AI时代云厂商仓皇的文件存储

Original 王太平 太平说存储 2024-06-03 07:20 四川

引言

在过去10年，云计算突飞猛进，已经成为不可忽视的基础设施平台，这一切源自于互联网厂商的狂奔，传统IT厂商难以望其项背。在那个时代有很多值得说的故事。当年为了去O,主要还是成本问题，传统oracle和EMC高端存储在这个场景成本高到极致。尤其是规模和性能越大，更是指数级增长。以国内为例，亿级的分布式应用的服务架构中间件，只有经过实践的云厂商有了解并且可以提供给创业企业使用，所以大数据、数仓、消息队列、分布式数据库，这一套针对互联网的海量业务中间件直接打造了云的腾飞。

过去十年间，云存储是如何发展的呢?大多数自研的云存储厂商，都可以追溯到大数据，阿里云为了支持自研的大数据系统替代Hadoop，发展了自己的盘古文件系统（主要还是为了给上层提供HDFS的能力）。而整套架构的思想参考了Google GFS、bigtable、HBase这样的组合。所以Google的三驾马车中的gfs带来了现代云计算存储的原旨教义。

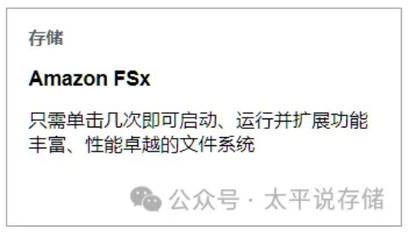
而AWS推出S3服务，则定义了互联网时代非结构化数据的新模式，serverless的云原生化服务，互联网可达的全球访问，形成了事实标准。在很长一段时间内，我们都可以看到，互联网云服务商对于NAS的态度是非常冷漠的。

随着云服务从互联网逐渐蔓延到不同的领域，很多传统存储的需求出现了，比如说企业上云。数据库上云已经被云上的rds替代，所以配套传统数据库的SAN存储需求很多都消失了，变成了云数据库的存储服务(客户不可见)，大多数由云厂商的分布式块存储服务提供。

很多传统的业务架构需要基于文件存储构建，对于windows、企业权限、多协议访问，带来了enterprise NAS的需求，这些生态适配的问题是传统IT厂商（windows、netapp、emc等）用几十年构建的，技术难度复杂（不是难，是复杂），投入大见效不一定大。所以在大多数云厂商内部对于enterprise NAS的需求都是忽视的，或者走最快捷的路线，那就是netapp on cloud等。

其次还有一个场景，就是HPC上云，对于HPC场景这个小众且狭窄的市场，一直以来云厂商投入并不大，对于HPC的存储需求，没有性能要求就使用云厂商自己的NAS服务，有性能需求就找第三方OEM。

AWS的文件存储服务



不数不知道一数一大跳，AWS提供了7款文件存储产品，看样子要集齐七龙珠。这里没有包含S3的网关产品、ossfs、mountpoint等可以通过转化S3到文件系统的产品。这里列出的全部是正式对外提供的独立存储服务。

Amazon FSx合作产品集

其实就是OEM产品组合，包含了四种广泛使用的文件系统之间进行选择：Lustre ， NetApp ONTAP ， OpenZFS 和 Windows File Server 。

但是好的是AWS提供了统一的Amazon FSx API：作为头部厂商，研发能力比较充足，AWS针对这四种存储产品研发了统一的Amazon FSx API，解决了客户使用过程中不需要关注底层服务差异化的问题，适配一次，只需要改变endpoint即可进行管理。

(1)Amazon FSx for NetApp ONTAP：高性能文件存储，它可可通过行业标准 NFS、SMB 和 iSCSI 协议从 Linux、Windows 和 macOS 计算实例广泛访问。可使用 ONTAP 的广泛采用的数据管理功能，例如快照、克隆和复制。支持压缩和数据去重。主要强调了针对企业NAS的场景提供和云下netapp一致的体验。

