Case Study

Intel® Optane™ Persistent Memory
Intel® PMDK
Artificial Intelligence

Intel® Optane™ Persistent Memory Empowers 4Paradigm's All-Process AI System with Both Capacity and Persistence

Enterprise-grade Artificial Intelligence (AI) systems are now playing an increasingly important role in the day-to-day business operations. They provide businesses with the unique ability to extract and construct high or ultra-high dimensional features/models from a sea of business data, allowing for the delivery of exceptional online inferencing services, thereby assisting with more efficient and sophisticated decision-making in business scenarios such as fraudulent transaction identification and personalized recommendation.

The explosive growth in the scale of data represents both an opportunity and challenge for AI systems. Businesses can exploit the data to construct even more massive high-dimensional features/models, but this also requires more space and computing power to store and process them. If Dynamic Random Access Memory (DRAM) is used exclusively as memory to hold such data, it will significantly increase the Total Cost of Ownership (TCO). This also results in excessively long data recovery time when power outage or downtime inevitably happens, leading to problems such as online inferencing interruption and degradation in the quality of service (QoS).

To help enterprise users overcome these challenges, Beijing 4Paradigm Technology Co., Ltd. (4Paradigm), a leading AI platform and service provider, has worked with Intel to introduce Intel® Optane™ persistent memory (PMem) to its proprietary SageOne AI computing platform. The optimization takes advantage of PMem's higher storage density, data persistence, DRAM-level I/O performance, and lower cost, creating a new HyperCycle Machine Learning all-process AI system for enterprise users.

By taking advantage of Intel Optane PMem's two primary operating modes (namely Memory mode and APP Direct or AD mode), the new system not only features a brand-new re-designed workflow, but also a thoroughly optimized HyperPS trillion-dimensional parameter server and OpenMLDB machine learning database[1]. Tests verified that while maintaining high-performance inferencing as was with pure DRAM, data recovery time with the new system has been significantly shortened, and the number of memory servers used is reduced. The solution also excelled in terms of service continuity and TCO. If the system hardware is further upgraded to the latest 3rd Generation Intel® Xeon® Scalable processor (formerly codenamed Ice Lake), used in combination with Intel Optane PMem 200 series, the new AI system will offer even greater performance advantages.

"The implementation of all-process AI system requires both high performance and high availability to shorten the data recovery time in the case of anomaly and to guarantee the quality of online services. The high density and data persistence of Intel® Optane™ persistent memory can help users meet these requirements all while lowering TCO significantly."

Zhao Zheng
Vice President
4Paradigm

Advantages that Intel Optane PMem has brought to 4Paradigm's brand new AI system:

- **Greater storage density:** Intel Optane PMem greatly reduces the unit price of memory resources, and thus helps lower the TCO for system users. In some scenarios, the cost in construction can be cut by approximately 60%[2];
- **Data persistence:** Intel Optane PMem empowers a breakthrough in data recovery speed for systems: service recovery time can be reduced from several hours to several minutes[3], providing better assurance of service continuity.
- **Queue performance:** The synergy created with the persistence feature can significantly increase the throughput of individual servers and optimize the performance of message queue on Kafka servers. The number of servers needed can thus be greatly reduced under the same throughput bandwidth.
- **Further enhancements to service performance:** Switching to the new 3rd Generation Intel Xeon Scalable processor and Intel Optane PMem 200 series provides a further boost to OpenMLDB's request/sending latency and throughput performance. Request latency can be reduced by up to 23.5%, while throughput increased by up to 27.6%[4];
- **Better price–performance:** Having an I/O performance similar to that of pure DRAM configuration means the performance of the new system based on Intel Optane PMem is only slightly lower than the pure DRAM-based solution, yet it's completely sufficient to fulfill the expectations of enterprise users.

Background: Service continuity and TCO — two challenges facing enterprises seeking to develop high-quality online inferencing services

Online AI inferencing services are gaining more importance in the management and decision-making processes across industries including financial, healthcare, retail, and manufacturing. Massive amounts of business data are the cornerstone of high-quality online inferencing. Their high-dimensionality and sparse nature are two important factors that enterprise users must consider when designing related AI systems.

