

IEEE NVMSA '21

Designing a persistent-memory-native storage engine for SQL database systems

August 19, 2021

Shohei Matsuura
Yahoo Japan Corporation



I Outline

1. Background
2. Our Work
3. Evaluation
4. Conclusion & Future Work

History & Figures about Yahoo Japan Corporation

- Yahoo Japan Corp. is one of the major internet companies in Japan since 1996.
- It has **52 million user logins** monthly and offers **100+ internet services** ranging from e-commerce and online new media.

Transforming People's Daily Lives with Cutting-Edge Information Technologies

Since its founding in 1996 at the dawn of the Internet, the Yahoo Japan Group has been the industry pioneer and the key driving force behind the proliferation of the Internet in Japan. Through unceasing provision of innovative services, the Group has made significant impacts on the quality of people's daily lives.

OUR HISTORY

1996

Exploring Information Via the Internet



Launched Yahoo JAPAN

Launched Yahoo JAPAN as the first commercial Japanese-language search engine, providing a database of Japanese-language websites and a directory-like site search feature as well as keyword search services.

1996

Accessing the Latest Information on the Internet



Launched Yahoo News, Yahoo Weather Information (Now Yahoo Weather)

Started the Group's first information service in collaboration with media companies and other content partners. Users were provided the access to the latest news and daily weather information via Yahoo JAPAN.

1999

Easy, Convenient and Rapid Online Shopping Experiences



Launched Yahoo Shopping

Started Yahoo Shopping as a virtual shopping mall featuring stores of prominent Japanese retailers providing easy, convenient, and rapid online shopping experiences.

1999

Opportunity for Everyone to Engage in Online Transactions



Launched Yahoo Auctions (Now YAHU!UKU) Service

Began offering Yahoo Auctions (currently YAHU!UKU), an on-line auction service. Provided an opportunity for everyone, including both individuals and companies, to make a listing or bid on various items.

2001

Onset for Internet Use Proliferation in Japan



Launched Yahoo BB Service

Launched Yahoo BB commercial high-speed Internet connection service, from ¥2,280 per month. Fueled the proliferation of Internet connectivity throughout Japan, changing ways of Internet use.

2008

Optimization of User Interface for iPhone



Optimized Yahoo JAPAN Service for Newly Launched iPhone 3G

Also optimized major web services of the iPhone 3G model in Japan, optimized all Yahoo JAPAN services for easy iPhone 3G access. Optimized UI of 65 service top pages and created icons for iPhone 3G start-up screen.

2013

E-Commerce Made More Accessible through Elimination of Fees



Announced New E-Commerce Strategy

In our "Yahoo! Shopping" service, completely eliminated store, tenant fees and system use fees and made the transition from a fee-based revenue model to one based on advertising. Similarly, decided to eliminate store tenant fees in YAHU!UKU.

2018

Promotion of Convenient and Stress-Free Cashless Society



Launched PayPay Service

Entered the mobile payment business and jointly established PayPay Corporation with SoftBank Corp. Rapidly increased users and merchants through large-scale campaigns, etc. PayPay grew to become a central payment to connect of business owners in Japan.

