云可信 创未来

Lustre 在国产化浪潮下的探索与实践之路

彭荣耀—中国电子云存储团队

2025/10/17

CEC 中国电子 | 中国电子云

CONTENT

01 中国电子云的产品及应用

02 中国电子云在 Lustre 的探索和实践

03 国产软硬件平台协同适配与融合机制研究

04 中国电子云信创 Lustre 开源生态建设倡议

01 中国电子云｜承载集团战略 聚焦主责主业

为国防、政府、金融、能源、交通等重要行业关键领域服务

重构计算产业体系 形成五大产业布局

计算产业是核心　集成电路是基础

网络安全是保障　数据应用是目标　高新电子是高地



芯片	设备	软件	平台	场景
CPU	服务器	操作系统	云平台	国防　政务
GPU	PC整机	数据库	AI边缘平台	金融　能源
XPU	交换设备	中间件	HPC使能平台	交通　教育
BMC	存储设备	办公软件	工业互联网平台	医疗　安防
FPGA	备份设备	工控软件		
MCU	移动终端	工业软件		

自主计算产业体系

网络安全	应急响应	内生安全	态势感知	可信计算	网络攻防	网络靶场
系统应用	软件行业第一品牌		金融全栈全域替代		粤省事 粤商通 粤政易	
	数字与信息服务		移动办公政企首选		国内最全医疗数据	
计算平台	数据治理	数据安全与数据要素化平台	云计算	中国信创云	智能制造	国家级“双跨”工业互联网平台
终端产品	计算整机		新型显示	出货量全球第一	网络整机		PLC	工控领域打破垄断
支撑软件	操作系统		数据库		存储	领先一代的分布式数据基础设施	
基础芯片	CPU		FPGA	国产化民用市场第一	内存接口芯片	全球领先	交换芯片	自主可控 国内第一
生态构建	PKS生态大会	直播浏览人次 2000万+	PKS生态伙伴	近万家 生态合作伙伴	产教融合	首批63家国家产教融合企业之一		

02 一云多芯、一云多算”的一体化算力平台

CEC 中国电子 | 中国电子云

CECSTACK 一体化算力平台，为客户提供集通用计算、智能计算和高性能计算于一体的统一算力底座，实现多元异构算力的统一抽象、灵活调度和即取即用，输出高安全、高性能全栈云服务能力，深化赋能政府和行业数字化的效能跃升



03 中国电子云 | 算力基础设施产品体系

政务 | 央企 | 金融 | 能源 | 医疗 | 教育 | 交通

服务数字中国建设
全面助力数字化转型

把握数字时代机遇
抢占计算产业制高点

聚焦核心平台技术
打造产业核心竞争力

专属云

- CECSTACK 全栈云
- CECSTACK 敏捷版

服务器虚拟化

- CeaSphere 企业版
- CeaSphere 涉密版

存算一体机

- CeaStor 分布式存储
- CeaCube 智算一体机

一云多芯

信创原生

原生安全

CCOS云操作系统

一套调度，管理所有服务

云原生 + AI原生

“云原生+AI 原生”的设计理念

雨燕架构

硬件加速，极致性价比

AI+

高性能存储

开放合作

CCLinux 操作系统

04 CeaStor | 技术领先的高性能全闪分布式存储

架构领先

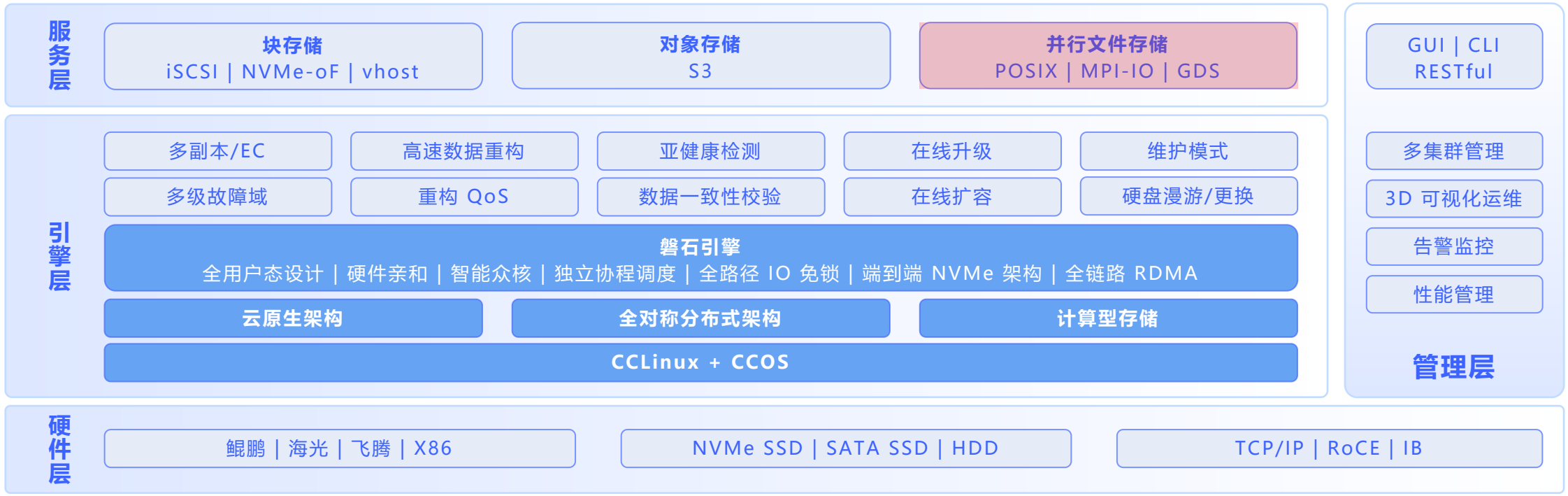
- 全闪架构
- 自研“磐石引擎”
- 云原生+全对称分布式架构
- 计算型存储

