standard Application Benchmarks Version 3: PRIMEFLEX for SAP HANA and Fujitsu Server PRIMERGY RX4770 M5 and Fujitsu Server PRIMEQUEST 3800B2
- Platform used for performance comparison tests: PRIMEFLEX for SAP HANA and Fujitsu Server PRIMERGY RX4770 M5
- Intel Optane Persistent Memory 100 Series used for all benchmarks and tests

What Is PMEM and Why Do Companies Have Use for It?

Whether we look at latest-gen OLTP systems, in-memory databases, real-time or near real-time BI inquiries, or HCI solutions – all of these major IT advancements from the past ten to 15 years have something very important in common: They all support the idea that in today's data centers nothing matters as much as the ability to bring and more importantly keep as much data as possible as close to the main memory and CPU as possible. This has led to a refreshed interest in so-called non-volatile memory (NVM) solutions that could help to further minimize the latencies that still impede a computer's ability to transfer data from storage media to processors and memory modules and back at the best rate possible. Non-volatile memory, or more specifically NVRAM/NVDIMM components – such as Intel Optane PMEM– can help solve this dilemma.

3D XPoint architecture

Historically, Intel Optane PMEM began life sometime in the mid-2000s and was first introduced to a broader public in July 2015 by its developers Intel and Micron, who had officially joined forces on the project three years earlier. Back then, the technology was called 3D XPoint (read: "3-D cross point"), a name that succinctly described the physical characteristics of the new components, as can be seen in Fig. 1, which also explains the functional design of the new modules.

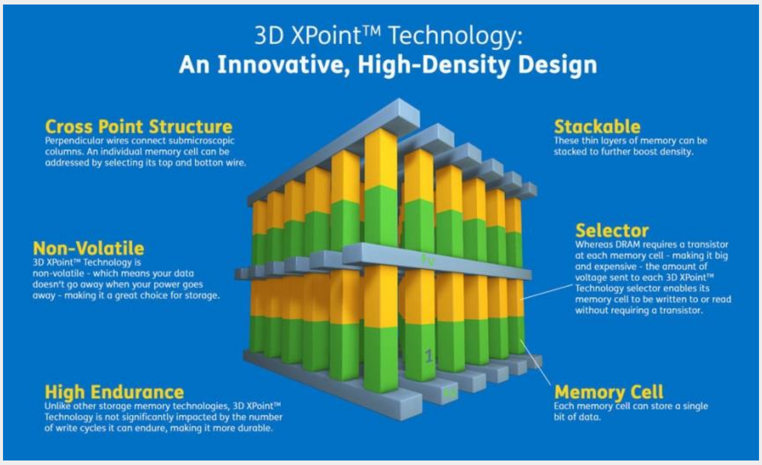


Fig. 1: The original 3D XPoint architecture, as presented by Intel and Micron in 2015

As may be easily concluded from the above description, the idea was to create a type of memory that is not only non-volatile, but also highly durable while working considerably faster than regular flash memory. In essence, the engineers designed a new class of modules that allowed ICT vendors to supersede the limitations of the classic three-tier storage model for "hot, warm and cold" data shown in Figure 2.