2018

New Solutions Created by Unlocking the Potential of Data

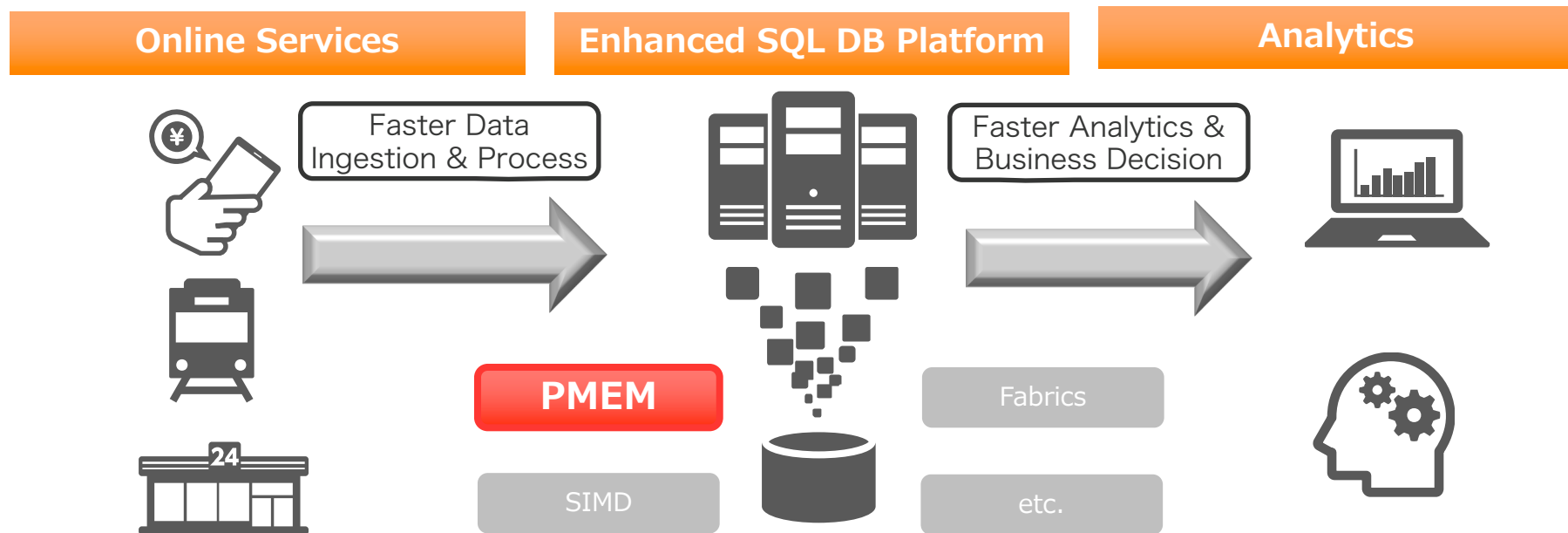


Announced DATA FOREST Initiative

By combining the data possessed by corporations, local governments, and research institutions with our big data and AI technologies, the DATA FOREST Initiative aims to create new businesses that support the problem-solving efforts of our clients. Full-scale proof-of-concept trials began for this purpose.

Research Motivation

- To Extend & enhance the capabilities of Yahoo! Japan's SQL database platform with the latest hardware technologies (i.e. Persistent Memory, or PMEM)



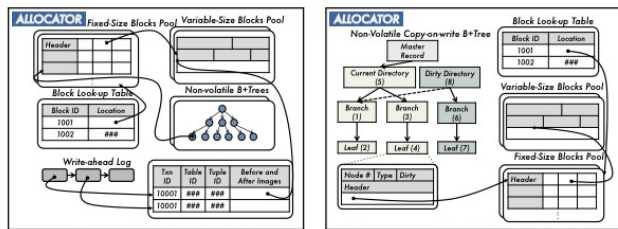
Existing Research & Application in Academia

Design Ideas for a PMEM-native Storage Engine

■ Arulraj (2018), Arulraj & Pavlo (2019)

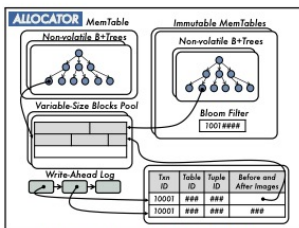
- Present 3 design ideas for the storage engine: (1) NVM-InP, (2) NVM-CoW, (3) NVM-Log
- Emulator-based performance study of the 3 designs with YCSB workloads