性能领先

- 全链路 RDMA，端到端 NVMe
- 千万级 IOPS、数十 TB 级带宽
- 时延低至 200 微秒
- 上万节点、EB 级容量、千亿对象

安全领先

- 底座+软件全栈自研，全国产硬件
- 专业团队，环境加固漏扫
- 6 个 9 高可用，故障跌零 < 10 秒
- 数据隔离、访问鉴权、国密加密



05 CeaStor | 为核心应用场景提供坚实数据基础设施

云及虚拟化

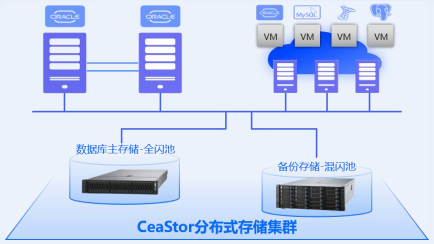


企业级云平台：上万节点，EB 容量，千万 IOPS，6个9可用性

精准匹配业务：性能池/均衡池/容量池，结构化/非结构化

智能管理：全自动资源调配、故障自愈、3D 可视化运维

企业级应用



交易型/分析型数据库，各类企业级业务应用

端到端 NVMe 架构，亚毫秒级时延

16条路径负载均衡，故障时IO跌零＜10秒，重构速率＞6TB/h

人工智能/高性能计算



数十 TB 聚合带宽、亚毫秒级时延

POSIX 接口、支持 MPI-IO，支持 GDS

跨异构对象存储智能数据流动，冷热分层

海量非结构化数据

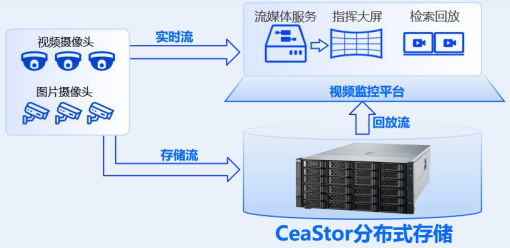


单桶千亿对象，海量文件共享存储

ROW 聚合写，智能预读技术，小文件高性能访问

数据生命周期管理，热温冷数据分层，均衡性能和成本

视频监控



