面向持久性内存的数据库发展与研究现状

华东师范大学 数据学院 胡卉芪

hqhu@dase.ecnu.edu.cn

数据库系统关注热点

其他数据管理方法

分布式数据库

新硬件

DB+AI

云数据库 DB-as-Service

NoSQL 多模态

性能

更低的成本实现更低的延迟、更高的吞吐

易用性

数据库低运维，简易实用

数据库系统的需求

DB	RDBMS	OLTP	OLAP	NoSQL	Cloud&NewSQL	? ?
1960s	1970s	1980s	1990s	2000s	2010s	2020s
• 64: concept • 65: network • 68:hierarchical	• 70: relation • 74: system R • 74: ingres • 76: ER • 79: oracle	• 83: DB2 • 85: OO • 86: PG • 88: SQLServer	• 93:OLAP • 95:MySQL	• 00:SQLite • 04:CStore • 07:Neo4j • 08: Cassandra • 09:Mango • 09: Redis	• 10:Hive • 13:F1 • 14: Spark SQL • 14: Aurora • 14: snowflake • 15: cosmos	

新型硬件对数据库系统的发展驱动

硬件红利，低延迟、高带宽

内存持久化能力
微妙级低延迟网络

当下：现有数据库的适配与优化

未来？新型硬件对数据库的架构性颠覆

数据存储

减少写放大，减少IO栈，降低访问延迟

内存计算

扩大缓存容量，减少内存成本，大内存计算

持久化与高可用

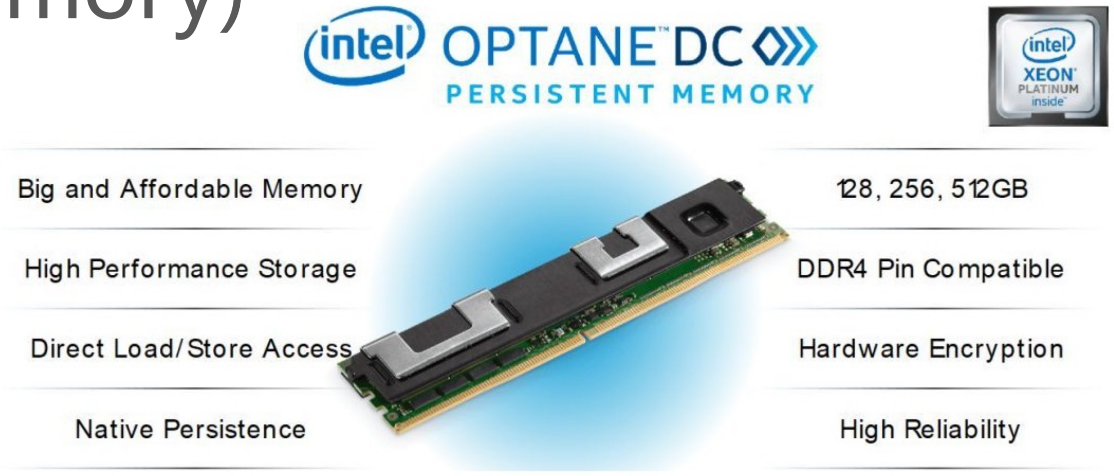
结合RDMA，低延迟的日志持久化与同步

新型硬件与设施 RDMA、持久性内存、GPU、云存储等对数据库的变革

新型硬件-持久化内存

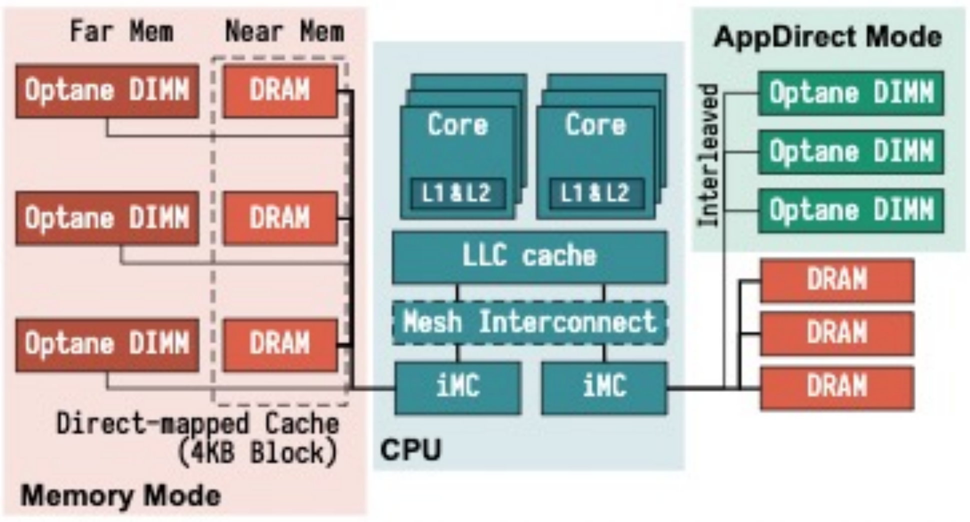
持久化内存(Persistent Memory)