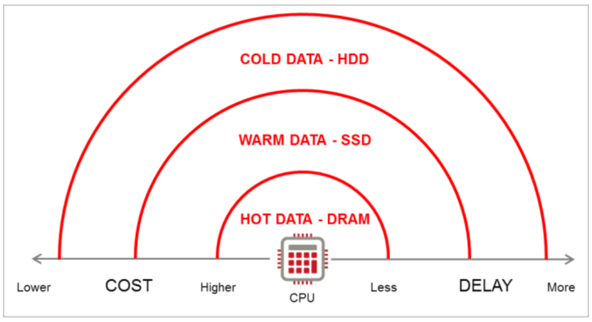
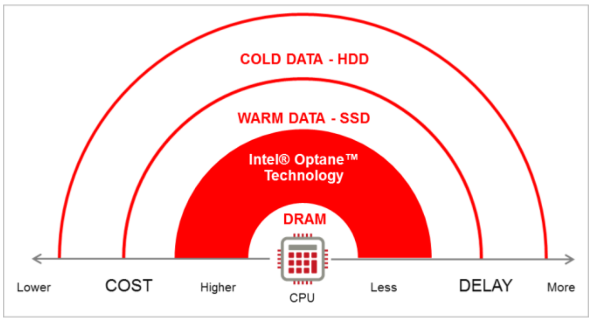
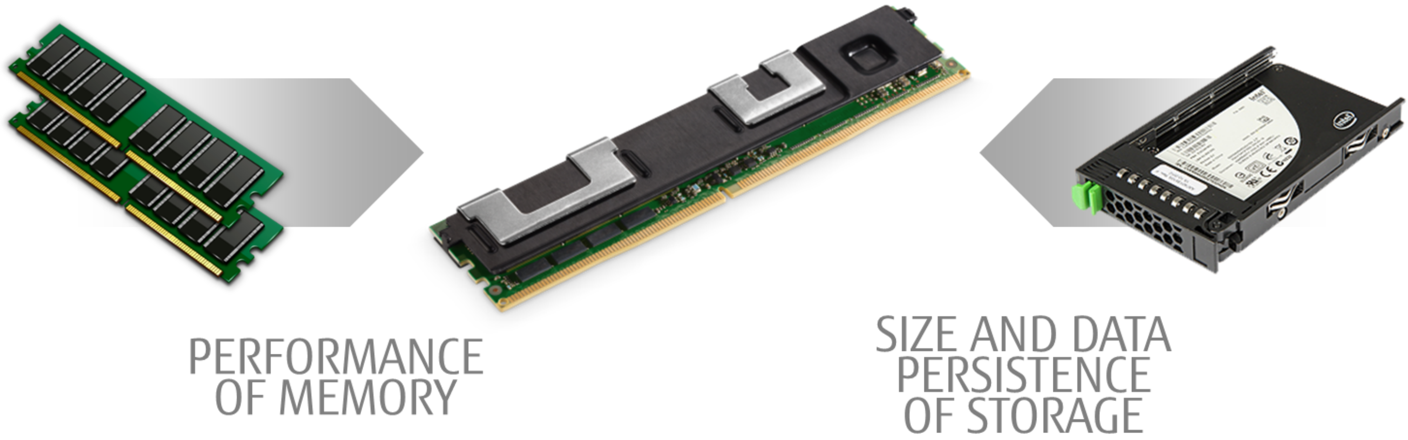


Fig. 2: Classic 3-tier data storage model



Instead, they could now introduce a fourth layer for "semi-hot" data, i.e. the kind of information users would like to find ready at their fingertips upon system startup, but usually don't – for example customer databases underpinning OLTP processes or transaction logs enabling real-time analytics. This new NVRAM layer was to be inserted directly in between the two innermost tiers or "rings" of data, namely the larger amount kept on SSDs and the considerably smaller chunk that must be loaded into the DRAM and CPU to start an application. Figure 3 shows the new 4-tier architecture as implemented by Intel through its Optane modules. (The Optane brand name was introduced after Intel and Micron parted ways again so each firm could market 3D XPoint technology individually.)

Fig. 3: The 4-tier data storage model used by Intel Optane technology



Fundamentally, Intel uses its PMEM to redefine what we've come to accept as the regular memory and storage hierarchy: The modules combine the size and persistence of typical storage media with performance levels that are similar to those found in standard DRAM. In practical terms, this means that vital information required to run actual applications can be accessed within hundreds of nanoseconds versus 10s of microseconds for SSD and millisecond for HDDs. Figuratively speaking, the difference can be compared to a production process in which intermediate goods ship to a manufacturer within a few hours or days instead of being readily stored on site. To make the technology even more attractive, Intel Optane PMEM fit into the same standard memory slots as DRAM does.

