云可信 创未来

Lustre 在私有云 AI 场景中的优化和实践

彭荣耀—中国电子云存储团队

2024/11/08

CEC 中国电子 | 中国电子云

CONTENT

05 中国电子云 | CECSTACK 一体化算力平台

CEC 中国电子 | 中国电子云

CECSTACK V5 一体化算力平台采用云原生架构，以自主研发的新型大规模分布式云操作系统 CCOS 为基础，为通用计算、智能计算和高性能计算等应用场景提供大规模、高可靠和可扩展的计算、存储、网络服务以及安全和灾备能力，具备完善的运营、运维、安全防护等云服务能力。

云管理

运营			
组织管理	定价管理	计量计费	日志管理
服务管理	资源管理	菜单管理	认证鉴权

运维			
作业管理	容量管理	巡检中心	全链路监控
AIOps	CMDB	监控告警	分析支撑

迁移
云下 to 云上
3rd to 电子云

云服务

超算			
虚拟机超算	裸金属超算	GPU 超算	超算应用

智算			
机器学习	深度学习	大模型服务	智算应用

通用计算								
弹性云服务	云服务器镜像服务	弹性伸缩	虚拟私有云 VPC	云解析	弹性负载均衡	并行文件存储	对象存储	云服务器备份
云容器引擎 CKE	容器镜像仓库	裸金属服务	安全组	NAT 网关	弹性公网 IP	云硬盘服务	云硬盘备份	文件和数据库备份

云操作系统CCOS

集群调度	多集群调度	安装部署	云应用市场	多级云管理

基础设施

多样算力				
NVIDIA GPU	天数智芯	飞腾 CPU	昇腾	Intel x86 CPU
海光 CPU	沐曦 GPU	鲲鹏 CPU	申威 CPU	

存储	
SSD	HDD

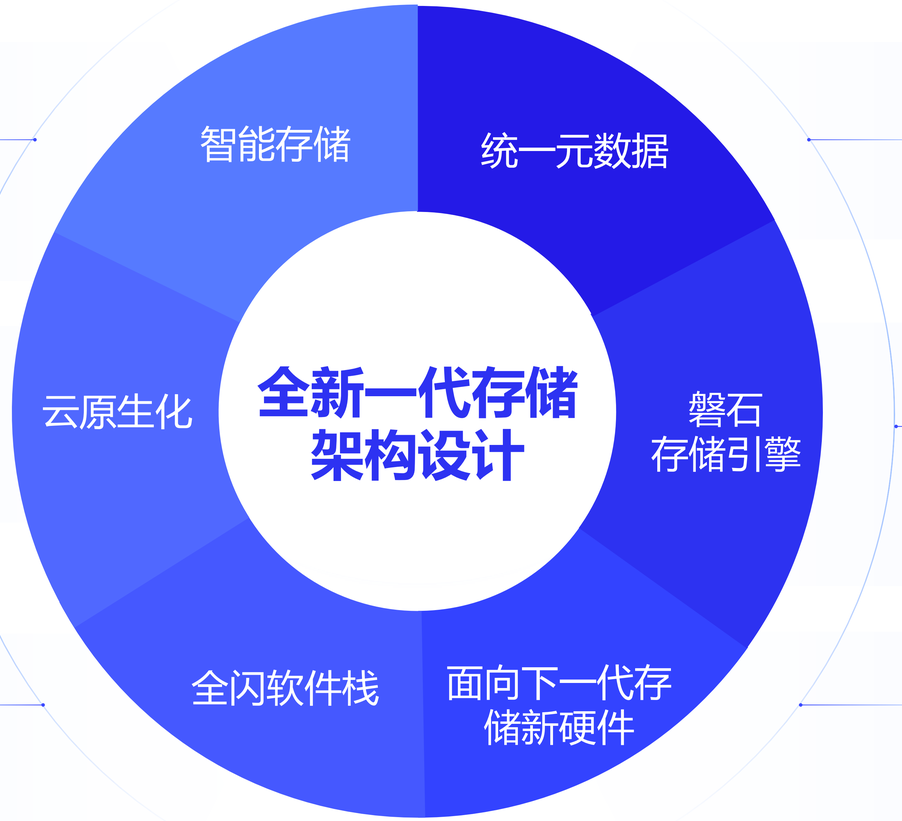
网络	
IB	RoCE

08 中国电子云 | CeaStor 新一代存储架构，行业领先

智能存储
智能运维
智能加速
智能预测
智能诊断

云原生化
持续交付
微服务
容器化

全闪软件栈
IO 栈超轻量化
IO 栈全用户态
Zero-Copy 网络零拷贝
Run to Complete 设计模式
SPDK 和 DPDK 技术结合



统一元数据
统一元数据管理
对象、文件协议互通

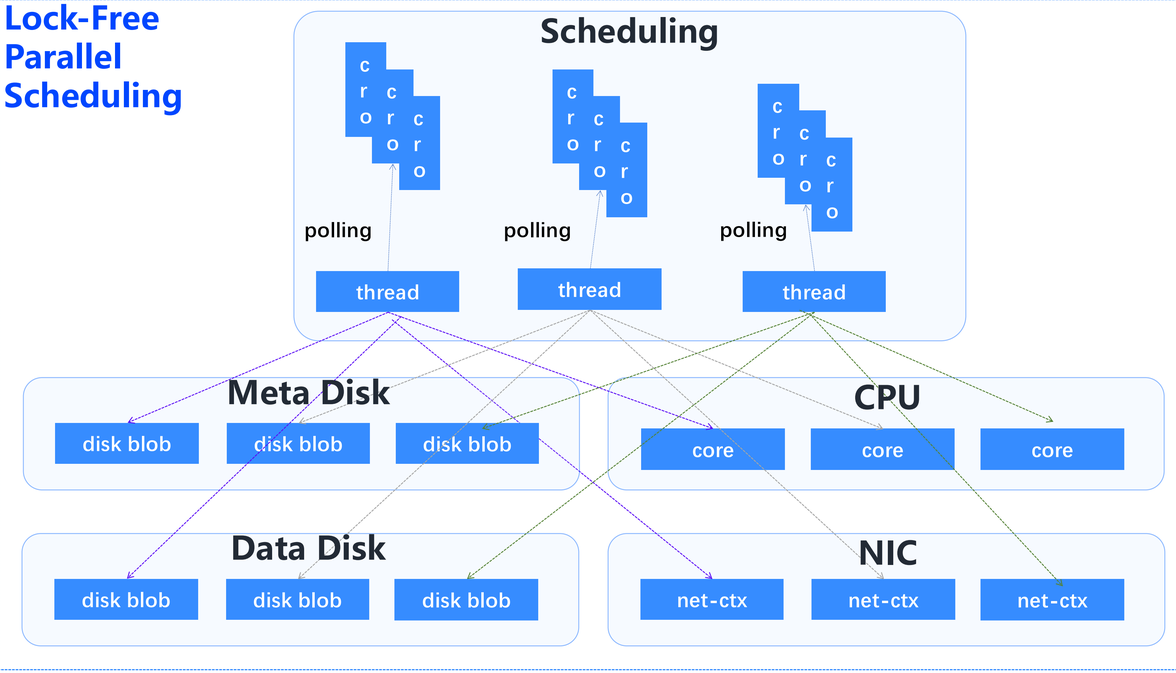
磐石存储引擎
智能数据聚合
智能缓存加速
全局重删
动态 EC
无损快照
容量与负载动态均衡

下一代存储新硬件
SPU（Storage Process Unit）
PCIe Gen 4.0/5.0、PMEM / SCM
open-channel-SSD/ZNS、 NVMe 2.0
NVMe-oF、RoCE v2、IB

