DapuStor

2020年代的SSD研究潮流

吕涛，大普微电子

2023/07/03

DapuStor

SSD新形态

SR-IOV SSD

- SR-IOV是PCIe规范的扩展，它允许一个PCIe物理设备在一个根端口下呈现为多个独立的物理设备，包括一个物理功能（PF）和多个虚拟功能（VF）
- SR-IOV SSD可以让虚拟机直接访问VF，绕过VM kernel层，从而降低延迟和提高CPU效率

ZNS SSD

- ZNS是一种允许数据根据使用和访问频率分组并顺序存储在SSD中独立区域内的技术
- 显著减少写入操作的次数，降低写放大因子(WAF)，从而延长SSD的寿命
- ZNS SSD可以消除过度配置的需要，充分利用SSD的容量

KV SSD

- 在SSD上直接存储和检索变长的key-value对
- 显著减少写入操作的次数，降低写放大因子(WAF)，从而延长SSD的寿命
- 消除翻译层和读写放大

CSD

- CS是一种将计算能力部署到存储设备上的技术，可以在数据源处执行数据处理，减少数据传输提高效率
- CSD可以分为两种类型：CSP和CSM，前者在SSD控制器上集成处理器或加速器，后者在闪存介质上集成内存或加速器

CXL

- CXL是一种计算表达链路（CXL）的接口规范
- CXL可以分为三种类型：CXL.io、CXL.cache和CXL.mem，分别用于I/O、缓存和内存操作

文章分类

1. Zoned storage and namespaces

- ZNS: Avoiding the Block Interface Tax for Flash-based SSDs (ATC'21, Western Digital)
- ZNS+ (OSDI'21, Samsung Electronics)
  - 将ZNS主机端垃圾回收的数据拷贝卸载到盘内
- ZNSwap: un-Block your Swap (ATC'22, Western Digital)
- RAIZN: Redundant Array of Independent Zoned Namespaces (ASPLOS'23, CMU & Western Digital)
  - RAIZN 是一个 Linux 设备映射器逻辑卷，它在阵列中配置 ZNS 设备并向主机公开单个 ZNS 设备，透明地提供冗余和条带化。RAIZN 的新颖之处在于公开了在不支持覆盖的物理 ZNS 设备上运行的逻辑 ZNS 卷。

2. Log-structured file systems and write amplification

- IPLFS: Log-Structured File System without Garbage Collection (ATC'22, KAIST)
- Separating Data via Block Invalidation Time Inference for Write Amplification Reduction in Log-Structured Storage (FAST'22, Alibaba)
- Pattern-Guided File Compression with User-Experience Enhancement for Log-Structured File System on Mobile Devices (FAST'21, Nanjing University of Science and Technology)
  - FPC 对写入密集型且高度可压缩的 SQLite 文件应用前台压缩，并对可执行文件应用后台压缩，以重新组织读取关键块，以实现快速应用程序启动和节省空间。
- Beating the I/O Bottleneck: A Case for Log-Structured Virtual Disks (EuroSys'22, NetApp Inc.)

文章分类

3. In-storage computing and processing

- FusionFS: Fusing I/O Operations using CISCOps in Firmware File Systems (FAST'22, Rutgers University)
- **Hardware/Software Co-Programmable Framework for Computational SSDs to Accelerate Deep Learning Service on Large-Scale Graphs (FAST'22, KAIST)**
- RM-SSD: In-Storage Computing for Large-Scale Recommendation Inference (HPCA'22, CUHK)
- INSPIRE: IN-Storage Private Information REtrieval via Protocol and Architecture Co-design (ISCA'22, UCSB)
- SmartSAGE: Training Large-scale Graph Neural Networks using In-Storage Processing Architectures (ISCA'22, KAIST)
- **λ-IO: A Unified IO Stack for Computational Storage (FAST'23, Tsinghua University)**

4. Hardware-software co-design and optimization

- FlashShare: Punching Through Server Storage Stack from Kernel to Firmware for Ultra-Low Latency SSDs (OSDI'18, Yonsei University)
  - 一种整体跨堆栈方法，使 ULL SSD 设备能够直接向用户提供低延迟优势并满足不同的服务级别要求
- IODA: A Host/Device Co-Design for Strong Predictability Contract on Modern Flash Storage (SOSP'21, CMU)
  - I/O 确定性闪存阵列，它引入了三种主要技术来增强 IOD 接口并促进确定性主机/SSD 协同设计，并结合降级 -无缝读取