内置视频流直存模块，摄像头数据直写

单节点 300+ 路视频并发写入，按需增加节点扩容

硬件故障对业务无感，视频不丢帧，数据不丢失

广电媒资



单客户端性能高达 2GB/s，轻松应对 4K 高清的非编要求

持单桶千亿对象承载能力，有效解决海量文件共享存储问题

通过主流 ISV 联合测试，业务应用无缝对接

CONTENT

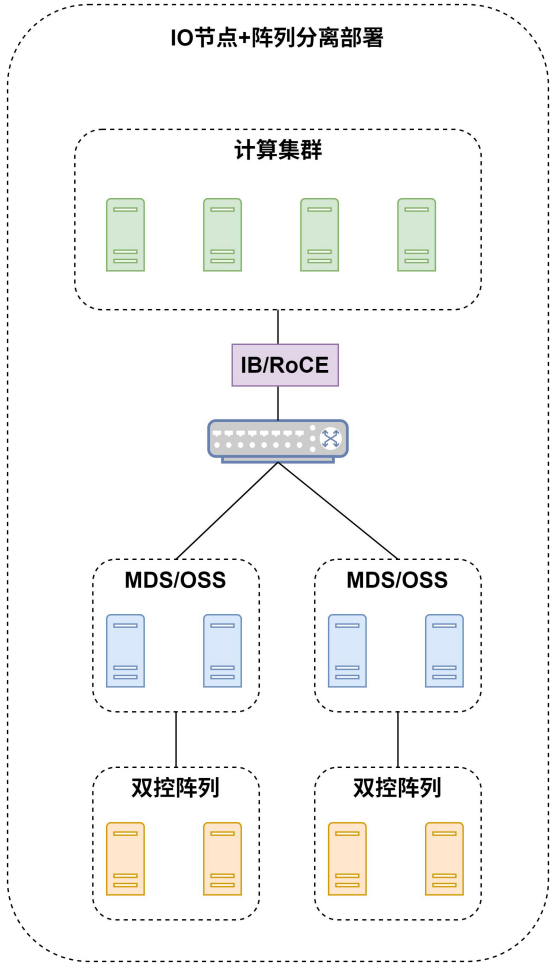
01 中国电子云的产品及应用

02 中国电子云在 Lustre 的探索和实践

03 国产软硬件平台协同适配与融合机制研究

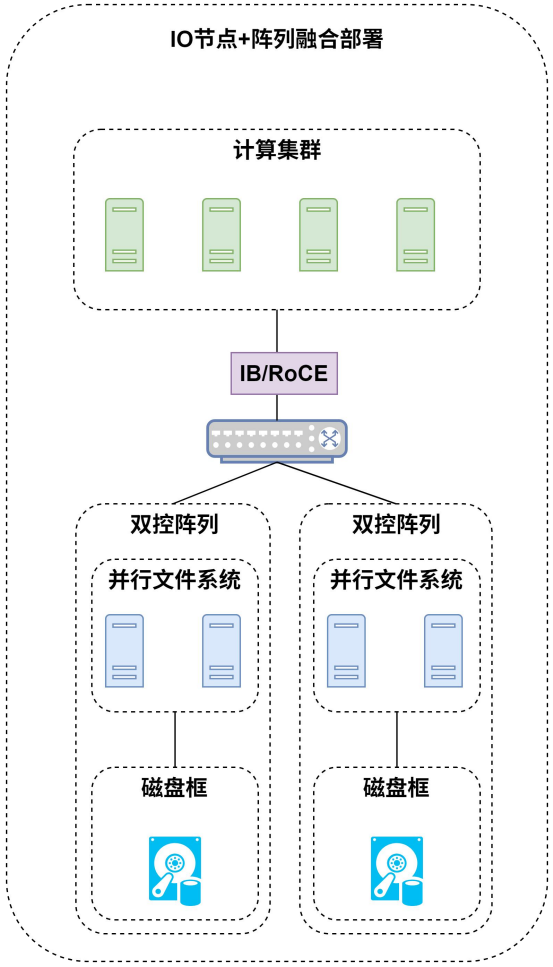
04 中国电子云信创 Lustre 开源生态建设倡议

06 业界典型的 Lustre 技术架构挑战



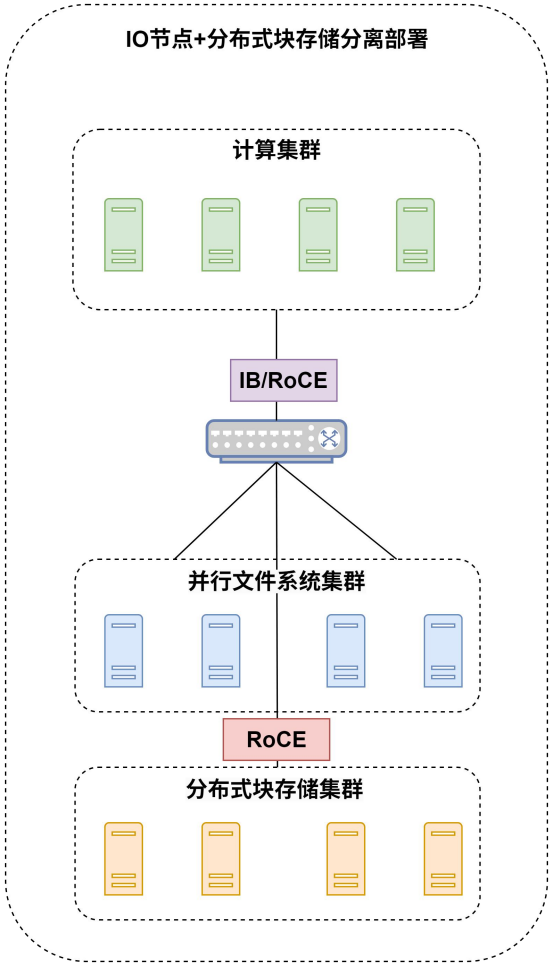
性能一般

需要独立的 IO 节点和翻倍的 RDMA 网卡，性能依赖标准 NVMeOF 协议



性能好

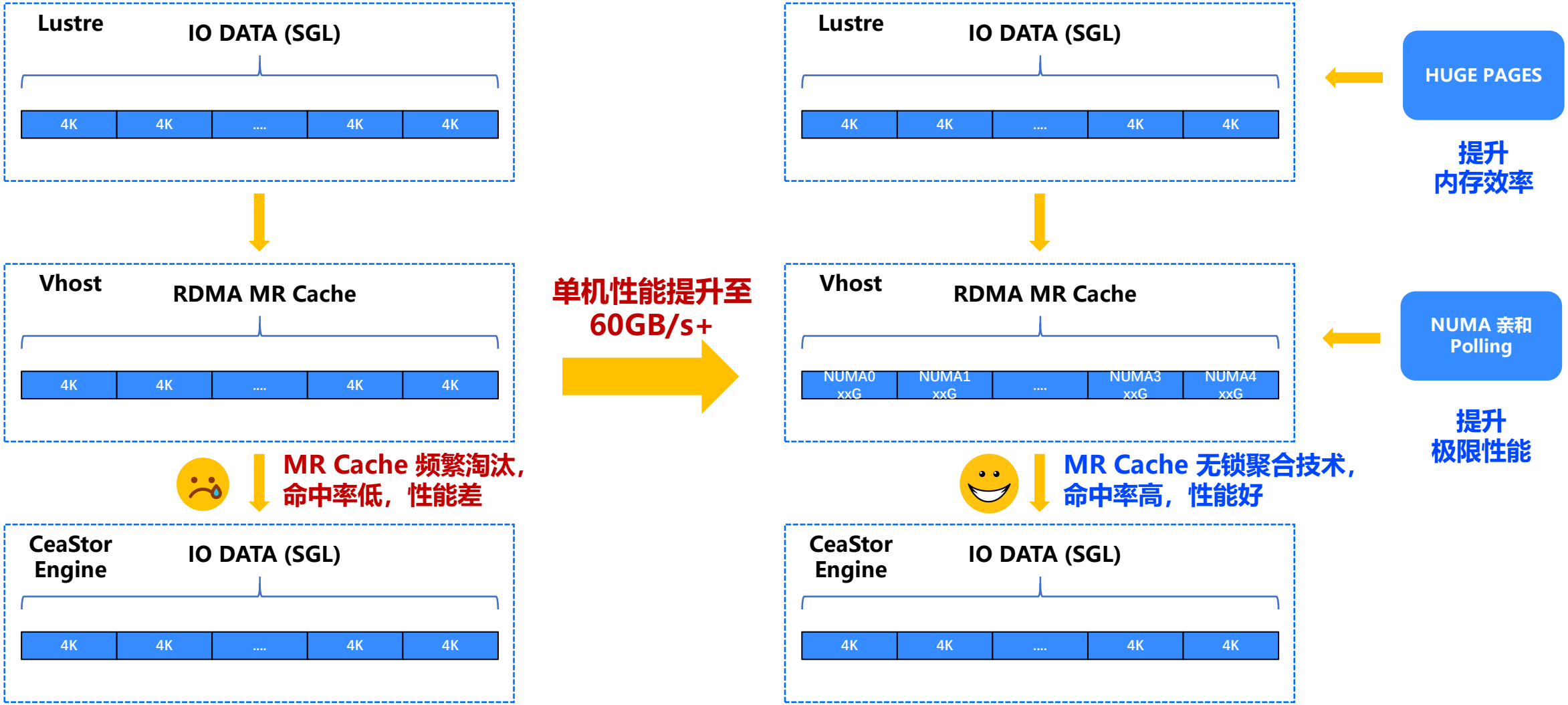
无需独立的IO节点，定制化硬件，支持私有 RAID 块协议



性能一般

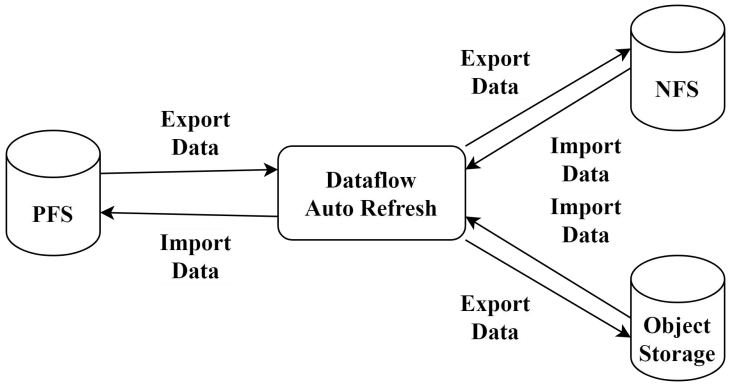
需要独立的 IO 节点和翻倍的 RDMA 网卡，性能依赖标准 NVMeOF 协议