(2)Amazon FSx for Lustre：FSx for Lustre 使启动和运行流行的高性能 Lustre 文件系统变得简单且经济高效。您可以将 Lustre 用于速度至关重要的工作负载，例如机器学习、高性能计算 （HPC）、视频处理和财务建模。（人工智能+HPC场景）FSx for Lustre 符合 POSIX 标准，因此您可以使用当前基于 Linux 的应用程序，而无需进行任何更改。FSx for Lustre 提供本机文件系统接口，与任何文件系统在 Linux 操作系统上的工作方式一样。它还提供先写后读的一致性，并支持文件锁定。（强一致性，文件锁机制）。

(3)Amazon FSx for OpenZFS：基于开源 OpenZFS 文件系统构建的高度可靠、可扩展、性能和功能丰富的文件存储。它将这些功能与完全托管的 AWS 服务的敏捷性、可扩展性和简单性相结合。使用行业标准 NFS 协议（v3、v4.0、v4.1、v4.2）从 Linux、Windows 和 macOS 计算实例和容器广泛访问。强大的 OpenZFS 数据管理功能，包括数据压缩、近乎即时的时间点快照和数据克隆，专为与 Amazon FSx API 配合使用而设计，使您可以轻松地将本地文件服务器替换为 AWS 存储，从而提供熟悉的文件系统功能，并且无需执行冗长的资格认证以及更改或重新构建现有应用程序或工具。FSx for OpenZFS 使您能够构建和运行高性能、数据密集型应用程序。

(4)FSx for Windows File Server：Amazon FSx for Windows File Server 提供完全托管的 Microsoft Windows 文件服务器，由完全原生的 Windows 文件系统提供支持。FSx for Windows File Server 具有轻松将企业应用程序直接迁移到 AWS 云的功能、性能和兼容性。借助 Amazon FSx 上的文件存储，Windows 开发人员和管理员目前使用的代码、应用程序和工具可以继续保持不变地工作。适用于 Amazon FSx 的 Windows 应用程序和工作负载包括业务应用程序、主目录、Web 服务、内容管理、数据分析、软件构建设置和媒体处理工作负载。

(5)Amazon File Cache：Amazon File Cache 是 AWS 上完全托管的高速缓存，用于处理文件数据，无论数据存储在何处。Amazon File Cache 可作为存储在本地文件系统、AWS 文件系统和 Amazon S3存储桶中的数据的临时高性能存储。Amazon File Cache 将链接数据集中的数据呈现为一组统一的文件和目录。它以始终如一的高速为在 AWS 上运行的应用程序提供亚毫秒级延迟的缓存中的数据 - 高达数百 GB/s 的吞吐量和高达数百万的每秒操作，从而加快了工作负载完成时间并优化了计算资源消耗成本。Amazon File Cache 在首次访问数据时自动将数据加载到缓存中，并在未使用时释放数据。

Amazon EFS自研文件系统

其实就是AWS自研的文件系统，可以理解为一个通用的文件系统，这个文件系统的研发思路类似于S3，提供统一的服务以及分层的存储类。

(1)Amazon EFS 标准：Amazon EFS 标准存储类基于高速 SSD 存储构建，EFS 标准专为经常访问或修改且需要高耐久性和可用性的数据而设计。EFS 标准适用于各种使用案例，包括容器化和无服务器应用程序的存储、应用程序开发、机器学习训练、量化投资研究、Web 服务和内容管理、主目录和数据库备份。它还适用于数据分析、模拟和媒体转码等应用程序。（总结，全闪存NAS，为有性能需求的文件共享场景，或者性能较低的超算或者AI场景）

(2)Amazon EFS IA：EFS 不频繁访问存储类针对每季度仅访问几次且不需要 EFS 标准亚毫秒延迟的数据进行成本优化。与 EFS 标准相比，EFS IA 提供的存储价格最多可低 95%。（低频类型的EFS）