Taking credit card risk control as an example, each set of user data can be categorized by status (e.g., whether it is linked to a debit card; whether it is a key account, etc.), behavior (usage frequency, location, etc.), and other attributes. Each category in turn can be categorized further into a number of sub-categories. For an Individual user, there can be even a greater variety of personalized attributes (e.g., age, gender), the combination of which can render hundreds of millions or even billions of data dimensions. Amidst the data, the hit rate for a target feature (e.g., malicious cash-out behavior) can be extremely low, leading to the typical sparse nature of the data.

Outside of financial risk management, online inferencing systems used in online recommendation and medical screening are also facing similar problems. Indeed, inference accuracy can be improved with high-dimensional and ultra-high dimensional feature engineering and modeling based on high-dimensional, sparse data. If the data volume is exceedingly large, however, this imposes more demands on the design and computing power of the system framework and algorithms used. Moreover, memory, a module less of a concern in the past, will rise to a key, or even the technical focus in the all-process AI system.

In the past, enterprises often adopted a distributed computing architecture using DRAM to satisfy the memory requirements of real-time data processing, and of ultra-high dimensional features and models. To overcome the data volatility of DRAM, multiple cluster replication was used to guarantee continuity and quality of service; however, the continued expansion in the scale of business also meant a rapid growth in the size and quantity of high-dimensional features and models. As a result, the TCO and memory hardware resources required for AI systems running online inferencing services see an exponential growth as well.

The old approach of backing up data to external storage devices such as SSDs and HDDs ran into new problems, too. Basically, in the case of system anomaly, such as software/hardware failure or downtime, the high-dimensional models need to be copied from external storage devices with lower performance back into the memory, following which the workflow goes into the recovery process based on the requirements of specific service scenarios; however, as these models grow larger in size, this process can take several hours. In some scenarios, this may even lead to a serious long-tail latency. For industries such as finance, healthcare, and e-commerce that require extremely high online service quality, this will seriously undermine QoS and user satisfaction.

Solution: An improved all-process AI system based on Intel Optane persistent memory

To fully address the above-mentioned challenges, 4Paradigm has partnered with Intel to develop the HyperCycle Machine Learning all-process AI system on its proprietary SageOne AI computing platform. With a design integrating hardware and software, the new system based on Intel Optane PMem features higher performance, higher service continuity and lower TCO, delivering much better online inferencing services for enterprise users. 4Paradigm expects the system to provide the following advantages:

- **Better performance:** The system shall be seamlessly compatible with 4Paradigm's proprietary high dimensionality machine learning framework and other open-source frameworks such as TensorFlow, delivering performance highly comparable to that of the pure DRAM configuration and capable of handling hundreds of millions of KV queries per second.