Storage Engine Design Ideas



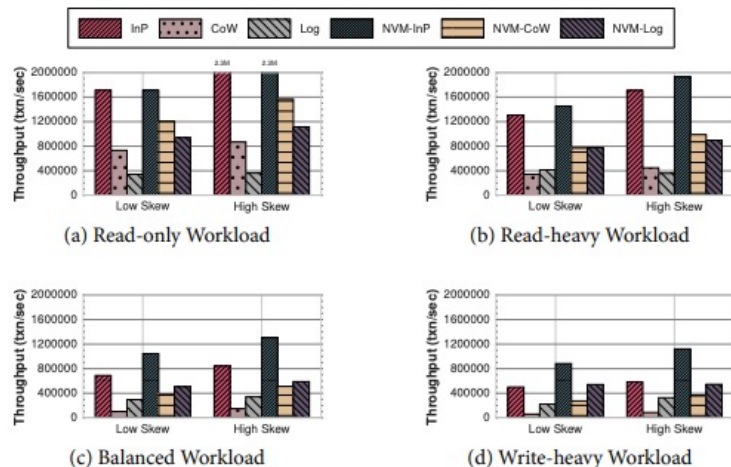
(a) In-place Updates (NVM-InP)

(b) Copy-on-Write Updates (NVM-CoW)



(c) Log-structured Updates (NVM-Log)

Emulator-based Performance Study



(a) Read-only Workload

(b) Read-heavy Workload

(c) Balanced Workload

(d) Write-heavy Workload

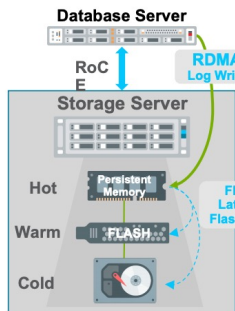
Existing Research & Application in Industry

Transaction Logging & Database Buffer Extension

■ Oracle Exadata*

- Accelerate Transaction Logging with PMEM

Exadata X8M Persistent Memory Commit Accelerator



- Log Write latency is critical for OLTP performance
 - Faster log writes means faster commit times
 - Any log write slowdown stalls the whole database
- Automatic Commit Accelerator
 - Database issues one-way RDMA writes to PMEM on multiple Storage Servers
 - Bypasses network and I/O software, interrupts, context switches, etc.
 - Up to 8x faster log writes

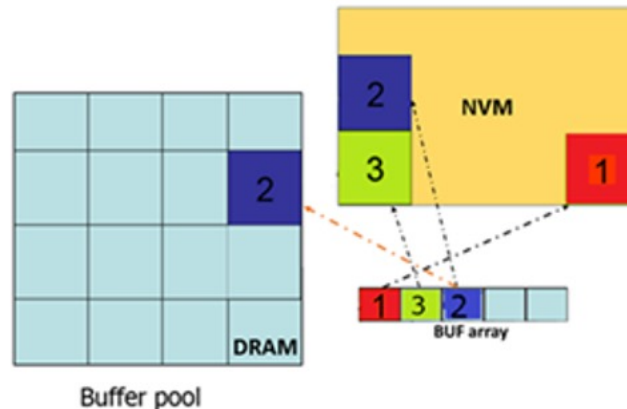
Enabled with Exadata System Software 19.3 and Database Software 19c

27

■ Microsoft SQL Server**

- Extend buffer pool & evicted pages direct read from PMEM (NVM)

Buffer pool with Hybrid Buffer Pool



* J.Shi, "Exadata with Persistent Memory: An Epic Journey." SNIA Persistent Memory Summit 2020.
https://www.snia.org/sites/default/files/PM-Summit/2020/presentations/11_PMEM_Jia_Shi_final_PM_Summit_2020_v2.pdf

** Microsoft Corporation, "SQL Server Hybrid Buffer Pool."
<https://docs.microsoft.com/en-us/sql/database-engine/configure-windows/hybrid-buffer-pool?view=sql-server-ver15>

Research Goals

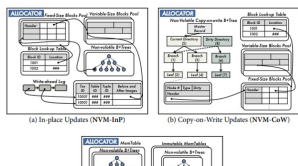
To state requirements for a practical storage engine that natively uses persistent memory for SQL database systems, and to illustrate how to design such a storage engine

Existing Research & Application in Academia

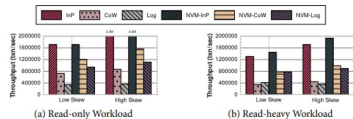
Design Ideas for a PMEM-native Storage Engine