(3)Amazon EFS Archive：EFS Archive 存储类针对一年仅访问几次或更少且不需要 EFS 标准亚毫秒延迟的数据进行成本优化。与 EFS 不频繁访问相比，EFS Archive 的存储价格最多可降低 50%，从而为很少访问的冷数据提供更具成本优化的体验。

存储类	设计用于	首字节读取延迟	持久性（设计）[1]	可用性 SLA	可用区	每个文件的最低账单费用[2]	最低存储持续时间
EFS 标准	需要快速亚毫秒级延迟性能的活跃数据	亚毫秒级	99.999999999%（11 个 9）	99.99%（区域性） 99.9%（单区）	=>3（区域性） 1（单区）	不适用	不适用
EFS 不频繁访问	每季度仅访问几次的非活动数据	几十毫秒		99.99%（区域性） 99.9%（单区）	=>3（区域性） 1（单区）	128 KiB	不适用
EFS Archive	每年访问几次或更少次的非活动数据	几十毫秒		99.9%（区域性）	=>3（区域性）	128 KiB	90 天

AWS的文件存储布局分析

场景	产品推荐
通用文件共享	EFS的三种存储类
企业NAS	Amazon FSx for NetApp ONTAP
特定场景： windows、NFS	FSx for Windows File Server Amazon FSx for OpenZFS
HPC&ML	Amazon FSx for Lustre

可以看出，AWS自研的NAS只能满足通用的业务场景，对于企业NAS、windows&nfs特定生态场景、HPC&ML高性能场景都是无法满足的。因此引入了大量的第三方存储系统来满足业务需求。

AZure的文件存储服务

微软的文件存储服务主要分为三种：自研的Azure Files、oem netapp的Azure NetApp Files、为HPC&AI场景引入的Azure Managed Lustre。

Azure Files

Azure 文件共享可由 Windows、Linux 和 macOS 的云或本地部署同时装载。还可以使用 Azure 文件同步将 Azure 文件存储共享缓存在 Windows Server 上，以便在使用数据的位置附近进行快速访问。由于微软自己在windows环境上的技术储备，因此他可以满足通用文件共享以及windows环境的适配。

Azure NetApp Files

在云中运行性能密集型和延迟敏感型文件工作负载可能很困难。借助 Azure NetApp 文件，企业业务线 （LOB） 和存储专业人员可以轻松迁移和运行复杂的基于文件的应用程序，而无需更改代码。在各种方案中，Azure NetApp 文件被广泛用作基础共享文件存储服务。其中包括迁移（直接迁移）符合 POSIX 的 Linux 和 Windows 应用程序、SAP HANA、数据库、高性能计算 （HPC） 基础架构和应用程序以及企业 Web 应用程序。

Azure Managed Lustre

Azure 托管 Lustre 是适用于高性能计算 （HPC） 和 AI 工作负载的即用即付托管文件系统。通过专门构建的托管服务简化操作、降低设置成本并消除复杂的维护。与 Azure 服务（如 Azure HPC 计算、Azure Kubernetes 服务和 Azure 机器学习）配合使用。