- 字节操作
- 持久化能力
- 更高存储密度

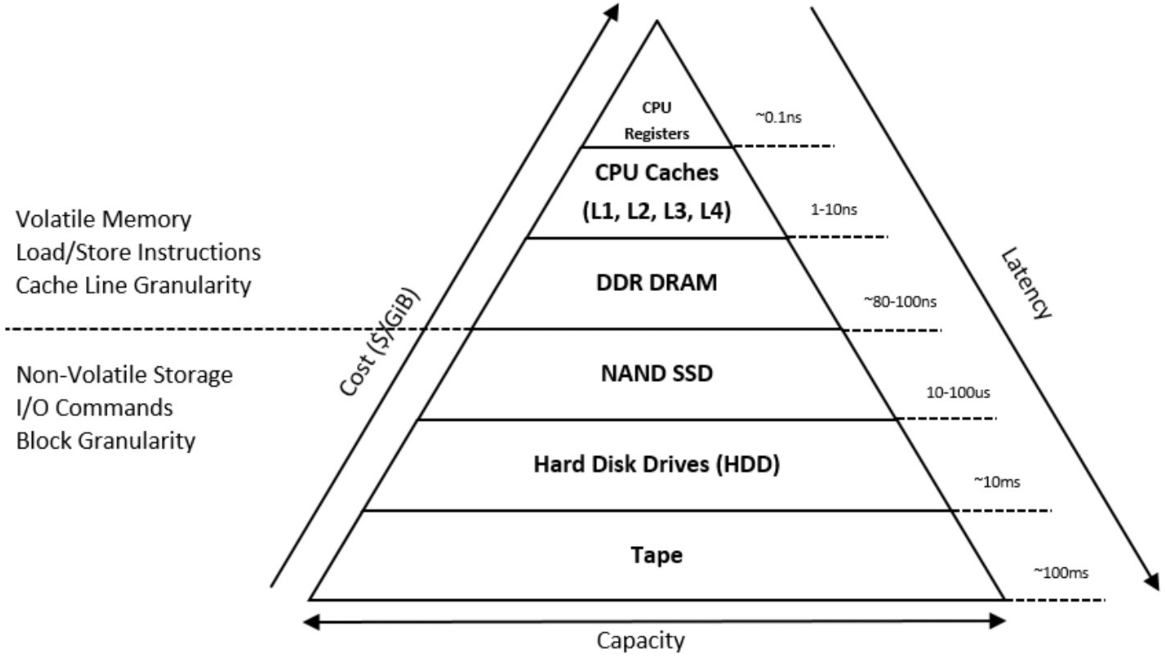


新型硬件-持久化内存

- 更高的性能（相较于SSD）
- 单位存储更低的价格（相较于DRAM）



持久性内存的两种模式[1]