Intel Optane PMEM and FUJITSU Server PRIMERGY and PRIMEQUEST

Intel Optane PMEM modules were released in April 2019 along with the highly anticipated second generation of Intel Xeon Scalable processors. Thanks to our long-standing partnership, Fujitsu was among the first companies to integrate the CPUs and memory technology into the PRIMERGY M5 generation and PRIMEQUEST 3000 series second generation systems. The servers offer the following technical highlights:

- Support new 2nd Generation Intel Xeon Scalable Processors
- Intel Optane persistent memory support (coming in sizes of 128, 256 and 512 GB) delivering a unique combination of affordable large capacity and persistence
- Converged data center management that provides organizations centralized control over the entire infrastructure that includes servers, storage, networking, cloud management software as well as power and cooling using a single user interface.PRIMERGY servers come with a wide variety of such robust security features and combine these capabilities with the best quality and efficiency, and more agility in daily operations helps to turn IT into a business advantage faster. Flexible configuration options: Dedicated base units, broad OS selection, broad selection of controllers and devices

The Fujitsu Intel-based servers benefit from Fujitsu's pioneering work in the NEXTGenIO project which forms part of EU Horizon 2020, the European Union's biggest ever research and innovation program. Fujitsu developed the NEXTGenIO system and systemware to close the gap between memory and storage by leveraging Intel Optane Persistent Memory Modules. The success of this EU project made it possible for Fujitsu to introduce these unprecedented I/O capabilities to the PRIMERGY and PRIMEQUEST servers.

Intel Optane Persistent Memory Operating Modes

Customers who opt to buy the latest FUJITSU Server PRIMERGY and PRIMEQUEST systems equipped with 2nd Generation Intel Xeon Scalable processors and Intel Optane PMEM can choose between three distinct operational modes to accommodate individual use cases – specifically, Memory Mode (MM), App Direct Mode (AD), and Mixed Mode. The names describe the role of PMEM in each setup and are almost self-explanatory.

- In Memory Mode, the entire PMEM capacity acts as 'classic' system memory, with all the consequences this entails. In particular, it means they become volatile just like regular DRAM modules. This, in turn, affects the way in which the OS and applications perceive and address both types of units as a pool of volatile memory in which PMEM is treated as main memory and DRAM as cache. In such a setup, the processor's memory controller handles all cache management operations, meaning that when data is requested from memory, it first checks the DRAM cache, and if the data is present, transfer speed and latency are identical to DRAM. If not, the data in question is read from the PMEM modules with slightly longer latency. MM is the operational mode of choice in scenarios that benefit from an increase in main memory capacity. For example, in highly virtualized environments PMEM modules can host more VMs and at the same time provide more memory per VM, all for less money than you'd have to spend for a comparably efficient DRAM configuration. The same goes for highly I/O-oriented workloads such as extremely data-rich analytics and BI applications as well as a wide range of legacy software.
- By contrast, App Direct Mode enables applications to straightforwardly communicate with a percentage of Intel Optane PMEM. Thus, all crucial data is stored in persistent memory and kept in close proximity to the processor, and therefore immediately available whenever a piece of software boots up. In this mode, the OS recognizes two separate pools of memory and treats them accordingly: the PMEM part reserved for AD is used as persistent memory, while DRAM functions as main memory. A further advantage of such configurations is that AD-enabled PMEM is persistent like classic storage media, but at the same time byte-addressable like DRAM. On top of that, the data from the PMEM App-Direct space is cached in the CPU caches just like the data from the DRAM and is therefore part of the system coherence domain, which ensures that all CPU cores in the system immediately see the same data. In order to simplify the use of App-Direct, SNIA has developed an NVM programming model that closely follows the well-known storage APIs. This makes it easier for developers of new applications to take advantage of App-Direct storage. In addition, this approach also allows App-Direct Space to be mounted directly by the OS as a super-fast SSD, making it accessible to unmodified applications. In other words, AD Mode makes the most of PMEM in terms of performance, capacity, flexibility, and ease of use. It is therefore no surprise that experts consider it the operational mode of choice for scenarios where resilience of systems and applications as well as fast boot/restart times are of paramount importance. Likewise, users will profit from higher endurance and larger bandwidths than one would find in NAND-based SSDs. PMEM in AD configurations work especially well with all applications that rely heavily on memory capacity and performance, such as in-memory databases, in-memory analytics frameworks, and superfast storage applications.
- Mixed Mode, as the name implies, allows for Intel Optane PMEM to be configured in such a way that one part runs in Memory Mode and the other in App Direct Mode. At first glance, this may not seem to deviate much from the end result of an AD configuration; however, the key difference here is that the "split" in capacity and functionality is meant to be there from the get-go so as to support applications with differing requirements that must be run in parallel. – In Mixed Mode, applications can take advantage of high performance storage without the latency of moving data to and from the I/O bus.