GCP的文件存储

Filestore企业级文件存储

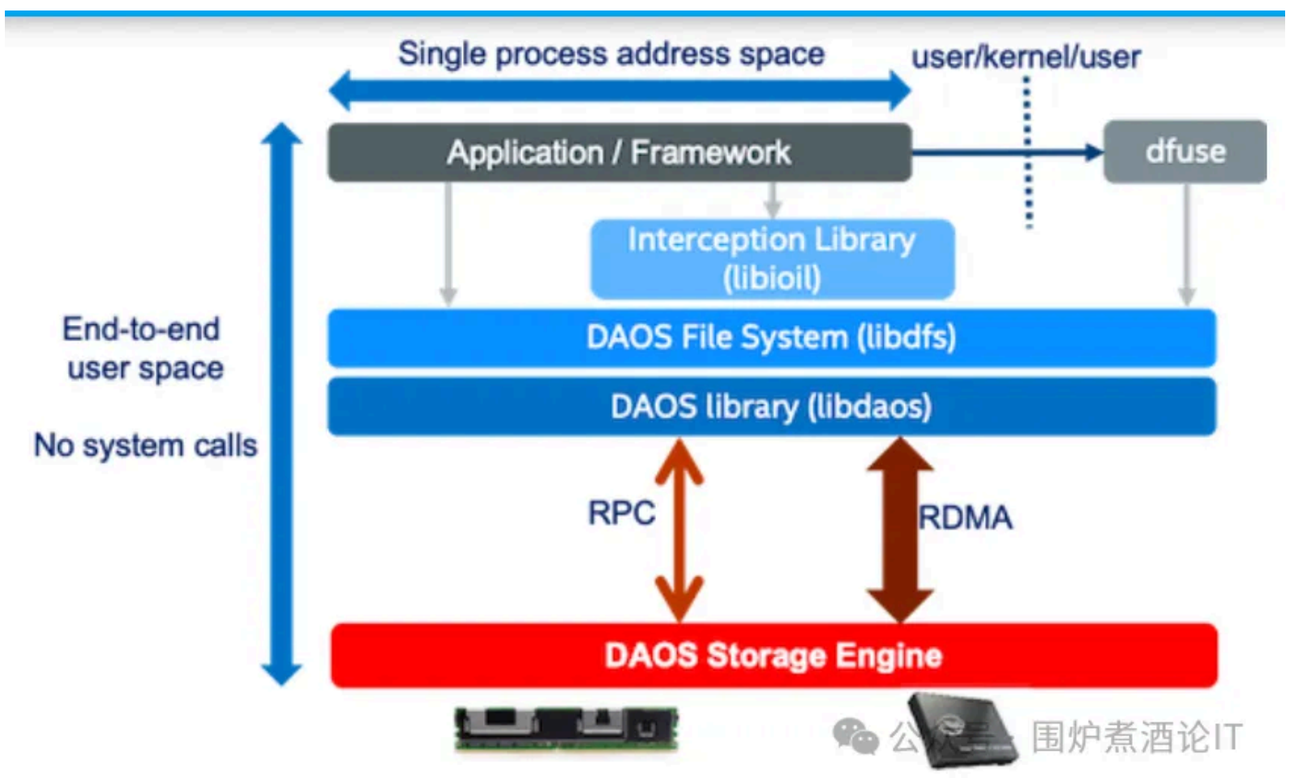
高性能、完全托管的文件存储，扩展以满足高性能工作负载的需求

Filestore 为应用程序提供低延迟存储操作。对于延迟敏感的工作负载，如高性能计算、数据分析或其他元数据密集型应用程序，Filestore 支持高达 100 TB 的容量、25 GB/s 的吞吐量和 920K IOPS。99.99% 区域可用性 SLA 支持企业应用

重点强调了SAP，Enterprise application migrations (SAP)，许多本地应用程序都需要文件系统接口。我们通过完全托管的存储服务，让您轻松地将企业应用程序迁移到云中。Filestore Enterprise 专为需要区域可用性和具有非结构化 NFS 数据要求的关键应用程序而构建。