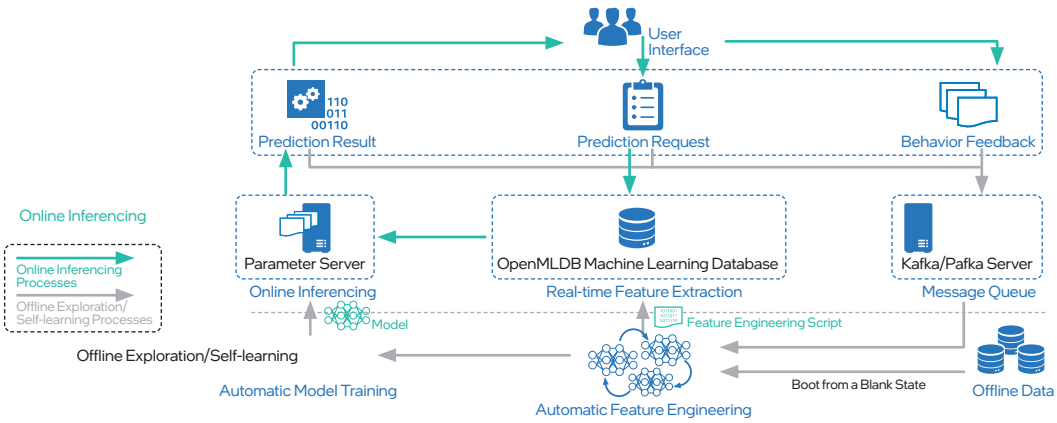

Figure 1. HyperCycle Machine Learning all-process AI system architecture

- **Better service continuity:** In the event of power outage, physical malfunction or software crash, the system shall ensure better online service continuity by shortening the data recovery time of the online inferencing system to several minutes.
- **Lower TCO:** The system shall lower TCO by effectively helping enterprise users reduce their system procurement, deployment, O&M, and administration costs.

To realize the three goals, as shown in Figure 1, the new all-process AI system architecture is broken down into two function zones: "online inferencing" and "offline exploration/self-learning". The workflow can be divided into the following parts:

- **Offline exploration:** When the system is booted from a blank state (meaning no models previously accumulated), it starts by carrying out automatic feature engineering (including auto-combining and auto-trimming of features) and automatic model training (including operations such as auto model selection and parameter adjustment) using offline data. The initial feature engineering script and model are thus generated.
- **Online inferencing:** Once initialization is completed, the user can submit online prediction requests through web or app interfaces. The system will carry out real-time feature extraction using the feature engineering script and the historical data stored in the feature engineering database (e.g., historical transaction records). After that, the results are sent to the parameter server for online inferencing, and the final predictions are then returned to the user.
- **Self-learning:** Using the Kafka server, the system can also form the prediction results, prediction requests, and behavioral feedback data into a message queue. Based on the pre-defined threshold or frequency, the data in the queue is channeled into the automatic feature engineering and automatic model training component. The feature engineering script and model are then updated to the online service flow to form a complete loop in the online inferencing system.

Workflow	Application Requirements	PMem Operating Mode	Application Highlights
Offline Exploration/ Self-learning	High-dimensional model processing requires higher memory capacity.	Memory mode	▪ In memory mode, Intel Optane PMem can work alongside with DRAM to greatly increase the memory capacity (PMem can provide up to 512GB memory on a single DIMM), with no need to adjust the system architecture or software. ▪ In the new solution, 4Paradigm chose to use Intel Optane PMem in combination with Intel Xeon Scalable platform. This provides a robust hardware infrastructure to support the massive computing involved in automatic feature engineering and model training.
Message Queue	Data writing performance needs to meet the requirement of feature engineering and training tasks.	AD mode	▪ The innovative storage media of Intel Optane PMem enables a huge performance breakthrough. In AD mode, PMem can access persistent memory address space through API, which delivers high read/write performance, while reducing the latency from moving data in and out of the I/O bus. ▪ Intel Optane PMem in AD mode features not only the similar persistence to conventional SDDs and HDDs, but its I/O performance also approaches that of DRAM. ▪ The new solution adopts Intel Optane PMem in AD mode to perform data writing in message queue, which has effectively reduced the I/O bottleneck.
Online Inferencing	Online inferencing needs to both satisfy the large memory capacity requirement of Open MLDB and parameter server and accelerate data recovery to guarantee service continuity.	AD mode	▪ As the OpenMLDB and parameter server are built upon memory solutions, they need larger memory capacity with the rapid growth of businesses. Intel Optane PMem in AD mode can be inserted directly into standard DIMM to provide the above-mentioned components with more cost-effective large-capacity memory support. ▪ The data persistence of Intel Optane PMem in AD mode allows the system to save data in persistent memory, without the need of reloading in the event of power outage or downtime. This significantly speeds up service recovery. In some typical scenarios, the overall recovery time can be shortened from several hours to several minutes. At the same time, the combination of high density and persistence features means there is no longer any need to back up data in external storage devices with lower I/O performance. This helps with the problem of long tail latency.

Table 1. Intel Optane persistent memory can be used in different modes to satisfy the requirements of different tasks.

High performance, high density, and persistence enhance the efficiency of all-process AI system processes

To optimize the performance and satisfy the requirements on data continuity and TCO, 4Paradigm has partnered with Intel, using Intel Optane persistent memory in different modes to tailor to the needs and characteristics of the various components involved in the aforementioned processes (see Table 1).