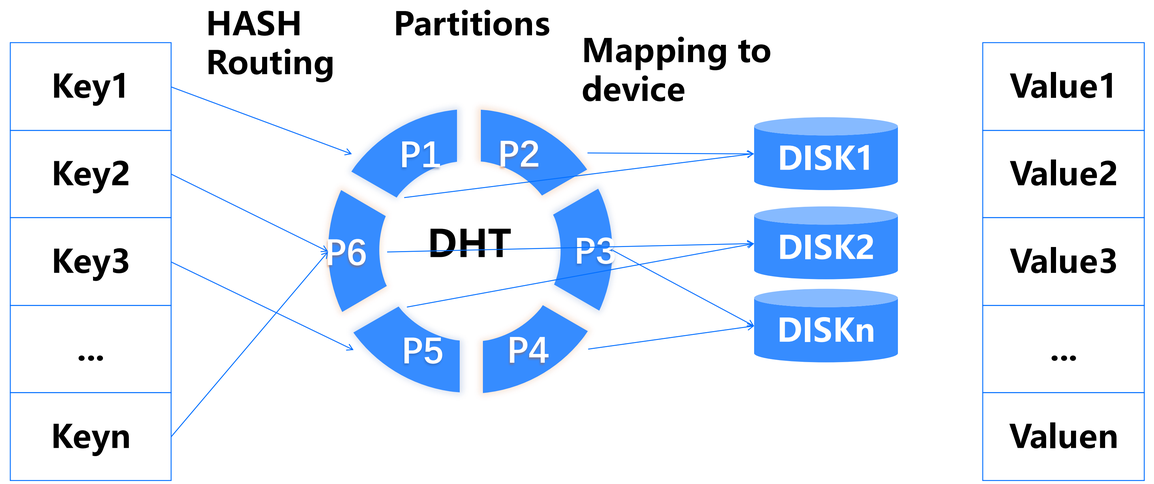
CONTENT

09 高性能“磐石”存储引擎架构

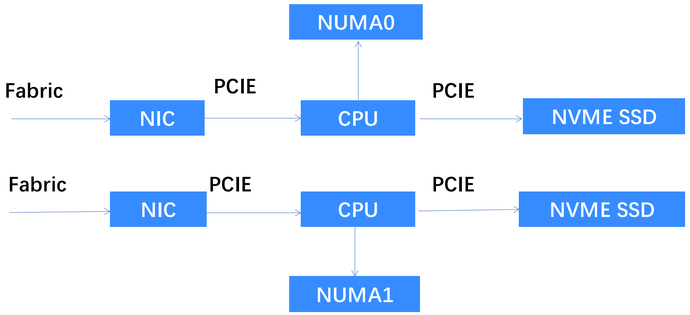
Lock-Free Parallel Scheduling



KV-Based DHT



Hardware Affinity



- 基于**硬件亲和的端到端无锁并行设计**，能够充分释放全闪存的性能潜力。
- 基于**全局去中心化、可扩展的 KV-Based DHT 数据均衡算法**。
- 基于 CRDB 启发的**去中心化 Epoch-Based 分布式一致性算法**，结合全闪存亲和的 **Append Only 架构**，通过简化设计显著提升系统性能。
- 基于 **BW-Tree 理念优化的单机全闪 KV Store 架构**，有效降低 GC 的放大效应。
- 2023年，**SPC-1 基准测试中荣获“全球领先”的认证**。

[1] https://dl.acm.org/doi/10.1145/3318464.3386134
[2] https://dl.acm.org/doi/10.1109/ICDE.2013.6544834
[3] https://dl.acm.org/doi/10.1145/3448016.3457553
[4] https://www.spcresults.org/benchmarks/results/spc1-spc1e

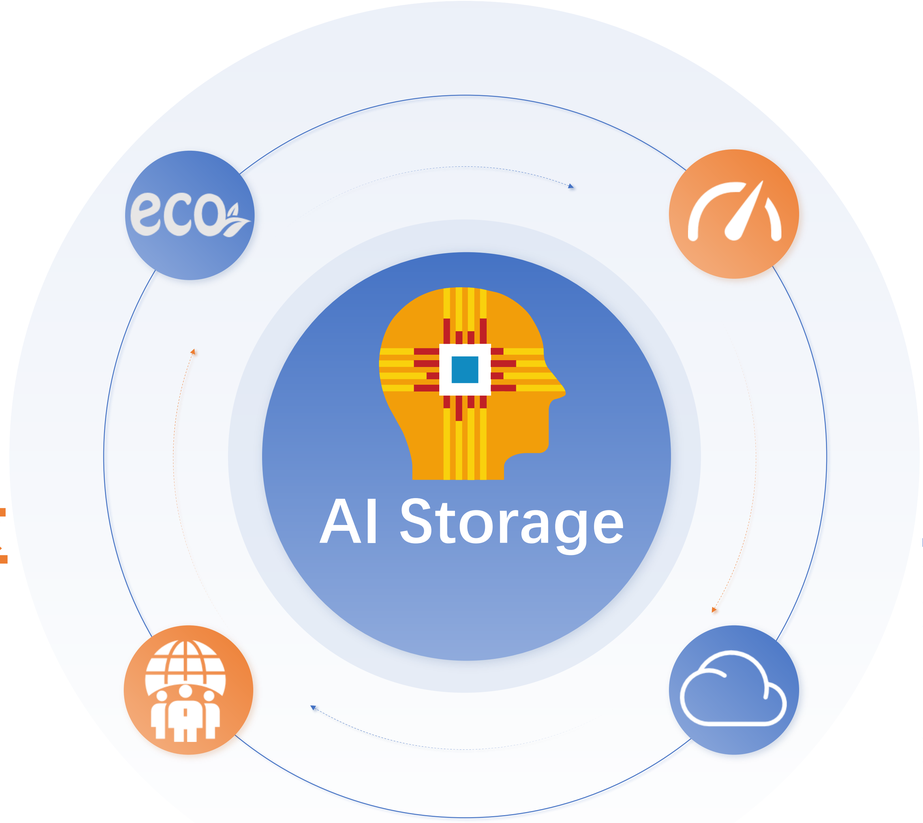
10 Why Lustre ?

生态

➢HPC 市场：**Lustre 46%**，GPFS 22%，PNFS 8%，BeeGFS 6%。

➢TOP8 云厂商：**Lustre 5家**，GPFS 3家，PNFS 1家，DAOS 1家。

Lustre 是并行文件存储领域的领导者。



社区

➢NFS 社区不活跃，且相对不开放，**厂商推进 NFS v3 并行插件极其困难**。

➢BeeGFS 并不能完整兼容 POSIX 协议，且商用需要授权协议。

Lustre 社区一直秉持厂商中立的态度，拥有一支由 DDN、HPE、AWS、Azure 和 Oracle 等知名企业成员组成的活跃领导团队，他们能够迅速响应社区提出的问题。