07 CeaStor | 低成本和高性能的融合部署加速技术实践

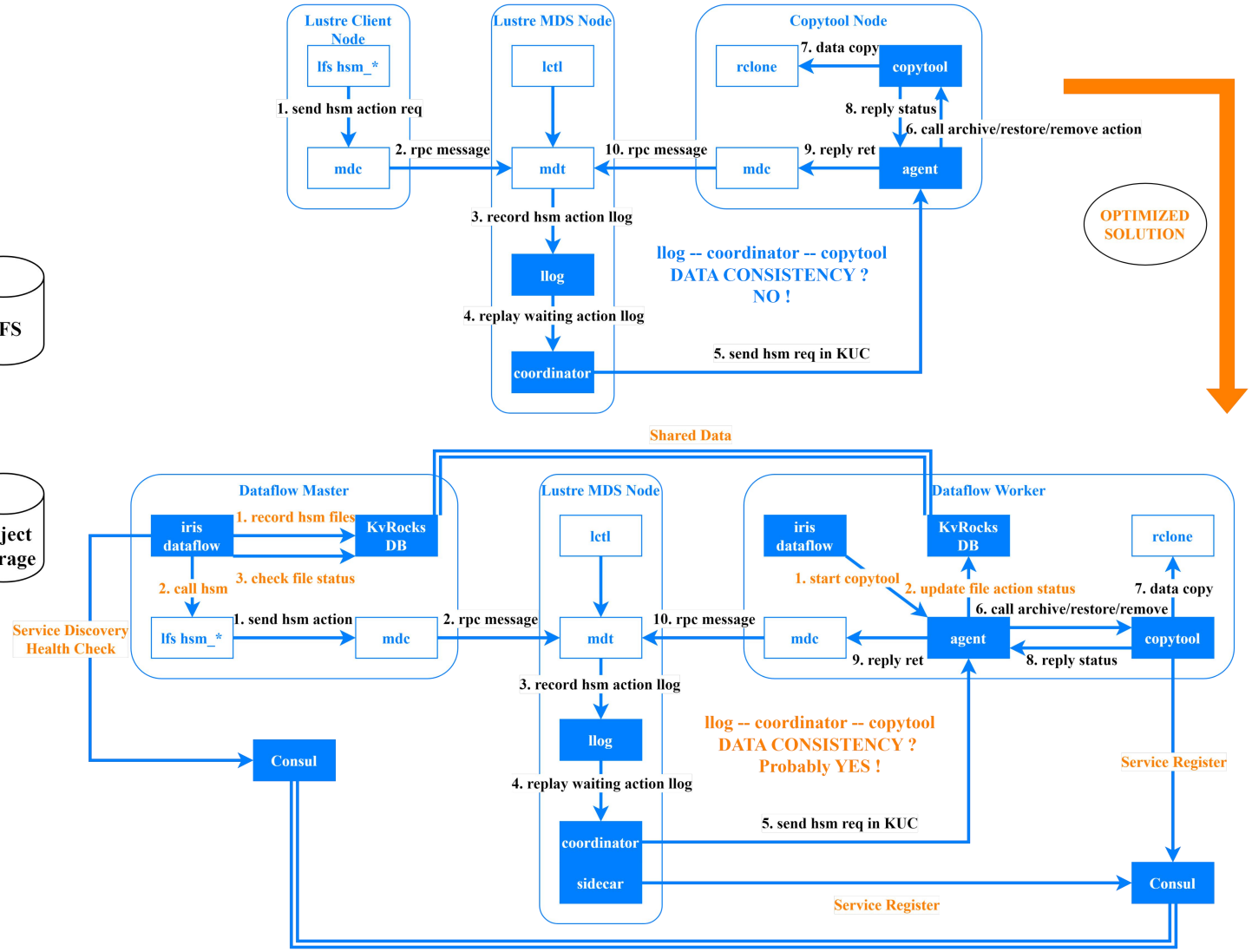


08 CeaStor | Lustre 数据流动机制优化实践

Lustre 数据流动基本流程



Lustre 数据流动核心机制



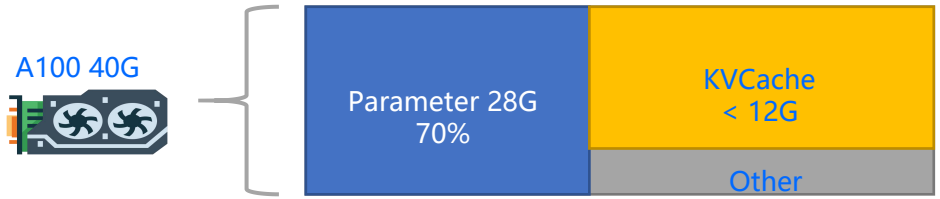
分布式大规模

最终一致

秒级自动
断点续传

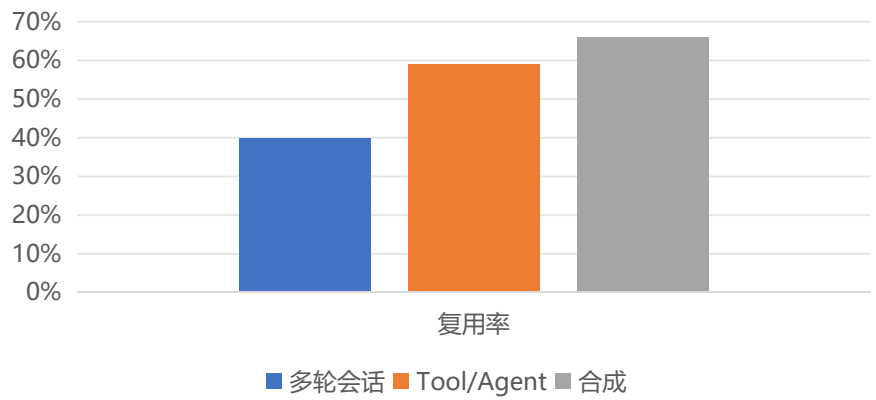
09 CeaStor | KVCache 在 Lustre 中的融合研究

显存容量明显不足



- Qwen3-14B 存储 1k token 的键值缓存需要大约 1G 显存
- Qwen3-14B 最大支持的上下文长度达 131k，意味着单轮对话最大显存占用高达 131G

推理过程大量 KV 可被复用





多
10 倍级上下文窗口扩展

快
TTFT 最高可降低 90%

省
Token 经济性提升 2 倍 +

10 CeaStor | 基于 AI 的 Lustre 缺陷与知识管理实践



找出2.15版本没有解决的内存泄漏问题，用表格展示

根据您提供的来源，以下是 Lustre 2.15 版本发行时**没有解决**的，或受 2.15 版本影响但修复版本为 2.16.0 或更高版本的内存泄漏、内存耗尽或关键内存错误问题。

需要注意的是，如果问题的修复版本（Fix Version/s）是 `Lustre 2.16.0` 或更高版本（例如 2.17.0），则意味着该问题在 2.15 版本中未得到解决，或者它是一个影响 2.15 版本并在后续补丁版本中解决的问题（但由于您要求的是“没有解决的”，我们通常关注那些在 2.15 周期内未被修复的问题）。

下表列出了满足这些条件的内存相关问题：

Key (编号)	Summary (问题摘要)	Affects Version/s (受影响版本)	Fix Version/s (修复版本)	Rationale (原因)	Source (来源)
LU-13594	register OOM callback in Lustre (在 Lustre 中注册 OOM 回调)	Lustre 2.14.0	None	这是一个改进项，旨在处理内存不足（OOM）情况 [1]。该问题的“修复版本”列为 None，但状态为 Resolved [1]，表示其不是作为 2.15 的修复内容完成。 [1]	
LU-14503	o2iblnd: clean up zombie connections on shutdown (o2iblnd: 在关机时清理僵尸连接)	N/A	Lustre 2.16.0	这是一个 Bug，其修复版本被设置为 Lustre 2.16.0 [2]。 [2]	
LU-15189	lnet: fix memory mapping (lnet: 修复内存映射)	N/A	Lustre 2.16.0	该问题的修复版本最初曾被标记为 Lustre 2.15.0，但随后被移除，并最终在 Lustre 2.16.0 中修复 [3]。 [3]	
LU-15427	Lustre client hangs under memory pressue (Lustre 客户端在内存压力下挂起)	Lustre 2.12.8	None	这是一个优先级为 **Critical** (关键) 的 Bug，状态为 Open (开放)，且 Fix Version/s 为 None (无) [4]。 [4]	
LU-15598	page leak on cfs_crypto_hash_init() error path in tgt_checksum_niobuf_t10pi (在 tgt_checksum_niobuf_t10pi 中 cfs_crypto_hash_init() 错误路径上的页面泄漏)	N/A	Lustre 2.16.0	这是一个页面泄漏 Bug，修复版本被设置为 Lustre 2.16.0 [5] [6]。 [5] [6]	
LU-15963	sanityn test_56b: OSS OOM with ZFS (sanityn test_56b: ZFS 导致 OSS OOM)	N/A	Lustre 2.17.0	该问题涉及 OSS 内存不足（OOM），尽管早期曾被考虑在 2.15.1 中修复，但最终的 Fix Version/s 为 Lustre 2.17.0 [7]。 [7]	
LU-16154	**memory leak** in class_config_llog_handler() (在 class_config_llog_handler() 中的内存泄漏)	N/A	Lustre 2.16.0	修复版本为 Lustre 2.16.0 [8] [9]。 [8] [9]	
LU-16575	UDSP **memory leak** / lnet: **memory leak** in copy_ioc_udsp_descr (UDSP 内存泄漏 / lnet: 在 copy_ioc_udsp_descr 中的内存泄漏)	N/A	Lustre 2.16.0	这是一个 Bug，修复版本为 Lustre 2.16.0 [10] [11]。 [10] [11]	