- Arulraj(2018), Arulraj & Pavlo (2019)
 - Present3 design ideas for the storage engine: (1) NVM-InP, (2) NVM-CoW, (3) NVM-Log
 - Emulator-based performance study of the 3 designs with YCSB workloads

Storage Engine Design Ideas



Emulator-based Performance Study



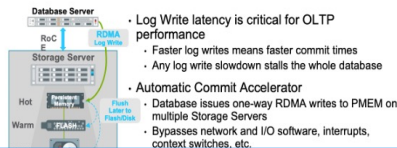
Making them practical is an open question!

Existing Research & Application in Industry

Transaction Logging & Database Buffer Extension

- Oracle Exadata
 - Accelerate Transaction Logging with PMEM
- Microsoft SQL Server
 - Extend buffer pool & direct page read from PMEM

Exadata X8M Persistent Memory Commit Accelerator



Buffer pool with Hybrid Buffer Pool



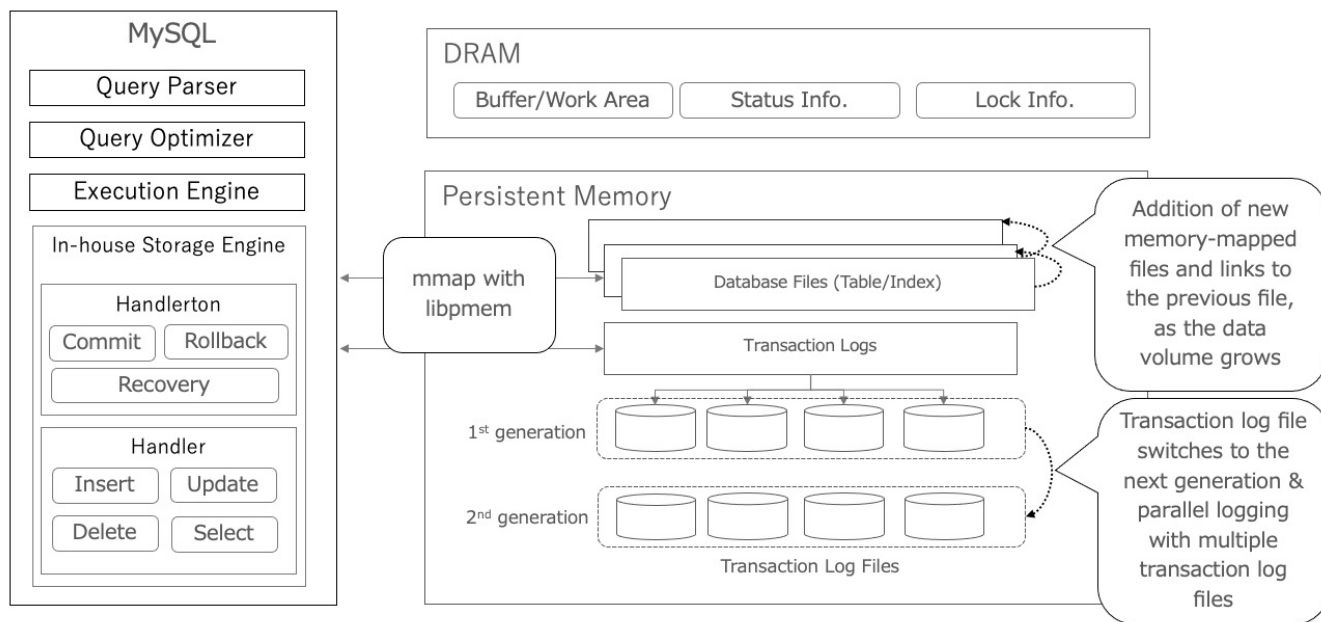
Partial use of persistent memory for database operations