性能

➢DDN Lustre 在 **NVIDIA SuperPOD 认证中性能排名 TOP3**。

➢NVIDIA 在 LUG 2024 中分享其利用 DDN Lustre 支撑 576 台 H100 GPU 服务器训练 340B 大模型。

➢Oracle CloudWorld 2024 提到其 20 PB 的 Lustre 为其一 AI 大客户服务。

云原生

➢**GPFS 的重客户端扩缩容极慢**，且生产环境只支持单集群 128 节点。

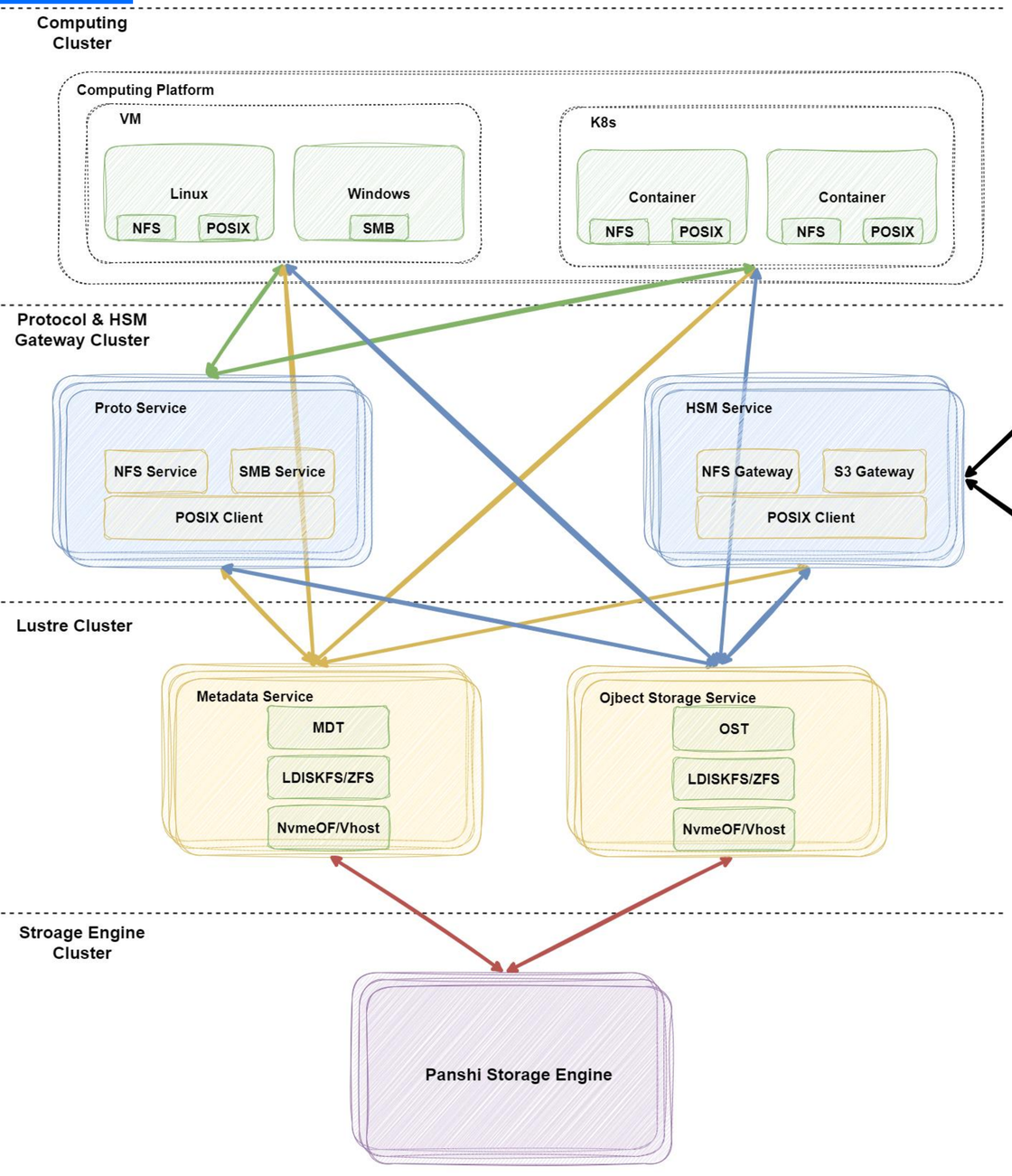
➢**NFS v3 不是并行文件协议**，由于其是非状态协议，**厂商私有的 NFS v3 并行插件不是“True Parallel”**，且存在数据一致性问题。

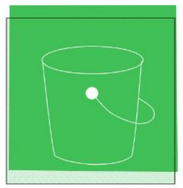
➢NFS v4+ 协议和 K8s 存在较多兼容性问题。

Lustre 支持容器客户端，完备的一致性协议，“True Parallel”架构，可提供高效的大规模数据共享和并发访问能力。

[1] https://www.hpcwire.com/2022/11/08/hyperion-paints-a-positive-picture-of-the-hpc-market/
[2] https://cloud.361way.com/ecology/provider/aaghot/
[3] https://blocksandfiles.com/2024/09/03/nvidia-superpod-storage-certification/
[4] https://wiki.lustre.org/images/d/dd/LUG2024-AI_Workload_Optimization_with_Lustre-Dauchy-Degremont.pdf
[5] https://blogs.oracle.com/cloud-infrastructure/post/building-storage-systems-for-future-oci-roadmap
[6] https://www.youtube.com/watch?v=j9HpfikV_ik
[7] https://wiki.lustre.org/images/c/c9/LUG2024-Community_Release_Update-Jones.pdf
[8] https://wiki.opensfs.org/Lustre_Working_Group

11 AI 文件存储架构





S3

性能

基于高性能的“磐石” 存储引擎和业界领先的 Lustre 并行文件系统，**端到端“True Parallel”架构**，支持**客户端本地盘分布式缓存加速**，单节点支持**数十 GB 级带宽**，单集群支持**百亿文件、数十 TB 级带宽和千万级 IOPS** 。