SAP HANA and Intel Optane PMEM

As outlined above, there is a wealth of applications that stand to benefit from PMEM capabilities, especially if the new type of memory is set to operate in App Direct Mode. One of the key candidates in this context is SAP HANA, the popular in-memory database that was introduced at the start of the decade and nowadays underpins SAP's complete enterprise software suite including SAP S/4HANA. The main advantage users could draw from the new technology was that it enabled them to execute OLTP and OLAP applications simultaneously, while at the same time simplifying data analytics. Better still, they only needed to run one instance of one single database to get there – so they no longer needed to cope with the complexities of cluster configurations in which data is spread across several nodes.

To achieve this, SAP HANA keeps large chunks of data in a column store, where it can be exposed to vector processing for complex analytic queries. Likewise, it uses a row-oriented differential store to log data changes that result from transactions and incorporate them into the columnar organization. Moreover, HANA also supports data tiering capabilities to assign data to the right level of memory/storage, so that customers may further optimize database and application performance and gain a competitive edge, along with realizing cost benefits.

Thanks to its underlying design, SAP HANA practically lends itself to running on server systems fitted with Intel Optane PMEM. In particular, it will benefit from the fact that the main portion of data typically kept in the column store is transferred to the PMEM, whereas frequently accessed or "hot" information (row store data, the delta part of column store data, and the so-called "work area") remains in DRAM (cf. Fig. 4 and 5 below). Moreover, it could also use the larger PMEM capacity to its advantage, as more data is kept in closer proximity to the CPU in DRAM than in regular systems.

Fig. 4: Traditional DRAM memory layout and usage

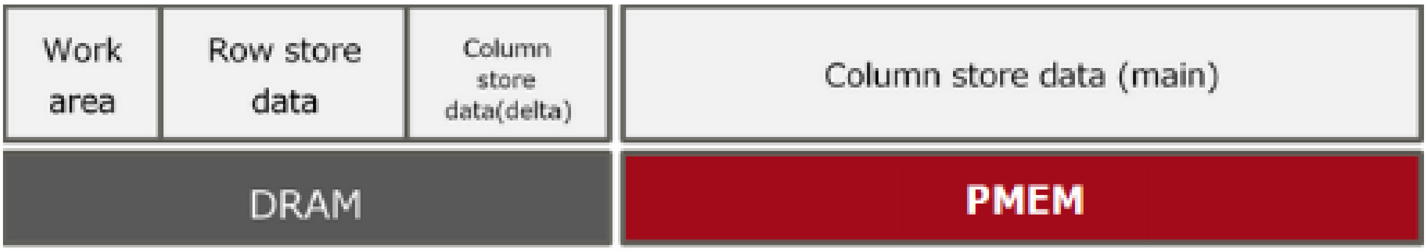
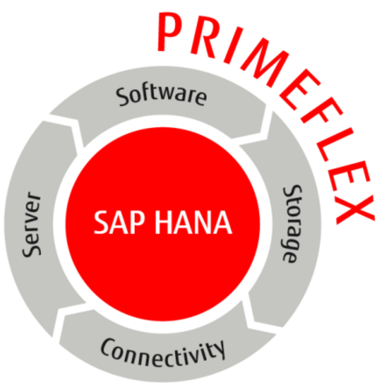


Fig. 5: PMEM-based memory layout and usage

FUJITSU Integrated System PRIMEFLEX for SAP HANA

SAP HANA is an innovative data platform optimized for performing real-time analytics and handling real-time transaction workloads. It is based on innovative in-memory technology which allows real-time access to massive amounts of data, by holding live data in the quickly accessible main memory.
The optimal solution is that Fujitsu's portfolio includes PRIMEFLEX, a pre-defined and pre-tested solution for SAP HANA that integrates servers, storage, network components and software. PRIMEFLEX for SAP HANA enables simplified, fast and secure implementation and operation of SAP HANA. The infrastructure solution is based on SAP-certified components and supplemented by a broad services portfolio. It covers everything, from pre-installed scale-up systems via VMware virtualized platforms and individual scale-up and scale-out concepts in line with the SAP HANA Tailored Data Center Integration (TDI) approach right through to customized disaster-tolerant set-ups. PRIMEFLEX for SAP HANA helps customers to fully exploit the potential of SAP HANA and to accelerate and innovate their business processes.