CONTENT

01 中国电子云的产品及应用

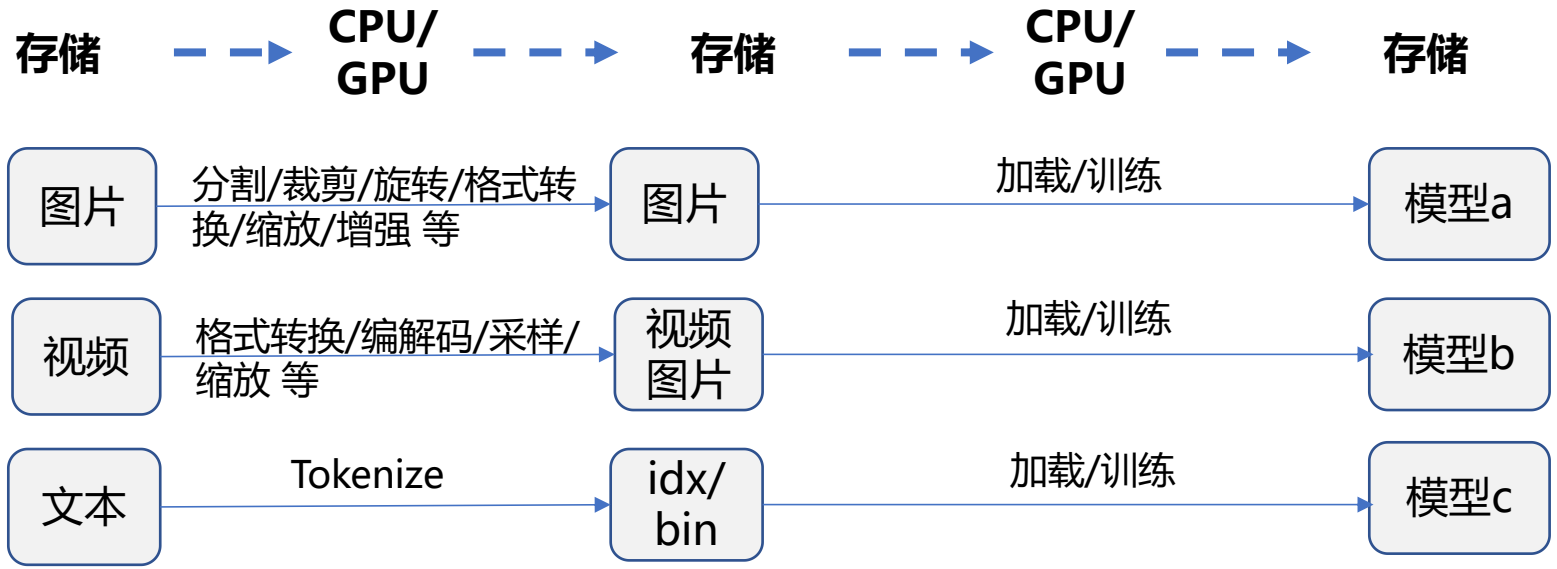
02 中国电子云在 Lustre 的探索和实践

03 国产软硬件平台协同适配与融合机制研究

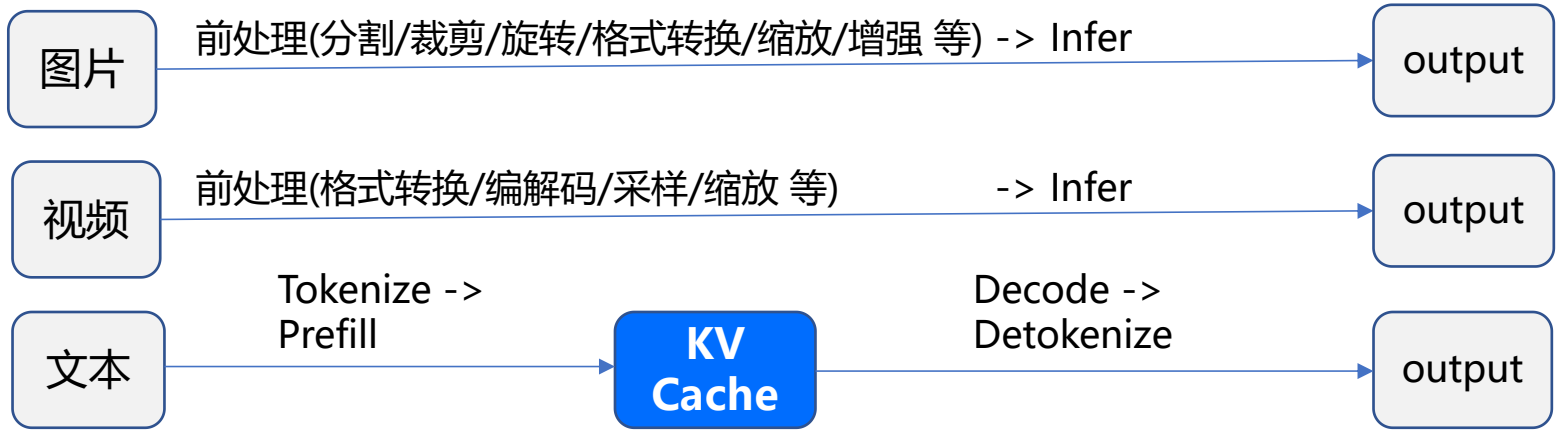
04 中国电子云信创 Lustre 开源生态建设倡议

11 过去一年，存储在 AI 场景中新的需求

- **训练**
  - 数据集预处理
  - 数据集加载



- **推理**
  - 数据集加载
  - **KVCache 读写**



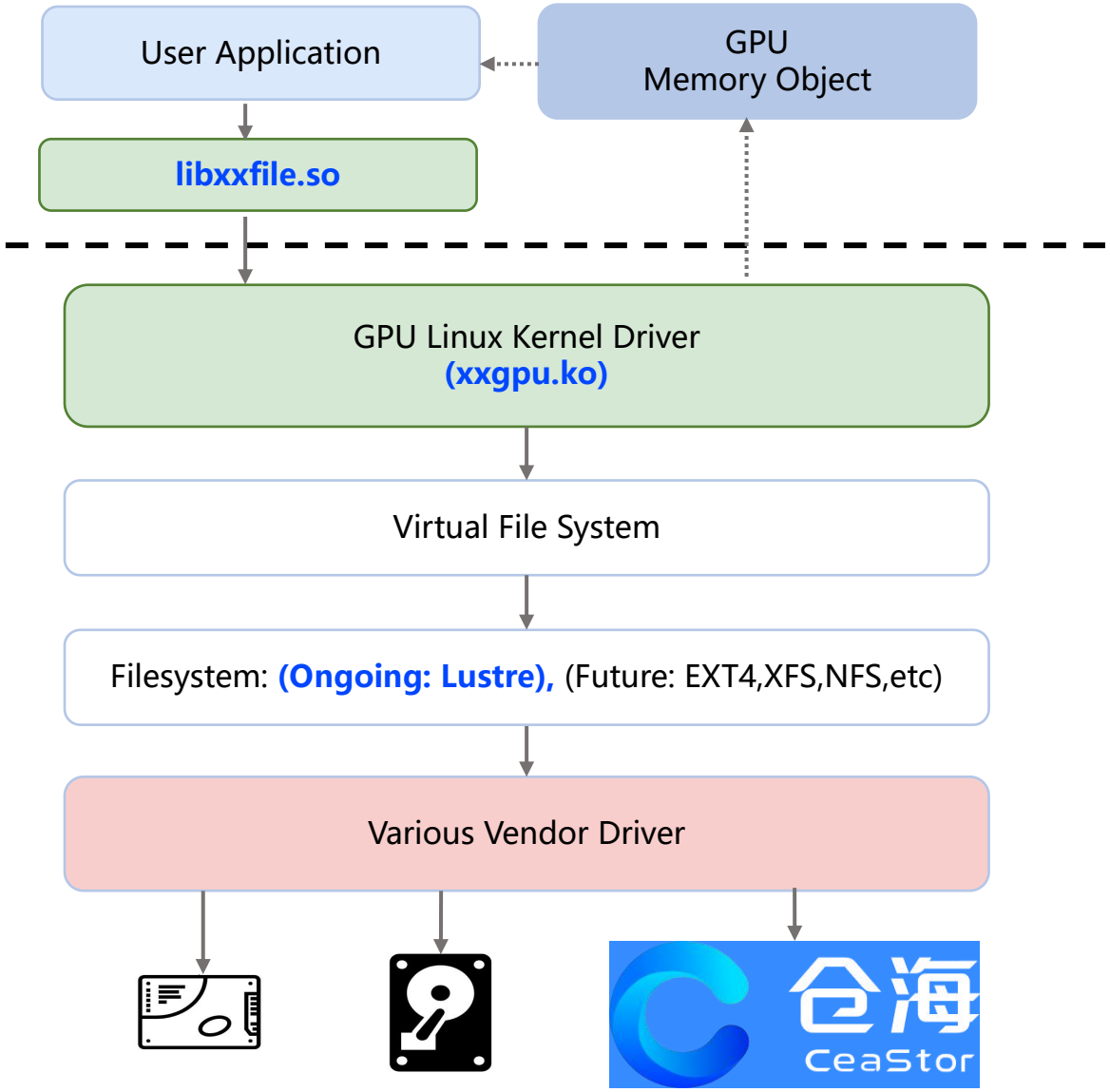
12 推理场景下，KVCache 对存储带宽提出更高要求

DeepSeek 实测统计数据，平均每台 H800 的吞吐量为：对于 Prefill 任务，输入吞吐约 73.7k Tokens/s（含缓存命中）；对于 Decode 任务，输出吞吐约 14.8k Tokens/s。56.3% 命中 KVCache 硬盘缓存。假设 KVCache Load/Save 耗时占端到端推理耗时 5% 。

类别	DeepSeek + H800	DeepSeek + 老一代国产 GPU
Prefill 并行	MLA TP1/PP1/DP32, MoE EP32	MLA TP8/PP4/DP1, MoE EP1
Decode 并行	MLA TP1/PP1/DP144, MoE EP144	MLA TP1/PP1/DP144, MoE EP144
数据类型	**FP8**	**FP16**
单token 单层 kv size	768 bytes (256 align)	1280 bytes (256 align)
KVCache Load 带宽 **Prefill**	73.7k*4*61*768*56.3%**/0.05=148.29 GB/s**	73.7k*4*61*1280*56.3%**/0.05=249.67 GB/s+**