NFS

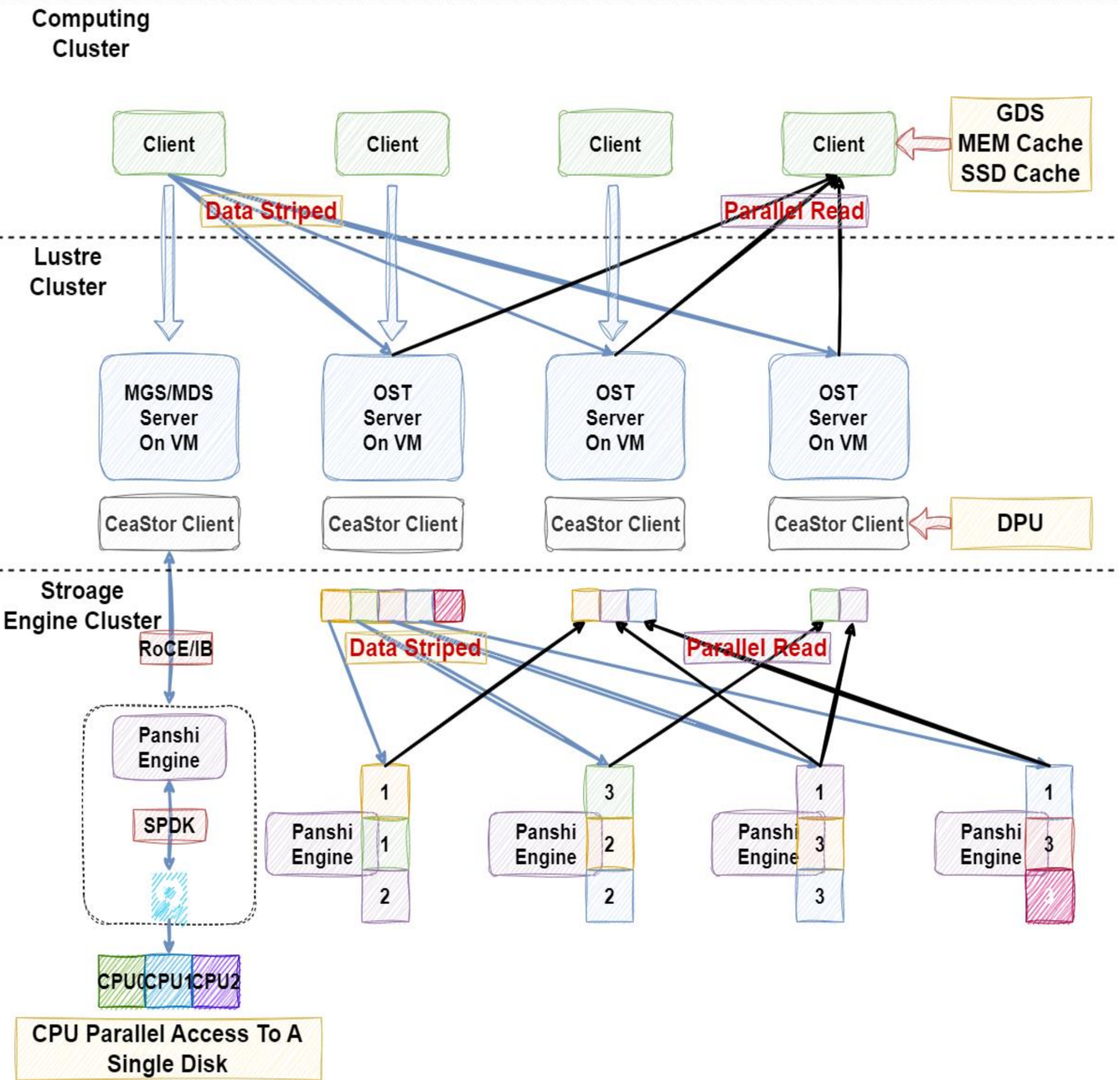
可靠性&扩展性

基于中国电子云 **K8s MicroVM 技术**，不仅可实现**资源高利用率和服务强隔离**，还可实现**秒级故障切换恢复和大规模横向扩展**。

智能数据流动

支持并行文件系统和**异构 S3 、NFS 的智能数据流动**，不仅可实现高效数据流转，还可实现将**数据分层到冷存储**，降低存储成本。

12 AI 文件存储高性能架构

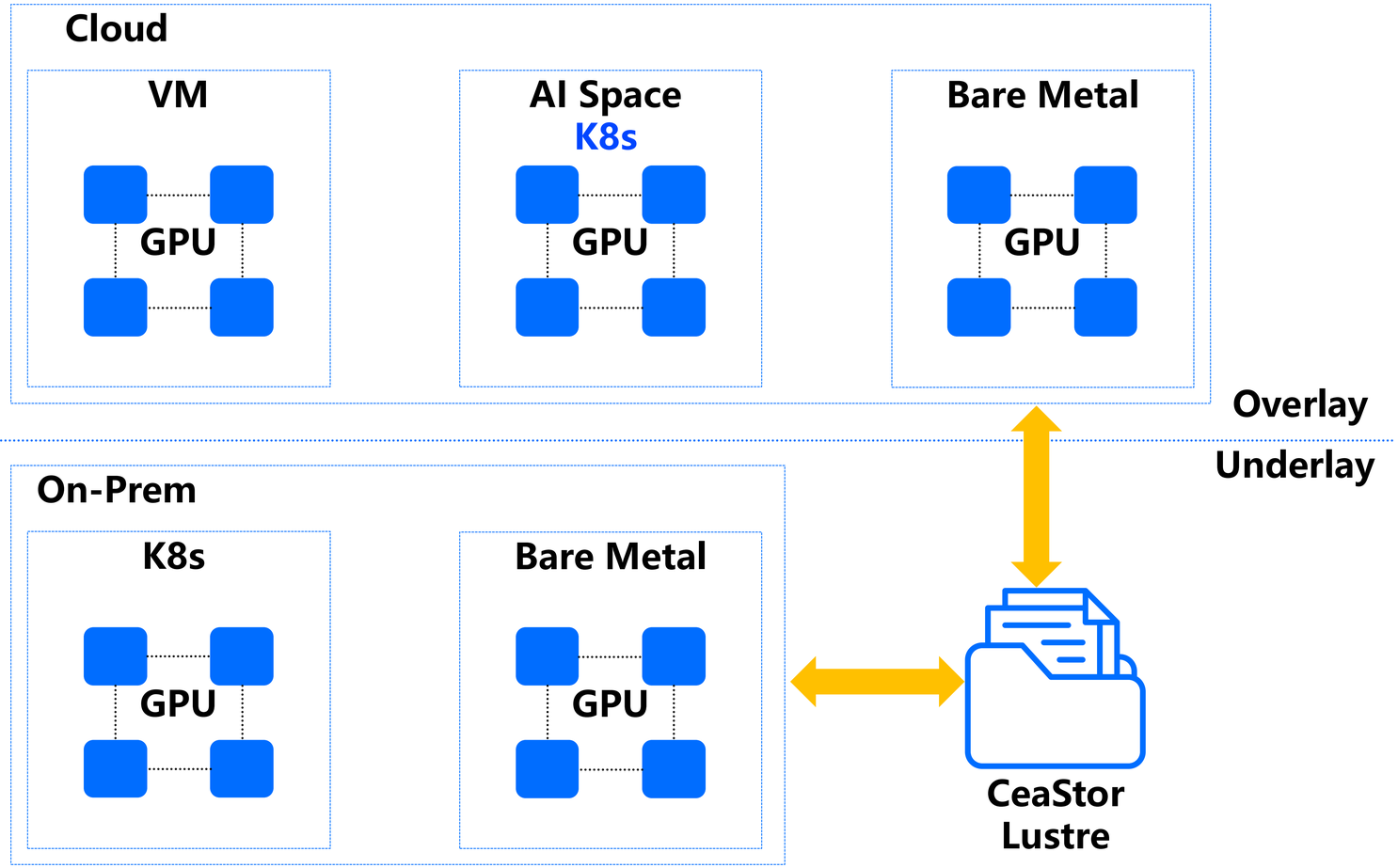


- CeaStor 软件优化：
  - **RDMA：**降低网络时延和 CPU 消耗。
  - **SPDK：**降低磁盘读写时延和 CPU 消耗。
  - **Append Only Write：**最大限度利用磁盘吞吐。
  - **客户端加速：**GDS 加速、内存缓存加速、SSD 缓存加速。
  - **Vhost User Block：**减少网络跳数，并支持 DPU 卸载。
  - **E2E “True Parallel”架构**：充分利用多机并发能力。
  - **E2E Zero Copy**：降低 CPU 消耗，提升数据吞吐。
  - **磁盘分区绑核：**充分利用多 CPU 发挥 NVME 磁盘极致性能。
  - **优化线程模型，使用无锁数据结构：**IO 路径避免线程切换开销，降低毛刺。
  - **慢盘隔离：**自动检测慢盘，数据迁移。
- CeaStor 硬件优化：
  - 联合CEC旗下中国长城，支持 5 代 Intel X86 CPU，支持 PICE5，支持 DDR5，支持 DPU卸载，支持 NVME 硬盘直通，支持 400G 网络等新硬件。