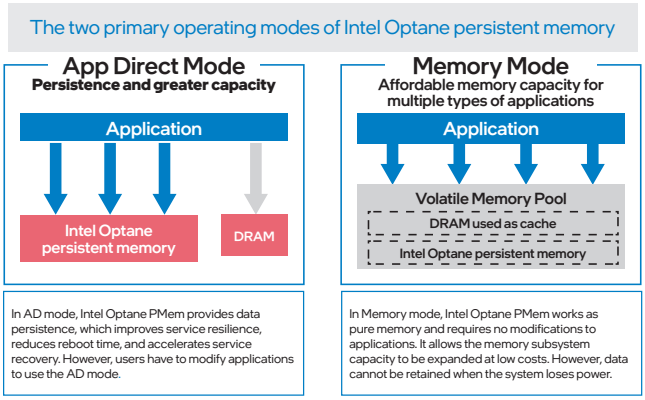

Figure 2. Two primary modes of Intel Optane persistent memory

HyperPS trillion-dimensional parameter server, OpenMLDB, and Pafka, a message queue-optimized version of Kafka

After the introduction of Intel Optane PMem, 4Paradigm continued working with Intel to optimize the HyperPS trillion-dimensional parameter server, OpenMLDB and Kafka message queue.

Increasing the efficiency of the parameter server

In addition to imbuing the new parameter server with features such as high-performance underlying serialization framework, link sharing, multi-level memory access optimization, double-ended parameter combination, and dynamic capacity scaling, 4Paradigm has also leveraged the advantages of Intel Optane PMem, using the programing models, environments, and extensive function libraries and tools provided by Intel® Persistent Memory Development Kit (PMDK) to optimize the underlying system architecture and access performance on both hardware and software levels. These optimizations include:

- A new storage engine specially designed for the shards in the node. It uses persistent hash table as the storage feature of the underlying data structure, which ensures a high level of parallelism and optimizes the data organization format based on Intel Optane PMem, securing the hash table performance approaching that of pure DRAM.
- Disaster recovery mechanism of the parameter server tuned based on Intel Optane PMem. Persistent smart pointers are used to record the root pointer and the core data structure of the hash table to achieve an industry-leading real-time recovery capability of the parameter server.
- Ensured data consistency in Intel Optane PMem. The transaction mechanism provided by PMDK and the pmempool function library have been used for the allocation and management of persistent memory space to ensure data consistency.
- Finally, persistent operations that may affect performance are strategically reduced to further increase the parameter server performance.

Increasing the efficiency of OpenMLDB

After the introduction of Intel Optane PMem, 4Paradigm further made innovative optimizations on the design of OpenMLDB. In addition to efficient computing, read/write separation, high parallelism, and high query performance, the optimized OpenMLDB now takes full advantage of PMem's high density and persistence feature. Performance losses from persistent operations are minimized, so that users could upgrade their legacy data persistence architecture based on conventional external storage devices such as SDDs and HDDs to a new data persistence architecture based on Intel Optane PMem in AD mode. These innovative design optimizations include:

- The introduction of persistent smart pointers, which cleverly leverage the unused lower 4 bits from the 64-bit pointer of the 64-bit operating system to flag whether data in the target address is persistent.
- A new "flush-before-read" concept introduced to the Compare-And-Swap (CAS) command, to fix the problem in the original CAS command where the lack of persistent semantics made persistent atomic operations impossible to perform directly in Intel Optane PMem.

Pafka: Kafka optimized for Intel Optane persistent memory

Pafka (PMem accelerated Kafka), an optimized version of Kafka based on Intel Optane persistent memory, can modify the data structure on existing storage, allowing segments that previously could only be stored on conventional persistent devices, such as SDDs and HDDs, to be stored on Intel Optane PMem through persistent operations with PMDK. For enterprise users, this offers the following advantages:

- Significantly improved throughput and latency per socket per server compared with conventional SSDs and HDDs;
- Higher single-node throughput, which will in turn reduce the hardware investment that a business must make in message queue clusters;
- Pafka is a modified version of Kafka, so users have no need to change their existing code when migrating to the new system.