KVCache Save 带宽 **Decode**	14.8K*18*61*768/0.05 = **238.04 GB/s**	14.8K*18*61*1280/0.05 = **396.74 GB/s**

单台 H800 带宽需求高达 40GB/s
单台国产 GPU 带宽需求高达 80GB/s+

13 最佳实践1：国产 GPU+Lustre GDS 的生态闭环构建

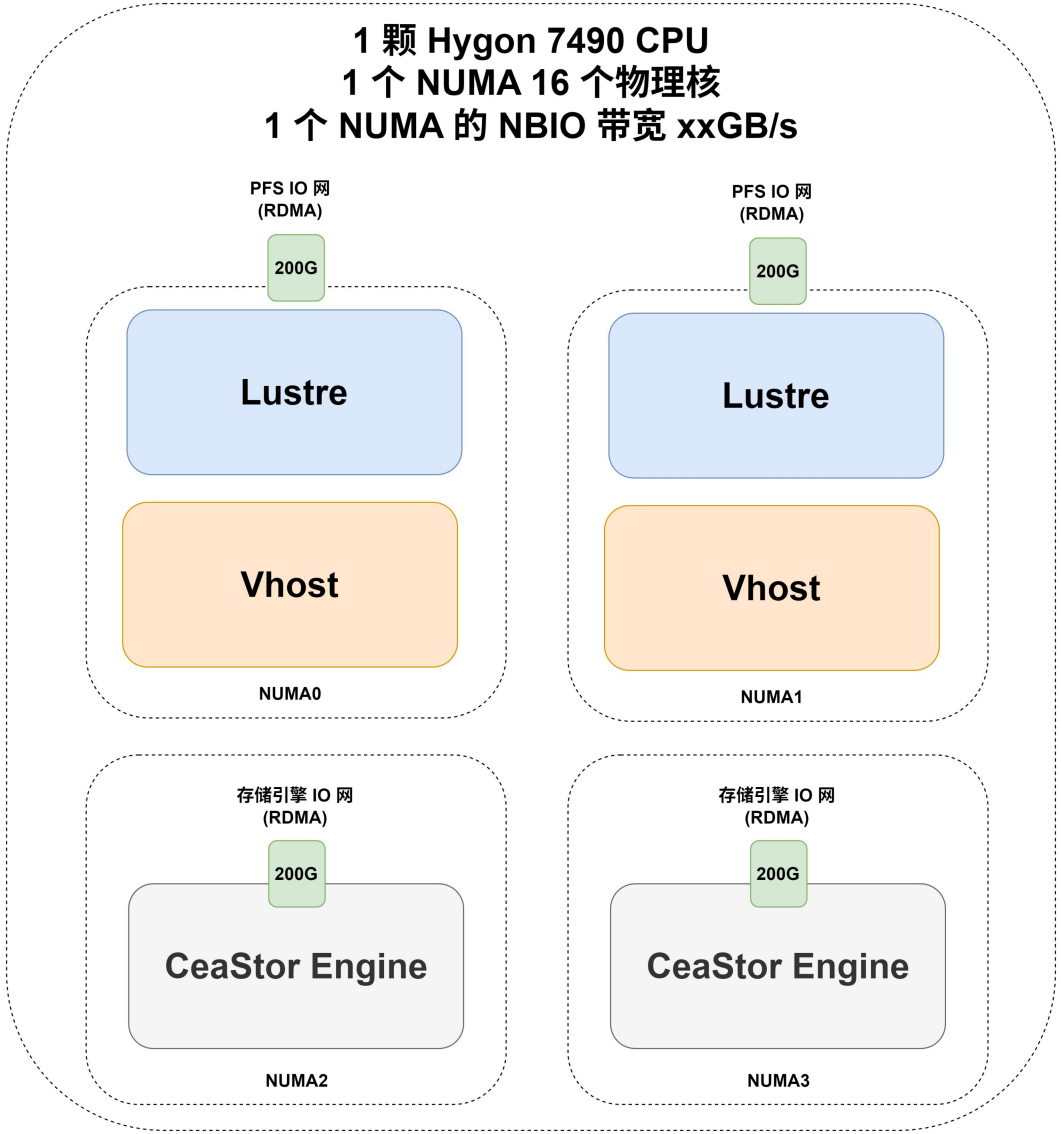


基于 ZONE_DEVICE 提供 *struct page mem_map* 服务

1. GPU 驱动加载时调用**devm_memremap_pages**() 将显存物理地址构建成 struct page。
2. GPU 内核驱动通过 struct bio_vec 将显存物理地址对应的struct pages 传递给 VFS。
3. VFS 将显存对应的 struct pages 传递给文件系统内核驱动。
4. 文件系统内核驱动使用 struct pages 对 DMA 引擎进行编程，最终 DMA 引擎完成数据搬移。

国产 GPU 的 GDS 带宽突破 150 GB/s+
全面释放国产 GPU 的 Lustre 生态潜能

14 最佳实践2：国产 CPU 海光平台的存储性能跃升



- **PCIE 链路优化：**针对海光 CPU 多 NUMA 架构下 PCIE 设备多路径传输的并行性限制，减少事务等待时间，提升数据传输并行效率。
- **软硬件协同优化：**通过软硬件适配调整，配合 HPT 工具分析带宽占用，优化代码减少无效数据访问，实现软硬件效能协同释放。
- **内存配置优化：**优化内存配置，最大化内存带宽，为高性能计算提供基础支撑。
- **编译优化：**基于二进制优化工具，通过插桩采样与布局重构，提升二进制文件在海光平台的运行效率。
- **NUMA 亲和：**CPU+内存+网卡+ NVME 盘使能 NUMA 亲和，降低 IO 时延和传输损耗。