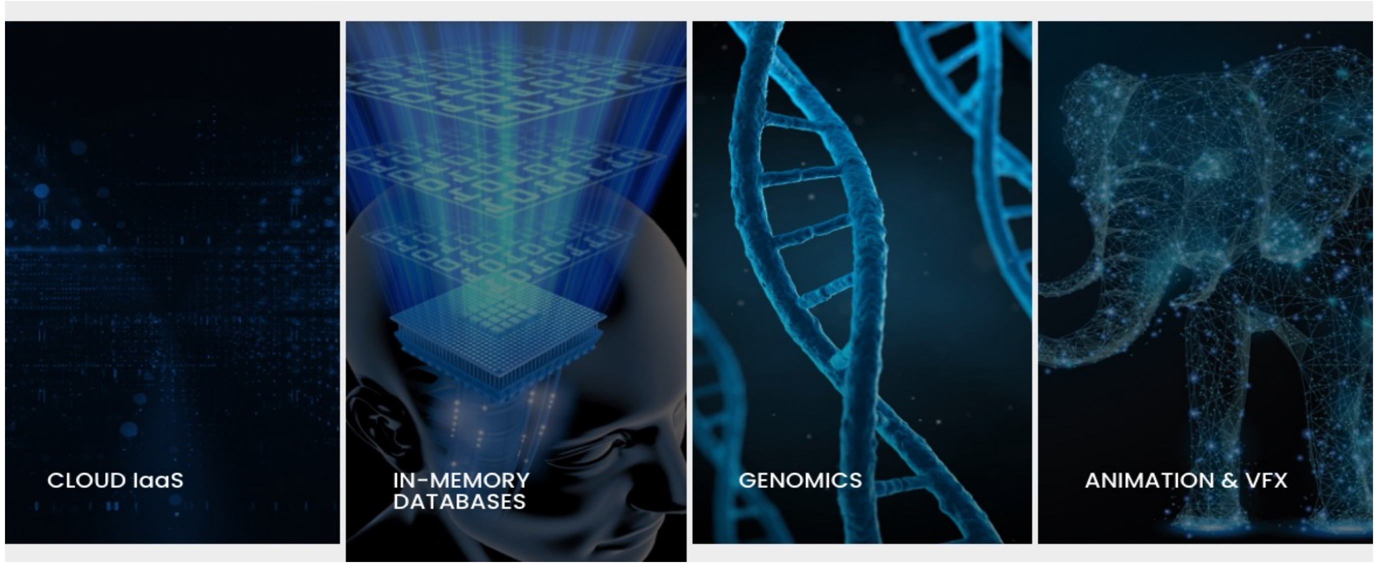


1. An Empirical Guide to the Behavior and Use of Scalable Persistent Memory. FAST 20

研究&Startups

- 存储
  - 文件系统
  - 键值存储

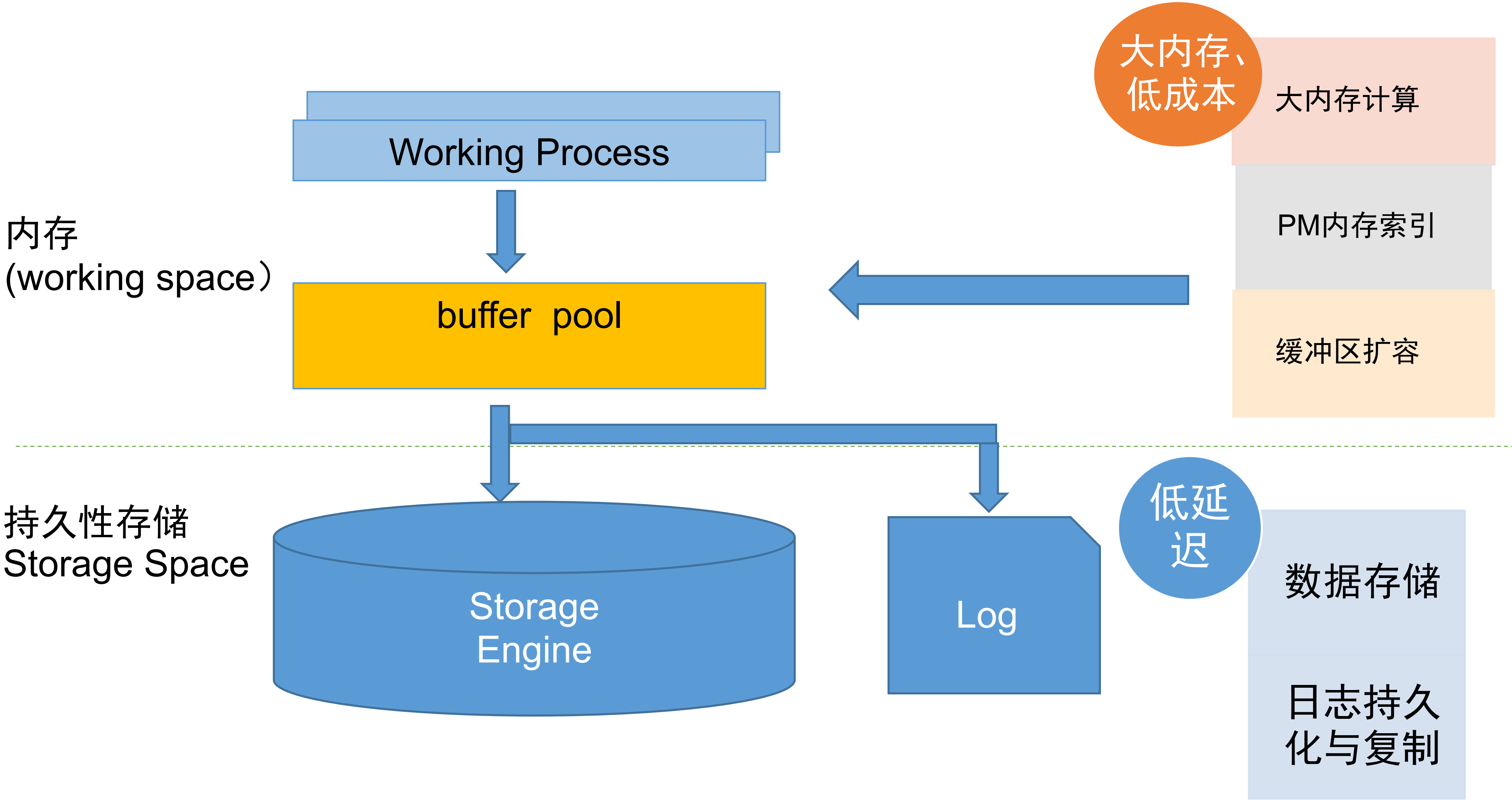
MemVerge



MemVerge & Big Memory

https://memverge.com/

PM在关系数据库中的应用范畴



研究范畴

ORACLE DATABASE

Microsoft SQL Server

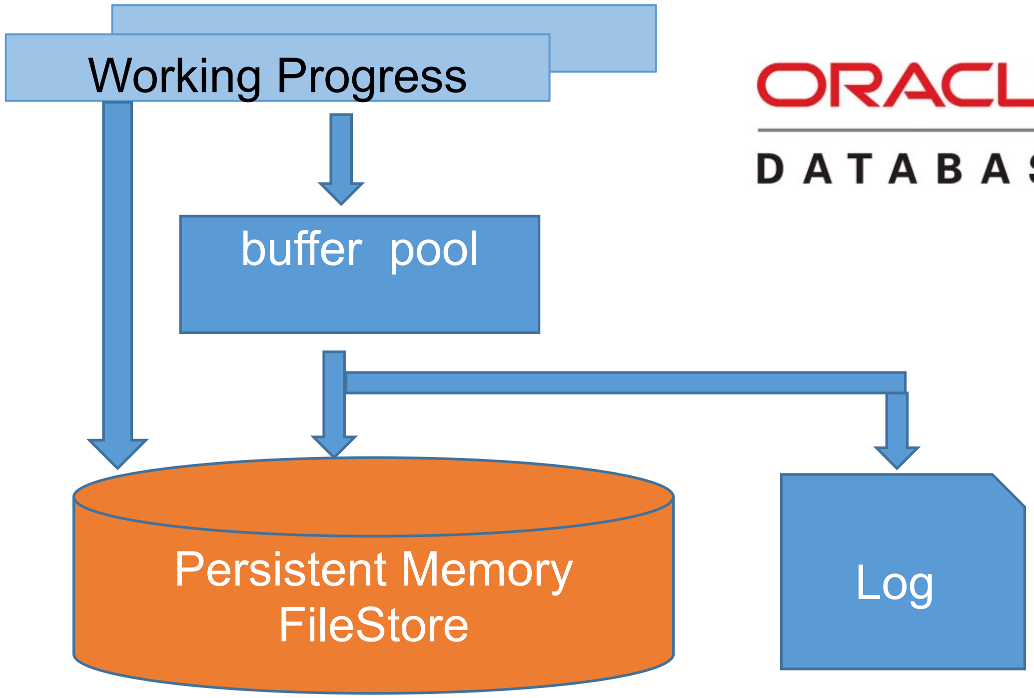
SAP HANA

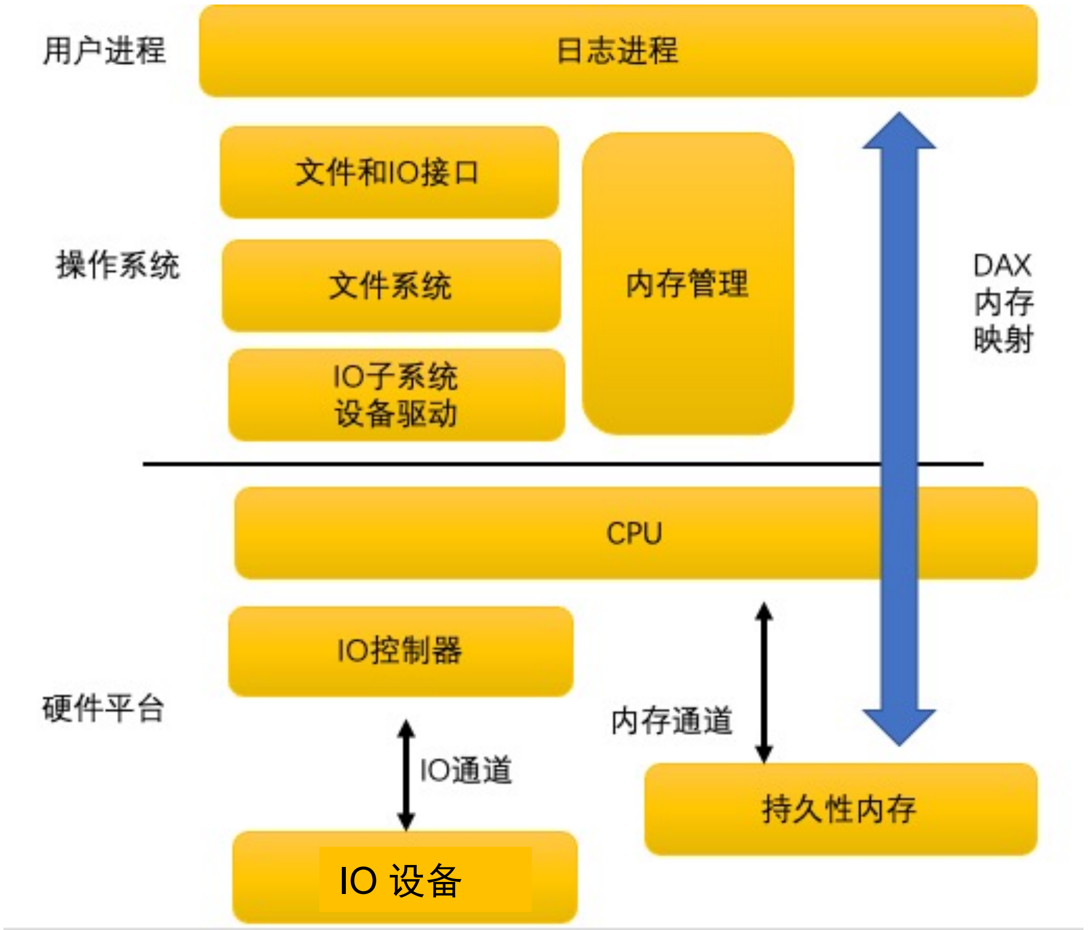
商用数据库中的应用

Storage Space

数据存储（一）

- ORACLE. 21c新特性： Persistent Memory Filestore