Parallelstore并行文件存储

GCP今年推出的 Parallelstore 并行文件存储基于 Intel DAOS，与竞争对手的 Lustre 产品相比，读取吞吐量性能提高了 6.3 倍。



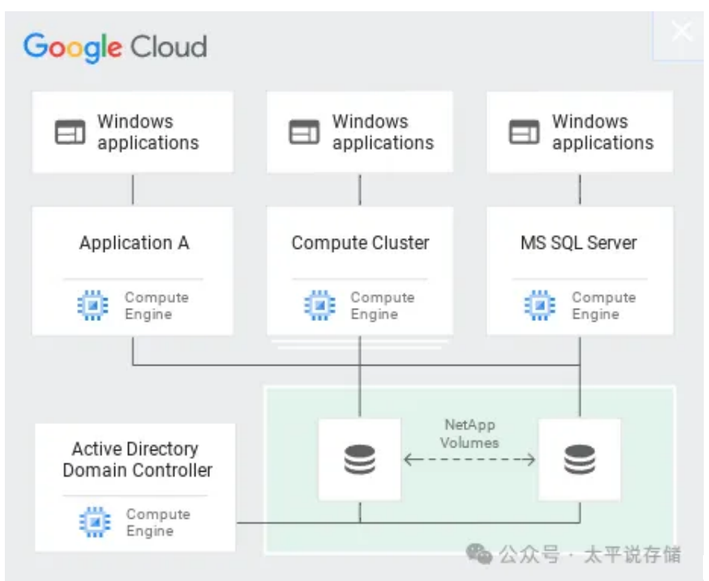
主打的就是AI/ML training 人工智能/机器学习训练场景：对于需要高速访问非常大的文件或数百万个小文件的系统，并行文件系统的性能可以优于 NFS 和对象存储。并行文件系统提供的延迟明显低于其他选项，这可能会影响最大 IOPS，这使得 Parallelstore 成为 AI/ML 工作负载暂存空间的绝佳选择。

量化交易分析：Parallelstore 非常适合在高性能定量分析和交易中提供中间数据的临时分析，也非常适合其他复杂、高速的金融服务用例，例如欺诈检测。
计算机辅助工程（CAE）：Parallelstore 是进行计算流体动力学 (CFD)、复杂建模、碰撞模拟、化学工程等的制造、汽车、航空航天和生物医学应用的不错选择。

Google Cloud NetApp Volumes

针对企业NAS的场景，还是免不了OEM netapp,基于文件系统的企业级功能还是需要很多年的积淀，云厂商的投入和企业NAS在云中的占比注定了不可能得到很多的投入。安全且高性能的文件存储，支持基于 SAP、Microsoft 和 Linux 的应用程序，并且无需重构或重新设计即可轻松迁移。

NetApp Volumes 支持 Windows 应用程序的数据共享，因此对于用户和组共享、非结构化数据的应用程序共享、SAP 共享文件、VDI 以及 MS-SQL 的共享存储非常有用。NetApp Volumes 支持 Linux 应用程序的数据共享，非常适合非结构化数据的应用程序共享；SAP共享文件：二进制文件、日志文件、配置文件；用户和组共享；共享机器学习数据；EDA——共享芯片设计数据；和 PACS 图像。



服务能力及价格

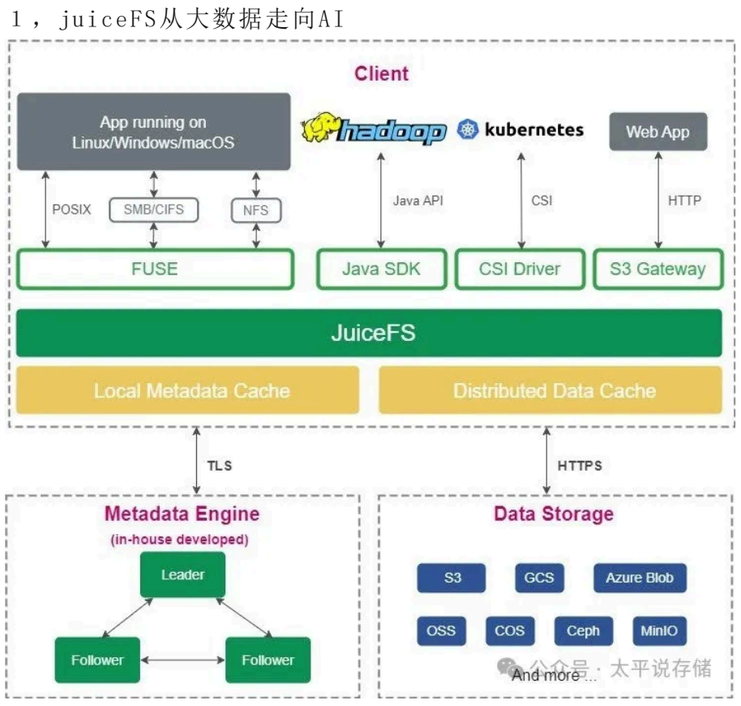
Service levels	Service level fit is evaluated by performance—as a combination of throughput, R/W mix, and latency.		
Service level	Standard	Premium	Extreme
Performance	16 MiB/sec per TiB (Throughput)	64 MiB/sec per TiB, Max 4.5 GiBps (Throughput)	128 MiB/sec per TiB, Max 4.5 GiBps (Throughput)
Price	$0.20/GiB	$0.29/GiB	$0.39/GiB
Volume replication	$0.11-$0.14/GiB (depending on RPO)	$0.11-$0.14/GiB (depending on RPO)	$0.11-$0.14/GiB (depending on RPO)
Regional availability	15 Google Cloud regions	15 Google Cloud regions	15 Google Cloud regions
Uptime / SLA	99.9%	99.95%	99.95%
Protocols	NFSv3/v4.1, SMB, Dual (SMB/NFS)	NFSv3/v4.1, SMB, Dual (SMB/NFS)	NFSv3/v4.1, SMB, Dual (SMB/NFS)