Please refer to https://github.com/4paradigm/pafka for more information

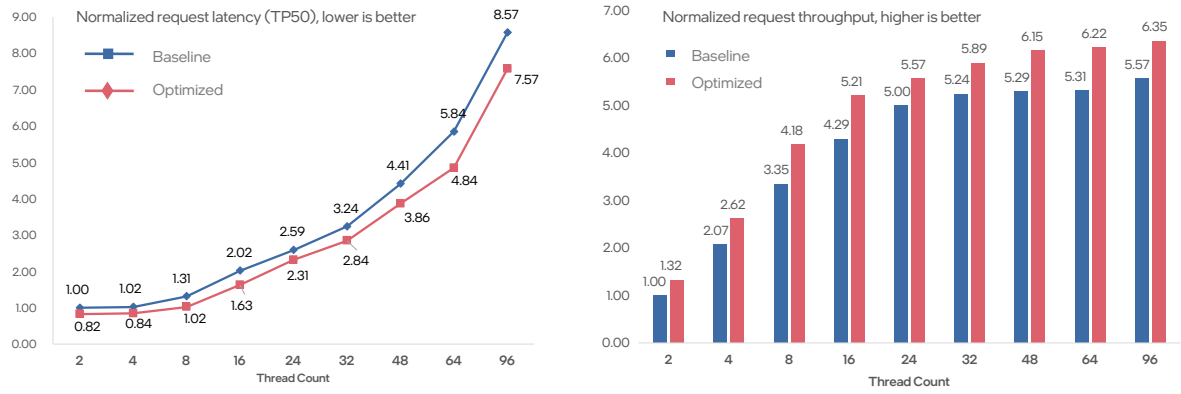

Figure 3. The new hardware configuration has helped boost the request performance of OpenMLDB

User benefits: Enhanced performance and significantly reduced data recovery time and TCO

Up till now, 4Paradigm's new all-process AI system has already been deployed and used in a number of fields. Its outstanding online inferencing performance has been well received among users. In particular, with the support from the newly released 3rd Generation Intel Xeon Scalable processor, the new system is supercharged with new power to enhance its service performance. When used in combination with Intel Optane persistent memory 200 series featuring advanced hardware architecture and software optimization techniques, the new system has been further optimized with regard to the efficiency of its various functional modules, leading to an overall service performance improvement. The details of performance gain are as below.

Service performance: Compared with the previous generation, the introduction of 3rd Gen Intel Xeon Scalable processor and Intel Optane PMem 200 series has seen significant performance improvements in the core OpenMLDB in terms of the latency and throughput performance in request and sending.

- Request latency (TP50): Figure 3 (left) compares the request latency (TP50) performance under different thread counts between the optimized configuration with 3rd Gen Intel Xeon Scalable processor and Intel Optane PMem 200 series against the baseline configuration consisting of 2nd Gen Intel Xeon Scalable processor and Intel Optane PMem 100 series. Performance was increased by 10.5% to 23.5%[5].
- Request throughput: Figure 3 (right) compares the throughput performance under different thread counts between the optimized hardware configuration against the baseline. Performance was increased by 11.9% to 27.6%[6].
- Sending latency (TP99): Figure 4 (left) compares the sending latency (TP99) performance under different thread counts between the optimized hardware configuration and the baseline. Performance was increased by 5.8% to 43.8%[7].
- Sending throughput: Figure 4 (right) compares the throughput performance under different thread counts between the optimized hardware configuration against the baseline. Performance was increased by 21.4% to 36.9%[8].