  1. 通过内存拷贝的方式持久化数据，比I/O操作更加高效
  2. 读操作不必一定经过缓存，提高了缓存利用率



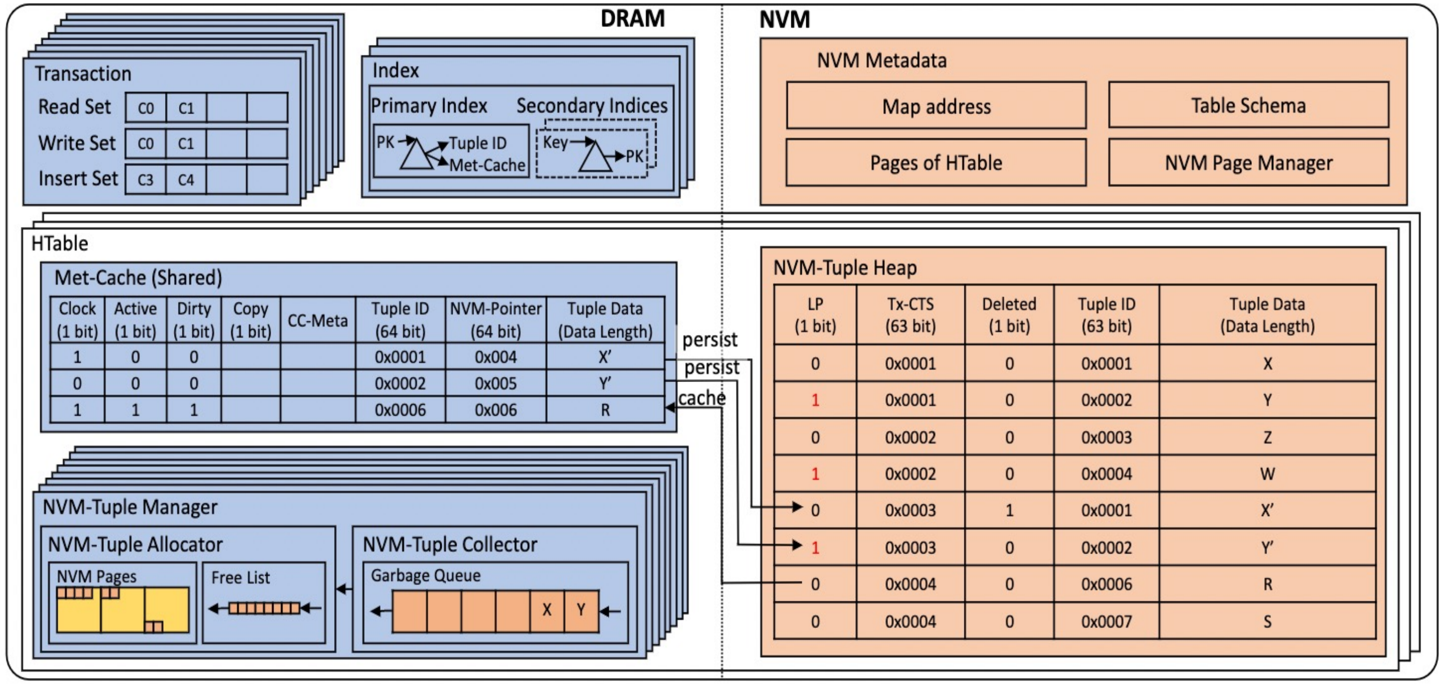


持久化内存操作比I/O操作更高效

https://docs.oracle.com/en/database/oracle/oracle-database/21/nfcon/persistent-memory-database-258797846.html

数据存储（二）-轻日志/无日志存储引擎

数据持久化在PM上，降低日志代价，轻日志甚至无日志的存储引擎



- 无日志

数据先持久化，事务再提交

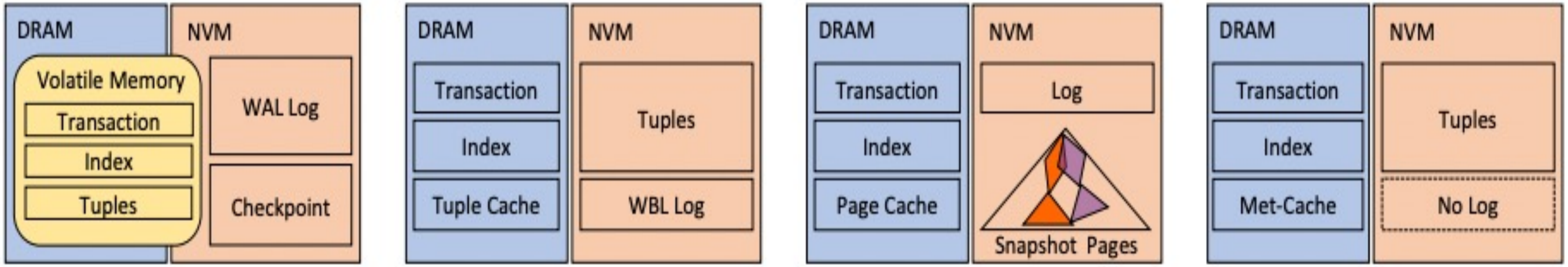
- 数据行粒度

Gang Liu, Leying Chen, Shimin Chen: Zen: a High-Throughput Log-Free OLTP Engine for Non-Volatile Main Memory. Proc. VLDB Endow. 14(5): 835-848 (2021)