I Requirements for a Practical PMEM-native Storage Engine

- Based on the current requirements for SQL database systems at Yahoo! Japan, we impose the same and the following requirements for the storage engine:
 1. Scale with Data
 2. Transaction Support
 3. Continuous Operation
 4. Performance
 5. MySQL Compatibility

Design for a Practical PMEM-native Storage Engine

- Place database & transaction log files on PMEM
- Storage expansion with linked database files, and transaction log switches for continuous operation
- WAL & Aries-based transaction support



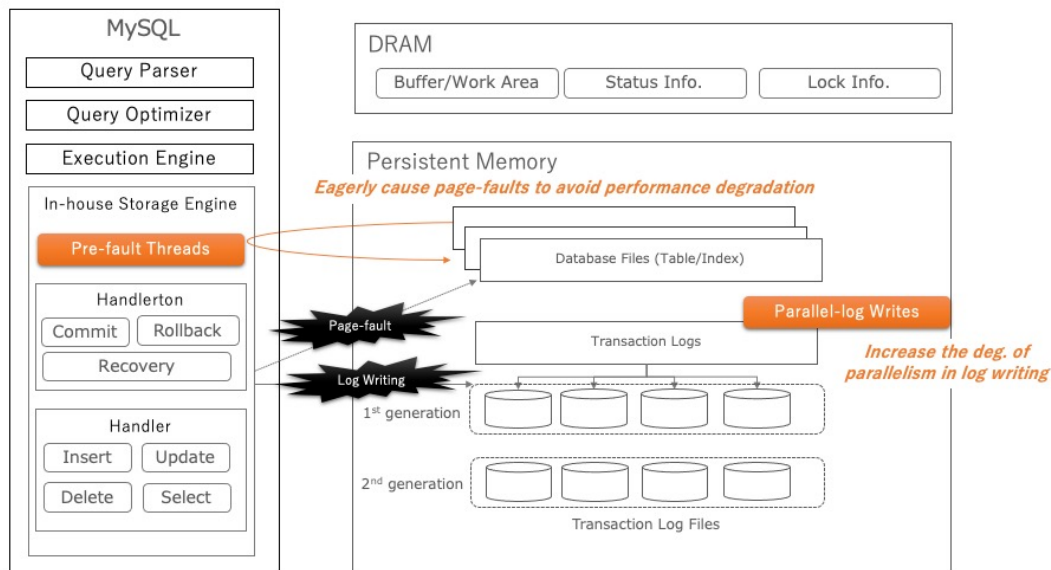
Two Important Design Features for Performance in the Storage Engine

1. Pre-fault

- Page-fault causes significant performance degradation when accessing mmap files [Choi & Kim, 2017]
- To avoid it during query processing, implement designated threads (pre-fault threads) to cause pre-fault before the storage engine main threads access the mmap files (database files & transaction log files)

2. Parallel-logging

- the state-of-the-art “parallel write-ahead logging algorithm” to increase the deg. of transaction log writing [Tanabe et al., 2018]



Evaluation: Environment & Workload

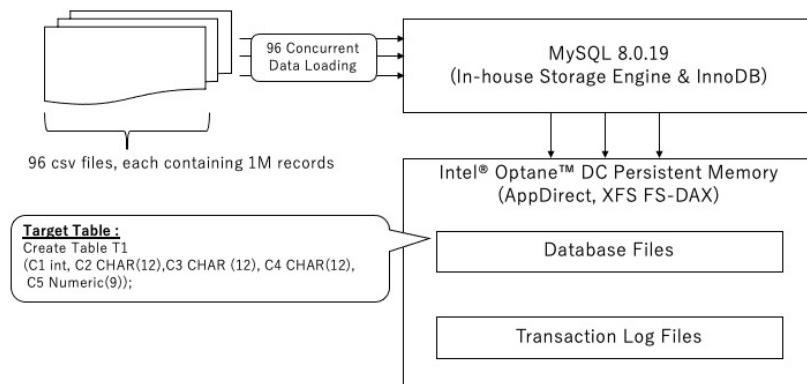
Evaluation Environment

- 2-Socket Server with 104-core
- Equipped with Intel® Optane™ DCPMM

CPU	Intel Xeon Gold 6230R 2.1GHz x 2 (Total 104 Cores)
DRAM	DDR4-192GB
Persistent Memory	Intel® Optane™ DC Persistent Memory
SSDs	SATA SSD 1.92TB (OS Boot, Load Data)
OS	CentOS 7.8
DBMS	MySQL 8.0.19 with the In-house Storage Engine