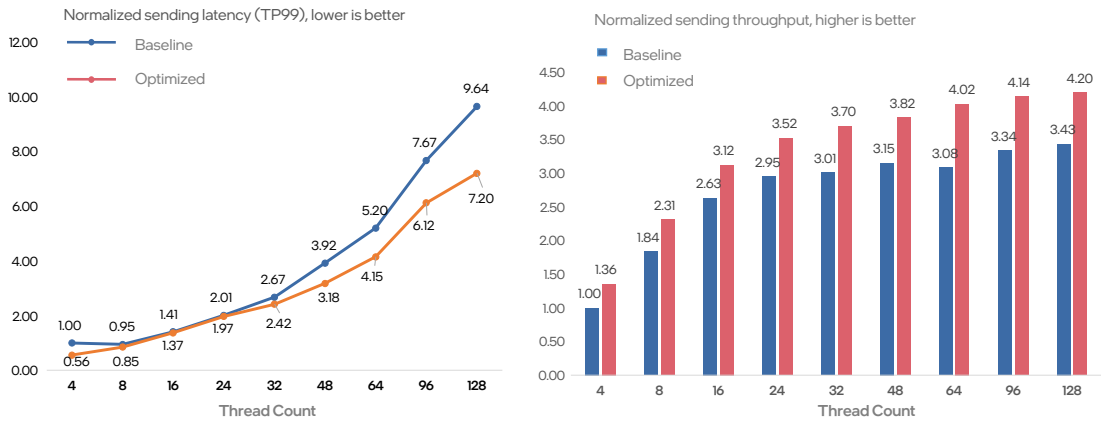

Figure 4. The new hardware configuration has helped boost the sending performance of OpenMLDB

The service performance of the optimized hardware configuration is no less performant compared with the baseline pure DRAM configuration. In some tests, the TP50, TP99, and TP9999 latency performance of OpenMLDB based on the new hardware configuration is shown highly comparable or even exceeding that of pure DRAM[9].

Service continuity: The optimized parameter server and OpenMLDB feature much shortened data recovery time. In some tests for the parameter server, data can be recovered within milliseconds[10]. The recovery time of the optimized OpenMLDB can also be reduced from hours to minutes in certain scenarios.

Figure 5 shows the result of a user verifying this new online inferencing system in real-world deployment. It was determined that the data recovery time of the OpenMLDB has been shortened from 373.33 minutes (using pure DRAM before optimization) to only 1.07 minutes (using Intel Optane PMem after optimization), a reduction of up to 99.7%[11].

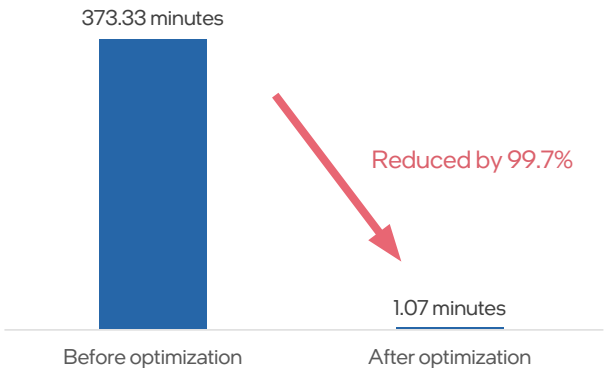

Figure 5. OpenMLDB data recovery time in a user's real-world deployment has been greatly reduced with Intel Optane persistent memory

TCO: The new solution based on Intel Optane PMem can help enterprise users achieve significant cost savings without sacrificing performance and QoS. In an anti-fraud deployment used by a bank, as shown in table 2, servers required for the deployment significantly reduced from up to 70 (with pure DRAM configuration) to 24 (with the introduction of the new solution with Intel Optane PMem), a drop of 65.7%[12]. It is estimated that procurement costs alone can be cut by about 60%[13].

Online Estimation System Components	Total Memory Required	No. of Servers Configured with DRAM	No. of Servers Configured with Intel Optane PMem
Automatic Feature Engineering + Automatic Model Training	8 TB	20	6
Parameter Server	8 TB	20	9
OpenMLDB	10 TB	25	8
Kafka Server	-	5	1
Total	**-**	**70**	**24**

Table 2. Cost-benefit comparison of using Intel Optane persistent memory and DRAM

Looking ahead

The latest generation online inferencing system from 4Paradigm has won unanimous praise from users in different industries thanks to its superior performance in real-world applications. With the success of the project, 4Paradigm and Intel are now planning to engage in more in-depth collaboration in the following areas:

- Further explore AI solutions optimized with Intel Optane PMem in large-scale and commercial applications to accelerate the convergence of advanced software/hardware products, AI technologies, and industrial implementations;
- Further optimize the performance of Intel Optane PMem-based online evaluation systems on 3rd Generation Intel Xeon Scalable processors.
- Promote the open-source development of Kafka, HyperPS trillion-dimensional parameter server, and openMLDB technologies optimized based on Intel Optane PMem. Specific tasks involve:
  - Pafka (Kafka optimized for Intel Optane persistent memory): Enable the persistent storage of message queue data on Intel Optane PMem through optimized data structure using Intel PMDK, to eliminate performance bottlenecks and reduce hardware costs. For more information, please visit: *https://github.com/4paradigm/pafka*.
  - PmemStore, the underlying persistent data structure for OpenMLDB based on Intel Optane PMem: This was decoupled by 4Paradigm for third parties to use as an independent storage engine. Meanwhile, the new data structure is also supported by Intel PMDK. For more information, please visit: *https://github.com/4paradigm/pmemstore*.

In the future, both parties will engage in further collaborations that combine Intel's innovative software/hardware products and technologies with 4Paradigm's technological edge in AI R&D to accelerate the innovation and implementation of AI on enterprise scale.

[1] For more information on OpenMLDB, please visit: https://github.com/4paradigm/OpenMLDB

[2, 12, 13] Data based on internal testing and verification by 4Paradigm. DRAM-based server configured with 512 GB DRAM + 4* SATA SDDs. Intel Optane PMem-based server was configured with 1.5TB (12 * 128GB). For more information, please visit 4Paradigm's website: https://www.4paradigm.com

[3] Data based on internal testing and verification by 4Paradigm. For more information, please visit 4Paradigm's website: https://www.4paradigm.com

[4, 5, 6, 7, 8.] Test configuration: Baseline configuration: Single-node 2S Intel Xeon Platinum 8260M processor @ 2.40 GHz, 24 cores, 48 threads. Hyperthreading and Turbo Boost are enabled. Memory: DRAM 32GB*12 (DDR 2666MHz), Intel Optane PMem 128GB*12 (2666 MHz). BIOS: SE5C620.86B.02.01.0012.070720200218 (microcode: 0x5002f01). Operating system: CentOS Linux release 8.2.2004 (Core). Kernel: 4.18.0-193.el8.x86_64 Additional configuration information: jdk1.8.0_121, libpmemobj-cpp-1.12, pmdk Ver. 1.9.2, pmemkv Ver. 1.4 Optimized configuration: Single-node 2S Intel Xeon Platinum 8352Y processor @ 2.20GHz, 32 cores, 64 threads. Hyperthreading and Turbo Boost are enabled. Memory: DRAM 16GB*16 (DDR 3200MHz), Intel Optane PMem 200 series 128GB*16 (3200 MHz). BIOS: WLYDCRB1.SYS.0020.P96.2104060045 (microcode: 0xd000280). Operating system: CentOS Linux release 8.2.2004 (Core). Core: 4.18.0-193.el8.x86_64. Additional configuration information: jdk1.8.0_121, libpmemobj-cpp-1.12, pmdk Ver. 1.9.2, pmemkv Ver. 1.4.

[9] For more information, please see: Cheng Chen, Jun Yang, Mian Lu, Taize Wang, Zhao Zheng, Yuqiang Chen, Wenyuan Dai, Bingsheng He, Weng-Fai Wong, Guoan Wu, Yuping Zhao, and Andy Rudoff,"Optimizing In-memory Database Engine For AI-powered On-line Decision Augmentation Using Persistent Memory"https://dl.acm.org/doi/10.14778/3446095.3446102

[10] Results sourced from public media report: https://www.163.com/tech/article/FGCFSO4N00099A7M.html

[11] Results sourced from public media report: https://newsroom.intel.cn/news-releases/the-joint-research-results-of-intel-and-4paradigm-were-selected-into-the-vldb-international-conference/#gs.yai3k2