阶梯式性能跃升，层层释放硬件潜能

- **客户端 1M 大小数据块读性能提升 2.5 倍：**客户端 1M 数据块读取带宽从 6.3GB/s 跃升至 15.9 GB/s，效率提升高达 152%。
- **写性能飞跃 10.4 倍：**服务器端 1M 数据块写入带宽更是从 1.4GB/s 升至14.6GB/s，实现 940% 的惊人提升。
- **实测单机 IOPS 性能提升至 2000K+。**
- **海光战略合作伙伴，SPC1 和 MLPerf Storage 打榜筹备中。**

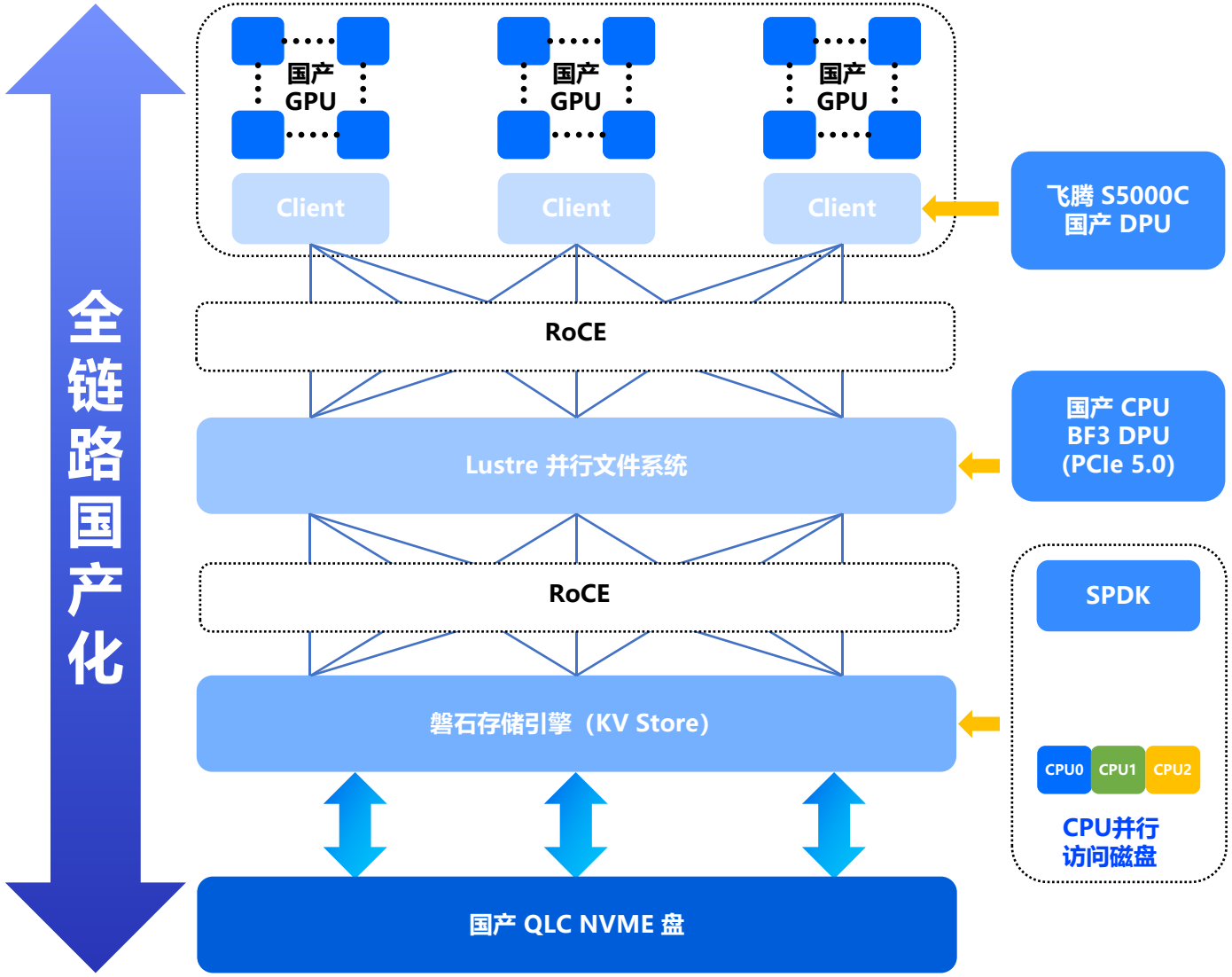
15 最佳实践3：国产 CPU 飞腾平台在 RDMA 场景下的适配挑战

飞腾 S5000C 在处理跨 NUMA 节点的 DMA-Write 事务时存在一致性缺陷：当目标内存位于与发起设备不同的 NUMA 域时，硬件无法保证数据写回与描述符更新之间的全局顺序（Write-Completion vs. Descriptor-Update Ordering）。该违例表现为 PCIe 写完成报文（WC）可能晚于描述符回写被 CPU 观测到，从而引发 CPU 在数据尚未落内存时即读取，导致“Stale-Data”类致命错误。

场景	场景具体描述	涉及的补丁模块
内核 DMA	TCP/UDP 读写 NVME 磁盘	Kernel
RDMA	RDMA 是程序（用户态）直接操作网卡来访问内存	用户态库：rdma-core 网络模块 大页管理模块
SPDK	SPDK 是程序（用户态）直接操作 NVME 盘来访问内存	SPDK 库