Cost for us-central1 region in USD.

云上文件存储的创业赛道

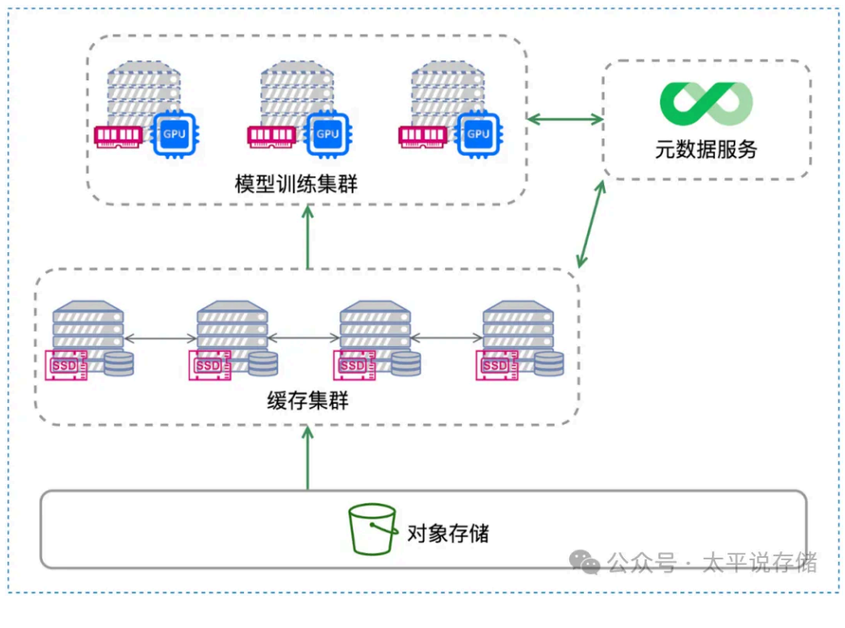
正因为各家云厂商NAS存储的拉跨表现，开启了各家文件系统创业厂商的云上春天。

1，juiceFS从大数据走向AI



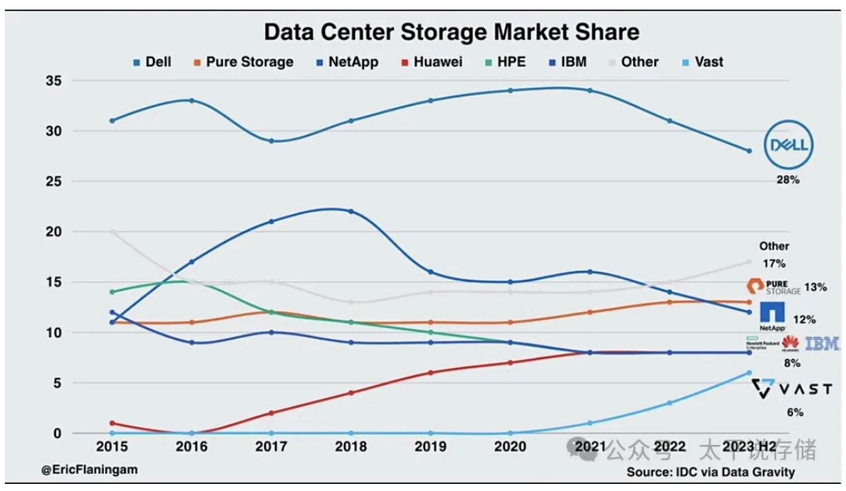
从去年开始，juicefs的宣传就从以前的大数据场景快速进化到了AI场景，主要宣传自己的FUSE文件系统posix能力。