Benchmark Achievements on PRIMEFLEX for SAP HANA with PMEM

Fujitsu started a series of SAP BW edition for SAP HANA Standard Application Benchmarks Version 3 using PMEM since it has been introduced by Intel with the 2nd generation Xeon Scalable Family processors (Cascade Lake-SP).
The first benchmark publication with 9.1 billion initial records was on FUJITSU Server PRIMERGY RX4770 M5, equipped with 4 processors / 112 cores / 224 threads Intel Xeon Platinum 8280L processor, 4.5 TB of memory – 1.5TB DRAM and 3.0 TB PMEM. This was followed by three further publications with very large data volumes of up to 31.2 billion initial records on FUJITSU Server PRIMEQUEST 3800B2 with 8 processors / 224 cores / 448 threads Intel Xeon Platinum 8280L processor and up to 18.0TB of memory – 6.0TB DRAM and 12.0 TB PMEM. Figure 6 below shows the benchmark configurations and table 1 the benchmark publications and achievements.

FUJITSU SERVER PRIMERGY RX4770 M5

- 4 processors, Intel® Xeon® Platinum 8280L, 112 cores, 224 threads
- 1,536 GB DRAM plus 3,072 GB Persistent Memory
- SUSE Linux Enterprise Server 12
- SAP Netweaver 7.50
- SAP HANA 2.0

FUJITSU SERVER PRIMEQUEST 3800B2

- 8 processors, Intel® Xeon® Platinum 8280L, 224 cores, 448 threads
- 3,072 GB DRAM plus 6,144 GB Persistent Memory resp. 6,144 GB DRAM plus 12,288 GB Persistent Memory
- SUSE Linux Enterprise Server 12
- SAP Netweaver 7.50
- SAP HANA 2.0

Fig. 6: SAP BW edition for SAP HANA Standard Application Benchmark Version 3 configurations

Fujitsu Server	CPU	Memory size used for SAP HANA			Initial records (billions)	Phase 1 Runtime of last data set (seconds)	Phase 2 Query executions per Hour	Phase 3 Runtime of complex query phase (seconds)	Certification number
		DRAM	PMEM	Total					
PRIMERGY RX4770 M5	4x Intel 8280L	1.5TB	3.0TB	4.5TB	9.1	25111	3478	171	2019012
PRIMEQUEST 3800B2	8x Intel 8280L	3.0TB	6.0TB	9.0TB	15.6	18267	4698	147	2019058
PRIMEQUEST 3800B2	8x Intel 8280L	3.0TB	6.0TB	9.0TB	20.8	19635	3621	168	2019059
PRIMEQUEST 3800B2	8x Intel 8280L	6.0TB	12.0TB	18.0TB	31.2	22129	2676	205	2020016

Table 1: SAP BW edition for SAP HANA Standard Application Benchmark Version 3 achievements

All BWH benchmarks were certified on a 3-tier configuration and are listed at https://www.sap.com/dmc/exp/2018-benchmark-directory/#/bwh .

In particular the PRIMEQUEST 3800B2 results show the scale-up capabilities of a single-node setup in such scenarios up to 31.2 billion records and 18TB of memory. This opens up new opportunities to consolidate existing scale-out configurations to reduce complexity and costs.

The SAP BW edition for SAP HANA Standard Application Benchmark Version 3 consists of 3 phases:

- Data load phase (phase 1)
  The data flow starts with a data load from the source object into the corporate memory layer. The source object is shipped with the backup and contains 1.3 billion records (= 1 data set). It is possible to load this data set of 1.3 billion records multiple times. The data load phase is a combination of CPU- and IO-intensive load.
- Query throughput phase (phase 2)
  The queries for the throughput phase must be executed via an ABAP program with a variant containing 190 query steps. Users execute the set of navigation steps in random order (via asynchronous RFCs). The queries contain typical query patterns which can be found in BW productive systems of customers. The query throughput phase runs one hour and is CPU bound.
- Query runtime phase (phase 3)
  For the query runtime phase the same ABAP program as for the throughput phase is used with a different variant. The variant contains 10 queries which are executed sequentially. These queries are used to measure the runtime. They contain complex query patterns which are executed in BW productive systems of customers but which are typically not executed by many users in parallel but selectively by some power users. Therefore they are executed sequentially. The query runtime phase takes a short time and generates a small load. Only a few processors cores are used, single thread performance is important for short runtimes.