Joy Arulraj, Matthew Perron, Andrew Pavlo: Write-Behind Logging. Proc. VLDB Endow. 10(4): 337-348 (2016)

Hideaki Kimura: FOEDUS: OLTP Engine for a Thousand Cores and NVRAM. SIGMOD Conference 2015: 691-706

数据存储（二）-轻日志/无日志存储引擎



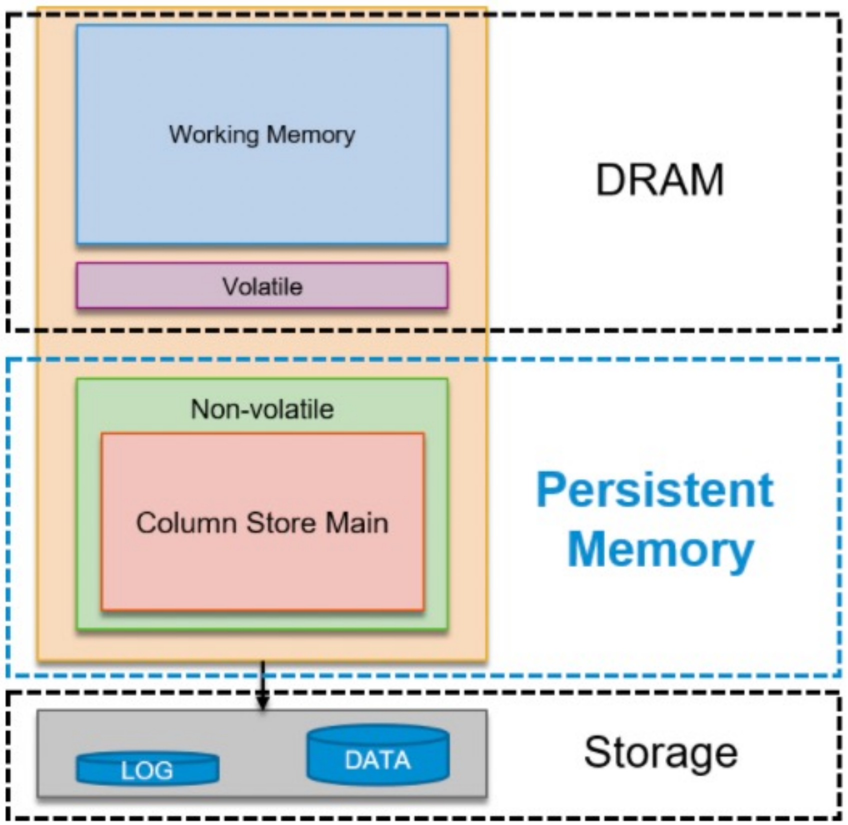
(a) MMDB with NVM Capacity (b) WBL (c) FOEDUS (d) Zen (our proposal)

OLTP engine designs for NVM	日志	减少NVM写入次数	事务	粒度	读写放大
MMDB with NVM Capacity	✓ (Undo log/Redo log & checkpoint)	×	DRAM NVM	tuple	✓
WBL	✓ (时间戳C_p, C_d)	✓	DRAM NVM	tuple	×
FOEDUS	✓ (Redo log)	✓	DRAM	page	✓
Zen	× (缓存中的元数据替代)	✓	DRAM	tuple	×