文章分类

5. Multi-streamed SSDs

- LightNVM: The Linux Open-Channel SSD Subsystem (FAST'17, CNEX Labs)
- FStream: Managing Flash Streams in the File System (FAST'18, Samsung Electronics)
- Fully Automatic Stream Management for Multi-Streamed SSDs Using Program Contexts (FAST'19, Samsung Electronics)

6. Key-value stores

- PinK: High-speed In-storage Key-value Store with Bounded Tails (ATC'20, MIT)
  - 一种基于 LSM 树的键值引擎，它克服了基于哈希的 KV-SSD 和传统 LSM 树实现的问题。PinK 在读取延迟、吞吐量和写入放大方面优于现有的 KV-SSD。
    - 哈希问题是DRAM不够大的时候需要访问闪存，开销大，导致不一致的性能
    - 传统LSM树的问题是bloom filter导致的概率性尾部延时，以及写放大
- Vigil-KV: Hardware-Software Co-Design to Integrate Strong Latency Determinism into Log-Structured Merge Key-Value Stores (ATC'22, KAIST)
- p2KVS: a Portable 2-Dimensional Parallelizing Framework to Improve Scalability of Key-value Stores on SSDs (EuroSys'22, HUST)

文章分类

7. Flash-based Caches

- Flashield: a Hybrid Key-value Cache that Controls Flash Write Amplification (NSDI'19, Stanford)
- Kangaroo: Caching Billions of Tiny Objects on Flash (SOSP'21, CMU)
  - 为了避免大型 DRAM 索引，Kangaroo 将大部分缓存容量组织为集合关联缓存，称为 KSet。
  - 为了减少闪存写入，Kangaroo 在 KSet 前面放置了一个小型日志结构缓存，称为 KLog。
- CacheSack: Admission Optimization for Google Datacenter Flash Caches (ATC'22, Google)

8. SSD failures and reliability

- Error Characterization, Mitigation, and Recovery in Flash Memory Based Solid-State Drives (PIEEE'17, Carnegie Mellon University)
- Improving 3D NAND Flash Memory Lifetime by Tolerating Early Retention Loss and Process Variation (Sigmetrics'18, Carnegie Mellon University)
- Design Tradeoffs for SSD Reliability (FAST'19, Seoul National University)
- NVMe SSD Failures in the Field: the Fail-Stop and the Fail-Slow (ATC'22, Shanghai Jiao Tong University)
- GuardedErase: Extending SSD Lifetimes by Protecting Weak Wordlines (FAST'22, Seoul National University)
- Efficient Bad Block Management with Cluster Similarity (HPCA'22, National Taiwan University)
- Multi-view Feature-based SSD Failure Prediction: What, When, and Why (FAST'23, Samsung Electronics)

文章分类

9. RAID and parity-based systems

- StRAID: Stripe-threaded Architecture for Parity-based RAIDs with Ultra-fast SSDs (ATC'22, HUST)
- Improving the Reliability of Next Generation SSDs using WOM-v Codes (FAST'22, Google)

10. Read-only file systems

- EROFS: A Compression-friendly Readonly File System for Resource-scarce Devices (ATC'19, Huawei)

11. eBPF and kernel storage functions

- XRP: In-Kernel Storage Functions with eBPF (OSDI'22, Columbia University) XRP: In-Kernel Storage Functions with eBPF (OSDI'22, Columbia University)
  - 通过 SPDK 等库的完全内核旁路允许应用程序直接访问底层设备，但此类库也迫使应用程序实现其文件系统，放弃隔离和安全性，并轮询 I/O 完成情况，这会在 I/O 时浪费 CPU 周期利用率低。
  - 使用 Linux eBPF 的高性能存储数据路径。XRP 在 NVMe 驱动程序的中断处理程序中使用钩子，从而绕过内核的块、文件系统和系统调用层。

文章分类

12. L2P Mapping

- Integrated Host-SSD Mapping Table Management for Improving User Experience of Smartphones (FAST'23, Seoul National University)
- LeaFTL: A Learning-Based Flash Translation Layer for Solid-State Drives (ASPLOS'23, UIUC)