SAP BW edition for SAP HANA Standard Application Benchmark results with different number of data sets cannot be compared.

Performance Test on PRIMEFLEX for SAP HANA with PMEM

Internal tests by Fujitsu have proved that performance losses/performance lowering of SAP HANA are/is in fact minimal where servers with PMEM configurations are used. The results have been confirmed by SAP. The tests were carried out on two PRIMERGY RX4770 M5 server models, one equipped with DRAM only and one with PMEM and DRAM. For the exact configurations, please refer to the graphics and tables below.

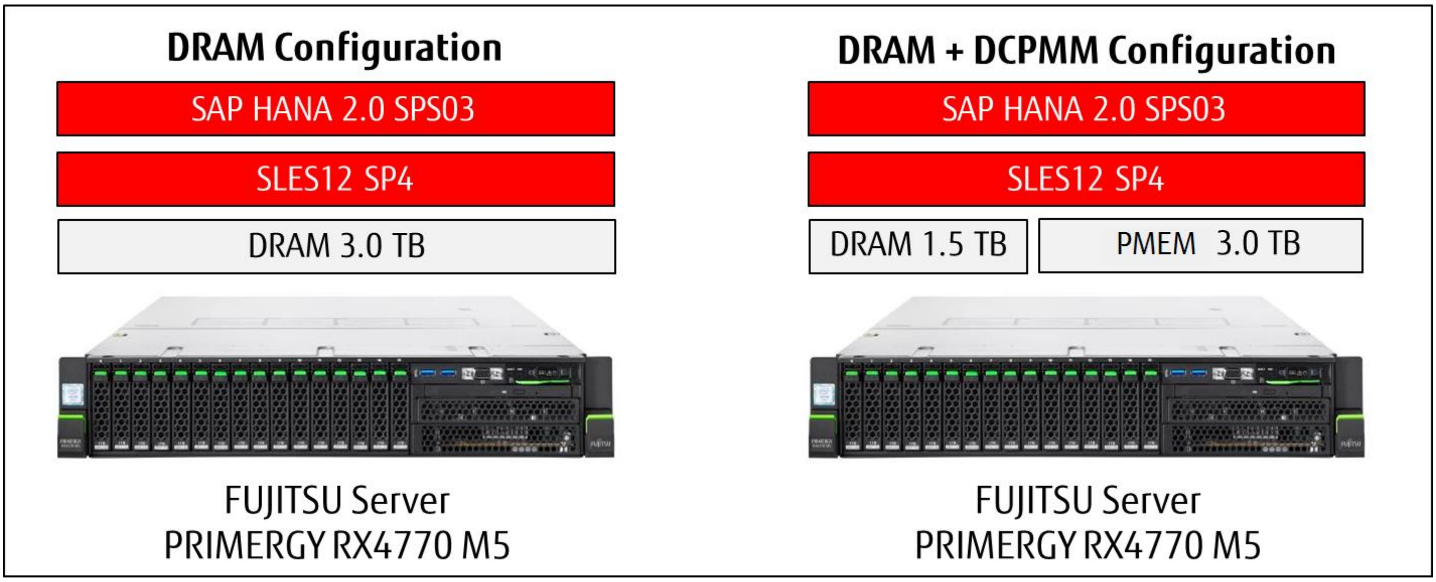


Fig. 7: Tested server configurations

	SAP HANA server with DRAM	SAP HANA server with DRAM+PMEM
Server	FUJITSU Server PRIMERGY RX4770 M5	FUJITSU Server PRIMERGY RX4770 M5
Processor	Intel Xeon Platinum 8280L @ 2.70 GHz x 4	Intel Xeon Platinum 8280L @ 2.70 GHz x 4
Memory	DRAM 64GB x 48 (3.0TB)	DRAM 64GB x 24 (1.5TB)
Non-volatile memory	-	PMEM 128GB x 24 (3.0TB)
OS	SUSE Linux Enterprise Server 12 SP4	SUSE Linux Enterprise Server 12 SP4
Database	SAP HANA 2.0 SPS03 Rev36	SAP HANA 2.0 SPS03 Rev36