其次，是构造一个计算端+存储服务端的整体缓存集群(社区版仅支持单节点缓存，而分布式缓存则需要企业版付费才支持)。这样可以给大多数厂商提供要给不被云厂商lock-in的架构。



2，VAST Data新架构

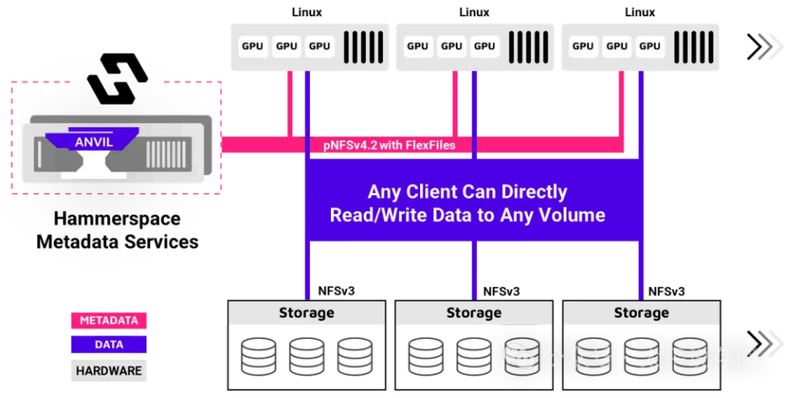
Vast Data 公司成立时间不长，但是已经迅速在存储市场形成了自己的优势，尤其是AI存储市场。通过使用QLC+SCM，并且构建了一个他们所称的通用存储（Universal Storage），构建了一个shareeverything的全新分布式存储架构，在当前的主要AI存储中占据了极大的份额。（之前有传言其全闪存存储的市场占比达到了6%）



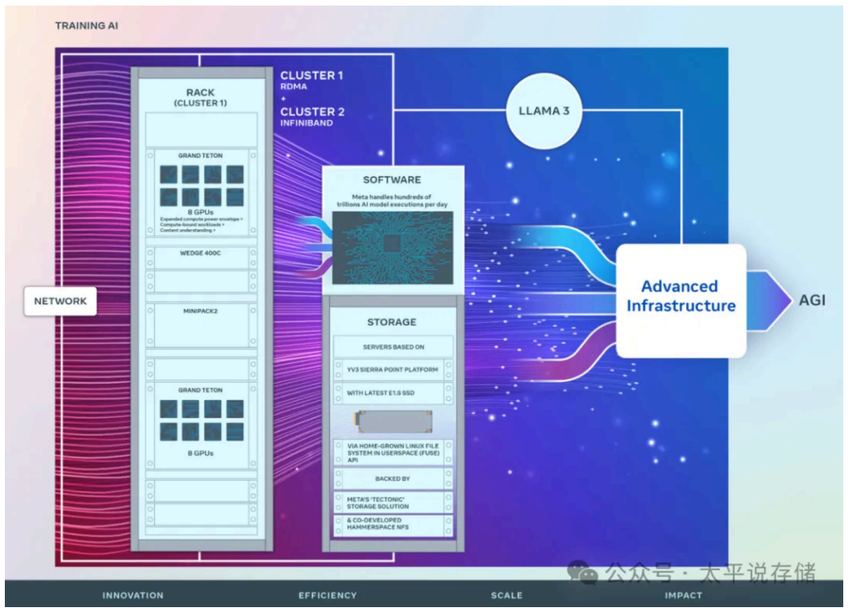
随着其业务的增加，vast data逐步向数据处理领域侵蚀，VAST 提供其QLC 闪存层并行、横向扩展基于文件的存储系统，并在此基础上构建了软件层：数据目录、全局命名空间、数据库和即将推出的数据引擎。