13. SSD Emulator

- NVMeVirt: A Versatile Software-defined Virtual NVMe Device (FAST'23, Ajou University)

14. Tiered memory and CXL

- TPP: Transparent Page Placement for CXL-Enabled Tiered Memory (ASPLOS'23, University of Michigan)
  - 该论文建议使用 CXL 来缓解内存成为超大规模数据中心中一项重大基础设施开支的问题。
  - 讨论了 Linux 内核的内存管理机制不是为 CXL-Memory 等异构 CPU 连接的仅 DRAM 系统设计的问题，其中平均内存访问延迟因内存层而异。
  - 提出了一种操作系统级透明页面放置机制（TPP），可以有效地将页面放置在分层内存系统中，从而使相对热的页面保留在快速内存层中，而冷页面则移动到慢速内存层中。

文章分类

15. SSD Virtualization

- BM-Store: Hardware-Assisted Local Storage Architecture for Bare-Metal Clouds (HPCA'23)
  - 现有的相关工作包括LeapIO和FVM，提出了裸机云中本地存储虚拟化的硬件解决方案。但这些解决方案需要安装定制驱动程序，无法部署在裸机实例上。现有的本地存储解决方案无法解决管理和维护挑战。
  - BM-Store架构的关键思想是：
    1. 裸机实例大规模部署的透明架构，无需额外驱动即可使用标准NVMe驱动访问虚拟本地存储资源。
    2. 通过将虚拟化层卸载到硬件来实现高性能存储虚拟化，以最小的开销实现极致的性能。
    3. 带外管理机制和维护功能，用于管理和维护裸金属云中的本地存储，提高可用性。
- LightIOV: Storage Virtualization in Data Centers

从前沿研究看SSD发展潮流

DapuStor

Zoned storage and namespaces

- ZNS: Avoiding the Block Interface Tax for Flash-based SSDs (ATC'21, Western Digital)
- ZNS+ (OSDI'21, Samsung Electronics)
- ZNSwap: un-Block your Swap (ATC'22, Western Digital)
- RAIZN: Redundant Array of Independent Zoned Namespaces (ASPLOS'23, CMU & Western Digital)

Log-structured file systems and write amplification

- IPLFS: Log-Structured File System without Garbage Collection (ATC'22, KAIST)
- Separating Data via Block Invalidation Time Inference for Write Amplification Reduction in Log-Structured Storage (FAST'22, Alibaba)
- Pattern-Guided File Compression with User-Experience Enhancement for Log-Structured File System on Mobile Devices (FAST'21, Nanjing University of Science and Technology)
- Beating the I/O Bottleneck: A Case for Log-Structured Virtual Disks (EuroSys'22, NetApp Inc.)

In-storage computing and processing

- FusionFS: Fusing I/O Operations using CISCOps in Firmware File Systems (FAST'22, Rutgers University)
- Hardware/Software Co-Programmable Framework for Computational SSDs to Accelerate Deep Learning Service on Large-Scale Graphs (FAST'22, KAIST)
- RM-SSD: In-Storage Computing for Large-Scale Recommendation Inference (HPCA'22, CUHK)
- INSPIRE: IN-Storage Private Information REtrieval via Protocol and Architecture Co-design (ISCA'22, UCSB)
- λ-IO: A Unified IO Stack for Computational Storage (FAST'23, Tsinghua University)

Key-value stores

- PinK: High-speed In-storage Key-value Store with Bounded Tails (ATC'20, MIT)
- Vigil-KV: Hardware-Software Co-Design to Integrate Strong Latency Determinism into Log-Structured Merge Key-Value Stores (ATC'22, KAIST)
- p2KVS: a Portable 2-Dimensional Parallelizing Framework to Improve Scalability of Key-value Stores on SSDs (EuroSys'22, HUST)

Tiered memory and CXL

- TPP: Transparent Page Placement for CXL-Enabled Tiered Memory (ASPLOS'23, University of Michigan)

SSD Virtualization

- BM-Store: Hardware-Assisted Local Storage Architecture for Bare-Metal Clouds (HPCA'23)

DapuStor

THE END

Making Data Storage Smarter