Table 2: Test configuration details

Test Scenario

The test was designed to determine how using PMEM instead of DRAM modules influences server and DB performance. In particular, we wanted to establish the following characteristics:

- Time to load memory
- Performance impact in parallel queries
- Performance impact in sequential queries
- Time reduction of SAP HANA systems restart

Based on these aims, we took the following steps:

- Test1 measured memory load times for large data sets
- Test2 measured throughput in parallel queries
- Test3 measured the execution time of sequential queries
- Test4 measured the restart times of SAP HANA systems

Test Results

Table 3 below shows the performance figures for PMEM configurations in comparison with the values for DRAM configurations. For tests 1, 2, and 3, we noted the relative performance losses or gains compared to DRAM configurations. With regard to test 4, we give you the restart time in minutes. Fig. 8 shows the results for tests 1 through 3 in bar graphs for easier comparison.

	Memory size used for SAP HANA			Test1 Memory load speed for large data sets	Test2 Throughput in parallel queries	Test3 Execution speed of sequential queries	Test4 Restart time
	DRAM	PMEM	Total				
DRAM configuration (DRAM 3.0TB)	1,850GB	–	1,850GB	100%	100%	100%	11 min.
DRAM+PMEM configuration (DRAM 1.5TB + PMEM 3.0TB)	780GB	1,147GB	1,927GB	88%	96%	101%	2 min.
Remarks				Higher values are better	Higher values are better	Higher values are better	Shorter times are better

Table 3: Results of performance tests 1 through 4

Results of performance test 1 through 3 on DRAM and PMEM configurations

As you can see, performance during the memory load and parallel queries tests is influenced by the speed at which data in different layers/tiers of memory can be accessed. As a result, the relevant relative performance of the PMEM configuration was lower than that of its DRAM counterpart; specifically, load times turned out to be 12% longer, while throughput difference amounted to 4%. Consequently, the relative performance of the PMEM configuration in this test was similar to that of DRAM configuration – actually even slightly better, as shown in Table 3.

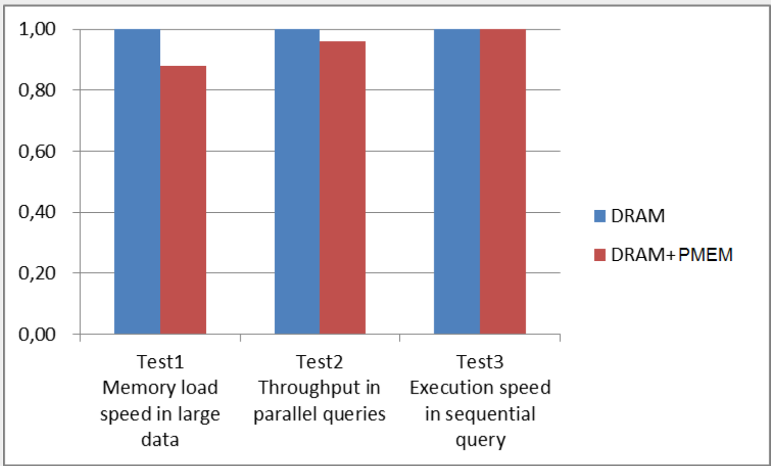


Fig. 8: Results of performance test 1 through 3 on DRAM and PMEM configurations

Test 4 was carried out to determine the restart times of both configurations following a power failure or regular restart. In this scenario, the PMEM configuration only took 2 minutes to restart SAP HANA, coming in far ahead of the DRAM setup with 11 minutes. In this case, keeping all relevant data stored in PMEM means that the system does not have to "re-fetch" this information from an SSD or HDD and load it into main memory. Instead, it will be available immediately after the boot process has finished, leading to a massive performance advantage.