首家全栈适配飞腾 S5000C RDMA，
全链路时延较 TCP 骤降 50%

16 国家级工信部重点项目：分布式加速服务器



千卡 GPU 集群

自研 GPU 软件栈

TB 级 存储带宽

存储成本 下降 30%

17 可计算卸载 QLC 数据分层技术探索，大幅缩减建设成本

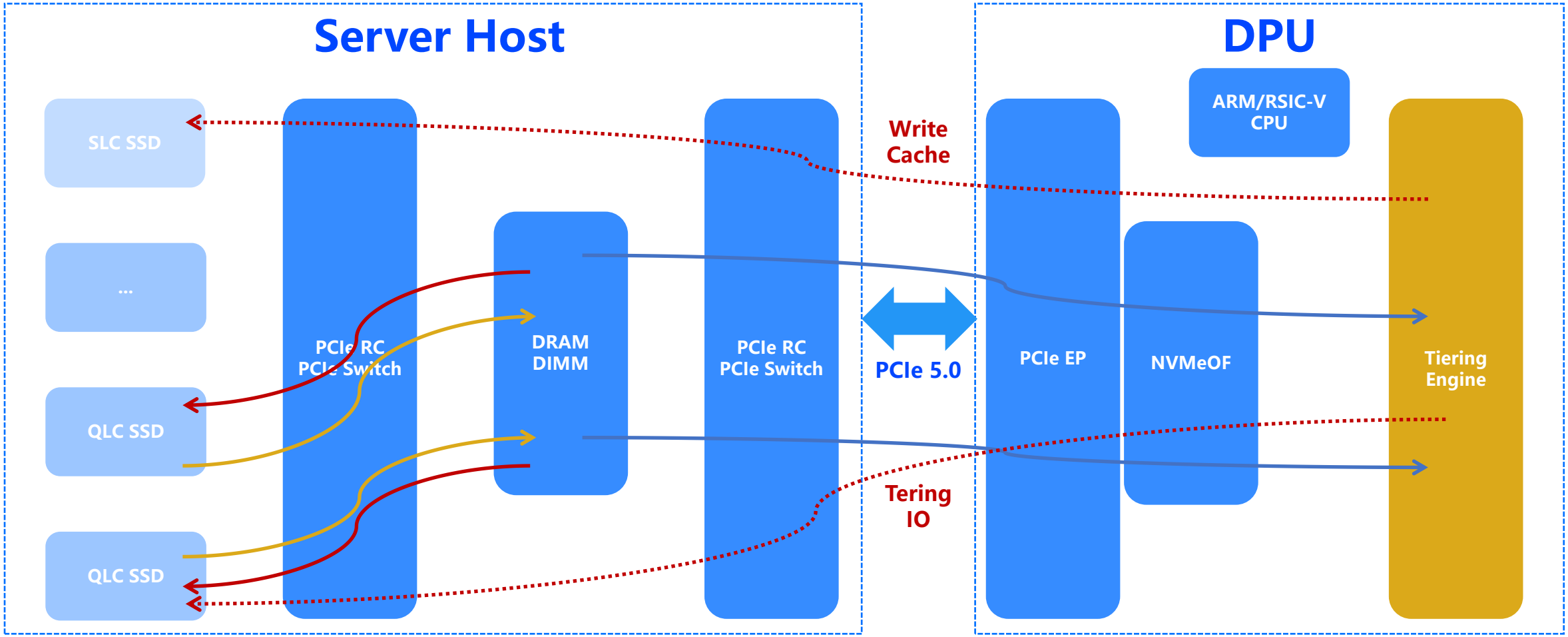
Note: command path (NVMe SQ/CQ/Interrupt) always pass through DPU

Write data for the Tering Engine process

Write Big IO: data from host to QLC SSD directly

Write Small IO: take use of PCIe peer to peer transfer

Read Data Path: data from SSD to host directly



CONTENT

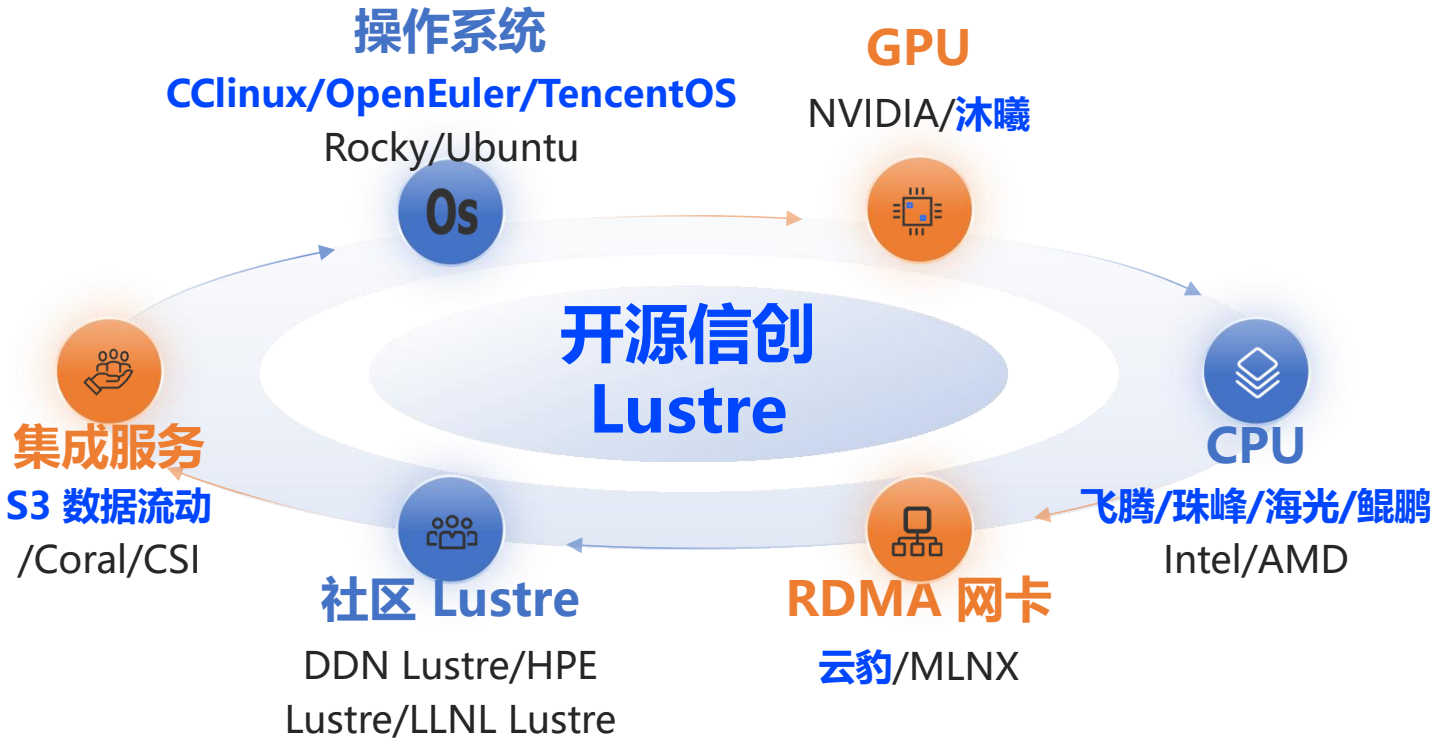
01 中国电子云的产品及应用

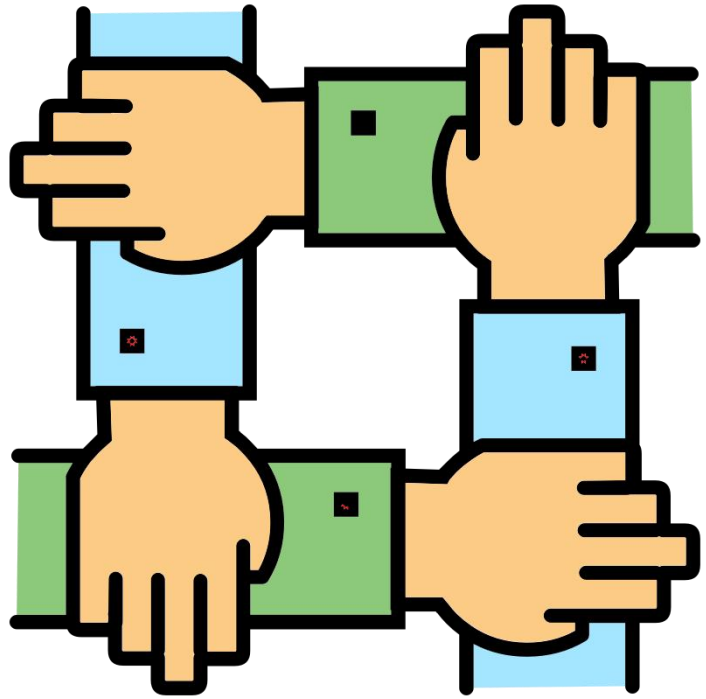
02 中国电子云在 Lustre 的探索和实践

03 国产软硬件平台协同适配与融合机制研究

04 中国电子云信创 Lustre 开源生态建设倡议

18 中国电子云信创 Lustre 开源生态建设倡议





联系我们

致力于打造全栈开源、自主可控的 Lustre 信创生态，交付企业级稳定与极致性能。我们诚邀全国团队共建、共享、共赢

CEC 中国电子 | 中国电子云

THANKS