预计在 2025 年 Q1 发布 MLPerf Storage 基准测试报告。

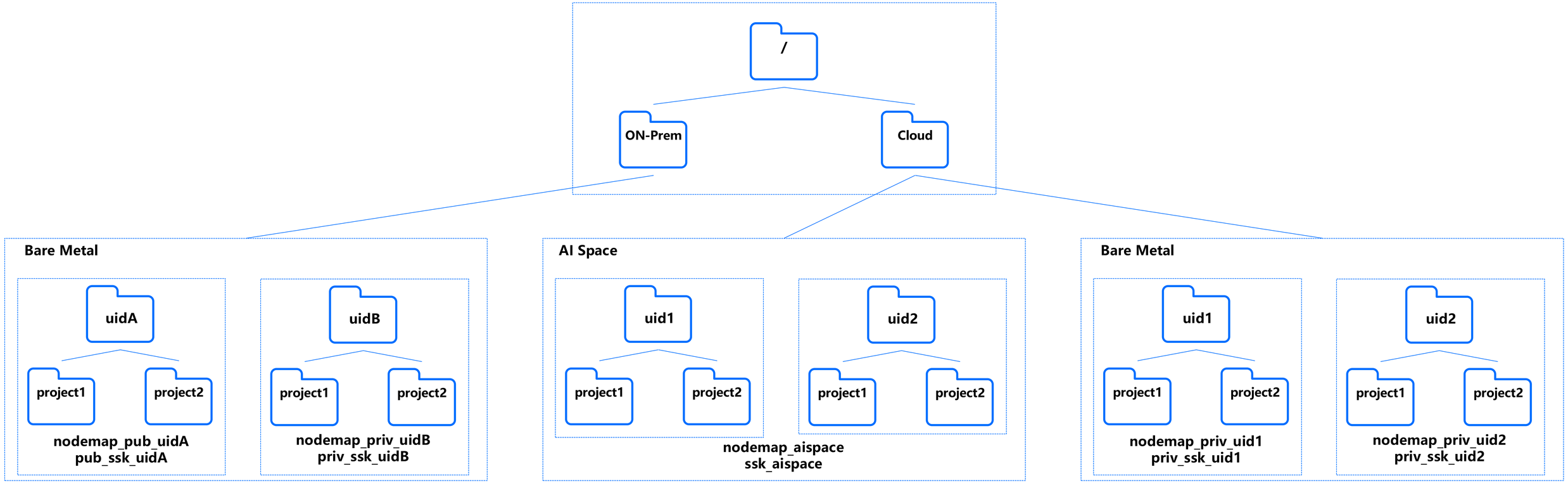
CONTENT

13 Lustre 在私有云 AI 场景的需求



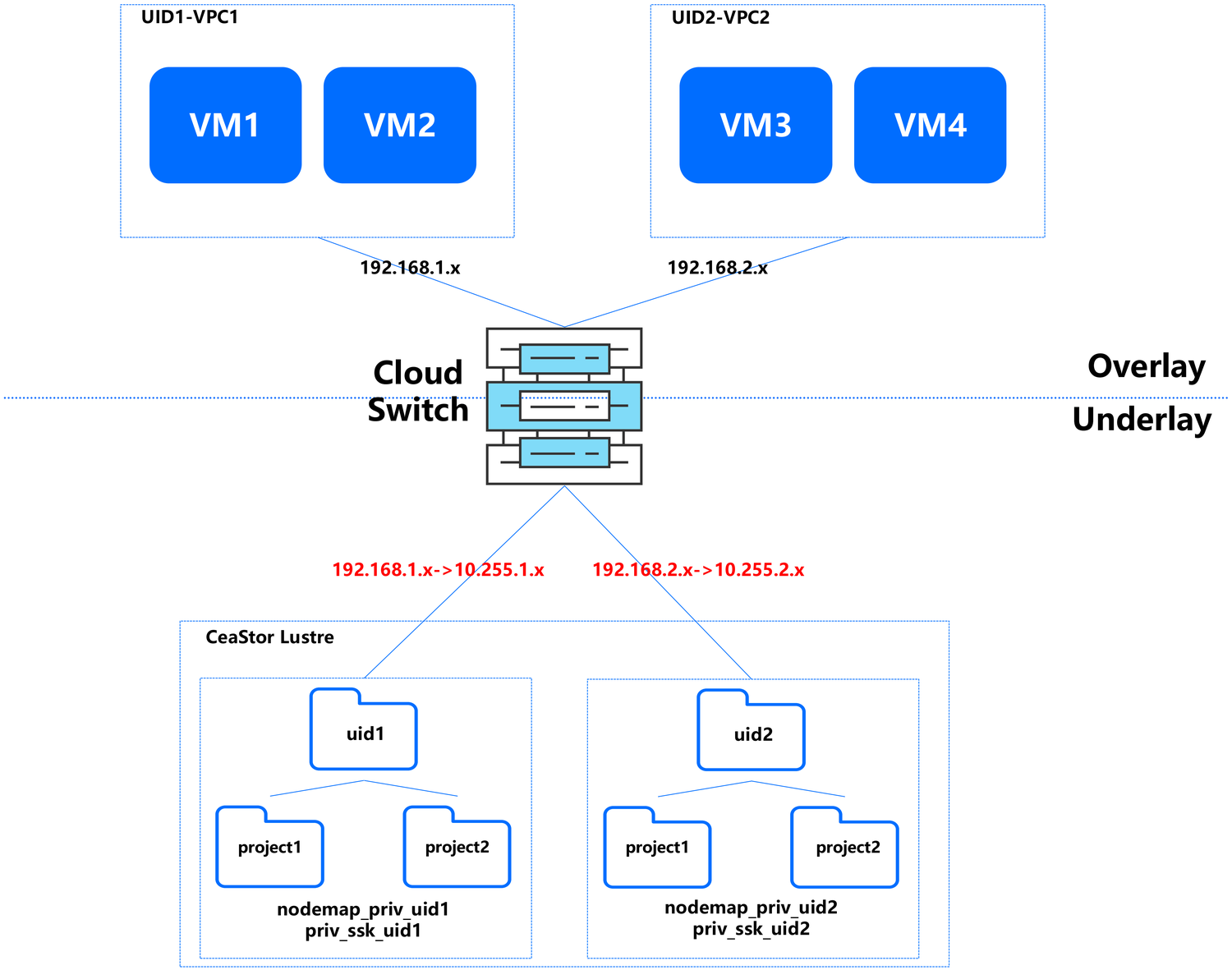
- **网络兼容性：**Lustre 需适配客户复杂的云和私有环境，支持 Overlay 和 Underlay 网络，但目前无法与云 VPC 网络隔离。
- **容量管理：**需要对多 Volume 进行容量 和 Inode 的Quota 管理，包括 Root 用户，但 Lustre LTS 2.15.5 不支持 Root 用户 Quota。
- **QoS能力：**Lustre 需支持全面的端到端 QoS，目前仅提供有限的 QoS 功能。
- **数据均衡与风险控制：**Lustre 需支持多 Volume 数据均衡和风险控制，而不仅仅是运维层面的数据分布控制。