数据存储（三）

数据分布与成本一致

- DRAM+PM+SSD+磁盘的混合存储
- 热/温/冷数据
  - 检测方法
  - 放置方法



https://blogs.sap.com/2018/12/03/sap-hana-persistent-memory/

数据存储（四）-LSM-tree存储

- 很多新的数据库采用LSM-tree架构作为其存储（TiDB，Oceanbase等、阿里数据库）

PM在LSM-tree存储中的可能性

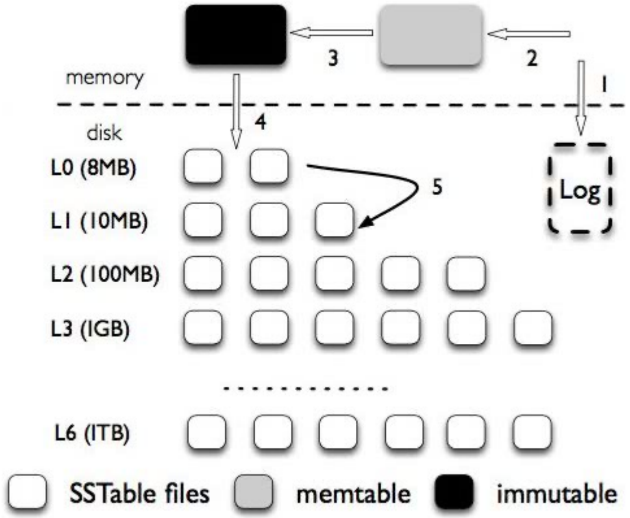
(1) Logging

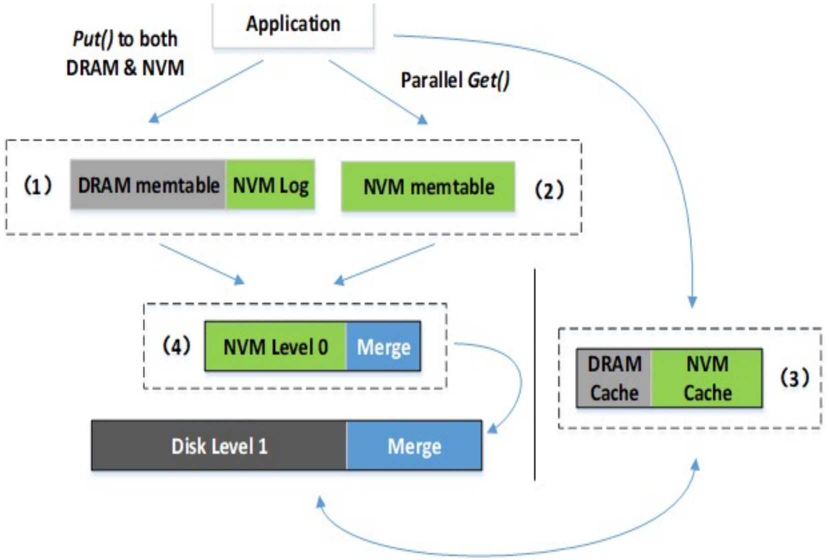
(2) Memtable

(3) Cache （扩大缓存容量）

(4) Level 0 （快速compaction）

(5) Level 1-N

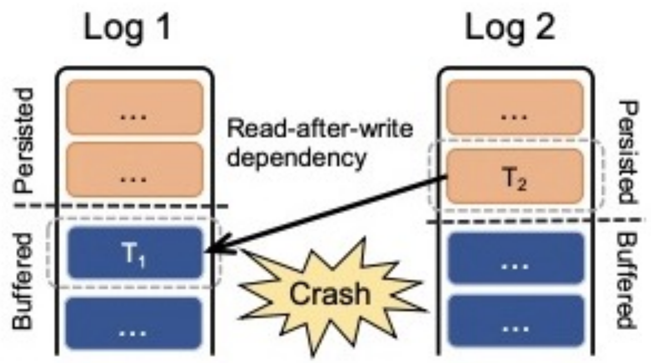


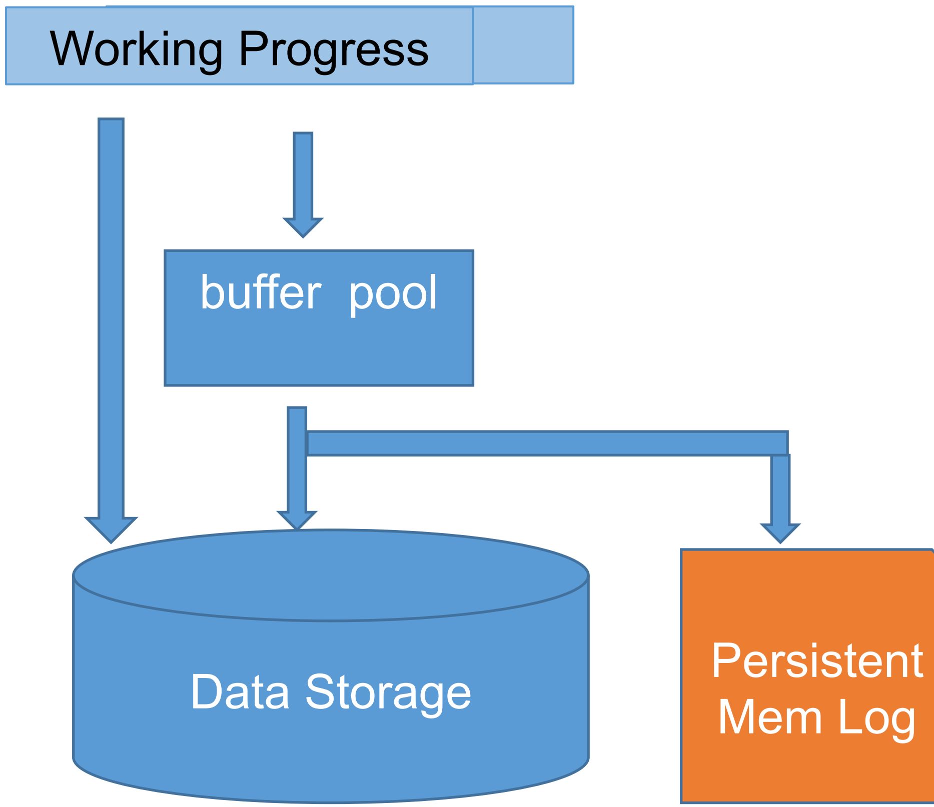


- Baoyue Yan, Xuntao Cheng, Bo Jiang, Shibin Chen, Canfang Shang, Jianying Wang, Kenry Huang, Xinjun Yang, Wei Cao, Feifei Li: Revisiting the Design of LSM-tree Based OLTP Storage Engine with Persistent Memory. Proc. VLDB Endow. 14(10): 1872-1885 (2021)
- Ting Yao, Yiwen Zhang, Jiguang Wan, Qiu Cui, Liu Tang, Hong Jiang, Changsheng Xie, Xubin He: MatrixKV: Reducing Write Stalls and Write Amplification in LSM-tree Based KV Stores with Matrix Container in NVM. USENIX Annual Technical Conference 2020: 17-31
- Sudarsun Kannan, Nitish Bhat, Ada Gavrilovska, Andrea C. Arpaci-Dusseau, Remzi H. Arpaci-Dusseau: Redesigning LSMs for Nonvolatile Memory with NoveLSM. USENIX Annual Technical Conference 2018: 993-1005
- Revisiting the Design of LSM-tree Based OLTP Storage Engine with Persistent Memory. VLDB 21