Evaluation Workload

- 96 concurrent data loading
- Good workload to observe the effects of the pre-fault & the parallel logging as it always accesses a new region of a mmap file and generates transaction logs



Evaluation: Effects of Pre-fault and Parallel-logging Features

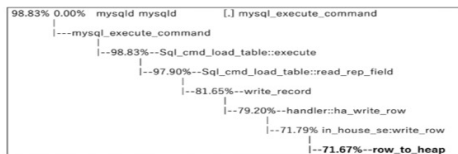
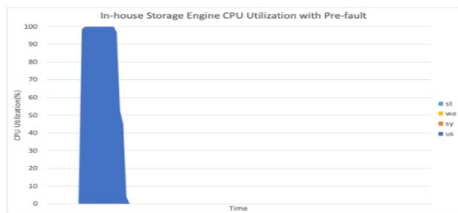
Pre-fault Feature

- More than 5x performance improvement
- Significant improvement in CPU utilization

Normalized Data-loading Time

Storage Engine	In-house Storage Engine with Pre-fault Feature	In-house Storage Engine without Pre-fault Feature
Loading Time	1	5.86

CPU Utilization & Perf Output

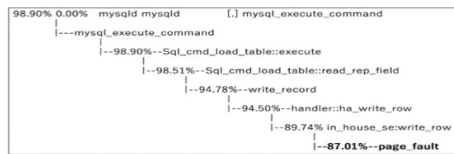
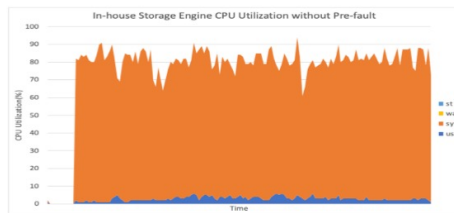


Parallel-logging Feature

- 30%+ performance improvement
- Increasing the deg. of parallel log write too much also causes performance degradation

Normalized Data-loading Time

Number of Parallel Log Write	1	2	4	8	10	12	16
Loading Time	1.31	1.06	1.00	1.17	1.18	1.05	1.28



Evaluation: Overall Performance Improvement by the Storage Engine

Overall Performance Comparison

- More than 50x performance improvement with our in-house storage engine than InnoDB running on PMEM
- In-house storage engine run with the pre-fault feature enabled and the parallel log write=4

Normalized Data-loading Time

Storage Engine	In-house Storage Engine with Pre-fault & Parallel-logging Features	InnoDB on Persistent Memory
Loading Time	1	58.29

Conclusion & Future Work

■ Conclusion:

- ✓ Presented & discussed five requirements for a practical PMEM-native storage engine that satisfies industry requirements
- ✓ Two important design features, (1) pre-fault and (2) parallel-logging, to make a storage engine performant on PMEM
- ✓ Overall, our designed in-house storage engine achieves 50x+ performance in write-workload on PMEM compared to InnoDB on PMEM

■ Future Work:

- ✓ Data Tiering to handle more data than PMEM capacity
- ✓ High-Availability feature to ensure database operations can continue even in the case of a data center failure

Trademarks & Registered Trademarks

- Oracle and MySQL are registered trademarks of Oracle and/or its affiliates.
- Microsoft and SQL Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.
- Intel® and Intel® Optane™ are trademarks of Intel Corporation or its subsidiaries.

Thank you very much! 😊

Please send us your feedback and questions to:

Contact: Shohei Matsuura

Email: shmatsuu@yahoo-corp.jp

Yahoo Japan Corporation

YAHOO!
JAPAN