14 Lustre 在私有云 AI 场景的实践



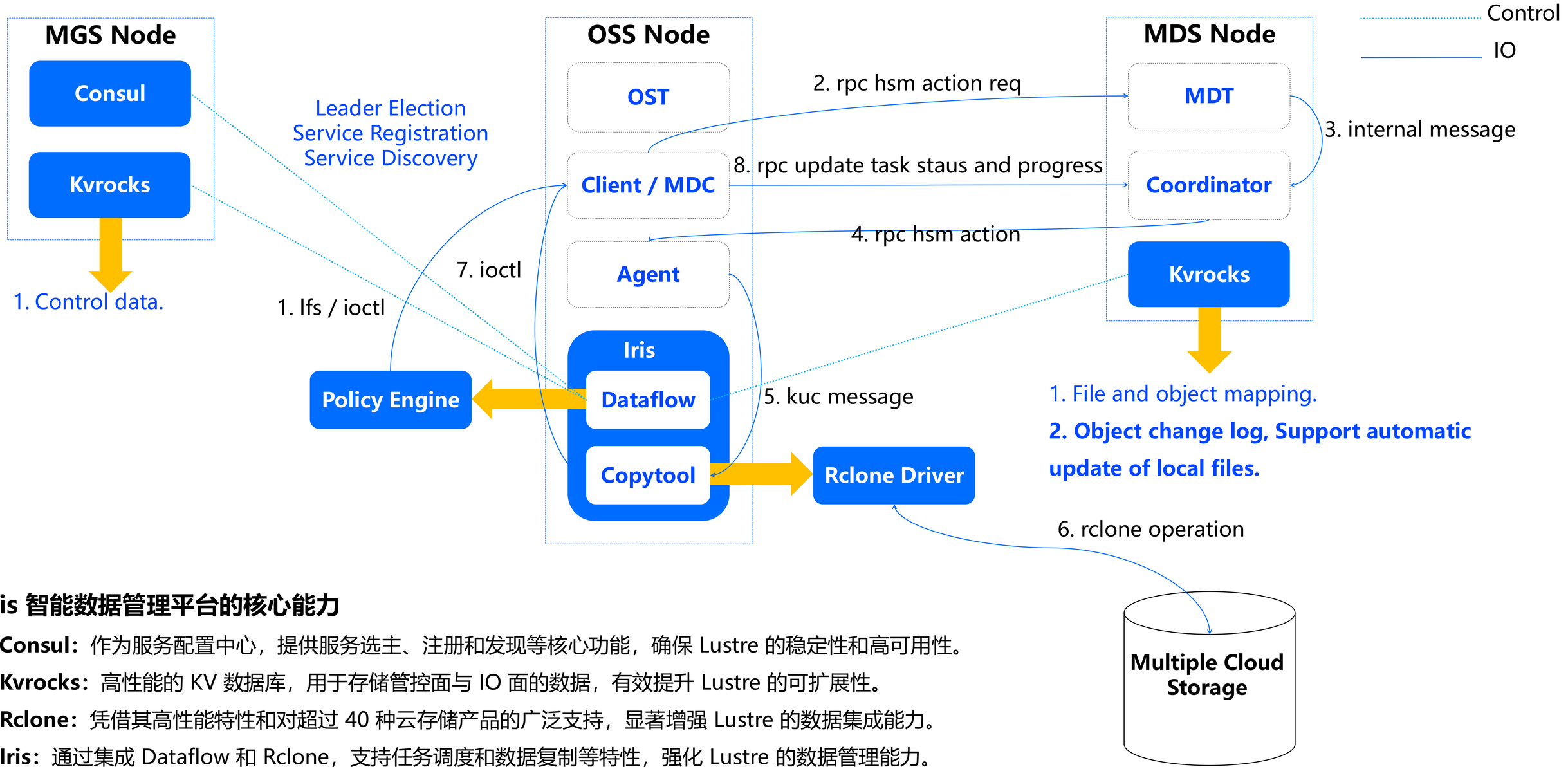
- **数据安全隔离：**分开云服务和客户私有环境的访问权限，保护数据安全。
- **多租户数据隔离：**使用 Lustre 的 Fileset、NodeMap、Quota 和 SSK 功能来隔离不同租户的数据目录。
- **Root 用户 Quota 支持：**将 LU-16415 补丁集成到 Lustre LTS 2.15.5，以便对 Root 用户的容量进行限制。
- **流量控制：**利用 Lustre QoS 功能对客户端流量进行优先级控制。
- **数据均衡与风险控制：**合理分配单一租户的数据分布，同时分散多租户的数据存储，以保持数据的均衡，从而降低数据集中可能引发的风险。

15 Lustre 的 VPC 网络隔离的实践



- **网络整合方案：**采用 Cloud Switch 技术将 VPC 网络映射至经典网络，同时借助 Lustre 的 NodeMap 和 SSK 功能实现租户访问路径的隔离，确保网络隔离和数据安全性。
- **性能与实现挑战**：由于流量需经网关处理，导致网络性能下降。Cloud Switch 的实现复杂，且在支持 RDMA 网络方面面临重大挑战。
- **性能优化策略：**计划开发 Lustre 的 Access Point 功能，实现 RDMA 网络直连 VM，以提升网络性能并简化 RDMA 网络的支持。

16 Lustre 和 S3 数据流动的实践



- **Iris 智能数据管理平台的核心能力**
  - **Consul**：作为服务配置中心，提供服务选主、注册和发现等核心功能，确保 Lustre 的稳定性和高可用性。
  - **Kvrocks**：高性能的 KV 数据库，用于存储管控面与 IO 面的数据，有效提升 Lustre 的可扩展性。
  - **Rclone**：凭借其高性能特性和对超过 40 种云存储产品的广泛支持，显著增强 Lustre 的数据集成能力。
  - **Iris**：通过集成 Dataflow 和 Rclone，支持任务调度和数据复制等特性，强化 Lustre 的数据管理能力。