日志持久化与复制（一）

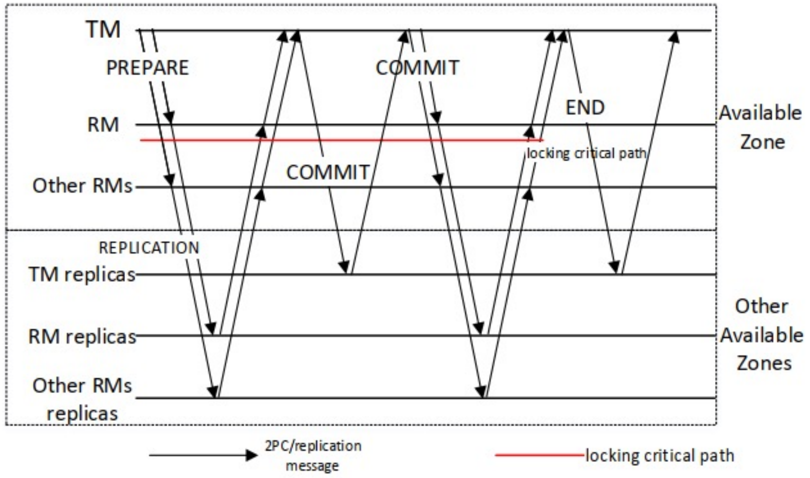
- 使用PM作为日志存储
  - 效果？？
  - 内存数据库（Tps 1M/s），产生大量日志
    - 并行日志：多设备提交日志

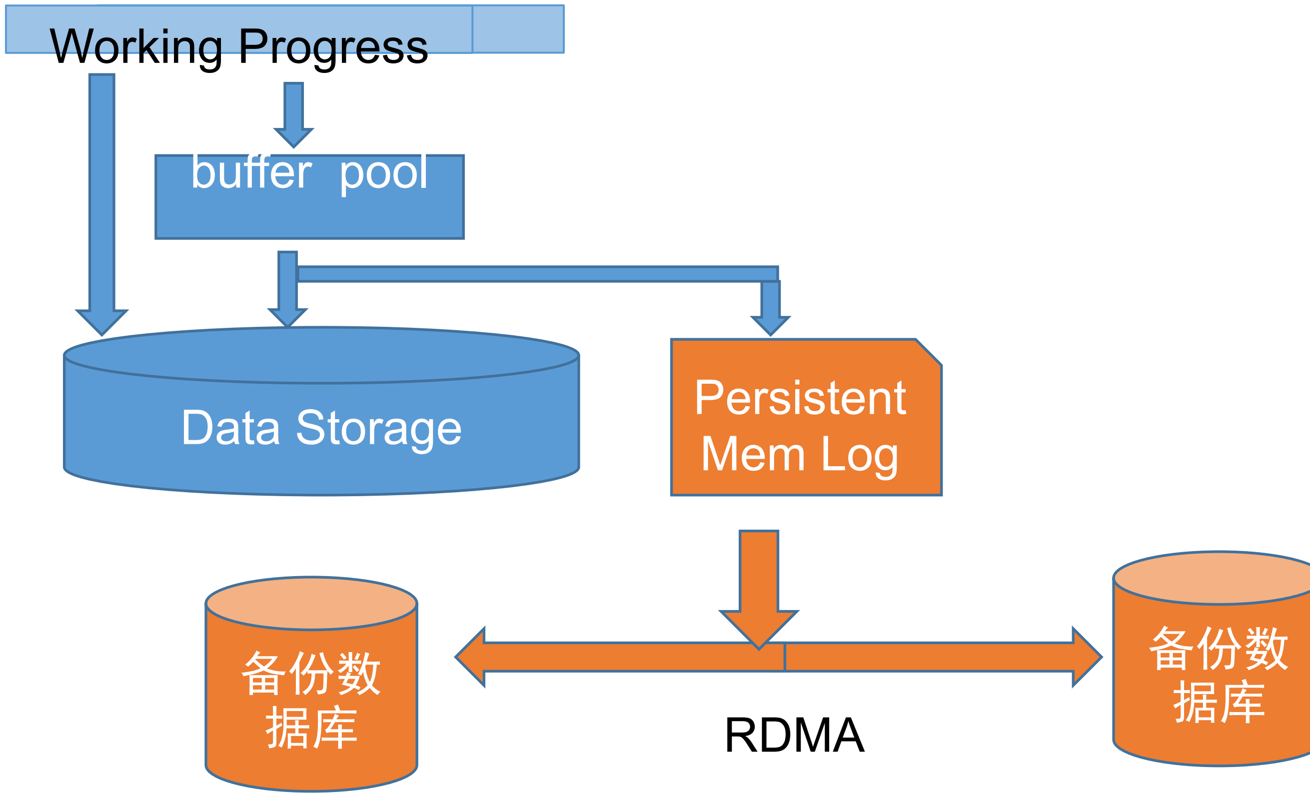




- Scalable Logging through Emerging Non-Volatile Memory. VLDB 2014
- Taurus：Lightweight Parallel Logging for In-Memory Database Management Systems
- Plover: parallel logging for replication systems. Frontiers Comput. Sci. 14(4) (2020)
-

日志持久化与复制（二）



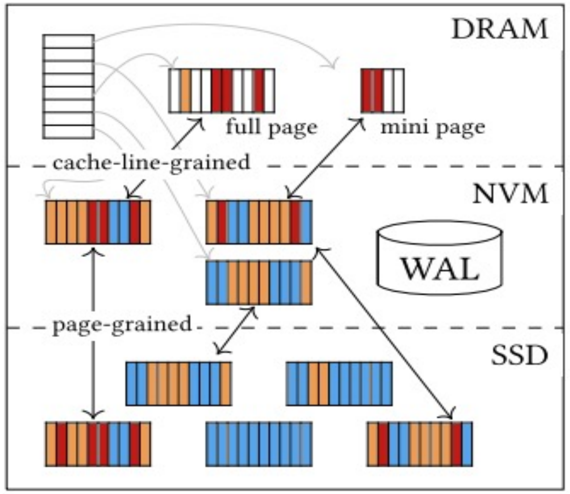


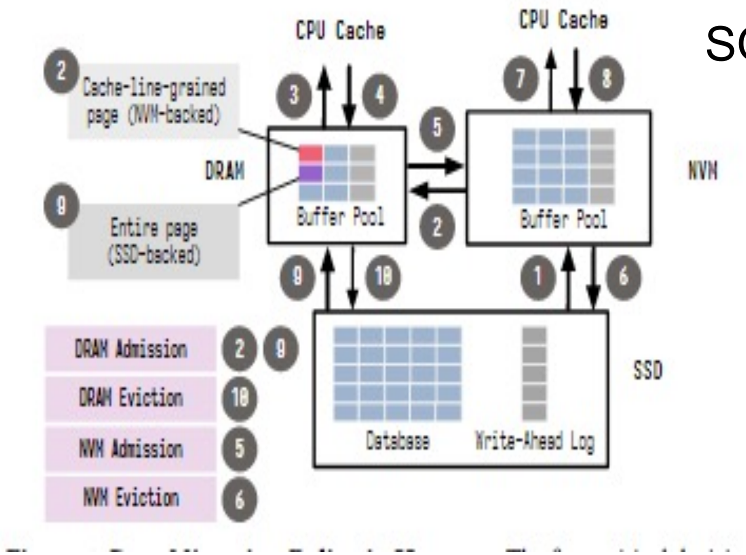
- 分布式数据库中数据库事务提交需要多轮通讯
- RDMA+PM
  - 应用于备机日志复制，减少事务提交延迟
  - RDMA操作在远程PM持久化的问题
  - 如何解决异地备份的问题?