3，Hammerspace pNFS架构

Hammerspace文件系统提供标准的并行文件系统服务，并且主要承载元数据的服务，对于数据服务，可以广泛的利旧传统NFS V3存储，这和我们看到的juicefs的架构很类似。不过juicefs的fuse客户端性能要差很多。



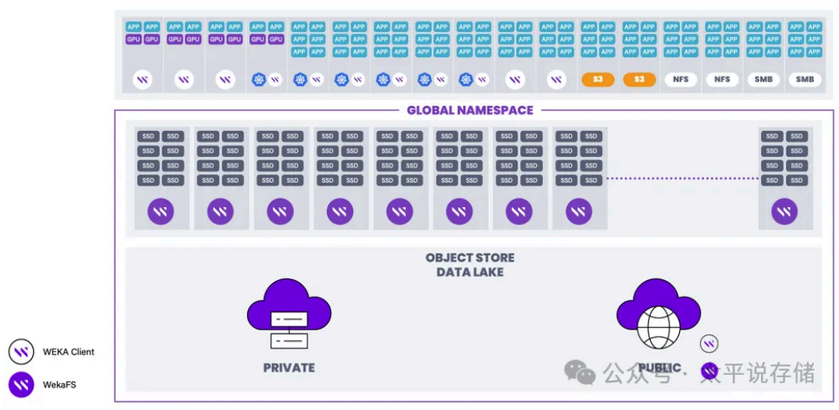
Meta还与Hammerspace合作，共同开发并落地并行网络文件系统（NFS）部署，以满足该AI集群的开发者体验要求。除了其他优势之外，Hammerspace 还使工程师能够使用数千个 GPU 对作业执行交互式调试，因为环境中的所有节点都可以立即访问代码更改。



4，WEKA文件系统

文件系统创业厂商weka最近融资了一笔1.4亿美元的融资，估值达到了16亿美金，是存储领域新的独角兽企业。实际上，这些现金并不需要用于支付任何业务扩张或产品开发费用，WEKA表示将增加其“可观的现金储备”，使其在扩大投资的同时扩大业务规模，同时扩大对开发数据平台软件的投资，并为WEKA员工提供流动性。所有其他轮次参与者都是现有投资者，其中包括英伟达。说明大部分投资者都非常看好weka，相当于通过这批投资给员工折现。

Weka当前增长非常迅速，在WEKA数据平台软件上运行了300多个世界上最大的GPU AI集群。



Weka良好的架构以及广泛适配性，给他带来了很多的合作伙伴，大多数服务器厂商都和weka有合作，帮助weka带货。Weka从传统线下的服务器到云上，已经逐步覆盖整个市场环境。它不仅克服了传统的存储扩展和文件共享限制，还允许通过 POSIX、NFS、SMB、S3 和 GPUDirect 存储进行并行文件访问。它提供了丰富的企业功能集，包括本地快照和云远程快照、克隆、自动分层、云爆发、动态集群重新平衡、私有云多租户、备份、加密、身份验证、密钥管理、用户组、带有建议的配额、软硬参数等等。

乍一看，上面没有什么差异化。其实不然，扩展性、统一存储、企业能力，这三者很难兼顾。这就是其优势。而且weka的posix支持不是用fuse，而是内核组件支持，WEKA 虚拟文件系统 （VFS） 内核驱动程序，它为应用程序提供 POSIX 文件系统接口。使用内核驱动程序提供的性能明显高于使用 FUSE 用户空间驱动程序所能实现的性能