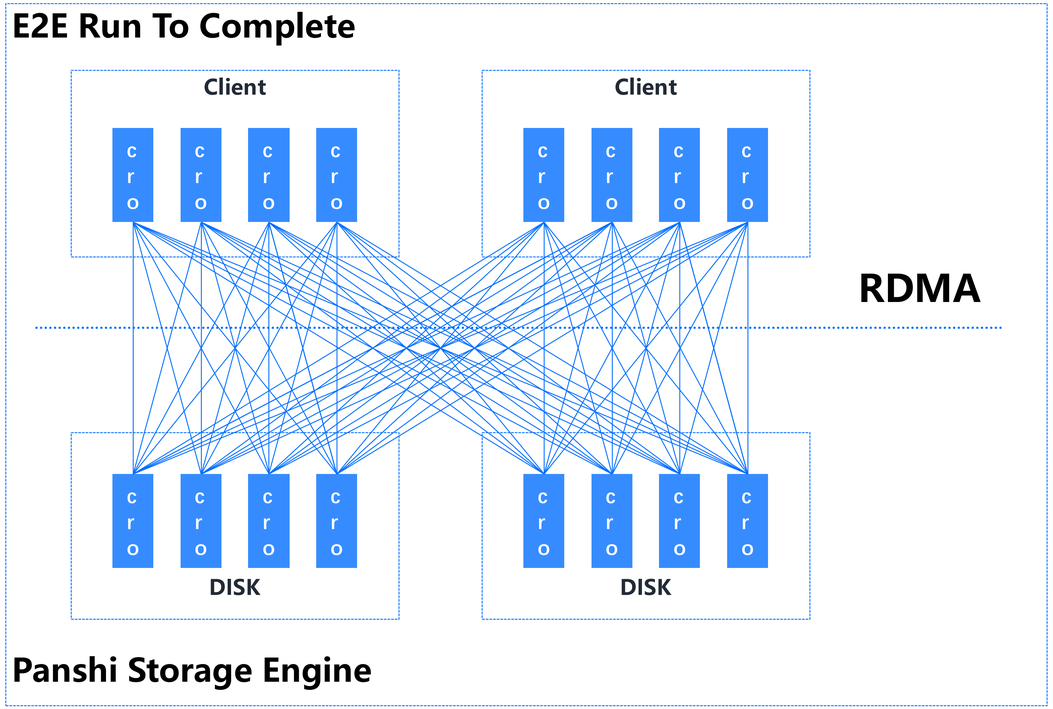
CONTENT

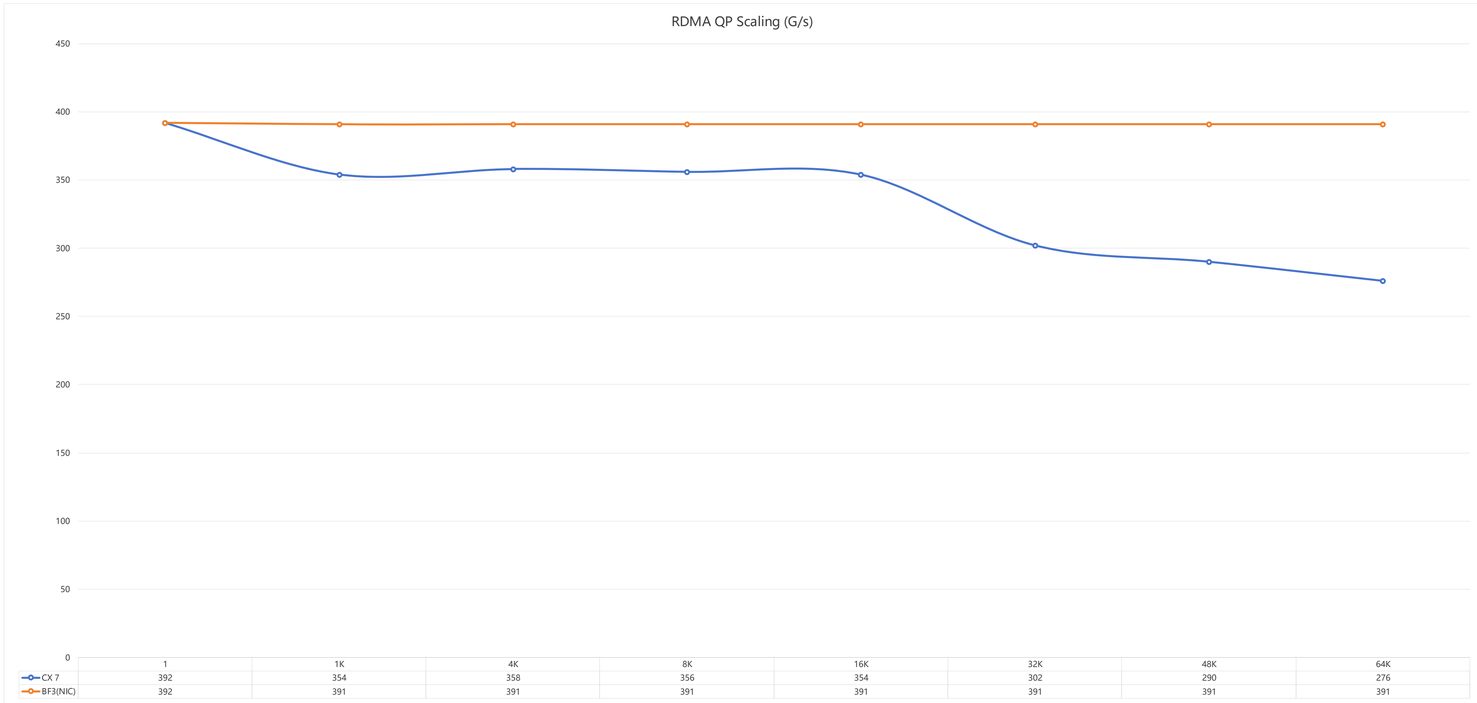
17 Lustre 的私有云稳定性的挑战

Lustre 的生产环境稳定性正在不断提升，我们将持续与社区合作，致力于推动 LTS 分支的稳定性发展。

- Lustre LTS 2.15.5
- Features
  - [LU-16415] Lustre quota enforcement for root files in project quota
- Bug Fixes
  - RDMA
    - [LU-18260] o2iblnd: graceful handling of unexpected RDMA_CM_EVENT_REJECTED
    - [LU-18275] o2iblnd: unable to handle kernel NULL pointer dereference in kiblnd_cm_callback when receiving RDMA_CM_EVENT_UNREACHABLE
    - [LU-18364] rdma_cm: unable to handle kernel NULL pointer dereference in process_one_work when disconnect
    - [LU-17480] lustre_rmmod hangs if a lnet route is down
    - [LU-16184] o2iblnd: set TX deadline when adding to peer queue
    - [LU-17632] o2iblnd: graceful handling of unexpected CM_EVENT_CONNECT_ERROR
    - [LU-17325] o2iblnd: graceful handling of CM_EVENT_UNREACHABLE on established connection
    - [LU-15885] o2iblnd: RDMA_CM_EVENT_UNREACHABLE may be received after conn clean-up
    - [LU-18426] reboot lustre server, crash in mlx5_del_flow_rules
  - MDS
    - [LU-18425] Failed to rmdir and ls shows "No such file or directory" after rebooting MDS
  - Others
    - [LU-16180] lustre 2.14.0_ddn54 + 5.15 kernel soft cpu lockups
    - [LU-13705] allow llstat to work properly on clients

18 RDMA 网络多 QP 的性能衰减的挑战





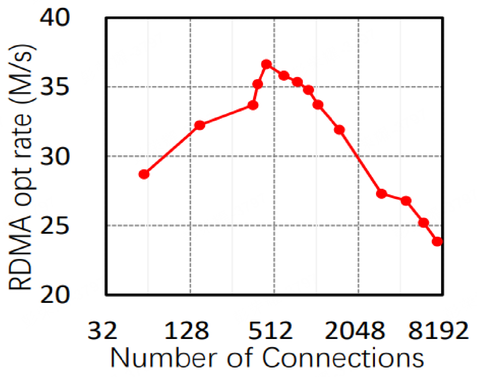


Fig. 1. Performance degrades when the number of concurrent connections grows on Mellanox ConnectX-6 Dx EN NIC.