- Query Fresh: Log Shipping on Steroids. VLDB 2017
- Xingda Wei, Xiating Xie, Rong Chen, Haibo Chen, Binyu Zang:
  Characterizing and Optimizing Remote Persistent Memory with RDMA and NVM.
- USENIX Annual Technical Conference 2021: 523-536
- Anuj Kalia, David G. Andersen, Michael Kaminsky:
  Challenges and solutions for fast remote persistent memory access. SoCC 2020: 105-119

Working Space

数据库缓冲区





SQL-Server. **Hybrid Buffer Pool**

1、将DRAM升级成更小粒度的缓存(64B)

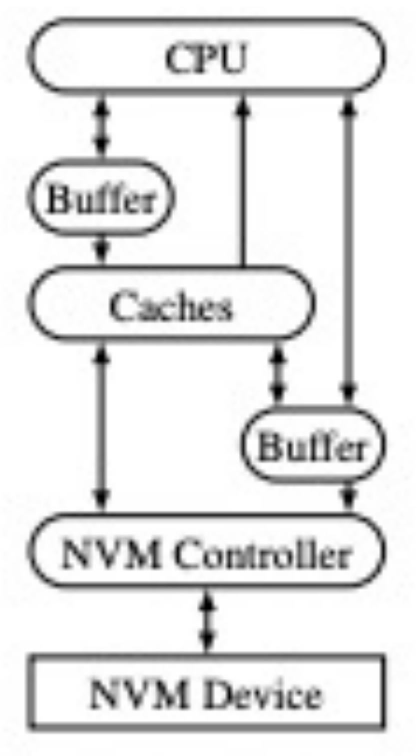
2、DRAM/PM 共存，灵活的替换策略

1. Managing Non-Volatile Memory in Database Systems. SIGMOD 2018

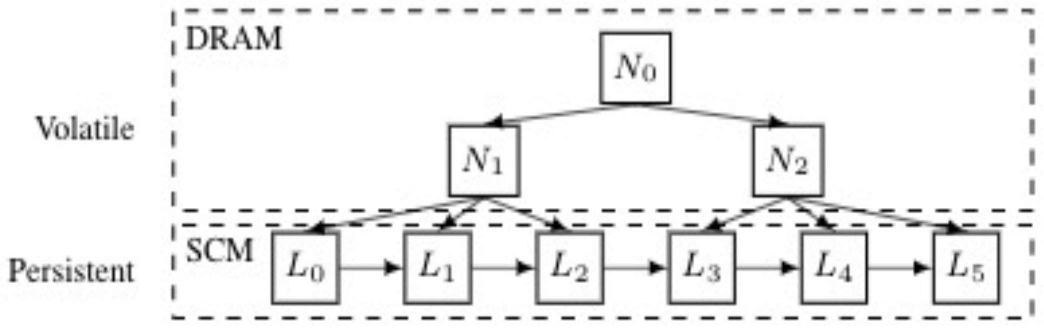
2. Spitfire: A Three-Tier Buffer Manager for Volatile and Non-Volatile Memory. SIGMOD 2021

PM内存索引

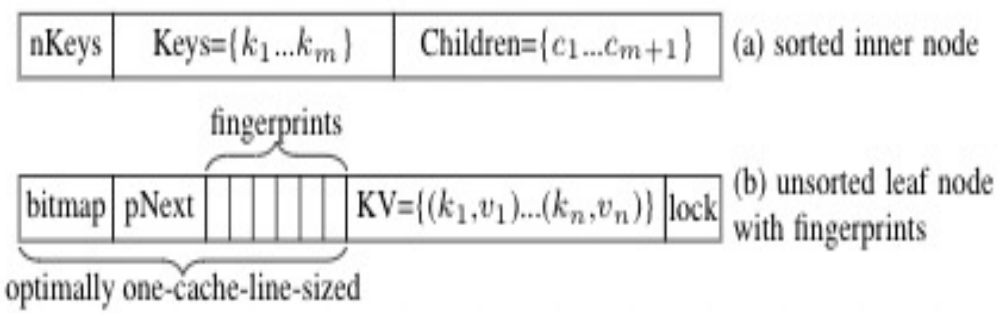
FPTree



宕机一致性



树形结构(数据和叶子在PM中)



叶子结点结构

应用场景

- 内存数据库
  - e.g. 华为openGauss内存引擎

索引结构

- 哈希表、B+树、ART

关键技术

- 宕机一致性
- 减少持久化内存的写入
- 减少读写放大

PM内存索引

基于B+树的PM索引

持久性索引结构	宕机一致性保证	减少NVM写入次数	读写放大	锁开销放大
PCM-friendly B+-Tree	✓ (日志)	✓	×	×
CDDS-Tree	✓ (多版本)	✓	×	×
wB+-Tree	✓ (原子操作、日志)	✓	×	×
NV-Tree	× 内部节点 ✓ 叶子节点 (原子操作、日志)	✓	×	×
FPTree	✓ (原子操作、日志)	✓	×	×
FAST+FAIR	✓ (无日志)	✓	×	×
LB+-Tree	✓ (原子操作)	✓	✓	×
BP+-tree	✓ 无锁日志	✓	✓	✓

总结

- 利用PM作为构建DRAM/PM/SSD等多级存储和缓冲区处理空间的数据库已经出现
  - 作为数据库的一个组成部分
- 在研究方面，技术已经渐渐成熟，在商业上，以PM特征而设计的，颠覆性架构的数据库系统还在发展阶段

谢谢聆听