Intel Select Solutions

SAP HANA server with DRAM+PMEM is verified as Intel Select Solutions. Intel Select Solutions are pre-defined, workload-optimized solutions designed to minimize the challenges of infrastructure evaluation and deployment. Solutions are developed by Fujitsu and verified by Intel, certified by ISVs as necessary. Intel and Fujitsu develop these solutions in extensive collaboration from hardware, software, and operating system perspectives. Every Intel Select Solution is a tailored combination of Intel data center compute, memory, storage, and network technologies that delivers predictable, trusted, and compelling performance.

Our Solution for SAP HANA Business Operations (based on PRIMEFLEX for SAP HANA) is based on the FUJITSU Server PRIMERGY RX4770 M5, the PRIMEFLEX for SAP HANA integrated solution and the Intel Select Solution specifications. The performance results of the solution have been verified by Intel.

INGREDIENT		Intel® Select Solution for SAP HANA Base Reference Architecture - 1 Node		
PLATFORM		Quad-socket server platform, FUJITSU Server PRIMERGY RX4770 M5		
PROCESSOR		4 x Intel® Xeon® Platinum 8260M* or 8276M or 8280M processors or 4 x Intel® Xeon® Gold 6238M* or 6240M* processors		
MEMORY		**4.5 TB (Ratio 1:2)** 24 x 128GB Intel® Optane Persistent Memory => 3072GB + 24 x 64GB 4Rx4 DDR4-2933 LR ECC => 1536 GB	**6.0 TB (Ratio 1:1)** 24 x 128GB Intel® Optane Persistent Memory => 3072GB + 24 x 128GB 4Rx4 DDR4-2933 LR ECC => 3072GB	**7.5 TB (Ratio 1:4)** 24 x 256GB Intel® Optane Persistent Memory => 6144GB + 24 x 64GB 4Rx4 DDR4-2933 LR ECC => 1536 GB
STORAGE	Boot Drive	2* SSD SATA 6G 240GB M.2 (mirrored)		
	HANA data & log, incl. hotspare	Internal disks or ETERNUS JX40 S2 (JBOD) - 16* TLC SSD 1.6TB (3DWPD)	ETERNUS JX40 S2 (JBOD) 20* TLC SSD 1.6TB (3DWPD)	ETERNUS JX40 S2 (JBOD) 16* TLC SSD 3.2TB (3DWPD)
NETWORK		Onboard High Performance Chip Intel LBG4 with flexible LAN connections (2*10GB / 4*1GB) optional: 2* 100 Gbit Ethernet Adapters (e.g. used in Scale-Out)		
SOFTWARE		OS (minimum or higher): SLES4SAP 12 SP4 or RHEL4SAP 7.6 SAP HANA 2.0 SPS 03 or higher Firmware/Software optimizations as described in SAP HANA manuals and recommended by SAP / Intel		

Summary

The SAP BW edition for SAP HANA benchmark achievements on FUJITSU Server with Intel Optane PMEM, in particular on PRIMEQUEST 3800B2, show the scale-up capabilities of a single-node setup up to 31.2 billion records and 18TB of memory.

Our performance comparison tests of SAP HANA on a FUJITSU Server PRIMERGY RX4770 M5 model equipped with Intel Optane PMEM have confirmed minimal performance disadvantages regarding memory load times for large data sets and throughput in parallel queries when compared to a DRAM-only configuration. Moreover, the PMEM setup did equally well or slightly better with regard to sequential queries and outperformed its counterpart with regard to restart times.

Intel Optane persistent memory is blurring the line between DRAM and persistent storage for in-memory computing. Unlike DRAM, Intel Optane persistent memory retains its data if power to the server is lost or the server reboots, but it still provides near-DRAM performance. In SAP HANA operation this results in tangible benefits:

- Speed up shutdowns, starts and restarts many times over – significantly reduce system downtime and lower operational costs
- Process more data in real-time with increased memory capacity
- Lower total cost of ownership by transforming the data storage hierarchy
- Improve business continuity with persistent memory and fast data loads at startup

Contact
FUJITSU Technology Solutions GmbH
E-mail: server-marketing@ts.fujitsu.com
Website: www.fujitsu.com/
2020-08-10