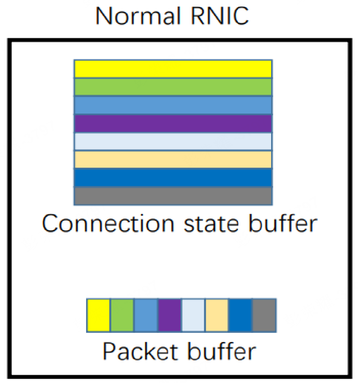


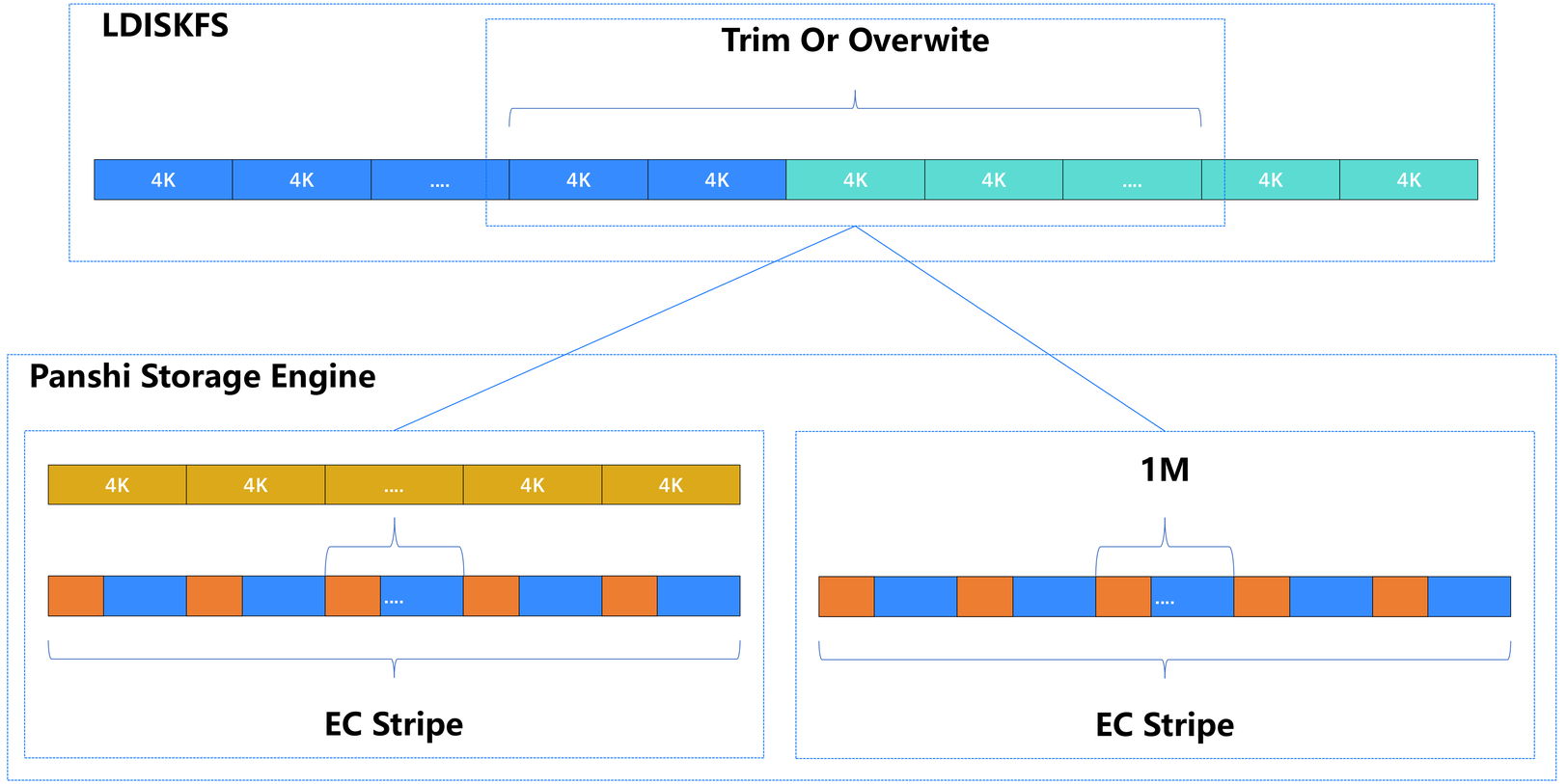
Fig. 2. State and packet buffer in normal RNICs.

➢**挑战**：在存储引擎中，客户端与服务端之间的 **IO 采用 Run To Complete 模型**。随着 **CPU 和磁盘资源的扩展，RDMA QP（Queue Pair）的数量会成倍增加**，这最终可能导致性能下降。

➢**策略**：随着 AIGC 技术的蓬勃发展，GPU 集群规模不断扩大，DPU 已成为 GPU 服务器的标准配置。同时，对存储集群的规模和性能要求也日益提高。在这种背景下，RDMA QP Scaling 不仅存在于计算领域，也涉及存储领域。因此，存储服务器应顺应技术发展的趋势，适时引入 DPU 卸载技术，以提升整体性能和效率。

[1] https://icnp21.cs.ucr.edu/papers/icnp21camera-paper30.pdf
[2] https://www.researchsquare.com/article/rs-4174332/v1.pdf
[3] https://cs.stanford.edu/~keithw/sigcomm2024/sigcomm24-final246-acmpaginated.pdf

19 全闪分布式存储 Trim 的挑战

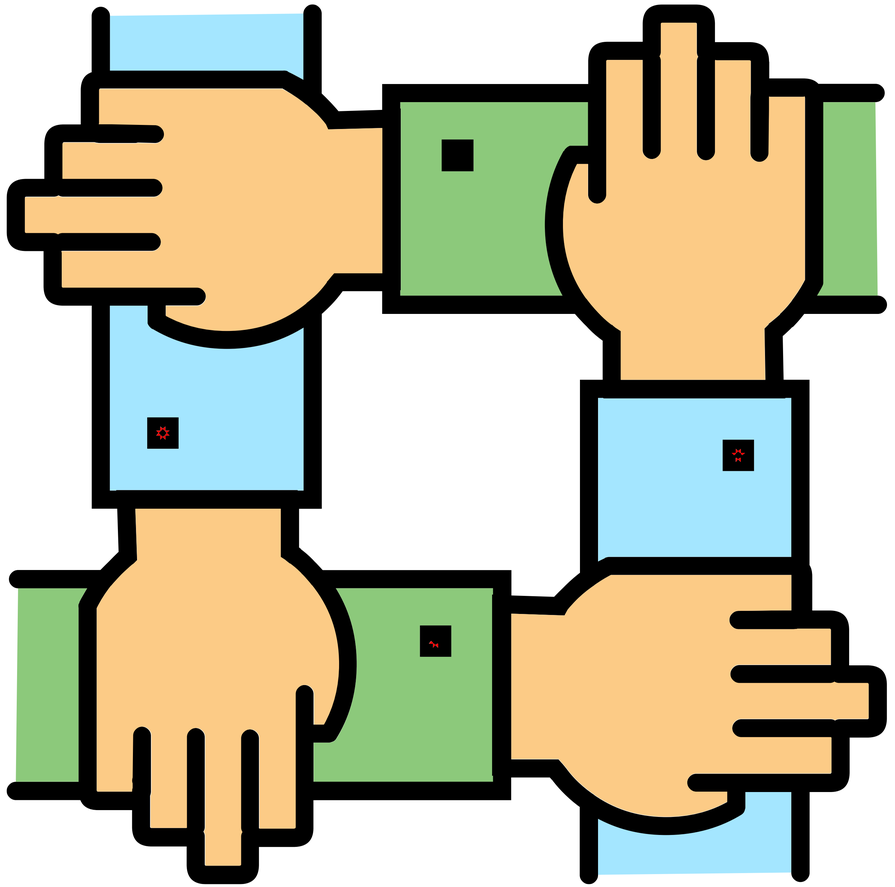


- **挑战：**存储引擎采用 Append Only 架构，在遇到非对齐的 Trim 操作或 Overwrite 数据时，全局垃圾回收（GC）的效率会显著降低，进而影响全闪存架构的性能。
- **策略**：
  - 改进 LDISKFS 的 Trim 策略，以确保 Trim 操作能够覆盖整个条带，从而提高效率。
  - 降低数据覆盖（Overwrite）的可能性，通过算法分析删除操作的频率，并实现实时的整条带 Trim，以优化存储管理。
  - 针对 AI 场景，优化全局垃圾回收（GC）算法以适应 Burst I/O 需求，确保在高负载情况下也能保持性能。

CONTENT

20 未来规划 & 社区合作

- **未来规划：**
  - Lustre 支持与 VPC 网络兼容的 Access Point 功能，以实现更高效的网络通信。
  - Lustre 支持 Fileset 的集群级 QoS，以确保不同工作负载的性能要求得到满足。
  - Lustre 支持对任意用户设置配额，优化存储资源分配。
  - Lustre 支持云 VM 的 RDMA 直通技术，提高数据传输效率。
  - Lustre 的自动化运维能力提升，简化存储管理流程。
- **社区合作：**
  - 持续投入资源以修复社区中发现的 Bug，确保 Lustre 系统的稳定性和可靠性。
  - 提升 Lustre 的云集成能力，优化其与云环境的兼容性和集成效率。
  - 加强 Lustre 的数据管理功能，特别是在数据流动和数据迁移方面，以提高数据处理的灵活性和效率。
  - 积极参与 Coral 项目的开发，以推动 Lustre 技术的进步和创新。



CEC 中国电子 | 中国电子云

THANKS